Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des actuaires
le 17/11/2025

Par : **Maxence Colin**

Titre : **Création de scores de santé tenant compte des risques émergents aux Etats-Unis**

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

Nom : Caroline HILLAIRET

Membres présents du jury de l'Institut des Actuaires

Entreprise : Milliman
Signature :

MILLIMAN SAS
14, Avenue de la Grande Armée
75017 PARIS
Tél. : +33 1 42 99 15 60
SIREN : 501 636 534

Directeurs du mémoire en entreprise :
Nom : Eve Elisabeth TITON
Signature :

Nom : Marie GANON
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

En raison du réchauffement climatique, les températures vont continuer d'augmenter avec des périodes de chaleur intense et, dans certaines régions, des épisodes de froid extrême. Parallèlement, la fréquence et l'intensité d'événements naturels tels que les inondations ou les tempêtes seront amenées à croître dans les prochaines années. Ce constat s'applique également aux pics de pollution de l'air. Or, il est désormais établi que le dérèglement climatique a un impact significatif sur la santé des populations exposées. La prise en compte du climat et de la pollution représente donc un enjeu majeur pour les acteurs de l'assurance santé. Si la littérature est abondante sur leurs effets sanitaires, elle est en revanche très pauvre en ce qui concerne leur intégration dans les modèles de score de santé ou de tarification en assurance santé.

Ce mémoire vise donc à explorer cette problématique et à proposer des pistes d'intégration de ces paramètres dans l'évaluation du risque santé. L'objectif est de développer des scores de santé intégrant les risques émergents liés au changement climatique et à la pollution de l'air afin d'analyser leur impact sur la santé. L'analyse portera à la fois sur la pertinence des scores construits et sur la capacité de ces modèles à quantifier l'influence du climat et de la pollution sur la santé. Un point d'attention sera porté à l'évaluation de l'équité des modèles, notamment vis-à-vis des différentes origines ethniques, afin d'identifier d'éventuels biais et de garantir une utilisation responsable de ces outils en assurance santé.

La première partie de ce mémoire sera consacrée à la présentation du cadre d'étude : le système d'assurance santé aux Etats-Unis, les scores de santé existants, ainsi que les effets du climat et de la pollution sur la santé. Dans une deuxième partie seront présentés les modèles de scores de santé développés dans le cadre de ce travail. Enfin, dans la dernière partie, ces modèles seront implémentés sur la base de données de santé d'assureurs américains auxquelles seront intégrées des données climatiques et de pollution afin de quantifier leur influence sur la santé.

Mots-clés : Assurance santé, Score de santé, Etats-Unis, Risque climatique, Pollution, GLM, Régression linéaire, XGBoost, Biais, Interprétabilité, Valeurs SHAP.

Abstract

Due to global warming, temperatures are expected to continue rising, with periods of intense heat and episodes of extreme cold in some regions. At the same time, the frequency and intensity of natural events such as floods and storms are set to increase in the coming years. This also applies to peaks in air pollution. It is now well established that environmental factors have a significant impact on the health of exposed populations. Taking climate and pollution into account is therefore a major challenge for health insurance providers. While literature is abundant on their health effects, there is very little information on their integration into health scoring or health insurance pricing models.

This thesis therefore aims to explore this issue and propose ways of integrating these parameters into health risk assessment. The objective is to develop health scores that incorporate emerging risks linked to climate change and air pollution in order to analyze their impact on health. The analysis will focus both on the relevance of the scores constructed and on the ability of these models to quantify the influence of climate and pollution on health. Attention will be paid to assessing the fairness of the models, particularly with regard to different ethnic origins, in order to identify any biases and ensure the responsible use of these tools in health insurance.

The first part of this thesis will be devoted to presenting the study framework : the health insurance system in the United States, existing health scores, and the effects of climate and pollution on health. The second part will present the health score models developed as part of this work. Finally, in the last part, these models will be implemented on the basis of health data from US insurers, to which climate and pollution data will be integrated in order to quantify their influence on health.

Keywords : Health insurance, Health score, United States, Climate risk, Pollution, GLM, Linear regression, XGBoost, Bias, Interpretability, SHAP values.

Remerciements

Je souhaite exprimer toute ma gratitude à Alexandre Boumezoued, Principal du département Recherche et Développement chez Milliman, pour m'avoir permis de travailler sur ce sujet au sein du département qu'il dirige et pour le suivi de qualité tout au long de mon alternance.

Je remercie chaleureusement mes tutrices, Eve Elisabeth Titon et Marie Ganon, qui m'ont fait confiance en me permettant de travailler à leurs côtés. Leur encadrement bienveillant et précieux m'a guidé tout au long de l'élaboration de ce mémoire et m'a permis de faire avancer le projet dans une direction aussi intéressante que stimulante.

Je tiens également à remercier François Hu, Consultant responsable du pôle IA, pour sa disponibilité, sa bienveillance et ses nombreux conseils.

Un grand merci à l'ENSAE pour la qualité de ses enseignements, à l'ensemble du corps professoral ainsi qu'à mes deux tuteurs pédagogiques, Caroline Hillairet et Emmanuel Gobet, pour leurs relectures attentives et leur disponibilité.

Je remercie sincèrement l'ensemble de l'équipe R&D de Milliman ainsi que tous les membres du cabinet pour leur accueil chaleureux et l'aide qu'ils m'ont apportée sur certains aspects de ce mémoire.

Je n'oublie pas mes proches, amis et famille, et plus particulièrement mes parents qui m'ont sans cesse soutenu dans l'élaboration de mon projet professionnel. Je tiens à remercier tout spécialement Camille, dont le soutien constant, la patience et les encouragements ont été d'une aide précieuse tout au long de la rédaction de ce mémoire.

J'ai enfin une pensée toute particulière pour mes grands-parents, qui m'ont appris l'importance de l'effort, de la curiosité et de la bienveillance et dont les valeurs et le soutien m'accompagnent toujours. Ce mémoire est aussi le fruit de tout ce qu'ils m'ont transmis.

À la mémoire de mes grands-mères.

Table des matières

Introduction

En raison du réchauffement climatique, les températures moyennes à la surface du globe vont continuer d'augmenter avec des périodes de chaleur intense, mais aussi l'apparition d'épisodes de froid extrême dans certaines régions. Parallèlement, la fréquence et l'intensité d'événements naturels tels que les inondations, les ouragans, les tempêtes ou encore les sécheresses prolongées seront amenées à croître dans les prochaines années, comme le soulignent de nombreux rapports du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC). Ce constat s'applique également aux pics de pollution de l'air, notamment en milieu urbain, où la concentration de particules fines et d'ozone atteint régulièrement des niveaux excédant les seuils recommandés par les organismes de santé publique.

Or, il est désormais établi par la communauté scientifique que le dérèglement climatique et la pollution atmosphérique ont un impact significatif sur la santé des populations exposées. En effet, de nombreuses études épidémiologiques ont montré qu'une exposition prolongée aux particules fines peut favoriser le développement de maladies cardiovasculaires, pulmonaires (asthme ou encore bronchopneumopathie chronique obstructive), de certains cancers ou encore de troubles de la reproduction. Aussi, la chaleur est à l'origine d'une augmentation de la prévalence des maladies chroniques, des maladies rénales chroniques et des maladies cardiovasculaires, en particulier chez les populations les plus vulnérables telles que les personnes âgées, les enfants en bas âge ou encore les individus présentant des comorbidités. La prise en compte du climat et de la pollution par les acteurs de l'assurance santé et par les autorités publiques représente donc un enjeu majeur, d'autant plus que leurs effets ne sont pas répartis de manière équitable au sein de la population : certaines catégories sociales, souvent les plus défavorisées, sont disproportionnellement exposées aux risques climatiques et à la pollution atmosphérique. Cette exposition inégale aux risques émergents et à leurs impacts sur la santé peut conduire à des disparités en termes de couverture des soins de santé, en particulier aux Etats-Unis, pays dans lequel il n'existe pas de système de santé universel.

Si la littérature est abondante sur les effets sanitaires du climat et de la pollution sur la santé, elle est en revanche très pauvre en ce qui concerne leur intégration dans les modèles de scores de santé ou de tarification en assurance santé. Ce mémoire s'inscrit donc dans une optique de création de score de santé, à partir de données d'assureurs santé américains, intégrant le climat et la pollution, dans le but de quantifier leurs impacts sur la santé. Il présente plusieurs modèles statistiques et de *machine learning* (ML) permettant de créer des scores pertinents et d'intégrer les risques émergents. L'objectif est ensuite de les implémenter et d'analyser, dans un premier temps, la pertinence des scores construits : le score doit être fidèle à la réalité, c'est-à-dire qu'il doit refléter la santé réelle des individus. En second lieu, l'analyse portera sur la capacité de ces modèles

à quantifier l'influence du climat et de la pollution sur la santé : les modèles de score doivent être facilement interprétables pour pouvoir mesurer directement l'impact de certains indicateurs climatiques ou de pollution sur la santé. Une attention particulière sera également accordée à l'évaluation de l'équité des modèles, notamment vis-à-vis des différentes origines ethniques, afin d'identifier d'éventuels biais et de garantir une utilisation responsable de ces outils en assurance santé.

Ce mémoire s'articule en trois parties :

— La **première partie** fournit des éléments de contexte nécessaires à la compréhension du sujet. Elle débute par une présentation détaillée du système d'assurance santé américain, qui, malgré les réformes récentes, peine à couvrir équitablement l'ensemble des citoyens. Elle décrit également l'utilisation et l'encadrement des scores de santé aux Etats-Unis, en insistant sur leur fonctionnement, leur encadrement réglementaire et les débats éthiques qu'ils suscitent. Cette partie s'achève par une revue de littérature approfondie des effets du climat et de la pollution sur la santé des populations exposées.

— La **deuxième partie** de ce mémoire est consacrée à la description des modèles permettant de créer des scores de santé intégrant explicitement des variables climatiques et de pollution. Elle détaille leur principe, leurs avantages et leurs limites. Les modèles retenus, tels que la régression linéaire, les modèles linéaires généralisés et XGBoost, satisfont le critère d'interprétabilité des impacts sous-jacents. Un volet spécifique est dédié dans cette partie aux méthodes qui permettront d'analyser l'équité des modèles développés.

— La **troisième partie** complète le mémoire avec la mise en œuvre concrète des modèles présentés dans la partie précédente. A partir d'une base de données de santé d'assureurs américains, enrichie de données climatiques et de pollution, cette partie s'attache à implémenter les différents scores de santé et à comparer leurs performances. Elle analyse la capacité de ces modèles à satisfaire à la fois les critères de performance et d'interprétabilité. Le cas échéant, une analyse fine des effets des indicateurs climatiques et de pollution sur les scores de santé développés est réalisée. Enfin, une attention particulière est portée à l'évaluation de l'équité des scores produits permettant d'identifier d'éventuels biais dans la modélisation.

Première partie

Contexte : assurance santé et risques émergents aux Etats-Unis

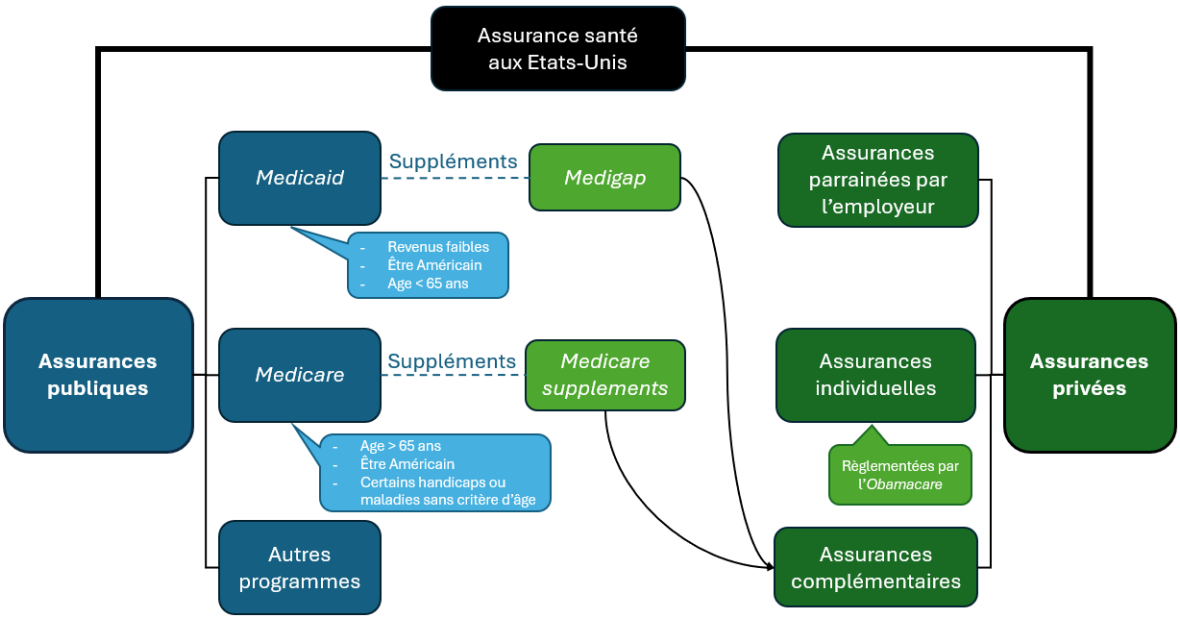
Chapitre 1

Assurance santé et données de santé aux Etats-Unis

1.1 Le système de santé américain

1.1.1 Le fonctionnement du système de santé américain

FIGURE 1.1 – Schéma du fonctionnement du système de santé américain



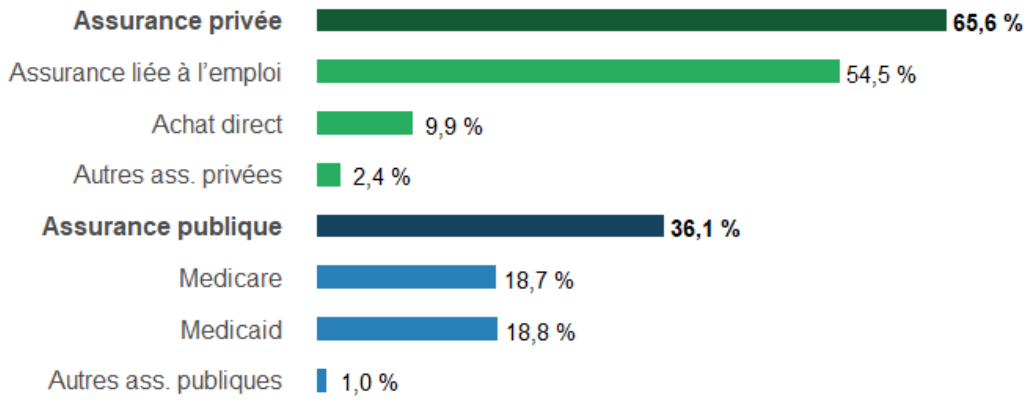
Aux Etats-Unis, le système de santé n'offre pas de couverture universelle obligatoire, contrairement au système français. En effet, ce sont les assurances privées qui couvrent les frais de santé de la majorité des Américains. Ces assurances sont principalement souscrites par l'intermédiaire des employeurs ou directement par les citoyens américains. L'*Affordable Care Act*[1] (ACA), appelé *Obamacare*, a notamment introduit à partir de 2014 une incitation exigeant les employeurs de plus de 50 salariés à fournir une couverture santé à au moins 95 % de leurs employés à temps plein sous peine d'amende

1. L'*Affordable Care Act* est une loi de réforme de la santé proposée par Barack Obama, adoptée aux Etats-Unis en 2010, visant à élargir l'accès à l'assurance santé, à réduire les coûts des soins de santé et à améliorer la qualité des soins

(247,5 dollars par mois par salarié non-couvert en 2024). Le coût à la charge du salarié ne doit pas dépasser un certain pourcentage de ses revenus (9,02 % en 2025 [1]). Les assurances privées coexistent avec des systèmes de couverture médicale financés par l'Etat, comme *Medicaid* et *Medicare*, destinés uniquement à une partie de la population en fonction de certaines caractéristiques (âge, revenus, invalidités). Malgré ces dispositifs, certains Américains restent sans couverture santé (7,9 % en 2022 [2]). La figure 1.1 schématise le système de santé américain, qui est décrit plus en détail dans les parties suivantes.

En 2022, 92,1 % des citoyens américains possédaient une assurance santé. 65,6 % d'entre eux étaient couverts par une assurance privée et 36,1 % par une assurance publique (voir figure 1.2). Il est possible de cumuler plusieurs assurances santé pour réduire le reste à charge pouvant être élevé avec un seul type d'assurance (voir figure 1.1).

FIGURE 1.2 – Type de couverture des Américains assurés en 2022 [2]



Note de lecture : 54,5 % des Américains assurés possèdent une assurance santé via leur employeur.

1.1.1.1 Les assurances privées

En 2022, comme l'indique la figure 1.2, 65,6 % des Américains assurés possèdent une assurance santé privée. Selon le recensement américain, environ 83 % d'entre eux sont couverts par un régime d'assurance maladie parrainé par l'employeur et 15 % souscrivent sans intermédiaire une assurance privée couvrant leurs frais de santé [2].

Pour les Américains assurés via leur employeur, il existe deux structures utilisées pour fournir une couverture :

— le régime auto-assuré (*self-insured plan* en anglais) : plan de santé réglementé par le gouvernement fédéral et généralement proposé par les grandes entreprises, dans lequel l'employeur collecte les cotisations des employés par le biais de retenues salariales et assume la responsabilité du paiement de toutes les demandes de remboursement de frais médicaux. L'entreprise fait alors appel à un tiers (compagnie d'assurance) pour gérer les services tels que l'inscription, le traitement des demandes de remboursement et les réseaux de prestataires ;

— le régime entièrement assuré (*fully-insured employer plan*) : régime de santé acheté par un employeur auprès d'une compagnie d'assurance. La compagnie d'assurance

se charge de payer les frais médicaux des souscripteurs en échange d'une prime versée par l'employeur.

Les contrats d'assurance maladie entre employés et employeurs se font majoritairement par l'intermédiaire d'organismes de gestion de soins (*Managed Care Organizations*, MCO). Les MCO sont des entités intégrées au système de santé américain, dont le but est de limiter les frais de santé tout en garantissant la qualité des soins. Ces organismes mettent l'accent sur la prévention. Les soins préventifs tels que les examens annuels, certains vaccins et les dépistages de routine sont totalement pris en charge pour inciter les affiliés à se soigner et réduire leurs frais de santé. Les MCO existent sous plusieurs formes :

— *Health Maintenance Organizations* (HMO) : ces programmes donnent à leurs assurés un accès à un réseau de prestataires spécifiques et proposent des formules prépayées. Les assurés ont l'obligation d'avoir un médecin traitant (PCP pour *primary care provider*) qui les oriente vers les spécialistes du réseau.

— *Preferred Provider Organizations* (PPO) : l'assuré a le choix entre des prestataires du réseau et des prestataires externes, mais aura des remboursements plus attractifs s'il fait appel à un personnel du réseau de fournisseurs du programme. Ces programmes n'obligent pas à choisir un PCP. Ils sont généralement plus onéreux que les HMO, car plus flexibles.

— *Exclusive Provider Organizations* (EPO) : ces organisations donnent accès à un réseau de fournisseurs spécifiques sauf en cas d'extrême urgence. Une recommandation par un PCP n'est pas nécessaire pour consulter un spécialiste, mais une autorisation est requise pour les grosses dépenses (étant donné que le médecin traitant n'est pas obligatoire). De plus, tous les frais engagés en dehors du réseau ne sont pas couverts. Le coût du programme est similaire à celui du HMO, car le fait d'avoir un réseau de soins fermé impose plus de contraintes aux assurés et à leurs parcours de soins.

— *Point of Service* (POS) : ce dispositif ressemble aux HMO mais est moins restrictif car il offre la possibilité de se soigner en dehors du réseau. Les POS requièrent d'avoir un médecin traitant pour toute initialisation de parcours de soins.

— *High-Deductible Health Plans* (HDHPs) : régimes d'assurance qui offrent des primes moins élevées, mais dont la franchise est plus grande. Les programmes HMO, PPO, EPO et POS sont considérés comme des HDHPs si la franchise qu'ils proposent dépasse une certaine limite. Les HDHPs peuvent être jumelés à des comptes d'épargne santé (*Health Saving Account* ou *Health Reimbursement Account*). Ces comptes ont pour fonction de permettre aux employés d'économiser de l'argent avant impôt pour les frais médicaux.

Les régimes de soins qui ne passent pas par l'intermédiaire d'organismes de gestion de soins (MCO) sont appelés régimes d'indemnisation (*indemnity plans*). Il s'agit de régimes de santé qui n'ont pas de réseaux de prestataires et qui remboursent simplement une partie des frais engagés pour tout service médical couvert. Les régimes d'indemnisation sont peu courants : en 2024, moins de 1 % des salariés américains bénéficiant d'une assurance maladie parrainée par leur employeur possèdent un régime d'indemnisation [3].

1.1.1.2 Les programmes d'assurances publics

Les Etats-Unis disposent de plusieurs programmes d'assurance publics gérés par le centre des services *Medicare* et *Medicaid* (*Centers for Medicare & Medicaid Services*, CMS)

et financés principalement par les impôts. Ces programmes sont uniquement dédiés à une partie de la population américaine respectant des critères spécifiques.

Medicare

Instauré en 1965 par Lyndon B. Johnson, le programme *Medicare* est destiné à toute personne âgée de plus de 65 ans et aux personnes de moins de 65 ans souffrant d'un handicap permanent ou d'une insuffisance rénale terminale. Il est financé par les taxes fédérales, les taxes payées par les employeurs ainsi que par les mensualités de ses bénéficiaires. Le programme est constitué de plusieurs parties :

— la partie A couvre les dépenses liées à une hospitalisation ;

— la partie B, facultative, couvre les visites chez le médecin (hors spécialiste) et est financée à hauteur de 25 % par l'assuré et 75 % par l'Etat fédéral ;

— la partie C, aussi connue sous le nom de *Medicare Supplement Plans*, est une couverture additionnelle à celle des parties A et B proposée par des assurances privées sous contrat avec le programme ; et

— la partie D est une couverture additionnelle à celle des parties A et B et couvre les coûts liés aux médicaments sur ordonnance (plafond de dépenses recouvrables à 2 400 dollars par an avec ticket modérateur de 25 % et au-dessus de 5 400 dollars ticket de 5 %).

En 2022, 18,7 % des Américains ont été couverts par *Medicare* (voir figure 1.2). Le coût annuel du programme est estimé à environ 944 milliards de dollars, soit 21 % des dépenses de santé nationales [4].

Medicaid

Instauré en 1965 dans le cadre de la lutte contre la pauvreté et géré par chacun des Etats, le programme *Medicaid* est destiné aux individus possédant de faibles revenus. Le seuil est différent selon les Etats. Le programme est financé par les Etats et par l'Etat fédéral qui accorde une subvention à chaque Etat en fonction de sa richesse propre. L'ACA a élargi l'éligibilité à *Medicaid* aux adultes de moins de 65 ans dont les revenus ne dépassent pas 138 % du seuil de pauvreté fédéral, afin de couvrir davantage d'Américains à faibles revenus. Cependant, les Etats ne sont pas dans l'obligation d'appliquer cette extension. A date, 10 Etats ne l'ont pas adoptée : l'Alabama, la Caroline du Sud, la Floride, la Géorgie, le Kansas, le Mississippi, le Tennessee, le Texas, le Wyoming et le Wisconsin.

Les Etats sont dans l'obligation (par eux-mêmes ou via des assureurs privés) de couvrir les dépenses liées à des consultations de médecine générale, aux hospitalisations, aux soins préventifs ou aux services liés à la santé mentale. Ils sont libres d'ajouter ou non d'autres actes de santé. En 2022, 18,8 % des Américains ont été couverts par *Medicaid* (voir figure 1.2). Parmi les bénéficiaires du programme en 2021, 38 % sont des enfants, 52 % des adultes à bas revenus âgés d'au plus 64 ans et 10 % des personnes âgées de plus de 65 ans. Par ailleurs, il est possible d'être couvert à la fois par *Medicare* et *Medicaid* (double éligibilité). Il est aussi possible de souscrire une assurance complémentaire privée, appelée *Medigap*, destinée à couvrir les frais non pris en charge par *Medicare*, comme les franchises, copaiements et tickets modérateurs. Le coût annuel du programme est estimé

à 806 milliards de dollars, soit 18 % des dépenses de santé nationales [4].

Children's Health Insurance Program

Financé à la fois par l'Etat fédéral et les Etats fédérés, le *Children's Health Insurance Program* (CHIP) a pour but de couvrir les dépenses de santé des enfants dont les parents ont des revenus trop importants pour bénéficier du programme *Medicaid* mais n'ayant toutefois pas les moyens suffisants pour souscrire une assurance privée. Une couverture de base, couvrant les visites chez le médecin ainsi que les médicaments sur ordonnance, doit être proposée dans tous les Etats. Cependant, une participation aux frais peut être demandée pour tout autre service plus spécifique. Aussi, certains Etats demandent une prime mensuelle ne pouvant excéder 5 % des revenus annuels de la famille pour la couverture CHIP.

Assurance pour les militaires

Des prestations de santé sont fournies aux militaires actifs, aux militaires retraités et à leurs proches par le système de santé militaire du ministère de la Défense (*Department of Defense Military Health System*) via deux réseaux : *TRICARE* et *Military Treatment Facilities*. Les anciens combattants sont également couverts par l'administration de la santé des anciens combattants (*Veterans Health Administration*).

1.1.1.3 Exemple d'une souscription

Un individu qui souscrit une assurance santé peut supporter plusieurs coûts incluant :

— les primes mensuelles ;

— les franchises : montants en dessous desquels l'assuré paie la totalité des coûts ;

— les copaiements ou la co-assurance : respectivement la somme ou la part fixe dont l'assuré doit s'acquitter après un acte de santé ;

— le plafond annuel de dépense : au-delà de ce plafond, la compagnie d'assurance prend en charge l'intégralité des frais de santé jusqu'à la fin de l'année. Les dispositifs de co-assurance et de copaiements ne s'appliquent donc plus.

1.1.1.4 Processus de déclaration des sinistres de santé : les *claims*

Les données de sinistres (*claims*) constituent une source essentielle d'information pour l'analyse des services de santé. Ces données sont générées à chaque interaction entre un patient et le système de santé, qu'il s'agisse d'une consultation médicale, d'une hospitalisation, d'une prescription médicamenteuse ou encore d'un acte de diagnostic. Le processus de création d'un *claim* repose sur plusieurs entités clés qui interagissent pour assurer la transmission, la validation et l'archivage des informations :

— **les fournisseurs de soins (*providers*)** : ce sont les hôpitaux, médecins, laboratoires, pharmacies et autres établissements de santé qui fournissent des services aux patients. Ils sont responsables de la création initiale des *claims* ;

— **les assureurs santé** : ils financent tout ou partie des soins de santé en fonction des contrats souscrits par les assurés ;

— **les *clearing houses*** : ces intermédiaires vérifient la conformité des *claims* avant de les transmettre aux assureurs. Ils s'assurent que toutes les informations requises sont bien renseignées ;

— **les entreprises de gestion du cycle des revenus** (RCM - *Revenue Cycle Management*) : elles assistent les fournisseurs de soins dans la gestion administrative et financière des *claims*, notamment en codifiant les actes médicaux et en assurant le suivi des paiements.

De la déclaration d'un sinistre, à la validation ou non de son paiement, le parcours d'un *claim* aux Etats-Unis suit plusieurs étapes pouvant parfois durer jusqu'à quelques semaines :

1. **Un acte médical est réalisé** : un patient reçoit les services d'un fournisseur de soin (*provider*).
2. **Création du *claim*** : le *provider*, ou une entreprise de gestion RCM, génère un *claim* sous un format standardisé. Ce fichier contient des informations détaillées sur l'acte médical, le patient et le coût du service.
3. **Vérification par un *clearing house*** : avant d'être envoyé à l'assureur, le *claim* passe par un *clearing house* qui s'assure qu'il est correctement rempli et conforme aux exigences de l'assureur.
4. **Traitement par l'assureur** : l'assureur examine le sinistre à travers un processus appelé *adjudication*, où il vérifie si :
   — le patient était bien couvert au moment des soins ;
   — l'acte médical est pris en charge par le contrat d'assurance ; et
   — des autorisations préalables étaient nécessaires.
5. **Décision et paiement** : l'assureur accepte, ajuste ou refuse le *claim*. Les paiements sont ensuite envoyés aux prestataires via une transaction électronique standardisée.
6. **Correction et révision** : si un *claim* est refusé (*denied*), il peut être corrigé et soumis à nouveau. Certains *claims* sont ajustés ou annulés (*reversed*) après paiement en cas d'erreur.
7. **Stockage des données** : une fois traités, tous les *claims* sont stockés dans des bases de données massives, comme celles utilisées dans le cadre de ce mémoire et décrites en partie 1.2.

1.1.2 Compléments sur l'*Affordable Care Act*

Le *Patient Protection and Affordable Care Act* (ACA), appelé aussi *Obamacare*, est une loi votée par Barack Obama le 23 mars 2010, entrée en vigueur le 1er octobre 2013 et mise en application en janvier 2014. Cette loi tend à créer, au niveau national, un système d'assurance santé universelle en visant à rendre accessible l'assurance santé à tous les Américains, sans pour autant imposer une couverture santé obligatoire, à l'exception des enfants.

La réforme du système de santé proposée par Barack Obama comprend de nombreuses mesures incitant les Américains à se couvrir :

— Au niveau national, il est interdit d'utiliser l'âge, le sexe, le lieu de résidence et l'état de santé de l'assuré lors de la signature du contrat d'assurance pour fixer son coût et son niveau de protection.

— L'éligibilité au programme *Medicaid* peut être étendue, comme mentionné dans la partie précédente. Cependant, le choix est laissé aux Etats d'appliquer ou non cette mesure.

— Les augmentations de primes sont encadrées et les résiliations de police individuelle sont interdites, sauf en cas de fraude.

— Les entreprises de plus de 50 employés sont incitées à proposer une couverture santé à leurs employés.

— Des subventions sont mises en place pour les individus se situant entre 100 % et 400 % du niveau de pauvreté fédéral.

Une plateforme en ligne, appelée *Health Insurance Marketplace* a été créée pour les individus et les familles n'ayant pas accès à des régimes d'assurance sponsorisés par leur employeur. C'est notamment le cas des chômeurs, des travailleurs indépendants et des salariés d'entreprises de moins de 50 employés. Elle permet de mettre en concurrence divers acteurs du marché qui proposent des plans basés sur le coût et les besoins, de centraliser l'information et les aides proposées. La plupart des Etats ont leur propre plateforme. Pour être éligible, il faut être citoyen américain vivant aux Etats-Unis, non-couvert par *Medicare*. Les plans d'assurance santé proposés sur la plateforme en ligne sont divisés en 4 catégories en fonction de leur valeur actuarielle, c'est-à-dire du pourcentage des dépenses de santé couvertes en moyenne par la police d'assurance souscrite :

— une police *Bronze* (Bronze) : entre 58 % et 62 % ;

— une police *Silver* (Argent) : entre 68 % et 72 % ;

— une police *Gold* (Or) : entre 78 % et 82 % ; et

— une police *Platinium* (Platine) : entre 88 % et 92 %.

Plus la valeur actuarielle de la police d'assurance souscrite augmente, plus les primes sont élevées, plus les franchises sont basses et donc plus le reste à charge des assurés est faible. En 2024, environ 21,5 millions d'Américains ont choisi ou ont été automatiquement réinscrits à une couverture d'assurance santé par le biais de la plateforme *Health Insurance Marketplace*, soit 31 % de plus qu'en 2023 [5]. 31 % d'entre eux ont une police *Bronze*, 54 % une police *Sliver*, 13 % une police *Gold*, 1 % une police *Platinium* et 1 % ont une police appelée *Catastrophic Plans*. Il s'agit d'une cinquième catégorie de police accessible uniquement aux individus de moins de 30 ans [5].

Après avoir choisi son niveau (Bronze, Argent, Or ou Platine), un individu doit déterminer le type d'assurance santé souhaité (HMO, POS, EPO ou PPO), comme proposé sur le marché de l'assurance santé privée.

1.1.3 Un système coûteux pour les Etats-Unis

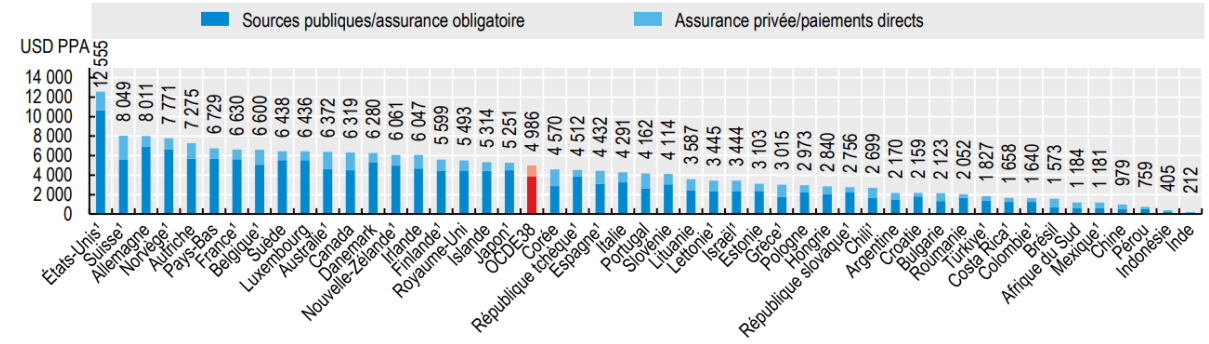
Les dépenses de santé correspondent à la consommation finale de biens et services de santé englobant :

— les dépenses engagées par tout dispositif de financement au titre de produits et de services médicaux ;

— les dépenses consacrées aux programmes de santé publique et de prévention et à l'administration des systèmes de santé.

En 2022, aux Etats-Unis, les dépenses de santé s'élèvent à 4 500 milliards de dollars, soit 16,6 % du PIB [6], devançant de loin l'Allemagne (12,7 %) et la France (12,1 %). Les Etats-Unis possèdent le système de santé le plus coûteux au monde avec 12 555 dollars dépensés par habitant sur l'année 2022 contre 4 986 dollars en moyenne dans les pays de l'OCDE, comme en témoigne la figure 1.3. A titre de comparaison, la France a dépensé 6 630 dollars par habitant en 2022.

FIGURE 1.3 – Dépenses de santé par habitant, 2022 (ou année la plus proche) [6]



1. Estimations de l'OCDE.
Source : Statistiques de l'OCDE sur la santé 2023 ; Base de données de l'OMS sur les dépenses de santé mondiales.

Note de lecture : les Etats-Unis ont dépensé 12 555 USD PPA par habitant en 2022. Cela indique que, en termes de pouvoir d'achat, chaque habitant de ce pays bénéficie de soins de santé équivalents à ce que 12 555 dollars pourraient acheter aux Etats-Unis.

Un tel écart entre les dépenses de santé aux Etats-Unis et la moyenne de l'OCDE peut être dû à plusieurs facteurs. Une liste non exhaustive des causes pouvant être à l'origine de cette disparité est proposée ci-dessous :

— le prix des prestations de santé aux Etats-Unis. Il faut compter en moyenne entre 100 et 200 dollars pour une consultation chez un médecin généraliste, entre 300 et 500 dollars pour une consultation chez un spécialiste et entre 1500 et 4000 dollars par jour pour une hospitalisation [7] ;

— le prix des médicaments : leur coût est, en moyenne, deux à trois fois plus élevé qu'en France [7] ;

— la crise du covid-19 a drastiquement augmenté la part des dépenses nationales allouées à la santé depuis 2020 [6] ;

— l'augmentation de la prévalence de maladies chroniques aux Etats-Unis, notamment l'obésité et le diabète : le taux d'obésité aux Etats-Unis (42,8 %) est quasiment deux fois supérieur au taux moyen des pays de l'OCDE [8].

Pour corroborer les points évoqués ci-dessus, les dépenses de santé nationales peuvent être analysées par type de prestataire de soins, mais également par type de maladie via le coût des traitements correspondants. La figure 1.4 montre qu'en 2022, 67,3 % des dépenses se concentrent sur cinq pôles de santé [4] :

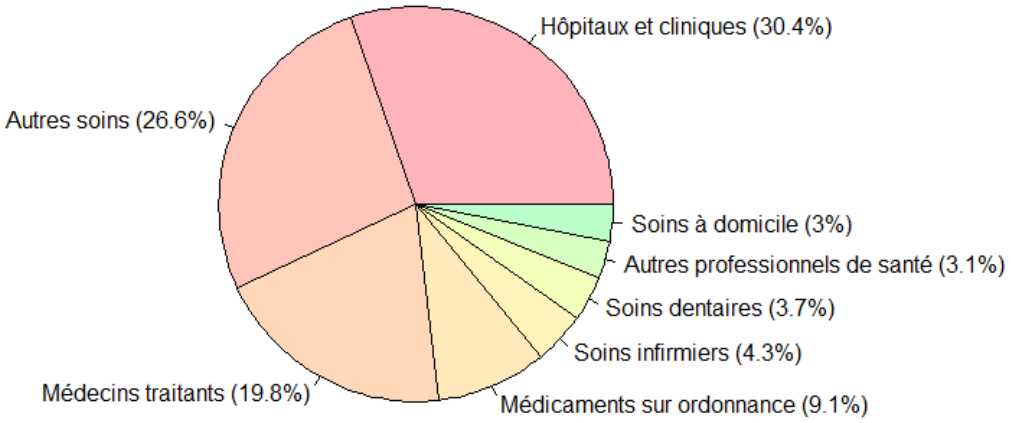
— les dépenses hospitalières, représentant près d'un tiers de l'ensemble des dépenses de santé en 2022 (30,4 %) ;

— les médecins traitants (19,8 %) ;

— les médicaments sur ordonnance (9,1 %) ;

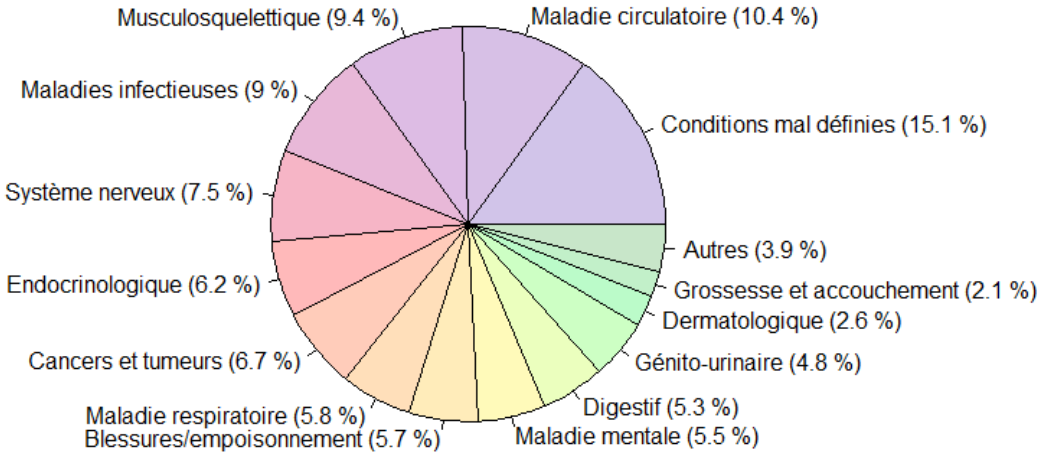
— les soins infirmiers (4,3 %) ;
— les soins dentaires (3,7 %).

FIGURE 1.4 – Répartition des dépenses de santé aux Etats-Unis en 2022 par type de prestataire de soins [4]



Note de lecture : en 2022 aux Etats-Unis, 3,7 % des dépenses de santé ont été allouées aux soins dentaires.

FIGURE 1.5 – Répartition des dépenses de santé liées aux coûts des traitements aux Etats-Unis en 2021 par type de maladie [9]



Note de lecture : 10,4 % des dépenses de santé liées aux traitements aux Etats-Unis ont été allouées aux maladies circulatoires.

Une autre façon d'étudier les dépenses de santé consiste à estimer le coût des traitements de certaines maladies plutôt que d'estimer les dépenses liées à un séjour à l'hôpital ou à une visite chez le médecin. Les chercheurs du *Bureau of Economic Analysis* (BEA) ont ainsi constaté que les catégories les plus importantes de dépenses de services médicaux comprennent [9] :

— le traitement des maladies circulatoires (10,4 %) ;

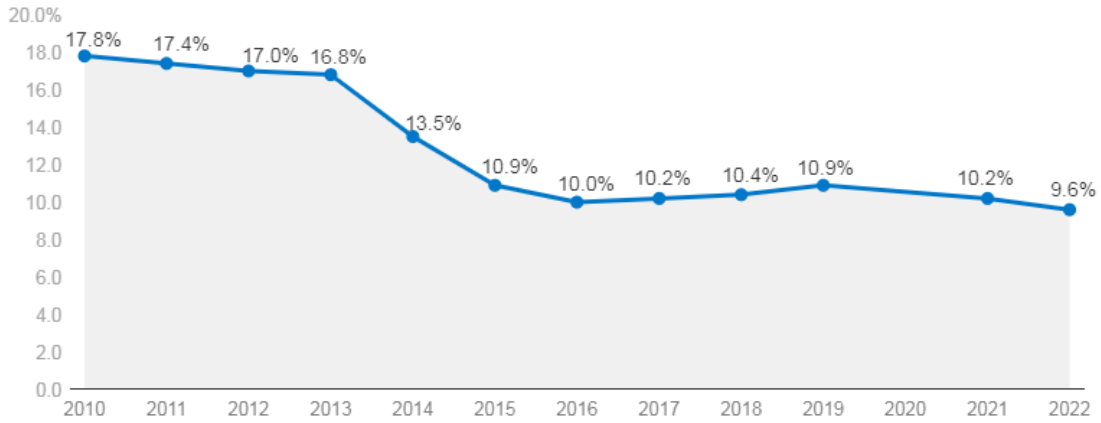
— les troubles musculo-squelettiques (9,4 %) ; et
— les maladies infectieuses (9,0 %).

Les chercheurs du BEA ont également mis en avant les « maladies mal définies », qui peuvent comprendre les examens de routine ou les soins de suivi qui ne sont pas facilement attribuables à une maladie particulière. Cette catégorie représente 15,1 % des dépenses liées aux traitements en 2021, s'élevant à 2 100 milliards de dollars. La figure 1.5, obtenue à partir des résultats du BEA montre la répartition des dépenses liées aux traitements divers des maladies aux Etats-Unis en 2021.

1.1.4 Une couverture assurantielle partielle des Américains

Le système de santé américain échoue à couvrir l'intégralité de la population dont une partie se retrouve sans assurance santé. En 2023, 7,6 % de la population américaine, soit plus de 25 millions d'habitants, ne sont pas couverts pour leurs frais de santé. La figure 1.6 représente l'évolution entre 2010 et 2022 de la part des Américains non-assurés âgés d'au plus 64 ans. Cette part est passée de 17,8 % à 9,6 % de 2010 à 2022, avec une forte baisse en 2014, date d'entrée en vigueur de l'ACA promulguée par Barack Obama. Dès son investiture le 20 janvier 2017, Donald Trump signe un décret relatif à l'ACA, prévoyant notamment des reports, des dérogations et des exemptions dans son application. Il demande également aux agences fédérales d'assouplir certaines contraintes qui pèsent sur les Etats en matière d'obligation de couverture. Ces mesures coïncident avec une légère hausse de la part d'Américains non-assurés entre 2016 et 2019. Depuis l'investiture de Joe Biden en janvier 2021, cette part ne cesse de diminuer pour atteindre 9,6 % en 2022, soit 26,5 millions d'Américains âgés d'au plus 64 ans.

FIGURE 1.6 – Part des Américains non-assurés âgés d'au plus 64 ans, 2010 - 2022 [10]



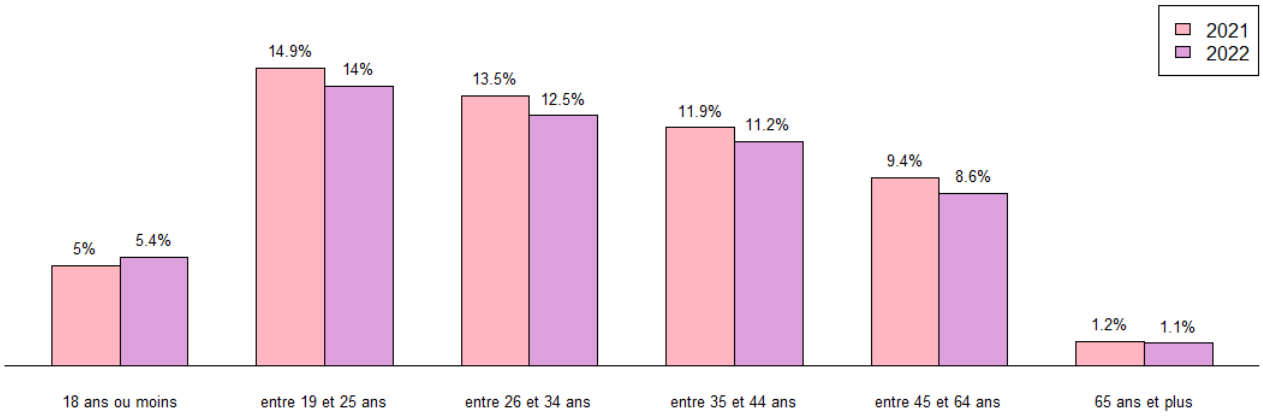
Note de lecture : en 2022, 9,9 % des Américains âgés d'au plus 64 ans ne possédaient pas d'assurance santé.

1.1.5 Un système de santé discriminatoire

Comme vu dans la partie précédente, environ 9,6 % des Américains âgés d'au plus 64 ans étaient dépourvus d'assurance santé en 2022. La figure 1.7 représente la part des Américains non-assurés par classe d'âge en 2021 et en 2022. D'une part, on remarque une baisse globale de la part d'individus non-assurés dans toutes les classes d'âges, sauf pour

les Américains de 18 ans ou moins. D'autre part, le taux d'individus sans couverture santé est plus important chez les jeunes adultes. Sur 100 jeunes âgés de 19 à 25 ans, 14 ne disposent d'aucune couverture maladie. Enfin, on observe l'efficacité du programme *Medicare* qui parvient à couvrir la quasi-totalité des Américains âgés de plus de 65 ans.

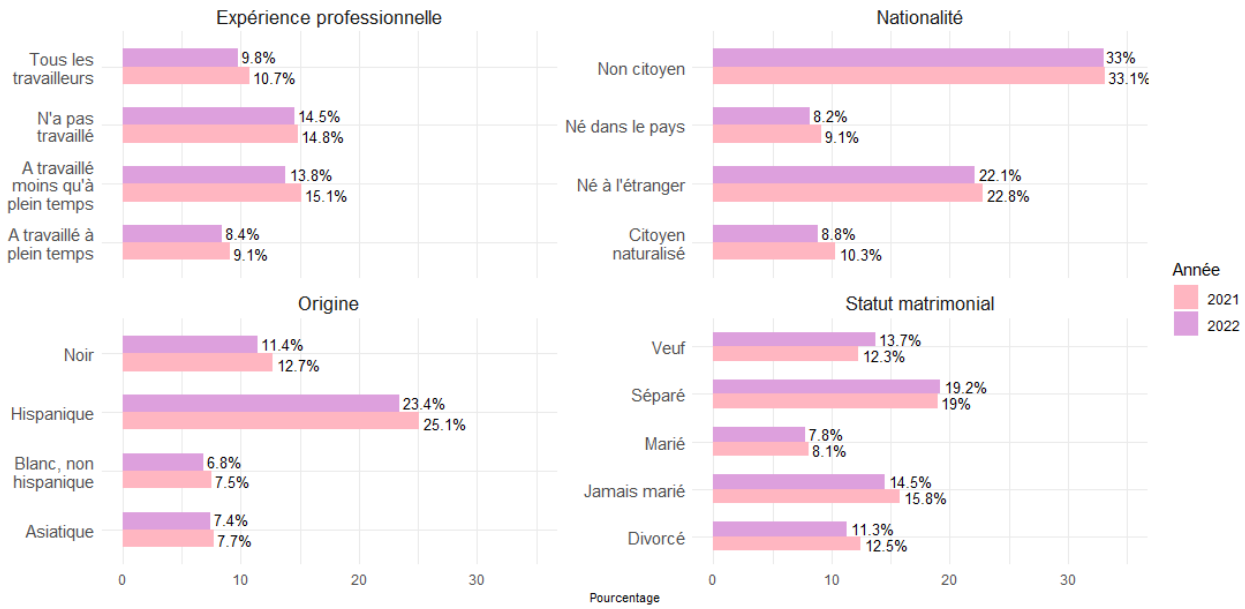
FIGURE 1.7 – Part des Américains non-assurés par classe d'âge en 2021 et 2022 [11]



Note de lecture : en 2021, 9,4 % des Américains âgés entre 45 et 64 ans ne possédaient pas d'assurance santé.

Aussi, malgré les programmes de couverture publics, le système de santé américain échoue à couvrir les personnes vivant dans les familles à faibles revenus. Les Américains dont les revenus sont inférieurs au seuil de pauvreté national représentent 23,6 % des non-assurés (19-64 ans) en 2022. 37,9 % des individus vivant dans les Etats n'ayant pas adopté l'extension de l'ACA ne possèdent aucune assurance santé [11].

FIGURE 1.8 – Part des Américains (19-64 ans) sans assurance santé par caractéristique en 2021 et 2022 [11]



Note de lecture : en 2021, 8,1 % des Américains mariés (19-64 ans) ne possédaient pas d'assurance santé.

Lorsque l'on s'intéresse à d'autres caractéristiques comme la situation professionnelle,

la nationalité, les origines ou encore le statut matrimonial des individus non-assurés, on remarque également des inégalités importantes (voir figure 1.8) :

— les travailleurs à temps plein sont plus couverts que les travailleurs à temps partiel ou les individus au chômage ;

— les personnes n'ayant pas la nationalité américaine ou nées à l'étranger sont 2 à 3 fois moins couvertes que les natifs et les citoyens naturalisés ;

— le système discrimine les minorités ethniques : 7,5 % des « Blancs » non-hispaniques en 2021 ne possèdent pas d'assurance santé contre 25,1 % des Hispaniques et 12,7 % des « Afro-Américains » ;

— les personnes mariées sont davantage assurées que les individus seuls (veufs, jamais mariés), séparés ou divorcés.

1.2 Les bases de données *MedInsight*

Milliman MedInsight, fondée en 1998, est une entreprise spécialisée dans la fourniture de données et d'analyses dans le domaine de la santé, à destination des assureurs publics et privés, des prestataires de soins, des organisations de soins coordonnés, des employeurs et des agences gouvernementales. Elle est considérée comme un leader dans le domaine de l'analyse des données de santé : plus de 300 organisations de santé lui font confiance. *Milliman MedInsight* permet à ces organisations de valoriser leurs données de santé afin de faciliter la prise de décision, d'optimiser les résultats cliniques et financiers et de réduire les gaspillages financiers. L'entreprise s'appuie sur une expertise approfondie du secteur ainsi que sur des technologies avancées pour offrir des analyses détaillées concernant l'utilisation des soins, les coûts, la qualité et la performance des systèmes de santé et propose des produits analytiques basés sur le *cloud*, avec des rapports et des configurations de données personnalisables, afin de répondre aux besoins spécifiques de chaque organisation.

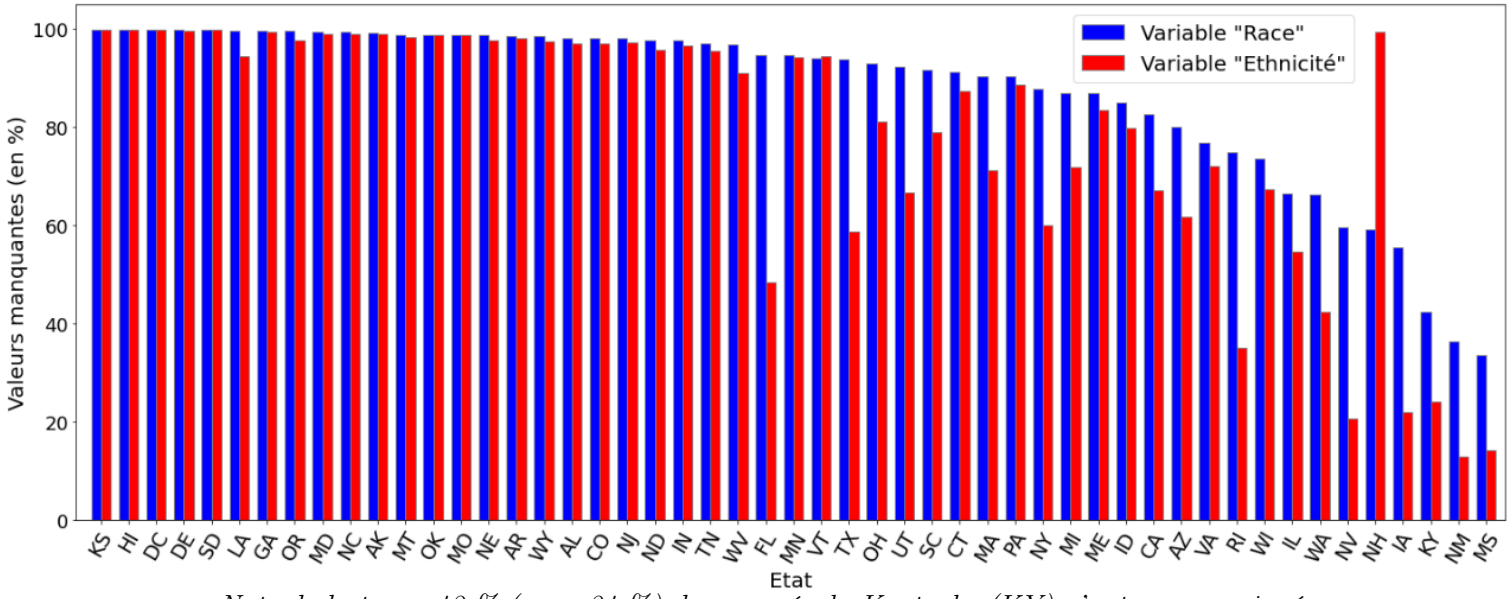
Dans le cadre de ce mémoire, *Milliman MedInsight* a fourni trois bases de données contenant des informations sur 48,5 millions d'assurés aux Etats-Unis ayant souscrit une assurance santé entre le 1er janvier 2017 et le 31 décembre 2023. Ces données relatives aux demandes de remboursement issues de sources privées correspondent aux services de santé évalués et réglés par les différents assureurs, tels que les organismes d'assurance privés, les plans de santé publics, les administrateurs tiers ou les gestionnaires de prestations pharmaceutiques, sur l'ensemble du territoire américain. Près de 80 organismes de santé ont transmis chaque mois ces données, qui rassemblent plus de 7,3 milliards d'enregistrements médicaux et pharmaceutiques. Les données renseignent notamment sur le statut de couverture, les frais de santé, les caractéristiques démographiques, l'âge, le sexe, l'ethnicité et l'état de santé de chaque individu de la base éligible à des prestations de santé privées et/ou publiques sur la période. Les trois bases de données fournies sont décrites dans les sections suivantes.

1.2.1 Base « assurés »

Une première base « assurés » donne des caractéristiques propres à l'assuré comme le genre, l'âge, l'origine ethno-raciale ainsi que l'Etat et la zone statistique métropolitaine (MSA) de résidence de l'individu. Aux Etats-Unis, une distinction est faite entre « race » et ethnicité. La première notion est une classification sociale basée principalement

sur des caractéristiques physiques comme la couleur de peau. Dans la base « assurés », les modalités de cette variable sont : *Caucasien, Afro-Américain, Asiatique, Autre ou Inconnue*. La seconde notion concerne l'origine culturelle et nationale. Dans la base de données étudiée, l'ethnicité est divisée en deux groupes principaux : *Hispanique ou Latino, Non-Hispanique et Non-Latino*. Chaque assuré de la base possède un identifiant interne unique facilitant le lien avec les autres bases de données. L'un des défis majeurs de cette base de données est le taux élevé de valeurs manquantes concernant l'origine ethno-raciale des assurés. Toutefois, cette absence d'information n'est pas homogène sur l'ensemble du territoire comme en témoigne la figure 1.9. En effet, l'analyse par Etat révèle des disparités significatives. Si, dans de nombreux cas, les données sur l'origine ethno-raciale sont largement absentes (plus de 80 % de valeurs manquantes dans 40 Etats), six Etats se démarquent avec plus de 40 % de valeurs renseignées. Parmi eux, le Mississippi, le Nouveau-Mexique et le Kentucky affichent des taux de complétion de la variable « Race » relativement élevés avec respectivement 66 %, 64 % et 58 % de valeurs renseignées. Ces disparités s'expliquent par les politiques locales et les pratiques administratives propres à chaque Etat.

FIGURE 1.9 – Valeurs manquantes par Etat pour les variables ethno-raciales



Note de lecture : 42 % (resp. 24 %) des assurés du Kentucky (KY) n'ont pas renseigné leur « race » (resp. « ethnicité »).

Dans ces données, les Etats sont divisés en zones de codes postaux américains à trois chiffres, appelés *ZIP Code 3-Digits* (ZIP3 par simplification) représentant les trois premiers chiffres d'un code postal : le premier chiffre d'un code postal à cinq chiffres divise les Etats-Unis en une zone équivalente à 10 grands groupes d'Etats, numérotés de 0 dans le Nord-Est à 9 dans l'extrême ouest. A l'intérieur de ces zones, chaque Etat est ensuite découpé en zones géographiques plus petites, identifiées par le deuxième et le troisième chiffre. Pour illustrer ces propos, un exemple de division en code postal à trois chiffres de l'Etat du Kentucky est donné dans la figure 1.10 réalisée à l'aide du logiciel QGIS, système d'informations géographiques. L'Etat du Kentucky comprend 27 codes postaux à trois chiffres numérotés de 400 à 427 (le code postal 419 n'existe pas). Pour chaque individu, dans la base « assurés » est renseigné son ZIP3 de résidence.

Pour résumer les informations précédentes, une liste des variables présentes dans la base « assurés » est donnée ci-après avec une description détaillée :

— L'identifiant unique de la personne à travers tous les enregistrements. Ce champ est dérivé des dossiers d'inscription et vise à identifier de manière cohérente les individus dans les données sources.

— Le genre du membre. Les valeurs possibles sont `F` (femme), `M` (homme) ou `U` (genre inconnu).

— La « race » du membre, codée selon les valeurs suivantes :
1 = Asiatique, 2 = Noir, 3 = Caucasien et 4 = Autre ou Inconnu.

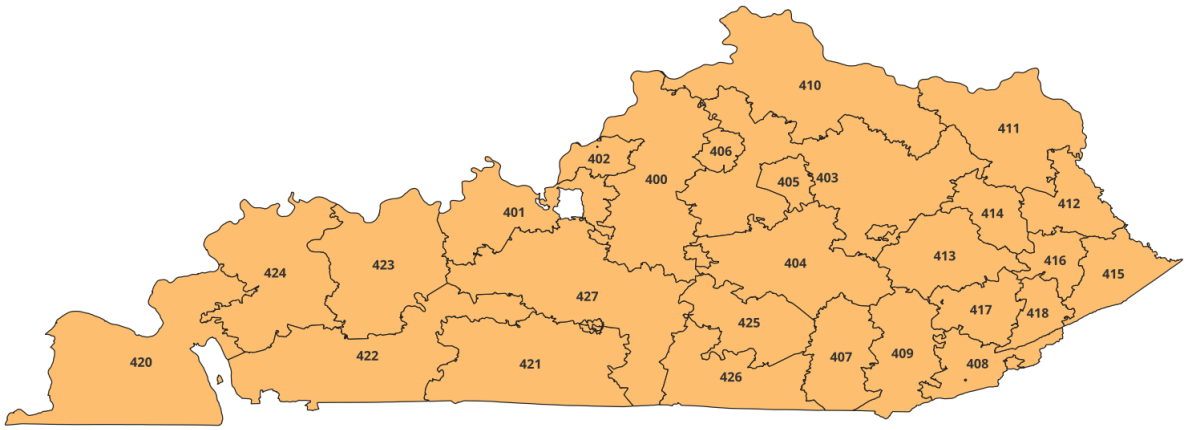
— L'ethnicité du membre, codée comme suit :
1 = Hispanique ou Latino, 2 = Non Hispanique ou Latino et 3 = Ethnicité inconnue

— Les trois premiers chiffres du code postal de résidence du membre. Cette variable sera nommée ZIP3 dans la suite du mémoire.

— La zone statistique métropolitaine (MSA) de résidence du membre.

— L'Etat de résidence du membre.

FIGURE 1.10 – Carte des ZIP3 du Kentucky (KY)



Note de lecture : le ZIP Code 3-Digits 420 est le code postal à trois chiffres le plus à l'ouest du Kentucky.

1.2.2 Base « souscriptions »

Une base « souscriptions » donne des informations mensuelles entre le 1er janvier 2017 et le 31 décembre 2023 sur les types d'assurances souscrites par chaque individu de la base. Chaque observation correspond à une personne donnée pour un mois donné, ce qui permet de suivre finement l'évolution de la couverture d'assurance au fil du temps. Pour chaque individu et chaque mois, sont renseignés :

— la date de début du mois d'enregistrement, permettant de situer précisément la période de couverture ;

— l'âge de l'individu au moment de la souscription ;

— le lien de parenté avec l'assuré principal, par exemple conjoint ou personne à charge ;

— la zone géographique de résidence, identifiée au niveau de la zone statistique métropolitaine (MSA) ;

— le type de programme d'assurance auquel l'individu est affilié : *Medicaid*, *Medicare*, assurance commerciale ou complémentaire *Medicare* ;

— le type de couverture souscrite : médicale, dentaire, vision ou médicamenteuse (prescription) ;

— la durée mensuelle de couverture pour chaque type de prestation (en nombre de mois couverts) ;

— la forme de l'organisation de soins choisie, telle que les réseaux HMO, PPO ou EPO ;

— le statut de l'individu au regard des prestations : actif, retraité ou bénéficiaire d'un régime de continuation comme le COBRA ;

— le groupe de pathologies chroniques associé à la souscription mensuelle, chaque individu étant classé dans une catégorie de diagnostic selon ses conditions médicales principales. Une description détaillée de ces groupes est disponible en annexe A.

Cette base de données contient environ 5 milliards d'observations sur les souscriptions des 48,5 millions d'assurés.

1.2.3 Base « sinistres »

Une base « sinistres » recense l'ensemble des événements médicaux donnant lieu à remboursement ou à facturation pour la période allant du 1er janvier 2017 au 31 décembre 2023. Elle contient plus de 7,5 milliards d'enregistrements individuels, chacun correspondant à une prestation médicale précise (consultation, examen, hospitalisation, délivrance d'un médicament, etc.). Chaque observation regroupe plusieurs dimensions d'information, que l'on peut organiser selon six grandes catégories :

— **Informations financières** : les coûts associés à chaque prestation sont détaillés, incluant le montant facturé, le montant autorisé après négociation, le montant payé par le régime d'assurance, ainsi que les restes à charge pour l'assuré (franchise, copaiement, quote-part, etc.).

— **Typologie des prestations** : la nature de chaque prestation est précisée via différents codages standards (codes de procédure, codes de diagnostic ICD-10 [2], codes de classification CCS [3], etc.), ainsi que par le type de formulaire utilisé (soins médicaux, dentaires, pharmaceutiques, hospitaliers, etc.).

— **Caractéristiques administratives du sinistre** : chaque prestation est rattachée à un identifiant de réclamation, à des dates clés (date de service, de paiement, d'admission et de sortie), à un statut de traitement (payée, rejetée, en attente, etc.), et à des informations sur les prestataires impliqués (prescripteur, facturant, soignant).

— **Informations cliniques** : la base permet de retracer les pathologies traitées, via les diagnostics (jusqu'à 30 par sinistre), les actes chirurgicaux ou thérapeutiques

2. La Classification internationale des maladies ou CIM (en anglais, *International Classification of Diseases* ou ICD) est une classification médicale codifiée classifiant les maladies et une très vaste variété de signes, symptômes, lésions traumatiques, empoisonnements, circonstances sociales et causes externes de blessures ou de maladies. Lien de la classification.

3. La *Clinical Classifications Software* (CCS) regroupe les codes de diagnostic et de procédure de la classification ICD en catégories cliniques agrégées, plus lisibles et exploitables statistiquement.

effectués, ainsi que les catégories cliniques auxquelles elles appartiennent. Les pathologies chroniques sont également identifiées par un système de regroupement diagnostique spécifique.

— **Caractéristiques des assurés** : pour chaque sinistre sont renseignés l'âge de l'individu au moment du soin, sa situation professionnelle vis-à-vis du régime d'assurance (actif, retraité, etc.), sa relation avec l'assuré principal (conjoint, enfant, etc.), ainsi que sa couverture d'assurance (type de programme et mode de gestion).

— **Cadre de soins** : la base distingue les contextes de prise en charge (ambulatoire, hospitalisation, pharmacie, etc.), ainsi que le statut du prestataire (en réseau ou hors réseau) et le lieu de prestation selon les standards de codage en vigueur.

1.3 Les enjeux du traitement d'un grand volume de données sensibles

Les compagnies d'assurance peuvent être amenées à manipuler des volumes de données colossaux, souvent composés d'informations sensibles. C'est notamment le cas des assureurs santé. Dans ce contexte, la confidentialité et la volumétrie des données représentent des enjeux majeurs qui influencent directement leurs choix technologiques et méthodologiques et qui ont un coût conséquent. La sécurisation des données est devenue une priorité, notamment avec l'essor des réglementations comme le Règlement Général sur la Protection des Données (RGPD) en Europe ou le *Health Insurance Portability and Accountability Act* (HIPAA) aux Etats-Unis. De plus, la gestion de très grands ensembles de données, comme les milliards d'observations considérées dans cette étude, impose des défis techniques et financiers considérables en matière de stockage, de traitement et d'analyse.

Pour garantir la confidentialité et la bonne gestion de grands volumes de données, des plateformes comme ***Databricks*** et des technologies comme ***Apache Spark*** ont été développées. Elles ont été très largement utilisées pour réaliser ce mémoire.

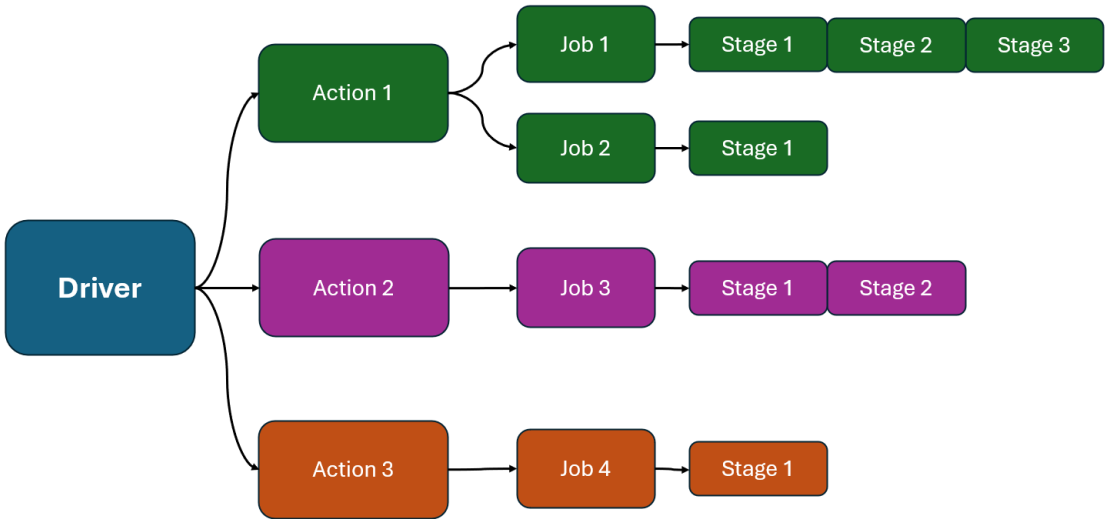
1.3.1 Volumétrie des données : enjeux et défis

Les données de santé exploitées sont particulièrement massives. Lorsqu'il s'agit de traiter des milliards de lignes d'observations, comme c'est le cas dans cette étude, la tâche devient bien plus ardue que pour des bases de données classiques contenant quelques milliers, voire quelques millions de lignes. En effet, ces données ne peuvent pas être traitées avec les méthodes computationnelles usuelles. Aussi, les bases de données ne peuvent être directement stockées sur l'ordinateur en raison de leur taille importante. Les données utilisées dans le cadre de ce mémoire ont donc été stockées sur une plateforme en ligne appelée ***Databricks*** et sont exploitées avec ***Apache Spark***.

Spark (ou ***Apache Spark***) est une architecture *open source* de calcul distribué développée à l'université de Californie à Berkeley. ***Spark*** est un cadre applicatif de traitements des mégadonnées, aussi appelées *Big Data*, conçu pour effectuer des analyses complexes à grande échelle. Sa principale force réside dans sa capacité à exécuter des calculs en temps quasi réel ou par traitements par lots, répartis sur plusieurs clusters. Dans ***Spark***, une tâche, appelée *job*, représente une unité de travail distincte qui est soumise au calcul

sur un cluster. Chaque *job* est orchestré par un *driver* (le pilote), qui gère l'application de traitement de données pour un cas d'utilisation spécifique. Ce *driver* coordonne la distribution des tâches et l'exécution des calculs auprès des différents nœuds du cluster. Chaque *job* est divisé en étapes individuelles, appelées *stage*, représentant une unité de travail pouvant être exécutée en parallèle, sur différentes partitions des données. Enfin, ces *stages* sont eux-mêmes divisés en une ou plusieurs tâches (*tasks*). Ce sont des unités de calcul élémentaires réalisées par les exécutants du cluster. Les résultats intermédiaires et finaux sont produits lorsque les *tasks* ont terminé leurs calculs. Le *job* s'achève lorsque tous les *stages* et toutes les *tasks* ont été exécutés. La figure 1.11 ci-dessous résume le fonctionnement de l'architecture Spark.

FIGURE 1.11 – Fonctionnement de l'architecture Spark



Note de lecture : une Action Spark se divise en plusieurs jobs qui se divisent eux-mêmes en plusieurs stages.

Concrètement, dans les *jobs*, les données sont partitionnées sur plusieurs nœuds pour permettre un traitement parallèle. Ainsi, chaque *job* exploite un sous-ensemble des données disponibles, ce qui permet une distribution et un calcul efficaces. Grâce à cette approche de calcul distribué, ***Spark*** peut traiter rapidement de grands ensembles de données et évoluer horizontalement à mesure que les ressources deviennent disponibles. A titre d'exemple, compter un nombre d'observations en fonction de la valeur spécifique d'une variable se réalise en un seul *job* contenant plusieurs étapes :

- **Stage 1** : lecture des données nécessaires pour la réalisation de l'opération.
- **Stage 2** : regroupement (`groupBy`).
- **Stage 3** : comptage du nombre d'observations au sein du groupe sélectionné.
- **Tasks** : chaque *stage* est divisé en plusieurs tâches traitant des partitions différentes des données, réparties sur plusieurs nœuds du cluster et traitées en parallèle pour optimiser la performance.

Grâce à l'architecture distribuée de ***Spark***, cette agrégation est réalisée efficacement, même sur une base de données massive.

1.3.2 Confidentialité et sensibilité des données

Comme mentionné précédemment, les trois bases de données présentées en partie 1.2 sont directement stockées sur la plateforme ***Databricks***. Il s'agit d'une entreprise fondée en 2013 aux Etats-Unis par les créateurs de ***Spark***, qui propose une plateforme *cloud* unifiée dédiée à la gestion des données. La plateforme permet à ses utilisateurs de traiter à grande échelle des données variées pour des usages en sciences des données, en apprentissage automatique et, plus récemment, en IA générative.

Databricks met un accent particulier sur la sécurité, la gouvernance et la confidentialité des données, en particulier pour les organisations manipulant des données sensibles comme les secteurs de la santé, de la finance, de l'administration publique ou encore des assurances. Aujourd'hui, plus de 10 000 organisations dans le monde utilisent ***Databricks*** pour analyser leurs données et les mettre au service de leurs projets d'intelligence artificielle. Elle respecte les normes strictes de sécurité, notamment FedRAMP, un programme de conformité à l'échelle du gouvernement fédéral des Etats-Unis qui fournit une approche standardisée pour l'évaluation de la sécurité, l'autorisation et la surveillance continue des produits et services *cloud*. Elle intègre des fonctionnalités de contrôle d'accès, de traçabilité, de chiffrement et de supervision des flux de données, permettant une gestion fine et conforme des données confidentielles. Ainsi, pour accéder aux données utilisées dans le cadre de ce mémoire et au cluster de calcul, un identifiant ainsi qu'un mot de passe sont nécessaires. Le cluster est sécurisé avec notamment un temps d'inactivité limité à 20 minutes au-delà duquel il se réinitialise. Il est également impossible d'exporter des données de la plateforme vers un environnement externe : la visualisation, le traitement ainsi que l'analyse des données de santé exploitées se font exclusivement sur la plateforme sécurisée.

Cependant, cette solution reste relativement coûteuse. En effet, dès l'utilisation de la plateforme (mise en route du noyau), la facturation débute, et ce, indépendamment du volume de données et de stockage utilisé. L'exploitation de ***Databricks*** est facturée aux DBU (*Databricks Units*) consommées. Les DBU représentent une unité de mesure créée par la plateforme en ligne. Elle prend notamment en compte le type de ressource utilisée, le type de cluster, ainsi que la durée d'exécution. Ainsi, le coût augmente avec la puissance et la spécificité des ressources allouées. A ces frais initiaux peuvent s'ajouter des coûts dus à des fonctionnalités supplémentaires, telles que la gestion de la sécurité, de la confidentialité, le support technique ou l'utilisation d'outils spécifiques à l'intelligence artificielle.

Dans le cadre de ce mémoire, malgré les solutions permettant de traiter les enjeux imposés par la base de données, les coûts d'utilisation des données restent élevés. Par conséquent, **les bases d'étude seront restreintes** et **les modèles complexes écartés** pour éviter un surcoût conséquent.

Chapitre 2

Scores de santé, climat et pollution

Le but de ce mémoire est de créer des scores de santé à partir des données fournies par *Milliman MedInsight*, intégrant des variables climatiques ainsi que des variables de pollution. Ce chapitre décrit les scores de santé utilisés aux Etats-Unis et leur réglementation. Un point d'attention est accordé à la justification de l'intégration des variables environnementales dans les modèles de score de santé qui seront décrits dans la partie II et implémentés dans la partie III.

2.1 Les scores de santé aux Etats-Unis

2.1.1 Scores de santé élaborés par les organismes publics

Aux Etats-Unis, les scores de santé sont des outils incontournables pour résumer la situation sanitaire d'un territoire, d'une population ou d'un système de soins, à partir d'une combinaison d'indicateurs statistiques. Ces scores permettent, à différentes échelles, de comparer la santé des populations en mettant en lumière les inégalités territoriales et sociales. Ils éclairent notamment les décideurs politiques qui s'appuient sur ces indicateurs de santé pour cibler les interventions et allouer les ressources nécessaires aux territoires les plus à risque. Les organismes publics utilisent ces scores pour la veille sanitaire et pour l'évaluation des impacts des mesures prises par les responsables politiques locaux. En tant qu'outils de communication, ils permettent de sensibiliser les citoyens et les acteurs locaux et de contribuer à mobiliser les communautés autour des enjeux de santé. Parmi les scores de santé les plus connus aux Etats-Unis se distinguent le *County Health Rankings* élaboré par l'Université du Wisconsin, l'*America's Health Rankings* basé sur le travail de la *United Health Foundation* ou encore les indices développés par les Centres pour le contrôle et la prévention des maladies (CDC) qui forment ensemble la principale agence fédérale des Etats-Unis en matière de protection de la santé publique. A titre d'exemple, la *United Health Foundation* a publié en 2021 un rapport sur les disparités en matière de santé. S'appuyant sur plus de 30 ans de données, ce rapport fournit un portrait complet de l'existence des disparités en santé à travers le pays, dans le but d'orienter les actions en faveur de l'équité en santé. Le document dépeint notamment les disparités à l'échelle nationale, dans les 50 Etats ainsi que dans le District de Columbia, en fonction du genre, de la géographie, du niveau d'éducation, ainsi que de la « race » et de l'origine ethnique. Les conclusions de ces travaux ont pour but principal d'éclairer les dirigeants et les décideurs sur les défis auxquels sont confrontés les Américains aujourd'hui.

2.1.2 Scores de santé individuels à des fins assurantielles

Aux Etats-Unis, le secteur de l'assurance santé s'appuie également sur des outils statistiques pour créer des scores de santé visant à évaluer le risque individuel des assurés. Les scores de santé individuels sont des instruments essentiels pour les assureurs privés et publics (*Medicare*, *Medicaid*, compagnies d'assurance privées). Ils fournissent une mesure quantifiable du risque santé, éclairant ainsi les stratégies de tarification et les plans proposés. Les scores de santé individuels permettent d'approcher la probabilité qu'un individu développe une maladie ou nécessite des soins coûteux et d'adapter les primes, franchises, plafonds et conditions de couverture en fonction. Des modèles statistiques ou d'intelligence artificielle permettent alors de prédire le risque d'occurrence d'événements sanitaires majeurs comme une hospitalisation, des complications ou encore le risque de mortalité ou de prédire les coûts futurs d'un assuré en particulier. Pour l'assureur, l'enjeu est de maîtriser les coûts en anticipant les dépenses de santé tout en segmentant au mieux le portefeuille d'assurés pour proposer des offres adaptées et inciter à la prévention. La construction de scores de santé individuels est strictement encadrée aux Etats-Unis par l'*Affordable Care Act*, comme décrit dans la section 2.1.3.

Avant l'entrée en vigueur de l'ACA en 2014, dans la plupart des Etats, les tarifs des polices d'assurance santé étaient basés sur le niveau de risque de l'individu ou du portefeuille assuré grâce à la sélection médicale. Les assureurs utilisaient des outils tels que des questionnaires médicaux détaillés, des examens médicaux et l'analyse des niveaux de sinistres encourus pour déterminer le tarif approprié pour chaque nouvel assuré ou chaque renouvellement individuel ou de groupe.

2.1.3 Encadrement de l'utilisation des scores de santé par l'ACA

L'*Affordable Care Act*, entré en vigueur en 2014, a pour objectif d'augmenter le nombre de personnes couvertes par une assurance maladie, en créant une couverture d'assurance à la fois complète et abordable, et en garantissant des niveaux minimaux de couverture et des primes facilement comparables. L'objectif était de faciliter la compréhension du marché par les consommateurs et de leur fournir un cadre mieux adapté à la comparaison des régimes. Trois éléments clés ont été inclus dans l'ACA pour atteindre ces objectifs du point de vue du consommateur :

— un mandat individuel qui introduit une pénalité financière en cas de non-achat d'une assurance maladie ;

— des subventions aux primes et une réduction des frais à la charge des familles à faible revenu ; et

— des réformes de la tarification pour garantir que les primes soient plus cohérentes entre les individus, y compris la tarification unisexe, la compression de la courbe d'âge pour les primes d'assurance maladie et la réduction des frais à la charge des familles à faible revenu.

L'ACA a notamment **interdit la tarification fondée sur l'état de santé individuel** en supprimant les exclusions pour conditions préexistantes : la variation des primes entre individus n'est autorisée qu'en fonction de l'âge, de la région géographique, de la composition familiale et de la consommation de tabac. La tarification en fonction de l'âge est également encadrée : l'écart de prix entre l'âge le plus élevé et l'âge le plus bas est

limité à un ratio de 3/1 même si, en réalité, les coûts réels de santé sont plus élevés chez les personnes âgées. Par exemple, si la prime pour le premier groupe d'âge est de 100 $, alors la prime pour le dernier groupe d'âge ne peut pas dépasser 300 $. Ainsi, les adhérents plus jeunes subventionnent désormais les plus âgés et les adhérents en meilleure santé subventionnent ceux en moins bonne santé. Ces mesures visent à rendre l'assurance maladie plus uniformément abordable.

Cependant, elles exposent les assureurs au risque de devoir gérer des portefeuilles à hauts risques sans pouvoir ajuster la tarification des polices souscrites en conséquence. Pour diminuer ce risque supporté par les assureurs, l'ACA a mis en place un programme d'ajustement du risque (*risk adjustment*), qui repose sur le calcul de scores de santé. Ces scores, déterminés à partir de données cliniques et socio-démographiques, servent exclusivement à rééquilibrer financièrement les assureurs selon le profil de risque de leurs assurés et à mutualiser les risques. Ils ne peuvent servir à fixer les primes individuelles. Contrairement aux scores individuels, les scores de santé du programme d'ajustement du risque peuvent utiliser des variables cliniques individuelles comme les antécédents médicaux, les diagnostics, les traitements et l'utilisation de médicaments et s'appuient sur des méthodes statistiques comme la régression linéaire. Le modèle de *risk adjustment* de l'ACA a pour objectif de garantir que la concurrence entre assureurs porte sur la qualité et la gestion des soins, plutôt que sur la sélection des risques. Ce mécanisme, strictement encadré par les autorités fédérales, permet ainsi d'éviter les dérives, de garantir l'équité en matière d'assurance santé et de favoriser l'accès à une couverture santé au plus grand nombre.

Par exemple, *Medicare* utilise un modèle appelé CMS-HCC pour produire les scores de santé relatifs à l'ajustement du risque. Ce modèle attribue un score à chaque assuré qui se base sur :

- des données socio-démographiques (âge, genre ou encore statut *Medicaid*) ;
- la présence de certaines maladies chroniques, identifiées à partir des codes ICD (par exemple, diabète, insuffisance cardiaque, cancer) ;
- un indicateur d'éligibilité à *Medicaid* ;
- un indicateur d'ouverture des droits à *Medicare* pour cause d'invalidité ; et
- six interactions de maladies.

Un score élevé indique un risque accru de générer des dépenses de santé importantes. Plus un assuré présente un nombre élevé de conditions chroniques, plus son score va augmenter. Ce score global est ensuite utilisé pour ajuster les transferts financiers entre assureurs, afin de compenser ceux qui couvrent une population plus risquée et d'éviter la sélection adverse.

Bien que le concept d'un score de santé soit pertinent et utile, il pose des défis majeurs. La précision et la fiabilité du score dépendent majoritairement de la qualité des données, de la robustesse des algorithmes et de la pertinence contextuelle des mesures. Les considérations éthiques, la confidentialité des données et les biais potentiels doivent donc être pris en compte pour garantir une utilisation responsable des scores de santé.

2.2 Comparaison avec le système de santé français

2.2.1 Un système de santé universel

En France, il existe un système universel fondé sur la solidarité et la mutualisation appelé la Sécurité sociale. Créée en 1945, la Sécurité sociale « est la garantie donnée à chacun qu'en toutes circonstances, il disposera des moyens nécessaires pour assurer sa subsistance et celle de sa famille dans des conditions décentes » [1]. Elle a pour objectif de protéger les individus des conséquences de divers événements ou situations, généralement qualifiés de risques sociaux. La Sécurité sociale couvre différentes catégories socioprofessionnelles se caractérisant par des modalités de gestion et de prise en charge différentes : le régime général, qui prend en charge la majorité de la population ; le régime agricole, qui prend en charge les exploitants et salariés agricoles et de nombreux régimes spéciaux, comme celui des marins, de la SNCF, de la RATP, d'EDF-GDF, de l'Assemblée nationale, ou encore du Sénat. La Sécurité sociale se compose de 6 branches [2] :

— la branche maladie (maladie, maternité, invalidité, décès) ;
— la branche famille (dont handicap et logement...) ;
— la branche accidents du travail et maladies professionnelles ;
— la branche retraite (vieillesse et veuvage) ;
— la branche autonomie ;
— la branche cotisations et recouvrement.

L'Assurance Maladie regroupe la branche maladie ainsi que la branche accidents du travail et maladies professionnelles de la Sécurité sociale. Elle a pour mission de garantir à tous les assurés une protection contre les risques liés à la maladie, à la maternité, à l'invalidité, ainsi qu'aux accidents du travail, aux maladies professionnelles et au décès. Chaque employé français cotise, ce qui permet de financer la prise en charge partielle ou totale des dépenses de santé de la population selon un barème défini. L'Assurance Maladie intervient ainsi dans le remboursement des soins médicaux, des médicaments, des hospitalisations, mais aussi dans le versement d'indemnités journalières en cas d'arrêt de travail ou d'une pension en cas d'invalidité. Elle assure également la gestion des droits des assurés, le versement des prestations dues et la conduite d'actions de prévention. Pour faciliter l'accès aux soins, le système repose sur la carte Vitale, qui permet une gestion informatisée et rapide des remboursements, et sur le dispositif du tiers-payant qui limite l'avance de frais pour les patients. Toutefois, les garanties de base de l'Assurance Maladie ne couvrent pas l'intégralité des frais médicaux ; c'est pourquoi la majorité des assurés souscrivent une complémentaire santé (ou mutuelle), qui prend en charge le reste à charge (ou ticket modérateur), les dépassements d'honoraires ou certains soins non remboursés. Ce modèle, fondé sur la solidarité nationale et la mutualisation des risques, vise à garantir un accès équitable aux soins pour l'ensemble de la population et ne nécessite pas l'utilisation de score de santé.

1. Ordonnance du 4 octobre 1945, texte fondateur.
2. Une branche est une entité qui a à sa charge la gestion d'un ou plusieurs « risques » définis comme des événements qui peuvent, au cours d'une vie, porter atteinte à la sécurité économique d'une personne.

2.2.2 Complémentaire santé : une utilisation des données de santé encadrée

La complémentaire santé est un contrat facultatif ayant pour but de compléter, en totalité ou partiellement, les remboursements de l'Assurance Maladie. Ces contrats permettent une prise en charge de tout ou partie des restes à charge en fonction du contrat choisi. Toute personne peut souscrire une complémentaire santé à titre individuel (contrat individuel) ou via son entreprise (contrat collectif) et potentiellement en faire bénéficier d'autres membres de sa famille, comme les enfants. Toute entreprise privée a l'obligation de proposer une couverture complémentaire santé collective à ses salariés. Les complémentaires se souscrivent auprès d'une compagnie d'assurance, d'une mutuelle, d'une institution de prévoyance ou d'un établissement bancaire. Il existe plusieurs types de garanties offertes par une complémentaire santé, qui varient selon le contrat souscrit :

— une complémentaire santé de base couvre généralement le ticket modérateur pour les actes médicaux simples, sans dépassement d'honoraires ;

— une couverture intermédiaire propose une meilleure prise en charge, notamment pour les consultations et soins médicaux, même en cas de dépassement d'honoraires, les médicaments sur ordonnance, les frais d'hospitalisation classiques ainsi que les soins et prothèses dentaires et les soins optiques de gamme moyenne ;

— une couverture optimale offre une meilleure couverture des dépassements d'honoraires et des soins spécialisés, comme les appareils auditifs, l'orthodontie pour adultes, ou encore les médecines douces ;

— la complémentaire santé solidaire permet de faire bénéficier les personnes à faibles ressources d'une complémentaire santé qui ne coûte rien ou bien moins d'un euro par jour et par personne.

La tarification des mutuelles repose à la fois sur le principe de mutualisation des risques et sur l'équité contributive. Ce système de tarification permet d'aligner les cotisations sur le profil de risque de chaque assuré, tout en maintenant une certaine solidarité entre les membres. Ainsi sont évités les phénomènes d'anti-sélection, où seuls les individus à haut risque souscriraient une assurance, mettant en péril l'équilibre du système. Les facteurs déterminants dans la tarification des mutuelles santé peuvent être :

— l'âge (complémentaire individuelle) ou la structure par âge (complémentaire collective) des assurés : plus l'assuré est âgé, plus ses cotisations seront élevées en raison d'une consommation de soins statistiquement plus importante ;

— le niveau de couverture sélectionné par l'assuré ;

— la durée des délais de carence : période durant laquelle l'assuré paye ses cotisations, mais n'active pas ses garanties ;

— le revenu ;

— la composition familiale et le nombre de bénéficiaires ;

— la zone géographique de l'assuré (complémentaire individuelle) ou de l'employeur (complémentaire collective) ;

— les coûts des soins locaux ;

— l'état de santé de l'assuré actuel et/ou passé (uniquement pour les contrats individuels).

Concernant l'utilisation de données de santé pour la tarification des contrats individuels, les exigences diffèrent selon les acteurs du marché. En effet, l'article L 112-1 du Code de la mutualité [3] **interdit aux mutuelles de tarifer selon l'état de santé**. En revanche, le Code des assurances [4] ne mentionne aucune interdiction sur le sujet pour les compagnies d'assurance ou les institutions de prévoyance. Toutefois, pour encourager des pratiques plus équitables, des incitations fiscales et sociales à ne pas utiliser l'état de santé des assurés ont été instaurées : l'article 63 de la loi de finances rectificative pour 2001, codifié aux paragraphes 15 et 16 de l'article 995 du Code général des impôts, institue une exonération en faveur des contrats d'assurance maladie dont les tarifs ne sont pas fixés selon l'état de santé de l'assuré. Ces contrats sont appelés contrats responsables. En dehors de ce cadre, l'utilisation des données de santé par les organismes pour tarifer des complémentaires santé individuelle est strictement encadrée par le Règlement général sur la protection des données (RGPD), entré en application en mai 2018. Ce règlement qualifie les données de santé de données « sensibles » (article 9) et en interdit le traitement, sauf exceptions strictement définies, notamment lorsque la personne concernée a donné son consentement explicite ou lorsque ce traitement est nécessaire à la gestion d'un contrat d'assurance dans un cadre légal et proportionné. Les assureurs, bancassureurs et instituts de prévoyance peuvent donc collecter des données de santé pour tarifer les complémentaires individuelles dans des conditions très encadrées et selon les principes rigoureux suivants :

— **Pertinence** : les questions doivent être en lien direct avec le risque assuré.

— **Proportionnalité** : interdiction de poser des questions excessives (habitudes sexuelles, origine ethnique, appartenance religieuse).

— **Consentement explicite** : l'assuré doit autoriser clairement l'utilisation de ses données, auquel cas l'assureur peut refuser de proposer un contrat.

— **Confidentialité** : seul un personnel habilité (souvent médical) peut accéder aux données nominatives.

— **Transparence** : l'assuré doit être informé de ses droits et de l'usage des données.

En cas de manquement de la part des organismes proposant des complémentaires santé individuelles, la Commission nationale de l'informatique et des libertés (CNIL) peut prononcer des sanctions lourdes.

2.2.3 Autres exemples d'utilisation des données de santé en France

Tout comme les Etats-Unis, la France utilise des scores de santé dans le but de mesurer la santé globale de sa population. Le **score EPICES** (Evaluation de la précarité et des inégalités de santé dans les Centres d'examens de santé) est l'un des plus reconnus. Construit en 1998 et financé par l'Assurance maladie, ce score socio-sanitaire est un indicateur individuel de précarité. Grâce à un questionnaire de 42 questions qui prend en compte plusieurs dimensions de la précarité : emploi, revenus, niveau d'étude, catégorie socio-professionnelle, logement, composition familiale, liens sociaux, difficultés financières, évènements de vie ou encore santé perçue. Il varie de 0 (absence de précarité) à 100 (maximum de précarité). Le score EPICES est performant pour détecter et quanti-

3. `https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006074067/`
4. `https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006073984/`

fier la précarité : le seuil de 30 est considéré comme le seuil de précarité selon EPICES [5].

Parmi les dispositifs français d'observation de la santé en population, le **réseau Sentinelles** constitue un outil historique et central de veille épidémiologique. Créé en 1984, ce réseau est animé conjointement par l'Inserm et Sorbonne Université. Il regroupe, sur la base du volontariat, un panel de médecins généralistes et pédiatres libéraux, appelés « médecins sentinelles », répartis sur l'ensemble du territoire métropolitain. Ces spécialistes fournissent au réseau les données des patients consultés. Le réseau collecte ainsi en continu des données de médecine générale et de pédiatrie, afin de constituer de grandes bases de données utiles à la recherche et à la surveillance sanitaire. Il contribue au suivi des maladies infectieuses, au développement d'outils de détection et de prévision épidémique, ainsi qu'à la mise en place d'études cliniques et épidémiologiques. Les données recueillies sont anonymisées et permettent non seulement de mieux comprendre les dynamiques sanitaires à l'échelle nationale, mais aussi de piloter plus efficacement les politiques de santé publique, notamment en cas d'épidémie saisonnière ou émergente. Le réseau ne construit pas de score de santé individuel, mais des indicateurs géographiques pour la surveillance.

Dans le secteur assurantiel, l'utilisation des données de santé, en dehors des complémentaires santé, reste autorisée sous certaines conditions, notamment en assurance emprunteur. Depuis le 1er juin 2022, l'article L113-2-1 du Code des assurances interdit toutefois de solliciter un questionnaire ou un examen médical lorsque :

— la part assurée par personne est inférieure ou égale à 200 000 euros ; et

— le remboursement total du prêt est prévu avant l'âge de 60 ans.

Cette évolution permet à des emprunteurs considérés à risque pour raisons de santé de ne plus supporter de surprimes ou d'exclusions de garanties. En dehors de ce cadre, un questionnaire médical demeure requis. Il peut inclure :

— des informations personnelles de l'assuré : taille, âge, poids ;

— des antécédents médicaux : maladies chroniques, traitements en cours, hospitalisations ;

— l'état de santé actuel ;

— le mode de vie : tabagisme, consommation d'alcool, pratique de sports à risque ; et

— l'historique familial : maladies héréditaires, décès prématurés dans la famille.

Les assureurs peuvent s'appuyer sur ces informations pour créer un score de santé individuel et ainsi ajuster les tarifs ou moduler les garanties, dans le respect du RGPD et du principe de proportionnalité.

Dans le cadre de ce mémoire, les données de tabagisme et de consommation d'alcool disponibles sur Internet ne correspondent ni à la maille géographique, ni à la période temporelle recherchées. Par conséquent, elles ne seront pas utilisées dans les modèles développés.

5. `https://www.nouvelle-aquitaine.ars.sante.fr/`

2.3 Justification de l'intégration des risques émergents dans la construction de scores de santé individuels

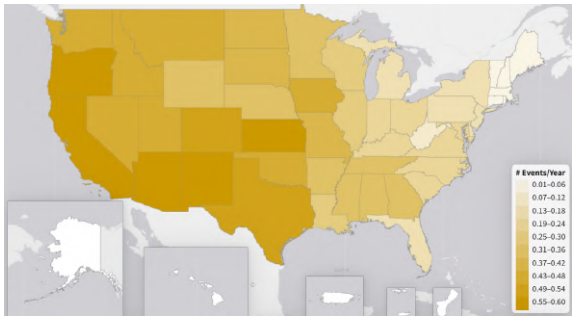
2.3.1 Les Américains inégalement exposés aux risques climatiques et à leurs impacts sur la santé

Bien que les Etats-Unis soient l'un des pays les plus riches du monde, se classant au premier rang en termes de PIB, les disparités en matière de santé demeurent largement répandues à travers le pays, comme le dépeint la figure 1.8. Ces disparités se manifestent sous diverses formes, touchant principalement la géographie, la « race » et l'ethnicité, ainsi que les inégalités de richesse ou de revenus. Ces disparités sont accentuées par le risque climatique qui affecte la population américaine de manière hétérogène.

2.3.1.1 Disparité géographique de l'exposition aux risques émergents

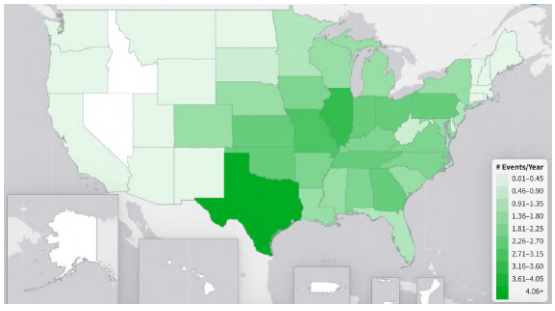
Les catastrophes naturelles ne touchent pas les Etats des Etats-Unis de la même manière. Le centre national des informations environnementales des Etats-Unis a recensé le nombre de catastrophes naturelles ayant touché le pays entre 2000 et 2024, ainsi que leur coût [12]. Les figures 2.1 et 2.2 représentent respectivement le nombre moyen de sécheresses et de tempêtes sévères par année et par Etats aux Etats-Unis entre 2000 et 2024. Les sécheresses touchent davantage l'ouest du pays : l'Oregon, l'Arizona, le Nouveau-Mexique, le Kansas et la Californie sont les Etats les plus touchés. Au contraire, les tempêtes s'abattent plus sur l'est du pays : le Missouri, l'Illinois et le Texas sont les Etats les plus touchés. Lorsque l'on s'intéresse aux inondations, ce sont les Etats du centre du pays qui sont les plus durement impactés. La Louisiane est l'Etat le plus touché par les inondations entre 2000 et 2024.

FIGURE 2.1 – Nombre moyen de sécheresses par année aux Etats-Unis par Etats entre 2000 et 2024 [12]



Note de lecture : les Etats de l'ouest sont les plus touchés par les sécheresses sur la période observée.

FIGURE 2.2 – Nombre moyen de tempêtes par année aux Etats-Unis par Etat entre 2000 et 2024 [12]

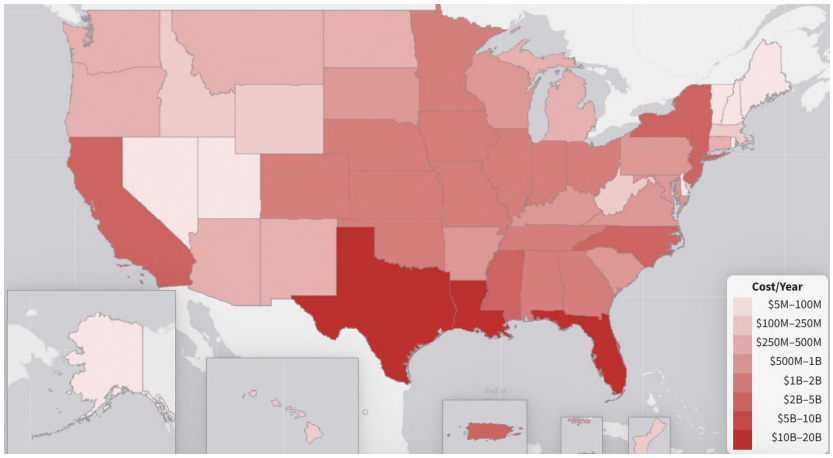


Note de lecture : les tempêtes extrêmes affectent majoritairement les Etats de l'est et du centre.

De plus, les conséquences économiques des catastrophes naturelles ne sont pas homogènes parmi les Etats. La figure 2.3 représente les coûts moyens par année et par Etat des catastrophes naturelles aux Etats-Unis entre 2000 et 2024. Le Texas, la Louisiane et la Floride sont les Etats ayant subi les coûts les plus élevés (entre 10 et 20 milliards de dollars) entre 2000 et 2024 en raison notamment des cyclones et des tempêtes sévères qui

ont touché ces Etats sur la période. Au contraire, le Nevada et l'Utah ont en moyenne dépensé entre 5 et 100 millions de dollars par année à la suite de catastrophes naturelles (sécheresses et feux de forêt majoritairement).

FIGURE 2.3 – Coûts moyens par année des catastrophes naturelles aux Etats-Unis par Etat entre 2000 et 2024 [12]



Note de lecture : le Texas, la Louisiane et la Floride sont les Etats ayant subi les coûts les plus élevés (entre 10 et 20 milliards de dollars par an) entre 2000 et 2024, principalement à cause des cyclones et tempêtes sévères.

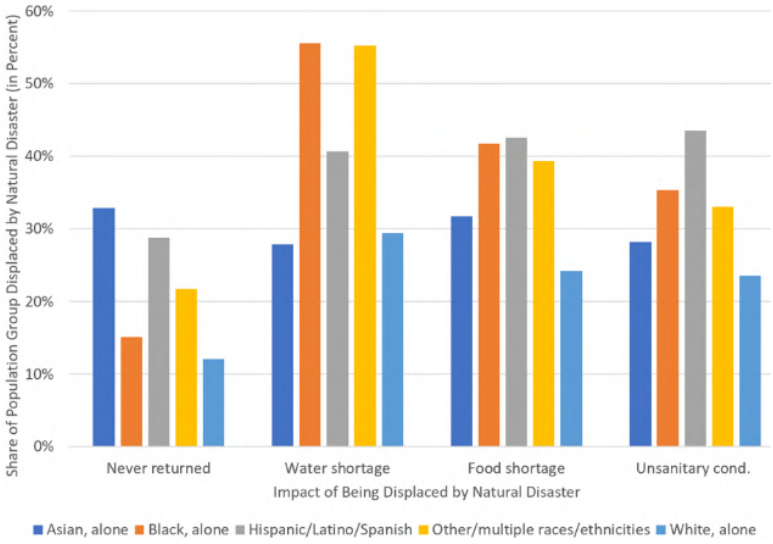
Ainsi, la population américaine n'est pas exposée de manière homogène au risque catastrophe et donc à son impact sur la santé : certains Etats sont plus touchés par des évènements extrêmes que d'autres, et subissent en conséquence des coûts plus élevés.

2.3.1.2 Disparité ethnique de l'exposition aux risques émergents

Depuis le début de la pandémie de covid-19, les Etats-Unis ont collecté des informations sur la situation financière, la santé et le logement des ménages avec son enquête *Household Pulse Survey* [13]. Cette enquête a notamment demandé aux individus sondés s'ils avaient été déplacés par une catastrophe naturelle. En moyenne, 1,5 % des personnes interrogées au cours des six premiers mois de 2023 ont déclaré avoir été déplacées au cours des douze mois précédents. Cependant, cette proportion était de 1,8 % pour les Latinos, 2,4 % pour les Afro-Américains et 2,3 % pour les personnes d'autres « races » et ethnicités ou de « races » et ethnicités multiples. En d'autres termes, s'ils avaient eu la même probabilité de déplacement que les familles caucasiennes, il y aurait eu 236 000 Afro-Américains déplacés en moins en juin 2023. Des centaines de milliers d'Afro-Américains et de Latinos supplémentaires, ainsi que des personnes d'autres « races » et ethnicités multiples, ont ressenti les effets extrêmes du changement climatique en raison des inégalités raciales et ethniques omniprésentes. Ce n'est pas seulement que la probabilité d'être touché par des catastrophes naturelles est plus grande, mais les conséquences d'être déplacé sont également plus graves. Par exemple, 21,8 % des Latinos, 28,7 % des personnes d'autres « races » et ethnicités multiples et 15,1 % des Afro-Américains ne sont jamais retournés chez eux, tandis que c'était le cas pour 12,1 % des Caucasiens (figure 2.4). De plus, les Afro-Américains et les Latinos ainsi que les adultes d'autres « races » et ethnicités multiples étaient beaucoup plus susceptibles de connaître des pénuries alimentaires, un manque d'eau et des conditions insalubres, parmi d'autres effets du déplacement, que les

adultes caucasiens (figure 2.4). Non seulement la probabilité d'être déplacé est presque deux fois plus élevée pour les Afro-Américains, mais le constat est similaire pour la probabilité de souffrir de pénuries alimentaires et de manque d'eau lorsqu'ils sont déplacés, par rapport aux Caucasiens. Comme c'est le cas dans de nombreux autres aspects de la vie quotidienne, les Afro-Américains, les Latinos et les personnes d'autres « races » et ethnicités multiples font face à des coûts supplémentaires massifs par rapport aux adultes caucasiens lorsqu'une catastrophe survient. Ainsi, les impacts du risque catastrophe sur la santé aux Etats-Unis ne sont pas homogènes au sein de la population lorsqu'on se focalise sur l'ethnicité des sinistrés.

FIGURE 2.4 – Impacts d'un déplacement géographique suite à une catastrophe naturelle par race/ethnicité aux Etats-Unis en 2023 [13]



Note de lecture : les Afro-Américains et les Latinos, ainsi que d'autres groupes ethniques, présentent une probabilité plus élevée d'être déplacés et de subir des conséquences graves (pénuries alimentaires, conditions insalubres) suite à une catastrophe naturelle, comparativement aux Caucasiens.

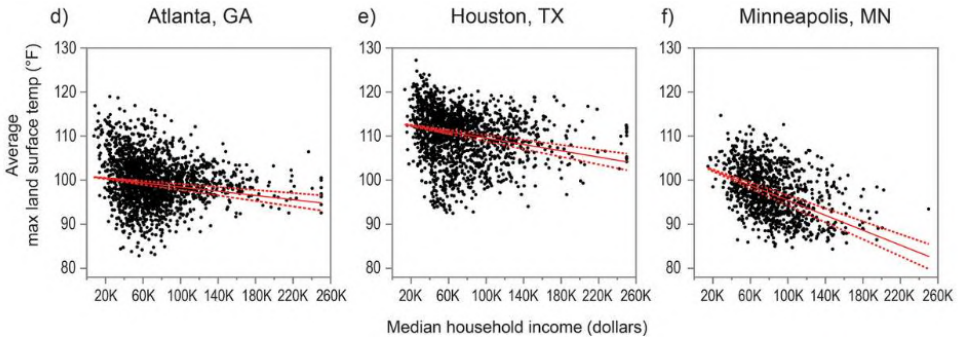
J.D. Sharp et ses co-auteurs ont analysé les données de mortalité de 1999 à 2018 aux Etats-Unis pour mesurer les disparités ethniques de l'exposition au risque catastrophe [14]. Les catastrophes naturelles et les conditions météorologiques extrêmes ont causé 27 335 décès aux Etats-Unis entre 1999 et 2018. Bien que les personnes caucasiennes non-hispaniques représentaient 68 % de la mortalité totale due aux catastrophes naturelles et aux conditions météorologiques extrêmes, le taux de mortalité pour 100 000 habitants parmi les personnes afro-américaines non-hispaniques était 1,87 fois plus élevé (0,71) que parmi les personnes caucasiennes non-hispaniques (0,38). Le ratio s'élève à 7,34 concernant les personnes amérindiennes ou autochtones de l'Alaska non-hispaniques (2,79). Pour tous les groupes raciaux et ethniques, l'exposition à la chaleur et au froid extrême sont les deux principales causes de mortalité dues aux catastrophes naturelles et aux conditions météorologiques extrêmes sur la période étudiée.

2.3.1.3 Effet revenu sur l'exposition aux risques émergents

Des chercheurs de l'ESRI se sont intéressés au lien entre les vagues de chaleur et le revenu des ménages dans les zones concernées [15]. Ils ont étudié les populations de trois

grandes villes des Etats-Unis, à savoir, Atlanta (GA), Houston (TX) et Minneapolis (MN). Ils ont mis en évidence de manière statistiquement significative que les ménages les plus riches résidaient dans les espaces urbains où la température à la surface de la Terre est moins élevée. Plus généralement, ils ont montré une corrélation négative entre le revenu des ménages et la température à la surface terrestre dans leur lieu de résidence. Ainsi, dans ces trois villes des Etats-Unis, les ménages les plus aisés sont moins sujets aux températures extrêmes, car ils résident dans des lieux davantage protégés et aménagés.

FIGURE 2.5 – Température à la surface de la terre en fonction des revenus des ménages de trois villes des Etats-Unis [15]



Note de lecture : les ménages les plus riches vivent dans des zones urbaines où la température à la surface de la Terre est plus basse, illustrant une protection accrue contre les vagues de chaleur.

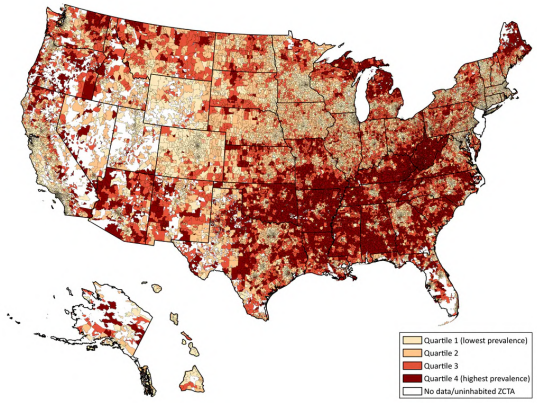
Traductions :
Average max land surface temp = température moyenne à la surface de la terre
Median household income (dollars) = Revenu médian des ménages (dollars)

En plus du lien entre risque catastrophe et revenu, il existe un lien entre les revenus et l'état de santé, accentué aux Etats-Unis par le système de santé non-universel mis en place. Benavidez et ses co-auteurs ont décidé d'étudier le lien entre la prévalence de maladies chroniques par Etats et des caractéristiques socio-économiques [16]. En tant qu'indicateur de la prévalence des maladies chroniques, ils ont créé un score en se basant sur les 10 maladies chroniques les plus répandues et les plus coûteuses aux Etats-Unis : obésité, hypertension, hypercholestérolémie, maladie coronarienne, bronchopneumopathie chronique obstructive, asthme, maladie rénale chronique, diabète, cancer et dépression. Ils ont ensuite créé trois catégories de prévalence des maladies chroniques en utilisant les seuils de quartiles (dans le 25e percentile inférieur, entre les 25e et 75e percentiles, et dans le 25e percentile supérieur) : les zones géographiques à la plus faible prévalence (score < 6), les zones à prévalence modérée (score de 7 à 13), et à la plus forte prévalence (score > 13). La figure 2.6 résume leurs résultats.

Les résultats de Benavidez et de ses co-auteurs ont été comparés à un rapport de santé nationale datant de 2023 étudiant notamment la richesse de la population américaine par Etats [17]. La figure 2.7 présente ainsi la proportion de la population américaine par Etats dont le revenu des ménages est inférieur au seuil de pauvreté entre 2020 et 2022. En comparant les figures 2.6 et 2.7, nous remarquons que les zones dans lesquelles les scores de prévalence des maladies chroniques sont les plus élevés sont aussi les zones contenant la plus grande proportion d'Américains vivant sous le seuil de pauvreté. Sans accès aux données, il est impossible d'établir un lien précis entre revenus des ménages

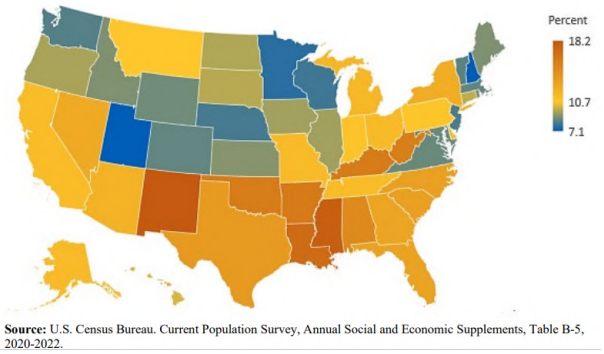
et scores de prévalence des maladies chroniques, mais la comparaison de ces deux cartes permet de corroborer les travaux de recherche qui ont déjà montré la corrélation négative entre état de santé et revenus [18, 19].

FIGURE 2.6 – Carte choroplèthe des Etats-Unis montrant la répartition géographique des scores de prévalence des maladies chroniques par quartile. [16]



Note de lecture : les zones avec un score de prévalence des maladies chroniques élevé se situent principalement dans le sud et certaines parties du Midwest.

FIGURE 2.7 – Proportion de la population américaine par Etats dont le revenu est inférieur au seuil de pauvreté, 2020-2022 [17]



Note de lecture : les Etats du sud et du sud-est présentent les plus fortes proportions de population sous le seuil de pauvreté.

Conclusions

Il apparaît ainsi indispensable d'**intégrer les risques climatiques et environnementaux dans la construction des scores de santé individuels**. En effet, cette section a illustré le fait que chaque individu est exposé différemment à ces risques en fonction de sa localisation géographique, de son origine ethnique, de son niveau de revenu et de ses conditions de vie. Ignorer ces facteurs reviendrait à sous-estimer ou à surestimer le risque réel encouru par chacun, et à perpétuer ou aggraver les inégalités de santé déjà existantes. L'intégration de ces variables permet ainsi de mieux refléter la réalité des expositions, d'améliorer la précision et l'équité des scores de santé, et de proposer des outils d'évaluation du risque plus justes, adaptés aux enjeux contemporains du secteur assurantiel.

Aussi, les résultats sur la mortalité ainsi que sur les populations déplacées en raison des évènements extrêmes aux Etats-Unis mettent en exergue la disparité ethnique de l'exposition au risque catastrophe aux Etats-Unis et de son impact sur la santé et sur la mortalité. Ils confirment l'inégalité ethnique de la couverture offerte par le système de santé américain illustrée dans la figure 1.8. Un point d'attention particulier sera donc porté dans ce mémoire à **l'évaluation de l'équité des modèles**, notamment vis-à-vis des différentes origines ethniques, afin d'identifier d'éventuels biais et de garantir une utilisation responsable des scores de santé dans le domaine assurantiel.

2.3.2 Conséquences du climat et de la pollution sur la santé des populations concernées

Il est aujourd'hui largement reconnu que les facteurs environnementaux, en particulier le climat et la pollution atmosphérique, exercent une influence significative sur la santé humaine. De nombreuses études épidémiologiques ont mis en évidence l'association entre l'exposition aux polluants atmosphériques (tels que les particules fines et les gaz irritants), les variations extrêmes de température (froid ou chaleur intense), ainsi que les événements météorologiques inhabituels (précipitations, tempêtes, inondations) et une augmentation des risques de maladies cardiovasculaires, respiratoires, infectieuses, voire de troubles mentaux. Aux Etats-Unis, ces expositions environnementales ne sont pas réparties de façon homogène au sein de la population et contribuent à des inégalités de santé, tant en termes d'incidence de maladies que d'accès aux soins. Le chapitre suivant proposera une revue de littérature détaillée sur les mécanismes par lesquels le climat et la pollution affectent la santé, ainsi que sur les principales variables environnementales à considérer pour la construction de scores de santé individuels, en lien direct avec la problématique de ce mémoire.

2.3.3 Prise en compte des données environnementales dans les scores de santé individuels

L'intégration des données environnementales dans les scores de santé individuels répond à plusieurs enjeux majeurs dans le secteur de l'assurance santé. Tout d'abord, elle permet de prendre en compte l'exposition individuelle aux polluants atmosphériques, à la chaleur, aux basses températures ou encore aux catastrophes naturelles (inondations, tempêtes, etc.) et donc d'affiner la tarification et la gestion du risque et d'offrir une assurance santé personnalisée, adaptée à la réalité de vie de chaque assuré. Par ailleurs, l'inclusion de données environnementales dans les scores de santé offre la possibilité d'anticiper l'impact des évolutions liées au changement climatique, tant sur la santé des assurés que sur la sinistralité du portefeuille. Les effets du climat et de la pollution sur la santé seront détaillés dans le chapitre 3, qui propose une revue de littérature détaillée sur le sujet. Enfin, certaines populations cumulent des expositions environnementales plus importantes que d'autres, comme le montre la section 2.3.1. Prendre en compte ces informations dans un score de santé garantit une approche plus juste et équitable dans l'évaluation du risque et la couverture des soins de santé.

2.4 Problématique du mémoire

Le premier chapitre de ce mémoire a présenté le système d'assurance santé américain, qui, malgré les réformes récentes, peine à couvrir équitablement l'ensemble des citoyens. Il a également introduit les trois bases de données fournies par *Milliman MedInsight* (bases « assurés », « souscriptions » et « sinistres ») ainsi que les enjeux économiques et opérationnels liés au traitement d'un grand volume de données sensibles qui vont contraindre les modèles implémentés et la taille des échantillons d'étude. Le chapitre 2 a décrit l'utilisation des scores de santé aux Etats-Unis ainsi que l'encadrement strict mis en place par l'ACA pour la tarification des assurances santé. Il a aussi mis en lumière l'exposition inégale des Américains aux risques climatiques et environnementaux et à leurs impacts

sur la santé. Les effets de la pollution atmosphérique et du climat (chaleur, froid, précipitations, humidité) sur la santé sont avérés : une revue de littérature détaillera toutefois plus précisément les conséquences sur la santé dans le chapitre 3.

PROBLÉMATIQUE :

Ce mémoire s'inscrit donc dans une optique de création de scores de santé, à partir des données d'assureurs santé américains fournies par *Milliman MedInsight* intégrant les dimensions du climat et de la pollution dans le but de quantifier leurs effets sur la santé. Les scores développés devront être fidèles à la réalité, c'est-à-dire qu'ils devront refléter la santé réelle des individus, et les modèles utilisés devront être facilement interprétables afin de mesurer directement l'impact des indicateurs climatiques et de pollution sur la santé. La création et la sélection des indicateurs climatiques et de pollution feront l'objet d'une démarche rigoureuse, guidée par une revue de littérature approfondie, présentée dans le chapitre suivant. Un point d'attention particulier sera également porté sur l'évaluation de l'équité des modèles, notamment vis-à-vis des différentes origines ethniques, afin d'identifier d'éventuels biais et de garantir une utilisation responsable de ces outils en assurance santé.

Les contraintes économiques et opérationnelles liées au traitement des trois bases de données à disposition imposent de restreindre le nombre de modèles implémentés et leur complexité ainsi que la taille de l'échantillon utilisé. Pour ce faire, l'étude sera restreinte à des sous-échantillons de données provenant de l'**Etat du Kentucky**, situé au centre-est des Etats-Unis. Les échantillons sélectionnés seront décrits dans la partie III de ce mémoire. Ce choix s'appuie, d'une part, sur les statistiques de la figure 1.9, qui mettent en exergue le faible taux de valeurs manquantes concernant l'ethnicité et la « race » des assurés de l'Etat du Kentucky, facilitant l'analyse de l'équité, et d'autre part, sur les caractéristiques climatiques de cet Etat. En effet, le climat modéré du Kentucky est caractérisé par des étés chauds, des hivers froids et des précipitations relativement fréquentes tout au long de l'année, permettant ainsi de limiter certains biais liés à des situations extrêmes.

Toutefois, le Kentucky n'est pas nécessairement représentatif de l'ensemble des Etats-Unis. En effet, le territoire américain est très étendu : 4 500 km séparent la côte atlantique à l'est et la côte pacifique à l'ouest et il faut parcourir 2 500 km pour relier le sud du Canada au nord du Mexique. Ainsi, le pays présente une grande diversité en termes de climat, de pollution, de démographie et de systèmes de santé locaux. Par exemple, le climat de type continental humide du Kentucky diffère du climat méditerranéen de la Californie. Ces différences peuvent influencer, limiter, voire annihiler certains effets des variables environnementales sur les scores de santé développés. Il est donc important de garder à l'esprit que les conclusions tirées dans ce mémoire seront difficilement généralisables car spécifiques aux caractéristiques du Kentucky.

Par ailleurs, pour limiter le nombre de calculs lourds et coûteux, deux types de modèles de score de santé seront implémentés dans ce mémoire à partir des sous-échantillons sélectionnés :

— un **score annuel** : score individuel de santé basé sur la prédiction du nombre annuel de conditions chroniques par assuré. Seules les données issues de la base « souscrip-

tions » (voir section 1.2.2) seront exploitées, enrichies de données climatiques et de pollution ;

— un **score mensuel** : score individuel de santé basé sur la prédiction des frais de santé mensuels de chaque assuré. Les données exploitées proviennent de la base « sinistres » (voir section 1.2.3) qui recense l'ensemble des événements médicaux donnant lieu à un remboursement ou à une facturation. Ces données seront complétées par des indicateurs climatiques et de pollution.

Plusieurs modèles statistiques et de *machine learning* (ML) vont permettre de créer ces scores pertinents et d'intégrer les risques émergents. Pour le score annuel, les **modèles linéaires généralisés** (GLM et GLMM) et le **modèle XGBoost** seront utilisés. Concernant le score mensuel, une **régression linéaire**, un GLM ainsi que deux modèles de Gradient Boosting seront implémentés. Ces modèles, décrits dans les chapitres 4 et 5, satisfont le critère d'interprétabilité des impacts sous-jacents et permettent de faire face aux enjeux opérationnels et financiers soulevés précédemment.

Chapitre 3

Impacts du climat et de la pollution sur la santé

Il est aujourd'hui bien établi que le climat et la pollution de l'air ont un impact à court et long terme sur la santé des individus exposés. Dans la continuité de la problématique présentée précédemment, ce chapitre propose une revue de littérature étendue pour comprendre les effets du climat et de la pollution sur la santé et mesurer les évolutions de ces phénomènes aux Etats-Unis. Cette analyse vise non seulement à mesurer l'ampleur de ces impacts, mais également à guider la sélection des variables environnementales les plus pertinentes pour la construction des scores de santé individuels développés dans ce mémoire.

3.1 Impacts de la pollution de l'air sur la santé

3.1.1 Effets des particules fines ($PM_{2.5}$)

Les particules fines sont une catégorie de particules en suspension dans l'air ambiant, d'un diamètre inférieur à 2,5 microns ($PM_{2.5}$). Les particules fines peuvent être d'origine naturelle ou anthropique. Les particules d'origine naturelle proviennent des éruptions volcaniques, de l'érosion, du vent ainsi que de la végétation (pollens). Les particules issues de l'activité humaine sont les conséquences du chauffage (principalement au bois), de la combustion de carburants, des centrales thermiques et des procédés industriels. Par leurs dimensions et leur persistance durable à l'état d'aérosols, les particules fines pénètrent en profondeur dans les voies respiratoires, augmentant ainsi le risque de maladies cardiaques et pulmonaires. Selon leur degré de concentration et de toxicité, elles peuvent provoquer à court ou long terme des pathologies qui vont de la simple inflammation aux affections plus graves.

Une exposition prolongée aux particules fines contribue au développement de maladies cardiologiques, pulmonaires, d'asthmes, de cancers ou encore de troubles de la reproduction.

Dans leur revue systématique, Krittanawong et al. ont montré qu'une augmentation de l'exposition à long terme aux $PM_{2.5}$ est associée à une hausse de la mortalité toutes causes confondues (HR [1] 1,08), des maladies cardiovasculaires (HR 1,09) et de la mortalité

1. Le *hazard ratio* (HR) est une mesure utilisée pour comparer le risque d'un événement entre deux

due à ces maladies (HR 1,12) [20]. Tsao et al. mettent en évidence que les concentrations de $PM_{2.5}$ sont fortement associées à des augmentations de la fréquence cardiaque, du débit cardiaque et de la vitesse de l'onde de pouls brachiale [2] (VOP) [21]. Les recherches menées au cours des dernières décennies montrent que chaque augmentation de 1 m/s de la VOP correspond à une augmentation de 12 à 14 % de la mortalité cardiovasculaire [22]. Une exposition accrue aux $PM_{2.5}$ entraîne une inflammation systémique, une coagulation et une détérioration de la fonction endothéliale [3], augmentant ainsi la charge de travail cardiaque et la résistance vasculaire. Ces effets peuvent accroître les risques de complications cardiovasculaires.

Les particules fines jouent également un rôle dans la prévalence des maladies pulmonaires et respiratoires. Tian et al. ont mis en relation les quantités de particules fines dans l'air de Pékin entre 2010 et 2012 et les hospitalisations pour asthme. Ils ont montré qu'une augmentation de la concentration de $PM_{2.5}$ de 10 µg/m^3 entraîne une hausse des visites hospitalières de 0,67 %, des consultations externes de 0,65 %, et des visites aux urgences de 0,49 % pour problèmes respiratoires (asthme) [23]. Kyung et al., dans une méta-analyse de 2015, ont montré qu'un accroissement de la concentration de $PM_{2.5}$ de 10 µg/m^3 multiplie le risque de cancer du poumon par 1,09 [24]. Le risque est plus élevé chez les fumeurs exposés aux particules fines. Une autre méta-analyse a mis en exergue une hausse de l'incidence de la pneumonie chez les enfants de 1,8 % lorsque la concentration de $PM_{2.5}$ augmente de 10 µg/m^3 [25]. Il a également été montré que cette même augmentation multiplie le taux de mortalité des maladies respiratoires chez des patients d'Amérique latine par 1,02 [26] et le taux de mortalité par fibrose pulmonaire d'un échantillon français par 7,93 [27].

Dans leur méta-analyse, Fu et al. montrent que la survenance de maladies neurologiques et psychiatriques est influencée par l'exposition aux particules fines. L'exposition à court et à long terme est corrélée à un risque accru d'accident vasculaire cérébral, de démence et de maladie de Parkinson. Aussi, être fortement exposé aux $PM_{2.5}$ augmente la probabilité d'anxiété et de dépression [28]. Il a aussi été prouvé sur 11 cohortes européennes que l'exposition aux $PM_{2.5}$ peut augmenter les risques de développer un cancer ou une maladie gastro-intestinale [29]. Ce risque est plus important chez les hommes que chez les femmes.

3.1.2 Effets des gaz polluants (NO_2, SO_2, O_3)

Les principaux gaz polluants présents dans l'air aujourd'hui sont l'ozone (O_3), le dioxyde d'azote (NO_2) et le dioxyde de soufre (SO_2). Tout comme les particules fines, ils peuvent être d'origine naturelle ou anthropique :

— **Sources anthropiques.** L'ozone troposphérique n'est pas émis directement dans l'air. Il se forme par des réactions photochimiques impliquant des oxydes d'azote et des composés organiques volatils en présence de la lumière solaire. Ces précurseurs

groupes. Il représente le rapport des taux de risque entre ces groupes à un instant donné.

2. La vitesse de l'onde de pouls est la vitesse à laquelle l'onde de pression artérielle se propage le long de l'arbre artériel. Elle est reconnue comme la référence pour l'évaluation de la rigidité des artères en routine clinique.

3. La fonction endothéliale désigne la capacité des cellules qui tapissent l'intérieur des vaisseaux sanguins à réguler la circulation sanguine, la coagulation et la réponse inflammatoire.

proviennent principalement des émissions des véhicules à moteur, des centrales électriques et des industries. Le NO_2 est principalement émis par la combustion des carburants fossiles dans les véhicules à moteur, les centrales électriques, les chaudières industrielles et les processus de chauffage domestique. Les moteurs diesel sont particulièrement importants comme source de NO_2. Le SO_2 est principalement émis par la combustion du charbon et du pétrole dans les centrales électriques, les raffineries de pétrole, les usines de ciment et d'autres industries lourdes.

— **Sources naturelles.** Bien que l'ozone soit principalement un polluant résultant de l'activité humaine, des éclairs et certaines réactions chimiques naturelles peuvent également contribuer à sa formation dans la basse atmosphère. Les éclairs et les feux de forêt peuvent également produire des oxydes d'azote : le SO_2 naturel provient majoritairement des volcans tandis que le NO_2 naturel est principalement issu des éclairs et des feux de forêt.

Maladies cardiaques

Une étude réalisée par Nicolle-Mir a suivi 57 053 habitants de Copenhague, d'Aarhus et de leurs banlieues, âgés de 50 à 64 ans, sur une période moyenne de 13 ans pour évaluer l'effet de l'exposition annuelle au dioxyde d'azote sur la santé. Les résultats montrent que les expositions au NO_2 augmentent le risque d'insuffisance cardiaque de 11 % pour les plus longues périodes d'exposition [30]. Des effets modificateurs du sexe, de la pression artérielle et du statut diabétique ont été observés, avec des risques plus élevés chez les hommes, les hypertendus et les diabétiques. Une méta-analyse de Zong et al. montre que l'exposition à court terme à l'ozone est associée à une baisse de certains indicateurs de variabilité de la fréquence cardiaque chez les adultes, ce qui suggère que le système nerveux autonome cardiaque pourrait être affecté après une exposition à l'ozone, contribuant à un risque accru de maladies cardiovasculaires [31].

Maladies respiratoires

Newell et al. ainsi que Badida et al. ont relevé à travers leurs méta-analyses l'effet négatif de l'exposition à court et long terme au dioxyde d'azote, au dioxyde de soufre et à l'ozone sur les maladies respiratoires [32, 33]. Les deux études s'accordent sur la classification de ces polluants en fonction de leur impact (NO_2>SO_2>O_3). Ces dernières ont également montré une augmentation du risque de maladie cardiovasculaire à la suite d'une forte exposition aux dioxydes de soufre et d'azote. Zhang et al. ainsi que Badida et al. ont observé une association positive entre l'exposition à court terme à l'ozone, aux dioxydes d'azote et de soufre et les hospitalisations pour maladies pulmonaires [33, 34] tandis que Huangfu et al. ont montré que l'exposition à long terme à l'ozone et au dioxyde d'azote augmente significativement les risques de mortalité due aux maladies pulmonaires obstructives chroniques et aux infections des voies respiratoires inférieures [35].

Maladies urinaires

Zhang et al. examinent la corrélation entre l'exposition à court terme au dioxyde d'azote (NO_2) et les hospitalisations pour maladie rénale chronique (MRC) à Lanzhou, en Chine. Les résultats montrent qu'une hausse de 10 µg/m^3 de NO_2 est associée à une multiplication par 1,034 du risque relatif d'hospitalisation pour MRC, avec des effets plus

prononcés chez les femmes, les personnes âgées de 65 ans et plus, et pendant la saison froide [36]. Il a également été démontré que l'exposition à des niveaux de NO_2 supérieurs à 0,015 ppm accroît significativement le risque de détérioration du taux de filtration glomérulaire chez les patients atteints de maladie rénale chronique, aggravant leur situation [37].

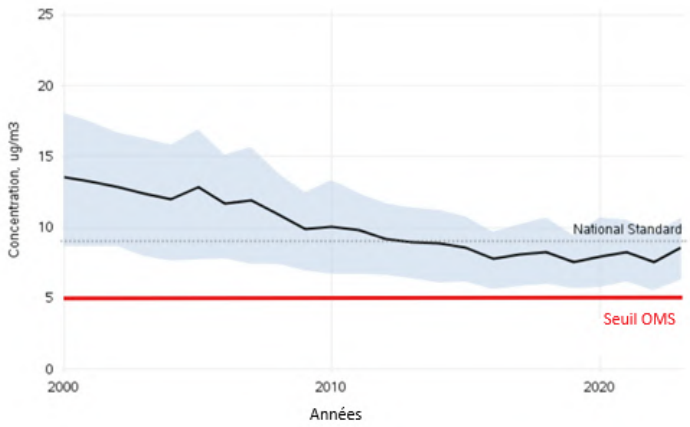
Maladies mentales

Enfin, des associations ont été trouvées entre l'exposition aux gaz polluants et le risque de développer des maladies psychiques et mentales. Oudin et al. ont montré qu'une hausse de la concentration moyenne annuelle en NO_2 de 10 µg/m³ multiplie le risque de prescription de psychotropes [4] chez l'enfant et l'adolescent par 1,09 [38]. Szyszkowicz et al. ont trouvé une association positive significative entre la concentration en NO_2 dans l'air dans la semaine et le risque de visites aux urgences pour dépression [39].

3.1.3 Seuils fixés par les organismes locaux et internationaux

La pollution atmosphérique constitue un enjeu de santé publique mondial. Selon l'Organisation mondial de la santé (OMS), chaque année, environ 7 millions de décès prématurés sont dus aux effets de la pollution de l'air, dont plus de 4 millions en lien avec l'air ambiant. Une revue exhaustive de la littérature des 15 dernières années a conduit l'OMS à publier de nouvelles lignes directrices pour la qualité de l'air ambiant plus exigeantes que celles publiées en 2005 : la concentration annuelle moyenne maximale acceptable de $PM_{2.5}$ est de 5 µg/m³ (ligne en rouge sur la figure 3.1), avec une limite de 15 µg/m³ par période de 24 heures. Aux Etats-Unis, la concentration moyenne annuelle de $PM_{2.5}$ était de 8,55 µg/m³ en 2023. Même si depuis 2011, la concentration nationale est inférieure à 10 µg/m³, la limite fixée par l'OMS n'a jamais été atteinte par les Etats-Unis.

FIGURE 3.1 – Concentration moyenne annuelle de $PM_{2.5}$ aux Etats-Unis entre 2000 et 2023 [40]



Note de lecture : en 2023, la concentration moyenne annuelle de $PM_{2.5}$ aux Etats-Unis était de 8,55 µg/m³ soit 0,05 µg/m³ sous le seuil préconisé par la NAAQS (pointillés noirs) mais supérieur au seuil fixé par l'OMS (ligne rouge) de 3,55 µg/m³.

4. Substance chimique qui altère les ressentis, la perception et les comportements. En agissant, sur le système nerveux, elle est utile dans le traitement de l'anxiété, la bipolarité, la dépression ou encore la schizophrénie.

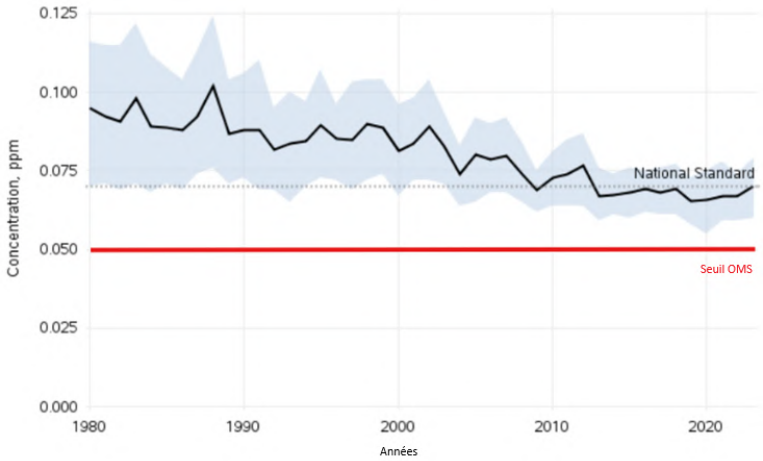
Les Normes Nationales de Qualité de l'Air Ambiant des Etats-Unis (NAAQS) sont des limites imposées sur les concentrations atmosphériques de six polluants responsables du smog, des pluies acides et d'autres dangers pour la santé. Ces normes ont été établies par l'Agence de Protection de l'Environnement des Etats-Unis (EPA) en vertu de la Loi sur la Qualité de l'Air et s'appliquent à l'air extérieur dans tout le pays [40]. Les six polluants de l'air concernés sont l'ozone, les particules fines, le plomb, le monoxyde de carbone (CO), les oxydes de soufre et les oxydes d'azote. Selon ces limites, la concentration annuelle de $PM_{2.5}$ ne doit pas excéder 9 µg/m^3 en moyenne sur 3 ans (ligne en pointillés sur la figure 3.1) avec une limite de 35 µg/m^3 par période de 24 heures, soit 20 µg/m^3 de plus que le seuil fixé par l'OMS. Ainsi, depuis 2013, la concentration de $PM_{2.5}$ dans l'air aux Etats-Unis est inférieure au seuil fixé par l'EPA.

La revue de littérature effectuée sur les particules fines conduit à exploiter les variables suivantes :

- le nombre de mois dans l'année où la concentration mensuelle moyenne de $PM_{2.5}$ dépasse 9 $\mu g/m^3$;
- les concentrations annuelles moyennes de $PM_{2.5}$ (moyenne effectuée sur les concentrations mensuelles) ; et
- les concentrations mensuelles maximales de $PM_{2.5}$ sur une année.

L'OMS fixe un seuil maximal à 0,05 ppm d'exposition à l'ozone à court terme (pendant 8 heures). Bien que la concentration moyenne annuelle d'ozone aux Etats-Unis ait baissé de 26 % depuis 1980, comme le montre la figure 3.2, les concentrations mesurées restent supérieures à la limite fixée par l'OMS. Lorsque l'on considère la moyenne de la concentration moyenne journalière maximale d'O_3 sur 8 heures au cours des six mois consécutifs où la concentration moyenne d'O_3 a été la plus élevée, cette limite descend à 0,03 ppm.

FIGURE 3.2 – Concentration d'ozone moyenne sur une période de 8 heures aux Etats-Unis de 1980 à 2023 [40]



Note de lecture : en 2023, la concentration d'ozone annuelle nationale était de 0,07 ppm soit exactement le taux préconisé par la NAAQS (pointillés noirs) mais supérieure au seuil fixé par l'OMS (ligne horizontale rouge).

Selon la NAAQS, la concentration d'ozone ne doit pas excéder 0,12 ppm en moyenne sur une heure et 0,07 ppm sur une période de 8 heures (seuil calculé en prenant la quatrième valeur la plus élevée de chaque année et en faisant la moyenne de ces valeurs sur une période de trois ans) [40]. Ainsi, depuis 2013, la concentration d'ozone dans l'air aux Etats-Unis est inférieure au seuil fixé par l'EPA.

Concernant le dioxyde d'azote, la NAAQS fixe une limite annuelle à 0,053 ppm (100 µg/m^3) et une limite à 0,1 ppm (188 µg/m^3) sur une période d'une heure. Cette dernière correspond au 98$^{\text{ème}}$ percentile des valeurs maximales quotidiennes sur une heure, calculée en moyenne sur une période de trois ans. L'OMS fixe des seuils moins optimistes : 10 µg/m^3 pour la moyenne annuelle.

Les données de pollution aux Etats-Unis ne sont disponibles que **mensuellement** pour la maille recherchée (ZIP3). Ainsi, la revue de littérature effectuée sur l'effet des gaz polluants sur la santé conduit à sélectionner les variables suivantes :

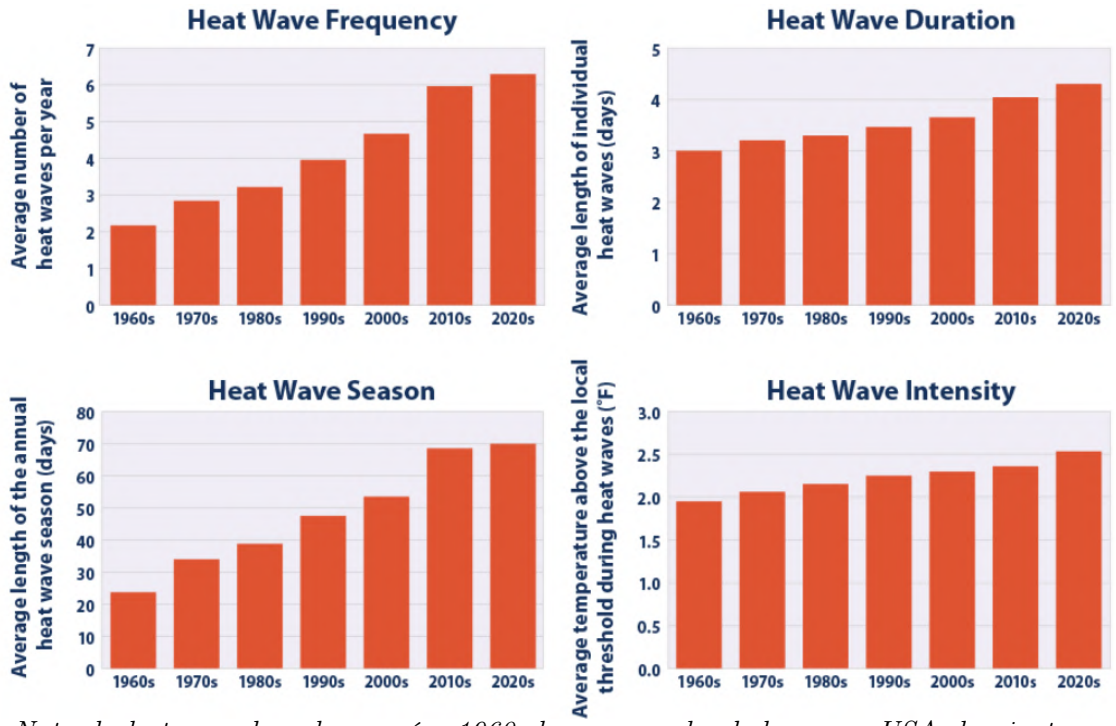
- le nombre de jours dans l'année où la concentration mensuelle moyenne de NO_2 dépasse 10 $\mu g/m^3$;
- le nombre de jours dans l'année où la concentration mensuelle moyenne de O_3 dépasse 60 $\mu g/m^3$;
- les concentrations annuelles moyennes de O_3 et de NO_2 (moyenne effectuée sur les concentrations mensuelles) ; et
- les concentrations mensuelles maximales de O_3 et de NO_2 sur une année.

3.2 Effets de la chaleur sur la santé

3.2.1 Définition et évolutions des vagues de chaleur aux Etats-Unis

L'Agence de protection de l'environnement des Etats-Unis (EPA) définit une vague de chaleur comme une période de deux jours ou plus consécutifs durant laquelle la température apparente minimale quotidienne dans une ville particulière dépasse le 85$^{\text{ème}}$ percentile des températures historiques de juillet et août (1981-2010) pour cette ville. Par exemple, dans la plupart des villes d'Amérique du Nord, une vague de chaleur est définie par un minimum de deux jours consécutifs avec une température maximale de 32 °C. D'après l'EPA, qui recense notamment les évènements climatiques extrêmes aux Etats-Unis, le nombre annuel de vagues de chaleur ne cesse d'augmenter depuis les années 1960 [41]. Il en est de même pour la durée des vagues de chaleur. Si dans les années 1960, les vagues de chaleur duraient en moyenne 3 jours, elles duraient en moyenne 4 jours dans les années 2010. Enfin, leur intensité ne cesse également d'augmenter. La différence moyenne entre les températures mesurées pendant les vagues de chaleur et le seuil local s'accroît de décennies en décennies depuis 1960 (voir figure 3.3).

FIGURE 3.3 – Caractéristiques des vagues de chaleur par décennies aux Etats-Unis de 1961 à 2023 [41]



Note de lecture : dans les années 1960, les vagues de chaleur aux USA duraient en moyenne 2 jours contre 4 jours dans les années 2010.

3.2.2 Mécanismes de l'impact de la chaleur sur le corps humain et conséquences sur la santé

Les épisodes de chaleur intense ont des conséquences importantes sur la santé des individus concernés. En effet, la quantité de chaleur stockée dans le corps humain est déterminée par :

— une absorption de chaleur externe provenant de l'environnement ;

— une incapacité à éliminer la chaleur générée en interne par les processus métaboliques en raison du stress thermique environnemental (par exemple, haute température, forte humidité, faible vent, rayonnement thermique élevé) ; et

— des vêtements créant une barrière à la perte de chaleur.

L'incapacité du corps à réguler la température interne et à éliminer le gain de chaleur dans de telles conditions augmente le risque de coup de chaleur et d'épuisement dû à la chaleur. La pression exercée sur le corps lorsqu'il tente de se refroidir stresse également le cœur et les reins. En conséquence, les extrêmes de chaleur peuvent aggraver les risques sanitaires liés aux conditions chroniques (cardiovasculaires, mentales, respiratoires et liées au diabète) et provoquer des lésions rénales aiguës [42]. En effet, Zhiwei et al. (2024) ont montré qu'une augmentation de la température moyenne est associée à une hausse des probabilités d'hospitalisation, surtout en présence de maladies chroniques, et ce, particulièrement chez les personnes âgées de plus de 65 ans, les hommes et les non-autochtones. Par exemple, chez les personnes âgées avec 0, 1, 2 ou 3 maladies chroniques, les risques

relatifs d'hospitalisation en raison de l'exposition à la chaleur ambiante étaient respectivement de 1,00, 1,06, 1,08 et 1,13. Parmi les maladies chroniques, les maladies rénales chroniques, les maladies cardiovasculaires et l'asthme, qu'elles soient présentes seules, ensemble ou en combinaison avec d'autres maladies, présentaient les probabilités les plus élevées d'hospitalisation en cas d'exposition à la chaleur ambiante [43].

Les périodes de fortes températures augmentent également le nombre d'hospitalisations pour des troubles mentaux [44]. Les décès et les hospitalisations déclenchés par des températures extrêmement élevées se produisent rapidement (le jour même et les jours suivants), ce qui signifie que les interventions doivent également être rapides lorsqu'une alerte de chaleur est émise [43].

A noter également que l'ampleur et la nature des impacts de la chaleur sur la santé dépendent du moment, de l'intensité et de la durée d'un épisode de chaleur, ainsi que du niveau d'acclimatation et d'adaptabilité de la population locale, des infrastructures et des institutions au climat dominant.

En accord avec la revue de littérature, il est pertinent de prendre en compte le nombre mensuel et annuel de vagues de chaleur, en prenant comme définition de vague de chaleur une période de 2 jours consécutifs où la température maximale dépasse les 32 °C. Des indicateurs mensuels de températures minimales, moyennes et maximales seront également créés.

3.3 Effets du froid sur la santé

Une vague de froid, également connue sous le nom de coup de froid, est un phénomène météorologique caractérisé par une chute rapide de la température de l'air. Selon le National Weather Service des Etats-Unis, une vague de froid se produit lorsqu'une baisse rapide de température sur une période de 24 heures nécessite une protection accrue pour l'agriculture, l'industrie, le commerce et les activités sociales. Les critères précis incluent le taux de chute de la température et la température minimale atteinte, qui varient selon la région géographique et la période de l'année. Aux Etats-Unis, une période de froid est définie lorsque la température nationale journalière maximale descend en dessous de $-7\,°C$ pendant plus de deux jours consécutifs. Une vague de froid suffisamment intense et prolongée peut être classée comme une « éruption d'air froid ». Durant la vague de froid nord-américaine de janvier 2024, Dillon (Montana) a atteint un record avec une température minimale de $-41\,°C$ tandis qu'une température de $-43\,°C$ a été mesurée à Bozeman (Montana).

L'exposition à de telles températures fait entrer en jeu le mécanisme thermorégulateur qu'est la thermogenèse, à savoir la production de chaleur. Cette production de chaleur découle de trois mécanismes : l'augmentation de l'activité musculaire via des frissons par exemple, la libération d'hormones et enfin la destruction des graisses via la lipolyse. Les faibles températures ont un impact direct sur la santé qui s'observe dans les 3 à 21 jours suivant l'exposition au froid, selon Santé publique France.

3.3.1 Impact indirect du froid sur la mortalité

Au-delà de son impact direct sur la mortalité, l'exposition au froid peut être indirectement liée à la mortalité à travers les maladies cardiovasculaires, respiratoires et les infections. Dans leur méta-analyse de 9 articles, Ryti et al. (2016) font l'état d'un impact significatif des vagues de froid sur la mortalité cardiovasculaire et respiratoire, entre 1979 à 2009 pour la Chine, la Russie, Taiwan, la République Tchèque et les Pays-Bas [45].

3.3.2 Conséquences sur les maladies cardiovasculaires et cérébro-vasculaires

Les mécanismes de thermogenèse entraînent la vasoconstriction, c'est-à-dire la contraction des vaisseaux sanguins, qui à son tour augmente la pression artérielle et l'activité cardiaque, aboutissant à une concentration plus élevée que la normale en protéines et en globules rouges. Cette hémoconcentration augmente la viscosité du sang et les risques de thrombose. En effet, Neild et al. (1994) et Keatinge et al. (1984) montrent que l'hémoconcentration est un facteur de risque pour les thromboses artérielles [46, 47]. Pitsavos et al. (2004) identifient une augmentation de 5 % d'admissions à l'hôpital pour des maladies coronariennes pour chaque baisse de la température quotidienne moyenne de 1 °C dans la région d'Athènes entre janvier 2001 et août 2002 [48]. Tsao et al. (2023) révèlent que les basses températures en hiver ont des effets négatifs sur la santé cardiovasculaire. En hiver, ils ont observé une augmentation significative des composants de la pression artérielle et de la résistance vasculaire systémique (SVR), accompagnée d'une diminution de la fréquence cardiaque, du volume systolique [5] et du débit cardiaque. Le froid provoque une vasoconstriction périphérique pour conserver la chaleur, augmentant ainsi la pression artérielle et la résistance vasculaire, ce qui peut entraîner des évènements cardiovasculaires [21].

3.3.3 Effets sur les maladies respiratoires

Mäkinen et al. (2009), estiment sur la base d'une population de 892 hommes militaires finlandais entre juillet 2004 et janvier 2006 qu'une baisse de 1 °C augmente respectivement le risque d'infection des voies respiratoires supérieures, de rhume, de pharyngite et d'infection des voies respiratoires inférieures de 4,3 %, 2,1 %, 2,8 % et 2,1 % [49]. Le froid affecte le système immunitaire [50]. En particulier, il fragilise l'épithélium trachéo-bronchique, diminuant la résistance du système immunitaire et permettant ainsi le développement d'infections broncho-pulmonaires [51].

3.3.4 Impacts du froid sur les infections virales

Les études épidémiologiques suggèrent que les basses températures jouent un rôle important dans l'efficacité de la transmission des virus. Jaakkola et al. (2014), montrent qu'une baisse de 1 °C de la température augmente le risque d'occurrence de la grippe de 11 % [52]. Dans une étude sur 20 062 cas entre octobre 2010 et juillet 2013 en Suède, Sundell et al. (2016) identifient que les basses températures ont été associées à l'incidence

5. Le volume systolique est la quantité de sang éjectée par le ventricule gauche à chaque contraction. Une diminution du volume systolique peut conduire à une insuffisance cardiaque.

hebdomadaire de la grippe, du virus respiratoire syncytial, du métapneumovirus [6], du bocavirus [7] et de l'adénovirus [8], tandis que l'incidence du rhinovirus [9] humain et de l'entérovirus [10] était indépendante de la température. Ils relèvent par ailleurs que la baisse hebdomadaire de la température moyenne par rapport à la semaine précédente est fortement associée à l'incidence de la grippe enregistrée la semaine suivante [53].

Bien que certains virus se conservent mieux par temps froid, le froid peut également être un facteur de confusion. En effet, les températures basses influencent les comportements humains. Notamment, le froid incite les gens à se rassembler dans des espaces clos et à augmenter l'utilisation du chauffage. Lofgren et al. (2007) ont ainsi constaté que les épidémies de grippe en Europe sont favorisées par le fait que les gens passent plus de temps à l'intérieur en hiver. De plus, le chauffage entraîne une recirculation continue de l'air avec un taux d'humidité très faible, créant des conditions idéales pour la persistance des particules virales dans l'environnement [54].

En accord avec la revue de littérature précédente et le seuil fixé par l'EPA dans la définition d'une vague de froid, il est pertinent de prendre en compte le nombre mensuel et annuel de vague de froid sur la période hivernale, en prenant comme définition de vague de froid une période de 2 jours consécutifs où la température maximale n'excède pas -7 °C.

3.4 Impacts des précipitations sur la santé

Les précipitations jouent un rôle clé dans la dynamique des maladies, influençant à la fois la transmission des agents pathogènes et l'exposition des populations. Cette section explore les liens entre les précipitations extrêmes et divers impacts sanitaires, notamment l'augmentation des maladies infectieuses, la propagation des vecteurs de maladie et l'aggravation des pathologies respiratoires. Une meilleure compréhension de ces relations est essentielle pour anticiper les risques et adapter les politiques de santé publique face aux changements climatiques.

3.4.1 Augmentation des maladies infectieuses

La relation entre les précipitations extrêmes et les maladies d'origine hydrique (provoquant notamment diarrhée, douleurs abdominales, nausées et vomissements) aux Etats-Unis est bien établie dans la littérature, avec des associations significatives dans de nom-

6. Le métapneumovirus est un virus à ARN qui peut provoquer des infections respiratoires semblables à la grippe et des symptômes tels que la toux, la fièvre, le nez bouché ou un essoufflement. Le virus est responsable de pneumopathies ou d'infections des voies aériennes supérieures.

7. Le bocavirus humain est un type de virus qui peut provoquer des maladies respiratoires, en particulier chez les enfants et les nourrissons.

8. Les adénovirus sont des virus à ADN à l'origine de maladies ORL et respiratoires mais aussi de gastro-entérites ou de conjonctivites.

9. Les rhinovirus sont des virus qui contiennent de l'acide ribonucléique. Ils sont le plus souvent les agents responsables des infections virales respiratoires aiguës. Les rhinovirus sont les coupables de la rhinite, de la pharyngite et de la bronchite.

10. Les entérovirus sont des virus à ARN impliqués dans de nombreuses pathologies humaines, en particulier les méningites estivales de l'enfant.

breux domaines d'étude. Un effet sanitaire commun résultant du contact avec de l'eau contaminée est la maladie gastro-intestinale aiguë (GI). En effet, les systèmes d'eau potable vieillissants augmentent la probabilité de maladies d'origine hydrique, compte tenu du stress qu'ils subissent déjà en conditions normales de fonctionnement [55]. Des études traitant des facteurs de risque de maladies GI après des événements de précipitations extrêmes ont trouvé que ces maladies sont souvent liées à la contamination des approvisionnements en eau potable [56]. La plus grande épidémie de maladie d'origine hydrique documentée dans l'histoire des Etats-Unis a rendu malades plus de 400 000 personnes à Milwaukee en 1993 en raison d'un événement de précipitations importantes qui a compromis la source d'eau de surface et a permis aux oocystes de Cryptosporidium de passer à travers l'un des systèmes de filtration des stations de traitement de l'eau potable [57].

3.4.2 Propagation de vecteurs de maladie

La saisonnalité et la quantité de précipitations dans une région peuvent fortement influencer la disponibilité des sites de reproduction pour les moustiques et d'autres espèces ayant des stades de vie aquatiques. Pour les maladies qui sont à la fois vectorielles et zoonotiques (c'est-à-dire ayant des réservoirs vertébrés autres que les humains), le climat peut affecter la distribution et l'abondance des espèces hôtes, ce qui peut, à son tour, affecter la dynamique des populations de vecteurs et la transmission des maladies.

Une étude de Parmenter et al. (1999) examine la relation entre les précipitations et les cas de peste humaine au Nouveau-Mexique de 1948 à 1996. En analysant 215 cas humains de peste en corrélation avec les précipitations, les chercheurs ont constaté que les infections étaient 60 % plus fréquentes après des périodes de précipitations hivernales et printanières (octobre à mai) supérieures à la moyenne. Les résultats significatifs ont été obtenus uniquement à l'échelle locale, où les précipitations étaient en moyenne 113 % supérieures à la normale durant les années de peste [58].

3.4.3 Augmentation des maladies respiratoires

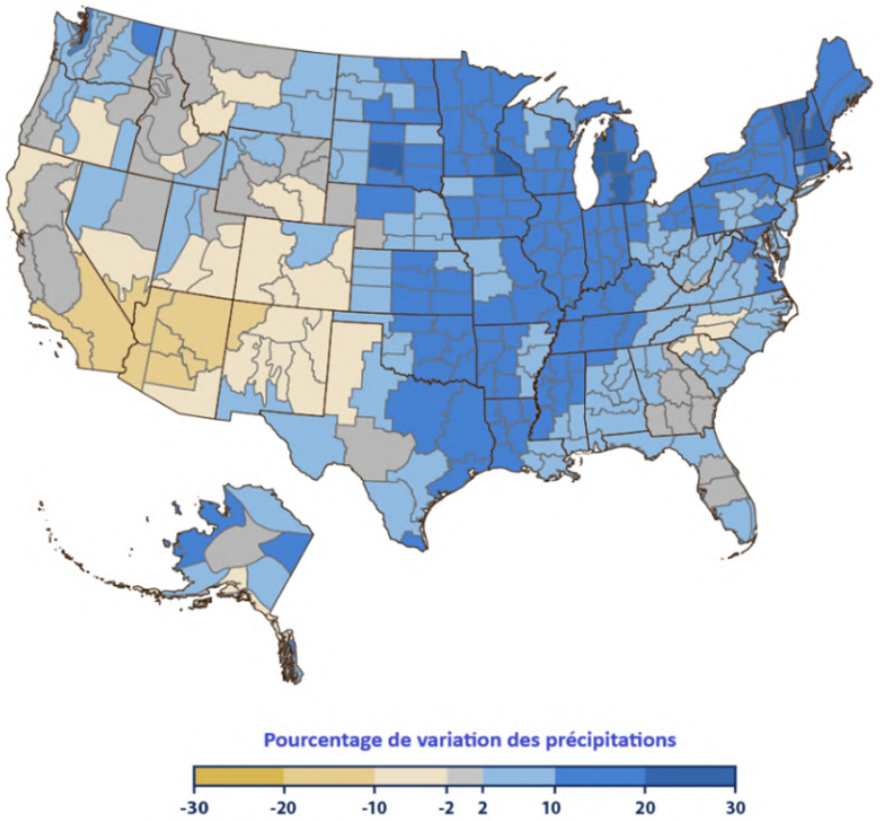
Les précipitations sont positivement associées aux niveaux d'oxyde nitrique exhalé fractionné (FeNO), un marqueur d'inflammation des voies respiratoires. Nassikas NJ et al. ont en effet montré que chaque augmentation de 2 mm de la moyenne mobile sur 7 jours des précipitations aux Etats-Unis était associée à une augmentation de 4,0 % des niveaux de FeNO chez les adolescents. Par ailleurs, il a été observé que le statut asthmatique modifiait cet effet : les précipitations sont associées à une capacité vitale forcée plus faible et à des niveaux de FeNO plus élevés chez les adolescents asthmatiques. Ces associations ne sont pas expliquées par l'humidité relative ou l'exposition à la pollution de l'air (variables de contrôles) [59]. Plusieurs études ont également trouvé une relation entre les orages et l'augmentation des admissions à l'hôpital pour asthme dues au pollen en Europe [60] et aux Etats-Unis [61].

3.4.4 Conclusion

La figure 3.4 présente la variation des précipitations aux Etats-Unis entre 1901 et 2023 [62]. Si la moitié ouest des Etats-Unis subit de moins en moins de précipitations depuis le début du XX$^{\text{ème}}$ siècle, la moitié est, quant à elle, subit une augmentation des

précipitations comprises entre 2 % et 30 % en fonction des comtés. Or, cette revue de littérature montre qu'une hausse des précipitations favorise les risques de développer des maladies respiratoires, vectorielles ou encore infectieuses.

FIGURE 3.4 – Variation des précipitations aux Etats-Unis, 1901–2023 [62]



Note de lecture : en Louisiane, entre 1901 et 2023, les précipitations ont augmenté en moyenne entre 10 % et 20 %.

Les précipitations ayant un impact significatif à long-terme sur la santé, il est pertinent de prendre en compte les précipitations journalières mensuelles et annuelles, captant ainsi les orages, les tempêtes et les fortes pluies.

3.5 Rôle de l'humidité dans la transmission et l'aggravation des pathologies

L'humidité dans l'air, souvent exprimée en pourcentage d'humidité relative, représente la quantité de vapeur d'eau présente dans l'atmosphère. Cette grandeur varie en fonction de plusieurs facteurs, notamment la température, la pression atmosphérique et la présence de sources d'eau à proximité. En général, l'air chaud peut contenir plus de vapeur d'eau que l'air froid, ce qui explique pourquoi les régions tropicales ont souvent une humidité élevée.

Tsao et al. (2023) ont montré que l'humidité relative ambiante a un impact significatif sur les fonctions cardiovasculaires. Une augmentation de l'humidité relative de 1 % est associée à une diminution de la pression artérielle diastolique de 0,5 mmHg ainsi qu'à

une augmentation de la fréquence cardiaque de 0,92 battements par minute. L'humidité élevée peut ainsi réduire l'efficacité du corps à dissiper la chaleur métabolique, ce qui entraîne une vasodilatation et une diminution de la pression artérielle, mais oblige le cœur à travailler plus dur pour maintenir la température corporelle, augmentant ainsi la charge de travail cardiovasculaire [21].

Jaakkola et al. (2014) montrent qu'une baisse de 0,5 g par m^3 de l'humidité absolue dans l'air augmente le risque d'occurrence de la grippe de 58 % [52]. Par ailleurs, Mäkinen et al. (2009) estiment qu'une baisse de 1 g/m^3 d'humidité absolue augmente respectivement le risque d'infection des voies respiratoires supérieures et de pharyngite de 10 % et 10,8 % [49]. L'asthme chez l'enfant et d'autres maladies allergiques pédiatriques sont des conséquences importantes de l'humidité. En Grèce, une étude a montré que l'humidité relative était une variable météorologique impliquée pour les jeunes enfants asthmatiques : une augmentation de 10 % de l'humidité est liée à une augmentation de 31 % de la probabilité d'avoir une admission pour asthme chez les enfants [63]. Une étude rétrospective en Amérique a indiqué qu'une augmentation intra-journalière de 10 % de l'humidité un ou deux jours avant l'admission était associée à environ une visite supplémentaire aux urgences pour asthme [64].

Timmermans et al. (2015) ont trouvé des associations significatives entre les douleurs articulaires et l'humidité moyenne quotidienne (beta = 0,004, $p < 0,01$) ainsi que l'humidité moyenne sur 3 jours (beta = 0,004, p = 0,01). Un effet d'interaction significatif a été observé entre l'humidité moyenne quotidienne et la température sur les douleurs articulaires. L'effet de l'humidité sur la douleur était plus fort dans des conditions météorologiques relativement froides [65].

Certaines maladies gastro-intestinales et urologiques sont inversement liées à l'humidité relative. Une étude menée au Japon a montré des relations négatives entre l'humidité relative et les gastro-entérites infectieuses [66]. En effet, dans des conditions de faible humidité relative, les virus conservent davantage leur intégrité. Aussi, plusieurs études ont montré qu'une baisse de l'humidité relative favorisait les calculs rénaux [67, 68].

Cette revue de littérature montre l'impact significatif de l'humidité sur la santé. Cependant, cette variable est fortement corrélée aux précipitations : plus l'humidité dans l'air est élevée, plus la vapeur d'eau est importante, et plus il est probable qu'il y ait de la pluie. Inversement, l'humidité peut augmenter avec une élévation de la température extérieure et/ou avec la quantité de pluie [69].

Ainsi, l'humidité ne sera pas prise en compte dans le cadre de cette étude, pour mesurer l'impact des risques émergents sur la santé des assurés américains.

Deuxième partie

Cadre théorique : modèles pour la construction de scores de santé

Chapitre 4

Modèles statistiques pour la construction de scores de santé

L'élaboration de scores de santé individuels ou collectifs repose sur la capacité à prédire de manière fiable une variable de santé « cible » à l'aide de différentes variables explicatives (facteurs individuels, environnementaux ou cliniques) et des indicateurs de santé. Ce chapitre présente les principaux modèles statistiques utilisés dans le cadre de ce mémoire pour la construction des scores de santé annuels et mensuels : la régression linéaire, les modèles linéaires généralisés (GLM), ainsi que les modèles linéaires mixtes généralisés (GLMM). L'ensemble de ces méthodes constitue une boîte à outils essentielle pour modéliser, prédire et interpréter les scores de santé à partir de données complexes et hétérogènes. Un point d'attention sera accordé à l'interprétabilité de ces modèles.

4.1 Régression linéaire

La régression linéaire est couramment utilisée en actuariat en raison de sa simplicité d'implémentation et d'interprétation. Ce modèle se place dans le cas d'une variable quantitative à expliquer, à partir de variables explicatives (réelles ou catégorielles). Ce chapitre vise à décrire plus en détail ce modèle.

4.1.1 Description du modèle

Soit Y une variable quantitative que l'on cherche à prédire, et $\mathbf{X} = (x_l)_{1\leq l\leq d}$ des variables explicatives, supposées quantitatives pour le moment. L'objectif est de trouver des coefficients constants $\boldsymbol{\beta} = (\beta_l)_{0\leq l\leq d}$ de sorte que $\beta_0 + \sum_{l=1}^{d} \beta_l x_l$ approxime au mieux la variable Y. Ainsi, le modèle se présente sous la forme :

$$Y = \beta_0 + \sum_{l=1}^{d} \beta_l x_l + \varepsilon$$

où ε est l'erreur commise par le modèle, c'est-à-dire la différence entre la prédiction et la véritable valeur prise par Y. On suppose généralement que ces erreurs suivent une loi normale centrée, de variance constante (homoscédastique) σ^2. Soient $Y = (y^{(i)})_{1\leq i\leq n}$ et $\mathbf{X} = (x_l^{(i)})_{1\leq i\leq n, 1\leq l\leq d}$ nos données où n représente le nombre d'individus. Sous forme matricielle, la régression linéaire se présente sous la forme :

$$Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon.$$

On trouve les coefficients $(\beta_l)_{0\leq l\leq d}$ du modèle par la méthode des moindres carrés en minimisant l'erreur commise par le modèle :

$$\boldsymbol{\beta} = \arg\min_{(\beta_l)_{0\leq l\leq d}} \sum_{i=1}^{n} \left(y^{(i)} - \beta_0 - \sum_{l=1}^{d} \beta_l x_l^{(i)} \right)^2.$$

Ce problème d'optimisation a une solution analytique simple. On trouve après écriture des conditions du premier ordre :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y.$$

La prédiction faite par le modèle linéaire pour $y^{(i)}$ est notée par la suite :

$$\hat{y}^{(i)} = \beta_0 + \sum_{l=1}^{d} \beta_l x_l^{(i)}, i \in \{1, ..., n\}.$$

4.1.2 Interprétabilité du modèle de régression linéaire

La régression linéaire est un modèle facilement interprétable. L'interprétation des coefficients estimés dépend du type des variables explicatives choisies (qualitatives binaires ou numériques). Prenons un exemple concret pour illustrer l'interprétabilité des sorties du modèle : prédire le coût d'une visite à l'hôpital pour un individu à l'aide d'une variable binaire, « être atteint d'une maladie chronique » (coefficient associé β_1) et d'une variable numérique, l'âge (coefficient associé β_2). Alors, pour la variable binaire, si l'individu a une maladie chronique, le coût estimé de sa visite à l'hôpital évoluera de β_2 (à la hausse ou à la baisse en fonction du signe de β_2) par rapport à l'estimation pour la même personne sans maladie chronique. Pour la variable numérique, si l'âge de l'individu augmente d'une année, alors l'estimation du coût de sa visite évoluera de β_1 euros. L'*intercept* β_0 représente la valeur de Y estimée par notre modèle lorsque toutes les variables explicatives sont à 0. Cette valeur n'a pas d'autre interprétation particulière.

D'autres éléments d'interprétation du modèle linéaire existent, à savoir :

— Le $\boldsymbol{R^2}$ (ou $\boldsymbol{R^2}$ **ajusté**) : pour savoir si les variables explicatives prédisent correctement Y, le R^2 mesure la part de variance de Y expliquée par l'ensemble des prédicteurs. La formule du R^2 est donnée par :

$$R^2 := \frac{\hat{Var}(\hat{Y})}{\hat{Var}(Y)} = \hat{Corr}(Y, \hat{Y})^2.$$

Le R^2 est compris entre 0 et 1 : plus il est proche de 0, moins le modèle expliquera les données. Inversement, plus il est proche de 1, plus le modèle ajuste convenablement les données. Le R^2 augmente nécessairement avec le nombre de variables explicatives, quel que soit la qualité de la variable ajoutée. Il peut alors être intéressant d'utiliser un élément qui pénalise les modèles si le nombre de paramètres est trop important : il s'agit du R^2 ajusté, noté R^2_{adj}. Il peut aussi être utilisé pour comparer des modèles qui ont un nombre de paramètres différents. Celui-ci est défini par :

$$R^2_{\text{adj}} = R^2 - (1 - R^2)\frac{d}{n - d - 1}.$$

— L'importance d'une variable x_l est définie par la **t-statistique** :

$$t_{\beta_l} = \left| \frac{\hat{\beta}_l}{\sigma_{\hat{\beta}_l}} \right|.$$

Elle représente le rapport entre le poids estimé et son écart-type. Ainsi, plus le poids d'une variable est grand, plus la variable est importante. De même, plus la variance du poids estimé est faible, plus la variable associée est importante.

— La **p-valeur** : de la t-statistique découle une p-valeur qui permet de juger la significativité de chaque variable. La p-valeur associée au coefficient β_l, notée p-valeur(t_{β_l}), est calculée à partir de la distribution de Student. Elle est donnée par :

$$\text{p-valeur}(t_{\beta_l}) = 2 \times (1 - F_t(t_{\beta_l}, \nu))$$

où $F_t(t_j, \nu)$ est la fonction de répartition cumulative de la distribution de Student avec ν degrés de liberté, et t_{β_l} la t-statistique. On dit qu'une variable est statistiquement significative au seuil de x % si la p-valeur associée est inférieure à x %. Les seuils de significativité usuels sont 1 %, 5 % et plus rarement 10 %.

4.1.3 Sélection des variables

La sélection rigoureuse des variables est importante pour évaluer la multicolinéarité [1] dans les modèles linéaires : elle assure la robustesse et l'interprétabilité du modèle linéaire. Les deux méthodes de sélection des variables décrites ci-après sont applicables à l'ensemble des modèles linéaires considérés dans ce mémoire (régression linéaire, GLM et GLMM). Dans la suite, elles sont uniquement illustrées sur la régression linéaire.

Régularisation de type Lasso

La régularisation Lasso (*Least Absolute Shrinkage and Selection Operator*) consiste à ajouter un terme de pénalité aux valeurs élevées des paramètres afin, d'une part, d'éviter le surapprentissage et, d'autre part, de rendre les modèles parcimonieux, en imposant un certain nombre de coefficients nuls. La régularisation Lasso permet donc une meilleure interprétation des modèles. Dans le cas de la régression linéaire, la fonction de coût L à minimiser est l'erreur quadratique, définie par :

$$\forall (y, \hat{y}) \in \mathbb{R}^d, \quad L(y, \hat{y}) = \frac{1}{2n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 = \frac{1}{2n} \sum_{i=1}^{n} \left(y^{(i)} - \beta_0 - \sum_{l=1}^{d} \beta_l x_l^{(i)} \right)^2.$$

La régularisation va ajouter une pénalisation à la fonction de coût. Cette pénalisation est proportionnelle à la norme des coefficients. Lorsque la norme utilisée est la norme ℓ_1, on parle de régularisation Lasso ; lorsqu'il s'agit de la norme ℓ_2, on parle de régularisation Ridge. L'avantage de la première est qu'elle annule les coefficients à retirer du modèle.

1. La multicolinéarité correspond à une situation dans laquelle certaines variables explicatives sont fortement corrélées entre elles.

C'est pour cette raison que, dans ce mémoire, cette régularisation sera exploitée. La fonction de coût dans ce cas s'écrit :

$$\forall (y, \hat{y}) \in \mathbb{R}^d, \quad L(y, \hat{y}) = \frac{1}{2n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \|\boldsymbol{\beta}\|_1 = \frac{1}{2n} \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2 + \lambda \sum_{l=1}^{d} |\beta_l|.$$

Le paramètre $\lambda \in \mathbb{R}$ est un paramètre de contrôle sur la force de la régularisation. Il est souvent choisi à l'aide de la méthode de validation croisée. Plus λ sera grand, plus le nombre de coefficients nuls sera élevé.

Facteur d'inflation de la variance (VIF)

Le facteur d'inflation de la variance (VIF, pour *Variance Inflation Factor*) est un indicateur couramment utilisé pour diagnostiquer la présence de multicolinéarité entre les variables explicatives d'un modèle linéaire. La multicolinéarité peut entraîner une instabilité dans l'estimation des coefficients $\boldsymbol{\beta}$ et une inflation de leur variance. Le calcul du VIF pour chaque variable permet notamment d'identifier les variables redondantes et de guider la sélection des variables à inclure dans le modèle final. Pour chaque variable explicative x_l, le VIF est défini comme suit :

$$\mathrm{VIF}(x_l) = \frac{1}{1 - R_l^2}$$

où R_l^2 est le coefficient de détermination obtenu en régressant la variable x_l sur l'ensemble des autres variables explicatives du modèle, c'est-à-dire :

$$x_l = \gamma_0 + \sum_{\substack{j=1 \\ j \neq l}}^{d} \gamma_j x_j + \xi$$

R_l^2 mesure la proportion de variance de x_l expliquée par les autres variables. Il est équivalent au R^2 décrit dans la section précédente, mais pour la régression de x_l sur l'ensemble des autres variables explicatives. Ainsi, plus R_l^2 est élevé, plus le VIF associé sera grand, car cela indique que x_l est bien expliquée par les autres variables. Un VIF supérieur à 5 (8 ou 10 selon certains usages) est généralement considéré comme révélateur d'une multicolinéarité problématique [70, 71]. Une telle valeur pour x_l suggère que cette variable peut être retirée ou remplacée afin d'améliorer la stabilité et l'interprétabilité du modèle.

Ces deux méthodes seront exploitées dans les chapitres 7 et 8 afin de sélectionner les variables pertinentes au sein des modèles statistiques développés pour l'élaboration des scores de santé, à savoir la régression linéaire, le GLM et le GLMM (ces deux derniers modèles sont décrits dans les sections suivantes).

4.2 Modèle linéaire généralisé

Les modèles linéaires généralisés (GLM) sont une extension des modèles de régression linéaire classiques permettant de traiter des variables dépendantes qui ne suivent pas nécessairement une distribution normale. Ils sont particulièrement utiles dans de nombreux

domaines, allant des sciences économiques et actuarielles à l'analyse environnementale, en passant par l'apprentissage automatique. Grâce à leur flexibilité, les GLM permettent de modéliser des relations entre une variable réponse qui suit une distribution appartenant à la famille exponentielle et un ensemble de prédicteurs tout en tenant compte de distributions variées, comme la loi Binomiale pour des données de type succès/échec ou la loi de Poisson pour des données de comptage. Les GLM se composent de trois éléments principaux : une fonction de lien, une distribution de la famille exponentielle et une structure linéaire pour les prédicteurs. Cette section s'appuie en grande partie sur le *Cours d'actuariat de l'assurance non-vie* de Nicolas Baradel [72].

4.2.1 Formulation du modèle

Afin d'estimer pour chaque assuré son nombre moyen de sinistres et/ou son coût moyen par sinistre, il est d'usage d'implémenter un modèle linéaire généralisé (GLM). En premier lieu, il convient de définir une famille de loi appelée **famille exponentielle** :

Définition 4.2.1 *Soit $(P_\theta)_{\theta\in\mathbb{R}}$ une famille de mesures de probabilité. Le modèle fait partie de la famille exponentielle si cette famille est dominée et que sa densité peut s'écrire sous la forme :*

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right),$$

où a, b, et c sont des fonctions mesurables, a est strictement positive, et ϕ est un paramètre de dispersion.

Un premier résultat important est la caractérisation des deux premiers moments de la variable aléatoire à prédire Y.

Lemme 4.2.2 *Si $Y \sim P_\theta$ et si b est une fonction de classe C^2, alors :*

$$\begin{cases} \mathbb{E}[Y] = b'(\theta), \\ \mathrm{Var}(Y) = b''(\theta)\cdot a(\phi). \end{cases}$$

Définition 4.2.3 (Fonction de lien canonique) *Soit $b' : \mathbb{R} \to A$ où $A \subset \mathbb{R}$ bijective. On a $\theta = (b')^{-1}(\mu)$. La fonction $(b')^{-1}$ est la fonction de lien canonique. Comme la relation est bijective, si $\theta = g(\mu)$ avec une fonction g, alors $g = (b')^{-1}$ est la fonction de lien canonique.*

La plupart des lois habituelles appartiennent à la famille exponentielle. Une liste non exhaustive de lois usuelles appartenant à la famille exponentielle est donnée ci-dessous à titre illustratif.

Exemple 4.2.4 (Loi de Bernoulli) *La loi de Bernoulli de paramètre $p \in]0,1[$ fait partie de la famille exponentielle et sa densité (par rapport à la mesure $\delta_0 + \delta_1$) s'écrit :*

$$f(y;\theta,\phi) = (1-p)^{1-y}p^y = \exp\left(y\log\frac{p}{1-p} + \log(1-p)\right).$$

On pose $\theta := \log\frac{p}{1-p}$ qui détermine la fonction de lien canonique. La densité se réécrit :

$$f(y;\theta,\phi) = \exp\left(y\theta - \log(1+e^\theta)\right).$$

On identifie également $b(\theta) = \log(1+e^\theta)$, $a(\phi) = 1$ et $c(y,\phi) = 0$.

Exemple 4.2.5 (Loi de Poisson) *La loi de Poisson de paramètre $\lambda > 0$ fait partie de la famille exponentielle et sa densité (par rapport à la mesure $\sum_{n\in\mathbb{N}} \delta_n$) s'écrit :*

$$f(y;\theta,\phi) = \frac{e^{-\lambda}\lambda^y}{y!} = \exp\left(y\log(\lambda) - \lambda - \log(y!)\right).$$

On pose $\theta := \log(\lambda)$ qui détermine la fonction de lien canonique. La densité se réécrit :

$$f(y;\theta,\phi) = \exp\left(y\theta - e^\theta - \log(y!)\right).$$

On identifie également $b(\theta) = e^\theta$, $a(\phi) = 1$ et $c(y,\phi) = -\log(y!)$. En pratique, on utilise les fonctions de lien $g(\lambda) = \log(\lambda)$ et $g(\lambda) = \sqrt{\lambda}$.

Exemple 4.2.6 (Loi Binomiale Négative) *La loi Binomiale Négative de paramètres $r > 0$ et $p \in]0,1[$ fait partie de la famille exponentielle et sa densité (par rapport à la mesure $\sum_{n\in\mathbb{N}} \delta_n$) s'écrit :*

$$f(y;\theta,\phi) = \frac{\Gamma(r+y)}{y!\Gamma(r)} p^r (1-p)^y.$$

Son espérance est $\mu = \frac{r(1-p)}{p}$. On paramétrise par (r,μ), la densité se réécrit :

$$\begin{aligned} f(y;\theta,\phi) &= \frac{\Gamma(r+y)}{y!\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^y \\ &= \exp\left(y\log\left(\frac{\mu}{r+\mu}\right) + r\log\left(\frac{r}{r+\mu}\right) + \log\left(\frac{\Gamma(r+y)}{y!\Gamma(r)}\right)\right). \end{aligned}$$

On pose $\theta := \log\frac{\mu}{r+\mu}$ qui détermine la fonction de lien canonique. La densité se réécrit :

$$f(y;\theta,\phi) = \exp\left(y\theta + r\log(1-e^\theta) + \log\left(\frac{\Gamma(r+y)}{y!\Gamma(r)}\right)\right).$$

On identifie également $b(\theta) = -r\log(1-e^\theta)$, $a(\phi) = 1$ et $c(y,\phi) = \log\left(\frac{\Gamma(r+y)}{y!\Gamma(r)}\right)$. En pratique, on utilise les fonctions de lien $g(\mu) = \log(\mu)$ et $g(\mu) = \sqrt{\mu}$.

Exemple 4.2.7 (Loi normale) *La loi normale de paramètre $\mu \in \mathbb{R}$ et $\sigma^2 > 0$ fait partie de la famille exponentielle et sa densité (par rapport à la mesure de Lebesgue) s'écrit :*

$$f(y;\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \exp\left(y\mu - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right).$$

On pose $\theta := \mu$ qui détermine la fonction de lien canonique. La densité se réécrit :

$$f(y;\theta,\phi) = \exp\left(y\theta - \frac{\theta^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right).$$

En supposant que l'on observe un vecteur aléatoire $\mathbf{X} \in \mathbb{R}^d$ constitué de d prédicteurs, que l'on dispose d'une fonction g bijective, le GLM est défini par l'équation suivante :

$$g(\mathbb{E}[Y|\mathbf{X}]) = \eta(\mathbf{X}) := \beta_0 + \mathbf{X}'\boldsymbol{\beta}$$

où :

— $g(\cdot)$ est la fonction de lien qui relie l'espérance de la variable réponse à la combinaison linéaire des prédicteurs. Il est possible d'utiliser la fonction de lien canonique du modèle ;
— $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^d$ est le vecteur des coefficients à estimer ; et
— $\eta(\mathbf{X})$ est appelé le prédicteur linéaire. Il vient :

$$\mathbb{E}[Y|\mathbf{X}] = g^{-1}(\eta(\mathbf{X})) = g^{-1}(\beta_0 + \mathbf{X}'\boldsymbol{\beta})$$

4.2.2 Estimation des paramètres par maximum de vraisemblance

Les paramètres du modèle, β_0 et $\boldsymbol{\beta}$, sont généralement estimés par la méthode du maximum de vraisemblance. Cette méthode cherche à maximiser la fonction de vraisemblance du modèle par rapport aux paramètres. On suppose que $Y \mid \mathbf{X}$ suit la loi $P_{\eta(\mathbf{X})}$. On note n le nombre d'observations et d le nombre de variables explicatives. En particulier, si on observe $(y^{(i)}, \boldsymbol{x^{(i)}})_{1\leq i\leq n}$ variables aléatoires indépendantes et identiquement distribuées (i.i.d.), où pour tout $1 \leq i \leq n$, $\boldsymbol{x^{(i)}} \in \mathbb{R}^d$, la log-vraisemblance, dans le cas de la fonction de lien canonique, s'écrit :

$$\begin{aligned}\ell(\beta_0, \boldsymbol{\beta}; (y^{(i)} \mid \boldsymbol{x^{(i)}})_{1\leq i\leq n}) &= \sum_{i=1}^{n} \log\Big(f(y^{(i)}; \beta_0 + \boldsymbol{x^{(i)\prime}}\boldsymbol{\beta}, \phi)\Big) \\ &= \sum_{i=1}^{n} \frac{y^{(i)}(\beta_0 + \boldsymbol{x^{(i)\prime}}\boldsymbol{\beta}) - b(\beta_0 + \boldsymbol{x^{(i)\prime}}\boldsymbol{\beta})}{a(\phi)} + c(y^{(i)}, \phi).\end{aligned}$$

Les paramètres du modèle, β_0 et $\boldsymbol{\beta}$ sont obtenus en résolvant les équations du premier ordre du problème :

$$(\hat{\beta}_0, \hat{\boldsymbol{\beta}}) \in \arg\max_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{d+1}} \ell(\beta_0, \boldsymbol{\beta}; (y^{(i)} \mid \boldsymbol{x^{(i)}})_{1\leq i\leq n}).$$

Ces équations ne permettent pas d'en déduire, en général, une solution explicite (en dehors de cas particuliers, comme le modèle gaussien avec fonction de lien canonique qui permet de retrouver le modèle linéaire avec bruit gaussien). Il est nécessaire de passer par des méthodes linéaires. Par exemple, la fonction `glm` dans R et le module `statsmodels` sous Python permettent d'estimer des modèles linéaires généralisés. Il convient ensuite d'éliminer les facteurs non-déterminants via des tests d'hypothèse.

Remarque 4.2.8 *L'estimation par maximum de vraisemblance de $(\beta_0, \boldsymbol{\beta})$ est indépendante de ϕ, qui peut être estimé dans un second temps. Ce résultat est également valide avec une autre fonction de lien.*

4.2.3 Interprétation du GLM

Il est facile d'interpréter les modèles linéaires généralisés, et notamment l'impact d'un changement d'une unité d'un prédicteur sur la prédiction. Considérons la prédiction y faite pour un certain vecteur de prédicteur $\mathbf{X} \in \mathbb{R}^d$. On a :

$$y = g^{-1}(\beta_0 + \mathbf{X}'\boldsymbol{\beta}).$$

Notons $\tilde{\mathbf{X}}$ le même vecteur que $\mathbf{X}$ sauf pour un certain $l \in \{1, \ldots, d\}$ où $\tilde{x}_l = x_l + 1$, i.e. :

$$\forall j \in \{1, \ldots, d\}, \quad \tilde{x}_l = x_l + 1_{\{j=l\}}.$$

Alors la prédiction $\tilde{y}$, faite pour $\tilde{x}$ est :

$$\tilde{y} = g^{-1}\left(\beta_0 + \tilde{\mathbf{X}}'\boldsymbol{\beta}^{-\boldsymbol{l}} + \beta_l\right),$$

où $\boldsymbol{\beta}^{-\boldsymbol{l}} \in \mathbb{R}^{d-1}$ est le vecteur $\boldsymbol{\beta}$ privé de β_l.

Donc l'écart de prédiction faite par le modèle lorsque la variable x_l augmente d'une unité est :

$$\tilde{y} - y = g^{-1}\left(\beta_0 + \tilde{\mathbf{X}}'\boldsymbol{\beta}^{-\boldsymbol{l}} + \beta_l\right) - g^{-1}\left(\beta_0 + \tilde{\mathbf{X}}'\boldsymbol{\beta}^{-\boldsymbol{l}}\right).$$

Dans les cas usuels suivants, on trouve, par exemple :

— si $g = \mathrm{Id}$ (fonction identité), un effet additif : $\tilde{y} - y = \beta_j$;

— si $g = \ln$, un effet multiplicatif : $\dfrac{\tilde{y}}{y} = e^{\beta_j}$.

4.2.4 Validation du modèle

Pour valider un modèle, il faut mesurer sa qualité. Pour cela, il existe plusieurs statistiques listées ci-dessous :

— **Déviance** : c'est une mesure qui compare la différence entre les valeurs prédites et celles observées. Elle peut être définie comme :

$$D_{mod} = 2 \ln\left(\frac{L_{\mathrm{SAT}}}{L_{mod}}\right),$$

où L_{SAT} est la vraisemblance du modèle saturé et L_{mod} la vraisemblance de notre modèle. La déviance est une quantité positive d'autant plus petite que le modèle est riche et s'ajuste bien. Elle est nulle dans le cas du modèle saturé, considéré comme le modèle offrant l'ajustement parfait. Il est supposé que le modèle est estimé à partir de n observations et d variables explicatives, avec la condition $d < n$. En pratique, si $\frac{D_{mod}}{n-d-1} \leq 1$, alors le modèle est considéré comme ayant une bonne qualité d'estimation.

— **Qualité d'ajustement** :

1. **Coefficient de détermination ajusté** R^2_α : la qualité d'ajustement d'un modèle peut être mesurée grâce au coefficient de détermination ajusté, construit par analogie avec le coefficient de détermination classique R^2. Pour cela, on compare la déviance du modèle nul (à un seul paramètre : *intercept*) D_0 avec celle du modèle que l'on veut évaluer D_{mod}. La formule du coefficient de détermination ajusté R^2_α est donnée par :

$$R^2_\alpha = \frac{D_0 - D_{mod}}{D_0}.$$

Plus il est proche de 1, meilleure est la qualité d'ajustement du modèle. La déviance est donc une statistique de qualité d'ajustement du modèle analogue à la somme des carrés des résidus dans les modèles linéaires classiques.

2. **Pseudo R^2 de Cox et Snell (R^2_{CS})** : cet indicateur repose sur la comparaison des vraisemblances du modèle nul et du modèle ajusté. Il mesure l'amélioration du modèle par rapport à un modèle sans covariables, mais il est borné strictement en dessous de 1 dans le cas des modèles de Poisson. Sa formule est donnée par :

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_{mod}}\right)^{\frac{2}{n}},$$

où n est le nombre d'observations, L_0 la vraisemblance du modèle nul (avec seulement l'*intercept*) et L_{mod} la vraisemblance du modèle estimé.

Le R^2_α s'appuie sur la déviance tandis que le R^2_{CS} utilise la vraisemblance. Les deux indicateurs permettent de comparer la performance de différents modèles, mais leur valeur numérique et leur interprétation diffèrent. Il est possible de les interpréter conjointement, sans pour autant les utiliser comme le R^2 classique de régression linéaire.

— **Statistique de Pearson** : elle compare les valeurs prédites $\hat{y}^{(i)}$ et les valeurs observées $y^{(i)}$:

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{\mathrm{Var}(Y)}.$$

— **AIC (Akaike Information Criterion)** :

$$\mathrm{AIC} = D_{mod} + 2r,$$

où r est le rang de la matrice de covariable. L'AIC est d'autant plus faible que la log-vraisemblance est élevée et que le nombre de paramètres est petit. Il permet donc d'établir un ordre sur les modèles en prenant en compte les deux contraintes. L'AIC seul est inutile. Il s'agit d'une mesure relative. Pour un jeu de données et un ensemble de modèles pour ce jeu de données, ceux qui ont un AIC plus petit sont « meilleurs » au sens de ce critère.

— **BIC (Bayesian Information Criterion)** : Le BIC est un critère d'information similaire à l'AIC, mais il pénalise plus fortement la complexité du modèle, ce qui le rend plus strict lorsque le nombre d'observations est élevé. Il est défini par :

$$\mathrm{BIC} = D_{\mathrm{mod}} + r \cdot \ln(n),$$

où D_{mod} est la déviance du modèle, r le nombre de paramètres estimés, et n le nombre d'observations. Comme pour l'AIC, plus le BIC est faible, meilleur est le modèle selon ce critère. Le BIC permet de comparer différents modèles ajustés sur le même jeu de données, en privilégiant celui ayant la plus petite valeur de BIC.

Pour toutes ces statistiques, il est nécessaire de supposer que $d < n$. Pour la validation de notre modèle, la déviance, le R^2_α, le R^2_{CS}, la statistique de Pearson, l'AIC ainsi que le BIC seront évalués.

4.3 Modèle linéaire mixte généralisé

4.3.1 Formulation du modèle

Les modèles linéaires mixtes généralisés (GLMM) représentent une extension des modèles linéaires généralisés. Ils permettent de traiter les données hiérarchiques, longitudi-

nales ou corrélées, où les observations ne sont pas indépendantes. Par exemple, lorsque plusieurs observations sont collectées par individus, cela crée une dépendance entre les observations du jeu de données. Une dépendance peut aussi être causée par les observations de patients suivis par le même médecin. Dans un GLMM, la relation entre la variable cible et les variables explicatives est modélisée par une combinaison linéaire d'effets fixes et aléatoires. En utilisant les notations du GLM, le GLMM est défini par l'équation suivante :

$$g(E[Y|X,Z]) = \eta(X,Z) := X\beta + Zu,$$

où :

— $g(\cdot)$ est la fonction de lien ;

— $X\beta$ représente les effets fixes, où β est le vecteur de coefficients à estimer ;

— Zu représente les effets aléatoires, où u est un vecteur de variables aléatoires suivant une distribution normale avec une moyenne nulle et une matrice de covariance G, $u \sim \mathcal{N}(0, G)$; et

— $\eta(X,Z)$ est le prédicteur linéaire intégrant les contributions des effets fixes et aléatoires.

Contrairement aux effets fixes, les effets aléatoires permettent de capturer la variabilité au sein de chaque groupe. Ainsi, l'ajout des effets aléatoires permet de modéliser la corrélation entre les observations au sein du même groupe. Cet ajout rend l'estimation des effets fixes plus précise et moins biaisée.

4.3.2 Estimation des paramètres

Si l'estimation des paramètres du GLM se fait assez simplement via la méthode du maximum de vraisemblance, pour le GLMM, l'estimation est plus délicate en raison de l'ajout des effets aléatoires. Plusieurs méthodes couramment utilisées permettent d'estimer les paramètres du GLMM :

— la **quasi-vraisemblance pénalisée** : méthode flexible et largement implémentée, mais qui peut être biaisée pour des grandes variances ou des petites moyennes ;

— l'**approximation de Laplace** : méthode plus précise que la quasi-vraisemblance pénalisée, mais plus lente et moins flexible ;

— la **quadrature de Gauss-Hermite** : méthode plus précise que l'approximation de Laplace, mais plus lente ;

— la **méthode de Monte Carlo par chaînes de Markov** : méthode très précise et flexible en raison du nombre arbitraire d'effets aléatoires, mais assez lente et techniquement complexe, nécessitant un cadre bayésien.

4.3.3 Interprétation du GLMM

Dans un modèle linéaire mixte généralisé (GLMM), nous avons une structure similaire à celle d'un GLM, mais avec l'ajout d'effets aléatoires pour capturer la variabilité intragroupe. La formule de prédiction pour un GLMM est donnée par :

$$y = g^{-1}(\beta_0 + X'\beta + Zu),$$

où :

— $X'\beta$ représente les effets fixes, comme dans un GLM.

— Zu représente les effets aléatoires, où u est un vecteur de variables aléatoires.

Pour interpréter un changement d'une unité d'un prédicteur x_l, nous pouvons examiner la différence de prédiction en tenant compte des effets aléatoires. Considérons la prédiction initiale y pour un vecteur de prédicteur X, et une prédiction modifiée $\tilde{y}$ pour un vecteur $\tilde{X}$ où $\tilde{x}_l = x_l + 1$. La prédiction modifiée est :

$$\tilde{y} = g^{-1}(\beta_0 + \tilde{X}'\beta + Zu).$$

L'écart de prédiction est alors :

$$\tilde{y} - y = g^{-1}(\beta_0 + \tilde{X}'\beta + Zu) - g^{-1}(\beta_0 + X'\beta + Zu).$$

Les cas usuels utilisés dans le chapitre précédent donnent :

— **Fonction de lien identité ($g = \mathbf{Id}$)** : l'effet est additif, comme dans un GLM :

$$\tilde{y} - y = \beta_l.$$

— **Fonction de lien log ($g = \ln$)** : l'effet est multiplicatif :

$$\frac{\tilde{y}}{y} = e^{\beta_l}.$$

Même si l'ajout des effets aléatoires peut influencer les prédictions au sein de chaque groupe, l'interprétation des effets fixes dans un GLMM reste similaire à celle des coefficients d'un GLM. Les coefficients du GLMM doivent donc être interprétés en tenant compte de cette variabilité supplémentaire.

Chapitre 5

Modèles de *machine learning* pour la construction de scores de santé

Dans un contexte où les modèles actuariels traditionnels tels que les GLM peuvent atteindre leurs limites, l'application des méthodes de *machine learning* en assurance santé ouvre de nouvelles perspectives, permettant la création de scores de santé plus précis et personnalisés en intégrant des variables complexes comme les facteurs environnementaux et climatiques. La construction d'un modèle de ML suit une série d'étapes chronologiques et méthodologiques, visant à garantir à la fois la performance et la robustesse du modèle. Ces étapes peuvent être résumées comme suit :

1. **Définition du problème et type d'apprentissage**
   On distingue principalement deux types de tâches : la régression (prédiction d'une variable quantitative) et la classification (prédiction d'une variable qualitative). Le choix de l'algorithme dépend de la nature de la variable cible y.
2. **Prétraitement des données**
   Le prétraitement des données constitue une étape cruciale, car la qualité des données conditionne directement la performance des modèles. Avant toute modélisation, il convient d'explorer, de nettoyer et de transformer les données afin d'en extraire le maximum d'information utile. Ce travail inclut généralement :
   - le traitement des valeurs manquantes (suppression ou imputation) ;
   - le traitement des valeurs aberrantes ;
   - la normalisation des variables ;
   - l'encodage des variables catégorielles (par exemple via du *one-hot encoding*) ;
   - la création de nouvelles variables pour enrichir les données pré-existantes via du *feature engineering* ;
   - l'analyse des corrélations pour détecter les redondances.
3. **Choix du modèle et des paramètres**
   Un modèle de *machine learning* peut s'écrire sous la forme d'une fonction $f(x;\theta)$, avec x les variables explicatives, y la variable cible et θ l'ensemble des paramètres à estimer. L'objectif est de trouver les paramètres optimaux qui minimisent une fonction de coût, choisie parmi les métriques suivantes :
   - la MAE (*Mean Absolute Error*) est la moyenne arithmétique des valeurs absolues des écarts entre les valeurs observées y_i et les valeurs prédites $\hat{y}_i$:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|;$$

— la MSE (*Mean Square Error*) est la moyenne des carrés des écarts entre les valeurs observées et prédites :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2;$$

— la RMSE (*Root Mean Square Error*) est la racine carrée de la MSE :

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

4. **Entraînement du modèle**
   L'apprentissage consiste à ajuster les paramètres θ pour minimiser la fonction de coût. L'algorithme de descente de gradient est couramment utilisé dans ce but.
5. **Evaluation et validation du modèle**
   Le jeu de données est divisé en plusieurs sous-ensembles :
   — l'ensemble d'entraînement est utilisé pour entraîner le modèle ;
   — l'ensemble de validation est utilisé pour ajuster les paramètres du modèle appelés aussi hyperparamètres ; et
   — l'ensemble de test est utilisé pour estimer la performance finale du modèle.

   Une méthode robuste d'évaluation consiste à utiliser la validation croisée *k-fold* : le jeu de données est divisé en k sous-groupes, le modèle est entraîné k fois sur $k-1$ sous-groupes, puis évalué sur le sous-groupe restant. L'erreur est alors moyennée sur les k partitions.
6. **Interprétation du modèle**
   Un point d'attention particulier de ce mémoire est porté sur l'interprétation des modèles. En effet, une fois le modèle entraîné et validé, il est important de savoir comment il procède. L'interprétation du modèle permet notamment d'identifier les variables les plus impactantes et de faciliter son analyse à ses utilisateurs. C'est pour cela que les modèles de *Deep Learning* ne sont pas exploités dans ce mémoire. La logique interne de ces modèles permettant d'atteindre la sortie n'est pas interprétable.

Le but de ce chapitre est de présenter les modèles de *machine learning* les plus couramment utilisés en actuariat (arbres de décision, forêts aléatoires et *Gradient Boosting*) pour justifier le choix du modèle XGBoost dans la cadre de ce mémoire.

5.1 Arbre de décision

Les arbres de décision (*Decision Tree* en anglais) sont des algorithmes très répandus utilisés dans les problèmes de régression et de classification. Les arbres de décision sont les éléments de base du modèle XGBoost qui sera détaillé dans la section 5.3. Le but de cette présente section est de décrire plus en détail cette famille d'algorithmes et d'en comprendre les fondements.

Leur principe repose sur l'algorithme CART proposé par Breiman en 1984, dont l'idée générale est de diviser l'espace des variables explicatives en plusieurs groupes mutuellement exclusifs, c'est-à-dire d'en réaliser une partition [73]. Une fois segmenté, il crée un ensemble de règles appelées séquences de décision uniques par groupe en vue de la prédiction d'un résultat ou d'une classe [73]. Ce cheminement crée une structure arborescente où chaque nœud interne représente un test sur une variable explicative, chaque branche représente le résultat d'un test et chaque feuille une prédiction. Le but est donc de partitionner l'espace des variables explicatives $\mathcal{X}$ en sous-espaces homogènes vis-à-vis de la variable cible y. L'algorithme CART est détaillé en annexe B.1.

Mathématiquement, un arbre de décision induit une fonction $f : \mathcal{X} \to \mathcal{Y}$ définie par morceaux. Chaque morceau correspond à une région R_m de l'espace des variables, à laquelle on associe une prédiction c_m. Cette prédiction est égale à une moyenne pondérée dans le cas d'une régression ou à la classe majoritaire pour une classification. La fonction f est de la forme

$$f(X) = \sum_{m=1}^{M} c_m \cdot \mathbb{I}_{\{X \in R_m\}},$$

où :

- $X = (x_1, ..., x_d)$ est l'ensemble des variables explicatives ;
- M est le nombre de feuilles de l'arbre ;
- c_m est la prédiction faite dans la région R_m ; et
- $\mathbb{I}_{\{X \in R_m\}}$ est l'indicatrice qui vaut 1 si X appartient à la région R_m, 0 sinon.

Dans le cas simplifié d'un arbre de classification binaire, on cherche à prédire une variable catégorielle Y à valeur dans $\{1, ..., M\}$. La construction de l'arbre est réalisée de manière récursive en séparant les individus successivement en deux sous-ensembles. Il convient alors de choisir à chaque étape la variable explicative et le seuil qui permettent de séparer au mieux les individus. Ces étapes sont réalisées jusqu'à ce qu'une nouvelle division n'apporte plus d'amélioration ou jusqu'à ce qu'un sous-groupe ne soit pas assez conséquent. Ainsi, il est nécessaire de définir un processus de sélection des variables ainsi qu'un critère d'arrêt.

La construction de l'arbre dit maximal nécessite la définition d'une fonction ϕ qui donne le degré d'impureté d'un nœud. En classification, l'entropie et l'indice de Gini constituent les exemples les plus classiques de fonction d'impureté. Cette fonction :

- prend son minimum lorsque le nœud est pur, c'est-à-dire lorsqu'une seule classe est représentée dans le nœud ;
- prend son maximum lorsque le nœud est impur, c'est-à-dire lorsque toutes les classes sont équiréparties dans le nœud ; et
- est symétrique par rapport aux classes.

La division d'un nœud est au cœur de la construction récursive d'un arbre de décision. Soit x_j avec $j \in \{1, ..., d\}$ une variable explicative candidate pour réaliser une division à partir d'un nœud B, et soit $s \in \mathbb{R}$ un seuil de coupure. Dans le cas de variables quantitatives, l'objectif est de scinder le nœud B en deux sous-nœuds, notés $B_1(z)$ et $B_2(z)$, selon la règle suivante :

$$B_1(z) = \{w \in B \,|\, x_j(w) \leq s\} \quad \text{et} \quad B_2(z) = \{w \in B \,|\, x_j(w) > s\}.$$

Le choix du seuil s repose sur la minimisation d'une fonction de perte basée sur une mesure d'impureté ϕ : l'objectif est de maximiser l'homogénéité des sous-nœuds, c'est-à-dire tendre vers une pureté parfaite où chaque sous-nœud ne contient que des observations d'une même classe. Dans le cas de la régression, on utilise un critère similaire visant à réduire la variance intra-nœud. La construction de l'arbre dit maximal consiste ainsi à appliquer ce processus récursivement à chaque nœud, en choisissant à chaque étape la variable x_j et le seuil s qui maximisent l'homogénéité des sous-nœuds. Deux conditions d'arrêt viennent compléter ce schéma :

— un nœud contenant un nombre d'observations inférieur à un seuil minimal fixé ne sera pas divisé davantage ;

— un nœud pur, c'est-à-dire ne contenant que des observations d'une même classe, devient automatiquement une feuille.

A l'issue de cet algorithme, l'arbre maximal obtenu est souvent sujet à un fort sur-apprentissage, bien qu'il présente une erreur d'apprentissage minimale. C'est pourquoi cet arbre maximal ne sera pas utilisé tel quel pour la prédiction. Il nécessite un élagage basé sur les notions conjointes de complexité et de taux d'erreur. Le sous-arbre retenu à partir de l'arbre maximal est celui pour lequel le taux d'erreur pénalisé par la complexité est minimum. Bien que très interprétables, les arbres de décision simples peuvent être instables et sensibles au bruit, d'où l'intérêt des méthodes d'agrégation comme les forêts aléatoires, détaillées dans la section suivante.

5.2 Forêts aléatoires

L'algorithme de Forêts aléatoires (*Random Forest*), proposé par Breiman (2001), appartient à la famille des méthodes d'agrégation de modèles [74]. Il s'appuie sur les arbres de décision (voir section 5.1) et vise à réduire la variance de ces derniers, souvent sujets au sur-apprentissage, en agrégeant un grand nombre d'arbres construits selon des stratégies aléatoires, comme le *bagging*. Le principe du *bagging* (*Bootstrap Aggregating*) consiste à générer plusieurs jeux de données d'apprentissage à partir du jeu initial, en tirant aléatoirement et avec remise des observations. Il est décrit plus en détail dans l'annexe B.2. Pour chaque jeu ainsi construit, un arbre de décision est généré indépendamment. L'agrégation des prédictions de ces arbres permet de réduire la variance du modèle global.

Soit (X, Y) un vecteur aléatoire où $X \in \mathbb{R}^d$ représente les variables explicatives et $Y \in \mathbb{R}$ la variable cible. On dispose d'un échantillon de taille n, noté $D_n = \{(X_i, Y_i)\}_{i=1}^n$. On cherche à estimer la fonction de régression $m(x) = \mathbb{E}[Y|X = x]$. La méthode du *bagging* consiste à agréger N estimateurs $\hat{m}_1, \ldots, \hat{m}_N$, chacun construit sur un échantillon issu d'un *bootstrap* différent. L'estimateur agrégé s'écrit alors :

$$\hat{m}(x) = \frac{1}{N} \sum_{k=1}^{N} \hat{m}_k(x).$$

Dans le cas de la classification, la prédiction finale est obtenue par vote majoritaire :

$$\hat{y}(x) = \text{mode}\{\hat{y}_1(x), \ldots, \hat{y}_N(x)\}$$

où $\hat{y}_k(x)$ est la prédiction de l'arbre k pour l'observation x et mode$\{\}$ désigne la fonction qui renvoie la valeur la plus fréquente de la liste donnée en abscisse.

Sous l'hypothèse que les estimateurs $\hat{m}_1, \ldots, \hat{m}_N$ sont indépendants et identiquement distribués (i.i.d.), on a :

$$\mathbb{E}[\hat{m}(x)] = \mathbb{E}[\hat{m}_1(x)] \qquad \text{et} \qquad \mathrm{Var}[\hat{m}(x)] = \frac{1}{N}\mathrm{Var}[\hat{m}_1(x)].$$

L'estimateur agrégé $\hat{m}$ conserve donc le même biais que chacun des estimateurs individuels, mais sa variance est divisée par N, ce qui améliore la stabilité et la capacité de généralisation du modèle. Une autre spécificité des forêts aléatoires réside dans l'introduction d'aléa lors de la construction des arbres. A chaque division d'un nœud, seul un sous-ensemble aléatoire de m variables parmi les d variables explicatives est considéré pour déterminer la meilleure coupure. Cette sélection aléatoire des variables permet de diminuer la corrélation entre les arbres et d'augmenter la diversité des modèles agrégés.

Les forêts aléatoires permettent également d'évaluer l'importance des variables explicatives, par exemple via la mesure de la diminution de l'impureté moyenne (Gini ou entropie) associée à chaque variable au sein des arbres de la forêt. Cependant, plus le nombre d'arbres utilisés augmente, plus le temps de calcul dans la phase d'apprentissage est important et plus il devient difficile de comprendre précisément comment le modèle prend ses décisions. On parle alors de « boîte noire ». Le modèle peut ainsi construire des scores pertinents, mais sans pouvoir aisément les interpréter. Le modèle XGBoost permet de concilier les deux via le *Gradient Boosting*.

5.3 XGBoost

XGBoost (pour *eXtreme Gradient Boosting*) est un modèle proposé par Tianqi Chen et Carlos Guestrin en 2016 [75]. Il s'agit d'une implémentation optimisée de l'algorithme de *Gradient Boosted Decision Trees* (GBDT). L'algorithme de *Gradient Boosting* est décrit en annexe B.3. Contrairement aux forêts aléatoires, dont le but est de construire des arbres de manière indépendante, XGBoost est un algorithme d'ensemble (*ensemble learning*) qui construit chaque arbre de façon séquentielle. Chaque nouvel arbre construit corrige ainsi les erreurs commises par la somme des arbres précédents. Cela a l'inconvénient de le rendre plus lent que les forêts aléatoires, mais lui permet de s'améliorer au fur et à mesure de la construction de la prédiction. Cette section s'appuie sur les travaux de Chen et Guestrin et reprend les notations utilisées dans leur papier.

5.3.1 Modèle d'ensemble additif

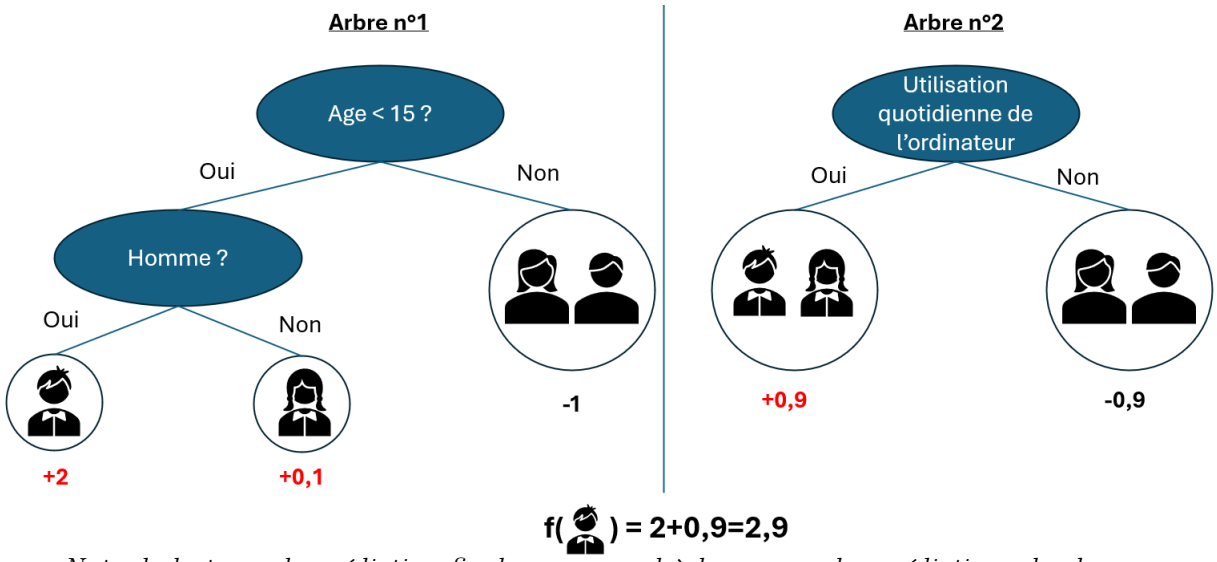
On définit $x_i \in \mathbb{R}^d$, un vecteur de d caractéristiques, y_i, la cible (valeur à prédire), et n le nombre d'observations. Soit un ensemble de données $\mathcal{P} = \{(x_i, y_i)\}_{i=1}^{n}$. Le modèle prédictif utilisé par XGBoost est un modèle d'ensemble additif : la prédiction finale correspond à la somme des prédictions de chaque arbre (voir figure 5.1). Il correspond à une somme de K fonctions f_k. Chacune de ces fonctions représente un arbre de décision :

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{5.1}$$

où $\mathcal{F}$ est l'ensemble des arbres de régression. On définit mathématiquement les arbres :

Définition 5.3.1 (Arbre) *Un arbre f_k est une fonction $f_k(x) = w_{q(x)}$ où $q : \mathbb{R}^m \to \{1, \ldots, T\}$ est une fonction qui assigne chaque échantillon à une feuille de l'arbre (par index), T est le nombre de feuilles, et $w_j \in \mathbb{R}$ est le poids associé à la feuille j.*

FIGURE 5.1 – Exemple d'un modèle d'ensemble additif



Note de lecture : la prédiction finale correspond à la somme des prédictions de chaque arbre.

5.3.2 Fonction objectif régularisée

L'apprentissage de l'ensemble des fonctions utilisées dans le modèle consiste à minimiser une fonction objectif régularisée. Intuitivement, l'utilisation d'une fonction objectif régularisée conduit à privilégier les modèles simples et prédictifs. La fonction objectif est sous la forme :

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{5.2}$$

avec :

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2,$$

où T est le nombre de feuilles de l'arbre, w_j est le score associé à la feuille j, et γ et λ sont les hyperparamètres de régularisation pouvant être fixés par validation croisée, par exemple.

La fonction objectif régularisée est composée de deux termes :

— une fonction de perte convexe l, qui mesure l'écart entre la prédiction $\hat{y}_i$ et la vraie valeur y_i ; et

— un terme de régularisation $\Omega(f_k)$, qui pénalise la complexité du modèle (c'est-à-dire des fonctions d'arbre de régression). Ce terme de régularisation supplémentaire permet de lisser les poids appris et d'éviter le surapprentissage. Lorsque le paramètre de régularisation est nul, la fonction objectif se ramène au *boosting* de gradient traditionnel.

5.3.3 Apprentissage par optimisation du gradient

Le modèle décrit dans l'équation (5.2) inclut des fonctions comme paramètres, empêchant d'utiliser les méthodes d'optimisation traditionnelles dans l'espace euclidien. Ainsi, XGBoost procède par ajout séquentiel d'arbres, en utilisant une approximation de Taylor au second ordre. A l'itération t, on cherche à ajouter une fonction f_t qui minimise :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t). \tag{5.3}$$

On ajoute donc la fonction f_t qui améliore le modèle selon l'équation (5.2). On peut développer au second ordre l'équation (5.3) pour optimiser l'objectif dans le cas général. On développe alors la fonction de perte autour de $\hat{y}_i^{(t-1)}$ et on obtient :

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t)$$

où :

$g_i = \left.\dfrac{\partial l(y_i, \hat{y}_i)}{\partial \hat{y}_i}\right|_{\hat{y}_i = \hat{y}_i^{(t-1)}}$ est le gradient de la fonction de perte par rapport à la prédiction précédente pour l'exemple i ;

$h_i = \left.\dfrac{\partial^2 l(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}\right|_{\hat{y}_i = \hat{y}_i^{(t-1)}}$ est la dérivée seconde (hessienne) de la fonction de perte pour l'exemple i.

On peut retirer les termes constants pour obtenir l'objectif simplifié à l'étape t :

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t). \tag{5.4}$$

5.3.4 Calcul optimal des scores des feuilles

Soit $I_j = \{i \mid q(x_i) = j\}$ l'ensemble des indices des exemples affectés à la feuille j. Nous pouvons réécrire l'équation (5.4) en développant Ω comme suit :

$$\begin{aligned}\tilde{\mathcal{L}}^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{i=1}^{n} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2\right] + \gamma T.\end{aligned}$$

En fixant $q(x)$, on peut calculer la valeur optimale des poids w_j^* pour la feuille j. Les scores de prédiction w_j n'interagissant pas entre eux, ils sont donc indépendants. En excluant la solution $w_j = 0$, on obtient :

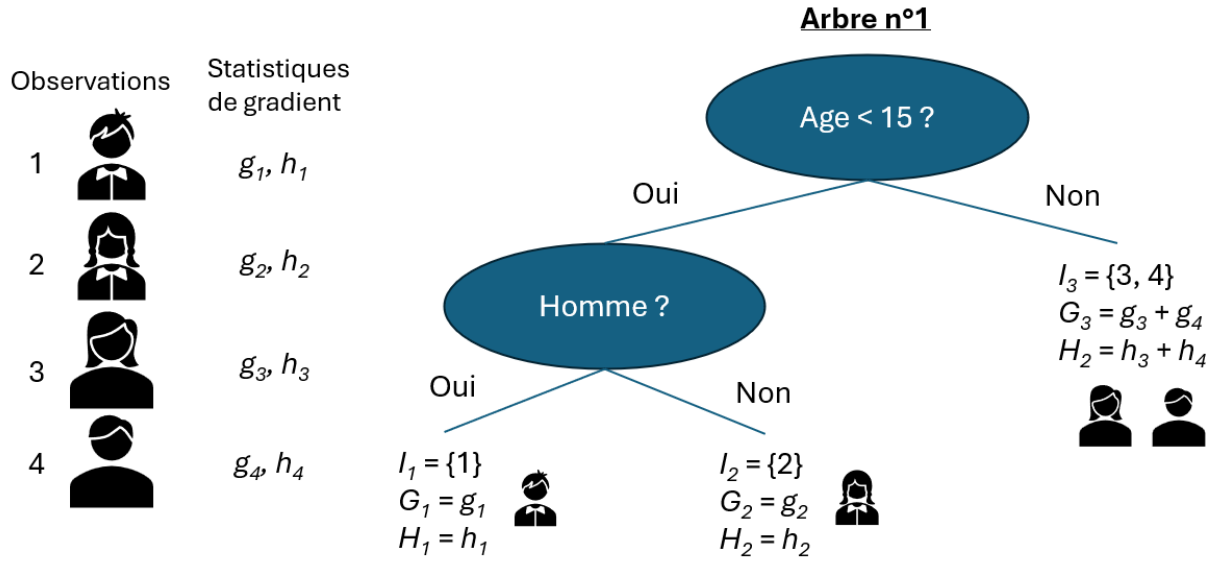
$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}.$$

Le score optimal correspondant de la structure d'arbre q est alors :

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T. \tag{5.5}$$

Cette dernière équation peut être utilisée pour mesurer la qualité d'une structure d'arbre q similaire au score d'impureté utilisé pour évaluer les arbres de décision, à la différence qu'il est dérivé pour un éventail plus large de fonctions objectif. La figure 5.2 ci-dessous, inspirée du travail des auteurs, illustre comment ce score peut être calculé. En reprenant l'arbre n°1 de la figure 5.1, et en appliquant la formule (5.5), on obtient un score optimal égal à $-\frac{1}{2}\sum_{j=1}^{T=3}\frac{G_j^2}{H_j+\lambda} + 3\gamma$, avec $G_j = \sum_{i\in I_j} g_i$ et $H_j = \sum_{i\in I_j} h_i$.

FIGURE 5.2 – Exemple de la structure de calcul du score dans le modèle XGBoost



Note de lecture : le score de qualité est obtenu en additionnant les statistiques de gradient et de gradient de second ordre pour chaque feuille, puis en appliquant la formule de score optimal (5.5).

Pour trouver la meilleure coupure dans un nœud, un algorithme qui commence à partir d'une seule feuille et qui ajoute itérativement des branches à l'arbre est utilisé. Si on pose I_G et I_D respectivement les ensembles d'exemples des nœuds gauche et droit après la séparation, et $I = I_G \cup I_D$, la réduction de la perte après séparation est donnée par :

$$\text{Gain}_{\text{split}} = \frac{1}{2}\left[\frac{(\sum_{i\in I_G} g_i)^2}{\sum_{i\in I_G} h_i + \lambda} + \frac{(\sum_{i\in I_D} g_i)^2}{\sum_{i\in I_D} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma.$$

5.3.5 Avantages du modèle XGBoost

XGBoost introduit de nombreuses innovations pour améliorer l'efficacité et la robustesse du modèle et possède de nombreux avantages qui justifient une utilisation accrue en science des données. En premier lieu, le modèle XGBoost est un modèle rapide comparativement au *Tree Gradient Boosting* classique. En second lieu, il possède de nombreux

hyperparamètres que l'on peut choisir pour optimiser ses performances. On peut notamment citer :

- le nombre maximal d'itérations du *boosting*, permettant l'ajustement du modèle ;
- la profondeur maximale pour chaque arbre utilisé dans le modèle ;
- *eta* : le taux d'apprentissage qui régule l'influence de chaque nouvel arbre dans l'ensemble en contrôlant la vitesse de convergence de l'algorithme ;
- le paramètre *gamma* qui correspond au gain de perte minimal requis pour autoriser une partition supplémentaire dans un arbre ;
- la part des variables sélectionnées à chaque arbre : ce paramètre introduit de la diversité dans les arbres générés ;
- le poids total minimal d'observations dans un nœud ; ou encore
- la part des données utilisée pour l'entraînement de chaque arbre : il agit comme une forme de régularisation et permet de réduire le temps d'exécution.

5.3.6 Synthèse

En résumé, XGBoost combine la puissance des arbres de décision avec l'efficacité du *boosting*, tout en intégrant des techniques avancées de régularisation et d'optimisation computationnelle. Contrairement aux forêts aléatoires (*bagging*), qui construisent des arbres de manière indépendante, XGBoost construit chaque arbre de façon séquentielle, corrigeant à chaque étape les erreurs précédentes : il possède ainsi une forte capacité de modélisation tout en contrôlant le risque de sur-apprentissage, grâce à ses nombreux hyperparamètres.

Pour limiter le nombre de modèles complexes implémentés, en raison des contraintes opérationnelles et financières liées au traitement des bases de données, nous avons retenu XGBoost comme modèle de *machine learning* pour créer des scores de santé mensuels et annuels à partir des données de souscription et de sinistres. Ce modèle permet de générer des scores réalistes et fiables, mais aussi d'interpréter les relations entre les prédicteurs et la variable « cible », notamment grâce à la méthode de SHAP, décrite dans la section suivante.

5.4 Contribution des variables dans la prédiction : valeurs de Shapley

Lorsqu'un modèle de *machine learning* effectue une prédiction, toutes les variables d'entrée n'ont pas la même influence sur le résultat final : certaines ont, par exemple, un fort impact positif tandis que d'autres n'ont qu'un faible effet négatif. Contrairement à la régression linéaire ou aux GLM, où il est relativement simple d'interpréter l'effet de chaque variable grâce aux coefficients et à leur significativité statistique, les modèles plus complexes comme XGBoost rendent cette interprétation plus délicate. Les scores construits dans ce mémoire doivent répondre à deux exigences : la pertinence des scores construits et la capacité de ces modèles à quantifier l'influence du climat et de la pollution sur la santé. Or, il est plus difficile pour un modèle de *machine learning* comme XGBoost

de satisfaire le deuxième critère. C'est précisément là qu'intervient **SHAP** : cet outil, basé sur les **valeurs de Shapley**, permet d'attribuer de manière rigoureuse et quantitative l'importance de chaque variable dans la décision produite par le modèle, rendant ainsi interprétables des modèles qui ne le sont pas intrinsèquement.

5.4.1 Origine des valeurs de Shapley : la théorie des jeux

Cette section reprend les travaux et les notations de Lloyd Shapley qui introduit en 1953 un concept central de la théorie des jeux coopératifs et, plus tard, de l'analyse de la contribution des prédicteurs en *machine learning* : la valeur de Shapley [76].

Soit un jeu de coopération défini par le couple (P, v) où :

— $P = \{1, \ldots, p\}$ représente un ensemble de p joueurs ($p \in \mathbb{N}^*$) ; et

— $v : \mathcal{P}(P) \to \mathbb{R}$ est une fonction caractéristique telle que $v(\emptyset) = 0$, où $\mathcal{P}(P)$ désigne l'ensemble des sous-ensembles de P.

Un sous-ensemble $S \subseteq P$, appelé *coalition*, correspond à un échantillon de joueurs de P. L'ensemble P est appelé la grande coalition. La fonction $v(S)$ mesure l'importance ou la valeur générée par la coalition S. L'objectif est d'attribuer à chaque joueur une part du gain total généré par la grande coalition, de manière équitable. Pour cela, on cherche une application φ qui, à chaque jeu (P, v), associe un vecteur de contributions $\varphi = (\varphi_1, \ldots, \varphi_p)$, où φ_i correspond à la part attribuée au joueur i.

Lloyd Shapley a proposé en 1953 une manière de définir cette répartition équitable à l'aide de quatre axiomes fondamentaux :

— **Efficacité** : la somme des gains attribués à tous les joueurs est égale à la valeur totale de la grande coalition :

$$\sum_{i=1}^{p} \varphi_i(v) = v(P).$$

— **Symétrie** : deux joueurs i et j contribuant de façon identique à toutes les coalitions auxquelles ils appartiennent reçoivent la même part :

$$\forall S \subseteq P \setminus \{i, j\},\ v(S \cup \{i\}) = v(S \cup \{j\}) \implies \varphi_i(v) = \varphi_j(v).$$

— **Nullité (joueur inutile)** : si un joueur i n'apporte aucune contribution supplémentaire à aucune coalition, sa valeur attribuée est nulle :

$$\forall S \subseteq P \setminus \{i\},\ v(S \cup \{i\}) = v(S) \implies \varphi_i(v) = 0.$$

— **Additivité** : pour deux jeux caractérisés par v et w, la valeur attribuée à chaque joueur dans le jeu somme est la somme des valeurs attribuées dans chaque jeu séparément :

$$\varphi(v + w) = \varphi(v) + \varphi(w)$$

où $(v + w)(S) = v(S) + w(S)$ pour tout $S \subseteq P$.

La valeur de Shapley est la seule répartition qui satisfait simultanément ces quatre propriétés. Il s'agit d'un théorème démontré par Shapley, qui gagna le prix Nobel en sciences économiques pour ces travaux. Elle est donnée par la formule suivante, pour tout joueur $i \in P$:

$$\varphi_i(v) = \sum_{S \subseteq P \setminus \{i\}} \frac{(p - |S| - 1)!\ |S|!}{p!} \left[v(S \cup \{i\}) - v(S) \right].$$

La formule peut s'interpréter de la manière suivante : lorsqu'un joueur rejoint la coalition, il demande une compensation équitable correspondant à sa contribution marginale, c'est-à-dire la différence entre la valeur de la coalition avec lui et sans lui, soit $v(S \cup \{i\}) - v(S)$. La valeur de Shapley pour un acteur donné est alors la moyenne de cette contribution marginale, calculée sur l'ensemble des différents ordres possibles dans lesquels la coalition peut être constituée. Ainsi, une autre formulation, équivalente, consiste à considérer toutes les permutations possibles des joueurs et à mesurer la contribution marginale de i lorsqu'il rejoint la coalition formée par ses prédécesseurs dans chaque ordre :

$$\varphi_i(v) = \frac{1}{p!} \sum_{\pi \in \mathrm{Perm}(P)} \left[v(\mathrm{Préc}_i(\pi) \cup \{i\}) - v(\mathrm{Préc}_i(\pi)) \right]$$

où $\mathrm{Perm}(P)$ désigne l'ensemble des permutations de P, et $\mathrm{Préc}_i(\pi)$ est l'ensemble des joueurs apparaissant avant i dans la permutation π.

5.4.2 Exemple illustratif en théorie des jeux : le jeu des gants

Le **jeu des gants** est un jeu de coopération dans lequel les joueurs possèdent des gants pour la main gauche ou droite, et l'objectif est de former des paires de gants. Considérons l'ensemble des joueurs suivant :

$$P = \{1, 2, 3\}$$

où les joueurs 1 et 2 possèdent chacun un gant pour la main droite, tandis que le joueur 3 possède un gant pour la main gauche. La fonction de valeur, qui attribue une valeur à chaque coalition, est définie par :

$$v(P) = \begin{cases} 1 & \text{si } P \in \{\{1,3\}, \{2,3\}, \{1,2,3\}\} \\ 0 & \text{sinon} \end{cases}$$

Autrement dit, une coalition reçoit la valeur 1 si elle permet de constituer au moins une paire de gants, et 0 sinon. La formule générale pour calculer la valeur de Shapley d'un joueur i dans ce jeu est rappelée ci-dessous :

$$\varphi_i(v) = \frac{1}{p!} \sum_{\pi \in \mathrm{Perm}(P)} \left[v(\mathrm{Préc}_i(\pi) \cup \{i\}) - v(\mathrm{Préc}_i(\pi)) \right]$$

où l'on rappelle que $\mathrm{Perm}(P)$ désigne l'ensemble des permutations de P, et $\mathrm{Préc}_i(\pi)$ est l'ensemble des joueurs apparaissant avant i dans la permutation π. Le tableau 5.1 présente les contributions marginales du joueur 1 selon chacun des ordres possibles d'arrivée. On constate donc que la valeur de Shapley pour le joueur 1 est :

$$\varphi_1(v) = \frac{1}{6} \times (1) = \frac{1}{6}.$$

Par symétrie, la valeur de Shapley du joueur 2, ayant les mêmes caractéristiques que le joueur 1, est identique :

$$\varphi_2(v) = \varphi_1(v) = \frac{1}{6}.$$

Enfin, selon l'axiome d'efficacité qui stipule que la somme des valeurs de Shapley doit être égale à la valeur totale possible, ici 1, la valeur de Shapley du joueur 3 est :

$$\varphi_3(v) = 1 - \varphi_1(v) - \varphi_2(v) = 1 - \frac{1}{6} - \frac{1}{6} = \frac{2}{3}.$$

TABLE 5.1 – Contributions marginales du joueur 1 selon l'ordre d'arrivée

Perm(P)	**Contribution marginale de 1**
(1, 2, 3)	$v(\{1\}) - v(\varnothing) = 0 - 0 = 0$
(1, 3, 2)	$v(\{1\}) - v(\varnothing) = 0 - 0 = 0$
(2, 1, 3)	$v(\{1,2\}) - v(\{2\}) = 0 - 0 = 0$
(2, 3, 1)	$v(\{1,2,3\}) - v(\{2,3\}) = 1 - 1 = 0$
(3, 1, 2)	$v(\{1,3\}) - v(\{3\}) = 1 - 0 = 1$
(3, 2, 1)	$v(\{1,3,2\}) - v(\{3,2\}) = 1 - 1 = 0$

Cet exemple illustre clairement la façon dont la valeur de Shapley permet de mesurer la contribution équitable de chaque joueur dans un jeu coopératif. Dans le jeu présenté ci-dessus, les joueurs 1 et 2, qui possèdent chacun un gant droit, ont une contribution marginale limitée, car former une paire n'est possible qu'en présence du joueur 3, qui possède le gant gauche. Ainsi, la valeur de Shapley attribue à chacun d'eux une part égale, mais relativement faible du gain total. En revanche, le joueur 3 joue un rôle central puisqu'il est indispensable à la formation de toute paire. Sans lui, impossible d'arriver à ses fins. Sa contribution marginale est donc beaucoup plus importante, ce que reflète sa valeur de Shapley quatre fois supérieure à celle des deux autres joueurs. Cette répartition met ainsi en lumière le fonctionnement de la valeur de Shapley : elle récompense davantage les joueurs qui sont essentiels à la réalisation du but commun. Plus la contribution d'un joueur est importante, plus la valeur de Shapley associée sera élevée.

5.4.3 Valeurs de Shapley et interprétabilité des modèles de *machine learning*

Plusieurs méthodes issues de la théorie des jeux coopératifs ont été développées pour attribuer l'importance des variables dans la prédiction d'un modèle de *machine learning*. Sont repris dans cette section les travaux de Scott M. Lundberg et Su-In Lee qui, en 2017, ont été les premiers à proposer l'application des valeurs de Shapley à l'analyse de la contribution des prédicteurs en *machine learning* [77]. Leurs travaux s'inscrivent dans la continuité des résultats formulés par Strumbelj et Kononenko en 2013 [78].

Méthode classique d'estimation des valeurs de Shapley

Supposons que l'on dispose d'un jeu de données constitué de n individus, chacun décrit par d variables explicatives $\mathcal{X} = \{X_1, \ldots, X_d\}$, une variable cible y, et un modèle de prédiction f entraîné à partir de ces données. Les méthodes classiques, comme les *Shapley*

regression values, consistent à évaluer, pour chaque variable X_j, l'effet de sa présence ou de son absence sur la prédiction du modèle. Pour cela, on note $\mathcal{D} = \{1, \ldots, d\}$ et on considère tous les sous-ensembles $S \subseteq \mathcal{D} \setminus \{j\}$. On compare ensuite la différence de prédiction du modèle entraîné avec X_j présent ($f_{S\cup\{j\}}$) et celle sans x_j (f_S). La contribution marginale de X_j pour une observation donnée s'écrit alors :

$$f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S).$$

La valeur de Shapley pour la variable x_j est obtenue en prenant la moyenne pondérée de ces contributions sur tous les sous-ensembles possibles :

$$\phi_j = \sum_{S\subseteq\mathcal{D}\setminus\{j\}} \frac{|S|!(p-|S|-1)!}{p!} \left[f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S)\right].$$

Cette approche, bien que théoriquement exacte, est rarement praticable en raison du nombre exponentiel de combinaisons à considérer. En effet, il nécessite d'entraîner le modèle pour chaque sous-ensemble de variables.

Valeurs SHAP (SHapley Additive exPlanation)

Pour contourner cette difficulté, Lundberg et Lee ont proposé d'estimer la contribution marginale de chaque variable par des méthodes d'échantillonnage, sans avoir à entraîner le modèle à chaque fois. L'effet de la suppression d'une variable est alors approximé en intégrant sur la distribution des valeurs observées dans le jeu de données. Ils ont formalisé en 2017 le cadre des méthodes dites additives d'attribution de l'importance des variables, appelées **méthodes SHAP** [77]. Ils ont montré qu'une seule méthode satisfait simultanément quatre propriétés fondamentales, issues du théorème de Shapley :

— **Efficacité** : la somme des contributions attribuées à chaque variable (plus une constante de base ϕ_0) reconstitue la prédiction du modèle pour l'individu considéré :

$$f(x) = \phi_0 + \sum_{j=1}^{d} \phi_j$$

où ϕ_0 est la prédiction moyenne du modèle sur l'ensemble des données.

— **Symétrie** : deux variables explicatives qui ont une influence identique sur la prédiction auront des valeurs de contributions identiques.

— **Nullité** : si une variable est absente (x_j manquant), sa contribution doit être nulle ($\phi_j = 0$).

— **Additivité** : si, dans un nouveau modèle f', la contribution marginale d'une variable x_j augmente ou reste constante pour toutes les combinaisons possibles, alors la valeur attribuée à x_j ne doit pas diminuer :

$$\forall S \subseteq \mathcal{X}\setminus\{j\},\ f'_{S\cup\{j\}}(x_{S\cup\{j\}}) - f'_S(x_S) \geq f_{S\cup\{j\}}(x_{S\cup\{j\}}) - f_S(x_S) \implies \phi_j(f') \geq \phi_j(f).$$

La méthode **SHAP** (*SHapley Additive exPlanations*) s'appuie sur ces principes et définit la contribution de chaque variable X_j pour une observation x comme la valeur de Shapley appliquée à la fonction de prédiction conditionnelle du modèle :

$$\phi_j(f,x) = \sum_{S \subseteq \mathcal{X} \setminus \{j\}} \frac{|S|!(p-|S|-1)!}{d!} \left[\mathbb{E}[f(x) \mid x_{S\cup\{j\}}] - \mathbb{E}[f(x) \mid x_S]\right]$$

où $\mathbb{E}[f(x) \mid x_S]$ désigne l'espérance de la prédiction du modèle lorsque seules les variables de S sont connues (les autres étant remplacées par leur distribution marginale ou conditionnelle). Ainsi, pour chaque individu, la prédiction du modèle peut être décomposée comme :

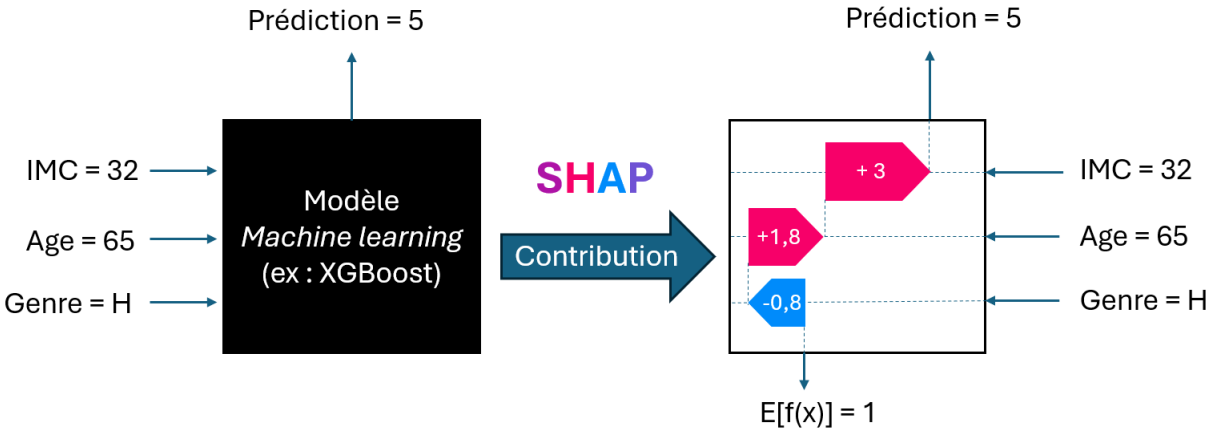
$$f(x) = \mathbb{E}[f(x)] + \sum_{j=1}^{d} \phi_j(f,x).$$

Les valeurs SHAP fournissent donc une attribution additive, rigoureuse et interprétable de l'importance de chaque variable dans la prédiction du modèle pour chaque observation.

5.4.4 Exemple concret d'utilisation en *machine learning*

Afin d'illustrer concrètement l'apport des valeurs de Shapley (ou valeur SHAP) à l'interprétabilité des modèles de *machine learning*, considérons le graphique ci-dessous, appliqué à un modèle XGBoost fictif. Les variables explicatives sont ici l'indice de masse corporelle (IMC), le genre et l'âge d'un individu, dont les valeurs sont respectivement 32 (IMC), Homme (genre) et 65 ans (âge). Le modèle fictif retourne pour cet individu une prédiction égale à 5.

FIGURE 5.3 – Exemple d'utilisation des valeurs de SHAP



Note de lecture : l'IMC de 32 contribue pour +3 à la prédiction finale, l'âge de 65 ans pour +1,8, tandis que le fait d'être un homme (genre : H) contribue de façon négative à hauteur de 0,8.

L'interprétation de cette prédiction via SHAP consiste à décomposer la valeur obtenue en une somme : d'une part, la valeur de référence (ou valeur attendue) du modèle, notée $\mathbb{E}[f(x)]$ (ici égale à 1), et d'autre part, les contributions attribuées à chaque variable. Ces contributions, appelées valeurs de Shapley, mesurent pour chaque variable l'effet marginal moyen de sa prise en compte dans toutes les combinaisons possibles de variables explicatives, conformément à la théorie des jeux coopératifs. Dans l'exemple représenté, l'IMC de 32 contribue pour $+3$ à la prédiction finale, l'âge de 65 ans pour $+1,8$ tandis que

le fait d'être un homme (genre : H) contribue de façon négative à hauteur de $-0,8$. La prédiction finale s'obtient ainsi en additionnant la valeur de référence et les contributions individuelles :

$$\text{Prédiction} = \mathbb{E}[f(x)] + \text{SHAP(IMC)} + \text{SHAP(Age)} + \text{SHAP(Genre)} = 1 + 3 + 1,8 - 0,8 = 5$$

Ce type de visualisation permet de comprendre de manière fine et transparente les facteurs qui expliquent la prédiction du modèle pour une observation donnée. On constate ici que c'est principalement l'IMC et l'âge qui tirent la prédiction vers le haut, tandis que le genre exerce un effet modérateur négatif. Cette décomposition additive respecte les propriétés fondamentales des valeurs de Shapley (additivité, nullité, cohérence) et offre une interprétabilité locale robuste, adaptée à l'analyse individuelle des prédictions issues de modèles complexes dits « boîte noire ».

5.5 Biais et équité des modèles de score

5.5.1 Constat des biais ethniques dans les modèles de score

Le système de santé américain et la construction de scores de santé peuvent être discriminatoires, notamment à l'égard des minorités ethniques (voir sections 1.1.5 et 2.1). Par ailleurs, il a été démontré que le dérèglement climatique ne touche pas tous les Américains de la même manière : la population afro-américaine, en particulier, souffre davantage des évènements climatiques que la population caucasienne (voir section 2.3). Ainsi, il semble important de garder à l'esprit le risque de développer des modèles discriminatoires. C'est pour cette raison que l'identification des biais ethniques constituera un critère important lors de la validation des modèles implémentés. **Ce mémoire se limite à l'identification des biais, sans traiter les méthodes de remédiation, qui ne relèvent pas de son périmètre**.

Dans un article publié dans *Science* en 2019, Obermeyer et ses co-auteurs ont analysé un algorithme couramment utilisé dans le système de santé américain pour identifier les patients à haut risque nécessitant un suivi renforcé [79]. Cet algorithme assigne un score de risque, similaire à un score de santé, basé principalement sur les coûts de santé futurs estimés à partir des données de facturation, sans inclure explicitement la variable indiquant l'ethnicité. Ils ont montré la présence d'un biais racial significatif : pour un score donné, les patients afro-américains sont considérablement plus malades que les patients caucasiens. En corrigeant cette disparité, le pourcentage de patients afro-américains recevant une aide supplémentaire passerait de 17,7 à 46,5 %. Se baser sur les coûts de santé futurs implique un biais en raison des barrières structurelles d'accès aux soins, de discriminations, ou de différences dans l'utilisation du système de soins : à niveau de santé égal, les patients afro-américains génèrent en moyenne moins de dépenses de santé que les patients caucasiens.

Dans la continuité de ces travaux, une étude réalisée par Erica Rode et Hans Leida pour Milliman en 2020 s'est concentrée sur les outils permettant de réduire les biais identifiés par Obermeyer, à savoir : le choix du modèle, la sélection de variables et le type de programme de santé concerné [80]. Ils ont également identifié les causes du décalage trouvé dans l'article de 2019 entre l'état de santé et les scores de santé. Selon eux, les

personnes afro-américaines reçoivent en moyenne des soins de moindre qualité et font moins confiance au système de santé, ce qui peut également limiter leur recours aux soins et la qualité du codage de leur dossier médical. Enfin, le fait que les algorithmes soient entraînés à prédire des coûts (et non la morbidité réelle) et utilisent des données liées aux actes passés ou aux coûts peut introduire des biais supplémentaires, sans pour autant améliorer la performance sur la santé réelle.

5.5.2 Calibrage et équité

Pour mesurer et identifier les biais dans les modèles prédictifs, la notion de calibrage joue un rôle central. Un modèle est dit calibré si, pour un score prédit donné, la fréquence observée de l'événement d'intérêt correspond à ce score. Dans les deux études mentionnées précédemment et dans un contexte d'équité algorithmique, le biais ethnique est mesuré à l'aide de courbes de calibrage faisant référence aux notions de ***calibration parity*** et de **calibrage par groupe**. Pour simplifier et illustrer la description de ces notions, nous nous plaçons dans un cadre de classification et l'on nomme :

- $X \in \mathcal{X} = \mathbb{R}^d$: l'ensemble de variables explicatives ;
- $S \in \mathcal{S} = \{1, \ldots, K\} = [K]$: un attribut sensible (comme le genre ou l'origine ethnique) ;
- $Y \in \mathcal{Y} = \{0, 1\}$: une variable réponse binaire ;
- $m : \mathcal{Z} \to [0, 1]$: un modèle prédictif estimant un score probabiliste, où $\mathcal{Z} = \mathcal{X} \times \mathcal{S}$;
- $\hat{Y} \in \{0, 1\}$: la variable réponse prédite après application d'une règle de décision (par exemple, un seuil) à $m(X, S)$.

Définition 5.5.1 (Calibrage : cas général [81]) *Un modèle $m : \mathcal{X} \to \mathcal{Y}$ est dit calibré si*

$$m(X) = \mathbb{P}_{Y|m(X)} \quad p.s.$$

La définition du calibrage général ci-dessus peut être appliquée à une classification binaire :

Définition 5.5.2 (Calibrage : classification binaire [82]) *Un modèle $m : \mathcal{X} \to [0, 1]$ (avec $\mathcal{Y} = \{0, 1\}$) est dit calibré si*

$$\mathbb{P}(Y = 1 \mid m(X)) = \mathbb{E}[Y \mid m(X)] = m(X) \quad p.s.$$

ou, de façon équivalente,

$$\mathbb{E}[Y \mid m(X) = p] = p, \quad \forall\, p \in [0, 1].$$

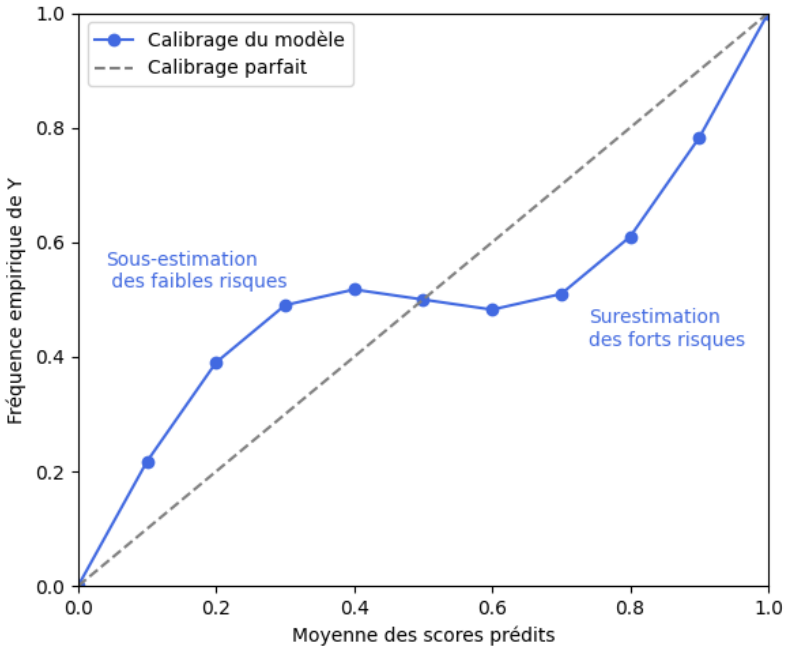
Pour évaluer le calibrage d'un modèle m, on peut tracer sa **courbe de calibrage** :

$$\forall p \in [0, 1], \quad g : \begin{cases} [0, 1] \to [0, 1] \\ p \mapsto g(p) := \mathbb{E}[Y \mid m(X) = p] \end{cases}$$

Un exemple de courbe réalisée avec des données simulées est donné en figure 5.4. Comme expliqué par Van Calster et ses co-auteurs en 2019, si le risque estimé de développer une maladie est de 20 %, alors, en réalité, 20 patients sur 100 doivent développer

l'évènement [83]. Si 40 patients sont trouvés atteints de la maladie, le risque est sous-estimé. Si, au contraire, 10 ont la maladie, le risque est surestimé.

FIGURE 5.4 – Exemple de courbe de calibrage



Note de lecture : à score prédit d'environ 0,2, la fréquence empirique est d'environ 0,4 indiquant une sous-estimation importante des faibles risques.

Dans le domaine de l'évaluation de la justesse et de l'équité des modèles prédictifs, plusieurs critères permettent de caractériser la relation entre les prédictions $\hat{Y}$, l'attribut sensible S et la variable d'intérêt Y : le **calibrage par groupe**, la ***sufficiency*** et la ***calibration parity*** occupent une place centrale pour analyser la performance et la neutralité des modèles vis-à-vis de groupes définis par une caractéristique sensible. Les définitions théoriques de ces trois notions fondamentales sont données ci-après.

Définition 5.5.3 (Sufficiency [81]) *Un modèle $m : Z \to Y$ satisfait le critère de sufficiency si la variable d'intérêt Y est indépendante de l'attribut sensible S conditionnellement à la prédiction $m(Z)$, pour la distribution $\mathbb{P}$ du triplet de variables (X, S, Y) :*

$$Y \perp\!\!\!\perp S \mid m(Z).$$

Définition 5.5.4 (*Calibration parity* [84]) *Un modèle binaire $m : Z \to [0, 1]$, où $Y \in \{0, 1\}$, satisfait la calibration parity pour l'attribut sensible S si, pour tout $s \in \mathcal{S}$ et pour toute valeur de $p \in [0, 1]$,*

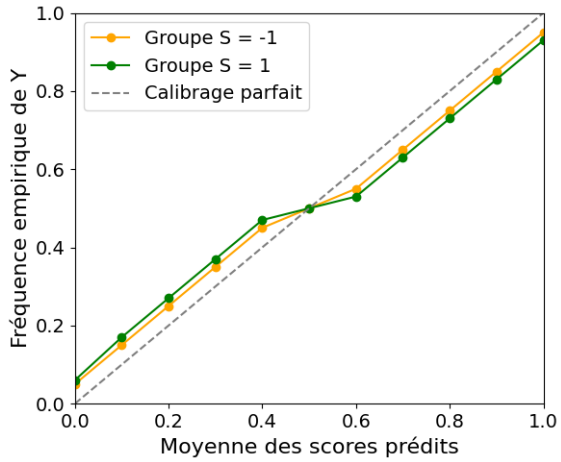
$$\mathbb{P}(Y = 1 \mid m(X, S) = p) = \mathbb{P}(Y = 1 \mid m(X, S) = p, S = s).$$

Définition 5.5.5 (Calibrage par groupe [85]) *Un modèle $m : Z \to [0, 1]$ binaire où $Y \in \{0, 1\}$ satisfait le calibrage par groupe pour l'attribut sensible S si pour tout $s \in \mathcal{S}$ et pour toute valeur de $p \in [0, 1]$,*

$$\mathbb{P}(Y = 1 \mid m(X, S) = p) = \mathbb{P}(Y = 1 \mid m(X, S) = p, S = s) = p.$$

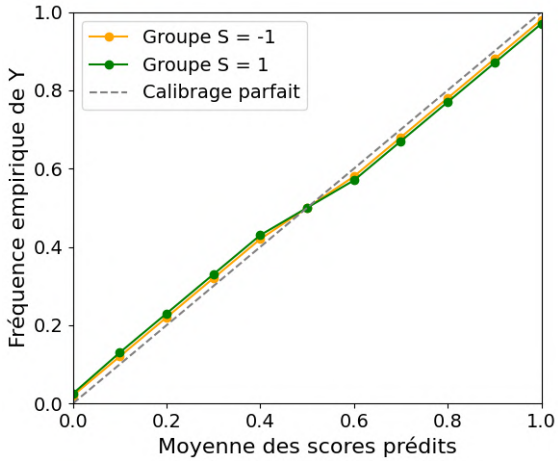
En reprenant les notations précédentes, on considère un modèle m estimant une probabilité. Le modèle dispose des variables explicatives (X, S) où $S \in \{-1, 1\}$ est une variable sensible (comme l'ethnicité). Dans le cas de la *calibration parity*, on souhaite que les erreurs de calibrage pour ces deux groupes protégés soient similaires. Un exemple de modèle respectant la *calibration parity* est donné en figure 5.5. Dans le cas du calibrage par groupe, on souhaite, d'une part, un calibrage similaire pour chaque groupe protégé, mais également un calibrage parfait (voir figure 5.6).

FIGURE 5.5 – Exemple d'un modèle respectant la *calibration parity*



Note de lecture : à score prédit d'environ 0,2, la fréquence empirique est d'environ 0,2 pour les deux groupes, avec sous-estimation des faibles risques.

FIGURE 5.6 – Exemple d'un modèle respectant le calibrage par groupe



Note de lecture : à score prédit d'environ 0,2, la fréquence empirique est 0,2 pour les deux groupes, avec estimation parfaite des risques.

5.5.3 Mesure des biais et outils d'évaluation

Pour évaluer les biais ethniques présents dans les modèles de scores implémentés, nous nous appuierons principalement sur l'analyse des courbes de calibrage précédemment présentées (figures 5.5 et 5.6). Ces courbes, tracées pour chaque groupe sensible, permettront de visualiser et de quantifier les écarts de calibrage entre deux groupes ethniques principaux : les Afro-Américains et les Caucasiens. Un point d'attention sera accordé au respect, *a minima*, de la *calibration parity*, et idéalement du calibrage par groupe, où le score prédit coïncide parfaitement avec la fréquence observée pour chaque groupe. Afin de quantifier les écarts de calibrage potentiels, nous utiliserons l'**erreur de calibrage espérée** (*Expected Calibration Error*), telle que proposée par Pakdaman Naeini et al. en 2015 [86]. Cette métrique, notée $ECE(m)$, mesure l'écart de calibrage parmi l'ensemble des groupes définis par la variable sensible S. Elle s'exprime comme suit :

$$\mathrm{ECE}(m) = \sum_{b \in [B]} \frac{n_b}{n} \cdot |\mathrm{freq}_b - \mathrm{conf}_b| \,,$$

où l'ensemble des données est partitionné en B intervalles (*bins*) selon les quantiles des scores prédits $m(x_i, s_i)$, n_b est le nombre d'observations dans le *bin* b, et n le nombre total d'observations. Pour chaque *bin* b :

— $\text{conf}_b = \frac{1}{n_b} \sum_{i \in b} m(x_i, s_i)$ est la confiance moyenne prédite ;

— $\text{freq}_b = \frac{1}{n_b} \sum_{i \in b} y_i$ est la fréquence empirique de l'événement.

5.5.4 Méthodes retenues pour l'identification des biais des modèles implémentés

L'identification des biais ethniques dans les modèles de score sera donc réalisée principalement à travers l'analyse du **calibrage par groupe** et de la ***calibration parity***. Ces métriques exigent l'indépendance entre la variable d'intérêt et l'attribut sensible à score prédit donné. Elles permettent ainsi de s'assurer que, pour un même niveau de risque estimé, les groupes protégés sont traités de façon équitable. Ainsi, l'évaluation des modèles s'appuiera à la fois sur les **courbes de calibrage** par groupe obtenues à l'aide de l'**ECE** et sur la vérification des critères de *calibration parity* ou de calibrage par groupe. Ils offrent un cadre robuste pour mettre en évidence d'éventuels biais dans les prédictions fournies par les algorithmes.

Troisième partie

Construction de scores de santé prenant en compte les risques émergents et résultats

Chapitre 6

Données climatiques et de pollution aux Etats-Unis

Ce chapitre fait directement suite à la revue de littérature du chapitre 3 présentant les effets du climat et de la pollution sur la santé ainsi que les indicateurs mensuels et annuels à intégrer dans nos modèles. Il décrit les données climatiques et de pollution américaines sélectionnées pour cette étude ainsi que la méthode utilisée pour agréger les données à la maille géographique des bases de *Milliman MedInsight*. Ce chapitre présente ensuite les indicateurs mensuels et annuels créés à partir des données disponibles en *Open Source*. Enfin, il détaille les résultats du regroupement géographique des zones étudiées en fonction de leurs caractéristiques climatiques et environnementales. L'ensemble de ces analyses vise à fournir une base solide pour l'intégration des facteurs climatiques et de pollution dans les modèles de score de santé développés dans les deux chapitres suivants.

6.1 Données climatiques

Les données climatiques relatives aux températures et aux précipitations utilisées dans ce mémoire ont été produites par la *National Oceanic and Atmospheric Administration* (NOAA). Des prévisions météorologiques quotidiennes aux alertes de tempêtes sévères, en passant par la surveillance du climat, la gestion des pêches, la restauration des zones côtières et le soutien au commerce maritime, les produits et services de la NOAA soutiennent la vitalité économique et influencent plus d'un tiers du produit intérieur brut (PIB) des Etats-Unis. La NOAA surveille le climat et la météo à l'échelle mondiale et collabore avec des partenaires du monde entier.

Le jeu de données spatiales *nClimGrid-Daily* est un ensemble de champs maillés journaliers et de moyennes de température à la surface et de précipitations couvrant les Etats-Unis à partir de 1951, mis à jour quotidiennement, et disponible publiquement sur leur site internet [1]. Avec une résolution d'environ 0,0417 degré de latitude et de longitude, soit environ une maille de 5 km, les données maillées offrent une représentation lissée des observations ponctuelles. Etant donné que la précision des estimations pour chaque maille et chaque jour peut être sensible à la variabilité spatiale locale, ainsi qu'aux techniques d'interpolation employées, la NOAA recommande l'utilisation du jeu

1. Données climatiques de la NOAA : `https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc%3AC01589/html#Coverage`

de données pour des applications impliquant une agrégation spatiale et/ou temporelle des estimations, comme les analyses de suivi climatique à l'échelle régionale ou nationale. Le jeu de données contient, pour chaque maille de 5 km, une valeur quotidienne de la température moyenne (*tavg*), minimale (*tmin*) et maximale (*tmax*) et de la somme des précipitations (*prcp*), du 1er janvier 2016 au 31 décembre 2023.

6.2 Données de pollution

Les données de pollution exploitées dans ce mémoire ont été produites par le Programme d'observation de la Terre de l'Union européenne *Copernicus*. Ce programme offre des services d'information basés sur l'observation de la Terre par satellite et les données in situ (non-spatiales). De vastes quantités de données mondiales provenant de satellites et de systèmes de mesure terrestres, aériens et maritimes sont collectées afin d'améliorer la qualité de vie des citoyens. Les services d'information fournis sont accessibles gratuitement et librement à leurs utilisateurs.

Le jeu de données spatiales *EAC4* est un ensemble de champs maillés mensuels, avec une résolution d'environ 0,75 degré, de moyennes de variables atmosphériques couvrant l'ensemble de la Terre de 2004 à 2024, et disponible publiquement sur leur site internet [2]. En lien avec la revue de littérature réalisée en amont, les moyennes mensuelles de concentrations d'ozone (O_3), de dioxyde d'azote (NO_2) et de $PM_{2.5}$ ont été extraites du jeu de données *EAC4* pour la période de janvier 2016 à décembre 2023.

Pour ce mémoire, il aurait été préférable d'utiliser des données quotidiennes de pollution aux Etats-Unis, à l'instar des données climatiques. Elles auraient permis de construire des indicateurs mensuels plus pertinents et précis. Cependant, ces données n'étaient pas disponibles à la fois à la maille géographique fine souhaitée et sur toute la période étudiée (2016-2023).

6.3 Agrégation des données climatiques et de pollution

Les données climatiques et de pollution ont été collectées à une résolution respective de 0,04° x 0,04° et de 0,75° x 0,75°. Un exemple de représentation des données journalières de précipitations est donné pour la journée du 28 juillet 2022 au Kentucky (voir figure 6.1). De fortes précipitations ont été observées dans l'est du Kentucky du 25 au 29 juillet. La région a reçu entre 10 et 20 centimètres de précipitations sur cette période, avec la rivière Kentucky atteignant un niveau record à Whitesburg (ZIP3 418) et Jackson (ZIP3 413). Le 2 août, environ 5 600 foyers et entreprises étaient toujours privés d'électricité et plus de 18 000 étaient toujours privés d'eau.

Dans le cadre de ce mémoire, ne possédant des informations géographiques que sur le ZIP3 des assurés, ces données climatiques et de pollution ont été agrégées pour correspondre à la maille du portefeuille étudié. L'agrégation des données a été réalisée à l'aide d'une moyenne pondérée, où chaque pixel de précipitations contribue proportionnellement à la surface qu'il occupe dans le ZIP3. Cette approche permet de calculer une

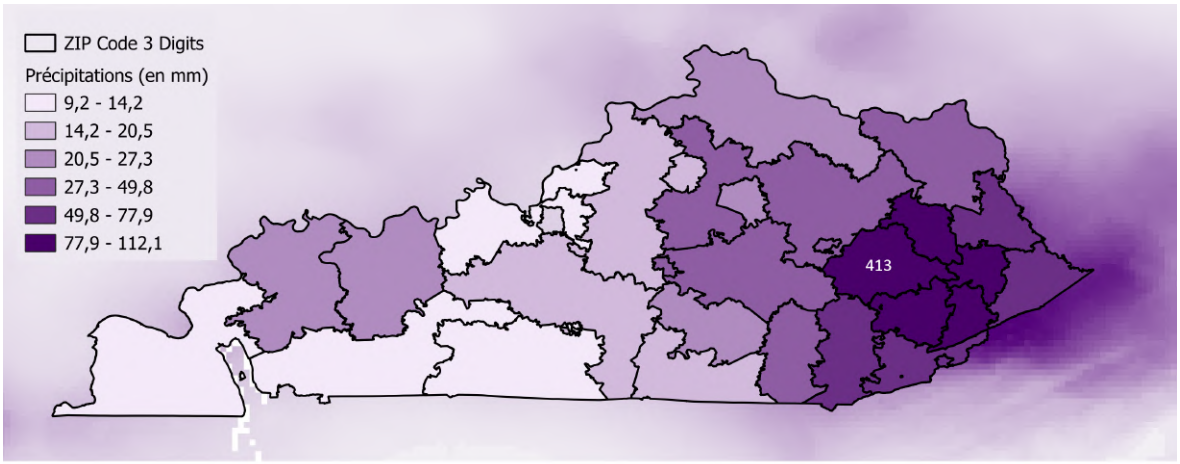
2. Données de pollution de *Copernicus* : `https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4-monthly?tab=overview`

moyenne représentative des précipitations pour chaque ZIP3, en tenant compte de la distribution géographique des données initiales. Ainsi, pour chaque ZIP3, une valeur de chaque variable climatique et de pollution est obtenue, quotidiennement pour les variables climatiques et mensuellement pour les variables de pollution. Une illustration est donnée par la figure 6.2 qui représente les valeurs obtenues en agrégeant les données de précipitations de la journée du 28 juillet 2022 utilisées dans la figure 6.1. Par exemple, dans le ZIP3 413, il est tombé en moyenne 103 millimètres de précipitations dans la journée du 28 juillet 2022.

FIGURE 6.1 – Précipitations le 28 juillet 2022 à une maille de 0,04°x0,04° au Kentucky

Note de lecture : le 28 juillet 2022, de forts cumuls de précipitations ont été observés dans l'est du Kentucky.

FIGURE 6.2 – Précipitations agrégées par *ZIP Code 3-Digits* le 28 juillet 2022 au Kentucky



Note de lecture : le 28 juillet 2022, il est tombé entre 77,9 et 112,1 mm de pluie dans les ZIP3 413, 414, 416, 417 et 418, à l'est du Kentucky, tandis qu'il n'est tombé qu'entre 9,2 et 14,2 mm de pluie dans le ZIP3 420, situé à l'extrême ouest.

6.4 Indicateurs climatiques et de pollution annuels et mensuels

6.4.1 Indicateurs annuels

Dans l'optique de construire des scores individuels de santé annuels, il a été nécessaire de construire, à partir des données climatiques et de pollution agrégées décrites dans les parties précédentes, des indicateurs annuels par ZIP3. Ces indicateurs annuels se basent sur les conclusions de la revue de littérature détaillée, effectuée dans le chapitre 3. A partir des données de températures, des indicateurs annuels de vagues de froid et de vagues de chaleur ont été définis. Aussi, pour chaque gaz, trois indicateurs annuels ont été construits, en notant $\overline{X}_m$ la concentration moyenne mensuelle du polluant X au mois m ($m = 1, \ldots, 12$) et S_X le seuil réglementaire associé :

— la concentration maximale observée parmi les moyennes mensuelles de l'année :

$$\max_{1 \leq m \leq 12} \left\{ \overline{X}_m \right\}$$

— la moyenne annuelle des concentrations mensuelles :

$$\frac{1}{12} \sum_{m=1}^{12} \overline{X}_m$$

— le nombre de mois où la concentration mensuelle dépasse le seuil réglementaire :

$$\sum_{m=1}^{12} \mathbb{I} \left(\overline{X}_m > S_X \right)$$

Les indicateurs retenus sont les suivants :

— `vague_froid` : nombre de vagues de froid dans l'année ;

— `vague_chaleur` : nombre de vagues de chaleur ;

— `somme_prcp` : somme annuelle des précipitations ;

— `max_annuel_NO2` : plus haute valeur mensuelle de la concentration moyenne de NO_2 sur l'année ;

— `moyenne_annuelle_NO2` : moyenne des concentrations mensuelles de NO_2 sur l'année ;

— `nb_mois_seuil_NO2` : nombre de mois où la concentration mensuelle de NO_2 dépasse le seuil réglementaire ;

— `max_annuel_O3` : plus haute valeur mensuelle de la concentration moyenne d'O_3 sur l'année ;

— `moyenne_annuelle_O3` : moyenne des concentrations mensuelles d'O_3 sur l'année ;

— `nb_mois_seuil_O3` : nombre de mois où la concentration mensuelle d'O_3 dépasse le seuil réglementaire ;

— `max_annuel_PM25` : plus haute valeur mensuelle de la concentration moyenne de $PM_{2.5}$ sur l'année ;

— `moyenne_annuelle_PM25` : moyenne des concentrations mensuelles de $PM_{2.5}$ sur l'année ; et

— `nb_mois_seuil_PM25` : nombre de mois où la concentration mensuelle de $PM_{2.5}$ dépasse le seuil réglementaire.

Un exemple de représentation de l'indicateur annuel `somme_prcp` est donné en figure 6.3 pour les ZIP3 401, situé au centre du Kentucky, 409 à l'est de l'Etat et 424 à l'extrême ouest. Dans chacune de ces régions, les précipitations augmentent depuis 2017 pour atteindre un pic en 2019-2020. En effet, plusieurs événements météorologiques ont contribué à l'augmentation des précipitations dans les différentes régions du Kentucky en 2019 :

— en mars, des tornades et des précipitations abondantes ont frappé l'ouest du Kentucky ;

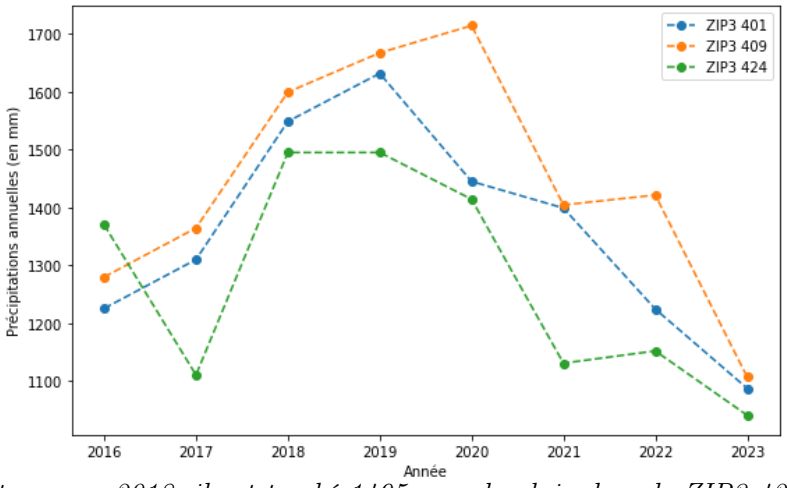
— le 21 juin, la fusion de deux lignes orageuses sur l'ouest et le centre du Kentucky a provoqué des dégâts importants et plusieurs tornades, s'accompagnant de pluies soutenues ;

— à l'automne, le passage des restes de la tempête tropicale Olga le 26 octobre a généré une ligne de précipitations intenses ;

— le 30 novembre, un épisode de basses pressions a apporté des pluies record dans le centre et le sud de l'Etat.

Par la suite, une diminution progressive des précipitations se manifeste jusqu'en 2023. Aussi, le ZIP3 409 enregistre des précipitations annuelles plus élevées que dans le ZIP3 401. Au contraire, le ZIP3 424 est moins sujet aux précipitations que les deux autres zones. Compte tenu de la localisation de ces régions, ces différences suggèrent des disparités climatiques d'est en ouest au sein du Kentucky, dont la largeur atteint 610 km.

FIGURE 6.3 – Graphique représentant l'indicateur annuel `somme_prcp` pour trois ZIP3 du Kentucky entre 2016 et 2024



Note de lecture : en 2018, il est tombé 1495 mm de pluie dans le ZIP3 424, 1599 mm dans le ZIP3 409 et 1549 mm dans le ZIP3 401.

6.4.2 Indicateurs mensuels

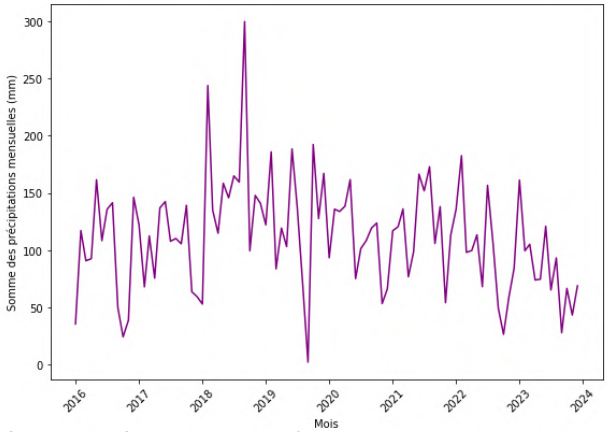
Dans l'optique de construire un score individuel de santé mensuel, il a été nécessaire de construire des indicateurs mensuels par ZIP3 à partir des données climatiques et de

pollution agrégées décrites dans les parties 6.1 et 6.2. Ces indicateurs se basent sur les conclusions de la revue de littérature détaillée, effectuée dans le chapitre 3. A partir des données de températures, des indicateurs mensuels de vagues de froid, de vagues de chaleur et de moyenne mensuelle des températures moyennes, minimales et maximales quotidiennes ont été définis. Aussi, pour chaque gaz, nous avons directement utilisé la valeur moyenne mensuelle des concentrations après agrégation des données de *Copernicus*. Les indicateurs mensuels créés sont les suivants :

— `vague_de_froid_mensuelle` : nombre de vagues de froid dans le mois ;
— `vague_de_chaleur_mensuelle` : nombre de vagues de chaleur dans le mois ;
— `somme_prcp_mensuelle` : précipitations mensuelles ;
— `moyenne_mensuelle_prcp` : valeur moyenne mensuelle des précipitations journalières ;
— `moyenne_mensuelle_tmin` : valeur moyenne mensuelle des températures minimales quotidiennes ;
— `moyenne_mensuelle_tmax` : valeur moyenne mensuelle des températures maximales quotidiennes ;
— `moyenne_mensuelle_tmoy` : valeur moyenne mensuelle des températures moyennes quotidiennes ;
— `moyenne_mensuelle_NO2` : valeur moyenne mensuelle des concentrations de NO_2 ;
— `moyenne_mensuelle_O3` : valeur moyenne mensuelle des concentrations d'O_3 ; et
— `moyenne_mensuelle_PM25` : valeur moyenne mensuelle des concentrations de $PM_{2.5}$.

Un exemple de représentation de l'indicateur `somme_prcp_mensuelle` est donné en figure 6.4. On y observe deux pics de précipitations mensuelles en février 2018 et en juillet 2018. En effet, le Kentucky a été frappé par deux périodes de précipitations intenses entraînant des inondations importantes dans le centre et l'ouest du pays. L'épisode orageux de février 2018 a duré plus de 5 jours avec des cumuls de pluie dépassant 250 mm à certains endroits sur cette période. *A contrario*, le mois de septembre 2019 fut particulièrement chaud et sec avec seulement 2,28 mm de précipitations tombées dans le mois et une température moyenne de 25,3 °C dans le ZIP3 405, ce qui en fait un des mois de septembre les plus chauds jamais enregistrés au Kentucky.

FIGURE 6.4 – Précipitations mensuelles dans le ZIP3 405 (2016-2024)



Note de lecture : il est tombé 2,28 mm de précipitations en septembre 2019 dans le ZIP3 405.

6.5 Interaction entre les températures et la pollution

Les températures et la pollution de l'air sont engagées dans des cycles de rétroaction [3] positifs et négatifs. Ces interactions sont à prendre en compte dans les modèles de scores de santé développés et dans les analyses des résultats.

6.5.1 Impact de l'augmentation de la pollution sur les températures

Les vagues de chaleur intenses favorisent l'émission naturelle et anthropique de polluants directement émis par une source appelés polluants primaires, ainsi que celle des polluants formés à partir de la réaction entre plusieurs polluants appelés polluants secondaires.

Les vagues de chaleur favorisent le développement de feux de forêt. Ces feux de forêt émettent des polluants primaires dégradant la qualité de l'air. En effet, l'Organisation mondiale de la météorologie dépeint le rôle des incendies de forêt observés en 2023 en Amérique du Nord dans la dégradation de la qualité de l'air. Ils auraient entraîné des anomalies de concentration en $PM_{2.5}$ allant jusqu'à $+18$ µg/m^3 par rapport à la période de référence 2003-2022 [87].

Par ailleurs, les températures élevées favorisent la formation de polluants secondaires, en accélérant les réactions chimiques nécessaires à leur formation. Par exemple, sous un rayonnement ultraviolet important, la formation d'ozone troposphérique, un polluant secondaire issu de la réaction entre les oxydes d'azote (NOx), le monoxyde de carbone et les composés organiques volatils, est accélérée. Le projet franco-italien *Climaera*, réalisé sous l'égide de l'Union européenne en 2020, a mis en évidence qu'à émissions de 2013 constantes, la météorologie de 2030 pourrait augmenter la concentration en PM_{10} de $11{,}5 \pm 3{,}1$ µg/m^3 par rapport à 2013 en Auvergne-Rhône-Alpes et dans le Piémont [88]. Les épisodes de chaleur augmentent ainsi non seulement l'émission de polluants atmosphériques, mais aussi leur concentration.

6.5.2 Impact de l'augmentation des températures sur la pollution

Réciproquement, la pollution peut amplifier ou atténuer la montée des températures. Certains polluants atmosphériques tels que l'ozone troposphérique s'avèrent être aussi des gaz à effet de serre (GES). Ces gaz renforcent l'effet de serre et accélèrent l'augmentation des températures, créant une rétroaction positive. Aussi, certaines particules en suspension (PM) qui ne sont pas des gaz à effet de serre, telles que le carbone noir, ont un effet de réchauffement en absorbant la lumière, tandis que les aérosols de sulfates et les nitrates refroidissent la planète en réfléchissant la lumière vers l'espace. J. Gao et ses co-auteurs ont étudié l'impact de l'augmentation du forçage radioactif [4] (ERF) sur les

3. Un cycle de rétroaction positive se produit lorsque l'augmentation d'un phénomène X contribue à l'aggravation d'un phénomène Y, et réciproquement, le phénomène Y amplifie à son tour le phénomène X, créant ainsi un cercle vicieux ou vertueux. A l'inverse, dans un cycle de rétroaction négative, l'augmentation d'un phénomène X entraîne celle d'un phénomène Y, mais ce dernier agit pour atténuer le phénomène X, stabilisant ainsi le système.

4. Le forçage radiatif, en climatologie, désigne la différence entre l'énergie radiative entrante (rayonnement solaire) et l'énergie radiative sortante (rayonnement infrarouge) au niveau de la surface terrestre,

températures moyennes [89]. Les résultats de leur étude sont résumés dans le tableau 6.1. A titre d'exemple, un ERF supplémentaire de 0,81 ± 0,92W/m^2 dû aux augmentations d'O_3 dans la basse troposphère a renforcé le réchauffement climatique de 0,07 ± 0,09 °C dans l'est de la Chine.

TABLE 6.1 – Caractéristiques des principaux polluants et leur effet sur les températures

Polluant	**GES ?**	**Polluant atmosphérique ?**	**Effet sur les températures**
Dioxyde de carbone	Oui	Non	+
Méthane	Oui	Non	+
Protoxyde d'azote	Oui	Non	+
Gaz fluorés	Oui	Non	+
Ozone	Oui	Oui	+
Carbone noir (PM)	Non	Oui	+
Aérosol de sulfates (PM)	Non	Oui	–
Nitrates (PM)	Non	Oui	–

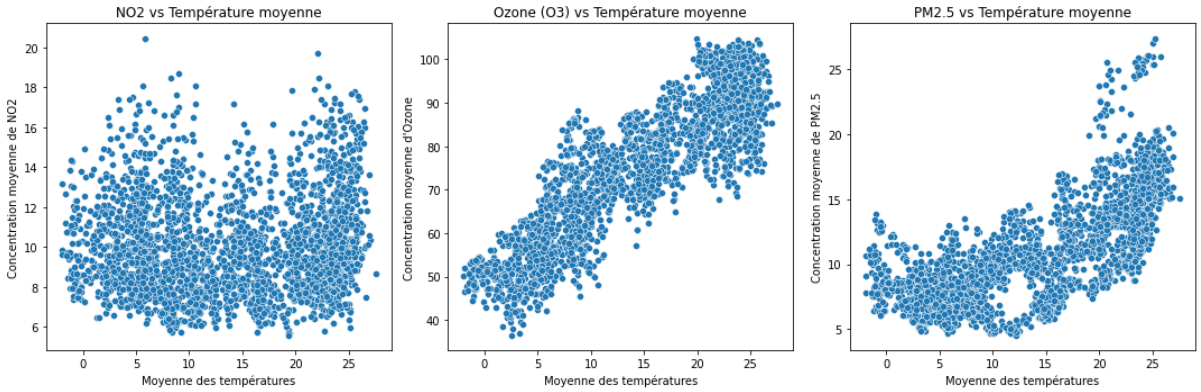
Ainsi, l'effet de la pollution sur la température n'est pas monotone, mais reste assez complexe puisque certains polluants ont un impact positif sur les températures tandis que d'autres entraînent des baisses de température.

6.5.3 Interactions observées dans les données utilisées dans le cadre de ce mémoire

Créée à partir des indicateurs mensuels décrits dans la section précédente, la figure 6.5 illustre le lien existant entre les températures et les concentrations moyennes de certains polluants, notamment l'ozone et les particules fines. Elle montre dans un premier temps une répartition des concentrations de NO_2 relativement dispersée selon la température moyenne, sans tendance claire. Ainsi, les données analysées ne permettent pas d'établir un lien entre la concentration moyenne mensuelle de NO_2 et les températures moyennes mensuelles. Dans un second temps, la figure met en évidence une corrélation positive entre la température et la concentration d'ozone. Les concentrations d'ozone tendent à augmenter avec la température moyenne, ce qui confirme le rôle des fortes chaleurs dans la formation de ce polluant secondaire, comme évoqué précédemment. Enfin, elle révèle une relation non-linéaire entre la concentration de $PM_{2.5}$ et la température. Cette courbe suggère que les mécanismes de formation et de dispersion des particules fines sont complexes et peuvent varier selon la saison ou les conditions météorologiques. En effet, les émissions dues au chauffage peuvent augmenter en hiver. A l'inverse, ces émissions diminuent généralement en été, mais les concentrations de $PM_{2.5}$ peuvent toutefois croître durant la saison estivale à cause des feux de forêt notamment ou de la stagnation de l'air lors des épisodes de canicule.

mesurée en watts par mètre carré (W/m^2).

FIGURE 6.5 – Relation entre les températures moyennes mensuelles et les concentrations moyennes mensuelles de NO_2, d'O_3, et de $PM_{2.5}$ au Kentucky



Note de lecture : les températures moyennes mensuelles croissent avec les concentrations moyennes mensuelles d'ozone.

En résumé, ces observations confirment que les interactions entre températures et pollution diffèrent selon les polluants, et qu'il est essentiel de les prendre en compte dans l'interprétation des impacts climatiques et sanitaires.

6.6 Regroupement géographique des ZIP3 en fonction des variables climatiques

Cette section vise à analyser plus finement les données climatiques et de pollution du Kentucky. L'objectif est d'extraire des caractéristiques propres à des groupes de ZIP3 au sein de l'Etat afin d'établir des profils climatiques et des profils de pollution atmosphérique. Les ZIP3 seront ainsi regroupés en plusieurs ensembles appelés clusters présentant chacun des caractéristiques climatiques spécifiques et donc potentiellement des profils sanitaires différents. Dans cette section, un cluster (groupe ou grappe en français) désigne un ensemble d'éléments similaires ou proches selon certains critères.

6.6.1 Description d'une méthode de partitionnement : le *DTW clustering*

Une méthode est particulièrement utilisée pour regrouper des éléments en fonction de séries temporelles multivariées. Il s'agit du ***DTW clustering*** ou partitionnement DTW. Cette approche compare les séries temporelles associées à chaque élément climatique à l'aide de la distance dynamique de type « *Dynamic Time Warping* » (DTW). Cette mesure quantifie la similarité entre deux séries temporelles, même si elles présentent des décalages dans le temps, ce qui n'est pas le cas dans cette étude. Le *DTW clustering* permet ainsi de mieux capturer les motifs communs ou divergents des séries temporelles associées à chaque variable climatique ou de pollution. A partir de la distance DTW est obtenue une matrice permettant ensuite de réaliser un regroupement à l'aide de la méthode *K-Means* par exemple. Des groupes d'éléments dont les séries temporelles présentent des évolutions similaires sont alors constitués.

Le nombre de clusters est un hyperparamètre important dans le partitionnement. Le Kentucky possédant 27 codes ZIP3 différents, nous aimerions idéalement obtenir entre 3 et 5 groupes climatiques distincts. Plusieurs méthodes existent pour trouver la valeur optimale du nombre de clusters ou pour s'assurer que la valeur choisie est satisfaisante. Nous décrirons ici la méthode de la **silhouette**. Cette méthode consiste à calculer les coefficients de silhouette de chaque point. Ces coefficients mesurent la similitude d'un objet avec son propre cluster (cohésion) par rapport à d'autres clusters (séparation).

L'indice de silhouette $s(i)$ pour un élément i est défini par :

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}$$

où :

— $a(i)$ représente la distance DTW moyenne entre i et les autres éléments de son propre cluster ; et

— $b(i)$ représente la plus petite distance DTW moyenne entre i et les éléments de tous les autres clusters (c'est-à-dire, le plus petit score moyen de dissimilarité entre i et un autre cluster.

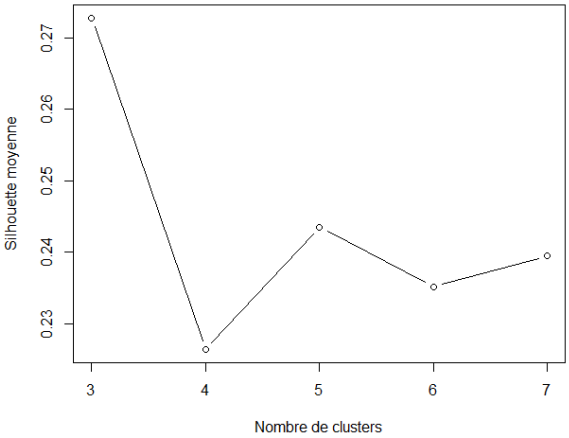
Une moyenne est ensuite effectuée sur tout l'échantillon pour obtenir le score de silhouette global (ou silhouette moyenne). Ce score est compris entre -1 et 1. Une valeur élevée indique que l'objet est bien adapté à son propre cluster et peu similaire aux clusters voisins. Une représentation graphique de ce score en fonction du nombre de clusters permet de déterminer la partition optimale, c'est-à-dire celle qui maximise la silhouette moyenne.

6.6.2 Résultat du partitionnement sur les données brutes

L'objectif de cette section est de présenter les résultats des partitionnements DTW des zones ZIP3 du Kentucky effectués à partir des données climatiques et de pollution brutes, décrites dans les sections 6.1 et 6.2. Un premier partitionnement DTW des ZIP3 a été implémenté sur les données quotidiennes climatiques (*tavg, tmin tmax* et *prcp*). Un deuxième partitionnement a été effectué à partir des niveaux de pollution mensuels de NO_2, d'O_3 et de $PM_{2.5}$.

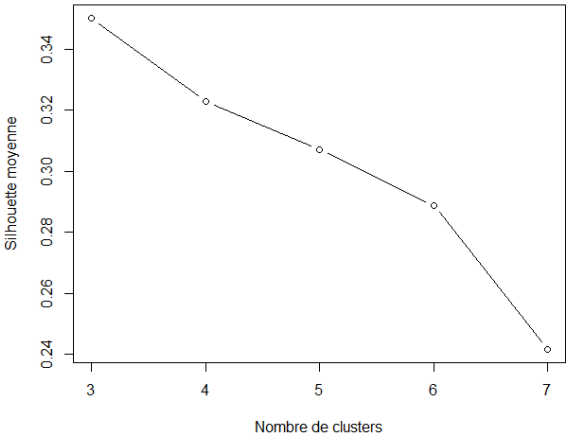
Pour chaque type de données, une analyse du coefficient de silhouette moyen a été réalisée afin de déterminer le nombre optimal de clusters à retenir pour le partitionnement. Différents nombres de clusters ont été testés (entre 3 et 7 inclus) et la silhouette moyenne a été calculée pour chacun. Comme illustré sur la figure 6.6, un nombre de clusters égal à 3 maximise la silhouette sur l'intervalle testé pour le partitionnement DTW à partir des données climatiques. De même, pour les données de pollution, la figure 6.7 suggère que trois clusters constituent une solution justifiée, puisque ce choix maximise la silhouette moyenne sur les tests effectués.

FIGURE 6.6 – Choix du nombre de clusters dans le partitionnement DTW avec les données climatiques



Note de lecture : pour un nombre de clusters égal à 3, la silhouette moyenne est de 0,273.

FIGURE 6.7 – Choix du nombre de clusters dans le partitionnement DTW avec les données de pollution

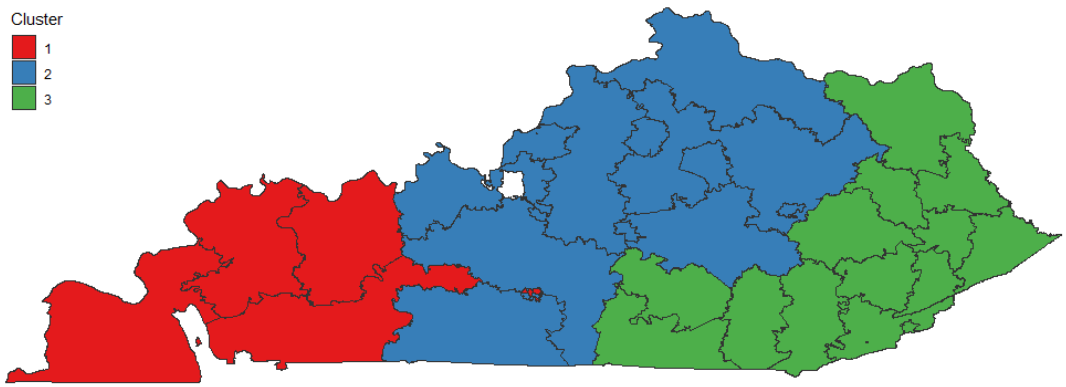


Note de lecture : pour un nombre de clusters égal à 3, la silhouette moyenne est de 0,350.

Ainsi, au regard des graphiques ci-dessus, il paraît pertinent de choisir 3 groupes de ZIP3 pour diviser le Kentucky en zones distinctes en fonction de leurs caractéristiques climatiques et de la qualité de l'air. La figure 6.8 représente les trois groupes obtenus à l'issue du *DTW clustering* sur les données journalières de températures et de précipitations. Le partitionnement sur les données mensuelles de pollution a donné exactement la même répartition. La carte de la figure 6.8 sera donc utilisée pour évoquer les deux regroupements réalisés. Ces résultats traduisent une structuration spatiale cohérente des dynamiques climatiques et de pollution à l'échelle des ZIP3. En effet, le partitionnement sépare clairement le Kentucky en trois couloirs verticaux :

— le **cluster 1** regroupe les ZIP3 situés à l'extrême ouest du Kentucky ;
— le **cluster 2** rassemble les ZIP3 situés au centre et s'étalant du sud au nord de l'Etat ; et
— le **cluster 3** est constitué des ZIP3 du sud-est et de l'extrême est du Kentucky.

FIGURE 6.8 – Partitionnement selon la méthode de *DTW Clustering* sur les données journalières de températures et de précipitations



Note de lecture : les ZIP3 420, 422, 423, 424 situés à l'extrême ouest du Kentucky constituent le cluster n°1 (en rouge).

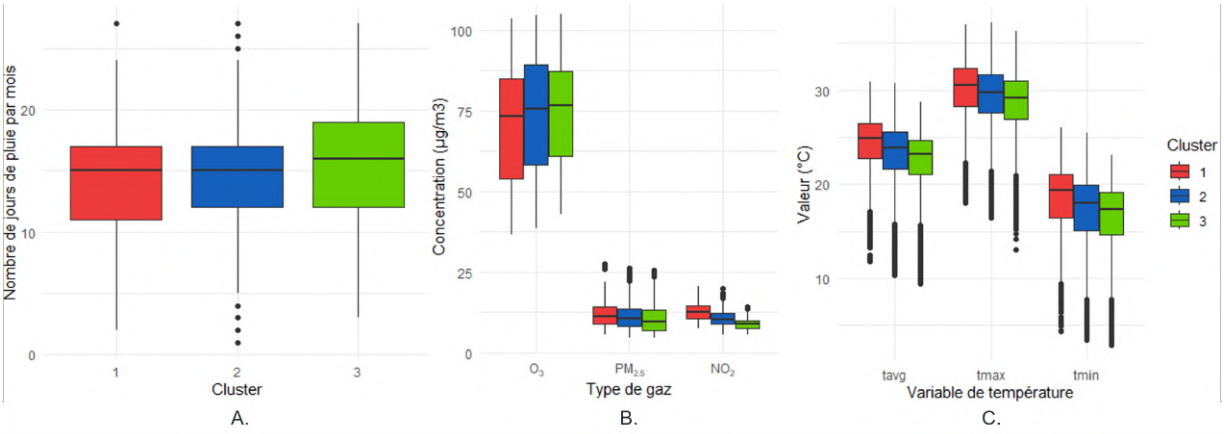
Après avoir identifié les groupes homogènes de ZIP3 à l'aide du partitionnement DTW, il est essentiel de caractériser chacun des clusters obtenus en termes de conditions climatiques et de pollution atmosphérique. Pour ce faire, les statistiques descriptives de chaque variable ont été calculées pour chaque cluster à l'aide de diagrammes en boîte. Ces graphiques permettent de comparer rapidement les clusters entre eux et de mieux comprendre leurs spécificités. Ces boxplots sont présentés en figure 6.9 et permettent de dresser, pour chaque cluster, une typologie climatique et environnementale précise :

— **Cluster 1** (ouest du Kentucky) : regroupant 4 ZIP3 à l'ouest du Kentucky, il est caractérisé par des températures moyennes plus élevées et des précipitations journalières plus faibles que les autres clusters. Il présente également des concentrations plus importantes de $PM_{2.5}$ et de NO_2, mais des niveaux plus faibles d'O_3. Le cluster 1 correspond donc à une zone plus chaude, moins pluvieuse et davantage exposée à la pollution, à l'exception de l'ozone.

— **Cluster 2** (centre et nord du Kentucky) : regroupant 10 ZIP3 au centre de l'Etat, il présente des températures moyennes modérées et des niveaux de précipitations intermédiaires. Les concentrations de $PM_{2.5}$ et de NO_2 sont, elles aussi, intermédiaires, tandis que les concentrations d'ozone sont plus élevées. Ce groupe correspond à une zone tempérée, où la pollution aux $PM_{2.5}$ et au NO_2 est moyenne. L'ozone y atteint toutefois des valeurs élevées.

— **Cluster 3** (sud-est et est du Kentucky) : regroupant 13 ZIP3, il se distingue par des températures moyennes plus basses et des précipitations plus abondantes que les deux premiers clusters. Les concentrations de $PM_{2.5}$ et de NO_2 y sont les plus faibles, avec des niveaux d'O_3 intermédiaires. Ce cluster représente donc une zone plus froide, plus humide, et globalement moins exposée à la pollution atmosphérique.

FIGURE 6.9 – Partitionnement selon la méthode de *DTW Clustering* sur les données journalières de températures et de précipitations



Note de lecture : A. Nombre de jours de pluie mensuels par cluster ; B. Distribution des concentrations de gaz par cluster ; C. Distribution des températures en été par variable et par cluster.

6.6.3 Résultat du partitionnement sur les indicateurs mensuels

Pour vérifier la cohérence avec les résultats mis en exergue dans la partie précédente, un *DTW clustering* a également été réalisé sur les indicateurs mensuels décrits en partie 6.4.2 excepté les indicateurs de pollution étant donné qu'ils correspondent aux données

brutes agrégées (voir partie 6.6.2 pour le partitionnement DTW sur ces données). Le partitionnement a donné exactement la même répartition que celle décrite par la figure 6.8. L'analyse plus détaillée des indicateurs mensuels par cluster est disponible en annexe (voir annexe C). Elle confirme les conclusions de la partie précédente en ce qui concerne les données climatiques.

Chapitre 7

Scores de santé annuels basés sur les données de souscriptions

Une première approche a été d'exploiter les bases de données « assurés » et « souscriptions » pour créer des indicateurs de santé annuels basés uniquement sur les informations transmises par les assurés lors de la souscription, sans utiliser les données de sinistres.

7.1 Base de données

7.1.1 Jeu de données

Les bases utilisées pour développer les scores de santé annuels sont les bases « assurés » et « souscriptions » concernant 551 335 assurés du Kentucky. Pour rappel, la base « souscriptions » fournit des informations mensuelles entre le 1er janvier 2017 et le 31 décembre 2023 sur les types d'assurances souscrites par chaque individu de la base et sur le groupe de pathologies chroniques associé à chaque souscription. Chaque observation correspond à une personne donnée pour un mois donné, ce qui permet de suivre finement l'évolution de la couverture d'assurance et des pathologies déclarées au fil du temps. Pour pallier les contraintes computationnelles et financières liées au traitement d'un grand volume de données, un échantillon représentatif de 50 000 assurés du Kentucky a été constitué comme suit :

— 50 % d'hommes et 50 % de femmes ;

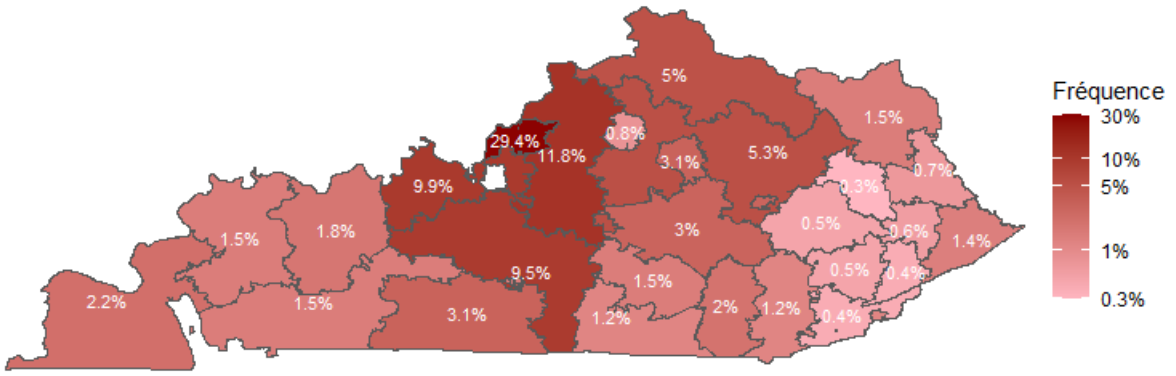
— 8 % d'Afro-Américains, 92 % de Caucasiens.

La figure 7.1 représente la répartition de l'échantillon représentatif du Kentucky par ZIP3. La majorité des assurés de l'échantillon réside dans le centre de l'Etat, qui regroupe les plus grandes villes. L'ouest et l'est du Kentucky sont moins représentés, en raison d'une plus faible densité de population.

La variable clé de la base « souscriptions » qui donne le groupe de maladies chroniques associé à la souscription mensuelle classant chaque individu dans une catégorie de diagnostic selon ses conditions médicales principales, constitue la variable de référence pour la construction du score annuel. En agrégeant les données mensuelles à la maille temporelle annuelle, il est possible de comptabiliser pour chaque assuré, le nombre de conditions chroniques différentes déclarées au cours de l'année. Puisqu'une seule pathologie peut être déclarée par mois, cet indicateur varie de 0 (aucune pathologie déclarée

dans l'année) à 12 (une pathologie différente chaque mois). L'objectif est d'utiliser cette variable comme proxy [1] du niveau de morbidité annuel, et constitue la cible d'un modèle de prédiction visant à attribuer un score de santé annuel à chaque personne.

FIGURE 7.1 – Répartition des membres de la base MedInsight résidant dans le Kentucky par Code ZIP3



Note de lecture : 29,4 % des individus du Kentucky la base MedInsight résident dans le Code ZIP3 402.

Pour développer ce modèle, il est nécessaire de construire des indicateurs annuels qui constitueront les variables prédictives potentielles du modèle implémenté. Ainsi, pour chaque année, de 2017 à 2023, les variables socio-démographiques, sanitaires et environnementales suivantes ont été collectées :

- `CC_N` (variable cible) : nombre de conditions chroniques à l'année N ;
- `CC_N_1` : nombre de conditions chroniques à l'année N-1 ;
- `ZIP3` : ZIP3 de résidence de l'assuré ;
- `age` : âge des individus dans l'échantillon ;
- `genre` : genre des individus :
  - 0-Femme
  - 1-Homme
- `vague_froid` : nombre de vagues de froid dans l'année N-1 ;
- `vague_chaleur` : nombre de vagues de chaleur dans l'année N-1 ;
- `somme_prcp` : somme des précipitations sur l'année N-1 ;
- `max_annuel_NO2` : concentration maximale mensuelle de NO_2 dans l'année N-1 ;
- `moyenne_annuelle_NO2` : concentration moyenne mensuelle de NO_2 dans l'année N-1 ;
- `nb_mois_seuil_NO2` : nombre de mois où le seuil de NO_2 a été dépassé dans l'année N-1 ;
- `max_annuel_O3` : concentration maximale mensuelle d'O_3 dans l'année N-1 ;
- `moyenne_annuelle_O3` : concentration moyenne mensuelle d'O_3 dans l'année N-1 ;

1. un proxy est une variable mesurable utilisée à la place d'une autre variable non observée ou difficile à mesurer, afin d'en estimer l'impact ou la valeur dans un modèle.

— `nb_mois_seuil_O3` : nombre de mois où le seuil d'O_3 a été dépassé dans l'année N-1 ;
— `max_annuel_PM25` : concentration maximale mensuelle de $PM_{2.5}$ dans l'année N-1 ;
— `moyenne_annuelle_PM25` : concentration moyenne mensuelle de $PM_{2.5}$ dans l'année N-1 ; et
— `nb_mois_seuil_PM25` : nombre de mois où le seuil de $PM_{2.5}$ a été dépassé dans l'année N-1.

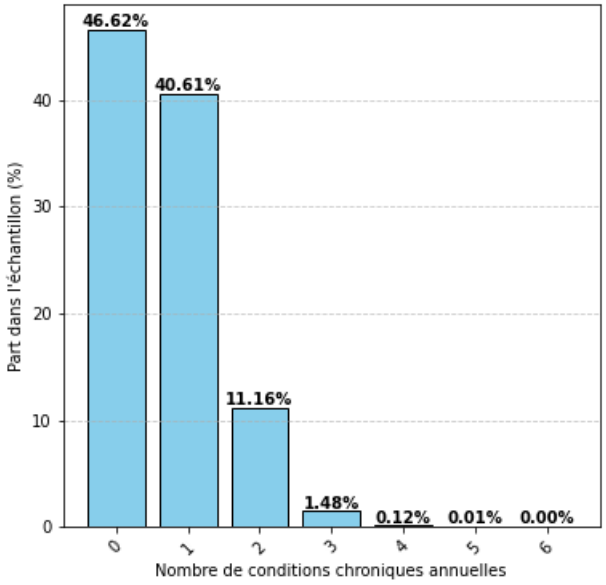
Une base de données annuelle contenant des informations annuelles sur chaque assuré de l'échantillon représentatif du Kentucky est ainsi obtenue. Elle comprend 179 205 observations. Il est important de noter que les indicateurs climatiques et de pollution sont décalés d'une année par rapport à la variable `CC_N` : autrement dit, pour chaque année N, la variable cible (`CC_N`) correspond au nombre de pathologies déclarées cette année, tandis que les variables environnementales et climatiques utilisées comme prédicteurs correspondent à l'année précédente (N-1). Cela permet d'étudier l'effet différé des conditions environnementales sur la santé. Un exemple fictif est donné dans le tableau 7.1 : il s'agit d'une femme de 65 ans, vivant dans le ZIP3 401 et qui, en 2020, présente deux conditions chroniques et en 2019, une seule condition chronique. Dans le ZIP3 401, la concentration mensuelle en NO_2 a dépassé les seuils 10 fois et trois vagues de chaleur ont été recensées en 2019.

TABLE 7.1 – Exemple d'une observation de la base annuelle créée

ID	ZIP3	age	genre	CC_N	CC_N_1	vague_chaleur	nb_mois_seuil_NO2	...
0000	401	65	1	2	1	3	10	...

La distribution de la variable cible dans la base annuelle créée et décrite ci-dessus est donnée dans la figure 7.2 : on comptabilise 46,6 % de valeurs nulles et 40,6 % de valeurs unitaires. Les valeurs supérieures ou égales à 2 représentent donc 12,8 % des observations.

FIGURE 7.2 – Distribution de la variable `CC_N`



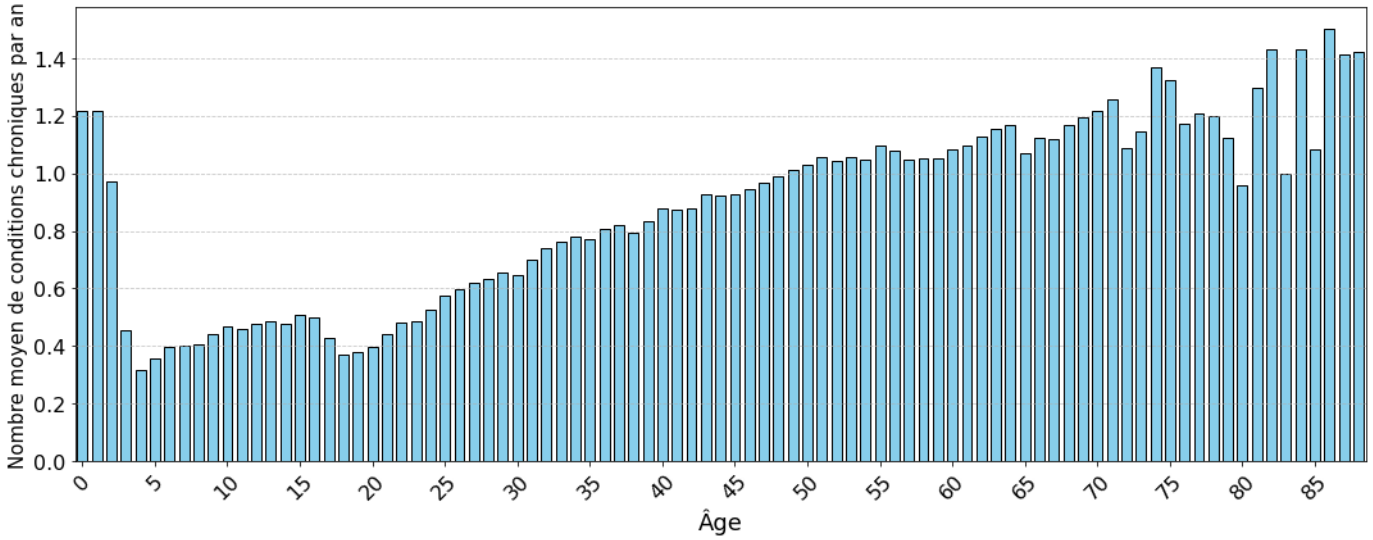
Note de lecture : 46,52 % des observations sont nulles pour la variable `CC_N`.

7.1.2 Statistiques descriptives en lien avec la variable cible

L'objectif de cette section est d'analyser le comportement et les principales caractéristiques de la variable cible en fonction des 15 variables prédictives (`age`, `genre`, `CC_N_1` ainsi que les 12 variables environnementales). Sont présentées ici les statistiques descriptives relatives à la variable cible de notre étude. Ces analyses descriptives permettent de mieux comprendre la structure des données et d'identifier d'éventuelles tendances. Elles constituent une étape essentielle pour interpréter les sorties des modèles de score implémentés par la suite.

La figure 7.3 illustre la relation entre l'âge des assurés et le nombre moyen de conditions chroniques déclarées annuellement. Le graphique met en exergue une relation très marquée : la fréquence des conditions chroniques est très élevée pour les très jeunes enfants (0 à 2 ans) suivie d'une chute rapide dès l'âge de 3 ans. Chez les enfants et les jeunes adultes, le nombre moyen de conditions chroniques reste relativement faible (autour de 0,4 en moyenne). Cependant, à partir de la trentaine, le nombre moyen de conditions chroniques augmente petit à petit avec l'âge, traduisant une vulnérabilité accrue des personnes plus âgées. Cette tendance s'accentue à la cinquantaine et se pérennise après 70 ans. La forme de la courbe, rappelant celle d'une courbe de mortalité, reflète ainsi la détérioration graduelle de l'état de santé au fil du vieillissement. Cette analyse justifie, d'une part, l'importance de l'âge comme facteur explicatif majeur de l'occurrence des conditions chroniques et, d'autre part, son inclusion parmi les variables prédictives pour la modélisation.

FIGURE 7.3 – Nombre moyen de conditions chroniques annuel en fonction de l'âge des assurés

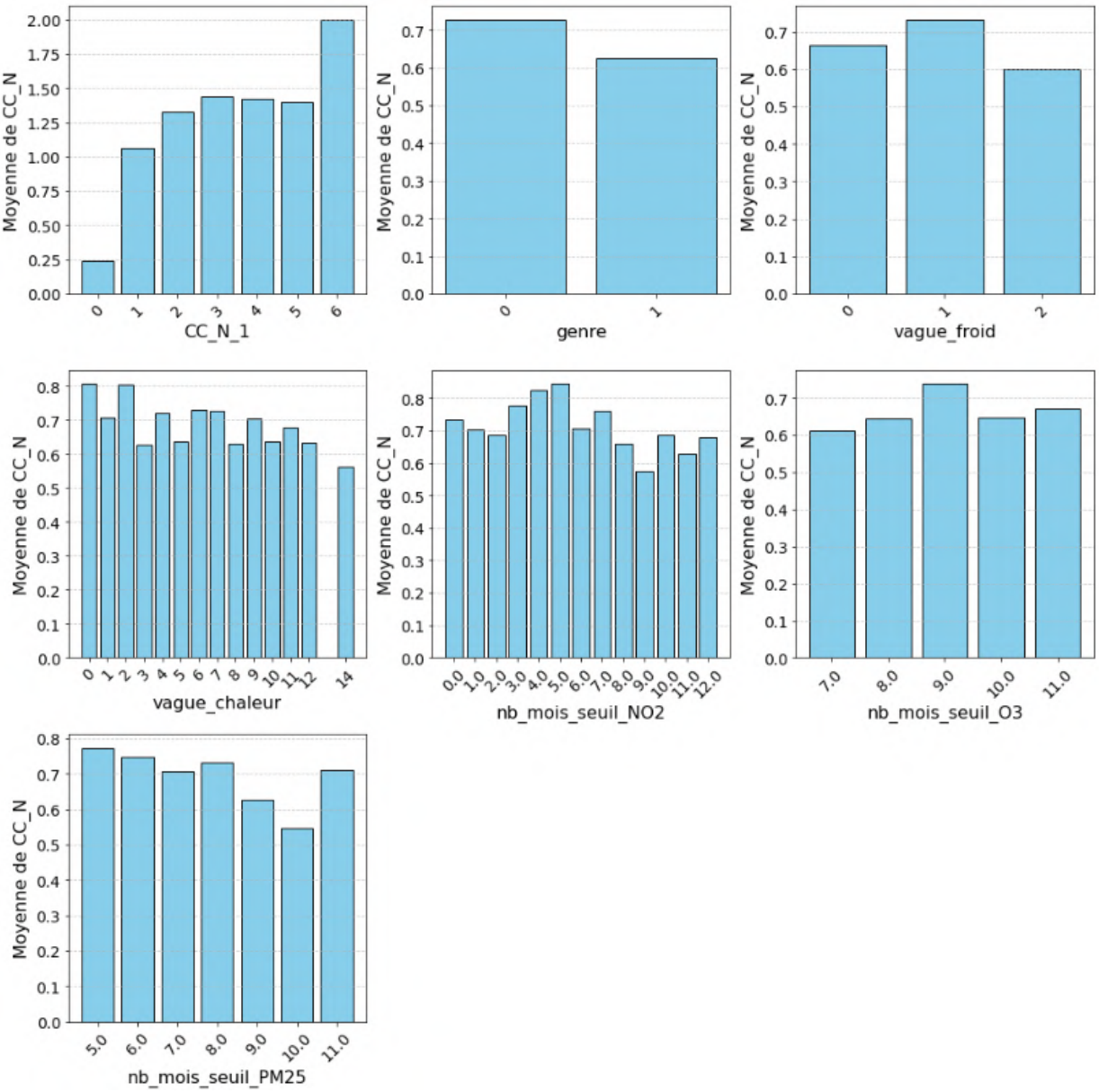


Note de lecture : un assuré de 48 ans a en moyenne 1 condition chronique renseignée par an dans la base « souscription ».

La figure 7.4 présente la distribution du nombre moyen annuel de conditions chroniques en fonction de différentes variables catégorielles, parmi lesquelles la variable cible (`CC_N_1`) variant de 0 à 6, le genre, certaines variables climatiques (`vague_froid` et `vague_chaleur`) et certaines variables de pollution atmosphérique (`nb_mois_seuil_NO2`, `nb_mois_seuil_O3` et `nb_mois_seuil_PM25`).

On observe d'abord que plus le nombre de conditions chroniques à l'année N-1 (`CC_N_1`) est élevé, plus le nombre de conditions chroniques l'année suivante est grand en moyenne. Concernant le genre, la moyenne du nombre de conditions chroniques semble légèrement plus élevée chez les femmes que chez les hommes.

FIGURE 7.4 – Analyse du nombre moyen de conditions chroniques annuel selon les variables catégorielles de l'année précédente



Note de lecture : les assurés n'ayant aucune condition chronique déclarée dans l'année ont en moyenne 0,25 condition chronique l'année suivante.

Pour les variables environnementales, les résultats apparaissent plus nuancés. Les graphiques relatifs à l'occurrence d'épisodes de vagues de froid ou de chaleur indiquent que les moyennes varient peu selon leur fréquence. Le nombre moyen de conditions chroniques a même tendance à diminuer en moyenne avec le nombre de vagues de chaleur, ce qui est contre-intuitif. Ces conclusions suggèrent un impact limité sur la variable cible à ce

stade de l'analyse. De même, les variables représentant le nombre de mois où les seuils réglementaires de pollution sont dépassés pour le NO_2, les $PM_{2.5}$ et l'O_3 montrent des différences de moyennes relativement faibles entre les catégories. Cependant, pour l'ozone, le nombre moyen de conditions chroniques tend à augmenter légèrement avec le nombre de mois concernés. Pour le NO_2 et les $PM_{2.5}$, l'analyse est plus délicate : si le nombre de conditions chroniques l'année N tend à diminuer légèrement avec le nombre de mois où le seuil réglementaire de $PM_{2.5}$ est dépassé l'année précédente, aucune tendance nette ne se dégage pour les seuils de NO_2.

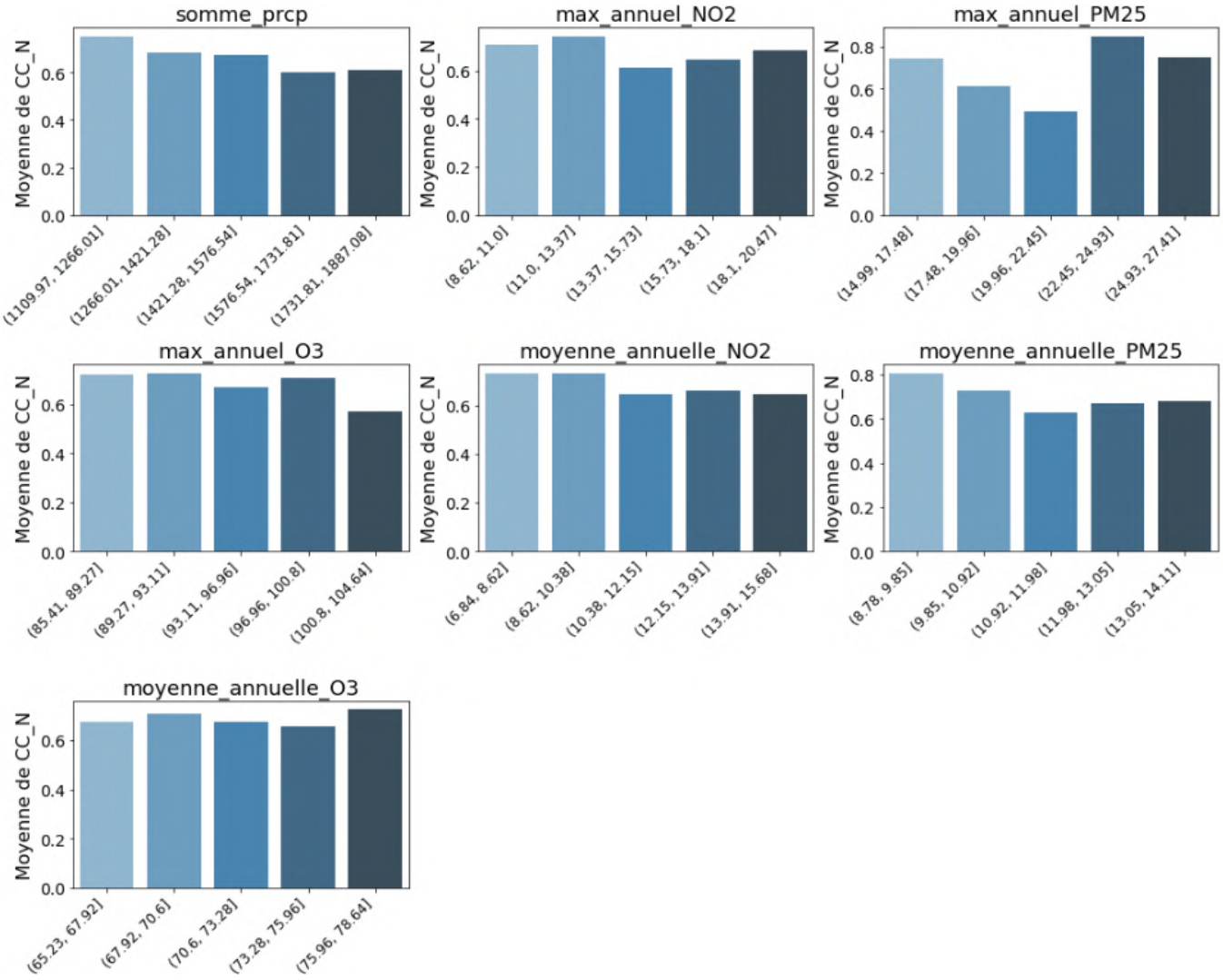
De manière générale, ces statistiques descriptives suggèrent que, parmi les variables catégorielles étudiées, peu présentent une association forte et évidente avec la variable cible, à l'exception de la variable `CC_N_1`, et plus faiblement des variables `genre`, `vague_chaleur`, `nb_mois_seuil_PM25` et `nb_mois_seuil_O3`. Les autres variables environnementales ont un effet modéré sur le nombre annuel de conditions chroniques, dans le cas de cette approche univariée. Il est ainsi nécessaire de recourir à des modèles statistiques ou de *machine learning* multivariés afin de mieux capter les effets de chaque variable sur les prédictions réalisées.

La figure 7.5 représente la distribution de la variable cible, à savoir le nombre moyen de conditions chroniques annuelles, en fonction des prédicteurs continus. Il s'agit de six variables environnementales mesurées l'année précédente : la somme des précipitations (`somme_prcp`) ainsi que les maxima mensuels et les moyennes mensuelles des trois polluants atmosphériques étudiés (NO_2, $PM_{2.5}$ et O_3). Le graphique relatif à la somme des précipitations montre une variation du nombre moyen de conditions chroniques en fonction des différentes plages de précipitations annuelles. On observe que pour les plages de précipitations les plus faibles, le nombre moyen de conditions chroniques tend à être légèrement inférieur à celui observé dans les plages les plus élevées. Le nombre annuel de conditions chroniques semble être inversement proportionnel à la somme des précipitations de l'année précédente. A ce stade, il est impossible de tirer des conclusions plus précises sur le lien entre ces deux variables.

Les trois polluants atmosphériques étudiés (NO_2, $PM_{2.5}$ et O_3) sont analysés à travers leurs valeurs maximales mensuelles et leurs moyennes mensuelles. Le graphique des maxima de NO_2 révèle une augmentation progressive du nombre moyen de conditions chroniques avec la hausse des concentrations de ce polluant. Cette corrélation positive est particulièrement marquée dans les catégories supérieures, suggérant que l'exposition à des niveaux élevés de NO_2 pourrait être un facteur aggravant pour l'apparition ou l'aggravation de maladies chroniques. Cependant, le graphique des concentrations mensuelles de NO_2 ne va pas dans ce sens : il semblerait que pour les valeurs les plus élevées, le nombre annuel de conditions chroniques soit plus faible que pour les concentrations plus basses. Concernant l'exposition aux $PM_{2.5}$, les individus sujets à des maxima plus élevés ($> 22{,}45$ µg/m^3) présentent en moyenne un nombre supérieur de conditions chroniques. Cette observation est cohérente avec la littérature scientifique, qui associe souvent la pollution particulaire à une augmentation des risques de maladies respiratoires et cardiovasculaires. Pour l'ozone, l'analyse est moins évidente. Néanmoins, une légère hausse du nombre moyen de conditions chroniques est perceptible pour les plages élevées des concentrations mensuelles de l'année précédente. Pour les maxima mensuels, la tendance est plutôt à la baisse, contrairement aux conclusions de la revue de littérature réalisée

dans le chapitre 3. Il est possible que l'effet de l'ozone soit modulé par d'autres variables environnementales ou socio-économiques non prises en compte dans cet examen univarié. Pour résumer, les variables environnementales continues ne semblent pas exercer une influence notable sur le nombre moyen de conditions chroniques observé l'année suivante. Ces statistiques descriptives ne permettent pas de mettre en exergue des facteurs proportionnels évidents. En effet, elles sont basées sur des moyennes par plage de valeurs qui ne permettent pas d'établir de lien de causalité direct. Il sera donc nécessaire d'analyser les sorties des modèles statistiques pour déterminer si ces variables ont réellement un impact sur la variable cible.

FIGURE 7.5 – Analyse du nombre moyen de conditions chroniques annuel selon les variables continues de l'année précédente



Note de lecture : les assurés résidant dans un ZIP3 où les précipitations de l'année N-1 s'élevaient entre 1266,01 et 1421,28 mm ont eu en moyenne 0,7 condition chronique l'année suivante.

Une autre approche d'analyse possible est une représentation géographique. La figure 7.6 illustre la moyenne de la variable cible `CC_N` pour chaque ZIP3. En se référant aux clusters mis en évidence dans le partitionnement effectué dans la section 6.6, on observe que les ZIP3 affichant les moyennes les plus élevées de la variable cible sont situés

dans le cluster 3. A contrario, les ZIP 3 avec une moyenne plus faible sont situés dans le cluster 1. Si l'agrégation est effectuée à l'échelle des clusters, on obtient une moyenne de

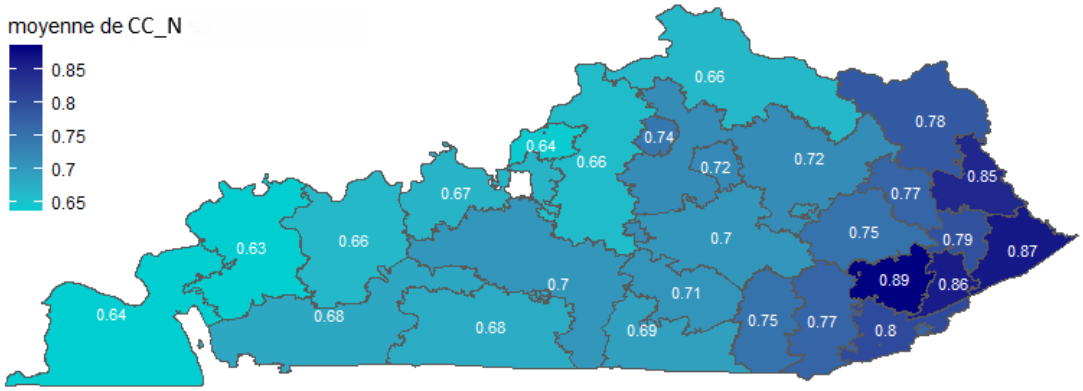
— 0,65 pour le cluster 1 ;

— 0,69 pour le cluster 2 ; et

— 0,79 pour le cluster 3.

Pour rappel, le cluster 1 est caractérisé par des températures moyennes plus élevées, des précipitations journalières plus faibles ainsi que des concentrations de NO_2 et de $PM_{2.5}$ supérieures à celles des autres clusters. Ainsi, les résultats de cette analyse géographique sont surprenants, mais sont à analyser avec prudence, car il s'agit d'une analyse univariée.

FIGURE 7.6 – Moyenne du nombre de conditions chroniques annuel (`CC_N`) par ZIP3 dans le Kentucky



Note de lecture : dans le ZIP3 413, la moyenne du nombre de conditions chroniques annuel est de 0,75.

7.2 Modélisation du score de santé par GLM

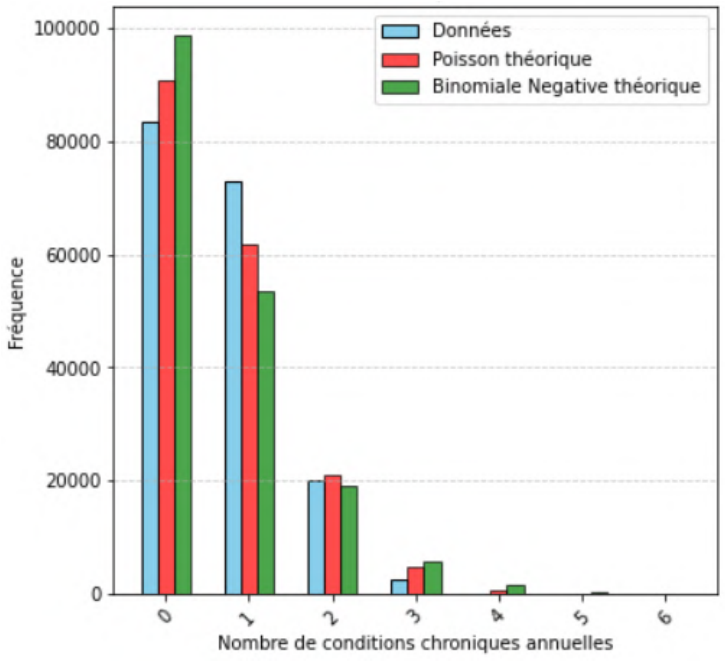
L'objectif de cette section est de construire un premier score de santé individuel à partir des données disponibles, en utilisant un modèle linéaire généralisé (GLM), tel que présenté dans la section 4.2. La mise en place d'un tel modèle nécessite le choix d'une loi adéquate pour modéliser la variable cible ainsi que d'une sélection minutieuse de variables afin de garantir la pertinence du modèle et d'éviter les problèmes de multicolinéarité. Les performances du modèle seront évaluées dans le but de les comparer ultérieurement aux différentes approches de *scoring* exploitées.

7.2.1 Justification de la loi de Poisson

Dans le cadre de la modélisation du score de santé individuel à l'aide d'un GLM, le choix de la loi suivie par la variable à modéliser est une étape cruciale. Dans cette étude, le score de santé correspond au nombre annuel de conditions chroniques par assuré prédit : il s'agit d'une variable discrète et positive. Il est donc essentiel de sélectionner une loi adaptée à cette variable cible. Deux lois sont possibles pour la modélisation d'une telle variable : la loi de Poisson et la loi Binomiale Négative. La figure 7.7 présente une comparaison entre la distribution observée de la variable cible et celles générées par les

deux lois candidates. La loi de Poisson a été générée avec un paramètre égal à la moyenne de la variable cible `CC_N`. La loi Binomiale Négative a la même moyenne et la même variance que la distribution cible. Après réalisation des tests du χ^2, aucune des deux lois n'ajuste parfaitement les données. L'analyse graphique permet toutefois de trancher en faveur de la loi de Poisson au vu de la similarité entre sa distribution et celle de la variable cible.

FIGURE 7.7 – Comparaison entre la distribution de la variable cible et celles de la loi de Poisson et de la loi Binomiale Négative



Note de lecture : générer une loi de Poisson de paramètre égal à la moyenne de la variable cible permet d'obtenir une fréquence de la valeur 2 très proche de celle observée dans la distribution réelle de la variable cible.

Pour prédire le nombre annuel de conditions chroniques par individu, un GLM Poisson est alors mis en œuvre, avec la fonction de lien canonique $g = ln$. Ainsi, en reprenant les notations de la section 4.2, et en supposant que le nombre de conditions chroniques annuel par individu Y suive une loi de Poisson, le GLM Poisson est défini par l'équation :

$$ln(\mathbb{E}[Y|\mathbf{X}]) = \eta(\mathbf{X}) := \beta_0 + \mathbf{X}'\boldsymbol{\beta}.$$

D'où :

$$\mathbb{E}[Y|\mathbf{X}] = e^{\beta_0 + \mathbf{X}'\boldsymbol{\beta}}.$$

7.2.2 Sélection des variables

Comme mentionné précédemment, la sélection rigoureuse des variables est importante pour évaluer la multicolinéarité dans le modèle GLM Poisson implémenté : elle assure la robustesse et l'interprétabilité du modèle linéaire. Pour ce faire, une première étape consiste à analyser les corrélations entre la variable cible et les variables prédictives ainsi que les corrélations entre les variables prédictives elles-mêmes. La matrice de corrélation

relative à l'ensemble des données disponibles est donnée en figure 7.8.

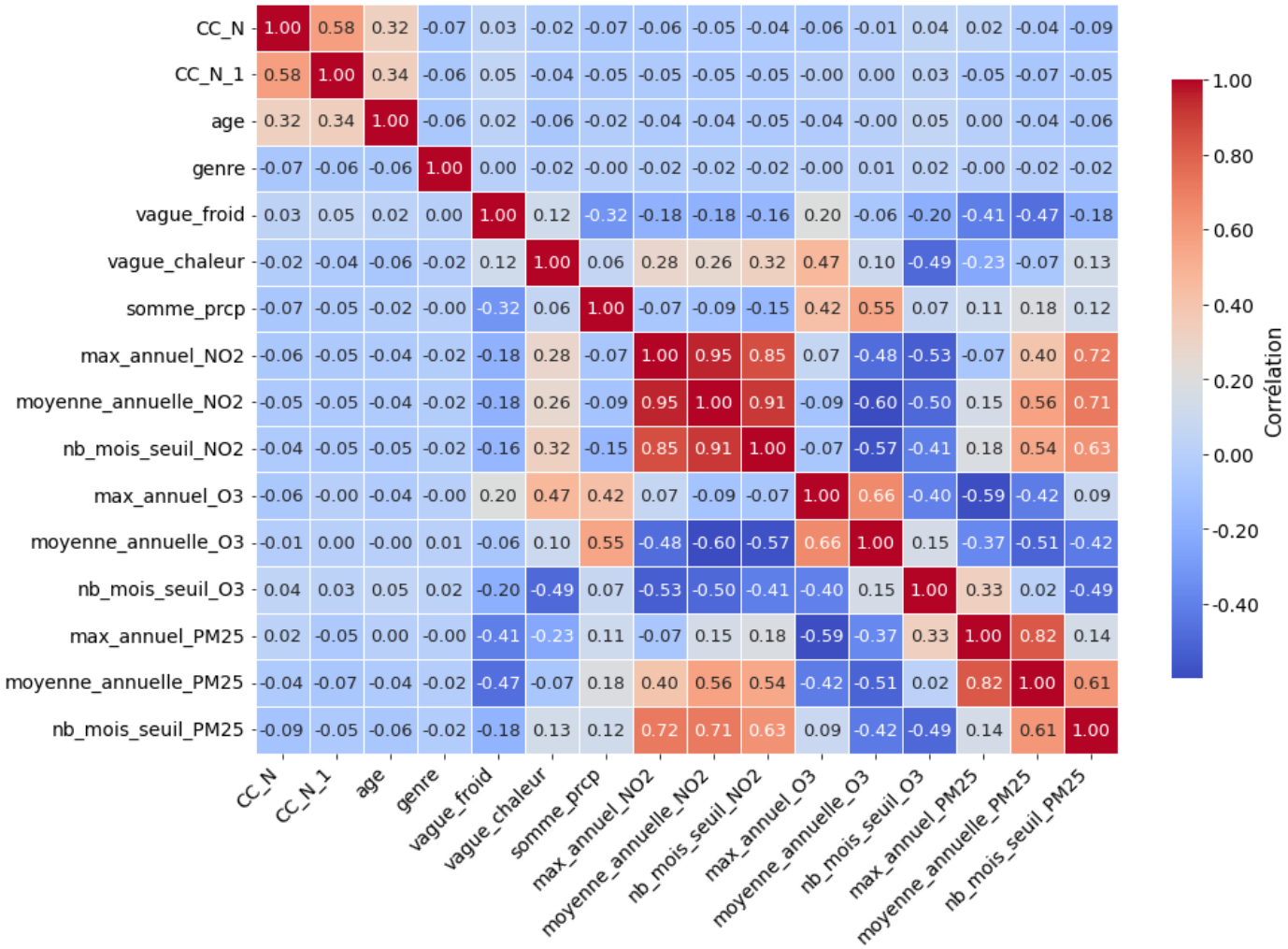
L'analyse de cette matrice révèle que la variable cible est uniquement corrélée au nombre de conditions chroniques de l'année précédente, sans que cette corrélation soit suffisante pour nuire aux performances du modèle. On considère dans la littérature qu'une corrélation absolue supérieure à 0,8 constitue un seuil au-delà duquel des problèmes de multicolinéarité peuvent survenir dans un modèle linéaire. Dans la matrice ci-dessous, on observe que :

— `max_annuel_NO2`, `moyenne_annuelle_NO2` et `nb_mois_seuil_NO2` sont fortement corrélées positivement ;

— `max_annuel_PM25` et `moyenne_annuelle_PM25` sont fortement corrélées positivement ; et

— aucune variable n'est fortement corrélée à la variable cible.

FIGURE 7.8 – Matrice de corrélation des variables prédictives



Note de lecture : la corrélation entre l'âge et le genre est égale à -0,06.

Afin d'approfondir l'analyse de la multicolinéarité et d'identifier les variables à conserver dans le modèle final, deux outils complémentaires sont utilisés : le facteur d'inflation de la variance (VIF) et la régression Lasso (voir chapitre 4). Pour rappel, le VIF permet

d'évaluer le niveau de redondance d'une variable explicative au regard des autres variables du modèle : nous considérerons dans cette étude qu'une valeur de VIF supérieure à 8 indique une multicolinéarité importante. Parallèlement, la régression Lasso, qui applique une pénalisation sur la somme des valeurs absolues des coefficients, favorise la sélection automatique des variables les plus pertinentes en contraignant les coefficients des variables moins informatives à s'annuler. Ces deux méthodes ont été testées sur l'ensemble des variables prédictives et les résultats sont donnés dans le tableau 7.2.

TABLE 7.2 – Facteurs d'inflation de la variance (VIF) et coefficients issus de la régularisation Lasso des variables explicatives

Variable	**VIF**	**Coefficient Lasso**
moyenne_annuelle_NO2	44,2	0,000
max_annuel_NO2	36,2	0,000
moyenne_annuelle_PM25	22,0	0,000
max_annuel_PM25	12,9	0,036
moyenne_annuelle_O3	8,3	0,000
nb_mois_seuil_NO2	7,4	0,000
max_annuel_O3	6,8	0,000
nb_mois_seuil_PM25	6,8	-0,037
somme_prcp	3,0	-0,018
nb_mois_seuil_O3	2,8	0,000
vague_froid	2,4	0,000
vague_chaleur	1,9	0,011
CC_N_1	1,1	0,384
age	1,1	0,096
genre	1,0	-0,014

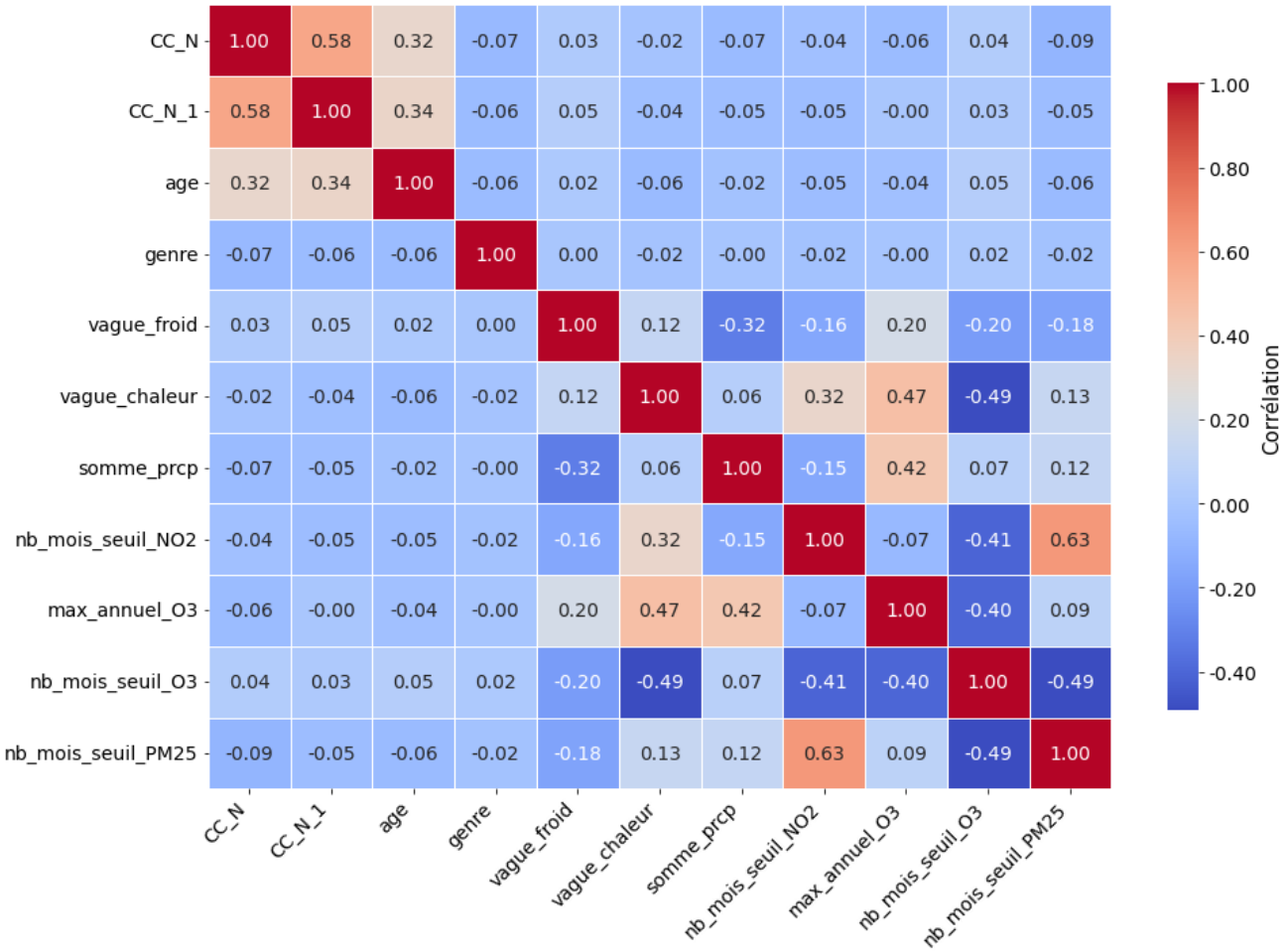
L'association de ces trois méthodes (matrice de corrélation, VIF, régularisation Lasso) permet donc de guider objectivement le choix des variables à exclure, en privilégiant celles qui contribuent le plus à la qualité prédictive du modèle tout en limitant la redondance d'information. Le VIF suggère la suppression des variables ayant un facteur supérieur à 8 : `max_annuel_NO2`, `moyenne_annuelle_NO2`, `max_annuel_PM25` `moyenne_annuelle_PM25` et `moyenne_annuelle_O3`. La régularisation Lasso, quant à elle, suggère en plus d'évincer `max_annuel_O3`, `nb_mois_NO2`, `nb_mois_O3` et `vague_froid`. Ces cinq variables ayant un VIF inférieur à 8, nous avons décidé de les intégrer dans le modèle, sous réserve d'une matrice de corrélation satisfaisante et d'une significativité statistique des coefficients associés.

7.2.3 Performance et résultats de la modélisation

Un GLM Poisson est alors implémenté sur l'ensemble des 10 variables sélectionnées dans la section précédente pour prédire le nombre annuel de conditions chroniques pour chaque assuré de la base afin de construire un score de santé individuel. La matrice de corrélation donnée en figure 7.9 confirme l'absence de colinéarité entre les variables prédictives choisies. Les résultats de ce GLM sont présentés dans le tableau 7.3.

La normalisation des variables n'est pas indispensable pour l'utilisation des modèles linéaires. Cependant, si les variables explicatives sont sur des échelles très différentes, cela peut d'une part aider à la convergence de l'algorithme d'optimisation et, d'autre part, permettre de comparer l'importance relative des variables à travers leurs coefficients. Toutefois, les résultats du modèle implémenté avec les données brutes sont essentiels pour interpréter numériquement l'impact d'une augmentation d'une variable prédictive sur le score créé. Ainsi, le tableau 7.3 présente à la fois les coefficients calculés à partir des données brutes et, entre parenthèses, ceux obtenus après normalisation (toutes les p-valeurs associées étant inférieures à 0,05).

FIGURE 7.9 – Matrice de corrélation pour le GLM à 10 variables prédictives



Note de lecture : la corrélation entre l'âge et le genre est égale à -0,06.

L'analyse des résultats du GLM visant à prédire le nombre annuel de conditions chroniques met en évidence plusieurs déterminants significatifs (p-valeur inférieure à 0,05), à la fois individuels et environnementaux. Le facteur prédictif le plus important est le nombre de conditions chroniques de l'année précédente (`CC_N_1`), ce qui reflète une forte inertie des trajectoires de santé : une augmentation d'une unité de `CC_N_1` multiplie le nombre de conditions chroniques l'année suivante par 1,82 ($e^{0,6002}$). L'âge joue également un rôle notable, confirmant l'effet cumulatif du vieillissement sur la santé chronique, mentionné dans l'analyse des statistiques descriptives. Entre un individu de 40 ans et un individu de 50 ans, le nombre annuel de conditions chroniques est multiplié en moyenne

par 1,08. Le genre présente également un effet significatif, mais modeste : être un homme multiplie le nombre de conditions chroniques annuel par 0,94 par rapport au fait d'être une femme. Plusieurs variables environnementales apparaissent significatives, bien que leurs effets soient d'intensité variable. Les vagues de chaleur sont associées à une légère augmentation du nombre de pathologies chroniques (facteur multiplicatif de 1,03), tandis que les vagues de froid présentent un effet inverse (facteur multiplicatif de 0,973), plus difficile à interpréter. Concernant la pollution atmosphérique, le nombre de mois dépassant le seuil pour le NO_2 est positivement corrélé à la variable cible, tandis que les dépassements pour l'ozone et le $PM_{2.5}$ affichent des effets négatifs inattendus. Enfin, les précipitations annuelles ont un effet très faible, mais significatif, probablement indirect. Dans l'ensemble, le modèle révèle la complexité des relations entre santé chronique, caractéristiques individuelles et environnementales, tout en soulignant la nécessité de creuser davantage les mécanismes sous-jacents, notamment en matière de pollution.

TABLE 7.3 – Résumé du modèle GLM - 10 prédicteurs avec coefficients normalisés

Variable	**Coef.** (Norm.)	**Ecart-type**	**t-stat**	**p-valeur**
Intercept	1,3445 (-0,5490)	0,138	9,717	0,000**
CC_N_1	0,6002 (0,4419)	0,004	158,976	0,000**
age	0,0086 (0,1529)	0,000	45,576	0,000**
genre	-0,0637 (-0,0317)	0,006	-9,815	0,000**
vague_froid	-0,0274 (-0,0140)	0,008	-3,255	0,001**
vague_chaleur	0,0132 (0,0343)	0,002	8,497	0,000**
somme_prcp	-8,129e-05 (-0,0148)	2,48e-05	-3,275	0,001**
nb_mois_seuil_NO2	0,0058 (0,0190)	0,001	5,031	0,000**
max_annuel_O3	-0,0172 (-0,0637)	0,001	-14,325	0,000**
nb_mois_seuil_O3	-0,0321 (-0,0214)	0,007	-4,685	0,000**
nb_mois_seuil_PM25	-0,0705 (-0,0912)	0,004	-19,775	0,000**

** p-valeur inférieure à 0,05

Les métriques de performance du GLM implémenté sont présentées dans le tableau 7.4. Elles sont comparées à celles du même modèle, mais dans lequel la variable relative à l'état de santé de l'assuré, `CC_N_1`, a été retirée. Ces résultats mettent en évidence l'importance du nombre antérieur de conditions chroniques dans la capacité prédictive du modèle. En effet, lorsque `CC_N_1` est incluse, tous les indicateurs d'erreur (MAE, MSE et RMSE) sont systématiquement plus faibles : la MAE passe de 0,578 à 0,511, la MSE de 0,483 à 0,424, et la RMSE de 0,695 à 0,651. Cette amélioration traduit une meilleure précision des prédictions du nombre annuel de pathologies chroniques, confirmant l'apport majeur de l'état de santé antérieur dans la modélisation du phénomène. Sans la prise en compte de cette variable qui constitue l'unique information sur la santé antérieure de l'assuré, le modèle peine à capter les dynamiques à partir des autres variables explicatives, qu'elles soient individuelles ou environnementales.

TABLE 7.4 – Comparaison, sur la base de test, du GLM avec ou sans la variable `CC_N_1`

Modèle GLM	**MAE**	**MSE**	**RMSE**
avec `CC_N_1`	0,511	0,424	0,651
sans `CC_N_1`	0,578	0,483	0,695

Les conclusions sont similaires au regard des indicateurs d'ajustement et de qualité présentés dans le tableau 7.5. La qualité du GLM implémenté avec la variable `CC_N_1` est modérée. En effet, les coefficients d'ajustement ($R^2_{CS} = 0{,}221$ et $R^2_\alpha = 0{,}263$) indiquent que le modèle explique environ un quart de la variabilité de la variable cible : il laisse donc une part importante inexpliquée (environ 75 %). La statistique de Pearson (90 631) et la déviance (100 196) sont relativement élevées, ce qui suggère que le modèle ne capture pas parfaitement la structure des données et qu'il subsiste des écarts notables entre les valeurs observées et prédites. Cependant, lorsque la variable relative aux conditions chroniques de l'année précédente est retirée, les variables prédictives restantes peinent à expliquer la variabilité du score de santé créé. En effet, le R^2_{CS} et le R^2_α sont très faibles (resp. 0,086 et 0,095) et les indicateurs de qualité (statistique de Pearson, déviance, AIC et BIC) sont tous plus élevés que le GLM avec `CC_N_1` indiquant un modèle de moindre qualité.

TABLE 7.5 – Indicateurs d'ajustement et de qualité du modèle GLM avec et sans la variable `CC_N_1`

Modèle GLM	$\mathbf{R^2_{CS}}$	$\mathbf{R^2_\alpha}$	**Stat. de Pearson**	$\mathbf{D_{mod}}$	**AIC**	**BIC**
avec `CC_N_1`	0,221	0,263	90 631	100 196	265 092	-1 601 854
sans `CC_N_1`	0,086	0,095	106 770	123 003	287 897	-1 578 059

Les résultats ci-dessus témoignent de la nécessité de prendre en compte l'historique médical pour la création de scores de santé individuels robustes et précis, en particulier dans le contexte de la santé chronique. L'intégration de la variable `CC_N_1` permet d'améliorer significativement la qualité, l'ajustement et la précision des scores prédictifs. Bien que l'utilisation des données de santé antérieures soit interdite pour la tarification individuelle des assurances santé aux Etats-Unis (voir section 2.1.3), ces variables seront exploitées dans ce mémoire pour maximiser la performance prédictive des modèles implémentés.

7.2.4 Calibrage du GLM

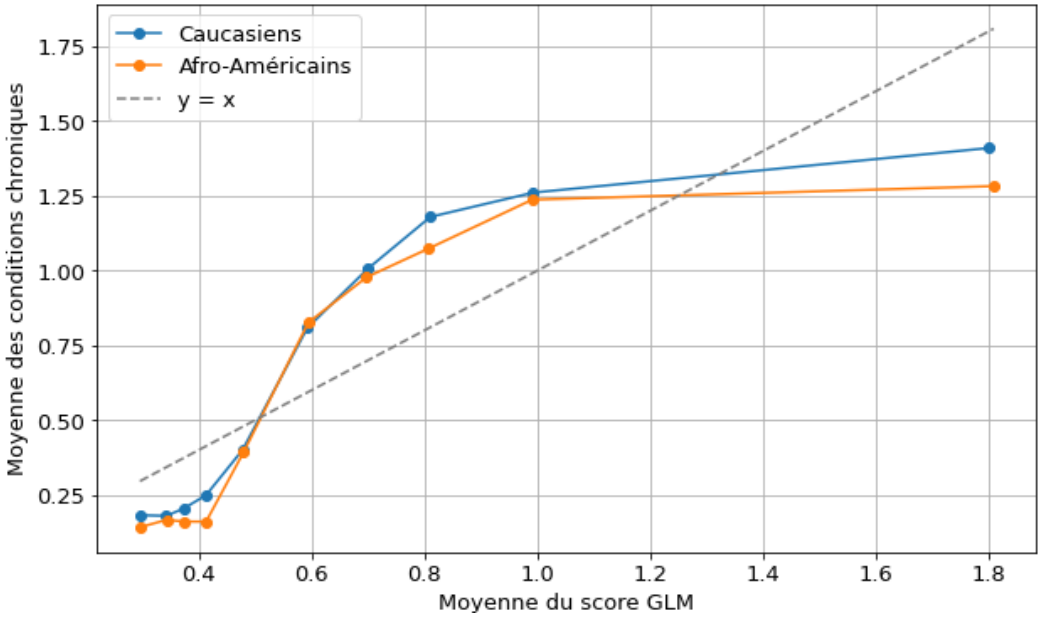
Une attention particulière est accordée dans ce mémoire à l'évaluation de l'équité des modèles, notamment vis-à-vis des différentes origines ethniques. L'objectif est d'identifier d'éventuels biais et de garantir une utilisation responsable des modèles de scores de santé en assurance. La section 5.5 montre la présence de biais dans les modèles de score existants et présente des outils permettant de mesurer à la fois le calibrage et l'équité des modèles développés : les courbes de calibrage et le calcul de l'ECE.

La figure 7.10 illustre la relation entre le score de santé annuel obtenu avec le GLM et la moyenne de maladies chroniques (`CC_N`) pour deux groupes ethniques : les Caucasiens (courbe bleue) et les Afro-Américains (courbe orange). On observe que, pour une même valeur de score prédit, le nombre de conditions chroniques diffère selon le groupe ethnique. En effet, la courbe de calibrage pour les Afro-Américains se situe au-dessous de celle des Caucasiens. Cela signifie qu'à score égal, les assurés afro-américains présentent en réalité un nombre moins élevé de maladies chroniques que les assurés caucasiens. Ce constat met en évidence un biais de calibrage du modèle, en contradiction avec les principes d'équité algorithmique évoqués dans la section 5.5, notamment la *calibration parity* et le calibrage par groupe. En effet, ici, l'écart constaté indique que le modèle sous-estime la gravité de

la morbidité chronique chez les Caucasiens par rapport aux Afro-Américains. Obermeyer avait observé le contraire dans son étude de 2019 [79].

Par ailleurs, les biais sont plus marqués pour les faibles et les hauts scores : cela peut avoir des conséquences importantes, car les décisions cliniques ou assurantielles basées sur ce score pourraient conduire à une inégalité dans la prise en charge des cas les plus graves, avec un risque de sous-traitement ou de sur-traitement selon l'origine ethnique. De plus, la distance entre les deux courbes de calibrage et la droite d'identité $y = x$ témoigne d'un calibrage peu satisfaisant du modèle. En effet, l'écart moyen entre la courbe de calibrage des assurés caucasiens et $y = x$ est égal à 0,2234 tandis que pour les Afro-Américains, il s'élève à 0,2431. Ces résultats révèlent un meilleur calibrage pour les assurés caucasiens que pour les assurés afro-américains, même si cette différence reste relativement faible.

FIGURE 7.10 – Calibrage par groupe ethnique du score annuel GLM



Note de lecture : pour le 5ème décile de score GLM (moyenne de score = 0,39), le nombre moyen de conditions chroniques est de 0,41 chez les Caucasiens et de 0,38 chez les Afro-Américains.

L'ECE (voir section 5.5.3) est calculée pour les assurés caucasiens, pour les assurés afro-américains et pour tous les assurés. Les résultats sont présentés dans le tableau 7.6. Les valeurs des ECE confirment le calibrage modéré du modèle sur toute la population (ECE=0,2250), qui reste donc perfectible.

TABLE 7.6 – ECE globale et par groupe ethno-racial

ECE globale	0,2250
ECE Caucasiens	0,2253
ECE Afro-Américains	0,2217

En somme, cette analyse de calibrage corrobore les conclusions de la partie précédente : bien que le GLM offre une interprétabilité des effets des prédicteurs sur le score créé, ses

performances en termes de prédiction et d'équité restent limitées, ce qui en fait un modèle de score peu satisfaisant dans ce contexte.

7.2.5 Conclusions et limites

Le modèle GLM développé présente une capacité explicative satisfaisante, comme en témoignent les indicateurs de performance. Néanmoins, des marges d'amélioration subsistent : l'introduction d'effets aléatoires ou d'interactions (via un GLMM par exemple) permettrait de mieux capturer l'hétérogénéité entre individus, tandis que des modèles d'apprentissage automatique plus avancés (tels que XGBoost) pourraient optimiser la représentation de relations complexes ou non-linéaires entre les variables.

Cependant, certains effets des variables environnementales sur le score de santé créé sont contre-intuitifs au regard de la littérature. Ils suggèrent d'éventuelles interactions non-modélisées ou une colinéarité avec d'autres variables polluantes. Aussi, l'agrégation à la maille temporelle annuelle écrase les effets potentiels des variables climatiques et de pollution sur le score créé. Par exemple, une vague de chaleur survenue en 2016 n'aura que peu d'impact en 2017. Il en est de même pour les vagues de froid ou les pics de pollution.

Le modèle GLM implémenté présente également des biais, couplés à un mauvais calibrage. Les modèles utilisés dans les sections suivantes devront toujours être mis en balance avec les enjeux éthiques et réglementaires liés à l'utilisation des données médicales, afin de garantir la pertinence et l'équité des scores de santé produits.

7.3 Score basé sur le GLMM

Un GLMM a ensuite été implémenté sur les mêmes données et variables que le GLM. Pour rappel, les modèles linéaires mixtes généralisés (GLMM) représentent une extension des GLM. Ils permettent de traiter, dans cette étude, les données longitudinales, où les observations ne sont pas indépendantes puisqu'elles sont collectées par individu : cela crée une dépendance entre les observations du jeu de données. La seule différence entre le GLM et le GLMM réside dans la prise en compte des effets aléatoires au sein de chaque groupe.

7.3.1 Performance et résultats du GLMM

Les résultats du GLMM sont présentés dans le tableau 7.7. Les coefficients sont du même signe que ceux obtenus avec le GLM. Comme pour le score de santé modélisé à l'aide d'un GLM, les résultats mettent en exergue l'effet prédictif majeur du nombre de conditions chroniques observé l'année précédente (`CC_N_1`), avec un coefficient normalisé de 0,365. L'âge a un effet significatif et positif, confirmant l'impact du vieillissement sur le cumul de pathologies chroniques observé avec le GLM. Le genre reste également un prédicteur significatif. Son effet est similaire à celui observé dans le GLM : être un homme est associé à un nombre légèrement inférieur de maladies chroniques par rapport à une femme.

Les variables environnementales conservent leur caractère significatif, à l'exception de la somme annuelle des précipitations enregistrées l'année précédente (`somme_prcp`), dont l'effet devient non significatif (p-valeur = $0,463$) dans le GLMM. L'impact des précipitations sur le score de santé individuel n'est donc plus interprétable dans ce cas. Les vagues de chaleur et de froid, ainsi que le nombre annuel de dépassements de seuils de pollution (`nb_mois_seuil_NO2`, `nb_mois_seuil_O3`, `nb_mois_seuil_PM2.5`), présentent des effets directionnels similaires à ceux du GLM, bien que leurs amplitudes soient légèrement atténuées. La variance intergroupe estimée (`Group Var`) est non-nulle, confirmant la pertinence de l'intégration d'un effet aléatoire individuel.

TABLE 7.7 – Résumé du modèle GLMM - 10 prédicteurs avec coefficients normalisés

Variable	**Coef.** (Norm.)	**Ecart-type**	**t-stat**	**p-valeur**
Intercept	1,880 (0,701)	0,068	27,552	0,000**
CC_N_1	0,497 (0,365)	0,004	129,751	0,000**
genre	-0,047 (-0,047)	0,003	-14,283	0,000**
age	0,006 (0,112)	0,000	58,533	0,000**
somme_prcp	-0,000 (-0,002)	0,000	-0,733	0,463
vague_froid	-0,018 (-0,008)	0,004	-4,417	0,000**
vague_chaleur	0,009 (0,026)	0,001	12,053	0,000**
nb_mois_seuil_NO2	0,004 (0,016)	0,001	6,750	0,000**
max_annuel_O3	-0,012 (-0,047)	0,001	-20,551	0,000**
nb_mois_seuil_O3	-0,025 (-0,017)	0,003	-7,233	0,000**
nb_mois_seuil_PM25	-0,045 (-0,062)	0,002	-25,880	0,000**
Group Var	0,016 (0,016)	0,003		

** p-valeur inférieure à 0,05

Les métriques de performance du GLMM (voir tableau 7.8) indiquent une amélioration par rapport au GLM. L'introduction des effets aléatoires permet de réduire la MAE de 0,511 à 0,451, la MSE de 0,424 à 0,345, et la RMSE de 0,651 à 0,587. Cette diminution des erreurs montre l'importance de la prise en compte, via le GLMM, de l'hétérogénéité non-observée. Le GLMM renforce ainsi la robustesse et la précision du score de santé développé.

TABLE 7.8 – Comparaison, sur la base de test, du GLM et du GLMM

Modèle	**MAE**	**MSE**	**RMSE**
GLM	0,511	0,424	0,651
GLMM	0,451	0,345	0,587

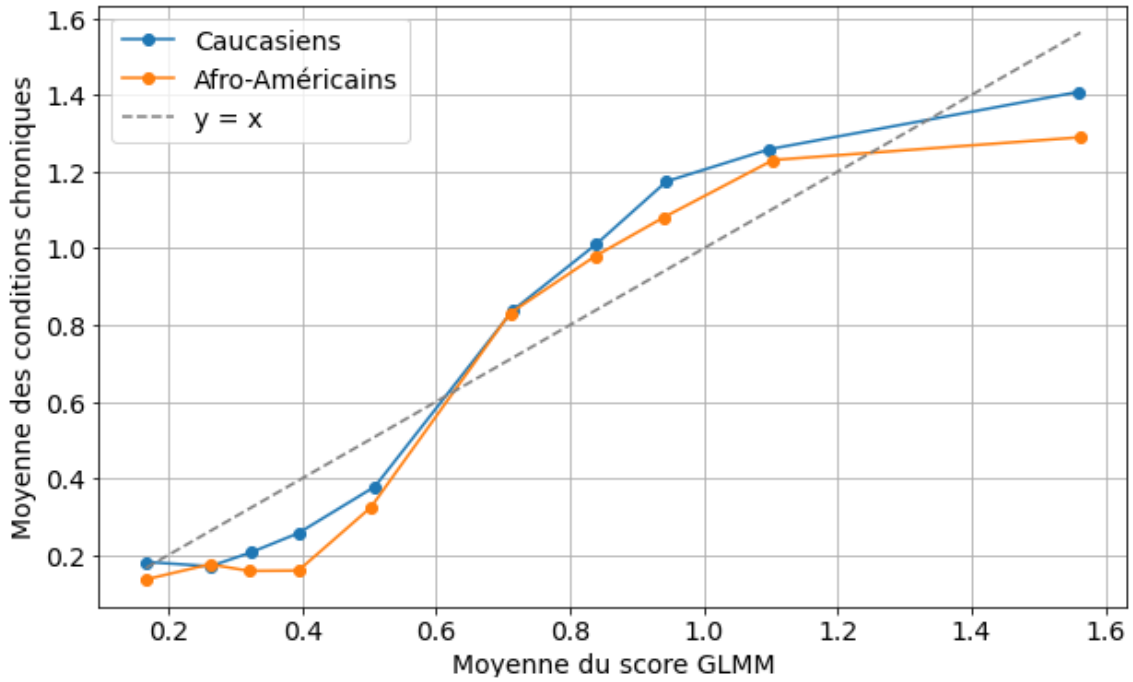
7.3.2 Calibrage du modèle

La figure 7.11 représente les courbes de calibrage du score de santé annuel obtenu avec le GLMM pour deux groupes ethniques : les Caucasiens (courbe bleue) et les Afro-américains (courbe orange). On observe que, pour une même valeur de score prédit, le nombre de conditions chroniques diffère selon le groupe ethnique. En effet, la courbe de calibrage pour les Afro-Américains se situe au-dessous de celle des Caucasiens, sauf pour les deux premiers déciles de score. Cela signifie qu'à score égal, les assurés afro-américains

présentent en réalité un nombre moins élevé de maladies chroniques que les caucasiens. Ce constat met en évidence un biais de calibrage du modèle. En effet, ici, l'écart constaté indique que le modèle sous-estime la gravité de la morbidité chronique chez les Caucasiens par rapport aux Afro-Américains. Cependant, même s'ils restent non-négligeables, les biais ethniques sont plus faibles que ceux observés avec le score GLM (voir figure 7.10).

Aussi, le GLMM est mieux calibré. En effet, l'écart moyen entre la courbe de calibrage des assurés caucasiens et $y = x$ est égal à 0,134 tandis que pour les Afro-Américains, il s'élève à 0,15. Ces résultats révèlent un meilleur calibrage pour les assurés caucasiens que pour les assurés afro-américains, même si cette différence reste relativement faible. Il correspond donc à un meilleur modèle en termes de calibrage et de biais ethnique que le GLM.

FIGURE 7.11 – Calibrage par groupe ethnique du score annuel GLMM



Note de lecture : pour le 5ème décile de score GLMM (moyenne de score = 0,51), le nombre moyen de conditions chroniques est de 0,38 chez les Caucasiens et de 0,32 chez les Afro-Américains.

En outre, il est pertinent de comparer les ECE du GLMM avec les ECE calculées pour le GLM. Cette comparaison est réalisée dans le tableau 7.9. L'ECE globale passe de 0,2250 avec le GLM à 0,1336 avec le GLMM, soit près de deux fois moins d'erreur moyenne de calibrage. Le calibrage est relativement similaire en moyenne entre les assurés caucasiens et les assurés afro-américains.

TABLE 7.9 – Comparaison des ECE globales et par groupe ethno-racial

	GLM	**GLMM**
ECE globale	0,2250	0,1336
ECE Caucasiens	0,2253	0,1339
ECE Afro-Américains	0,2217	0,1380

7.3.3 Conclusions et limites

En résumé, le GLMM s'avère supérieur au GLM pour la modélisation du nombre annuel de conditions chroniques, tant en termes d'ajustement que de capacité prédictive. Il permet de mieux modéliser la dynamique individuelle, en tenant compte de la structure longitudinale des données. Néanmoins, plusieurs limites persistent. Malgré l'amélioration du calibrage (ECE globale ramenée à 0,13), le modèle demeure imparfait : les écarts de calibrage entre groupes, bien que faibles, persistent, et certains effets non-linéaires ou interactions complexes entre variables ne peuvent être entièrement capturés par une approche linéaire, même mixte. Enfin, la performance prédictive, bien qu'améliorée, semble plafonner avec ces approches.

Dans ce contexte, il apparaît nécessaire d'explorer les modèles de *machine learning*, plus flexibles, capables de capturer des relations complexes et non-linéaires entre les variables (sans besoin de les sélectionner en amont). L'utilisation d'un modèle XGBoost pourrait permettre d'améliorer encore la précision et le calibrage du score de santé, tout en offrant des outils pour l'explication locale et globale des prédictions (via l'analyse des importances de variables ou les méthodes SHAP). Cette démarche s'inscrit dans une volonté de combiner la performance prédictive et la robustesse, tout en maintenant une attention particulière à l'équité du score entre groupes ethnico-raciaux.

7.4 Score de santé issu du modèle XGBoost

Nous avons constaté dans les sections précédentes que la prédiction du nombre annuel de conditions chroniques à l'aide des modèles linéaires (GLM et GLMM) présente certaines limites. En effet, bien qu'interprétables, ces modèles ne parviennent pas à expliquer avec précision les scores développés. Dans ce contexte, il est pertinent d'implémenter un modèle de *machine learning* qui respecte à la fois les critères de performance et les contraintes financières et opérationnelles liées au traitement d'un grand volume de données. Comme vu en détail dans la section 5.3, le modèle XGBoost est particulièrement efficace pour capturer les non-linéarités, les interactions, et gérer la présence de variables de différentes natures. Il permet également l'analyse des valeurs SHAP, offrant la possibilité d'interpréter le modèle en comprenant les facteurs sous-jacents aux prédictions.

L'objectif de cette section est ainsi d'évaluer la capacité prédictive du XGBoost pour la construction du score de santé individuel comparativement aux modèles implémentés précédemment. Le modèle est toujours appliqué à l'échantillon de 50 000 assurés du Kentucky, en utilisant l'ensemble des variables disponibles, puisque ce modèle ne nécessite pas une sélection préalable des prédicteurs. Les étapes de réglage des hyperparamètres, de validation croisée, ainsi que l'analyse des performances, du calibrage, de l'équité du modèle et de l'importance des variables dans la prédiction via les valeurs de Shapley sont détaillées dans les sous-sections suivantes.

7.4.1 Choix des hyperparamètres

Dans un premier temps, il est nécessaire d'ajuster finement ce modèle : cela est possible grâce à ses nombreux hyperparamètres. Dans ce qui suit, un point d'importance est donné à choisir les paramètres optimaux permettant d'obtenir les meilleures performances en

termes de prédiction. En s'inspirant du mémoire de Dimitri DELCAILLAU réalisé en 2019, l'utilisation de la validation croisée du type *kfold Cross-Validation* avec $k = 5$ est choisie dans ce mémoire pour ajuster les paramètres suivants [90] :

- *nrounds* : nombre maximal d'itérations du *boosting*. Une valeur élevée implique un ajustement plus fin du modèle, mais également un risque de surapprentissage.
- *max_ depth* : profondeur maximale pour chaque arbre utilisé dans le modèle. Une profondeur trop grande peut entraîner un surapprentissage.
- *eta* : le taux d'apprentissage $eta \in [0, 1]$ régule l'influence de chaque nouvel arbre dans l'ensemble. Une façon de réduire le risque de surapprentissage est de prendre des valeurs faibles pour le taux d'apprentissage.
- *gamma* : gain de perte minimal requis pour autoriser une partition supplémentaire dans un arbre. Une valeur élevée de *gamma* tend à rendre l'algorithme plus prudent en termes de complexité.
- *colsample_ bytree* : part des variables sélectionnées à chaque arbre. Ce paramètre permet d'introduire de la diversité dans les arbres générés.
- *min_ child_ weight* : poids total minimal d'observations dans un nœud. S'il est trop faible, le modèle peut surajuster. Plus cette valeur est élevée, plus le modèle est régularisé.
- *subsample* : part des données utilisée pour l'entraînement de chaque arbre. Cela agit comme une forme de régularisation et permet de réduire le temps d'exécution. Une valeur de 0,75 signifie que chaque arbre est construit avec trois-quarts des données.

En théorie, la meilleure approche consisterait à explorer toutes les combinaisons possibles de ces 7 hyperparamètres, en évaluant la performance du modèle à chaque essai via les métriques décrites au début du chapitre 5 (MSE, MAE, RMSE) avec l'utilisation de la validation croisée. Cependant, cette méthode est inenvisageable au vu de la puissance nécessaire pour exécuter l'ensemble des tests. En s'appuyant sur les travaux de Dimitri DELCAILLAU, la stratégie d'optimisation séquentielle permet alors de sélectionner les paramètres les plus influents à ajuster en priorité, puis d'ajuster les autres de façon progressive et ciblée.

TABLE 7.10 – Plages de recherche et valeurs testées pour l'ajustement des hyperparamètres du modèle XGBoost

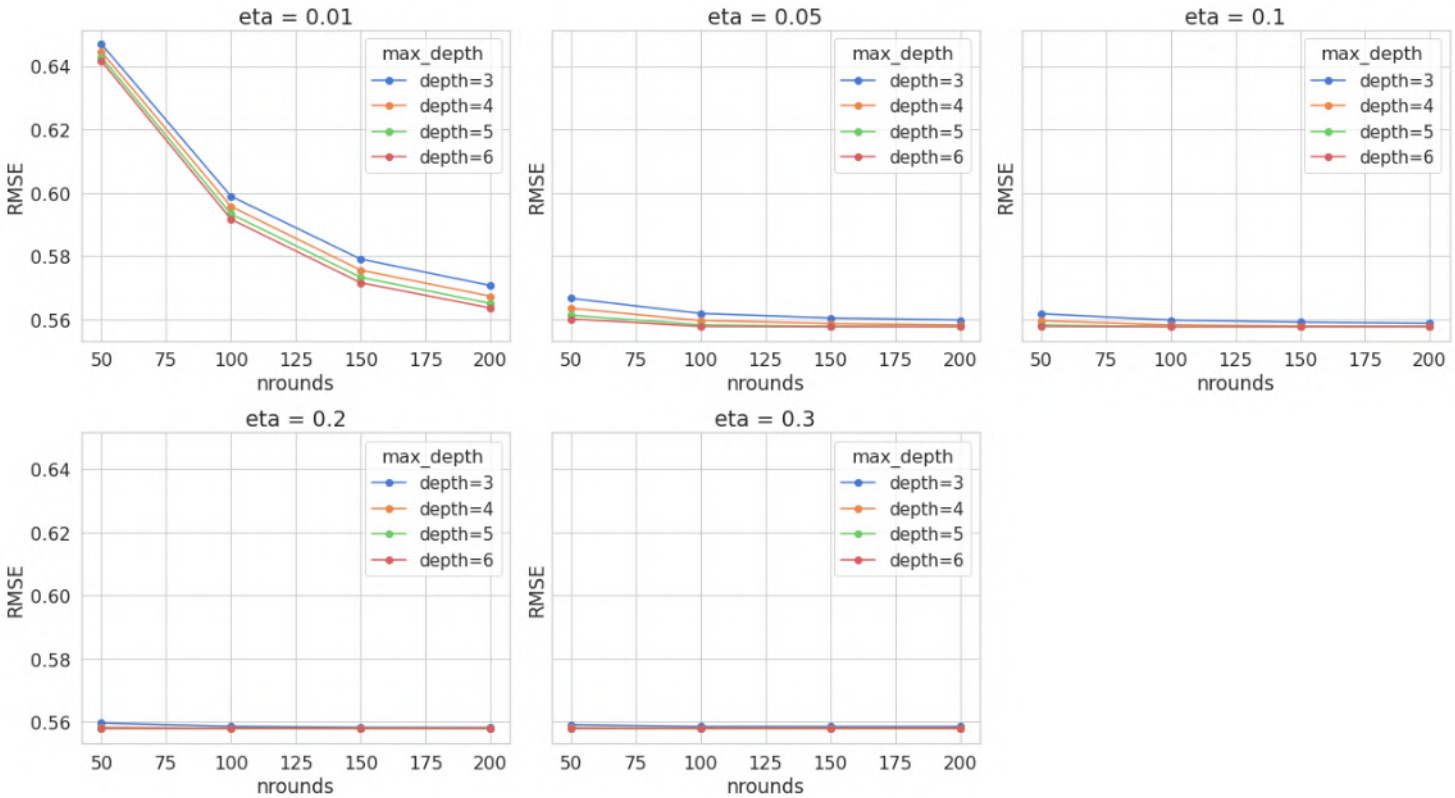
Paramètre	**Plage**	**Valeurs testées**
`eta`	[0,01 ; 0,3]	0,01 ; 0,05 ; 0,1 ; 0,2 ; 0,3
`max_depth`	[3 ; 6]	3 ; 4 ; 5 ; 6
`min_child_weight`	[1 ; 3]	1 ; 2 ; 3
`colsample_bytree`	[0,5 ; 1]	0,5 ; 0,75 ; 1
`subsample`	[0,5 ; 1]	0,5 ; 0,75 ; 1
`gamma`	[0 ; 1]	0 ; 0,2 ; 0,4 ; 0,6 ; 0,8 ; 1
`nrounds`	[50 ; 250]	50 ; 75 ; 100 ; 125 ; 150 ; 175 ; 200 ; 225 ; 250

Les travaux de Bartz-Beielstein et al. sur l'optimisation des paramètres des modèles de *machine learning* ont fait ressortir des plages usuelles pour chaque paramètre à ajuster ainsi que les interactions entre eux [91]. La première étape se concentre sur le choix du nombre d'itérations *nrounds*, du taux d'apprentissage *eta* et de la profondeur maximale

des arbres (*max_ depth*), qui sont intimement liés [91]. Un nombre maximal d'arbres de 250 est choisi d'une part pour éviter tout risque de surapprentissage, d'autre part pour optimiser le temps d'exécution. Il convient ensuite de trouver un taux d'apprentissage ainsi qu'une profondeur maximale des arbres associés à ce nombre d'arbres. En accord avec les travaux de Bartz-Beielstein et de ses co-auteurs, les valeurs de *eta* seront prises dans [0,01 ; 0,3] tandis que la plage utilisée pour *max_ depth* est [3, 6]. Les autres paramètres décrits ci-dessus sont fixés à leur valeur par défaut dans le package `xgboost` de Python. Une fois les paramètres cibles fixés, les étapes précédentes sont réitérées : sélection d'une plage pour le nouveau paramètre cible tandis que les autres sont fixés à la valeur par défaut ou à la valeur choisie en amont. Les plages choisies ainsi que les valeurs testées pour chaque paramètre sont indiquées dans le tableau 7.10.

La courbe de validation croisée de la figure 7.12 est alors obtenue, basée sur la RMSE moyenne dans les 5 *folds*, pour déterminer la valeur du taux d'apprentissage *eta*. Pour la plage d'itérations considérée, un taux d'apprentissage fixé à 0,1 permet de minimiser la RMSE moyenne. Le graphique suggère également de fixer *max_ depth* à 6.

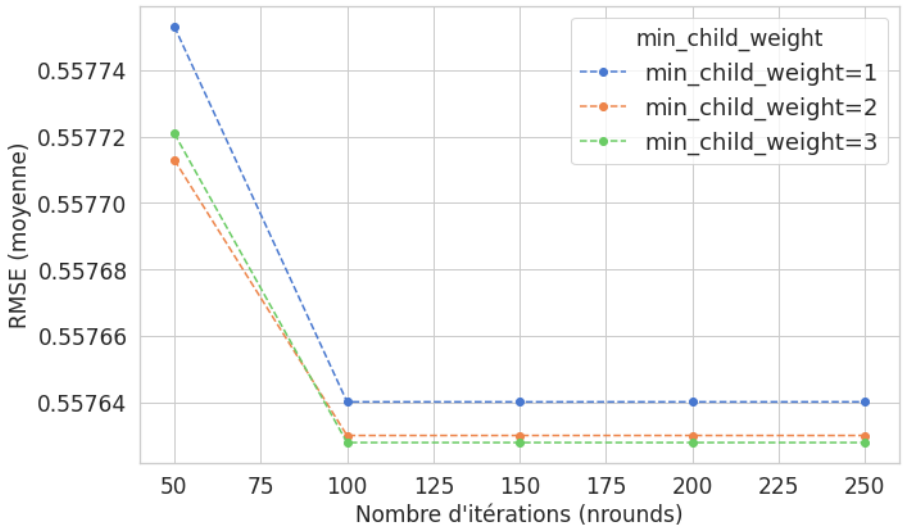
FIGURE 7.12 – Choix de *eta* à l'aide d'une validation croisée



Note de lecture : pour eta $= 0,1$*, max_ depth* $= 6$ *et nrounds* $= 100$*, la RMSE moyenne est d'environ 0,557640.*

En fixant ces deux paramètres, en faisant varier *nrounds* et *min_ child_ weight* et en fixant les paramètres restants, la figure 7.13 est obtenue. Ce graphique permet de fixer la valeur de *min_ child_ weight* à 3.

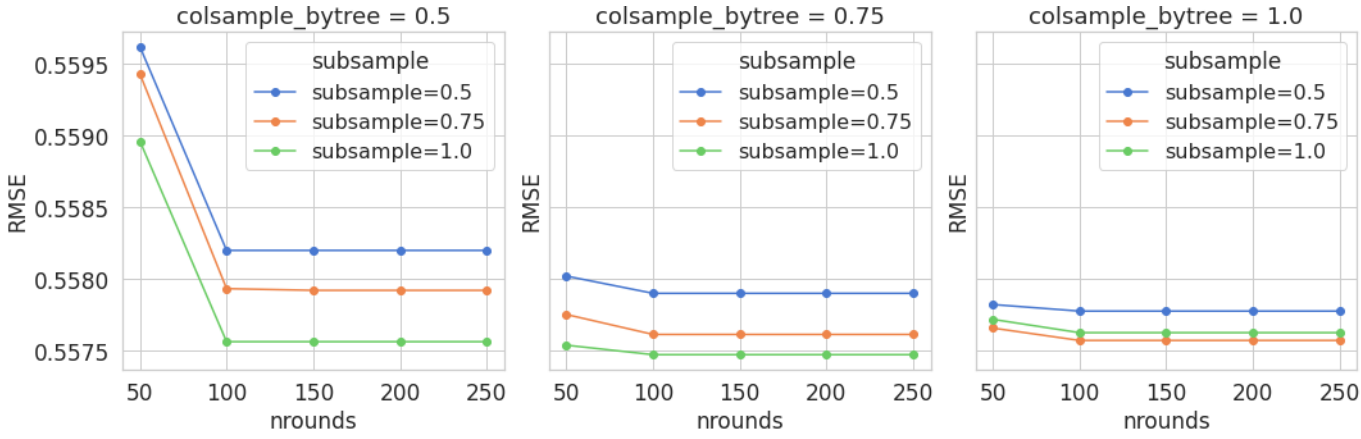
FIGURE 7.13 – Choix de *min_child_weight* et *max_depth* à l'aide d'une validation croisée



Note de lecture : pour eta $= 0,1$*, max_depth* $= 6$*, nrounds* $= 100$*, et min_child_weight* $= 3$*, la RMSE moyenne est d'environ 0,557628.*

Les paramètres *colsample_bytree* et *subsample* sont ensuite testés dans la plage $[0,5;1]$ en gardant la valeur des paramètres précédemment optimisés et en fixant les derniers à leur valeur par défaut. Les résultats sont donnés par la figure 7.14. Il convient de fixer *colsample_bytree* à 0,75 et *subsample* à 1.

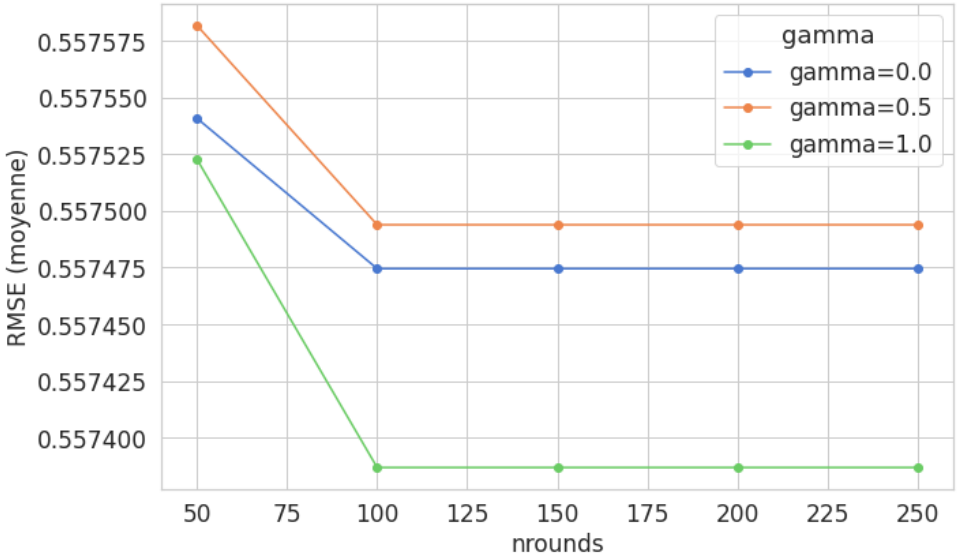
FIGURE 7.14 – Choix de *colsample_bytree* et *subsample* à l'aide d'une validation croisée



Note de lecture : pour eta $= 0,1$*, max_depth* $= 6$*, nrounds* $= 100$*, min_child_weight* $= 3$*, subsample* $= 1$ *et colsample_bytree* $= 0,75$ *la RMSE moyenne est d'environ 0,557475.*

Ensuite, la RMSE moyenne est analysée en fonction des valeurs de *gamma*. Les résultats sont donnés par la figure 7.15. Fixer la valeur de *gamma* à 1 permet le meilleur ajustement du modèle.

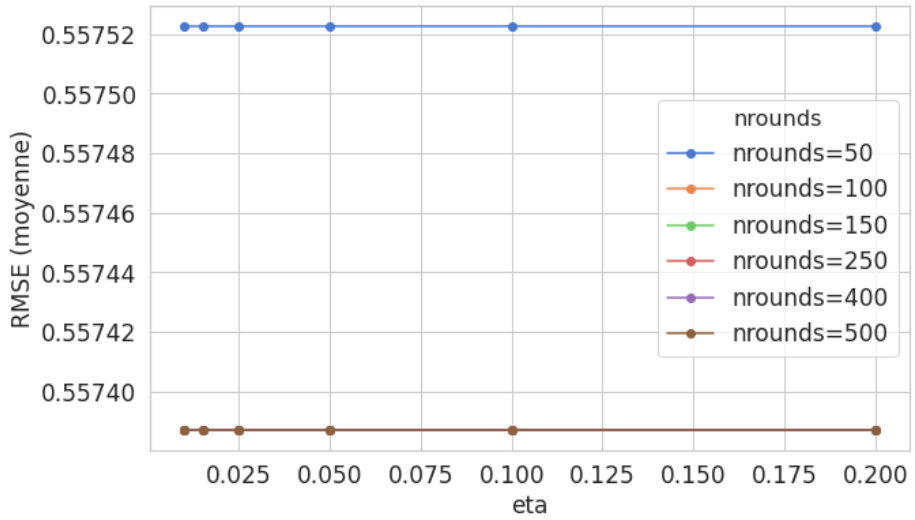
FIGURE 7.15 – Choix de *gamma* à l'aide d'une validation croisée



Note de lecture : pour eta $= 0,1$*, max_depth* $= 6$*, nrounds* $= 100$*, min_child_weight* $= 3$*, subsample* $= 1$*, gamma* $= 1$ *et colsample_bytree* $= 0,75$*, la RMSE moyenne est d'environ 0,557352.*

Enfin, la dernière étape consiste à ajuster les valeurs de *eta* et de *nrounds* en fixant les 5 autres paramètres aux valeurs précédemment choisies. La RMSE n'évolue que très peu en fonction des différentes valeurs de *eta* et de *nrounds* testées (voir figure 7.16). De plus, pour *nrounds* > 100, les points sont superposés. Ainsi, pour éviter le surapprentissage et pour être cohérent avec les résultats de la figure 7.12, la valeur du taux d'apprentissage est gardée à 0,1 et celle de *nrounds* à 100.

FIGURE 7.16 – Choix de *nrounds* et de *eta* à l'aide d'une validation croisée



Note de lecture : pour eta $= 0,1$*, max_depth* $= 6$*, nrounds* $= 100$*, min_child_weight* $= 3$*, subsample* $= 1$*, gamma* $= 1$ *et colsample_bytree* $= 0,75$*, la RMSE moyenne est d'environ 0,557387.*

Ainsi, les 7 paramètres du modèle XGBoost ont été fixés dans le but d'ajuster au mieux le modèle. Les valeurs choisies sont résumées dans le tableau 7.11 suivant.

TABLE 7.11 – Valeurs des hyperparamètres retenus pour le modèle XGBoost

Paramètre	**Valeur choisie**
`eta`	0,1
`max_depth`	6
`min_child_weight`	3
`colsample_bytree`	0,75
`subsample`	1
`gamma`	1
`nrounds`	100

7.4.2 Résultats et performance du modèle XGBoost

Après une sélection rigoureuse des hyperparamètres, le modèle XGBoost a été implémenté dans le but de prédire individuellement le nombre de conditions chroniques annuel à l'aide de l'ensemble des variables prédictives de la base de données. Les résultats de performance du modèle sur les ensembles d'entraînement et de test sont présentés dans le tableau 7.12.

TABLE 7.12 – Comparaison des performances du XGBoost sur les bases de test et d'entraînement

	MAE	**MSE**	**RMSE**
Base d'entraînement	0,414	0,304	0,552
Base de test	0,416	0,307	0,554

Pour les métriques considérées, à savoir la MAE, la MSE et la RMSE, l'écart de performance du XGBoost entre les ensembles de test et d'entraînement suggère un apprentissage satisfaisant du modèle : il n'y a ni surapprentissage, ni sous-apprentissage. Cela indique que le modèle a trouvé un bon compromis entre biais et variance : il est bien ajusté.

Le score issu du modèle XGBoost est à comparer aux deux autres scores implémentés précédemment à l'aide respectivement d'un GLM et d'un GLMM. La comparaison entre les trois modèles de score développés est réalisée dans le tableau 7.13. L'implémentation d'un XGBoost permet de réduire de 18,6 % la MAE, de 27,6 % la MSE et de 14,9 % la RMSE par rapport aux scores développés sur la base d'un GLM. Le XGBoost permet donc, en un temps raisonnable, d'améliorer la qualité et la pertinence des scores précédemment développés. Toutefois, pour s'assurer que le modèle est bien calibré et qu'il ne présente pas de biais ethniques, une analyse est menée ci-après.

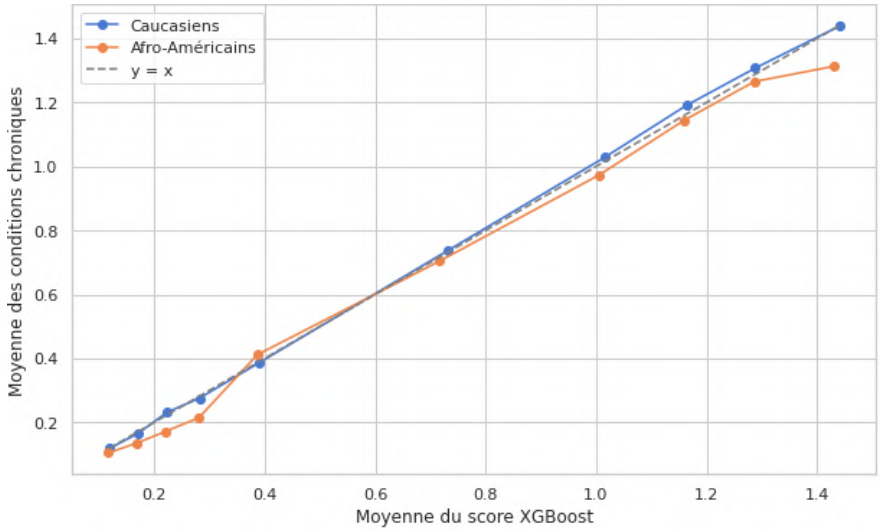
TABLE 7.13 – Comparaison, sur la base de test, des différents modèles mis en place pour modéliser le nombre annuel de conditions chroniques

Modèle	**MAE**	**MSE**	**RMSE**
GLM	0,511	0,424	0,651
GLMM	0,451	0,345	0,587
XGBoost	0,416	0,307	0,554

7.4.3 Calibrage du modèle XGBoost

La figure 7.17 illustre la relation entre le score de santé annuel obtenu avec le modèle XGBoost et la moyenne de maladies chroniques pour deux groupes ethniques : les Caucasiens (courbe bleue) et les Afro-Américains (courbe orange).

FIGURE 7.17 – Calibrage par groupe ethnique du score annuel XGBoost



Note de lecture : pour le 5ème décile de score XGBoost (moyenne de score = 0,39), le nombre moyen de conditions chroniques est de 0,41 chez les Caucasiens et de 0,38 chez les Afro-Américains.

On observe que, pour une même valeur de score prédit, le nombre de conditions chroniques diffère selon le groupe ethnique. En effet, la courbe de calibrage pour les Afro-Américains se situe légèrement au-dessous de celle des Caucasiens. Cela signifie qu'à score égal, les assurés afro-américains présentent en réalité un nombre moins élevé de maladies chroniques que les caucasiens. Ce constat met en évidence un biais de calibrage du modèle, qui s'avère toutefois beaucoup moins prononcé que les biais des scores GLM et GLMM (voir figures 7.10 et 7.11). Cependant, on observe que ces biais sont plus prononcés pour les faibles risques et, dans une moindre mesure, pour les très hauts risques. Aussi, le modèle est relativement bien calibré au vu de la proximité des deux courbes de calibrage avec la courbe $y = x$. En effet, l'écart moyen entre la courbe de calibrage des assurés caucasiens et la courbe $y = x$ est de 0,01 tandis qu'il est de 0,04 pour les assurés afro-américains. Le modèle est donc mieux calibré pour les assurés caucasiens. En comparant les ECE obtenues pour chaque modèle et pour chaque groupe d'assurés (voir tableau 7.14), XGBoost apparaît comme le meilleur modèle en termes de calibrage. Aussi, le score XGBoost corrige en partie les biais observés dans les scores GLM et GLMM.

TABLE 7.14 – Comparaison des ECE globales et par groupe ethno-racial

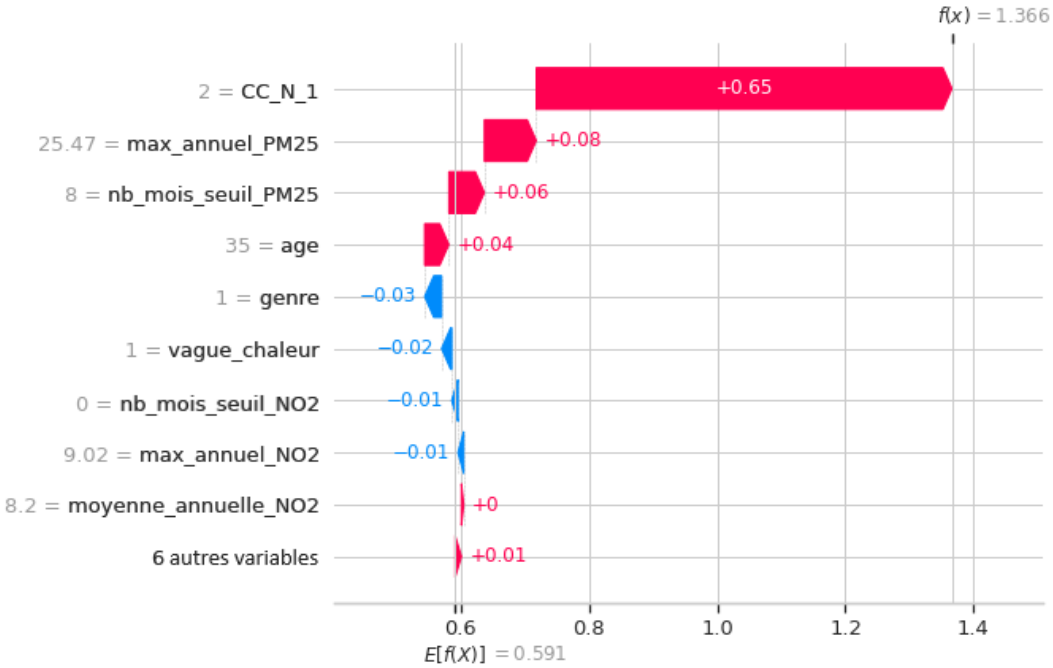
	GLM	**GLMM**	**XGBoost**
ECE globale	0,2250	0,1336	0,0087
ECE Caucasiens	0,2253	0,1339	0,0092
ECE Afro-Américains	0,2217	0,1380	0,0358

7.4.4 Importance des variables : analyse des valeurs SHAP

Comme mentionné dans la section 5.4.3, l'utilisation des valeurs SHAP est particulièrement intéressante pour interpréter les sorties d'un modèle de *machine learning*. En particulier, elles permettent d'attribuer à chaque variable une importance dans la prédiction de la variable cible. C'est cette valeur que nous allons analyser dans cette section et qui permettra de mesurer l'impact des variables environnementales sur la santé des assurés.

La figure 7.18 met en évidence l'impact de chaque variable explicative sur la prédiction individuelle du modèle. On observe que le nombre de conditions chroniques l'année précédente (`CC_N_1`) constitue le facteur le plus déterminant, avec une contribution positive notable de +0,65 à la prédiction finale. D'autres variables, telles que le pic annuel de particules fines (`max_annuel_PM25`), le nombre de mois dépassant le seuil de $PM_{2.5}$ (`nb_mois_seuil_PM25`) et l'âge, exercent également une influence positive, bien que moins marquée. A l'inverse, des variables comme le nombre annuel de vagues de chaleur (`vague_chaleur`) et le nombre de mois dépassant le seuil de NO_2 (`nb_mois_seuil_NO2`) contribuent à la diminution de la prédiction, traduisant un effet protecteur ou modérateur. Etre un homme (variable `genre`=0) a tendance à diminuer le nombre annuel de conditions chroniques. Ce résultat est cohérent avec les statistiques descriptives de la figure 7.4. Les autres variables ne jouent pas un rôle significatif dans la prédiction du nombre annuel de conditions chroniques.

FIGURE 7.18 – Graphique SHAP : importance des variables



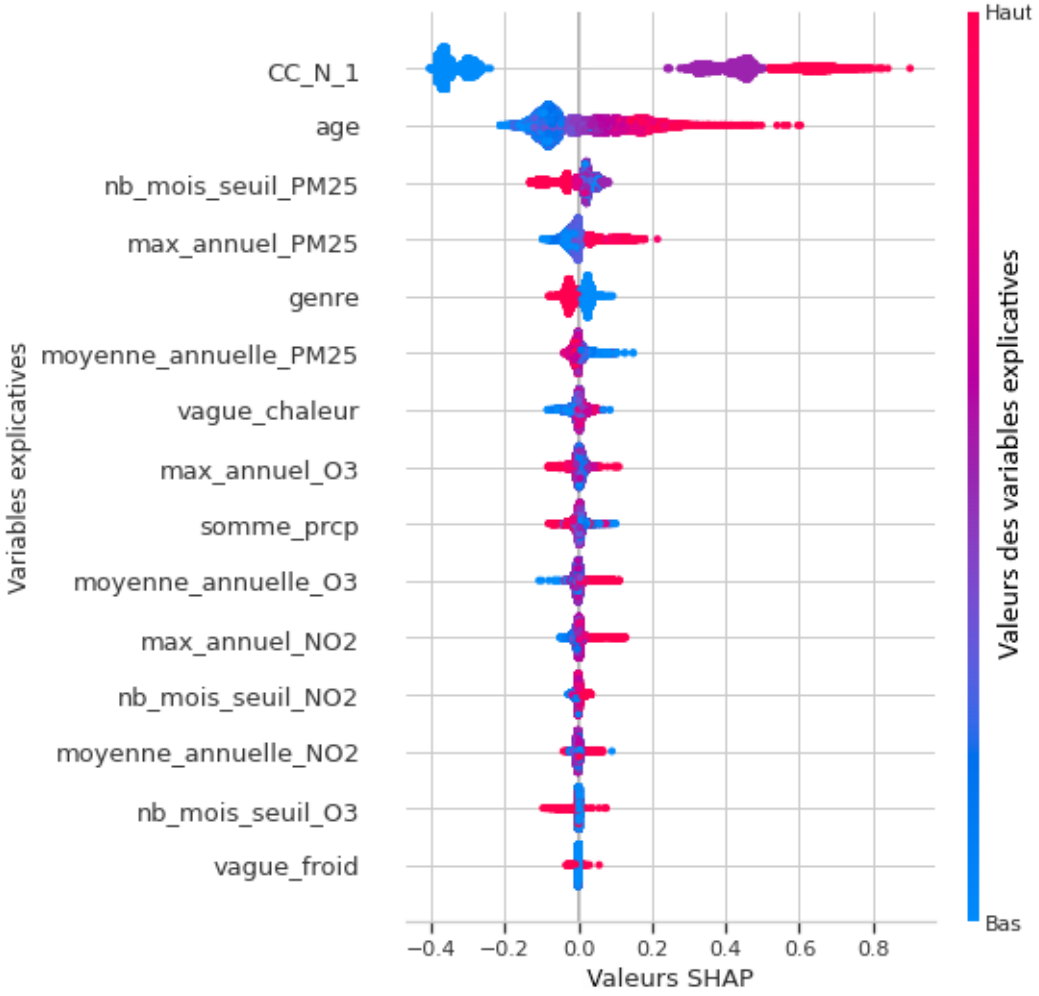
Note de lecture : l'importance de la variable `CC_N_1` dans la prédiction de `CC_N` est égale à +0,65.

La valeur de référence est égale à 0,591. Il s'agit de la valeur attendue sans information spécifique. La prédiction de 1,366 s'obtient en additionnant la valeur de référence et les contributions individuelles. La figure 7.18 permet ainsi d'identifier précisément les

facteurs ayant le plus d'influence sur le score de santé à l'échelle individuelle, et en facilite l'interprétation : ainsi, les indices environnementaux qui jouent le plus grand rôle dans la création du score de santé XGBoost sont les indicateurs relatifs aux émissions de $PM_{2.5}$ et aux vagues de chaleur.

Le graphique SHAP de la figure 7.19 constitue un outil supplémentaire dans l'analyse de l'importance relative des variables prédictives utilisées par le modèle XGBoost pour prédire l'issue étudiée. Les variables sont ordonnées de haut en bas selon leur valeur SHAP. Ce graphique corrobore ainsi les conclusions émanant de l'analyse de la figure 7.18 : `CC_N_1`, l'âge et l'exposition aux particules fines sont les facteurs les plus déterminants dans la prédiction du modèle. La dispersion des points le long de l'axe horizontal indique la variabilité de l'impact de chaque variable sur la prédiction individuelle : plus l'étalement est large, plus la variable influence fortement le score du modèle pour certains individus.

FIGURE 7.19 – Graphique SHAP : importance des variables en fonction de leurs valeurs



Note de lecture : les valeurs SHAP associées aux faibles valeurs de `CC_N_1` sont comprises entre -0,4 et -0,2.

La coloration des points, allant du bleu (valeurs faibles de la variable) au rouge (valeurs élevées), permet de visualiser le sens de l'effet. Par exemple, pour la variable `CC_N_1`, des valeurs élevées (en rouge) sont associées à des valeurs SHAP positives, ce qui signifie

qu'un nombre plus important de conditions chroniques tend à augmenter le score de santé développé. A l'inverse, certaines variables peuvent avoir un effet inverse ou plus nuancé selon la distribution de leurs valeurs. L'âge possède d'une part une bonne dispersion et d'autre part une bonne séparation des couleurs. Ainsi, des âges élevés sont liés à des valeurs SHAP positives, suggérant un impact croissant sur le score de santé. Les âges faibles provoquent l'effet inverse. Aussi, une transition nette du bleu au rouge indique une relation monotone entre le score et la variable prédictive, tandis qu'une disposition erratique des couleurs suggère une interaction plus complexe à identifier.

Ainsi, il est difficile de se prononcer sur les effets des variables suivantes sur le score créé, car elles ne présentent pas une grande dispersion et les points sont centrés sur la valeur SHAP nulle :

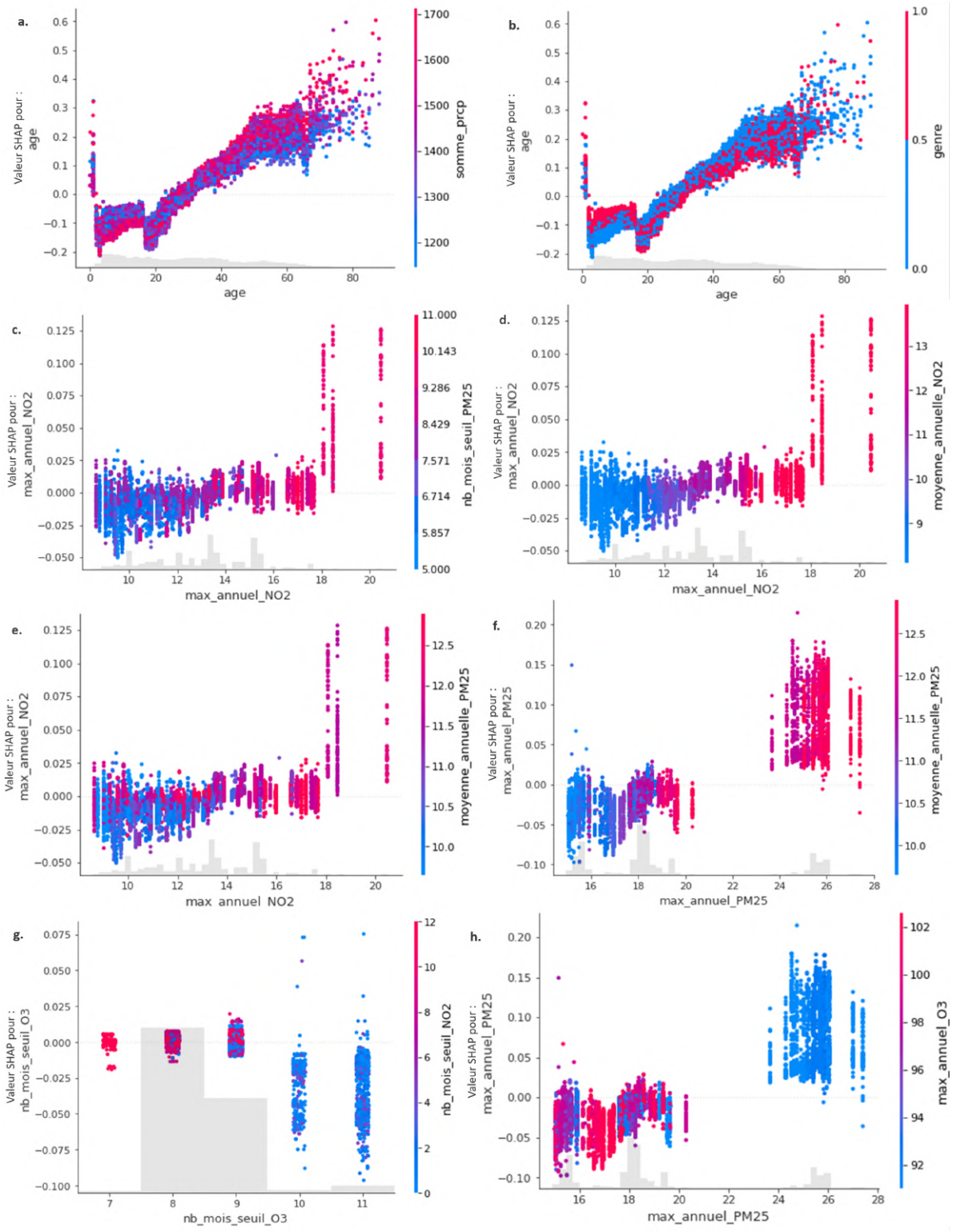
— `vague_froid` : nombre de vagues de froid dans l'année ;
— `vague_chaleur` : nombre de vagues de chaleur ;
— `somme_prcp` : somme des précipitations sur une année ;
— `moyenne_annuelle_NO2` : concentration moyenne mensuelle de NO_2 dans l'année ;
— `nb_mois_seuil_NO2` : nombre de mois où le seuil de NO_2 a été dépassé dans l'année ;
— `max_annuuel_O3` : concentration maximale mensuelle d'O_3 dans l'année ;
— `nb_mois_seuil_O3` : nombre de mois où le seuil d'O_3 a été dépassé dans l'année ;
— `max_annuel_PM25` : concentration maximale mensuelle de $PM_{2.5}$ dans l'année.

Certains indicateurs climatiques ont une importance, même si parfois minime, dans la construction du score de santé. Dans l'ordre d'importance de la figure 7.19, on retrouve :

— `nb_mois_seuil_PM25` : la dispersion des valeurs SHAP correspondant est assez limitée comparativement aux variables `CC_N_1` et `age`. L'effet observé est contre-intuitif puisque les grandes valeurs ont tendance à diminuer le nombre de conditions chroniques prédit.
— `max_annuel_PM25` : cette variable présente une dispersion des valeurs SHAP. Elle contribue donc au score de santé créé. Les fortes valeurs de concentrations en $PM_{2.5}$ peuvent donc avoir un impact sur le score, qui a tendance à augmenter.
— `moyenne_annuelle_O3` : même conclusion que pour la variable précédente. Les valeurs élevées sont associées à des contributions légèrement positives, indiquant que des niveaux moyens élevés d'ozone pourraient être un facteur aggravant.
— `max_annuel_NO2` : la variable montre une dispersion faible des valeurs SHAP autour de zéro indiquant une influence modérée sur la prédiction. Toutefois, les valeurs élevées sont associées à une importance positive sur les prédictions. Ainsi, une exposition à des maxima de NO_2 peut accroître le risque estimé par le modèle.

Pour approfondir l'analyse de l'impact du climat et de la pollution sur le score de santé développé à l'aide du modèle XGBoost, il est possible de réaliser des analyses croisées de l'importance des variables en fonction de leurs valeurs. Ces analyses sont présentées dans la figure 7.20. Les huit sous-figures représentent les valeurs SHAP associées à une variable prédictive donnée en fonction de la distribution de cette variable. Parallèlement sont ajoutées en couleur (du bleu au rouge) les valeurs d'une seconde variable explicative. Par exemple, la sous-figure **a.** illustre l'importance de la variable `age` en fonction de ses valeurs, avec une coloration selon les valeurs de `somme_prcp`.

FIGURE 7.20 – Graphique SHAP : analyse croisée de l'importance des variables en fonction de leurs valeurs



Note de lecture : les valeurs SHAP associées à l'âge 0, combinées à l'effet des précipitations, varient de 0 à 0,32 (sous-figure ***a.****).*

La sous-figure **a.** montre que les valeurs SHAP augmentent avec l'âge, comme observé précédemment. La forme de la courbe obtenue rappelle celle de la figure 7.3 qui repré-

sente la distribution de `CC_N` en fonction de l'âge. Elle rappelle la forme d'une courbe de mortalité : les nourrissons sont plus à risque que les jeunes enfants, puis les risques augmentent linéairement après 20 ans. A âge égal, à partir de 20 ans, une exposition à davantage de précipitations accroît l'occurrence de conditions chroniques l'année suivante (gradient de couleur du bleu au rouge pour un âge donné). Pour les âges inférieurs, la tendance s'inverse, sans explication évidente à ce stade. La sous-figure **b**. colore les points selon le genre (0 pour les femmes, 1 pour les hommes). On observe, en valeur absolue et à âge égal, qu'être une femme conduit à une plus grande importance de la variable `age`, ce qui signifie que, toutes choses égales par ailleurs, l'âge contribue davantage à la prédiction du score de santé chez les femmes, suggérant une vulnérabilité accrue liée à l'âge dans cette population.

La sous-figure **c.** met en exergue l'augmentation de l'importance de la variable `max_annuel_NO2` dans la prédiction de `CC_N` avec ses valeurs. Ainsi, plus le maximum annuel de la concentration en NO_2 augmente, plus il a un impact à la hausse sur le nombre individuel de conditions chroniques observé l'année suivante. Ce risque, lié aux maxima annuels de NO_2, est aggravé par l'exposition aux $PM_{2.5}$ comme le montrent les sous-figures **c.** et **e.**. La sous-figure **d.** confirme l'impact positif de l'exposition au NO_2 sur le score de santé développé avec le modèle XGBoost.

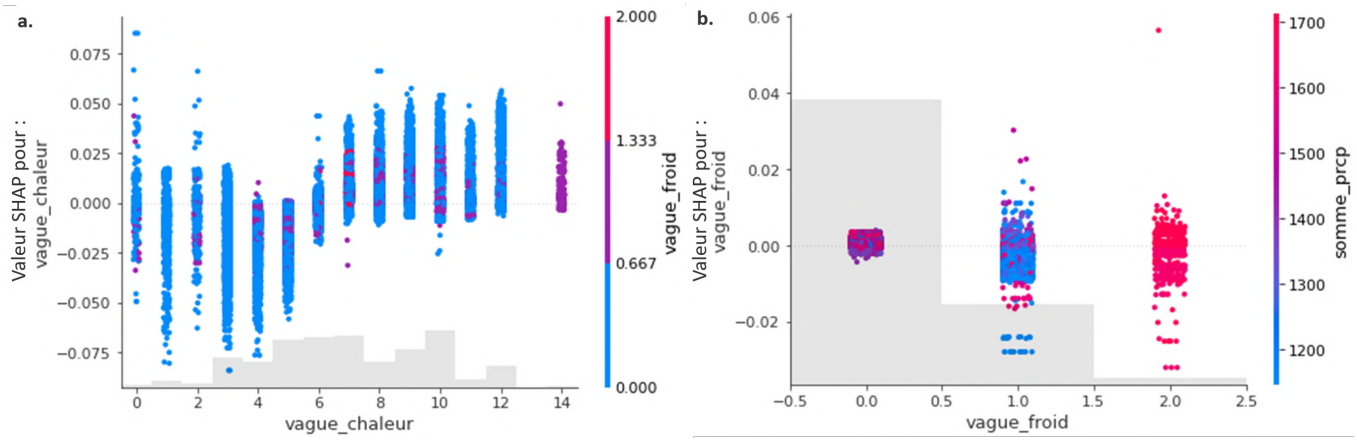
Sur la sous-figure **f.**, on observe que plus la valeur du maximum annuel de $PM_{2.5}$ est élevée, plus la valeur SHAP associée l'est aussi. Cela signifie que des pics annuels élevés de $PM_{2.5}$ augmentent le risque de développer des conditions chroniques l'année suivante. Pour un même niveau de maximum annuel, les fortes moyennes annuelles de $PM_{2.5}$ sont souvent associées à des valeurs SHAP plus élevées. La pente des valeurs SHAP s'accentue à partir d'un certain seuil de $PM_{2.5}$ (21 µg/m^3) : l'impact sur la santé devient beaucoup plus marqué au-delà de ce seuil. La forte dispersion verticale des points laisse sous-entendre que l'effet n'est pas porté uniquement par `max_annuel_PM25`.

Les sous-figures **g.** et **h.** confirment l'effet contre-intuitif de l'ozone sur le score de santé issu du XGBoost : une coupure nette sur les deux graphes indique que l'impact sur la santé devient beaucoup plus faible au-delà d'un certain seuil pour les variables `nb_mois_seuil_O3` et `max_annuel_O3`. A ce stade, aucune explication ne permet de justifier l'effet à la baisse des concentrations d'ozone sur le nombre individuel de conditions chroniques l'année suivante, relaté à la fois par le modèle XGBoost mais également par le GLM et le GLMM.

La figure 7.21 se focalise sur les variables dont l'importance est difficile à analyser : `vague_froid`, `vague_chaleur` et `somme_prcp`. On remarque sur la sous-figure **a.** que plus le nombre de vagues de chaleur annuel est élevé, plus la valeur SHAP associée l'est aussi. Ce graphique traduit un effet positif des vagues de chaleur sur le score : une hausse du nombre annuel de vagues de chaleur entraînerait un risque de développer davantage de conditions chroniques l'année suivante. Le constat n'est pas aussi clair pour les vagues de froid. En effet, les sous-figures **a.** et **b.** ne parviennent pas à démontrer un impact significatif de cette variable sur le score créé. En revanche, les grandes valeurs de `vague_froid` sont associées à des cumuls annuels de précipitations élevés, traduisant une relation entre ces deux variables. Concernant l'importance de la variable `somme_prcp`, la sous-figure **a.** de la figure 7.20 ainsi que la figure 7.19 soulignent un effet à la fois faible

et ambigu sur le score développé.

FIGURE 7.21 – Graphique SHAP : analyse croisée de l'importance des variables `vague_froid`, `vague_chaleur` et `somme_prcp` en fonction de leurs valeurs



Note de lecture : les valeurs SHAP associées à un nombre de vagues de froid égal à 2 (valeurs correspondant à des cumuls de précipitations annuels élevés) varient de -0,03 à 0,06.

L'ensemble de ces conclusions est à mettre au regard des faibles valeurs de Shapley associées aux variables environnementales, ce qui suggère que, bien que certains facteurs climatiques et de pollution de l'air impactent le score de santé XGBoost, leur poids reste limité comparativement aux variables propres à l'assuré (âge, genre et nombre de conditions chroniques antérieures). L'interprétation causale de l'effet des variables environnementales dans le modèle doit donc rester prudente. Ces conclusions constituent une première piste, mais soulignent la nécessité de travaux complémentaires pour mieux comprendre leur rôle dans la construction du score de santé développé.

7.4.5 Conclusions et limites

Le modèle XGBoost pour l'élaboration d'un score de santé basé sur le nombre annuel de conditions chroniques a permis d'atteindre des performances nettement supérieures à celles des modèles linéaires traditionnels (GLM et GLMM). Les indicateurs de performance (MAE, MSE et RMSE) témoignent de l'utilité d'optimiser rigoureusement les hyperparamètres. Cette optimisation a ainsi permis de mieux calibrer le modèle. Au-delà de la performance brute, le modèle XGBoost présente également une meilleure équité en termes de biais ethno-raciaux. Par rapport aux modèles linéaires, l'écart moyen de calibrage est nettement réduit, ce qui traduit une meilleure équité. Aussi, l'interprétabilité des résultats a été rendue possible grâce aux valeurs SHAP qui ont mis en lumière l'importance de l'état de santé antérieur, de l'âge et du genre, mais aussi l'impact significatif de certains facteurs environnementaux : l'exposition aux vagues de chaleur, aux particules fines ($PM_{2.5}$) ou au NO_2 traduit une augmentation du score de santé ainsi développé.

Cependant, le score de santé individuel développé à l'aide du modèle XGBoost n'est pas parfait et comporte plusieurs limites :

1. **Biais persistants** : les biais ethniques subsistent notamment pour les assurés présentant des scores très faibles ou très élevés. A score de santé égal, le nombre

de conditions chroniques réel de l'assuré est en réalité plus faible chez les Afro-Américains que chez les Caucasiens. L'équité du modèle pourrait être approfondie par des méthodes d'ajustement ou de post-traitement spécifiques. Ces méthodes ne constituent pas le cœur de ce mémoire et ne seront donc pas détaillées ici.

2. **Interprétabilité partielle** : les valeurs SHAP permettent une lecture globale et individuelle des facteurs influents, mais certaines relations restent difficiles à expliquer. Par exemple, l'effet contre-intuitif de l'ozone et des vagues de froid sur le score développé pourrait résulter de corrélations non-causales ou d'indicateurs pas assez représentatifs du phénomène qu'ils sont censés capter. Aussi, la maille annuelle imposée par le format des données écrase les informations contenues dans les indicateurs environnementaux. Il est possible qu'une vague de froid touchant le Kentucky en 2016 n'ait pas d'impact sur les assurés de l'Etat en 2017.
3. **Limites structurelles du XGBoost** : le modèle XGBoost n'est pas le seul modèle de *machine learning* qui permette un bon calibrage et qui offre la possibilité d'interpréter les résultats. D'ailleurs, il peut ne pas capter l'ensemble des interactions temporelles ou spatiales susceptibles d'influencer les trajectoires de santé. L'exploration de modèles alternatifs comme les réseaux de neurones ou encore les modèles séquentiels pourrait pallier ces limites, mais elle ne permettrait pas de respecter les contraintes financières et computationnelles de cette étude.
4. **Généralisation** : le modèle a été entraîné et validé sur un échantillon de 50 000 assurés du Kentucky. Son application à d'autres groupes d'assurés des Etats-Unis (sujets à davantage de vagues de froid par exemple) pourrait mettre en lumière des interactions différentes entre la variable cible et les prédicteurs.

Chapitre 8

Score de santé mensuel basé sur les données de sinistres

Une seconde approche a été d'exploiter conjointement les bases de données « assurés » et « sinistres » pour créer des indicateurs de santé mensuels s'appuyant à la fois sur les informations transmises par les assurés lors de la souscription, mais également sur les données de frais de santé disponibles sur la période d'étude. L'utilisation de la base « sinistres » ainsi que l'agrégation des données à la maille mensuelle (une observation par individu et par mois) impliquent des contraintes financières et computationnelles qui limitent la possibilité d'implémenter des modèles complexes et d'utiliser l'ensemble des assurés du Kentucky. Ce chapitre accordera une importance au respect et à la description de ces contraintes lors de l'exploitation des données.

8.1 Base de données

8.1.1 Description des bases de données exploitées

Les bases utilisées pour développer les scores de santé annuels sont les bases « assurés », « souscriptions » et « sinistres » concernant 1 000 assurés parmi les 50 000 assurés du Kentucky sélectionnés dans le chapitre précédent. Pour rappel, la base « souscriptions » fournit des informations mensuelles entre le 1er janvier 2017 et le 31 décembre 2023 sur les types d'assurances souscrites par chaque individu de la base et sur le groupe de pathologies chroniques associé à chaque souscription. La base « sinistres » recense l'ensemble des événements médicaux donnant lieu à remboursement ou à facturation pour la période allant du 1er janvier 2017 au 31 décembre 2023. Chaque enregistrement correspond à une prestation médicale précise (consultation, examen, hospitalisation, délivrance d'un médicament, etc.) et regroupe plusieurs dimensions d'informations comme les coûts associés à chaque prestation, le montant payé par le régime d'assurance, la typologie des prestations, les caractéristiques du sinistre ou encore les informations cliniques. Cette base est particulièrement volumineuse. A titre d'exemple, parmi les 551 335 individus du Kentucky présents dans la base « assurés » (et donc possédant une assurance), 513 082 ont au moins une observation dans la base « sinistres », soit environ 93 %. L'utilisation de la totalité des sinistrés du Kentucky aboutirait à plus de 160 millions d'observations. Un tel volume de données représenterait, d'une part, un coût d'utilisation de la plateforme *Databricks* trop élevé et nécessiterait, d'autre part, des temps de calcul particulièrement longs, même en utilisant *Spark*. C'est pourquoi, la taille de l'échantillon a été drastique-

ment réduite, d'autant plus que la sélection d'un nombre fini d'assurés dans une base de 160 millions d'observations est également chronophage : il a fallu compter près de deux heures pour sélectionner les sinistres des 1 000 individus à partir de leur identifiant. L'échantillon représentatif de 1 000 assurés du Kentucky est toujours constitué de :

— 50 % d'hommes et 50 % de femmes ;
— 8 % d'Afro-Américains, 92 % de Caucasiens.

La variable clé de la base « sinistres » qui donne le coût réel de chaque prestation par individu constitue la variable de référence pour la construction du score mensuel. En considérant un échantillon de 1 000 assurés du Kentucky, la base « sinistres » contient plus de 430 000 observations concernant 97,2 % d'entre eux. Lorsque les données sont agrégées à la maille temporelle mensuelle, il est possible de recenser la somme mensuelle des frais de santé de chaque assuré à partir de janvier 2017 et jusqu'à décembre 2023. L'objectif est d'utiliser cette variable comme proxy de l'état de santé des individus constituant alors la cible d'un modèle de prédiction visant à attribuer un score de santé mensuel à chaque personne. Cette agrégation permet par ailleurs de réduire considérablement le volume de la base de données (une observation par mois et par année pour chaque individu) tout en gardant un maximum d'informations.

8.1.2 Agrégation des données à la maille mensuelle

Pour développer le score de santé mensuel, il est nécessaire de construire des indicateurs mensuels qui constitueront les variables prédictives potentielles du modèle implémenté. Ainsi, pour chaque mois, de 2017 à 2023, les variables socio-démographiques, sanitaires et environnementales ci-après ont été collectées dans les trois bases de données de l'étude. A titre indicatif, il convient de préciser que, dans la suite, les variables dont le nom se termine par `lag1` correspondent à la valeur de l'indicateur considéré au cours du mois précédant la période d'observation (mois $t-1$), tandis que celles se terminant par `lag2` font référence à la valeur observée deux mois avant la période d'observation (mois $t-2$). Les autres variables sont relatives à la période d'observation. Par ailleurs, sauf mention contraire, les variables sont agrégées à la maille mensuelle. Les variables agrégées sont les suivantes :

— `mois_annee` : mois et année de l'observation ;
— `somme_montant_facture` (variable cible) : somme totale des frais de santé pour chaque assuré au mois et à l'année de l'observation ;
— `somme_montant_facture_lag1` : somme totale des frais de santé pour chaque assuré (mois précédent) ;
— `zip3` : ZIP3 de résidence de l'assuré ;
— `age` : âge des individus dans l'échantillon ;
— `genre` : genre des individus (0-Femme, 1-Homme) ;
— `somme_montant_autorise_lag1` : montant total autorisé par l'assureur ;
— `somme_montant_cob_lag1` : montant total pris en charge par la coordination des prestations ;
— `somme_montant_copaiement_lag1` : montant total des copaiements ;
— `somme_montant_franchise_lag1` : montant total de la franchise ;

— `somme_montant_coassurance_lag1` : montant total de coassurance ;
— `somme_montant_paye_lag1` : montant total effectivement payé par l'assureur ;
— `somme_montant_refuse_lag1` : montant total des prestations refusées ou non remboursées ;
— `nb_diagnostics_icd_01_lag1` : nombre de diagnostics ICD différents ;
— `nb_reclamations_medicaments_lag1` : nombre de réclamations liées aux médicaments ;
— `nb_reclamations_hopital_lag1` : nombre de réclamations hospitalières ;
— `prop_reclamations_reseau_lag1` : proportion de réclamations dans le réseau agréé ;
— `cout_total_medicaments_lag1` : somme totale des frais de médicaments ;
— `nb_renouvellements_rx_lag1` : nombre de renouvellements d'ordonnance ;
— `moy_mensuelle_prcp_lag1` : moyenne mensuelle des précipitations ;
— `moy_mensuelle_tavg_lag1` : moyenne mensuelle des températures moyennes ;
— `moy_mensuelle_tmin_lag1` : moyenne mensuelle des températures minimales ;
— `moy_mensuelle_tmax_lag1` : moyenne mensuelle des températures maximales ;
— `vague_chaleur_lag1` : nombre de vagues de chaleur ;
— `vague_froid_lag1` : nombre de vagues de froid ;
— `vague_froid_lag2` : nombre de vagues de froid (deux mois avant la période d'observation) ;
— `precipitation_totale_lag1` : cumul des précipitations ;
— `moyenne_pm25_lag1` : concentration moyenne de $PM_{2.5}$;
— `moyenne_o3_lag1` : concentration moyenne d'O_3 ; et
— `moyenne_no2_lag1` : concentration moyenne de NO_2.

Une base de données mensuelle contenant des informations chaque mois sur chaque assuré de l'échantillon représentatif du Kentucky est ainsi obtenue. Elle comprend 77 560 observations après nettoyage. Ce nombre d'observations obtenu avec seulement 1 000 assurés met en évidence les limites d'un échantillon trop important. Par exemple, sélectionner 50 000 assurés conduirait à une base de plus d'environ 4 millions d'observations en moyenne entraînant des temps de calcul déraisonnables.

Il est important de noter que l'ensemble des indicateurs climatiques et de pollution sont décalés d'un mois par rapport à la variable cible `somme_montant_facture` : autrement dit, pour chaque date d'observations, la variable cible correspond à la somme des frais de santé en un mois donné d'un assuré, tandis que les variables environnementales et climatiques utilisées comme prédicteurs correspondent au mois précédent. Cela permet d'étudier l'effet différé des conditions environnementales sur la santé à une maille plus fine que le score de santé créé précédemment (voir chapitres 3 et 7).

Un exemple fictif est donné dans le tableau 8.1 : il s'agit d'une femme de 65 ans vivant dans le ZIP3 401 qui, en janvier 2016, a eu des frais de santé s'élevant à 2 000 $. Le mois précédent :

— trois vagues de froid ont été enregistrées dans le ZIP3 401 ;

— ses frais de santé se sont élevés à 1 500 $; et
— ses frais de médicaments ont été de 500 $.

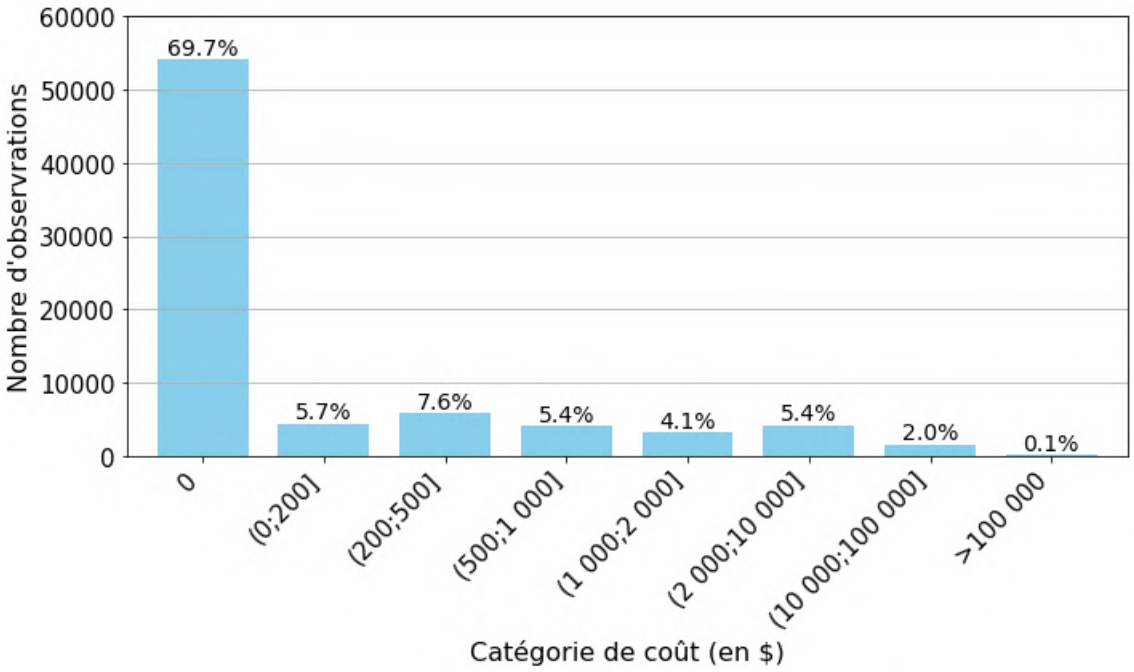
TABLE 8.1 – Exemple d'une observation de la base annuelle créée

Variable	Valeur
ID	0000
ZIP3	401
mois_annee	01-2016
age	65
genre	1
somme_montant_facture	2 000
somme_montant_facture_lag1	1 500
cout_total_medicaments_lag1	500
vague_froid_lag1	3
...	...

8.1.3 Description de la variable cible

La distribution de la variable cible `somme_montant_facture` dans la base mensuelle créée et décrite ci-dessus est donnée dans la figure 8.1.

FIGURE 8.1 – Distribution de la variable `somme_montant_facture`



Note de lecture : 69,7 % des coûts observés sont nuls.

On comptabilise 69,7 % de valeurs nulles. 5,7 % des frais mensuels sont compris entre 0 et 200 $ et 7,6 % entre 200 et 500 $. Les frais mensuels supérieurs à 1 000 $ correspondent à 11,6 % des observations. La distribution de `somme_montant_facture` présente une importante asymétrie, avec une grande majorité d'observations à 0 $.

Une autre approche d'analyse possible est une représentation géographique. La figure 8.2 illustre la moyenne de la variable cible `somme_montant_facture` pour chaque ZIP3. En se référant aux clusters décelés dans le partitionnement effectué dans la section 6.6, les conclusions sont similaires à celles mises en évidence pour le nombre de conditions chroniques annuel. En effet, on observe que les ZIP3 affichant les moyennes les plus élevées de la variable cible sont situés dans le cluster 3. *A contrario*, les ZIP3 avec une moyenne des frais mensuels plus faible sont situés dans le cluster 1. Si l'agrégation est effectuée à l'échelle des clusters, et non à l'échelle des ZIP3, on obtient des frais de santé mensuels moyen de :

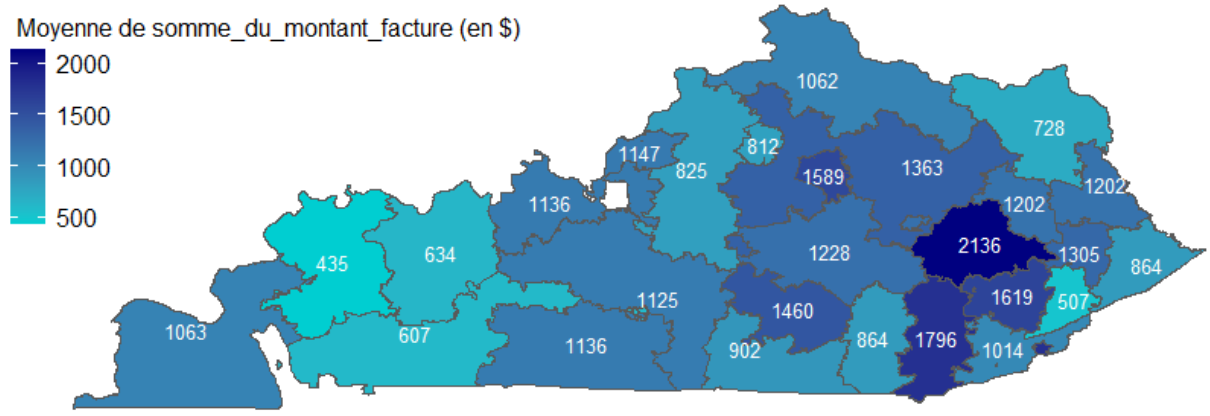
— 685 $ pour le cluster 1 ;

— 1 142 $ pour le cluster 2 ;

— 1 200 $ pour le cluster 3.

Pour rappel, le cluster 1 est caractérisé par des températures moyennes plus élevées, des précipitations journalières plus faibles ainsi que des concentrations de NO_2 et de $PM_{2.5}$ supérieures à celles des autres clusters. Ainsi, les résultats de cette analyse géographique sont en tout point similaires à ceux observés pour le nombre de conditions chroniques annuel moyen. Bien que surprenants, ils restent néanmoins cohérents avec les résultats antérieurs. Ces conclusions sont à analyser avec prudence, car il s'agit d'une analyse univariée.

FIGURE 8.2 – Moyenne des frais mensuels (`somme_montant_facture`) par ZIP3 dans le Kentucky



Note de lecture : dans le ZIP3 413, les frais moyens mensuels liés à des soins de santé s'élèvent à 2 136 $.

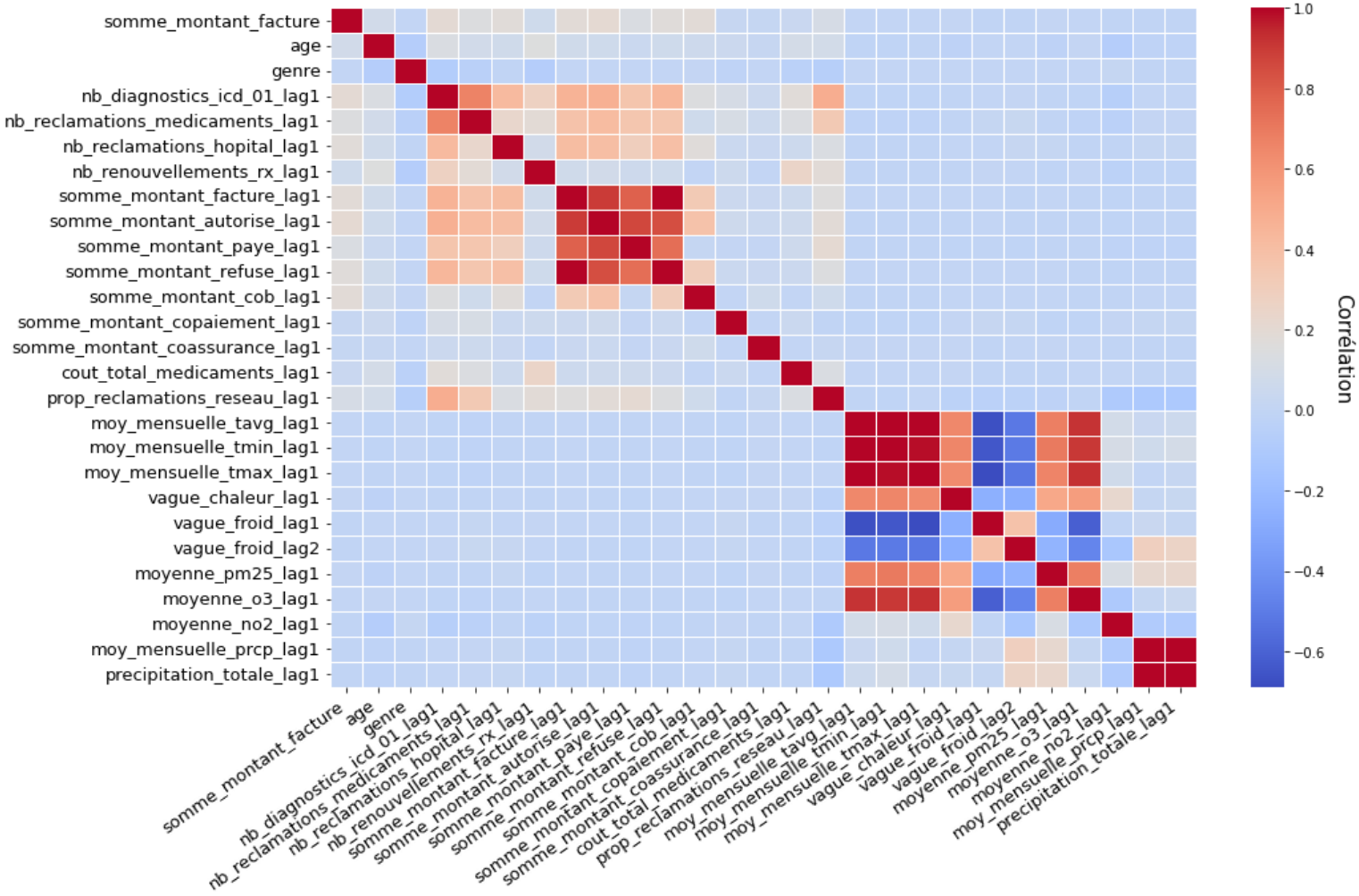
8.2 Score basé sur la régression linéaire

8.2.1 Analyse des corrélations entre variables explicatives

Comme mentionné dans le chapitre précédent, la sélection rigoureuse des variables est importante pour évaluer la multicolinéarité dans une régression linéaire : elle assure la robustesse et l'interprétabilité du modèle linéaire. Pour ce faire, une première étape consiste à analyser les corrélations entre la variable cible et les 28 variables prédictives ainsi que les

corrélations entre les variables prédictives elles-mêmes. La matrice de corrélation relative à l'ensemble des données disponibles est donnée en figure 8.3.

FIGURE 8.3 – Matrice de corrélation des 26 variables prédictives disponibles



Note de lecture : la variable `moy_mensuelle_prcp_lag1` *est fortement corrélée à* `moy_mensuelle_tavg_lag1`.

L'analyse approfondie de la matrice de corrélation révèle que la variable cible n'est fortement corrélée à aucune des variables prédictives disponibles. Elle met toutefois en évidence plusieurs relations significatives entre les variables de l'étude. On observe tout d'abord la présence de groupes de variables corrélées positivement (zones de couleur rouge foncé situées hors de la diagonale principale) :

— les quatre variables regroupant les coûts mensuels du mois précédant la date d'observation, à savoir `somme_montant_facture_lag1`, `somme_montant_autorise_lag1`, `somme_montant_paye_lag1` et `somme_montant_refuse_lag1` ;

— les moyennes de températures minimales, moyennes et maximales ;

— `moyenne_o3_lag1` et les moyennes de températures minimales, moyennes et maximales ; et

— `moy_mensuelle_prcp_lag1` et `precipitation_totale_lag1`.

A l'inverse, certains groupes de variables présentent des corrélations négatives marquées, indiquées par des carrés bleu foncé. On observe notamment que les indicateurs de températures ainsi que la concentration moyenne en ozone sont négativement corrélés aux nombres de vagues de froid (corrélations comprises entre -0,69 et -0,61).

L'analyse des multicolinéarités ainsi que les statistiques descriptives complètes (voir annexes D.2 et D.3) conduisent à ôter les indicateurs mensuels moyens de précipitations, de températures minimales, moyennes et maximales ainsi que les variables `somme_montant_autorise_lag1`, `somme_montant_paye_lag1` et `somme_montant_refuse_lag1` de la liste des prédicteurs.

TABLE 8.2 – Facteurs d'inflation de la variance (VIF) et coefficients issus de la régularisation Lasso des 19 variables explicatives potentielles

Variable	**VIF**	**Coefficient Lasso**
moyenne_o3_lag1	4,2	0,00
nb_diagnostics_icd_01_lag1	2,7	637,22
moyenne_pm25_lag1	2,4	15,12
nb_reclamations_medicaments_lag1	1,9	303,70
vague_froid_lag1	1,8	0,00
vague_chaleur_lag1	1,7	12,42
somme_montant_facture_lag1	1,5	423,26
vague_froid_lag2	1,5	-38,26
prop_reclamations_reseau_lag1	1,4	231,83
nb_reclamations_hopital_lag1	1,3	781,43
moyenne_no2_lag1	1,3	0,00
precipitation_totale_lag1	1,2	0,00
nb_renouvellements_rx_lag1	1,2	125,47
somme_montant_cob_lag1	1,1	1207,32
cout_total_medicaments_lag1	1,1	0,00
age	1,0	376,75
genre	1,0	93,46
somme_montant_copaiement_lag1	1,0	0,00
somme_montant_coassurance_lag1	1,0	-44,38

Afin d'approfondir l'analyse de la multicolinéarité et d'identifier les variables à conserver dans le modèle final, deux outils complémentaires sont utilisés : le facteur d'inflation de la variance (VIF) et la régression Lasso. A l'instar du processus de sélection utilisé pour le score de santé annuel, nous considérerons qu'une valeur de VIF supérieure à 8 indique une multicolinéarité importante. Parallèlement, la régression Lasso, qui applique une pénalisation sur la somme des valeurs absolues des coefficients, favorise la sélection automatique des variables les plus pertinentes en contraignant les coefficients des variables moins informatives à s'annuler. Ces deux méthodes ont été testées sur l'ensemble des variables prédictives. Les résultats sont donnés dans le tableau 8.2.

L'association de ces trois méthodes (matrice de corrélation, VIF, régularisation Lasso) permet donc de guider objectivement le choix des variables à exclure, en privilégiant celles qui contribuent le plus à la qualité prédictive du modèle tout en limitant la redondance

d'information. Le VIF ne suggère aucune exclusion de variables. En effet, les VIF obtenus sont tous inférieurs à 8. Par ailleurs, la variable `moyenne_o3_lag1` possède un coefficient de Lasso nul et est assez corrélée aux températures ainsi qu'aux vagues de chaleur et aux vagues de froid. Ainsi, elle a été évincée de l'étude. Aussi, même si les autres variables ont un coefficient Lasso nul, nous décidons de les garder. Elles n'auront pas d'impact néfaste sur la convergence de la régression linéaire puisque leur VIF est assez faible. Les corrélations obtenues (voir annexe D.1) sont toutes inférieures à 0,70 en valeur absolue. Ainsi, la matrice confirme l'absence de colinéarité entre les variables prédictives choisies.

8.2.2 Statistiques descriptives en lien avec la variable cible

L'analyse des corrélations entre variables prédictives est à mettre au regard des statistiques descriptives des prédicteurs sélectionnés. En effet, si une variable présente une distribution très déséquilibrée ou si elle n'a que très peu d'impact sur la distribution de la variable cible, son interprétation et sa contribution au modèle peuvent être contre-productives. C'est pourquoi l'analyse de la multicolinéarité a été complétée par une étude statistique des variables présélectionnées.

TABLE 8.3 – Pourcentage de valeurs nulles parmi les variables de réclamations et de frais différés.

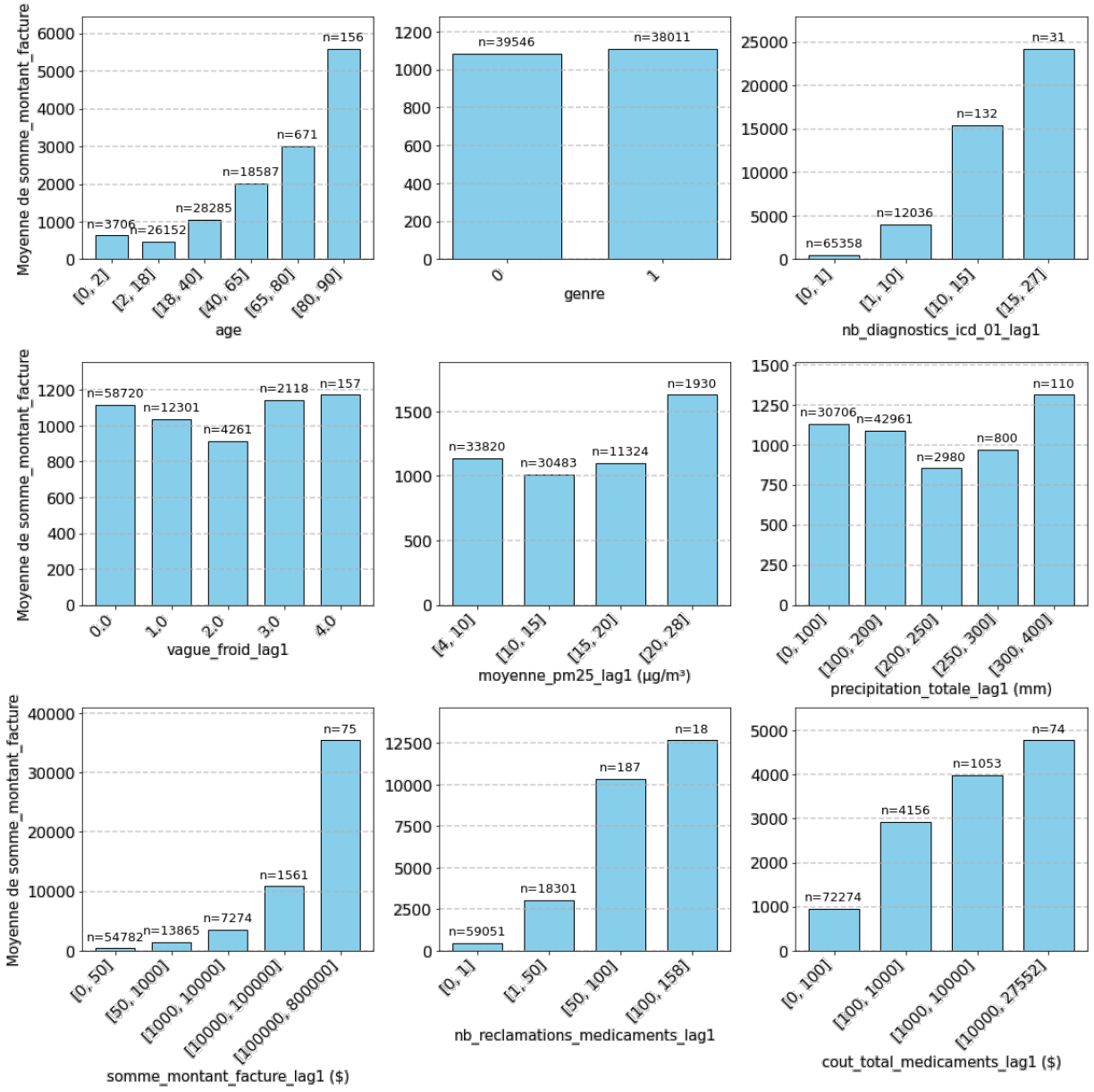
Variable	**Pourcentage de valeurs nulles**
nb_reclamations_hopital_lag1	97,1 %
somme_montant_cob_lag1	99,6 %
somme_montant_copaiement_lag1	99,4 %
somme_montant_coassurance_lag1	99,9 %

Le tableau 8.3 représente le pourcentage de valeurs nulles parmi les variables explicatives `nb_reclamations_ hopital_lag1`, `somme_montant_cob_lag1`, `somme_montant_ _lag1` et `somme_ montant_coassurance_lag1`. Pour chacune de ces variables, la majorité des observations se concentre sur une plage très basse de valeurs (en majorité des valeurs nulles) tandis que les autres tranches ne contiennent que très peu d'observations. Par exemple, 77 118 valeurs des frais mensuels de copaiement sont nulles. Cela représente 99,4 % des observations. Par conséquent, la variable `somme_montant_copaiement_lag1` ne pourra pas parvenir à expliquer le score de santé développé. Les distributions de ces quatre variables indiquent une répartition très déséquilibrée pouvant, d'une part, réduire la significativité statistique de ces variables dans la régression et, d'autre part, biaiser le modèle. Même si les tranches supérieures semblent avoir un impact sur les frais de santé mensuels individuels, elles sont basées sur un effectif faible, ce qui limitera la robustesse des estimations associées. Ainsi, pour garantir la qualité du modèle et éviter le surajustement, ces quatre variables ne seront pas intégrées à la régression linéaire implémentée pour créer le score de santé mensuel.

La figure 8.4 illustre, à travers une série de graphiques, la variation du montant moyen des frais de santé mensuels individuels selon différentes catégories de variables explicatives sélectionnées précédemment. Ces statistiques descriptives doivent permettre ainsi de mettre en lumière des tendances et d'identifier les facteurs associés à des coûts plus

ou moins élevés. Les graphiques liés à la totalité des prédicteurs sont disponibles en annexes D.2 et D.3.

FIGURE 8.4 – Evolution de la variable cible selon différentes variables explicatives



Note de lecture : les femmes (genre=1) ont en moyenne des frais de santé mensuels s'élevant à 1 100 $.

Le premier sous-graphique dégage une relation très marquée entre l'âge et les frais de santé mensuels facturés : ces frais sont plus élevés chez les personnes les plus âgées. Nous pouvons tirer les mêmes conclusions que dans le chapitre précédent : chez les très jeunes enfants (0 à 2 ans), les coûts sont plus élevés que chez les mineurs âgés de 3 à 18 ans. Après 18 ans, les frais de santé mensuels augmentent petit à petit avec l'âge, traduisant une vulnérabilité accrue des personnes plus âgées et une consommation plus importante du système de santé. Cette tendance s'accentue à la quarantaine et se pérennise après 65 ans. La forme du graphique rappelle, une fois encore, celle d'une courbe de mortalité.

Le genre n'affiche pas de disparité notable dans le montant moyen des frais mensuels, les deux modalités présentant des moyennes similaires et des effectifs équilibrés. Le sexe de l'assuré ne représente donc pas un facteur discriminant majeur du coût des soins et de l'utilisation des services de santé.

L'exposition aux vagues de froid ainsi qu'aux précipitations totales dans le mois précédant la date d'observation révèle une corrélation positive avec le montant moyen des frais de santé. Les valeurs les plus élevées de ces deux variables sont associées à des coûts plus élevés, traduisant ainsi une augmentation des frais liés au froid et aux précipitations. Cependant, la relation n'est pas linéaire : si certaines catégories de précipitations et de vagues de froid sont associées à des frais relativement élevés, les catégories médianes ([200,250] pour les précipitations et 2 pour les vagues de froid) affichent les frais moyens les plus faibles. Ainsi, le lien entre les vagues de froid, les précipitations et les frais mensuels de santé pourrait être modulé par d'autres facteurs.

Les variables liées à l'état de santé antérieur de l'assuré contiennent également de nombreuses informations pertinentes. Le nombre de diagnostics médicaux (`nb_diagnostics_icd_01_lag1`) a une influence positive sur la variable cible. Les individus ayant accumulé un plus grand nombre de diagnostics se distinguent par d'importants frais de santé. En effet, des assurés avec des pathologies nombreuses sont plus susceptibles de nécessiter des soins coûteux. Les conclusions sont similaires concernant le nombre de réclamations pour médicaments (`nb_reclamations_medicaments_lag1`) et les coûts mensuels en médicaments (`cout_total_medicaments _lag1`). En effet, les individus ayant effectué un grand nombre de réclamations supportent des frais de santé plus élevés le mois suivant. Ce résultat reflète le lien entre consommation médicamenteuse et pathologies chroniques multiples ou sévères. Aussi, la variable cible semble positivement liée aux frais mensuels de santé le mois précédant l'observation : plus la somme des frais de santé le mois M est importante, plus elle l'est le mois $M+1$. L'ensemble de ces variables sanitaires semble donc être pertinent pour prédire les frais de santé mensuels de chaque assuré.

Enfin, la qualité de l'air, mesurée par la concentration moyenne de particules fines (`moyenne_pm25_lag1`), s'avère également un déterminant important : les niveaux les plus élevés de pollution sont associés à des montants mensuels de facturation supérieurs. Ces conclusions laissent présager d'un impact sanitaire de la pollution atmosphérique sur la demande de soins et les coûts afférents.

8.2.3 Performance du modèle

Ainsi, une régression linéaire a été implémentée pour prédire les frais de santé mensuels par individu (`somme_montant_facture`) en utilisant les 14 prédicteurs sélectionnés dans la section précédente. La variable cible ainsi que les variables relatives à des frais de santé (`cout_total_medicaments_lag1` et `somme_montant_facture_lag1`) ont subi une transformation logarithmique pour diminuer l'asymétrie de leur distribution (voir figures 8.1 et 8.4) et limiter l'influence des valeurs extrêmes. Les résultats de la régression linéaire sont présentés dans le tableau 8.4.

TABLE 8.4 – Résumé du modèle de régression linéaire – coefficients bruts et normalisés

Variable	**Coef.** (Norm.)	**Ecart-type**	**t-stat**	**p-valeur**
Intercept	1,062 (2,013)	0,070	15,099	0,000**
age	0,007 (0,125)	0,001	11,547	0,000**
genre	-0,191 (-0,096)	0,021	-8,988	0,000**
nb_diagnostics_icd_01_lag1	0,159 (0,228)	0,013	12,063	0,000**
nb_reclamations_medicaments_lag1	0,546 (0,491)	0,029	19,088	0,000**
nb_renouvellements_rx_lag1	0,310 (0,187)	0,023	13,451	0,000**
somme_montant_facture_lag1 (log)	0,239 (0,755)	0,009	26,805	0,000**
cout_total_medicaments_lag1 (log)	0,080 (0,143)	0,009	8,555	0,000**
prop_reclamations_reseau_lag1	0,572 (0,202)	0,040	14,333	0,000**
vague_froid_lag1	0,114 (0,082)	0,016	6,987	0,000**
vague_froid_lag2	-0,044 (-0,032)	0,017	-2,607	0,009**
vague_chaleur_lag1	0,073 (0,076)	0,012	5,938	0,000**
precipitation_totale_lag1	0,0002 (0,010)	0,000	0,875	0,382
moyenne_pm25_lag1	0,002 (0,008)	0,003	0,623	0,533
moyenne_no2_lag1	-0,030 (-0,068)	0,005	-6,150	0,000**

** p-valeur inférieure à 0,05

Le coefficient de détermination R^2 ainsi que le R^2_α sont égaux à 0,314. Ainsi, les prédicteurs expliquent un peu moins d'un tiers de la variance de la variable cible. Bien que la majorité des coefficients soient statistiquement significatifs, le test de Shapiro-Wilk [1] ne valide pas l'hypothèse de normalité des résidus (statistique de Shapiro-Wilk égale à 0,81<1). La normalité des résidus est une hypothèse clé pour la validité des tests de significativité dans la régression linéaire (intervalles de confiance et p-valeurs). Ainsi, même si les prédictions moyennes restent valides, les tests statistiques et donc, les p-valeurs ne le sont plus. Ainsi, il est dangereux de se lancer dans l'interprétation des impacts sur le score mensuel obtenu puisque les p-valeurs ne sont pas robustes.

8.2.4 Conclusions et limites

Malgré une sélection minutieuse des prédicteurs et la transformation logarithmique appliquée à la variable cible pour réduire l'asymétrie observée dans sa distribution, la régression linéaire ne permet pas de fournir un score de santé mensuel individuel pertinent pour cette étude : le score ainsi développé est ni performant, ni interprétable. En parallèle de la régression linéaire, le modèle XGBoost a été optimisé sur la variable cible ainsi que sur la variable cible transformée (logarithme). Les résultats n'ont également pas été concluants en termes de prédiction, les modèles ne parvenant pas à intégrer convenablement l'asymétrie de la variable cible. Ces résultats sont disponibles en annexe E.

Ainsi, pour pallier les difficultés liées à la distribution de la variable cible, les montants de sinistres mensuels ont été discrétisés en quatre classes ordinales définies selon des seuils permettant d'obtenir une répartition des effectifs proche de celle attendue sous une loi de Poisson. Ce choix repose sur l'idée que la distribution des sinistres pourrait être mieux

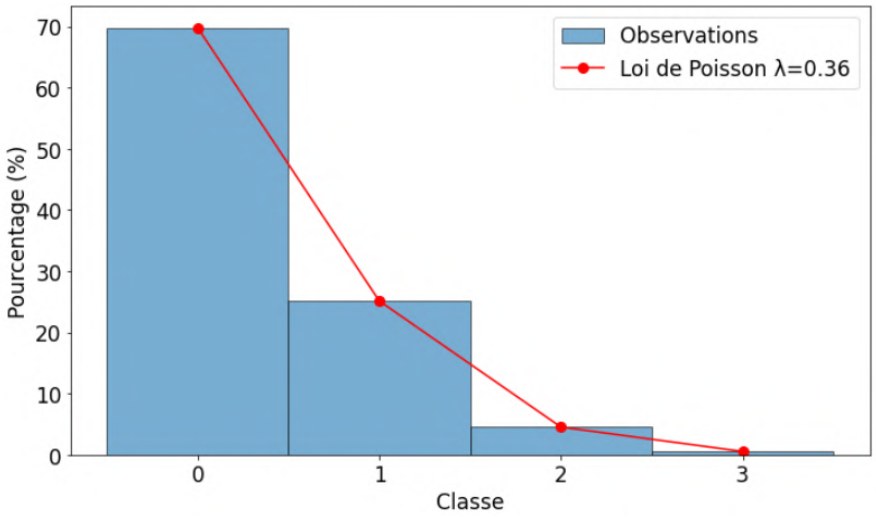
1. Le test de Shapiro-Wilk permet d'évaluer l'hypothèse de normalité d'un échantillon. Une statistique proche de 1 indique une distribution normale, tandis qu'une valeur plus faible suggère un écart à la normalité.

capturée par des modèles de comptage. Ainsi, deux types de modèles ont été implémentés : un modèle linéaire généralisé (GLM) avec une distribution de Poisson similaire en tout point au modèle implémenté dans la section 7.2 (voir section 4.2 pour la description du modèle), et des modèles de *Gradient Boosting*, plus précisément un XGBoost (voir chapitre 5 pour la description du modèle) et un LightGBM. L'implémentation de ces modèles et les résultats sont présentés dans la section suivante.

8.3 Classification des frais de santé

Pour discrétiser la variable cible en accordant un point d'attention à obtenir une répartition des effectifs proche de celle attendue sous une loi de Poisson, les bornes de chaque classe ont été choisies en fonction de l'effectif de la classe nulle, inaltérable. Or, 69,9 % de l'effectif est concentré dans cette classe. Cela peut correspondre à une loi de Poisson de paramètre 0,36.

FIGURE 8.5 – Discrétisation de la variable `somme_montant_facture`



Note de lecture : la classe 0 contient 69,8 % des effectifs de l'ensemble d'entraînement.

La figure 8.5 présente le résultat de la discrétisation obtenue en suivant les effectifs d'une telle loi de Poisson :

— la **classe 0** regroupe l'ensemble des valeurs nulles de la variable cible et représente 69,8 % de l'effectif total ;

— la **classe 1** rassemble les frais mensuels de 0 $ exclus à 3 350 $ inclus, représentant 25,1 % de l'effectif total ;

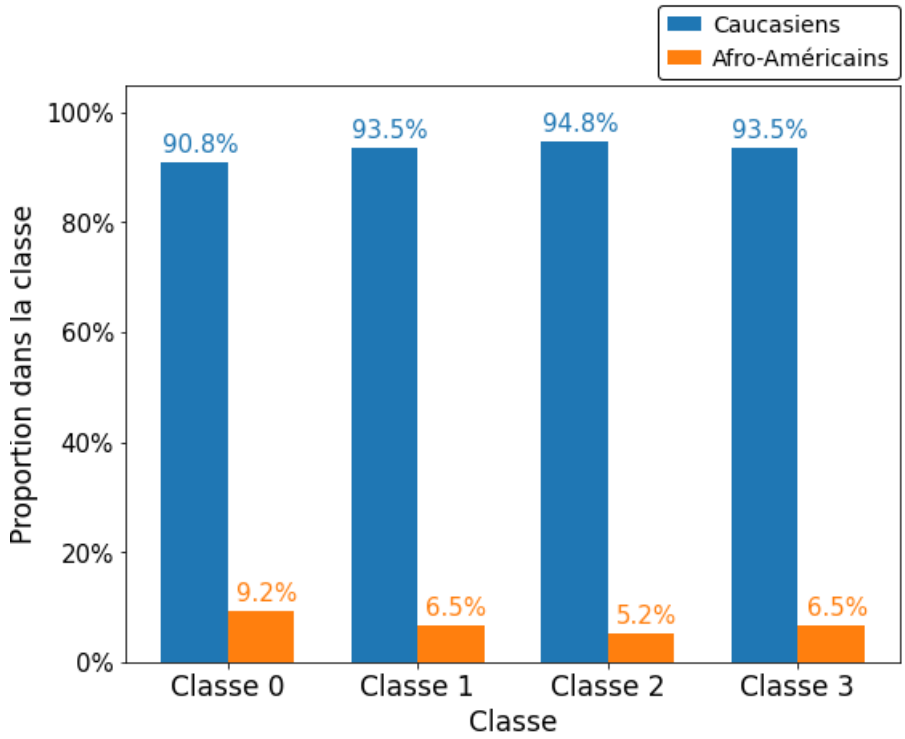
— la **classe 2** recense les frais mensuels de 3 350 $ exclus à 30 000 $ inclus (4,5 % de l'effectif total) ; et

— la **classe 3** est constituée des frais supérieurs à 30 000 $, regroupant les 0,6 % restants.

La classe 2 présente une amplitude importante, due à une distribution asymétrique possédant une queue fine sur la droite. Cette étendue importante vient de la similarité avec les effectifs d'une loi de Poisson, avec une majorité d'observations concentrées dans

les premières classes. D'autres classifications ont été testées, mais seule la présente discrétisation a donné les meilleurs résultats en termes de performances prédictives pour les deux modèles implémentés dans les sections suivantes. De plus, cette classification assure une répartition équilibrée selon l'ethnicité, comme l'illustre la figure 8.6. Pour rappel, dans le Kentucky, on dénombre environ 8 % d'Afro-Américains. Dans notre échantillon représentatif, il y a donc 8 % d'assurés afro-américains. On remarque, dans la figure 8.6, que les différentes classes de frais mensuels 0, 1, 2 et 3 regroupent respectivement 9,2 %, 6,5 %, 5,2 % et 6,5 % d'assurés afro-américains.

FIGURE 8.6 – Proportion d'assurés caucasiens et afro-américains par classe dans l'échantillon d'entraînement.



Note de lecture : la classe 0 contient 90,8 % d'assurés caucasiens et 9,2 % d'assurés afro-américains.

8.3.1 GLM Poisson

Un GLM Poisson a donc été implémenté pour prédire les classes de frais de santé qui sont traitées comme un score croissant reflétant le niveau moyen des frais mensuels individuels. Lorsque la classe augmente, le coût moyen augmente également, ce qui justifie l'utilisation de ce modèle. Le GLM Poisson inclut les variables démographiques, médicales et environnementales utilisées précédemment dans la régression linéaire. En théorie, les coefficients obtenus s'interprètent comme un effet multiplicatif sur la variable cible. Mais ici, la variable représente une classe et non un compte d'évènements au sens strict. Ainsi, l'interprétation des effets des prédicteurs est peu intuitive dans le cas de notre étude, car l'effet multiplicatif sur une variable de comptage n'est pas transposable en termes concrets dans le contexte de classes de coûts. C'est pour cela que l'analyse de ce modèle portera uniquement sur le signe des coefficients : un coefficient positif (resp. négatif) et significatif indique qu'une hausse du prédicteur est associée à une probabilité plus élevée

(resp. faible) d'appartenir à une classe supérieure.

8.3.1.1 Résultats du GLM

Les coefficients obtenus ainsi que leurs caractéristiques (écart-type, t-statistique et p-valeur) sont résumés dans le tableau 8.5. La plupart sont statistiquement significatifs, exceptés les coefficients associés aux variables `moyenne_pm25_lag1` et `precipitation_totale_lag1`. Leur impact sur la variable cible ne sera donc pas commenté. Le R^2_{CS} associé au GLM implémenté est égal à 0,205. Pour un modèle de classes ordonnées, cette valeur est raisonnable, même si elle peut être améliorée, en ajoutant des informations à l'aide de bases supplémentaires par exemple (IMC, consommation de tabac ou informations géographiques plus fines).

TABLE 8.5 – Résumé du modèle de régression linéaire – coefficients bruts

Variable	**Coef.**	**Ecart-type**	**t-stat**	**p-valeur**
Intercept	-1,6551	0,046	-36,050	0,000**
age	0,0033	0,000	8,714	0,000**
genre	-0,0954	0,014	-6,989	0,000**
nb_diagnostics_icd_01_lag1 (log)	0,1346	0,025	5,460	0,000**
nb_reclamations_medicaments_lag1	0,0032	0,001	4,322	0,000**
nb_renouvellements_rx_lag1	0,0036	0,001	3,747	0,000**
somme_montant_facture_lag1 (log)	0,1580	0,005	34,110	0,000**
cout_total_medicaments_lag1 (log)	0,0274	0,004	7,645	0,000**
prop_reclamations_reseau_lag1	0,1794	0,018	10,146	0,000**
vague_froid_lag1	0,0569	0,010	5,831	0,000**
vague_froid_lag2	-0,0175	0,010	-1,677	0,094*
moyenne_no2_lag1	-0,0158	0,003	-4,943	0,000**
vague_chaleur_lag1	0,0407	0,008	5,391	0,000**
precipitation_totale_lag1	0,0002	0,000	1,448	0,148
moyenne_pm25_lag1	-0,0002	0,002	-0,090	0,928

* p-valeur inférieure à 0,01
** p-valeur inférieure à 0,05

La moyenne et l'écart-type des résidus de Pearson associés au modèle GLM Poisson implémenté sont respectivement égaux à -0,02 et à 0,94. Ces deux indicateurs constituent une condition de validité et de qualité du modèle GLM, indiquant que les résidus sont bien centrés et correctement standardisés. Ainsi, la moyenne des résidus doit être égale à 0 et leur écart-type doit être proche de 1. Ces valeurs sont assez proches des seuils théoriques souhaités. Ainsi, les prédictions du modèle ne présentent pas de biais majeur. Aussi, la distribution centrée-réduite des résidus appuie la validité des inférences réalisées à partir des coefficients estimés et permet de valider leur interprétation.

Les métriques de performance du GLM implémenté sont présentées dans le tableau 8.6. Elles sont comparées à celles du même modèle, mais dans lequel les prédictions réalisées sur l'ensemble de test ont été arrondies à l'unité afin de correspondre aux classes discrètes de frais mensuels. Sans arrondi, la MAE vaut 0,37, la MSE est de 0,26 et la RMSE atteint 0,51. Lorsque les prédictions sont arrondies à l'unité, la MAE est nettement améliorée

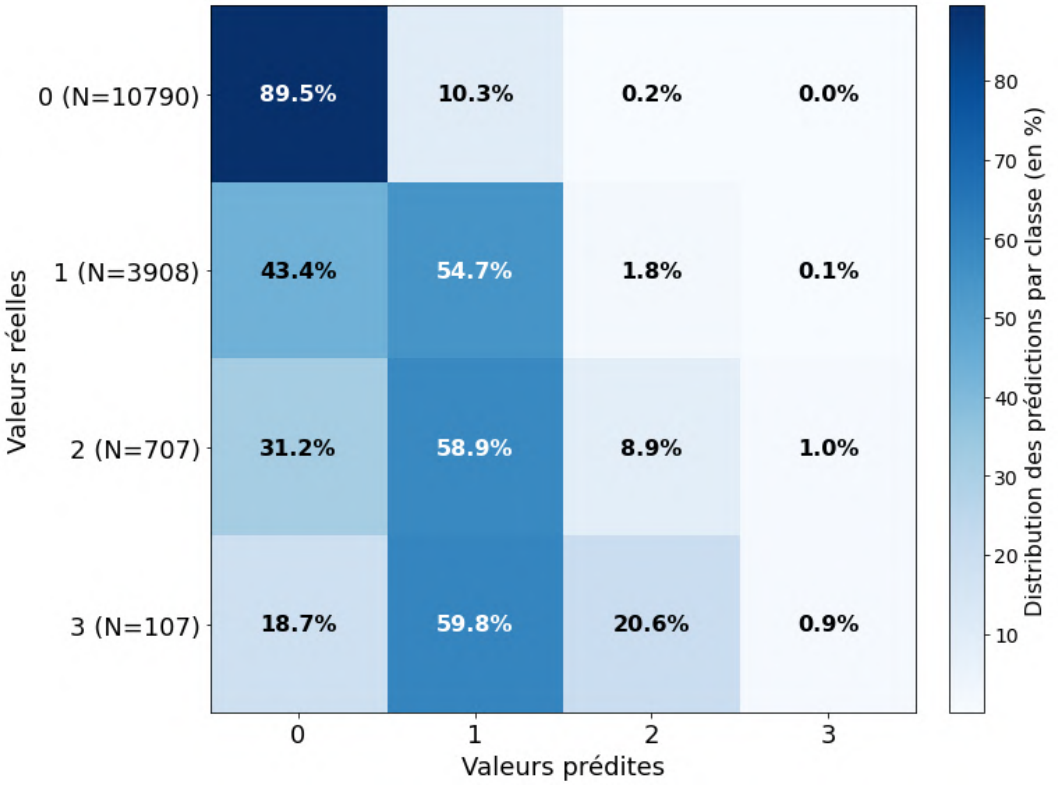
puisqu'elle chute à 0,26. Cette amélioration traduit une meilleure adéquation des valeurs prédites avec les classes réelles. Cependant, la MSE et la RMSE se dégradent légèrement en raison des plus grands écarts créés par l'effet de l'arrondi (mis au carré).

TABLE 8.6 – Comparaison, sur la base de test, du GLM avec ou sans arrondi des prédictions

Modèle GLM Poisson	**MAE**	**MSE**	**RMSE**
Prédictions brutes	0,37	0,26	0,51
Prédictions arrondies	0,26	0,31	0,56

La matrice de confusion suivante montre la distribution en pourcentage des prédictions obtenues pour chaque classe de frais mensuels dans l'ensemble de test (voir figure 8.7). Le nombre d'observations de chaque classe réelle dans la base de test est indiqué entre parenthèses sur l'axe vertical. La diagonale indique, pour chaque classe réelle, le taux de bonne prédiction. L'expression des résultats en pourcentage permet de mieux distinguer la performance relative du modèle pour chaque classe.

FIGURE 8.7 – Matrice de confusion (en proportion) pour le GLM Poisson



Note de lecture : pour la classe 0 qui contient 10 790 observations dans la base de test, 89,5 % des valeurs sont correctement prédites (cases diagonales), tandis que les autres sont réparties sur les classes 1 à 3 (hors diagonale).

La classe 0, qui contient le plus grand nombre d'observations, est prédite correctement dans 89,5 % des cas. Ainsi, le modèle est efficace pour détecter cette classe dans les données. Pour les classes davantage minoritaires, la proportion de bonnes prédictions diminue drastiquement. Pour la classe 1, 54,7 % des observations sont bien prédites,

pour la classe 2, cette valeur chute à 8,9 % tandis que le modèle ne parvient absolument pas à détecter la dernière classe (0,9 % de bonnes prédictions pour la classe 3). Leur sous-représentation dans la base (de test et d'entraînement) implique une performance moindre à prédire ces classes.

8.3.1.2 Interprétation des résultats

Similairement aux conclusions du chapitre 7, le genre est significativement associé à une augmentation de la classe de frais de santé mensuels et être un homme réduirait les coûts associés aux soins de santé par rapport au fait d'être une femme.

L'ensemble des variables médicales est statistiquement significatif. Le nombre de diagnostics (0,1346), le nombre de réclamations relatives aux médicaments (0,0032), le nombre de renouvellements d'ordonnances (0,0036), la somme des frais le mois précédant la période d'observation (0,1580) ainsi que les coûts totaux des médicaments (0,0274) sont tous positivement associés à la variable cible. La somme des frais totaux est une variable clé dans la prédiction, car la t-statistique associée est grande comparativement à celle des autres prédicteurs. Le taux de réclamations dans le réseau de prises en charge de l'assureur est également positivement significatif. L'effet observé est contre-intuitif, car on s'attend à une baisse des coûts si l'assuré choisit des prestataires de soins inclus dans son offre (voir section 1.1).

Concernant les variables climatiques, le bilan est mitigé : si les variables `vague_chaleur_lag1` et `vague_froid_lag1` semblent avoir un effet positif sur les classes de frais de santé prédites, l'inverse est observé pour `moyenne_no2_lag1` et `vague_froid_lag2`. Cet effet contre-intuitif peut être dû à des interactions non-modélisées ou des variables omises. Pour les vagues de froid, il est également possible que leur influence soit plus marquée un mois avant la période étudiée (lag1), plutôt que deux mois en amont (lag2).

8.3.1.3 Conclusion et limites

La classification de la variable cible a permis d'obtenir un modèle valide et interprétable. Cependant, le modèle linéaire généralisé implémenté peine, dans l'ensemble, à détecter les classes relatives aux frais de santé élevés. Il ne respecte pas les contraintes de qualité prédictive, car le score individuel mensuel créé ne reflète pas réellement la sinistralité subie par les assurés de la base. Dans ce contexte, et en lien avec le chapitre précédent, il apparaît nécessaire d'utiliser les avantages des modèles de *machine learning* pour prédire les classes de frais. Ils sont plus flexibles, capables de capturer des relations complexes et non-linéaires entre les variables. Des modèles de *Gradient Boosting* comme XGBoost et LightGBM (*Light Gradient Boosting Machine*) seront ainsi implémentés afin d'améliorer la précision et le calibrage du score de santé mensuel en captant davantage les classes minoritaires, tout en offrant des outils pour l'explication locale et globale des prédictions (valeurs SHAP). Enfin, alors que les biais n'ont pas été analysés précédemment en raison du mauvais calibrage du GLM, une attention particulière sera portée à l'équité du score entre les différents groupes ethnico-raciaux lors de l'évaluation des modèles de *machine learning*.

8.3.2 XGBoost et LightGBM

Dans cette sous-section sont décrits les deux modèles de *Gradient Boosting* implémentés pour obtenir un score de santé mensuel pertinent basé sur les coûts réels des soins de santé de chaque assuré : le XGBoost et le LightGBM, appelé LGBM par la suite. Comme expliqué précédemment dans ce mémoire, les modèles de *machine learning* sont capables de gérer les multicolinéarités, Ainsi, le XGBoost et le LGBM ont été implémentés en utilisant les variables utilisées pour le GLM Poisson auxquelles ont été rajoutées les variables `moyenne_o3_lag1` et `moy_mensuelle_tavg_lag1`. Tout comme le modèle de *machine learning* développé dans la section 7.4, le XGBoost ainsi que le LGBM implémentés ici sont configurés en tant que *classifier* : ils sont donc adaptés aux tâches de classification et prédisent une catégorie spécifique pour chaque ensemble de prédicteurs fourni. Ce choix méthodologique est bien adapté à la nature de la variable cible étudiée dans cette section.

Dans un premier temps, il est nécessaire d'ajuster finement ces modèles : cela est possible grâce à leurs nombreux hyperparamètres. Une première méthode d'optimisation possible est celle utilisée dans la sous-section 7.4.1, consistant à explorer systématiquement différentes valeurs possibles de chaque paramètre afin d'illustrer les possibilités offertes. Cependant, cette approche pas-à-pas s'est révélée rapidement trop chronophage (3 heures d'exécution) et coûteuse en ressources, car elle exige de tester l'ensemble des combinaisons. Pour pallier cette limite, nous avons eu recours à la méthode `RandomizedSearchCV` implémentée dans le module `scikit-learn` de Python, qui permet de sélectionner aléatoirement un sous-ensemble de combinaisons à partir d'une grille prédéfinie et d'évaluer leur performance par validation croisée. Cette stratégie, plus efficace, retourne la meilleure combinaison d'hyperparamètres et le score associé, tout en réduisant significativement le temps de calcul (quelques minutes). Les paramètres obtenus pour le XGBoost à l'issue de cette optimisation sont résumés dans le tableau 8.7, et leur rôle est détaillé dans la sous-section 7.4.1.

TABLE 8.7 – Valeurs des hyperparamètres retenus pour le modèle XGBoost

Paramètre	**Valeur choisie**
`eta`	0,05
`max_depth`	6
`min_child_weight`	5
`colsample_bytree`	0,8
`subsample`	1
`gamma`	0
`nrounds`	300

Le même processus a été mis en place pour l'implémentation du modèle LGBM. Ce modèle de *Gradient Boosting* est très similaire au modèle XGBoost. Cependant, il existe quelques différences notables qui incitent à tester les deux modèles pour le développement du score de santé mensuel :

— XGBoost s'appuie sur une méthode de croissance des arbres niveau par niveau tandis que LGBM privilégie une croissance feuille par feuille, plus rapide et plus efficace ;

— LGBM présente un risque de surapprentissage si les paramètres ne sont pas bien

choisis, contrairement à XGBoost qui supprime le surapprentissage via la croissance des arbres niveau par niveau ;

— LGBM a une mémoire plus importante et est capable de traiter de grands volumes de données.

Les paramètres obtenus pour le LGBM à l'issue de cette optimisation sont résumés dans le tableau 8.8.

TABLE 8.8 – Valeurs des hyperparamètres retenus pour le modèle LGBM

Paramètre	**Description**	**Valeur**
`num_leaves`	Nombre maximal de feuilles par arbre	63
`max_depth`	Profondeur maximale de l'arbre	10
`min_child_samples`	Nombre min. d'échantillons par feuille	10
`colsample_bytree`	Fraction de colonnes par arbre	0,8
`bagging_freq`	Fréquence du bagging	1
`feature_fraction`	Fraction de variables par itération	1
`learning_rate`	Taux d'apprentissage	0,01
`nrounds`	Nombre d'itérations	200

8.3.2.1 Résultats des modèles

Après une sélection rigoureuse des hyperparamètres, les modèles XGBoost et LGBM ont été implémentés dans le but de prédire individuellement les frais mensuels à l'aide de l'ensemble des variables prédictives de la base de données. Les résultats de performance des modèles sur les ensembles d'entraînement et de test sont présentés dans le tableau 8.9 et comparés avec le GLM Poisson implémenté précédemment.

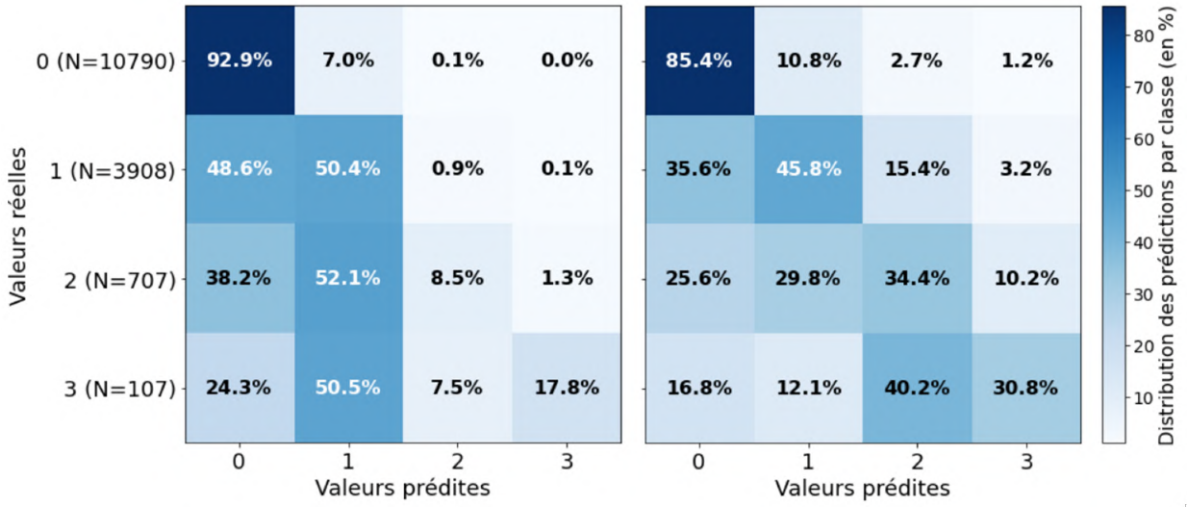
TABLE 8.9 – Comparaison, sur la base de test, des performances des modèles GLM, XGBoost et LGBM

	MAE	**MSE**	**RMSE**
GLM - Prédictions brutes	0,37	0,26	0,51
GLM - Prédictions arrondies	0,26	0,31	0,56
XGBoost - Ensemble de test	0,25	0,30	0,55
XGBoost - Ensemble d'entraînement	0,23	0,25	0,50
LGBM- Ensemble de test	0,33	0,45	0,68
LGBM - Ensemble d'entraînement	0,32	0,45	0,67

Les valeurs des métriques de performance (MAE, MSE et RMSE) sont assez proches entre l'entraînement et le test pour les deux modèles de *machine learning*. Cela signifie notamment qu'il n'y a pas de surapprentissage. Aussi, ces métriques sont assez bonnes pour le modèle XGBoost et sont similaires à celles obtenues avec le GLM Poisson développé dans la section précédente et dont les prédictions ont été arrondies à l'unité. Cependant, le GLM Poisson ne parvient pas à prédire convenablement les classes de frais les plus minoritaires. La matrice de confusion présentée dans la figure 8.8 permet de comparer les performances de prédiction entre le XGBoost optimisé et le GLM Poisson (voir figure 8.7 pour la matrice de confusion du GLM Poisson). Tout comme le GLM Poisson, le XGBoost ne parvient pas à prédire les classes 2 et 3 : les taux de bonnes prédictions

pour ces classes sont respectivement de 8,5 % et de 17,8 %. Même si le LGBM présente des MSE et RMSE plus élevées que les deux premiers modèles, sa matrice de confusion est plus satisfaisante dans le sens où les taux de bonnes prédictions pour les classes 2 et 3 sont beaucoup plus élevés (resp. 34,4 % et 30,8 %). Aussi, on remarque que, lorsque le LGBM se trompe, la valeur prédite se situe autour de la classe réelle, ce qui était moins le cas avec le GLM Poisson et le XGBoost.

FIGURE 8.8 – Matrice de confusion (en proportion) pour les modèles XGBoost (à gauche) et LGBM (à droite)



Note de lecture : pour la classe 2, qui contient 707 observations dans la base de test, 8,5 % des valeurs sont correctement prédites (cases diagonales) avec le modèle XGBoost, tandis que 34,4 % des valeurs sont correctement prédites pour cette même classe avec le modèle LGBM.

Bien que les modèles GLM Poisson et XGBoost offrent des performances satisfaisantes au regard des valeurs des MAE, MSE et RMSE, le LGBM a une capacité supérieure à distinguer les classes minoritaires. Or, dans un contexte de prédiction de frais de santé, l'identification précise des valeurs ou des cas extrêmes a un impact important sur la gestion des risques. Aussi, les prédictions sont davantage centrées autour de la valeur de la classe réelle traduisant une meilleure robustesse du modèle. C'est pour cela que le choix a été fait de sélectionner le LGBM pour interpréter l'effet de chaque variable explicative sur le score de santé individuel ainsi créé.

8.3.2.2 Importance des variables

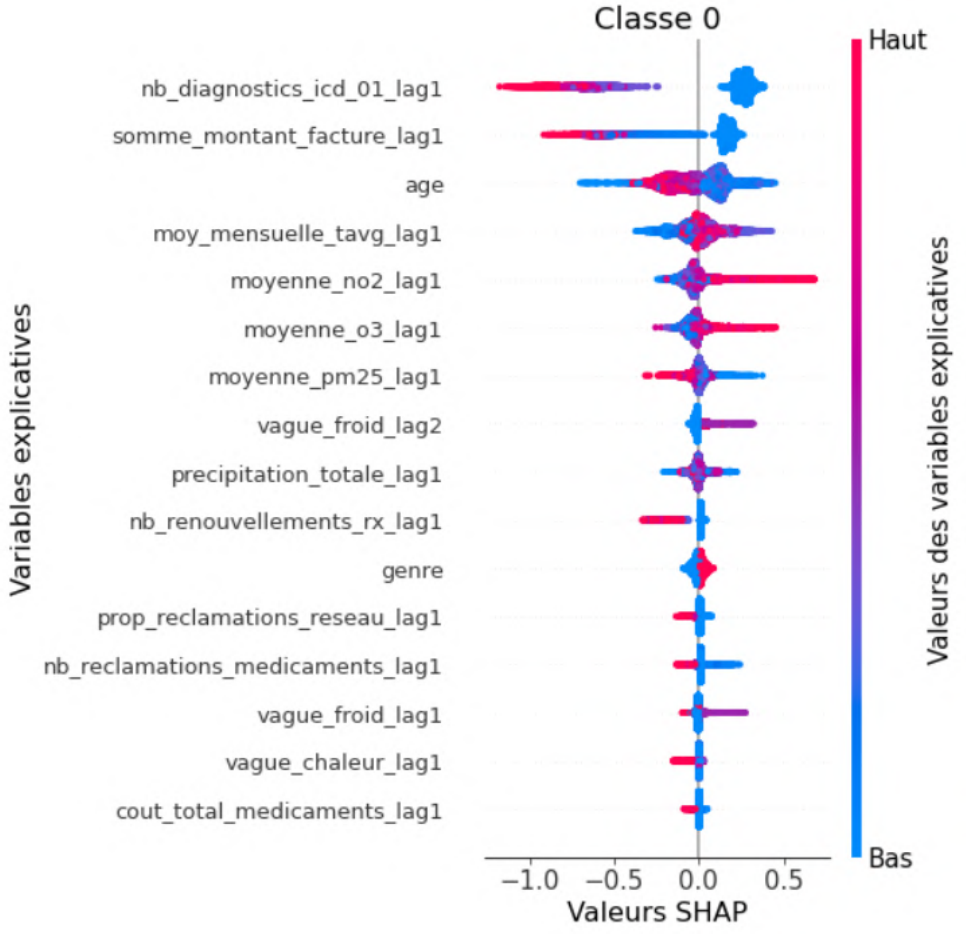
Comme mentionné dans la section 5.4.3, l'utilisation des valeurs SHAP est particulièrement intéressante pour interpréter les sorties d'un modèle de *machine learning*. En particulier, elles permettent d'attribuer à chaque variable une importance dans la prédiction de la variable cible. De manière analogue à ce qui a été fait dans la sous-section 7.4.4, nous allons analyser l'influence de chaque variable explicative sur les prédictions du LGBM.

Les figures 8.9, 8.10, 8.11 et 8.12 mettent en évidence l'impact de chaque variable explicative sur la prédiction individuelle de chaque classe. Ainsi, une valeur SHAP est attribuée pour chaque variable, chaque observation et pour chaque classe. Cette valeur correspond à l'influence de ladite variable sur la probabilité d'appartenir à la classe en

question. Plus cette valeur est grande et positive, plus elle a un effet important sur la probabilité d'appartenir à la classe étudiée. Les variables sont ordonnées de haut en bas selon leur valeur SHAP. Le classement des variables par importance dans la prédiction est effectué pour chaque classe. Ainsi, sur les quatre figures, on distingue que les quatre premières variables en termes d'importance sont, dans l'ordre :

- pour la classe 0 : `nb_diagnostics_icd_01_lag1`, `somme_montant_facture_lag1`, `age` et `moy_mensuelle_tavg_lag1` ;
- pour la classe 1 : `age`, `nb_reclamations_medicaments_lag1`, `somme_montant_facture_lag1` et `genre` ;
- pour la classe 2 : `somme_montant_facture_lag1`, `age`, `nb_reclamations_medicaments_lag1` et `genre` ; et
- pour la classe 3 : `age`, `somme_montant_facture_lag1`, `precipitation_totale_lag1`, et `moyenne_no2_lag1`.

FIGURE 8.9 – Analyse des valeurs SHAP pour la classe 0



Note de lecture : les valeurs SHAP associées à l'âge pour la prédiction de la classe 0 sont comprises entre -0,75 et +0,5.

La figure 8.9 indique les effets de chaque variable sur la prédiction de la classe 0 correspondant à des frais mensuels nuls. La couleur des points (du bleu au rouge) illustre l'effet des valeurs faibles ou élevées de chaque variable sur la prédiction de la classe 0.

Les variables les plus influentes, présentées en haut du graphique, sont l'âge, le nombre de diagnostics et le montant des frais de santé mensuels le mois précédant la période d'observation. Des valeurs élevées pour les deux variables médicales sont associées à des valeurs SHAP négatives. Ainsi, elles sont associées à des probabilités plus faibles d'être classées dans la dernière classe 0 des frais mensuels. Il en est de même pour `prop_reclamations_reseau_lag1` et `nb_reclamations_medicaments_lag1`, mais dans une moindre mesure. Pour l'âge, l'interprétation se fait en trois temps :

— premièrement, certains âges faibles (bébés ou enfants) sont associés à des valeurs SHAP négatives, ce qui traduit une probabilité plus faible d'avoir des coûts liés aux soins de santé nuls ;

— ensuite, les âges les plus élevés sont également associés à des valeurs SHAP négatives, car l'âge est négativement lié avec l'état de santé ; et,

— enfin, certains âges faibles sont liés à des probabilités plus élevées de faire face à des frais de santé nuls.

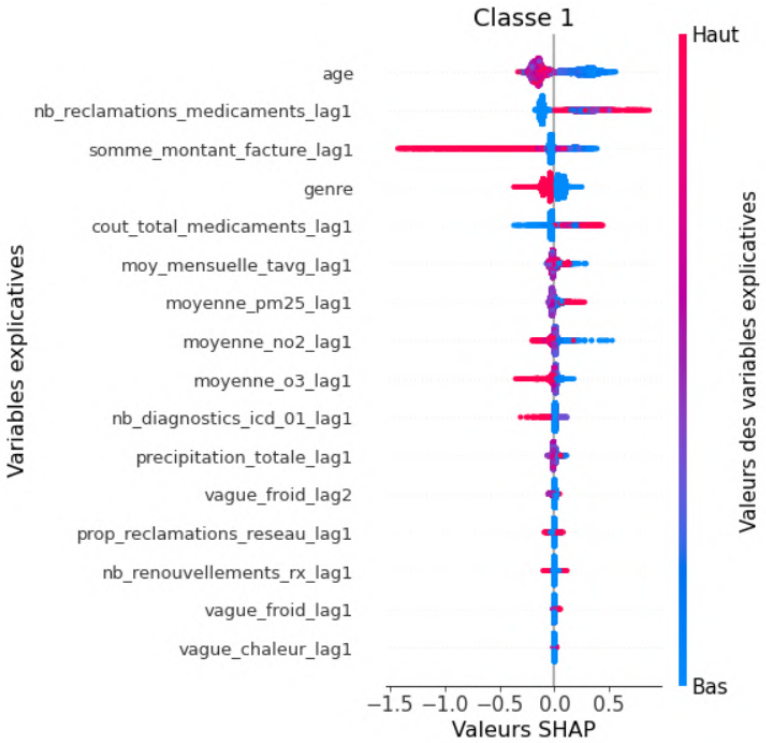
Ainsi, l'effet de l'âge apparaît non-linéaire, avec des comportements contrastés selon les tranches d'âge observées. L'annexe F, qui donne une analyse plus fine de l'importance de la variable `age` par classe de frais de santé, corrobore les conclusions précédentes.

Concernant les variables environnementales, seules les concentrations mensuelles de NO_2, d'O_3 et de $PM_{2.5}$ jouent un rôle important dans la prédiction de la classe 0. Une baisse des concentrations en particules fines semble impliquer une hausse de la probabilité de faire face à des coûts médicaux nuls. Cependant, les effets des deux premiers gaz polluants sont en contradiction avec la revue de littérature effectuée dans le chapitre 3, car une hausse des concentrations associées serait à l'origine d'une hausse de la probabilité de ne subir aucun frais de santé. Les autres variables environnementales sont très peu influentes dans la prédiction de la classe 0.

La figure 8.10 se concentre sur l'influence des variables médicales et environnementales sur la prédiction de la classe 1. Les variables qui ont un impact à la hausse sur la probabilité d'être classé dans cette catégorie de frais de santé sont le nombre de réclamations pour des médicaments, le coût total des médicaments prescrits ainsi que la concentration moyenne de $PM_{2.5}$ dans une moindre mesure. L'âge et le montant des frais le mois précédant la période d'observation ont un impact décroissant sur la probabilité de prédire la classe 1.

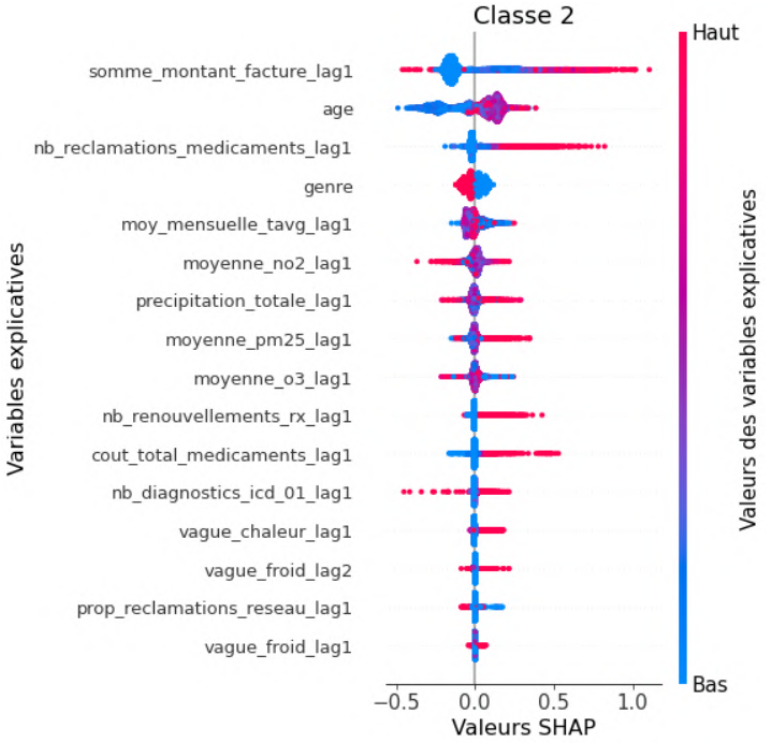
Le contraire est mis en évidence sur la figure 8.11, qui illustre l'importance de chaque variable sur le score de santé mensuel développé. L'âge, `nb_reclamations_medicaments_lag1`, `somme_montant_facture_lag1` et `cout_total_medicaments_lag1`, la concentration en $PM_{2.5}$ ainsi que le nombre de vagues de froid (lag2) et de chaleur (lag1) dans une moindre mesure, exercent une influence positive importante sur la probabilité de prédire la classe 2. Etre un homme (`genre` = 1) conduit à une plus forte probabilité d'être classifié dans les classes 1 ou 2. L'inverse est observé pour la classe de frais nuls, mais sans que les valeurs de SHAP soient assez élevées pour que l'influence du genre soit notable. Des résultats ambigus et non interprétables sont obtenus pour les moyennes mensuelles de températures, les concentrations de NO_2 et d'O_3, les précipitations totales ainsi que le nombre de vagues de froid (lag1).

FIGURE 8.10 – Analyse des valeurs SHAP pour la classe 1



Note de lecture : les valeurs SHAP associées à l'âge pour la prédiction de la classe 1 sont comprises entre -1 et +0,5.

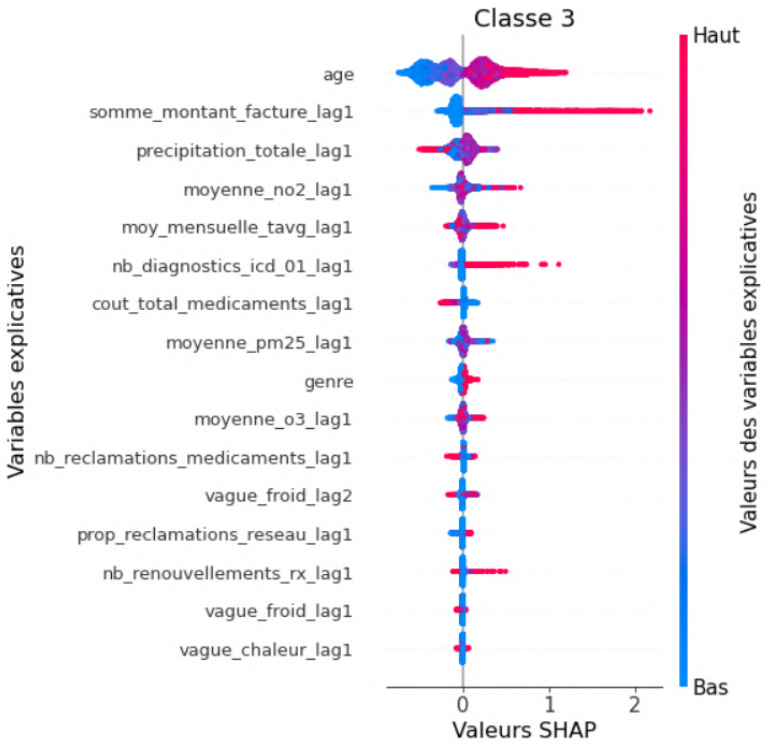
FIGURE 8.11 – Analyse des valeurs SHAP pour la classe 2



Note de lecture : les valeurs SHAP associées à l'âge pour la prédiction de la classe 2 sont comprises entre -0,5 et +0,5.

L'interprétation du graphique SHAP pour la classe 3, qui correspond aux frais liés à des soins de santé les plus élevés, synthétise l'influence de chaque variable sur la probabilité pour un assuré d'appartenir à la classe 3, à travers la distribution des valeurs SHAP selon les valeurs prises par chaque variable (voir figure 8.12).

FIGURE 8.12 – Analyse des valeurs SHAP pour la classe 3



Note de lecture : les valeurs SHAP associées à l'âge pour la prédiction de la classe 3 sont comprises entre -0,2 et +2,1.

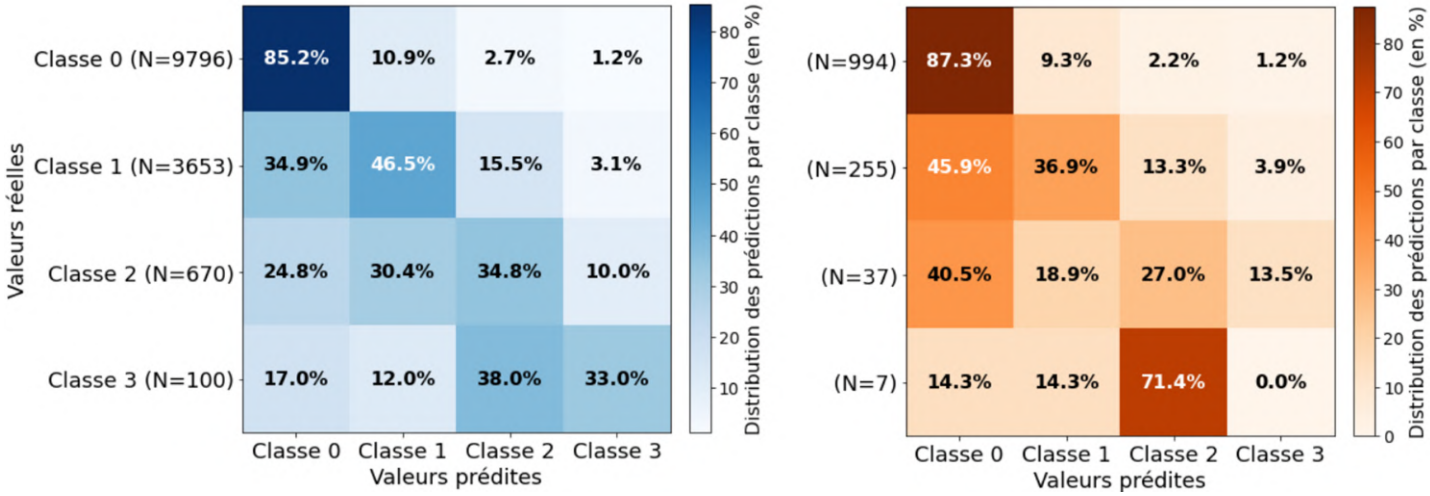
Les variables les plus influentes, présentées en haut du graphique, sont l'âge, le montant des frais de santé mensuels le mois précédant la période d'observation et les précipitations totales. Des valeurs élevées pour la variable `age` sont associées à des valeurs SHAP positives. Ainsi, l'âge est positivement associé à la probabilité d'être classé dans la dernière classe des frais mensuels. Il en est de même pour `somme_montant_facture_lag1`. Ce résultat met en évidence une persistance des dépenses de santé élevées d'un mois à l'autre. `nb_diagnostics_icd_01_lag1` et `nb_renouvellements_rx_lag1` sont des variables de santé ayant un impact positif sur la probabilité d'identifier une observation en classe 3. Plus le nombre de diagnostics et de renouvellements d'ordonnances est élevé, plus les chances de prédire une classe 3 sont grandes. Les autres variables médicales n'ont que très peu d'influence sur la prédiction de la classe 3.

Concernant les variables environnementales, les précipitations, les concentrations mensuelles de NO_2 ainsi que les moyennes mensuelles de température jouent un rôle important dans la prédiction de la classe 3. La première est négativement liée à la probabilité de prédire la classe 3, ce qui est contre-intuitif au vu de la revue de littérature effectuée dans le chapitre 3. Les deux autres variables sont associées à des probabilités positives, qui traduisent une augmentation de la probabilité d'être classé dans la $3^{\text{ème}}$ catégorie de frais de santé mensuels.

8.3.2.3 Analyse des biais ethniques

Pour les problèmes de classification non-binaire, il n'est pas pertinent d'analyser les courbes de calibrage et l'ECE pour étudier les potentiels biais ethniques du modèle développé. Ainsi, il peut être d'usage de comparer les matrices de confusion entre les individus du groupe sensible étudié. La figure 8.13 représente les matrices de confusion obtenues après la prédiction des classes basées sur les observations des assurés caucasiens (à gauche) et des assurés afro-américains (à droite).

FIGURE 8.13 – Matrices de confusion (en proportion) pour les assurés caucasiens (à gauche) et afro-américains (à droite)



Note de lecture : pour la classe 2, il y a 34,8 % de bonnes prédictions pour les assurés caucasiens contre 27 % pour les assurés afro-américains.

La diagonale de ces matrices indique la proportion de prédictions correctes (vrais positifs). Le taux de vrais positifs est similaire dans les deux groupes ethniques pour la classe 0, ce qui indique une certaine équité dans la prédiction de cette classe. Cependant, pour les classes 1, 2 et 3 de frais mensuels de soins de santé, la performance est beaucoup plus faible chez les assurés afro-américains que chez les assurés caucasiens. Par exemple, pour la classe 2, il y a 34,8 % de bonnes prédictions pour les Caucasiens contre 27 % pour les Afro-Américains. Ainsi, pour les classes de frais non-nuls, la prédiction discrimine fortement la population afro-américaine. On observe que les valeurs prédites pour les Afro-Américains sont souvent plus faibles que celles prédites pour les Caucasiens et sont plus fréquemment égales à 0 (classe de frais nuls). Autrement dit, à coûts mensuels identiques, le modèle tend à prédire une classe inférieure pour un assuré afro-américain que pour un assuré caucasien.

Les mêmes conclusions ont été mises en évidence avec le modèle XGBoost rejeté précédemment. La comparaison des matrices de confusion du modèle XGBoost est disponible en annexe G.

Bien qu'il soit très performant, le modèle LGBM présente un biais marqué : il sous-estime les frais de santé des assurés afro-américains par rapport aux assurés caucasiens, particulièrement pour les classes de frais non-nuls. Ce comportement, observé aussi avec le modèle XGBoost, évincé en raison de sa propension à minimiser le risque santé, met

en évidence des biais dans le processus de calcul des scores de santé mensuels basés sur le coût des soins de santé. Ces biais peuvent être liés :

— aux données : même si, dans cette étude, les classes de frais sont équitablement réparties entre Afro-Amércains et Caucasiens, certaines variables, comme le ZIP3 ou le type de couverture, peuvent être corrélées à la variable sensible ;

— aux modèles : le modèle peut sous-performer pour la classe minoritaire (les assurés afro-américains représentent 8 % de l'échantillon) ; et

— aux inégalités sociales : les différences historiques et systémiques dans la fréquence des diagnostics, l'accès aux spécialistes, la prévention ainsi que les revenus, les conditions de logement et l'exposition environnementale influencent indirectement les coûts liés aux soins de santé, mais ne sont pas pris en compte dans le modèle.

8.3.2.4 Conclusions et limites

Les modèles de *machine learning* XGBoost et LightGBM ont été implémentés dans le but d'améliorer la prédiction des classes de frais de santé mensuels par rapport au GLM Poisson, notamment pour mieux détecter les classes minoritaires. Seul le LGBM est parvenu à prédire convenablement les classes les plus élevées, critère essentiel pour la gestion du risque. L'analyse des valeurs SHAP relatives à ce modèle a permis d'identifier les principales variables influentes dans la prédiction des frais mensuels liés aux soins de santé. Ainsi, l'âge, le montant des frais antérieurs ainsi que le nombre de diagnostics et de réclamations médicamenteuses sont les principaux prédicteurs du score de santé individuel mensuel. Les variables environnementales sont moins influentes, car sûrement éclipsées par les variables médicales. Toutefois, les concentrations en NO_2 ainsi que les températures mensuelles semblent jouer un rôle non-négligeable dans la prédiction des classes les plus élevées. Le XGBoost développé pour le score annuel (voir chapitre 7) avait déjà mis en lumière l'impact du NO_2 sur la santé.

Bien que le modèle sélectionné présente de bonnes performances et soit interprétable, il met en évidence des biais ethniques notables : pour des niveaux de frais médicaux mensuels identiques, le modèle attribue des scores plus faibles aux Afro-Américains qu'aux Caucasiens. La présence de ces biais rappelle que, dans le domaine de la santé, il est important de les identifier et d'y remédier en intégrant des méthodes de correction lors du développement des scores de santé.

Conclusion

Les impacts du climat et de la pollution sur la santé ne sont plus à prouver : la revue de littérature effectuée dans ce mémoire a mis en lumière les effets néfastes sur la santé des vagues de froid et de chaleur, des précipitations, des fortes températures et de l'exposition aux polluants tels que l'ozone, le dioxyde d'azote ou encore les particules fines. Ainsi, la prise en compte de ces facteurs environnementaux semble indispensable, que ce soit pour la tarification des assurances santé ou pour la mutualisation des capitaux des sociétés d'assurance américaines. Ce mémoire s'est attaché à élaborer des scores de santé annuels et mensuels à partir de données issues d'assureurs américains, auxquelles ont été ajoutées des informations relatives au climat (vagues de chaleur et de froid, températures, précipitations) et à la pollution atmosphérique (concentrations d'ozone, de dioxyde d'azote et de $PM_{2.5}$), afin d'évaluer leur impact sur la santé des assurés américains. Un point d'attention a été accordé à l'équité des scores développés.

Pour le score annuel, trois modèles ont été testés pour prédire le nombre de conditions chroniques annuel par assuré : un GLM, un GLMM et un modèle de machine learning (XGBoost).

Bien que le modèle GLM ait présenté une capacité explicative satisfaisante, l'introduction d'effets aléatoires ou d'interactions via le GLMM a permis de mieux capturer l'hétérogénéité entre individus. Cependant, les erreurs de calibrage sont restées élevées et certains effets non-linéaires ou interactions complexes entre variables n'ont pu être entièrement capturés par une approche linéaire, même mixte. Aussi, la performance prédictive, bien qu'améliorée, a plafonné avec les modèles linéaires. L'utilisation d'un modèle XGBoost a permis d'atteindre des performances nettement supérieures. Au-delà de la performance brute, le modèle XGBoost a montré une équité satisfaisante en termes de biais ethno-raciaux (analyse des courbes de calibrage).

Les modèles linéaires (GLM et GLMM) sont intrinsèquement interprétables via leurs coefficients. Cependant, certains effets observés des variables environnementales restent contre-intuitifs au regard de la littérature, suggérant des interactions non-modélisées ou de la colinéarité. Dans ces modèles, seuls l'état de santé antérieur, l'âge, le genre, les vagues de chaleur et le dépassement des seuils annuels de NO_2 ont présenté des effets néfastes pour la santé. Avec XGBoost, l'usage des valeurs SHAP a confirmé l'importance de l'état de santé antérieur, de l'âge et du genre, tout en révélant un impact significatif de certains facteurs environnementaux : l'exposition aux vagues de chaleur, aux particules fines ($PM_{2.5}$) et au NO_2.

Pour le score mensuel, quatre modèles ont été testés pour approcher le coût mensuel des services de soins de santé par assuré : une régression linéaire, un GLM Poisson et deux modèles de machine learning (XGBoost et LightGBM).

Malgré une sélection minutieuse des prédicteurs et la transformation logarithmique

appliquée à la variable pour réduire l'asymétrie observée dans sa distribution, la régression linéaire et le XGBoost n'ont pas permis de fournir un score de santé mensuel individuel pertinent pour cette étude : les scores développés étaient ni performants, ni interprétables.

Pour pallier les difficultés liées à la distribution de la variable cible, les montants de sinistres mensuels ont été discrétisés en quatre classes ordinales et un GLM Poisson a tout d'abord été implémenté pour prédire ces classes et ainsi obtenir un score individuel de santé mensuel. La classification de la variable cible a permis d'obtenir un modèle valide qui peine cependant à détecter les classes de frais de santé élevées. Ce modèle a mis en lumière l'effet significatif du genre, de l'âge et des indicateurs médicaux antérieurs sur le score de santé. Concernant les variables environnementales, le bilan est mitigé : seules les variables liées aux vagues de froid et de chaleur semblent avoir un effet néfaste sur la santé.

Pour capturer davantage d'informations et parvenir à prédire les classes de frais les plus élevées, un LightGBM a ensuite été retenu. Parfois plus rapide et plus efficace que le modèle XGBoost il permet d'obtenir un score individuel de santé mensuel pertinent. L'analyse des valeurs SHAP relatives à ce modèle a permis d'identifier les principales variables influentes dans la prédiction des frais mensuels associés aux soins de santé. Ainsi, l'âge, le montant des frais antérieurs ainsi que le nombre de diagnostics et de réclamations médicamenteuses sont les principaux prédicteurs du score de santé individuel mensuel. Les variables environnementales sont moins influentes, car sûrement éclipsées par les variables médicales. Toutefois, les concentrations en NO_2 ainsi que les températures mensuelles semblent jouer un rôle non-négligeable dans la prédiction des classes les plus élevées. Bien que le modèle sélectionné présente de bonnes performances et soit interprétable, il met en évidence des biais ethniques notables : pour des niveaux de frais médicaux mensuels identiques, le modèle attribue des scores plus faibles aux Afro-Américains qu'aux Caucasiens.

Les scores développés dans ce mémoire n'ont pas permis d'identifier d'effets néfastes importants du climat et de la pollution sur la santé des assurés. Plusieurs facteurs peuvent expliquer ces résultats, en atténuer la portée ou même en fausser l'interprétation. Tout d'abord, les contraintes opérationnelles et financières liées au traitement des grands volumes de données utilisées dans le cadre de ce mémoire ont rendu la modélisation plus complexe. En effet, il a été impossible de travailler sur l'intégralité de la base de données et même sur l'ensemble des assurés d'un seul Etat. Il a donc été nécessaire de sélectionner des échantillons afin de pouvoir effectuer les calculs et exécuter les modèles dans des délais raisonnables. Toutefois, cet échantillonnage a pu entraîner une perte d'informations et être à l'origine de certains effets contre-intuitifs ou sous-estimés. Aussi, seuls les ZIP3 sont renseignés dans la base de données pour localiser les assurés. Ces territoires géographiques, assez étendus, contiennent à la fois des zones urbaines et des zones rurales. Cette maille géographique peut annihiler les effets de la pollution ou même de certaines variables climatiques sur la santé de la population assurée. L'importance de l'âge et de l'état de santé antérieur des assurés dans la modélisation des scores de santé annuels et mensuels a pu également contribuer à l'atténuation des effets du climat et de la pollution sur la santé. Ces variables restent néanmoins importantes pour respecter le critère de pertinence des scores créés. Enfin, l'agrégation aux mailles temporelles annuelle et mensuelle peut écraser les effets potentiels des variables environnementales sur le score de santé créé. Par exemple, une vague de chaleur survenue en 2016 n'aura que peu d'impact sur la santé des assurés en 2017. Il en est de même pour les vagues de froid ou les pics de pollution.

Ce mémoire a mis en exergue la difficulté de capter de manière fine les effets du climat et de la pollution sur la santé aux Etats-Unis à partir de données d'assurance. Les résultats invitent à poursuivre les recherches en mobilisant des bases de données plus exhaustives, une maille géographique plus précise et des méthodes capables de mieux intégrer les interactions complexes entre variables environnementales et médicales. Cependant, ces contraintes impliquent nécessairement des coûts non-négligeables. Au-delà de l'intérêt scientifique, l'approfondissement de ces travaux pourrait également aider les assureurs et les systèmes de santé à anticiper l'impact du changement climatique et de la pollution en ajustant la tarification des polices, les mécanismes de mutualisation ainsi que les stratégies de prévention.

Bibliographie

[1] The ACA Times, *IRS Increases Safe Harbor Affordability Threshold for 2025 Tax Year*.https://acatimes.com/irs-increases-safe-harbor-affordability-threshold-for-2025-tax-year/. Consulté le 24 octobre 2024.

[2] Katherine Keisler-Starkey, Lisa N. Bunch, Rachel A. Lindstrom. « Health Insurance Coverage in the United States : 2022 ». Census.gov. https://www.census.gov/library/publications/2023/demo/p60-281.html. Consulté le 18 novembre 2024.

[3] KFF, « 2024 Employer Health Benefits Survey ». https://www.kff.org/health-costs/report/2024-employer-health-benefits-survey/. Consulté le 9 octobre 2024.

[4] Centers for Medicare & Medicaid Services, *National Health Expenditure Data*. https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet. Consulté le 24 octobre 2024.

[5] CMS, « 2024 Marketplace Open Enrollment Period Public Use Files ». https://www.cms.gov/data-research/statistics-trends-reports/marketplace-products/2024-marketplace-open-enrollment-period-public-use-files. Consulté le 2 décembre 2024.

[6] OCDE (2023), Panorama de la santé 2023 : Les indicateurs de l'OCDE, Éditions OCDE, Paris. https://doi.org/10.1787/5108d4c7-fr.

[7] Europ Assistance, « Voyage aux Etats-Unis : comment fonctionnent les frais médicaux aux USA ? ». https://www.europ-assistance.fr/fr/conseils/voyager-aux-etats-unis-gare-aux-frais-medicaux. Consulté le 25 novembre 2024.

[8] Munira Z. Gunja, Evan D. Gumas, and Reginald D. Williams II. U.S. Health Care from a Global Perspective, 2022 : Accelerating Spending, Worsening Outcomes (Commonwealth Fund, Jan. 2023). https://doi.org/10.26099/8ejy-yc74.

[9] BEA, *Medical Services by Disease - Categories and Detailed Conditions, 2021*. https://apps.bea.gov/data/special-topics/health-care/viz/diseases/. Consulté le 24 ocotbre 2024.

[10] KFF, *Key Facts about the Uninsured Population*. https://www.kff.org/uninsured/issue-brief/key-facts-about-the-uninsured-population/. Consulté le 24 octobre 2024.

[11] Katherine Keisler-Starkey, Lisa N. Bunch, and Rachel A. Lindstrom *Health Insurance Coverage in the United States : 2022*, US Census Bureau. https://www.kff.org/uninsured/issue-brief/key-facts-about-the-uninsured-population/. Consulté le 24 octobre 2024.

[12] NOAA National Centers for Environmental Information (NCEI), U.S. Billion-Dollar Weather and Climate Disasters (2024), https://www.ncei.noaa.gov/access/billions/. Consulté le 21 décembre 2024.

[13] US Census Bureau, Household Pulse Survey : Measuring Emergent Social and Economic Matters Facing U.S. Households, https://www.census.gov/programs-surveys/household-pulse-survey.html. Consulté le 21 décembre 2024.

[14] Sharpe JD, Wolkin AF. The Epidemiology and Geographic Patterns of Natural Disaster and Extreme Weather Mortality by Race and Ethnicity, United States, 1999-2018. Public Health Rep. 2022 Nov-Dec ;137(6) :1118-1125. https://doi.org/10.1177/00333549211047235.

[15] Jay AK, Crimmins AR, Avery CW, Dahl TA, Dodder RS, Hamlington BD, Lustig A, Marvel K, Méndez-Lazaro PA, Osler MS, Terando A, Weeks ES, Zycherman A. Overview : Understanding risks, impacts, and responses. In : Crimmins AR, Avery CW, Easterling DR, Kunkel KE, Stewart BC, Maycock TK, editors. Fifth National Climate Assessment. Washington, DC, USA : U.S. Global Change Research Program ; 2023. https://doi.org/10.7930/NCA5.2023.CH1

[16] Benavidez GA, Zahnd WE, Hung P, Eberth JM. Chronic Disease Prevalence in the US : Sociodemographic and Geographic Variations by Zip Code Tabulation Area. Prev Chronic Dis. 2024 Feb 29 ;21 :E14. http://dx.doi.org/10.5888/pcd21.230267.

[17] Agency for Healthcare Research and Quality (US), 2023 National Healthcare Quality and Disparities Report : Portrait of American Healthcare, Rockville (MD), 2023 Dec, https://www.ncbi.nlm.nih.gov/books/NBK600454/. Consulté le 22 décembre.

[18] Carter KN, Blakely T, Collings S, Imlach Gunasekara F, Richardson K. What is the association between wealth and mental health ? J Epidemiol Community Health. 2009 Mar ;63(3) :221-6. https://doi:10.1136/jech.2008.079483.

[19] Kyriopoulos I, Machado S, Papanicolas I. Wealth-related inequalities in self-reported health status in the United States and 14 high-income countries. Health Serv Res. 2024 Dec ;59(6) :e14366. https://doi.org/10.1111/1475-6773.14366.

[20] Krittanawong C, Maitra NS, Qadeer YK, Wang Z, Fogg S, Storch EA, Celano CM, Huffman JC, Jha M, Charney DS, Lavie CJ. Association of Depression and Cardiovascular Disease. Am J Med. 2023 Sep ;136(9) :881-895. https://doi.org/10.1016/j.amjmed.2023.04.036.

[21] Tsao TM, Hwang JS, Chen CY, Lin ST, Tsai MJ, Su TC. Urban climate and cardiovascular health : Focused on seasonal variation of urban temperature, relative humidity, and PM2.5 air pollution. Ecotoxicol Environ Saf. 2023 Sep 15 ;263 :115358. https://doi.org/10.1016/j.ecoenv.2023.115358.

[22] Vlachopoulos C, Aznaouridis K, Stefanadis C. Prediction of cardiovascular events and all-cause mortality with arterial stiffness : a systematic review and meta-analysis. J Am Coll Cardiol. 2010 Mar 30 ;55(13) :1318-27. https://doi.org/10.1016/j.jacc.2009.10.061.

[23] Tian Y, Xiang X, Juan J, Sun K, Song J, Cao Y, Hu Y. Fine particulate air pollution and hospital visits for asthma in Beijing, China. Environ Pollut. 2017 Nov ;230 :227-233. https://doi:10.1016/j.envpol.2017.06.029

[24] Kyung SY, Jeong SH. Particulate-Matter Related Respiratory Diseases. Tuberc Respir Dis (Seoul). 2020 Apr ;83(2) :116-121. https://doi:10.4046/trd.2019.0025

[25] Nhung NTT, Amini H, Schindler C, Kutlar Joss M, Dien TM, Probst-Hensch N, Perez L, Künzli N. Short-term association between ambient air pollution and pneumonia in children : A systematic review and meta-analysis of time-series and case-crossover studies. Environ Pollut. 2017 Nov ;230 :1000-1008. https://doi:10.1016/j.envpol.2017.07.063.

[26] Fajersztajn L, Saldiva P, Pereira LAA, Leite VF, Buehler AM. Short-term effects of fine particulate matter pollution on daily health events in Latin America : a systematic review and meta-analysis. Int J Public Health. 2017 Sep ;62(7) :729-738. https://doi:10.1007/s00038-017-0960-y.

[27] Sesé L, Nunes H, Cottin V, Sanyal S, Didier M, Carton Z, Israel-Biet D, Crestani B, Cadranel J, Wallaert B, Tazi A, Maître B, Prévot G, Marchand-Adam S, Guillot-Dudoret S, Nardi A, Dury S, Giraud V, Gondouin A, Juvin K, Borie R, Wislez M, Valeyre D, Annesi-Maesano I. Role of atmospheric pollution on the natural history of idiopathic pulmonary fibrosis. Thorax. 2018 Feb ;73(2) :145-150. https://doi:10.1136/thoraxjnl-2017-209967.

[28] Fu P, Guo X, Cheung FMH, Yung KKL. The association between PM2.5 exposure and neurological disorders : a systematic review and meta-analysis. Science of the Total Environment 2019. 655 :1240-1248. https://doi.org/10.1016/j.scitotenv.2018.11.218.

[29] Nagel G, Stafoggia M, Pedersen M, Andersen ZJ, Galassi C, Munkenast J, Jaensch A, Sommar J, Forsberg B, Olsson D, Oftedal B, Krog NH, Aamodt G, Pyko A, Pershagen G, Korek M, De Faire U, Pedersen NL, Östenson CG, Fratiglioni L, Sørensen M, Tjønneland A, Peeters PH, Bueno-de-Mesquita B, Vermeulen R, Eeftens M, Plusquin M, Key TJ, Concin H, Lang A, Wang M, Tsai MY, Grioni S, Marcon A, Krogh V, Ricceri F, Sacerdote C, Ranzi A, Cesaroni G, Forastiere F, Tamayo-Uria I, Amiano P, Dorronsoro M, de Hoogh K, Beelen R, Vineis P, Brunekreef B, Hoek G, Raaschou-Nielsen O, Weinmayr G. Air pollution and incidence of cancers of the stomach and the upper aerodigestive tract in the European Study of Cohorts for Air Pollution Effects (ESCAPE). Int J Cancer. 2018 Oct 1 ;143(7) :1632-1643. https://doi.org/10.1002/ijc.31564.

[30] Nicolle-Mir, L. (2018). Impact de l'exposition à long terme au dioxyde d'azote et au bruit routier sur le risque d'insuffisance cardiaque. Environnement, Risques Santé, 17(4), 357-358.

[31] Zong Z, Zhang M, Xu K, Zhang Y, Hu C. Association between Short-Term Exposure to Ozone and Heart Rate Variability : A Systematic Review and Meta-Analysis. International Journal of Environmental Research and Public Health. 2022 ; 19(18) :11186. https://doi.org/10.3390/ijerph191811186

[32] Newell K, Kartsonaki C, Lam KBH, Kurmi OP. Cardiorespiratory health effects of particulate ambient air pollution exposure in low-income and middle-income countries : a systematic review and meta-analysis. Lancet Planet Health. 2017 Dec ;1(9) :e368-e380. https://doi:10.1016/S2542-5196(17)30166-3.

[33] Pavanaditya Badida, Arun Krishnamurthy, Jayapriya Jayaprakash, Meta analysis of health effects of ambient air pollution exposure in low- and middle-income countries, Environmental Research, Volume 216, Part 4, 2023, 114604, ISSN 0013-9351, https://doi.org/10.1016/j.envres.2022.114604.

[34] Zhang S, Li G, Tian L, Guo Q, Pan X. Short-term exposure to air pollution and morbidity of COPD and asthma in East Asian area : A systematic review and meta-analysis. Environ Res. 2016 Jul ;148 :15-23. https://doi:10.1016/j.envres.2016.03.008.

[35] Huangfu P, Atkinson R. Long-term exposure to NO2 and O3 and all-cause and respiratory mortality : A systematic review and meta-analysis. Environ Int. 2020 Nov ;144 :105998. https://doi:10.1016/j.envint.2020.105998.

[36] Zhang W, Ruan Y, Ling J. Short-Term Effects of NO2 Exposure on Hospitalization for Chronic Kidney Disease. Toxics. 2024 ; 12(12) :898. https://doi.org/10.3390/toxics12120898

[37] Wu YH, Wu CD, Chung MC, Chen CH, Wu LY, Chung CJ, Hsu HT. Long-Term Exposure to Fine Particulate Matter and the Deterioration of Estimated Glomerular Filtration Rate : A Cohort Study in Patients With Pre-End-Stage Renal Disease. Front Public Health. 2022 Apr 8 ;10 :858655. https://doi:10.3389/fpubh.2022.858655.

[38] Oudin A, Bråbäck L, Åström DO, Strömgren M, Forsberg B. Association between neighbourhood air pollution concentrations and dispensed medication for psychiatric disorders in a large longitudinal cohort of Swedish children and adolescents. BMJ Open. 2016 Jun 3 ;6(6) :e010004. https://doi:10.1136/bmjopen-2015-010004.

[39] Szyszkowicz, M., Tremblay, N. Case-crossover design : Air pollution and health outcomes. IJOMEH 24, 249–255 (2011). https://doi.org/10.2478/s13382-011-0034-y.

[40] United States Environmental Protection Agency (EPA). https://www.epa.gov/. Consulté le 24 mars 2025.

[41] United States Environmental Protection Agency (EPA). Climate Change Indicators : Heat Waves https://www.epa.gov/climate-indicators/climate-change-indicators-heat-waves. Consulté le 19 juin 2025.

[42] Ebi KL, Capon A, Berry P, Broderick C, de Dear R, Havenith G, Honda Y, Kovats RS, Ma W, Malik A, Morris NB, Nybo L, Seneviratne SI, Vanos J, Jay O. Hot weather and heat extremes : health risks. Lancet. 2021 Aug 21 ;398(10301) :698-708. https://doi:10.1016/S0140-6736(21)01208-3.

[43] Xu Z, Yi W, Bach A, Tong S, Ebi KL, Su H, Cheng J, Rutherford S. Multimorbidity and emergency hospitalisations during hot weather. EBioMedicine. 2024 Jun ;104 :105148. https://doi.org/10.1016/j.ebiom.2024.105148.

[44] Corvetto JF, Helou AY, Kriit HK, Federspiel A, Bunker A, Liyanage P, Costa LF, Müller T, Sauerborn R. Private vs. public emergency visits for mental health due to heat : An indirect socioeconomic assessment of heat vulnerability and healthcare access, in Curitiba, Brazil. Sci Total Environ. 2024 Jul 15 ;934 :173312. https://doi.org/10.1016/j.scitotenv.2024.173312.

[45] Niilo R.I. Ryti, Yuming Guo, and Jouni J.K. Jaakkola 2016 Global Association of Cold Spells and Adverse Health Effects : A Systematic Review and Meta-Analysis Environmental Health Perspectives 124 :1 CID : https://doi.org/10.1289/ehp.1408104

[46] Neild PJ, Syndercombe-Court D, Keatinge WR, Donaldson GC, Mattock M, Caunce M. Cold-induced increases in erythrocyte count, plasma cholesterol and plasma fi-

brinogen of elderly people without a comparable rise in protein C or factor X. Clin Sci (Lond). 1994 Jan ;86(1) :43-8. https://doi.org/10.1042/cs0860043.

[47] Keatinge WR, Coleshaw SR, Cotter F, Mattock M, Murphy M, Chelliah R. Increases in platelet and red cell counts, blood viscosity, and arterial pressure during mild surface cooling : factors in mortality from coronary and cerebral thrombosis in winter. Br Med J (Clin Res Ed). 1984 Nov 24 ;289(6456) :1405-8. https://doi.org/10.1136/bmj.289.6456.1405.

[48] Pitsavos C, Panagiotakos DB, Antonoulas A, Zombolos S, Kogias Y, Mantas Y, Stravopodis P, Kourlaba G, Stefanadis C ; Greek study of acute Coronary Syndromes study investigators. Epidemiology of acute coronary syndromes in a Mediterranean country ; aims, design and baseline characteristics of the Greek study of acute coronary syndromes (GREECS). BMC Public Health. 2005 Mar 16 ;5 :23. https://doi.org/10.1186/1471-2458-5-23.

[49] Tiina M. Mäkinen, Raija Juvonen, Jari Jokelainen, Terttu H. Harju, Ari Peitso, Aini Bloigu, Sylvi Silvennoinen-Kassinen, Maija Leinonen, Juhani Hassi, Cold temperature and low humidity are associated with increased occurrence of respiratory tract infections, Respiratory Medicine, Volume 103, Issue 3, 2009, Pages 456-462, ISSN 0954-6111, https://doi.org/10.1016/j.rmed.2008.09.011.

[50] Castellani JW, M Brenner IK, Rhind SG. Cold exposure : human immune responses and intracellular cytokine expression. Med Sci Sports Exerc. 2002 Dec ;34(12) :2013-20. https://doi.org/10.1097/00005768-200212000-00023.

[51] Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. The Eurowinter Group. Lancet. 1997 May https://pubmed.ncbi.nlm.nih.gov/9149695/.

[52] Jaakkola K, Saukkoriipi A, Jokelainen J, Juvonen R, Kauppila J, Vainio O, Ziegler T, Rönkkö E, Jaakkola JJ, Ikäheimo TM ; KIAS-Study Group. Decline in temperature and humidity increases the occurrence of influenza in cold climate. Environ Health. 2014 Mar 28 ;13(1) :22. https://doi.org/10.1186/1476-069X-13-22.

[53] Sundell N, Andersson LM, Brittain-Long R, Lindh M, Westin J. A four-year seasonal survey of the relationship between outdoor climate and epidemiology of viral respiratory tract infections in a temperate climate. J Clin Virol. 2016 Nov ;84 :59-63. https://doi.org/10.1016/j.jcv.2016.10.005.

[54] Lofgren E, Fefferman NH, Naumov YN, Gorski J, Naumova EN. Influenza seasonality : underlying causes and modeling theories. J Virol. 2007 Jun ;81(11) :5429-36. https://doi.org/10.1128/JVI.01680-06.

[55] Levin RB, Epstein PR, Ford TE, Harrington W, Olson E, Reichard EG. U.S. drinking water challenges in the twenty-first century. Environ Health Perspect. 2002 Feb ;110 Suppl 1(Suppl 1) :43-52. https://doi:10.1289/ehp.02110s143.

[56] Cann KF, Thomas DR, Salmon RL, Wyn-Jones AP, Kay D. Extreme water-related weather events and waterborne disease. Epidemiol Infect. 2013 Apr ;141(4) :671-86. https://doi:10.1017/S0950268812001653.

[57] Mac Kenzie WR, Hoxie NJ, Proctor ME, Gradus MS, Blair KA, Peterson DE, Kazmierczak JJ, Addiss DG, Fox KR, Rose JB, et al. A massive outbreak in Milwaukee of cryptosporidium infection transmitted through the public water supply. N Engl J Med. 1994 Jul 21 ;331(3) :161-7. https://doi:10.1056/NEJM199407213310304.

[58] Parmenter RR, Yadav EP, Parmenter CA, Ettestad P, Gage KL. Incidence of plague associated with increased winter-spring precipitation in New Mexico. Am J Trop Med Hyg. 1999 Nov ;61(5) :814-21. https://doi.org/10.4269/ajtmh.1999.61.814.

[59] Nassikas NJ, Rifas-Shiman SL, Luttmann-Gibson H, Chen K, Blossom JC, Oken E, Gold DR, Rice MB. Precipitation and Adolescent Respiratory Health in the Northeast United States. Ann Am Thorac Soc. 2023 May ;20(5) :698-704. https://doi.org/10.1513/AnnalsATS.202209-805OC.

[60] D'Amato G et al (2007b) Allergenic pollen and pollen allergy in Europe. Allergy 62 :976–990. https://doi.org/10.1111/j.1398-9995.2007.01393.x

[61] Grundstein A, Sarnat SE, Klein M, Shepherd M, Naeher L, Mote T, Tolbert P. Thunderstorm associated asthma in Atlanta, Georgia. Thorax. 2008 Jul ;63(7) :659-60. https://doi:10.1136/thx.2007.092882.

[62] United States Environmental Protection Agency (EPA). Climate Change Indicators : U.S. and Global Precipitation https://www.epa.gov/climate-indicators/climate-change-indicators-us-and-global-precipitation. Consulté le 19 juin 2025.

[63] Priftis KN, Paliatsos AG, Panagiotopoulou-Gartagani P, Tapratzi-Potamianou P, Zachariadi-Xypolita A, Nicolaidou P, Saxoni-Papageorgiou P. Association of weather conditions with childhood admissions for wheezy bronchitis or asthma in Athens. Respiration. 2006 ;73(6) :783-90. https://doi:10.1159/000093817.

[64] Mireku N, Wang Y, Ager J, Reddy RC, Baptist AP. Changes in weather and the effects on pediatric asthma exacerbations. Ann Allergy Asthma Immunol. 2009 Sep ;103(3) :220-4. https://doi:10.1016/S1081-1206(10)60185-8.

[65] Erik J. Timmermans, Laura A. Schaap, Florian Herbolsheimer, Elaine M. Dennison, Stefania Maggi, Nancy L. Pedersen, Maria Victoria Castell, Michael D. Denkinger, Mark H. Edwards, Federica Limongi, Mercedes Sánchez-Martínez, Paola Siviero, Rocio Queipo, Richard Peter, Suzan van der Pas and Dorly J.H. Deeg for the EPOSA Research Group The Journal of Rheumatology October 2015, 42 (10) 1885-1892 ; https://doi.org/10.3899/jrheum.141594.

[66] Onozuka D, Hashizume M. Weather variability and paediatric infectious gastroenteritis. Epidemiol Infect. 2011 Sep ;139(9) :1369-78. https://doi.org/10.1017/S0950268810002451.

[67] Lin KJ, Lin PH, Chu SH, Chen HW, Wang TM, Chiang YJ, Liu KL, Wang HH. The impact of climate factors on the prevalence of urolithiasis in Northern Taiwan. Biomed J. 2014 Jan-Feb ;37(1) :24-30. https://doi.org/10.4103/2319-4170.117888.

[68] Sirohi M, Katz BF, Moreira DM, Dinlenc C. Monthly variations in urolithiasis presentations and their association with meteorologic factors in New York City. J Endourol. 2014 May ;28(5) :599-604. https://doi.org/10.1089/end.2013.0680.

[69] F. Elie (2010). Humidité atmosphérique et précipitations

[70] Kutner MH, Nachtsheim CJ et Neter J. Applied Linear Regression Models. McGraw-Hill Irwin. 2004. ://doi:10.2307/1269508.

[71] Sheather S. A modern approach to regression with R. Springer. 2009.

[72] Baradel N, Cours d'actuariat de l'assurance non-vie. Octobre 2024.

[73] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Mathematics, 1984. https://doi.org/10.1201/9781315139470.

[74] Breiman L, Random Forests, Machine Learning 45, 5–32 (2001), https://doi.org/10.1023/A:1010933404324

[75] Tianqi Chen and Carlos Guestrin. 2016. XGBoost : A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[76] Shapley L.S., Kuhn H.W., Tucker A.W. (Eds.), A value for n-person games, Contributions to the Theory of Games (AM-28), vol. II, Princeton University Press, Princeton (1953), pp. 307-318, https://doi.org/10.1515/9781400881970-018

[77] Scott M. Lundberg and Su-In Lee, 2017, A unified approach to interpreting model predictions, In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777, https://doi.org/10.5555/3295222.3295230

[78] Štrumbelj E., Kononenko I., Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst 41, 647–665 (2014). https://doi.org/10.1007/s10115-013-0679-x.

[79] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25 ;366(6464) :447-453. https://doi.org/10.1126/science.aax2342.PMID:31649194.

[80] Baird E, Leida H. Milliman Advanced Risk Adjuster models for racial bias : Medicare model results. https://fr.milliman.com/fr-FR/insight/testing-milliman-advanced-risk-adjuster-models-for-racial-bias-medicare-model-results

[81] Barocas S, Hardt M, and Narayanan A. (2023). Fairness and Machine Learning : Limitations and Opportunities. Adaptive Computation and Machine Learning series. MIT Press. https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/.

[82] Schervish MJ (1989). A General Method for Comparing Probability Assessors. The Annals of Statistics, 17(4) :1856–1879. ://www.jstor.org/stable/2241668.

[83] Van Calster B, McLernon DJ, Van Smeden M, Wynants L, and Steyerberg EW (2019). Calibration : the achilles heel of predictive analytics. BMC medicine, 17(1) :1-7. https://doi.org/10.1186/s12916-019-1466-7.

[84] Chouldechova A. (2017). Fair prediction with disparate impact : A study of bias in recidivism prediction instruments. Big Data, 5(2) :153–163. PMID : 28632438. https://doi.org/10.1089/big.2016.0047..

[85] Kleinberg J, Mullainathan ., and Raghavan M (2016). Inherent trade-offs in the fair determination of risk scores. https://doi.org/10.48550/arXiv.1609.05807.

[86] Pakdaman Naeini M, Cooper G, and Hauskrecht M (2015). Obtaining well calibrated probabilities using bayesian binning. Proceedings of the AAAI Conference on Artificial Intelligence, 29(1) :2901–2907. https://doi.org/10.1609/aaai.v29i1.9602.

[87] World Meteorological Organization. Air quality and climate bulletin no. 4. Technical Report 4, World Meteorological Organization, Geneva, Switzerland, sep 2024.

[88] Climaera, Changement climatique : améliorer la planification territoriale des institutions publiques pour l'adaptation au changement. https://www.climaera.eu/. Consulté le 25 juin 2025.

[89] Gao J, Yang Y, Wang H, Wang P, Li H, Li M, Ren L, Yue X, Liao H. Fast climate responses to emission reductions in aerosol and ozone precursors in China during 2013–2017. Atmospheric Chemistry and Physics, 22(11) :7131–7142, 2022. https://doi.org/10.5194/acp-22-7131-2022.

[90] Delcaillau D, "Contrôle et Transparence des modèles complexes en actuariat" (2019)

[91] Bartz-Beielstein T, Chandrasekaran S, Rehbach F. Case Study II : Tuning of Gradient Boosting (xgboost). In : Bartz, E., Bartz-Beielstein, T., Zaefferer, M., Mersmann, O. (eds) Hyperparameter Tuning for Machine and Deep Learning with R. Springer, Singapore. https://doi.org/10.1007/978-981-19-5170-1_9

Table des figures

Liste des tableaux

Glossaire

TABLE - Liste des acronymes utilisés

Acronyme	**Signification**
ACA	*Affordable Care Act*
AIC	*Akaike Information Criterion*
BEA	*Bureau of Economic Analysis*
BIC	*Bayesian Information Criterion*
CART	*Classification and Regression Trees*
CCS	*Clinical Classifications Software*
CDC	*Centers for Disease Control and Prevention*
CHIP	*Children's Health Insurance Program*
CIM	Classification internationale des maladies
CMS	*Centers for Medicare & Medicaid Services*
CNIL	*Commission nationale de l'informatique et des libertés*
DBU	*Databricks Units*
DTW	*Dynamic Time Warping*
ECE	*Expected Calibration Error*
EDF	Électricité de France
EPA	*Environmental Protection Agency*
EPICES	Évaluation de la précarité et des inégalités de santé dans les centres d'examens de santé
EPO	*Exclusive Provider Organizations*
ERF	*End of Radioactive Fallout*
ESRI	*Environmental Systems Research Institute*
FeNO	Fraction expirée de monoxyde d'azote
GES	Gaz à effet de serre
GDF	Gaz de France
GI	Gastro-intestinale

Suite page suivante

Acronyme	**Signification**
GIEC	Groupe d'experts intergouvernemental sur l'évolution du climat
GLM	*Generalized Linear Model*
GLMM	*Generalized Linear Mixed Model*
HDHPs	*High-Deductible Health Plans*
HIPAA	*Health Insurance Portability and Accountability Act*
HMO	*Health Maintenance Organizations*
HR	*Hazard Ratio*
ICD	*International Classification of Diseases*
IMC	Indice de masse corporelle
MAE	*Mean Absolute Error*
MCO	*Managed Care Organizations*
ML	*Machine Learning*
MRC	Maladie rénale chronique
MSA	*Metropolitan Statistical Area*
MSE	*Mean Square Error*
NAAQS	*National Ambient Air Quality Standards*
NOAA	*National Oceanic and Atmospheric Administration*
OCDE	Organisation de coopération et de développement économiques
OMS	Organisme mondial de la santé
ORL	Oto-rhino-laryngologie
PCP	*Primary Care Provider*
PIB	Produit intérieur brut
PM	Particulate Matter
POS	*Point of Service*
PPA	Parité de pouvoir d'achat
PPO	*Preferred Provider Organizations*
RATP	Régie autonome des transports parisiens
RCM	*Revenue Cycle Management*
RGPD	Règlement général sur la protection des données
RMSE	*Root Mean Square Error*
SNCF	Société nationale des chemins de fer français
VIF	*Variance Inflation Factor*
XGBoost	*eXtreme Gradient Boosting*
ZIP	*Zone Improvement Plan*

Note de synthèse

Introduction

En raison du réchauffement climatique, les périodes de chaleur intense, les épisodes de froid extrême, la fréquence et l'intensité d'événements naturels tels que les inondations ou les tempêtes ainsi que les pics de pollution de l'air seront amenés à croître dans les prochaines années. Or, il est désormais établi que le dérèglement climatique a un impact significatif sur la santé des populations exposées. La prise en compte du climat et de la pollution représente donc un enjeu majeur pour les acteurs de l'assurance santé. Si la littérature est abondante sur leurs effets sanitaires, elle est en revanche très pauvre en ce qui concerne leur intégration dans les modèles de score de santé ou de tarification en assurance santé.

L'objectif de ce mémoire est donc de développer, à partir de données fournies par des assureurs américains, des scores de santé mensuels et annuels intégrant les risques émergents liés au changement climatique et à la pollution de l'air afin d'analyser leur impact sur la santé. La population américaine n'étant pas touchée équitablement par le dérèglement climatique, un point d'attention est également porté à l'évaluation de l'équité des modèles, notamment vis-à-vis des différentes origines ethniques, afin d'identifier d'éventuels biais et de garantir une utilisation responsable de ces outils en assurance santé [10–19, 79–86].

Données de l'étude

Données assurantielles : bases MedInsight

Dans le cadre de ce mémoire, *Milliman MedInsight*, une entreprise spécialisée dans la fourniture de données et d'analyses dans le domaine de la santé, a procuré trois bases de données contenant des informations sur 48,5 millions d'assurés aux Etats-Unis ayant souscrit une assurance santé entre le 1er janvier 2017 et le 31 décembre 2023. Ces bases rassemblent plus de 7,3 milliards d'enregistrements médicaux et pharmaceutiques. Les données renseignent notamment sur le statut de couverture, les frais de santé, les caractéristiques démographiques, l'âge, le sexe, l'ethnicité et l'état de santé de chaque individu de la base éligible à des prestations de santé privées et/ou publiques sur la période. Les trois bases de données fournies sont :

— une base « assurés » présentant des caractéristiques propres à l'assuré comme le genre, l'âge, l'origine ethno-raciale ainsi que l'Etat et le ZIP3 [1] de résidence de l'individu ;

1. Zones de codes postaux américains à trois chiffres qui divisent les Etats américains.

— une base « souscriptions » donnant des informations mensuelles entre le 1er janvier 2017 et le 31 décembre 2023 sur les types d'assurances souscrites par chaque individu de la base et les conditions chroniques dont ils souffrent ; et

— une base « sinistres », recensant l'ensemble des événements médicaux donnant lieu à remboursement ou à facturation pour la période allant du 1er janvier 2017 au 31 décembre 2023. Elle contient plus de 7,5 milliards d'enregistrements individuels, chacun correspondant à une prestation médicale précise (consultation, examen, hospitalisation, délivrance d'un médicament, etc.).

Données environnementales

Les données climatiques relatives aux températures et aux précipitations utilisées dans ce mémoire ont été produites par la *National Oceanic and Atmospheric Administration* (NOAA). Le jeu de données spatiales *nClimGrid-Daily* est un ensemble de champs maillés journaliers et de moyennes de température à la surface et de précipitations couvrant les Etats-Unis à partir de 1951, et disponible publiquement sur leur site internet [2]. Le jeu de données contient, pour chaque maille de 5 km, une valeur quotidienne de la température moyenne (*tavg*), minimale (*tmin*) et maximale (*tmax*) et de la somme des précipitations (*prcp*), du 1er janvier 2016 au 31 décembre 2023.

Les données de pollution exploitées dans ce mémoire ont été produites par le Programme d'observation de la Terre de l'Union européenne *Copernicus*. Le jeu de données spatiales *EAC4* est un ensemble de champs maillés mensuels, avec une résolution d'environ 0,75 degré, de moyennes de variables atmosphériques couvrant l'ensemble de la Terre de 2004 à 2024, et disponible publiquement sur leur site internet [3]. En lien avec une revue de littérature effectuée sur l'impact de la pollution sur la santé [20–40], les moyennes mensuelles de concentrations d'ozone (O_3), de dioxyde d'azote (NO_2) et de $PM_{2.5}$ ont été extraites du jeu de données *EAC4* pour la période de janvier 2016 à décembre 2023.

Ces données environnementales ont été agrégées à la maille géographique des ZIP3 pour correspondre aux données assurantielles exploitées dans ce mémoire. Ainsi, en lien avec une revue de littérature effectuée sur les effets du climat sur la santé [41–69], les valeurs quotidiennes de températures ont permis de créer des indicateurs mensuels et annuels de vagues de froid et de chaleur. Dans l'optique de développer des scores mensuels et annuels, des indicateurs climatiques et de pollution ont également été construits pour chaque maille temporelle.

Données agrégées pour la construction des scores de santé

A partir de ces données assurantielles et environnementales, deux bases de données ont été créées. Une première base de données contient des informations annuelles sur 50 000 assurés du Kentucky (âge, genre), leur santé (nombre de conditions chroniques sur deux années consécutives) et leur environnement géographique (nombre de vagues de froid et de chaleur, concentrations annuelles de NO_2, d'ozone (O_3) et de $PM_{2.5}$). Pour mesurer

2. Données climatiques de la NOAA : `https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc%3AC01589/html#Coverage`

3. Données de pollution de *Copernicus* : `https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4-monthly?tab=overview`

leur impact sur la santé, les indicateurs climatiques et de pollution sont décalés d'une année par rapport à la variable cible, qui correspond au nombre annuel de conditions chroniques `CC_N`. Une seconde base de données contient des informations mensuelles sur 1 000 assurés du Kentucky, leur santé et leur environnement géographique. La variable cible correspond ici aux coûts mensuels liés à des services de soins de santé.

Construction des scores de santé annuels

Méthodologies utilisées

Pour créer un score de santé annuel en approximant le nombre de conditions chroniques par an et par assuré, trois modèles ont été testés et optimisés pour obtenir des scores fidèles à la réalité et permettre une interprétation des sorties : un modèle linéaire généralisé (GLM), un modèle linéaire mixte généralisé (GLMM) et un modèle de *machine learning*, XGBoost (pour *eXtreme Gradient Boosting*). Ces trois modèles ont l'avantage de permettre la mesure des effets des prédicteurs sur le score créé. On note par la suite Y la variable cible et $X \in \mathbb{R}^d$ un vecteur de d prédicteurs.

GLM

Si l'on dispose d'une fonction g bijective, alors le GLM est défini par l'équation suivante :

$$g(\mathbb{E}[Y|\mathbf{X}]) = \beta_0 + \mathbf{X}'\boldsymbol{\beta}$$

où $g(\cdot)$ est la fonction de lien qui relie l'espérance de la variable réponse à la combinaison linéaire des prédicteurs et $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^d$ est le vecteur des coefficients à estimer.

Ici, g sera la fonction *log*. Ainsi, pour l'interprétation, si une variable augmente d'une unité, l'effet sur la variable cible est multiplicatif.

GLMM

Les modèles linéaires mixtes généralisés (GLMM) représentent une extension des modèles linéaires généralisés. Ils permettent de traiter les données hiérarchiques, longitudinales ou corrélées, où les observations ne sont pas indépendantes. Dans un GLMM, la relation entre la variable cible et les variables explicatives est modélisée par une combinaison linéaire d'effets fixes et aléatoires. L'interprétation des coefficients du GLMM est similaire à celle des coefficients du GLM. En utilisant les notations précédentes, le GLMM est défini par l'équation suivante :

$$g(E[Y|X, Z]) = X\beta + Zu$$

où :

- $g(\cdot)$ est la fonction de lien (g sera la fonction *log*) ;
- $X\beta$ représente les effets fixes, où β est le vecteur de coefficients à estimer ; et
- Zu représente les effets aléatoires, où u est un vecteur de variables aléatoires suivant une distribution normale avec une moyenne nulle et une matrice de covariance G, $u \sim \mathcal{N}(0, G)$.

XGBoost

XGBoost est un modèle proposé par Tianqi Chen et Carlos Guestrin en 2016 [75]. Il s'agit d'une implémentation optimisée de l'algorithme de *Gradient Boosted Decision Trees* (GBDT). Contrairement aux forêts aléatoires, dont le but est de construire des arbres de manière indépendante, XGBoost est un algorithme d'ensemble (*ensemble learning*) qui construit chaque arbre de façon séquentielle. Chaque nouvel arbre construit corrige ainsi les erreurs commises par la somme des arbres précédents. Cela a l'inconvénient de le rendre plus lent que les forêts aléatoires, mais lui permet de s'améliorer au fur et à mesure de la construction de la prédiction. On définit $x_i \in \mathbb{R}^d$, un vecteur de d caractéristiques, y_i la cible (valeur à prédire), et n le nombre d'observations. Soit un ensemble de données $\mathcal{P} = \{(x_i, y_i)\}_{i=1}^{n}$. Le modèle prédictif utilisé par XGBoost est un modèle d'ensemble additif : la prédiction finale correspond à la somme des prédictions de chaque arbre. Il correspond à une somme de K fonctions f_k. Chacune de ces fonctions représente un arbre de décision :

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

où $\mathcal{F}$ est l'ensemble des arbres de régression.

L'interprétation des sorties du modèle XGBoost peut se faire à l'aide des valeurs SHAP, qui fournissent une attribution additive, rigoureuse et interprétable de l'importance de chaque variable dans la prédiction du modèle [76–78].

Résultats obtenus

La performance des modèles implémentés pour créer un score de santé annuel basé sur le nombre individuel de conditions chroniques sera mesurée à l'aide de trois métriques :

— la MAE (*Mean Absolute Error*) est la moyenne arithmétique des valeurs absolues des écarts entre les valeurs observées y_i et les valeurs prédites $\hat{y}_i$:
$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$;

— la MSE (*Mean Square Error*) est la moyenne des carrés des écarts entre les valeurs observées et prédites : $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$; et

— la RMSE (*Root Mean Square Error*) est la racine carrée de la MSE.

Table 1 – Comparaison, sur la base de test, des différents modèles mis en place pour modéliser le nombre annuel de conditions chroniques

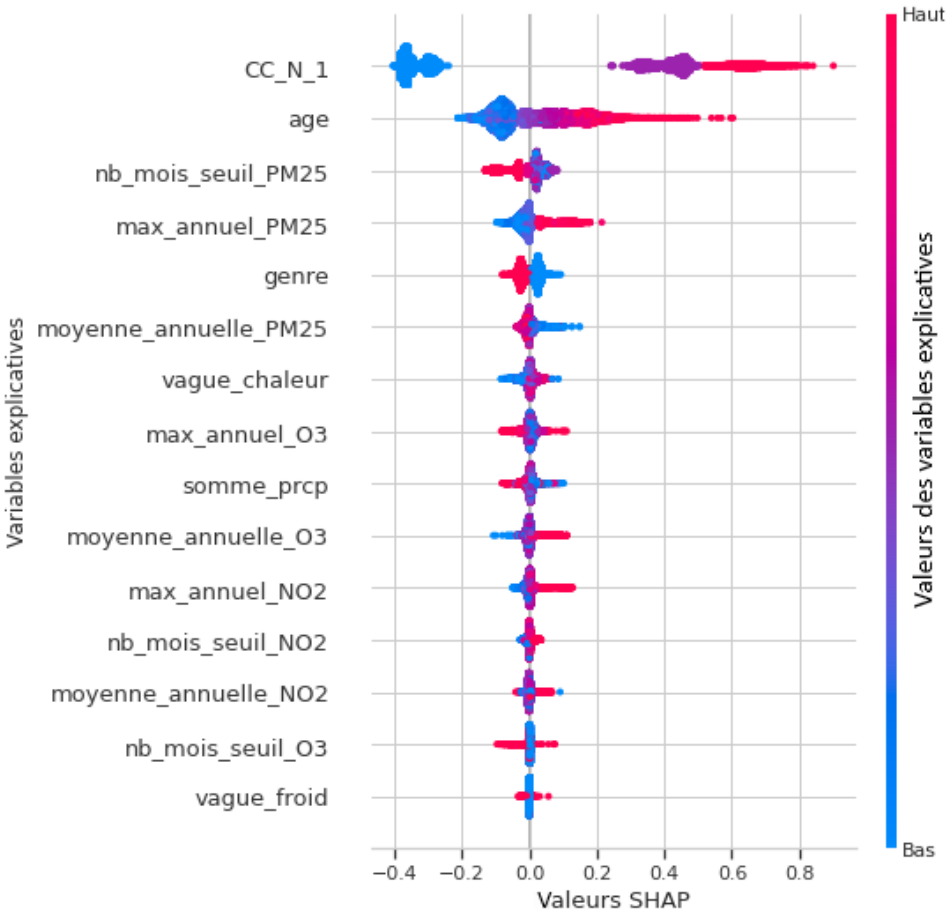
Modèle	**MAE**	**MSE**	**RMSE**
GLM	0,511	0,424	0,651
GLMM	0,451	0,345	0,587
XGBoost	0,416	0,307	0,554

Le tableau 1 résume les résultats des performances des trois modèles implémentés. Bien que le modèle GLM présente une capacité explicative satisfaisante, l'introduction d'effets aléatoires ou d'interactions via le GLMM permet de mieux capturer l'hétérogénéité entre individus. Cependant, les erreurs de performance restent élevées et certains

effets non-linéaires ou interactions complexes entre variables ne sont pas entièrement capturés par une approche linéaire, même mixte. Aussi, la performance prédictive, bien qu'améliorée, a plafonné avec les modèles linéaires. L'optimisation d'un modèle XGBoost a permis d'atteindre des performances nettement supérieures [90, 91]. Au-delà de la performance brute, le modèle XGBoost a montré une équité satisfaisante en termes de biais ethno-raciaux entre les assurés caucasiens et afro-américains.

Les modèles linéaires (GLM et GLMM) sont intrinsèquement interprétables via leurs coefficients. Cependant, certains effets observés des variables environnementales restent contre-intuitifs au regard de la littérature, suggérant des interactions non-modélisées ou de la colinéarité. Dans ces modèles, seuls l'état de santé antérieur, l'âge, le genre, les vagues de chaleur et le dépassement des seuils annuels de NO_2 (seuils fixés par l'*Environmental Protection Agency*) ont présenté des effets néfastes pour la santé. Avec XGBoost, l'usage des valeurs SHAP (voir figure 1) a confirmé l'importance de l'état de santé antérieur, de l'âge et du genre, tout en révélant un impact significatif de certains facteurs environnementaux : l'exposition aux vagues de chaleur, aux particules fines ($PM_{2.5}$) ou au NO_2. Les effets significatifs liés aux variables environnementales restent faibles, car ils sont probablement annihilés par les agrégations temporelle et géographique des données.

FIGURE 1 – Graphique SHAP : importance des variables en fonction de leurs valeurs



Note de lecture : les valeurs SHAP associées aux faibles valeurs de `CC_N_1` *sont comprises entre -0,4 et -0,2.*

Construction des scores de santé mensuels

Méthodologies utilisées

Pour créer un score de santé mensuel en approximant les frais liés à des soins de santé par mois et par assuré, quatre modèles ont été testés et optimisés pour obtenir des performances accrues et permettre une interprétation des sorties : une régression linéaire, un GLM Poisson ainsi que deux modèles de *machine learning* : XGBoost et LightGBM (LGBM). Ces quatre modèles ont l'avantage de permettre la mesure des effets des prédicteurs sur le score créé. Les modèles GLM Poisson et XGBoost ayant été décrits précédemment, seuls le LGBM et la régression linéaire sont détaillés ci-dessous.

Régression linéaire

En reprenant les notations précédentes, la régression linéaire se présente sous la forme matricielle suivante : $Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$. On trouve les coefficients $(\beta_l)_{0 \leq l \leq d}$ du modèle par la méthode des moindres carrés en minimisant l'erreur commise par le modèle. Ce modèle est intrinsèquement interprétable via les coefficients liés à chaque prédicteur.

LGBM

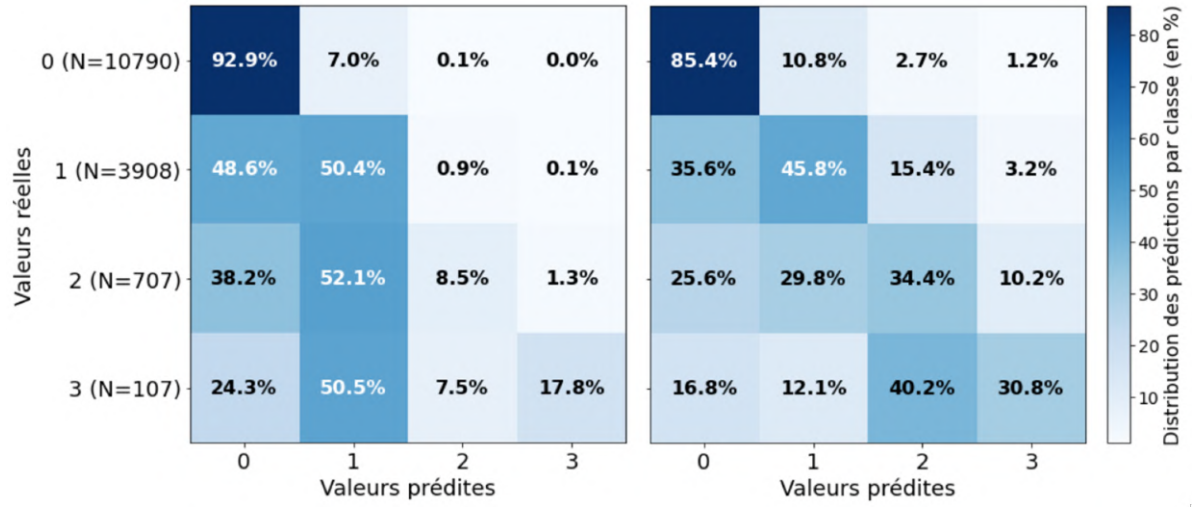
LGBM est un modèle de *Gradient Boosting*, très similaire au modèle XGBoost. Cependant, il existe quelques différences notables qui incitent à tester les deux modèles pour le développement du score de santé mensuel : XGBoost s'appuie sur une méthode de croissance des arbres niveau par niveau tandis que LGBM privilégie une croissance feuille par feuille, plus rapide et plus efficace. De plus, LGBM a une mémoire plus importante et est capable de traiter de grands volumes de données.

Résultats obtenus

Malgré une sélection minutieuse des prédicteurs et la transformation logarithmique appliquée à la variable pour réduire l'asymétrie observée dans sa distribution, la régression linéaire ne permet pas de fournir un score de santé mensuel individuel pertinent pour cette étude : le score ainsi développé est ni performant, ni interprétable. En parallèle de la régression linéaire, le modèle XGBoost a été optimisé sur la variable cible ainsi que sur la variable cible transformée (logarithme). Les résultats n'ont également pas été concluants en termes de prédiction, les modèles ne parvenant pas à intégrer convenablement l'asymétrie de la variable cible.

Ainsi, pour pallier les difficultés liées à la distribution de la variable cible, les montants de sinistres mensuels ont été discrétisés en quatre classes ordinales définies selon des seuils permettant d'obtenir une répartition des effectifs proche de celle attendue sous une loi de Poisson. Ainsi, deux types de modèles ont été implémentés : un GLM Poisson et deux modèles de *Gradient Boosting*, un XGBoost et un LGBM. Le GLM Poisson a montré des métriques de performance nettement inférieures et une matrice de confusion traduisant une capacité limitée à prédire les classes les plus élevées. Même si les métriques de performance du XGBoost sont meilleures, le LGBM présente une matrice de confusion plus pertinente : il a une capacité supérieure à distinguer les classes minoritaires (voir figure 2), ce qui est important dans la gestion du risque santé.

FIGURE 2 – Matrice de confusion (en proportion) pour les modèles XGBoost (à gauche) et LGBM (à droite)



Note de lecture : pour la classe 2, qui contient 707 observations dans la base de test, 8,5 % des valeurs sont correctement prédites (cases diagonales) avec le modèle XGBoost, tandis que 34,4 % des valeurs sont correctement prédites pour cette même classe avec le modèle LGBM.

L'analyse des valeurs SHAP relatives à ce modèle permet d'identifier les principales variables influentes dans la prédiction des frais mensuels associés aux soins de santé. Ainsi, l'âge, le montant des frais antérieurs ainsi que le nombre de diagnostics et de réclamations médicamenteuses sont les principaux prédicteurs du score de santé individuel mensuel. Les variables environnementales sont moins influentes, car sûrement éclipsées par les variables médicales. Toutefois, les concentrations en NO_2 ainsi que les températures semblent jouer un rôle non-négligeable dans la prédiction des classes les plus élevées et pourraient avoir des répercussions néfastes pour la santé des assurés américains.

Bien qu'il soit très performant, le modèle LGBM présente un biais marqué : il sous-estime les frais de santé des assurés afro-américains par rapport aux caucasiens, particulièrement pour les classes de frais non-nuls. Ce comportement, observé aussi avec le modèle XGBoost évincé en raison de sa propension à minimiser le risque santé, met en évidence des biais dans le processus de calcul des scores de santé mensuels basés sur le coût des soins de santé.

Conclusion

Les divers scores de santé développés dans ce mémoire intègrent le climat et la pollution dans le but de mesurer leur impact sur la santé des assurés américains. Les principaux facteurs ayant des effets sur les scores de santé créés sont l'état de santé antérieur, l'âge et le genre de l'assuré et, dans une moindre mesure, certains facteurs environnementaux comme l'exposition aux vagues de chaleur, aux particules fines ($PM_{2.5}$) ou au NO_2. Les résultats obtenus avec ces premières applications sont encourageants, bien que les approches présentées restent perfectibles : la maille temporelle ainsi que les zones géographiques pourraient être affinées pour mieux capter l'impact des risques émergents sur la santé.

Executive summary

Introduction

Due to global warming, periods of intense heat, episodes of extreme cold, the frequency and intensity of natural events such as floods or storms, and peaks in air pollution are set to increase in the coming years. It is now well established that climate change has a significant impact on the health of exposed populations. Taking climate and pollution into account is therefore a major challenge for health insurance providers. While literature is abundant on their health effects, there is very little information on their integration into health scoring or health insurance pricing models.

The objective of this thesis is therefore to develop monthly and annual health scores based on data provided by US insurers that integrate emerging risks related to climate and air pollution to analyze their impact on health. As the American population is not equally affected by environmental issues, attention is also paid to assessing the fairness of the models, particularly concerning different ethnic origins, to identify any biases and ensure the responsible use of these tools in health insurance [10–19, 79–86].

Data

Insurance data : MedInsight databases

For this thesis, Milliman MedInsight, a company specializing in providing data and analysis in the healthcare field, provided three databases containing information on 48.5 million insured individuals in the United States who took out health insurance between January 1, 2017, and December 31, 2023. These databases contain more than 7.3 billion medical and pharmaceutical records. The data includes information on coverage status, healthcare costs, demographic characteristics, age, gender, ethnicity, and health status for each individual in the database who was eligible for private and/or public healthcare benefits during the period. The three databases provided are :

— a database called "assurés [1]" presenting characteristics specific to the insured, such as gender, age, ethnic-racial origin, as well as the state and ZIP3 [2] of residence of the individual ;

— a database called "souscriptions [3]" providing monthly information between January 1, 2017, and December 31, 2023, on the types of insurance policies taken

1. "insured" in English.
2. Three-digit US postal code areas that divide US states.
3. "enrollment" in English.

out by each individual in the database and the chronic conditions they suffer from ; and

— a database called "sinistres [4]" listing all medical events giving rise to reimbursement or billing for the period from January 1, 2017, to December 31, 2023. It contains more than 7.5 billion individual records, each corresponding to a specific medical service (consultation, examination, hospitalization, dispensing of medication, etc.).

Environmental data

The climate data on temperatures and precipitation used in this thesis were produced by the National Oceanic and Atmospheric Administration (NOAA). The *nClimGrid-Daily* spatial dataset is a collection of daily gridded fields and averages of surface temperature and precipitation covering the United States since 1951, and is publicly available on their website [5]. For each 5 km grid cell, the dataset contains daily values for average temperature (*tavg*), minimum (*tmin*) and maximum (*tmax*) temperatures, and total precipitation (*prcp*) from January 1, 2016 to December 31, 2023.

The pollution data used in this thesis were produced by the European Union's Earth observation program *Copernicus*. The *EAC4* spatial dataset is a set of monthly gridded fields, with a resolution of approximately 0.75 degrees, of atmospheric variable averages covering the entire Earth from 2004 to 2024, and is publicly available on their website [6]. In connection with a literature review on the impact of pollution on health [20–40], the monthly averages of ozone (O_3), nitrogen dioxide (NO_2), and $PM_{2.5}$ concentrations were extracted from the *EAC4* dataset for the period from January 2016 to December 2023.

These environmental data were aggregated to the ZIP3 geographic grid to correspond to the insurance data used in this thesis. Thus, in connection with a literature review on the effects of climate on health [41–69], daily temperature values were used to create monthly and annual indicators of cold and heat waves. With a view to developing monthly and annual scores, climate and pollution indicators were also constructed for each time grid.

Aggregated data for the construction of health scores

Based on this insurance and environmental data, two databases were created. The first database contains annual information on the insured person (age, gender), their health (number of chronic conditions over two consecutive years) and their geographical environment (number of cold and heat waves, annual concentrations of NO_2, ozone (O_3) and $PM_{2.5}$). To measure their impact on health, climate and pollution indicators are shifted by one year relative to the target variable, which corresponds to the annual number of chronic conditions `CC_N`. A second database contains monthly information on the insured person, their health, and their geographical environment. The target variable here corresponds to the monthly costs associated with healthcare services.

4. "claims" in English.

5. NOAA Climate Data : `https://www.ncei.noaa.gov/metadata/geoportal/rest/metadata/item/gov.noaa.ncdc%3AC01589/html#Coverage`

6. Copernicus pollution data : `https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4-monthly?tab=overview`

Construction of annual health scores

Methodologies

To create an annual health score by approximating the number of chronic conditions per year per insured person, three models were tested and optimized to obtain scores that are true to reality and allow for interpretation of the outputs : a generalized linear model (GLM), a generalized linear mixed model (GLMM), and a machine learning model, XGBoost (for *eXtreme Gradient Boosting*). These three models have the advantage of allowing the effects of predictors on the created score to be measured. We then denote the target variable by Y and a vector of d predictors by $X \in \mathbb{R}^d$.

GLM

If we assume that we have a bijective function g, then the GLM is defined by the following equation :

$$g(\mathbb{E}[Y|\mathbf{X}]) = \beta_0 + \mathbf{X}'\boldsymbol{\beta}$$

where $g(\cdot)$ is the link function that connects the expectation of the response variable to the linear combination of predictors, and $(\beta_0, \boldsymbol{\beta}) \in \mathbb{R} \times \mathbb{R}^d$ is the vector of coefficients to be estimated.

Here, g will be the *log* function. Thus, for interpretation, if a variable increases by one unit, the effect on the target variable is multiplicative.

GLMM

Generalized linear mixed models (GLMM) are an extension of generalized linear models. They allow hierarchical, longitudinal, or correlated data to be processed, where observations are not independent. In a GLMM, the relationship between the target variable and the explanatory variables is modeled by a linear combination of fixed and random effects. The interpretation of GLMM coefficients is similar to that of GLM coefficients. Using the previous notation, the GLMM is defined by the following equation :

$$g(E[Y|X, Z]) = X\beta + Zu$$

where :

— $g(\cdot)$ is the link function (g will be the *log* function) ;

— $X\beta$ represents the fixed effects, where β is the vector of coefficients to be estimated ; and

— Zu represents the random effects, where u is a vector of random variables following a normal distribution with zero mean and covariance matrix G, $u \sim \mathcal{N}(0, G)$.

XGBoost

XGBoost is a model proposed by Tianqi Chen and Carlos Guestrin in 2016 [75]. It is an optimized implementation of the Gradient Boosted Decision Trees (GBDT) algorithm. Unlike random forests, whose goal is to build trees independently, XGBoost is an ensemble learning algorithm that builds each tree sequentially. Each new tree built thus corrects

the errors made by the sum of the previous trees. This has the disadvantage of making it slower than random forests, but allows it to improve as the prediction is constructed. We define $x_i \in \mathbb{R}^d$, a vector of d features, y_i, the target (value to be predicted), and n the number of observations. Let $\mathcal{P} = \{(x_i, y_i)\}_{i=1}^n$ be a data set. The predictive model used by XGBoost is an additive ensemble model : the final prediction corresponds to the sum of the predictions of each tree. It corresponds to a sum of K functions f_k. Each of these functions represents a decision tree :

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

where $\mathcal{F}$ is the set of regression trees.

The outputs of the XGBoost model can be interpreted using SHAP values, which provide an additive, rigorous, and interpretable attribution of the importance of each variable in the model's prediction [76–78].

Results

The performance of the models implemented to create an annual health score based on the individual number of chronic conditions will be measured using three metrics :

— MAE (*Mean Absolute Error*) is the arithmetic mean of the absolute values of the differences between the observed values y_i and the predicted values $\hat{y}_i$: $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$;

— MSE (*Mean Square Error*) is the average of the squares of the differences between the observed and predicted values : $\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$; and

— RMSE (*Root Mean Square Error*) is the square root of MSE.

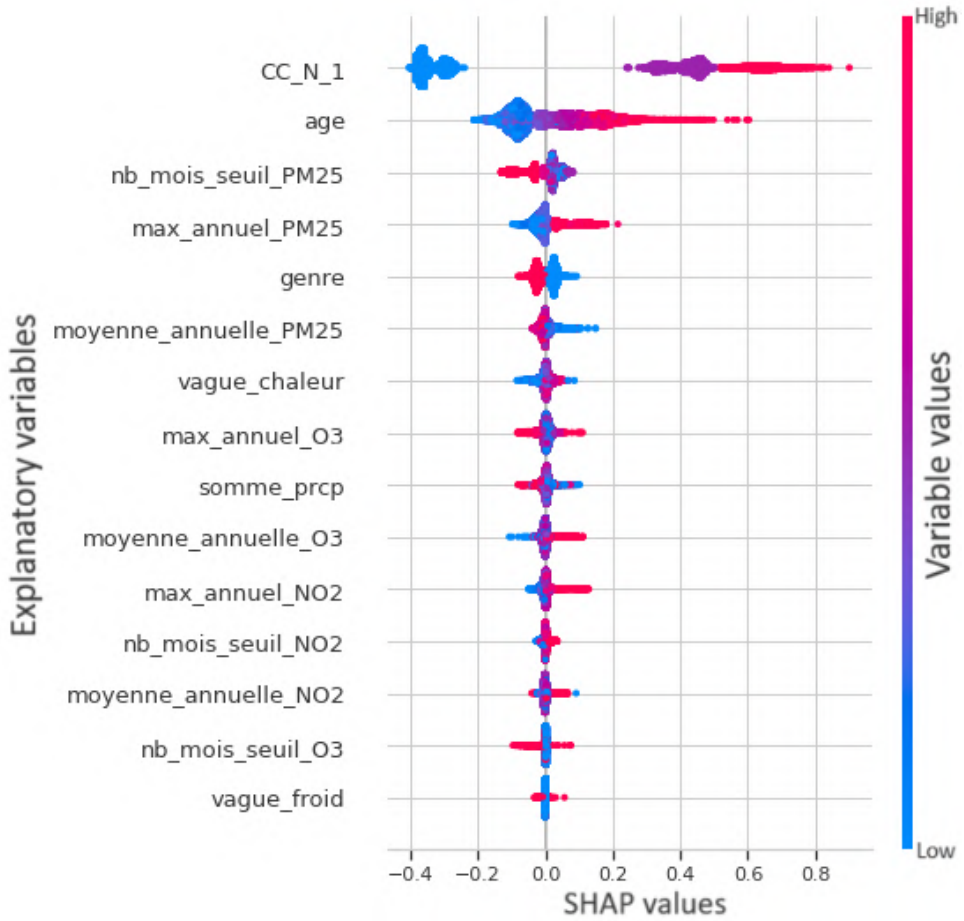
TABLE 1 – Comparison on test dataset of the different models implemented to aproximate the annual number of chronic conditions

Model	**MAE**	**MSE**	**RMSE**
GLM	0.511	0.424	0.651
GLMM	0.451	0.345	0.587
XGBoost	0.416	0.307	0.554

Table 1 summarizes the performance results of the three models implemented. Although the GLM model has satisfactory explanatory power, the introduction of random effects or interactions via the GLMM allows for better capture of heterogeneity between individuals. However, performance errors remain high, and certain nonlinear effects or complex interactions between variables are not fully captured by a linear approach, even a mixed one. Moreover, predictive performance, although improved, has plateaued with linear models. Optimizing an XGBoost model has made it possible to achieve significantly higher performance [90, 91]. Beyond raw performance, XGBoost has shown satisfactory fairness in terms of racial bias between Caucasian and African US policyholders.

Linear models (GLM and GLMM) are inherently interpretable via their coefficients. However, some observed effects of environmental variables remain counterintuitive in light of the literature, suggesting unmodeled interactions or collinearity. In these models, only previous health status, age, gender, heat waves, and exceeding annual NO_2 thresholds (thresholds set by the Environmental Protection Agency) had adverse health effects. With XGBoost, the use of SHAP values (see Figure 1) confirmed the importance of previous health status, age, and gender, while revealing a significant impact of certain environmental factors : exposure to heat waves, fine particulate matter ($PM_{2.5}$), or NO_2. The significant effects related to environmental variables remain weak, as they are likely offset by the temporal and geographical aggregation of the data.

FIGURE 1 – SHAP graph : importance of variables based on their values



Reading note : SHAP values associated with low values of `CC_N_1` *are between -0.4 and -0.2.*

Construction of monthly health scores

Methodologies

To create a monthly health score by approximating healthcare costs per month per insured person, four models were tested and optimized to improve performance and enable interpretation of the outputs : a linear regression, a Poisson GLM, and two machine learning models : XGBoost and LightGBM (LGBM). These four models have the advan-

tage of allowing the effects of predictors on the score created to be measured. As the Poisson GLM and XGBoost models have been described previously, only LGBM and linear regression are detailed below.

Linear regression

Using the previous notation, linear regression can be expressed in the following matrix form : $Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$. The coefficients $(\beta_l)_{0 \leq l \leq d}$ of the model are found using the least squares method by minimizing the error made by the model. This model is intrinsically interpretable via the coefficients linked to each predictor.

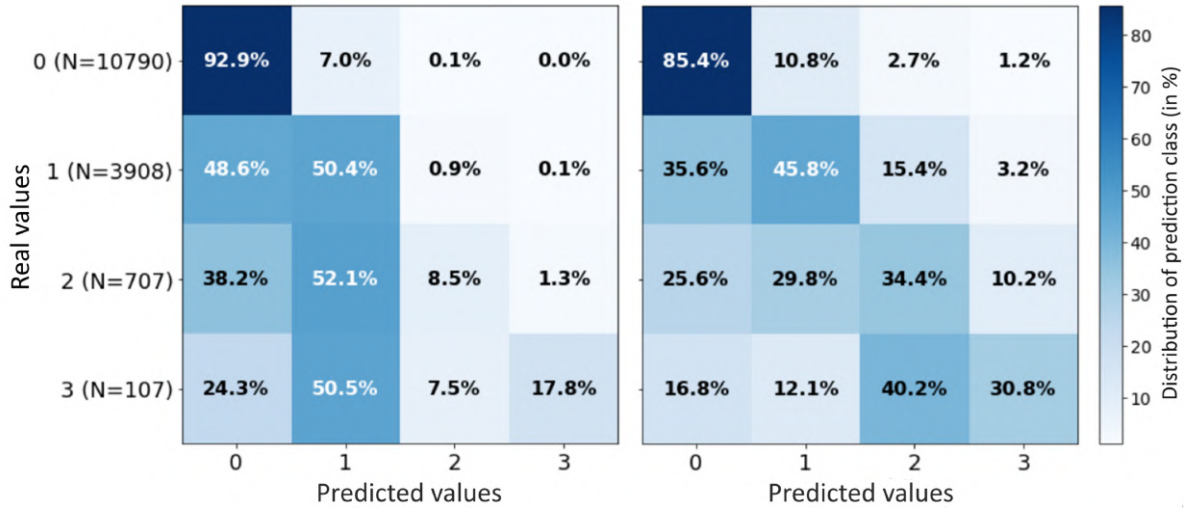
LGBM

LGBM is a Gradient Boosting model, very similar to the XGBoost model. However, some notable differences make it worthwhile to test both models for the development of the monthly health score : XGBoost relies on a level-by-level tree growth method, while LGBM favors a leaf-by-leaf growth method, which is faster and more efficient. In addition, LGBM has a larger memory and is capable of processing large volumes of data.

Results

Despite careful selection of predictors and logarithmic transformation applied to the variable to reduce the asymmetry observed in its distribution, linear regression does not provide a relevant individual monthly health score for this study : the score developed in this way is neither effective nor interpretable. In parallel with linear regression, the XGBoost model was optimized on the target variable as well as on the transformed target variable (logarithm). The results were also inconclusive in terms of prediction, as the models failed to adequately integrate the asymmetry of the target variable.

FIGURE 2 – Confusion matrix (in proportion) for the XGBoost (left) and LGBM (right) models



Reading note : for class 2, which contains 707 observations in the test database, 8.5 % of the values are correctly predicted (diagonal cells) with the XGBoost model, while 34.4 % of the values are correctly predicted for this same class with the LGBM model.

Thus, to overcome the difficulties associated with the distribution of the target variable, the monthly claim amounts were discretized into four ordinal classes defined according to thresholds that allowed for a distribution of frequencies close to that expected under a Poisson distribution. Two types of models were implemented : a Poisson GLM and two Gradient Boosting models, an XGBoost and an LGBM. The Poisson GLM showed significantly lower performance metrics and a confusion matrix reflecting a limited ability to predict the highest classes. Although the performance metrics of XGBoost are better, LGBM presents a more relevant confusion matrix : it has a superior ability to distinguish minority classes (see Figure 2), which is important in health risk management.

Analysis of the SHAP values for this model identifies the main variables influencing the prediction of monthly healthcare costs. Age, previous costs, and the number of diagnoses and medication claims are the main predictors of the individual monthly health score. Environmental variables are less influential, as they are likely overshadowed by medical variables. However, NO_2 concentrations and temperatures appear to play a significant role in predicting the highest classes and could have adverse health effects on US policyholders.

Although highly effective, the LGBM model has a marked bias : it underestimates the healthcare costs of African American insured individuals compared to Caucasians, particularly for non-zero cost classes. This behavior, also observed with the XGBoost model, which was rejected due to its tendency to minimize health risk, highlights biases in the process of calculating monthly health scores based on healthcare costs.

Conclusion

The various health scores developed in this thesis incorporate climate and pollution to measure their impact on the health of American insured individuals. The main factors affecting the health scores created are the insured person's previous health status, age, and gender and, to a lesser extent, certain environmental factors such as exposure to heat waves, fine particulate matter ($PM_{2.5}$), or NO_2. The results obtained with these initial applications are encouraging, although the approaches presented could be improved : the time grid and geographical areas could be refined to better capture the impact of emerging health risks.

Annexes

Annexe A

Description des pathologies chroniques possibles dans la base « souscriptions »

ID	Pathologies chroniques
101	Psychose majeure
102	Démence sévère
103	Cancer
104	Insuffisance rénale et/ou après une greffe de rein
105	Maladie du foie (Hépatite, Cirrhose) – après une greffe
106	VIH
107	Maladie rhumatismale sévère et autres maladies du tissu conjonctif
108	Insuffisance cardiaque sévère/transplantation/maladie cardiaque rhumatismale/maladie valvulaire non rhumatismale
109	Hémophilie et drépanocytose et troubles sanguins chroniques
110	À la fois maladie coronarienne et diabète
111	maladie coronarienne
112	Diabète
113	Hypertension (y compris AVC et maladie vasculaire périphérique)
114	MPOC
115	Asthme
116	Troubles neurologiques
117	Déficience intellectuelle
118	Anomalies congénitales
119	Fibrose kystique
120	Troubles musculosquelettiques chroniques/arthrose/ostéoporose
121	Dépression, toxicomanie et autres troubles de santé mentale
122	Troubles gastro-intestinaux
123	Troubles de la thyroïde
124	Troubles dermatologiques
125	Nouveau-nés et prématurés en mauvaise santé
126	Autres conditions chroniques
127	Nourrisson en bonne santé (0-1)
128	Enfant en bonne santé (2-5)
129	Garçon en bonne santé (6-15)
130	Homme en bonne santé (16-40)

Suite page suivante

ID	Pathologies chroniques
131	Homme en bonne santé (41-64)
132	Homme en bonne santé (65-69)
133	Homme en bonne santé (70-74)
134	Homme en bonne santé (75-79)
135	Homme en bonne santé (80-84)
136	Homme en bonne santé (85+)
137	Fille en bonne santé (6-15)
138	Femme en bonne santé (16-40)
139	Femme en bonne santé (41-64)
140	Femme en bonne santé (65-69)
141	Femme en bonne santé (70-74)
142	Femme en bonne santé (75-79)
143	Femme en bonne santé (80-84)
144	Femme en bonne santé (85+)
145	Autres en bonne santé (âge ou sexe inconnu)

Annexe B

Compléments sur les modèles de *machine learning*

Les compléments suivants sont basés sur le cours de « *Statistical Learning in Practice* » dispensé par Dr. Randolf Altmeyer à l'université *Imperial College London*, ainsi que sur le cours « Méthode de Gradient Boosting » dispensé par Laure Ferraris et Paul Liautaud à la Sorbonne Université.

B.1 Arbres de décision : algorithme CART

Dans cette section est détaillé l'algorithme standard pour calculer des fonctions de régression et des classificateurs basés sur les arbres : l'algorithme CART. Il partitionne l'espace des covariables $\mathbb{R}^p$ en un nombre fini de petites régions rectangulaires R_m, où les fonctions de régression ou les classificateurs prennent des valeurs constantes. Ils sont de la forme

$$x \mapsto \sum_{m=1}^{M} c_m 1(x \in R_m), \quad x \in \mathbb{R}^p.$$

Trouver la meilleure partition en termes de minimisation de l'erreur de prédiction est computationnellement infaisable. A la place, l'utilisation d'un algorithme est nécessaire pour séparer récursivement les régions existantes en deux régions plus petites jusqu'à ce qu'un certain critère d'arrêt soit atteint. Les données servent à guider les séparations. Chaque séparation concurrente correspond à une séparation alignée sur un axe d'une région existante R en deux sous-régions

$$R_l(j,t) = \{x \in R : x_j \leq t\}, \quad \text{et} \quad R_r(j,t) = \{x \in R : x_j > t\},$$

avec des variables de séparation $j \in \{1, \ldots, p\}$ et des points de séparation $t \in \mathbb{R}$. La meilleure séparation est décidée en minimisant une fonction de coût $Q(j,t)$ sur toutes les variables et points de séparation possibles. Ainsi, une variable peut être utilisée dans plusieurs séparations, une seule fois, ou jamais).

Concernant les arbres de régression CART, notons les observations (X_i, Y_i), $i = 1, \ldots, n$ avec $X_i \in \mathbb{R}^p$ et $Y_i \in \mathbb{R}$. L'arbre de régression est construit comme suit :

1. Initialiser l'ensemble des régions $\mathcal{R} = \{\mathbb{R}^p\}$.

2. Pour $R \in \mathcal{R}$, soit $(b_j, b_t) \in \arg\min_{j,t} Q(j,t)$, où nous utilisons la fonction de coût

$$Q(j,t) := \min_{c_l \in \mathbb{R}} \sum_{X_i \in R_l(j,t)} (Y_i - c_l)^2 + \min_{c_r \in \mathbb{R}} \sum_{X_i \in R_r(j,t)} (Y_i - c_r)^2 - \min_{c \in \mathbb{R}} \sum_{X_i \in R} (Y_i - c)^2.$$

3. Remplacer R par $(\mathcal{R} \setminus R) \cup \{R_l(b_j, b_t), R_r(b_j, b_t)\}$.
4. Répéter les étapes 2-3 jusqu'à ce que toutes les régions de $\mathcal{R}$ contiennent au plus (disons) 5 covariables X_i.
5. Avec $\mathcal{R} = \{R_1, \ldots, R_M\}$ et $N(R) := |\{i : X_i \in R\}|$, la fonction finale de l'arbre de régression est

$$\hat{f}_b(x) = \sum_{m=1}^{M} \hat{c}_m 1(x \in R_m), \quad \hat{c}_m = \frac{1}{N(R_m)} \sum_{X_i \in R_m} Y_i, \quad x \in \mathbb{R}^p.$$

Les régions qui ne sont plus séparées sont appelées feuilles, et la partition binaire résultante peut être organisée comme un arbre binaire avec les régions de séparation comme nœuds. La fonction de coût peut se réécrire :

$$Q(j,t) = \sum_{X_i \in R_l(j,t)} (Y_i - \hat{c}_l)^2 + \sum_{X_i \in R_r(j,t)} (Y_i - \hat{c}_r)^2 - \sum_{X_i \in R} (Y_i - \hat{c})^2$$

avec $\hat{c}_l, \hat{c}_r$ et $\hat{c}$ étant les valeurs moyennes des Y_i sur les régions $R_l(j,t)$, $R_r(j,t)$ et R. Cela signifie que minimiser $Q(j,t)$ réduit les variances intra-nœuds des réponses pour les covariables tombant dans les régions $R_l(j,t)$ et $R_r(j,t)$ comparé à la variance intra-nœud sur tout R. Il est possible de montrer que séparer les régions de cette manière ne peut jamais augmenter l'erreur d'entraînement. Etant donné le critère de coût ci-dessus, il suffit de restreindre les points de séparation s séparant les observations. Par exemple, pour chaque j, nous pouvons choisir comme points de séparation les milieux entre les X_{ij} adjacents pour $X_i \in R$. Cela permet également d'éviter les ex-æquo.

Concernant les arbres de classification, les données sont notées (X_i, Y_i), $i = 1, \ldots, n$ avec $Y_i \in \{1, \ldots, K\}$. Pour une région R on pose

$$\hat{p}_k(R) := \frac{1}{N(R)} \sum_{X_i \in R} 1(Y_i = k), \quad k = 1, \ldots, K,$$

et l'indice de Gini est défini comme suit : $G(R) := \sum_{k=1}^{K} \hat{p}_k(R)(1 - \hat{p}_k(R))$. La fonction de coût utilisée est la suivante :

$$Q(j,t) := \frac{N(R_l(j,t))}{N(R)} G(R_l(j,t)) + \frac{N(R_r(j,t))}{N(R)} G(R_r(j,t)) - G(R).$$

L'indice de Gini mesure l'impureté d'une région R. Si toutes les observations de R tombent dans la même classe, alors l'indice s'annule. L'indice de Gini est, en général, préféré à l'erreur de classification

$$\frac{1}{N(R)} \sum_{X_i \in R} 1(Y_i \neq \hat{b}_k(R)), \quad \hat{b}_k(R) \in \arg\max_k \hat{p}_k(R),$$

car cette dernière n'est pas différentiable (ce qui est utile pour l'optimisation numérique) et peut ne pas produire de solutions uniques à chaque séparation. Avec cette fonction de

coût, l'arbre de classification est construit comme ci-dessus. Le classificateur final avec les feuilles $R_1, \ldots, R_M$ est

$$h_{\text{Tree}}(x) = \sum_{m=1}^{M} \hat{c}_m 1(x \in R_m), \quad \hat{c}_m = \hat{b}_k(R_m).$$

CART peut produire des arbres très profonds, et plus l'arbre est profond, plus il est susceptible de sur-ajuster. Une stratégie pour réduire le sur-apprentissage et donc la variance est de n'accepter que les séparations en dessous d'un certain seuil pour $Q(j,t)$, mais cette stratégie est vouée à l'échec car une mauvaise séparation peut être suivie d'une bien meilleure plus tard. Une autre stratégie consiste à considérer la taille de l'arbre comme un paramètre de complexité pouvant être optimisé via la validation croisée. Au lieu de considérer tous les sous-arbres possibles de même taille, la stratégie préférée est de faire croître un grand arbre, puis de le réduire par élagage selon la complexité du coût pour trouver une suite de sous-arbres candidats, qui peuvent être comparés par validation croisée.

B.2 Forêts aléatoires

Les arbres de régression et de classification sont faciles à interpréter, mais sont aussi connus pour être instables : ils ont un faible biais, mais aussi une forte variance. En effet, une réalisation différente des données ou l'ajout de nouveaux points de données peut conduire à des séparations différentes dès le début, et ainsi générer un arbre très différent. Le *bagging* des arbres de décision consiste à tirer à plusieurs reprises des échantillons *bootstrap*, en construisant un arbre de décision sur chaque échantillon, et en agrégeant les différents arbres. C'est le principe de base derrière la construction des forêts aléatoires. Pour la régression, l'algorithme procède comme suit :

1. Pour $b = 1, \ldots, B$: Tirer n échantillons $(X_i^{(b)}, Y_i^{(b)})$ au hasard avec remise à partir des observations. Faire croître un arbre de régression maximal $f^{(b)}$ (sans élagage) à partir de l'échantillon *bootstrap*, mais avant chaque séparation, sélectionner aléatoirement sans remise m parmi les p variables comme candidates pour la séparation.
2. Agréger les arbres de régression pour former

$$f_{\text{RF}}(x) = \frac{1}{B} \sum_{b=1}^{B} f^{(b)}(x).$$

Pour la classification, les arbres de classification *bootstrapés* $h^{(b)}$ sont construits de la même manière en sous-échantillonnant les variables avant la séparation, puis en agrégeant les arbres de classification pour former le classificateur

$$h_{\text{RF}}(x) \in \arg\max_k \frac{1}{B} \sum_{b=1}^{B} 1(h^{(b)}(x) = k).$$

Sans sous-échantillonnage (donc $m = p$), une forêt aléatoire devient un arbre de décision *baggé*. Les valeurs typiques pour m sont $\sqrt{p}$ pour la classification et $p/3$ pour la régression. En général, m est un paramètre d'ajustement et peut être choisi par validation croisée.

Pour comprendre l'importance du *bagging*, notons que les arbres de régression $f^{(b)}$ ont tous la même distribution (dépendant des échantillons *bootstrap* et des données originales). Pour x fixé dans $\mathbb{R}^p$, la moyenne conserve donc l'espérance. Le biais est alors inchangé :

$$\mathbb{E}[f_{\mathrm{RF}}(x)] = \mathbb{E}[f^{(b)}(x)].$$

Pour $b \neq b'$, posons $\sigma^2 = \mathrm{Var}(f^{(b)}(x))$, $\rho = \mathrm{Corr}(f^{(b)}(x), f^{(b')}(x))$. Alors, la variance est de la forme

$$\begin{aligned}
\mathrm{Var}(f_{\mathrm{RF}}(x)) &= \frac{1}{B^2} \sum_{b,b'} \mathrm{Cov}\left(f^{(b)}(x), f^{(b')}(x)\right) \\
&= \frac{\sigma^2}{B^2} \sum_{b,b'} \mathrm{Corr}\left(f^{(b)}(x), f^{(b')}(x)\right) \\
&= \frac{1}{B^2}\left(B\sigma^2 + B(B-1)\rho\sigma^2\right) \\
&= \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2
\end{aligned}$$

Lorsque $B \to \infty$, l'erreur se réduit à $\rho\sigma^2$, et l'étape de sous-échantillonnage lors de la construction des arbres vise à réduire davantage la corrélation. Pour la classification, la réduction de la variance obtenue par le bagging peut être encore plus importante, car la variance influence l'erreur de prédiction de manière non linéaire.

On rappelle qu'un échantillon bootstrap contient une observation donnée avec une probabilité d'environ $1/e$. Cela signifie que pour chaque point de données (X_i, Y_i), il existe environ B/e arbres bootstrap qui ont été construits sans ce point de données. Soit $\hat{Y}_{\mathrm{b}i}$ la prédiction agrégée pour Y_i à partir de X_i parmi ces arbres. L'erreur *out-of-bag* (erreur OOB)

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_{\mathrm{b}i})^2$$

est alors une estimation de l'erreur de prédiction. Pour la classification, on utilise de façon similaire comme erreur OOB

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(Y_i \neq \hat{Y}_{\mathrm{b}i}).$$

Il est possible de montrer que l'erreur OOB tend vers l'erreur de validation croisée à n plis (Err_{CV}) lorsque $B \to \infty$.

Il n'y a généralement pas beaucoup de gain à ajuster la profondeur des arbres bootstrap individuels. En effet, les forêts aléatoires donnent souvent d'excellents résultats avec la configuration par défaut et sans aucun ajustement. Cela constitue un avantage significatif par rapport à d'autres méthodes. Les forêts aléatoires rencontrent des difficultés en haute dimension lorsqu'il n'y a que peu de variables pertinentes mais que m est trop petit. En revanche, le sous-échantillonnage réduit le risque de surapprentissage lorsqu'il y a de nombreuses variables bruitées.

B.3 *Gradient Boosting*

On dispose d'un jeu de données d'entraînement $D_n := \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$ contenant n paires $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ de variables aléatoires où $\mathcal{X} \subset \mathbb{R}^d$ et $\mathcal{Y} = \{-1, 1\}$ pour un problème de classification binaire ou $\mathcal{Y} = \mathbb{R}$ pour un problème de régression.

On suppose les données D_n indépendantes et identiquement distribuées (i.i.d.) et $\mathbb{E}[Y^2] < \infty$. Le but de cette section est d'estimer, à l'aide de l'algorithme de *Gradient Boosting*, la fonction $F^* : \mathcal{X} \to \mathcal{Y}$ solution du problème

$$F^* = \arg\min_{F \in \mathcal{F}} \mathbb{E}[L(Y, F(X))]$$

où $\mathcal{F}$ est l'espace des fonctions de $\mathcal{X} \to \mathcal{Y}$ et L une fonction de perte.

Pour ce faire, introduisons dans un premier temps **l'apprenant de base**.

B.3.1 Apprenant de base

On appelle *weak learner* un modèle de prédiction très simple qui performe légèrement mieux que le hasard. A l'inverse, un *strong learner* est un modèle robuste capable de donner des prédictions fiables. Le but du *Boosting* est de transformer un *weak learner* en un *strong learner* par incréments successifs. Chaque agrégat booste l'apprenant initial en concentrant son entraînement sur des erreurs passées. Il peut s'appliquer tant à un problème de régression que de classification.

Définition - Apprenant faible
On appelle apprenant de base ou apprenant faible un algorithme W pour lequel il existe $\epsilon_w < \frac{1}{2}$ et $\delta_w < 1$ tels que, étant donné l'ensemble D_n, alors W génère un résultat avec précision au moins $1 - \epsilon_w$ et probabilité au moins $1 - \delta_w$. On appellera $\mathcal{H}$ la classe de tels apprenants faibles.

Par exemple, les algorithmes suivants sont usuellement considérés comme des apprenants de base dans la littérature :

- **multi-layer perceptron** : un réseau avec peu de couches et un faible nombre de neurones (nœuds) sur chaque couche qui s'applique sur peu de variables d'entrée ;
- **classification naïve bayésienne** ;
- **arbres prédicteurs de profondeur faible** et un faible nombre $L \in \{0, 1, 2, 3\}$ de nœuds terminaux.

En fixant $0 < \epsilon < \frac{1}{2}$ et $0 < \delta < 1$, le but du *Boosting* est alors de produire un résultat d'une précision d'au moins $1 - \epsilon$ et ce avec une confiance $1 - \delta$ en utilisant itérativement les apprenants faibles W.

B.3.2 Descente de gradient

Le but de la descente de gradient est d'estimer une fonction reliant un vecteur de données (entrées) $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$ à une sortie Y. A partir d'un jeu d'apprentissage $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, l'objectif est de déterminer la fonction F^* qui modélise

au mieux l'interdépendance entre X et Y. Formellement, pour une classe de fonctions $\mathcal{F}$ choisie, un tel F^* est défini par

$$F^* = \arg\min_{F \in \mathcal{F}} \mathbb{E}\left[L(Y, F(X))\right]$$

où $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ est une fonction de perte convexe et intégrable.

En pratique, la loi jointe (X, Y) est inconnue. On approche ainsi l'équation précédente par sa version empirique :

$$F^* = \arg\min_{F \in \mathcal{F}} \sum_{i=1}^{n} L(Y_i, F(X_i))$$

Le *Boosting* introduit une approximation de F^* par une combinaison additive d'apprenants faibles :

$$F(X) = \sum_{m=0}^{M} \beta_m h(X, a_m)$$

où $\{\beta_m\}_{m=0}^{M}$ sont les poids attribués à chaque apprenant faible $\{h(\cdot, a_m)\}_{m=0}^{M}$. Chaque apprenant $h(\cdot, a_m) \in \mathcal{H}$ est caractérisé par un ensemble de paramètres $\{a_m\}_{m=0}^{M}$. Le problème d'optimisation initial est ainsi ramené à l'estimation des couples (a_m, β_m). On peut alors reformuler l'équation empirique précédente :

$$F^* = \arg \min_{\{a_m\}_0^M, \{\beta_m\}_0^M} \sum_{i=1}^{n} L\left(Y_i, \sum_{m=0}^{M} \beta_m h(X_i, a_m)\right)$$

L'optimisation conjointe de $2M$ paramètres est complexe. On adopte alors une approche dite *stage-wise* : elle scinde le problème en M étapes. On initialise la procédure avec une fonction F_0 puis, pour chaque itération $1 \leq m \leq M$, on résout :

$$a_m, \beta_m = \arg\min_{a, \beta} \sum_{i=1}^{n} L\left(Y_i, F_{m-1}(X_i) + \beta h(X_i, a)\right)$$

où F_{m-1} est définie récursivement par

$$F_m(X) = F_{m-1}(X) + \beta_m h(X, a_m)$$

Pour $1 \leq m \leq M$, la méthode *Gradient Boosting* optimise séparément les paramètres a_m et β_m :

1. $$a_m = \arg\min_{a} \sum_{i=1}^{n} \left(\frac{\partial L(Y_i, F_{m-1}(X_i))}{\partial F_{m-1}(X_i)} - h(X_i, a) \right)^2$$

2. $$\beta_m = \arg\min_{\beta} \sum_{i=1}^{n} L\left(Y_i, F_{m-1}(X_i) + \beta h(X_i, a_m)\right)$$

On peut interpréter les deux étapes ci-dessus comme une itération d'un algorithme de descente de gradient pour le problème d'optimisation sous contrainte dans l'espace fonctionnel $\mathcal{F}$. Pour le montrer, on peut définir pour $(x_1, y_1), \ldots, (x_n, y_n)$ de D_n, la fonction de perte L :

$$L : \mathcal{F} \longrightarrow \mathbb{R}$$
$$F \longmapsto L(F(x_1), \ldots, F(x_n)) = \sum_{i=1}^{n} L(y_i, F(x_i))$$

On cherche cette fois-ci à résoudre :

$$\min_{F \in \mathcal{F}} L(F)$$

sous la contrainte que $F \in \mathcal{H}$.

L'algorithme de descente du gradient est une méthode classique de résolution d'un tel problème. En ignorant la contrainte :

$$F_{m+1} = F_m - \beta \nabla L(F)\big|_{F=F_m}$$

où $\beta > 0$ est un pas d'apprentissage strictement positif. Une version sous contrainte peut être donnée par :

$$F_{m+1} = F_m + \beta h_m$$

où h_m joue le rôle d'un apprenant et β est le pas optimal. Pour $h_m \in \mathcal{H}$ (faible apprenant), alors F, construite par récurrence, est bien définie comme une combinaison de ces h_m.

Il reste à approcher l'opposé du gradient $\nabla L(F)$ par une fonction $h \in \mathcal{H}$ par rapport aux données d'entraînement D_n. Pour ce faire, comme $\nabla L(F)$ est un vecteur, on peut chercher $h \in \mathcal{H}$ de telle sorte que :

$$\|-\nabla L(F) - h\|_2^2 = \sum_{i=1}^{n} \left(-\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} - h(x_i) \right)^2$$

soit minimal.

On peut voir cette étape comme la projection de $-\nabla L(F)$ sur l'espace des contraintes $\mathcal{H}$. Comme h_m est paramétrique par définition, on peut chercher les paramètres a_m qui la caractérisent :

$$a_m = \arg\min_{a} \sum_{i=1}^{n} \left(\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} - h(x_i, a) \right)^2$$

Puis, le pas optimal β_m est déterminé comme suit :

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \beta h_m(x_i)\right)$$

Finalement, ces deux étapes correspondent bien à l'itération m décrite par les équations précédentes. On note, $\forall i \in \{1, \ldots, n\}$ et $\forall m \in \{1, \ldots, M\}$,

$$\tilde{y}_{i,m} = -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}$$

que l'on dénomme les *pseudo-résidus*. Ils constituent les nouvelles cibles à approcher à chaque itération.

Le pseudo code ci-dessous résume l'algorithme de *Gradient Boosting* :

Algorithm 1 Gradient Boosting

Require: Données d'entraînement $\{(x_i, y_i)\}_{i=1}^n$, nombre d'itérations M

1: **But :** minimiser $F^* = \arg\min_{F \in \mathcal{F}} \sum_{i=1}^{n} L(y_i, F(x_i))$

2: **Initialisation :** $F_0(x) = \arg\min_\gamma \sum_{i=1}^n L(y_i, \gamma)$ {meilleure constante d'approximation}

3: **for** $m = 1$ **to** M **do**

4: Calculer les pseudo-résidus pour $i = 1, \ldots, n$:

$$\tilde{y}_{i,m} \leftarrow -\frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}$$

5: Estimer les paramètres de l'apprenant faible :

$$a_m \leftarrow \arg\min_{a,\beta} \sum_{i=1}^{n} (\tilde{y}_{i,m} - \beta h(x_i; a))^2$$

6: Calculer le pas optimal :

$$\beta_m \leftarrow \arg\min_{\beta} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \beta h(x_i, a_m))$$

7: Mettre à jour le modèle :

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m)$$

8: **end for**

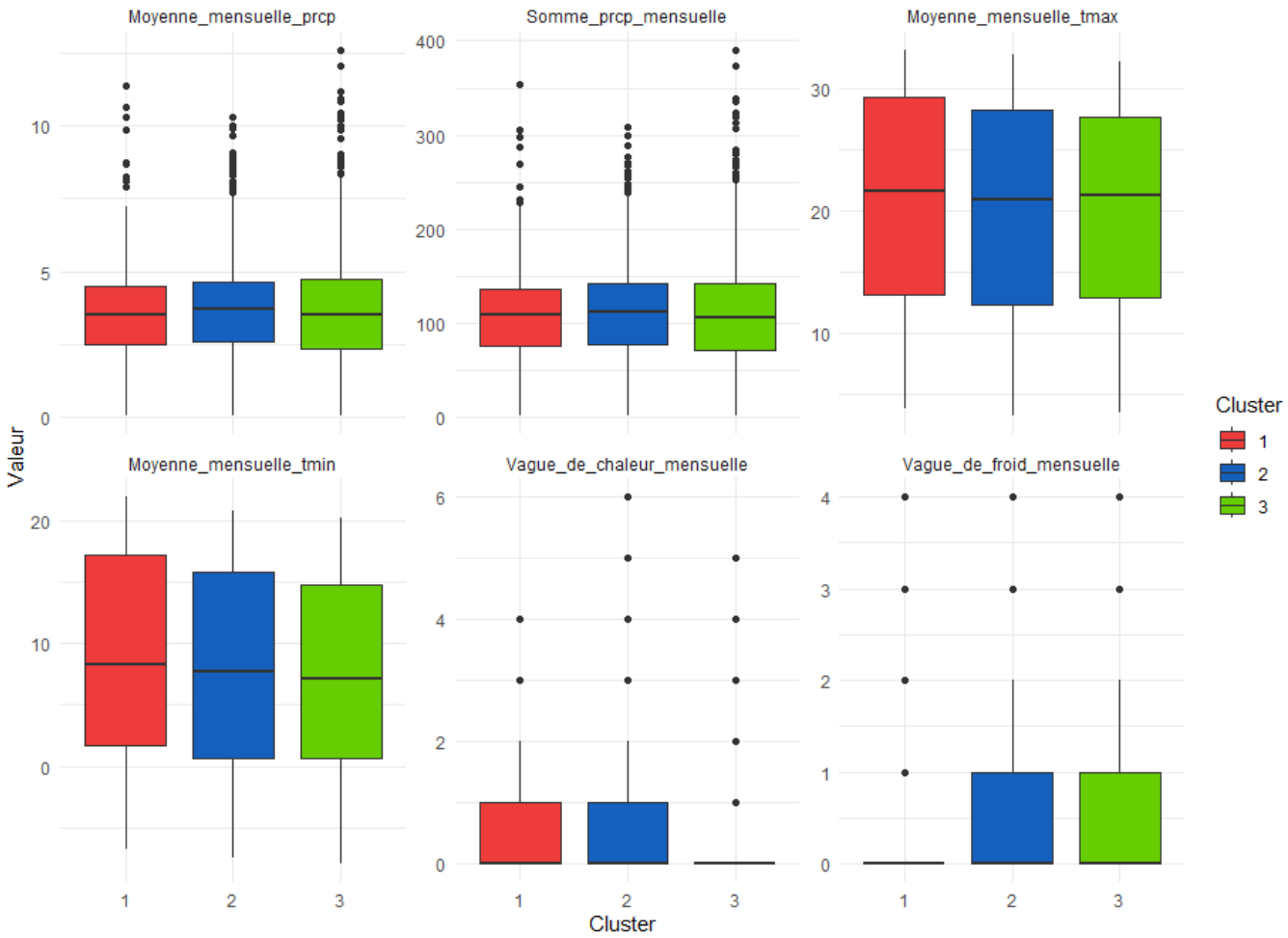
9: **Sortie :** F_M

B.3.3 *Gradient Tree Boosting*

Dans le cas du *Gradient Tree Boosting*, la classe $\mathcal{H}$ contient l'ensemble des arbres prédicteurs CART (voir annexe B.1).

Annexe C

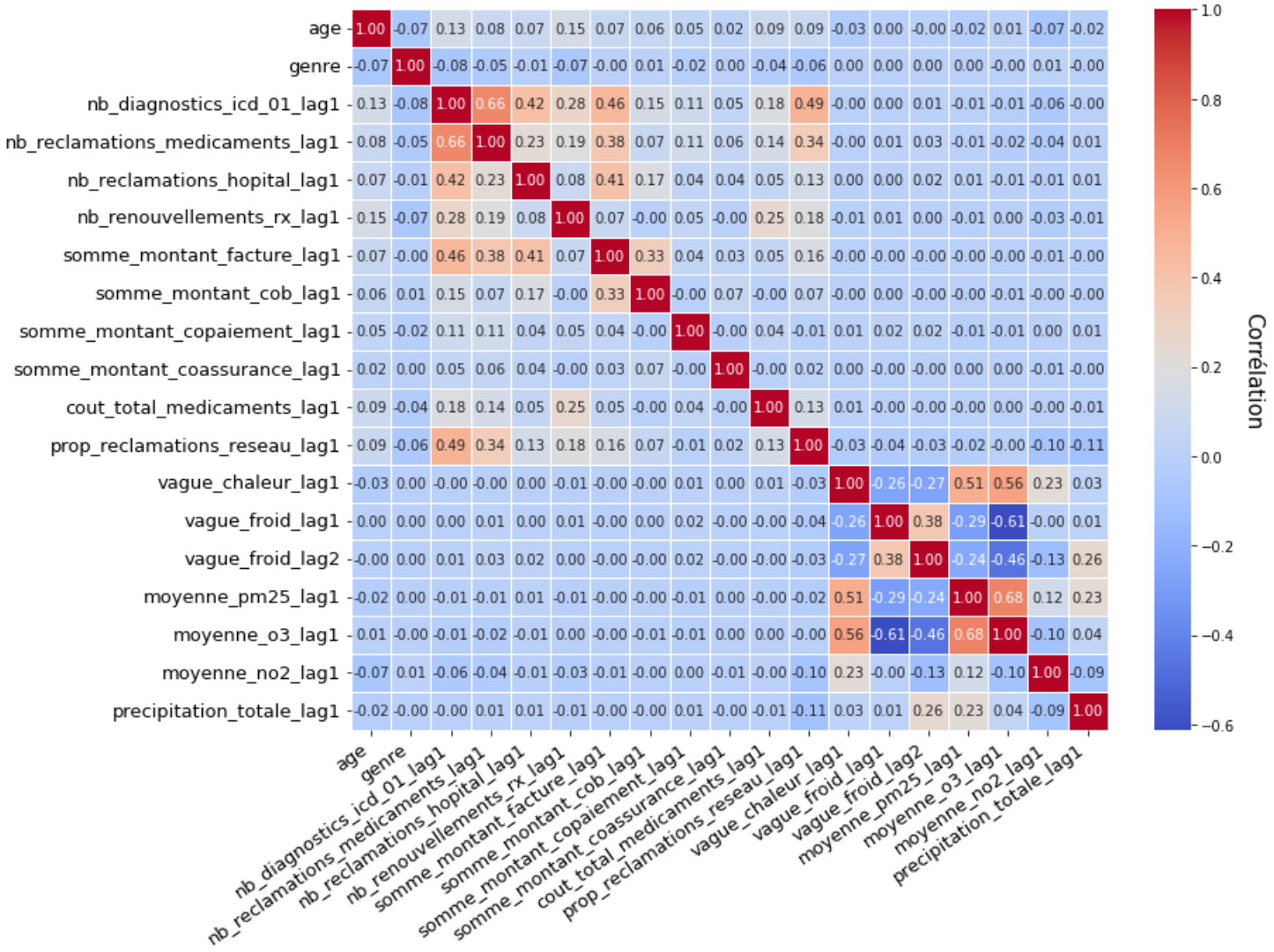
Distribution des différents indicateurs mensuels par cluster sous forme de Boxplots



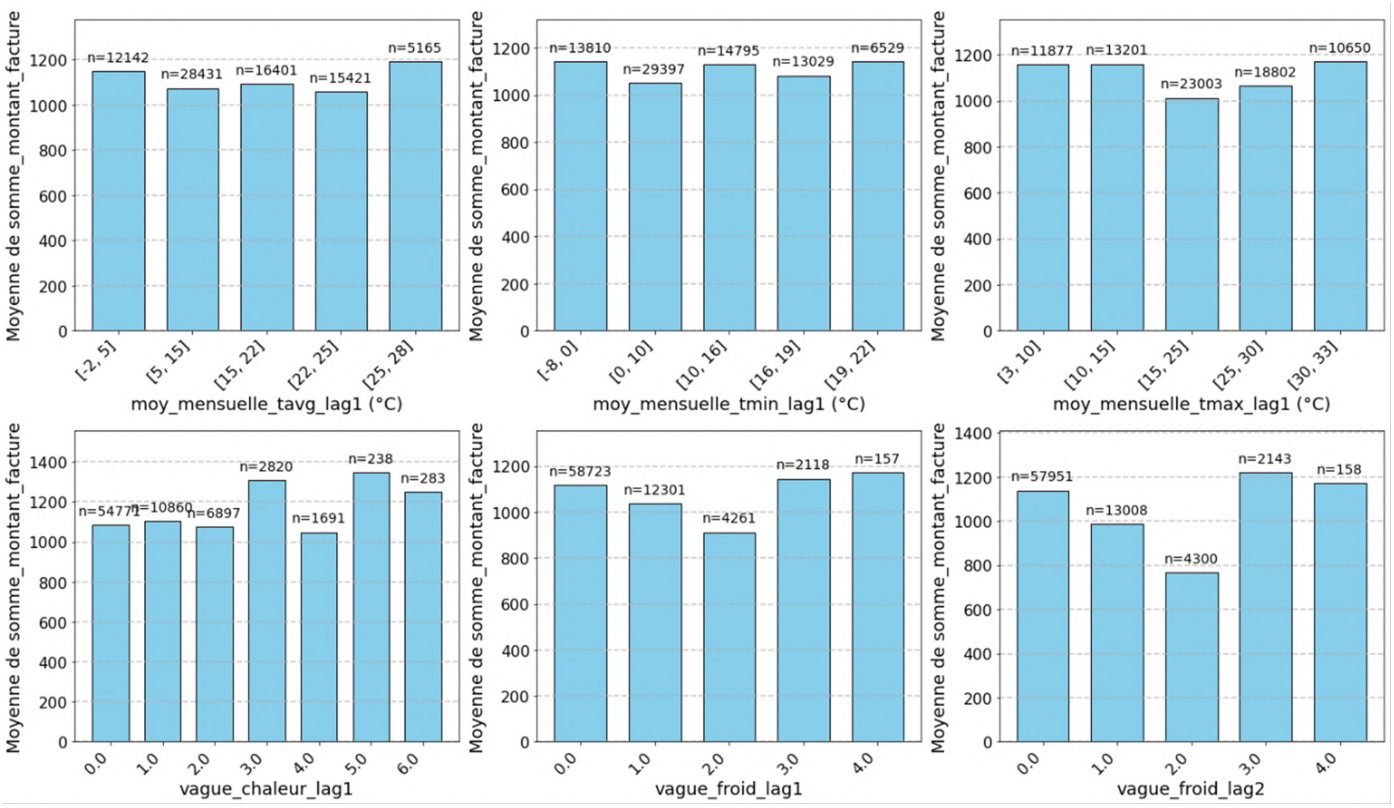
Annexe D

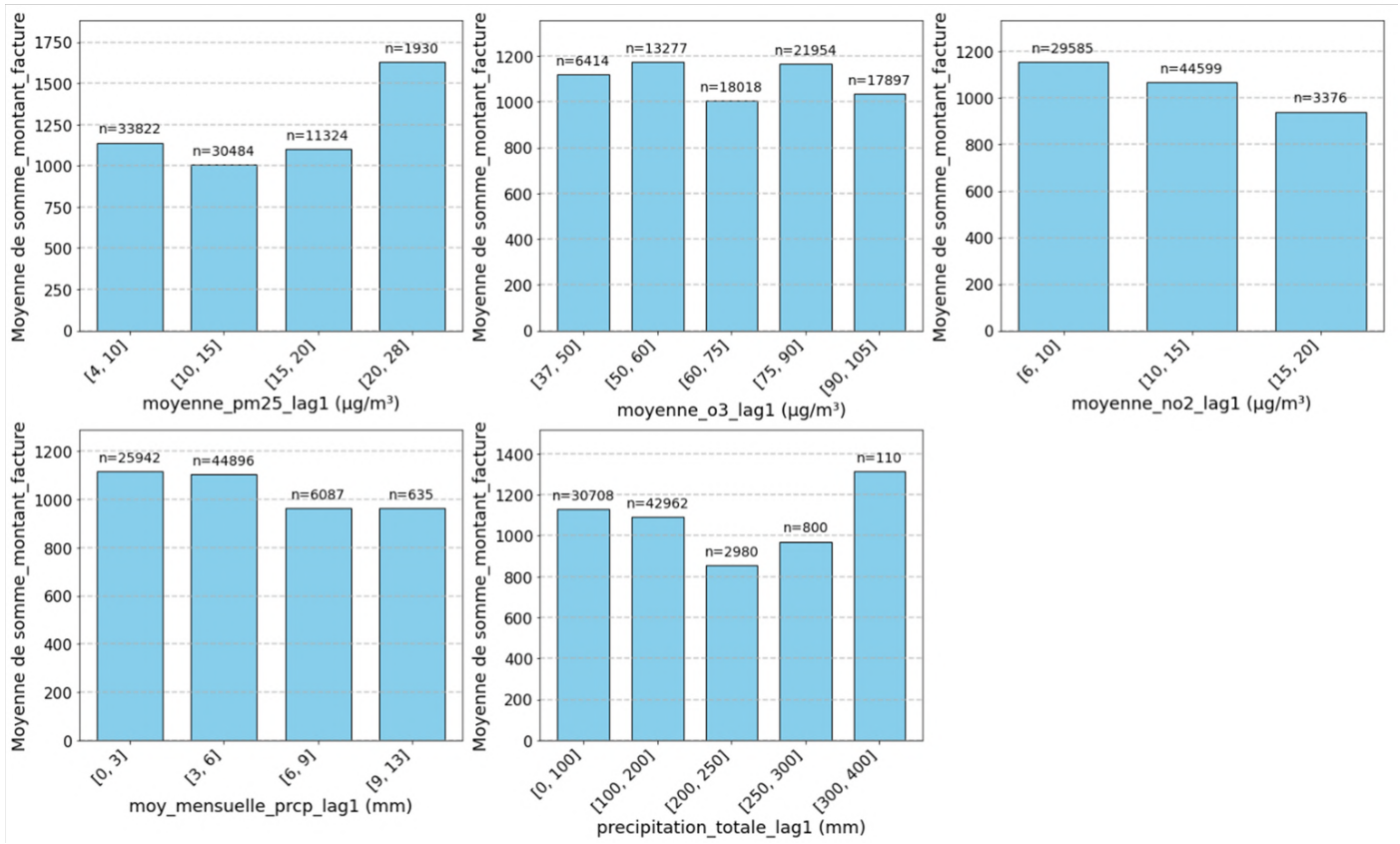
Compléments statistiques sur le score mensuel

D.1 Matrice de corrélation après sélection des variables

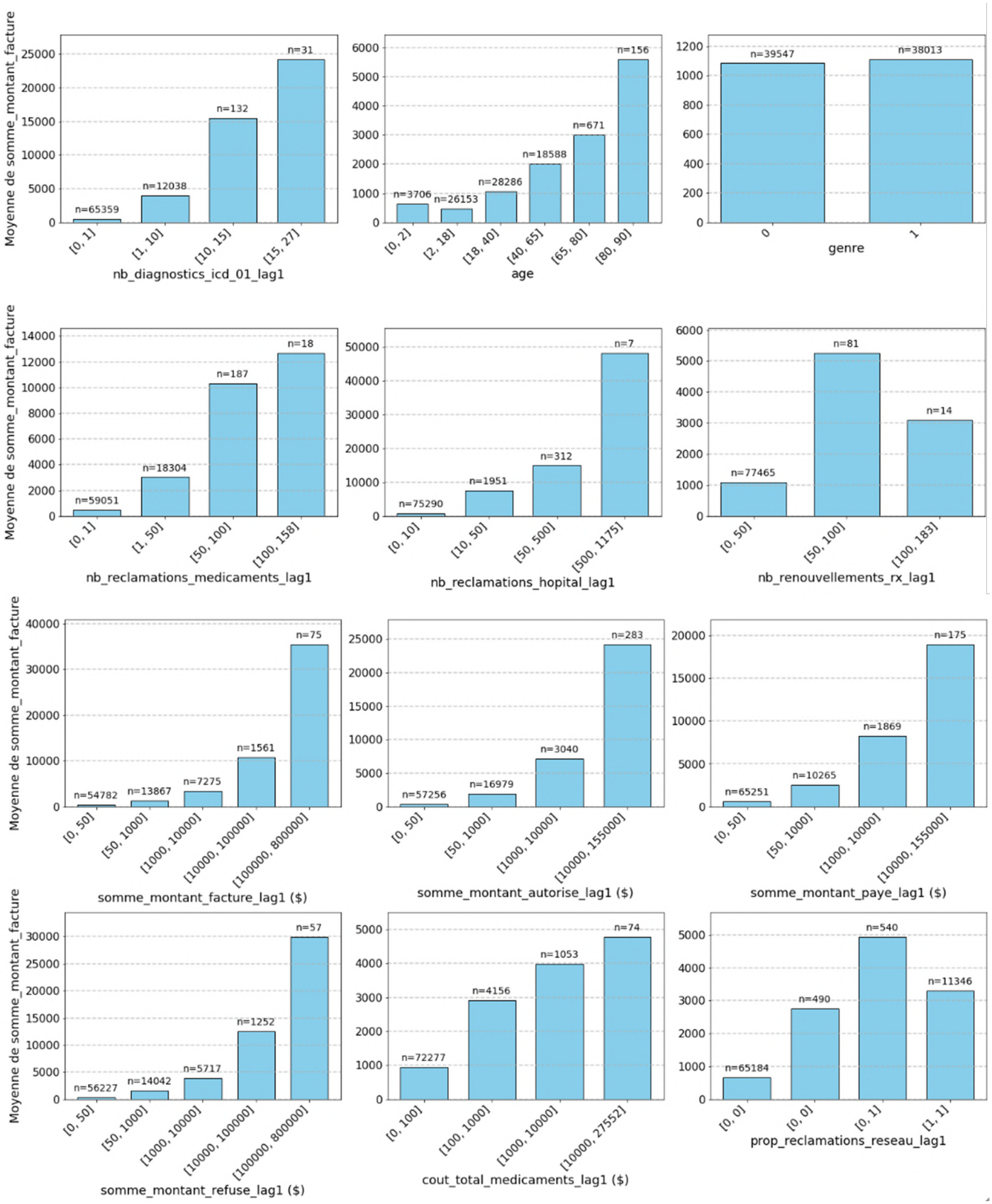


D.2 Statistiques descriptives des variables environnementales





D.3 Statistiques descriptives des variables personnelles et de santé

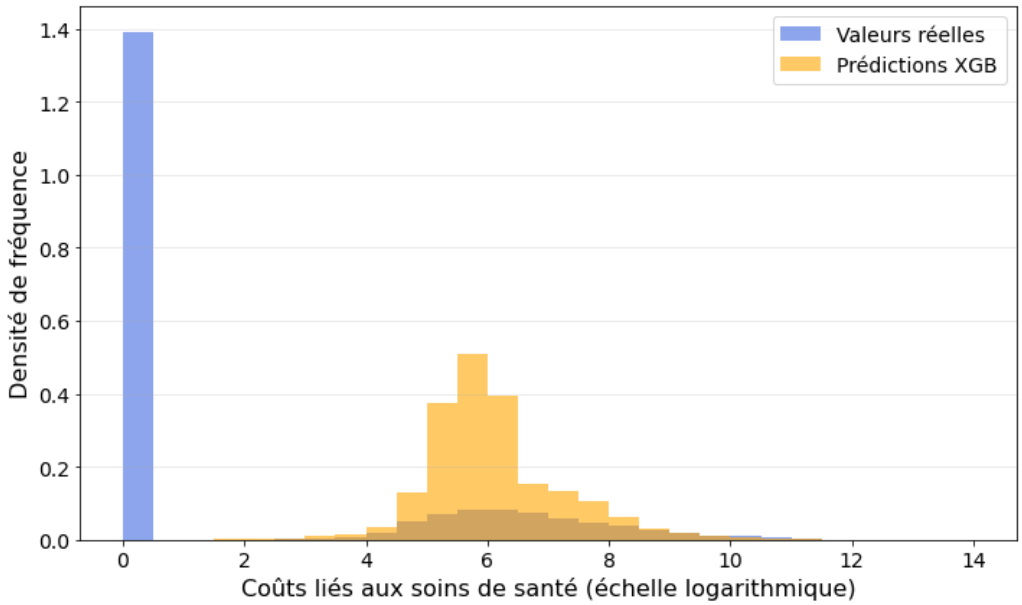


Annexe E

Compléments sur les modèles de scores basés sur la sinistralité

Dans un premier temps, le modèle XGBoost a été optimisé sur la variable cible sans transformation. Les résultats sont peu satisfaisants au vu de la valeur des métriques de performance (RSME de 8 957 $ et MAE de 1 570 $) et de la figure ci-dessous qui témoigne de lacunes en termes de prédictions. En effet, le modèle ne parvient à capter la grosse majorité des frais mensuels qui s'avèrent être nuls. Ce modèle ne peut donc pas être sélectionné car il crée des scores mensuels qui ne sont pas représentatifs de la santé réelle des assurés.

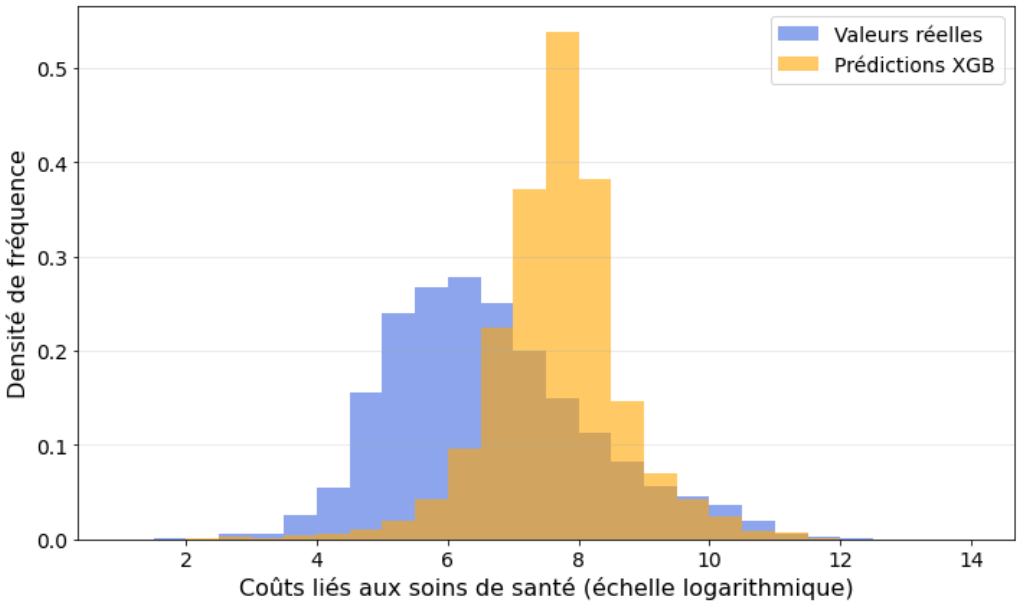
Comparaison des distributions de la variable cible et des valeurs prédites à l'aide du modèle XGBoost



Pour pallier ces problèmes de prédictions asymétriques, un modèle XGBoost a été optimisé sur la variable cible, ôtée des valeurs nulles. La RMSE obtenue est égale à 19 794 $ tandis que la MAE s'élève à 4670 $. La figure ci-dessous compare les distributions de la variable cible (valeurs strictement positives) et des valeurs prédites à l'aide du modèle XGBoost. La distribution des prédictions est fortement décalée sur la droite par rapport à la distribution de la variable cible. Les frais de santé ainsi prédits sont trop élevés par

rapport à la réalité. Les métriques de performance confirment ce point : la RMSE, très élevée, suggère des écarts de prédiction conséquents. A nouveau, les résultats ne sont pas concluants.

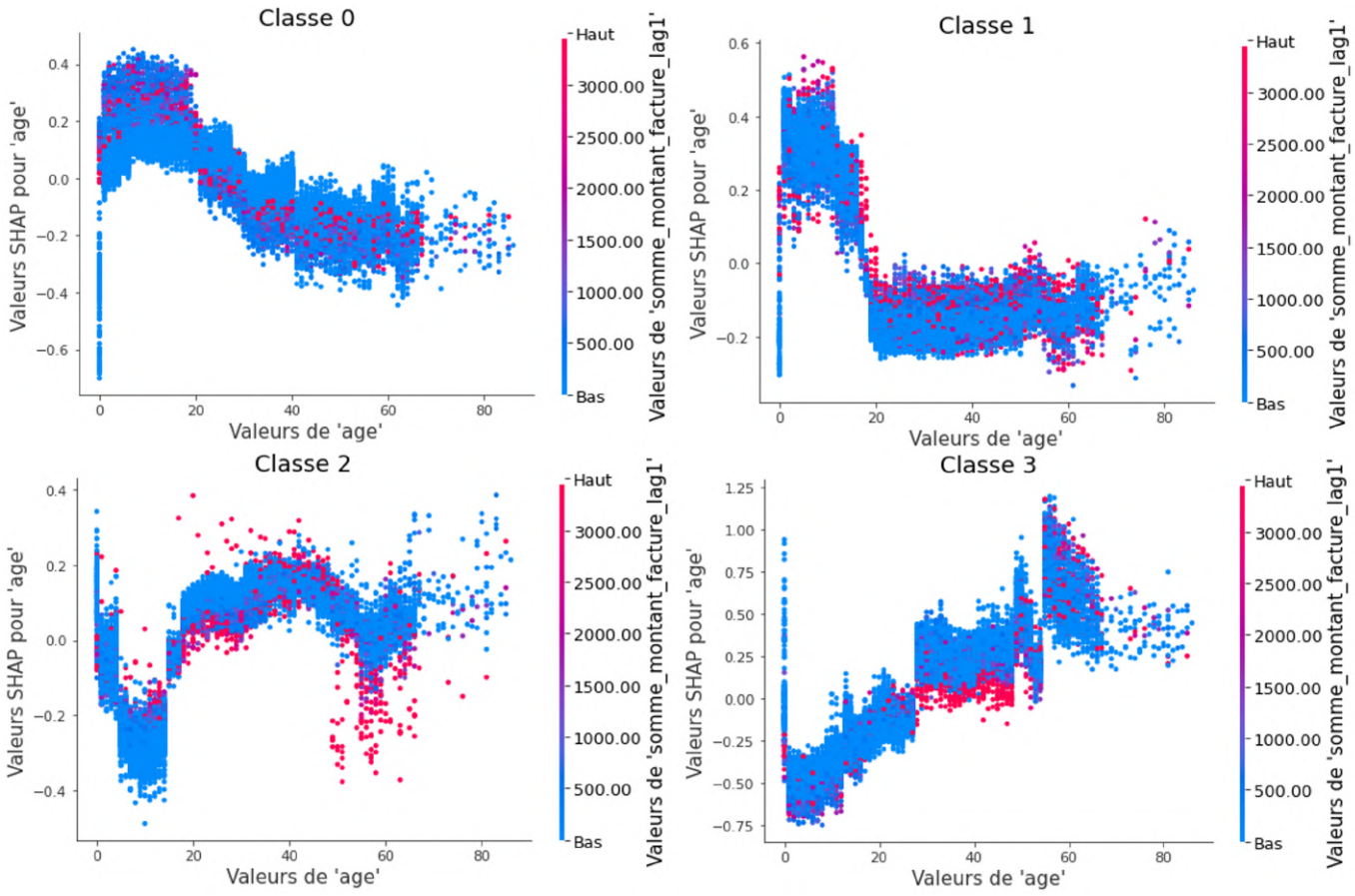
Comparaison des distributions de la variable cible et des valeurs prédites à l'aide du modèle XGBoost (uniquement sur les valeurs strictement positives)



Ainsi, les deux modèles ne parviennent pas à intégrer convenablement l'asymétrie de la variable cible. Il est ainsi nécessaire de discrétiser les frais mensuels afin de trouver des modèles permettant de créer des scores fiables et pertinents.

Annexe F

LGBM - Graphique SHAP : analyse croisée de l'importance de l'âge et des frais de santé antérieurs



Annexe G

Matrices de confusion par groupe ethnique issues du modèle XGBoost (prédiction des classes de coûts)