L'effet de l'arsenic aéroporté PM2.5 sur les maladies respiratoires chroniques : une analyse des données provenant des États-Unis

Mémoire

José Manuel Rodríguez Caballero

Sous la direction de:

M'Hamed Lajmi Lakhal Chaieb
Karim Barigou

Table des matières

Liste des tableaux

Liste des figures

Chapitre 1

État de l'Art

1.1 Introduction et importance de la problématique

La pollution de l'air est reconnue depuis plusieurs décennies comme un enjeu majeur de santé publique. Elle est associée à un large éventail d'effets néfastes sur la santé, dont une proportion importante de pathologies respiratoires chroniques (bronchites chroniques, asthme, cancer du poumon, etc.) (Bang et al., 2015; Donaldson et al., 2010; Mazurek et al., 2017). Les maladies respiratoires liées à l'exposition chronique ou aiguë à des polluants atmosphériques constituent en effet un défi épidémiologique et socio-économique, car elles mobilisent des ressources médicales importantes et engendrent des pertes de productivité.

Les études épidémiologiques mettent en évidence l'existence d'une corrélation robuste entre l'exposition à divers polluants de l'air (particules fines $PM_{2.5}$ et PM_{10}, gaz, fibres, etc.) et la survenue ou l'aggravation de maladies respiratoires (Cho et al., 2011; Gomes et al., 2014). Au-delà de la corrélation, l'établissement d'un lien de causalité exige une modélisation statistique et mécanistique précise, prenant en compte la dynamique d'exposition, les mécanismes biologiques sous-jacents et la présence de facteurs confondants (âge, statut socio-économique, co-expositions, etc.).

Parmi les travaux de grande ampleur, le *National Morbidity Mortality Air Pollution Study (NMMAPS)* illustre l'importance des méthodes novatrices, de leur implémentation dans R, et des outils de recherche reproductible pour produire des résultats solides (Dominici et al., 2003; Samet et al., 2000a,b,c; Dominici et al., 2002; Bell et al., 2004a; Peng et al., 2005; Dominici et al., 2007). Plus généralement, l'épidémiologie de la pollution atmosphérique recourt à quatre principaux types de plans d'études :

1. les *études temporelles* (*time series* dites "écologiques"),
2. les études de type *case-crossover*,
3. les études de panel,

4. les études de cohorte.

Nous présentons ci-dessous ces designs, leurs modèles statistiques habituels, ainsi que quelques exemples d'applications.

1.1.1 Études temporelles (Time Series Studies)

Les *études temporelles* relient les expositions et les issues de santé au cours du temps en exploitant des données agrégées au plan quotidien (p. ex. nombre de décès ou d'hospitalisations) et des niveaux de pollution journaliers mesurés par des stations fixes (Bell et al., 2004b). Ces travaux s'appuient généralement sur des modèles de régression, de type *Generalized Linear Models (GLM)* ou *Generalized Additive Models (GAM)*, afin d'estimer l'effet de la pollution sur la mortalité ou la morbidité tout en prenant en compte les fluctuations lentes (saisonnalité, tendance à long terme) via l'utilisation de fonctions de lissage (p. ex. splines cubiques ou loess) (McCullagh and Nelder, 1989; Hastie and Tibshirani, 1990).

Le *NMMAPS* constitue la plus vaste étude de ce type à ce jour (Dominici et al., 2003; Samet et al., 2000a,b,c; Dominici et al., 2007; Bell et al., 2004a; Peng et al., 2005; Dominici et al., 2002). Contrairement à la plupart des études temporelles antérieures, souvent mono-centriques, le *NMMAPS* inclut des données de dizaines de villes américaines et fournit des estimations spécifiques à chaque localité, à l'échelle régionale et nationale, notamment pour les particules inhalables (PM_{10}) et la mortalité. Des modèles hiérarchiques ("multiniveaux") y consolident les estimations issues de différentes localités (Dominici et al., 2002).

1.1.2 Études de type Case-Crossover

La conception *case-crossover* (Maclure, 1991; Jaakkola, 2003; Maclure and Mittleman, 2000) a été développée pour étudier des effets aigus et transitoires d'expositions intermittentes (Breslow et al., 1982; Schlesselman, 1982). Chaque individu qui développe l'événement (le "cas") sert de son propre témoin ("contrôle"), ce qui permet de diminuer l'impact des variables de confusion invariantes dans le temps (âge, sexe, tabagisme, etc.). Dans les études de pollution atmosphérique, ce design se révèle particulièrement adapté lorsque l'exposition présente de fortes variations journalières et que l'issue de santé (par exemple, une crise d'asthme) est aiguë (Maclure and Mittleman, 2000; Jaakkola, 2003).

Deux biais majeurs ont été identifiés :

1. la présence de tendance à long terme et de saisonnalité dans les concentrations de polluants, qui viole l'hypothèse de stationnarité de la série temporelle (Navidi, 1998; Bateson and Schwartz, 1999; Lumley and Levy, 2000; Bateson and Schwartz, 2001; Levy et al., 2001a),

2. l'"overlap bias" (recouvrement des périodes de référence) (AUSTIN et al., 1989; Lumley and Levy, 2000; Janes et al., 2005a,b). Malgré cela, plusieurs études mono-centriques et multi-centriques ont mis en œuvre cette approche (Lee and Schwartz, 1999; Neas et al., 1999; Peters et al., 2001; Levy et al., 2001b; D'Ippoliti et al., 2003; Symons et al., 2006; Zanobetti and Schwartz, 2005; Barnett et al., 2005, 2006; Medina-Ramón et al., 2006).

Par exemple, Levy et al. (Levy et al., 2001b) n'ont pas observé d'effet significatif d'une hausse à court terme des particules PM sur le risque d'arrêt cardiaque soudain chez des sujets sans antécédent cardiovasculaire, tandis que Peters et al. (Peters et al., 2004) ont mis en évidence un lien entre la pollution liée au trafic et la survenue d'infarctus du myocarde chez des sujets à risque.

1.1.3 Études de panel

Les *études de panel* consistent à suivre un même groupe d'individus pendant une période de temps, en mesurant régulièrement l'exposition et l'issue de santé (Peters et al., 1999b,a; Gauderman et al., 2002). Elles sont adaptées à l'étude d'effets aigus dans des sous-populations susceptibles (enfants, personnes âgées, patients atteints de BPCO, etc.). Les mesures de santé peuvent être répétées (p. ex. plusieurs mesures de la fonction respiratoire) et l'exposition peut être évaluée via des capteurs individuels ou des stations fixes. L'un des exemples les plus connus est la *Children's Health Study* en Californie du Sud (Peters et al., 1999a,b; Gauderman et al., 2002), qui a utilisé des modèles mixtes (*linear mixed models*) pour mettre en évidence un lien entre l'exposition chronique aux particules fines, au NO_2 et la réduction de la croissance pulmonaire chez l'enfant (Gauderman et al., 2002).

1.1.4 Études de cohorte

Les *études de cohorte* (Bell et al., 2004b) évaluent l'impact d'une exposition chronique ou cumulée sur la santé (p. ex. la mortalité à long terme). Elles peuvent être prospectives ou rétrospectives. Les cohortes multicentriques, couvrant plusieurs zones géographiques, permettent de mieux capturer la variabilité d'exposition mais génèrent davantage de facteurs de confusion potentiels (socio-économiques, culturels, etc.). L'approche statistique standard consiste alors à utiliser des modèles de survie, notamment le *modèle de Cox*.

Le "Harvard Six Cities Study" et l'"American Cancer Society (ACS) Study" illustrent ces approches (Dockery et al., 1993; Pope et al., 1995; Pope, 2007). Dans la première, un échantillon de plus de 8000 adultes de six villes américaines a été suivi pendant 14 à 16 ans, montrant une augmentation du risque de mortalité liée à l'exposition aux particules fines (Dockery et al., 1993). L'ACS Study, couvrant 151 zones métropolitaines, a confirmé cette association et suggéré un risque accru de mortalité avec l'élévation des particules fines (Pope et al., 1995). Des réanalyses ultérieures, intégrant de nouvelles covariables spatiales et des modèles

d'autocorrélation spatiale, corroborent l'effet délétère à long terme de la pollution (Laden et al., 2006; Krewski et al., 2009; Pope, 2007).

1.2 Comparaison des plans d'étude

Le choix d'un design d'étude dépend étroitement du type d'effet recherché (aigu vs. chronique), de la disponibilité des données, du type de population et de l'issue de santé (Vedal, 1997). Les études de panel, de cohorte et de type *case-crossover* permettent de mieux prendre en compte les facteurs individuels potentiellement modificateurs de l'effet, tandis que les *time series studies* s'appuient sur des données agrégées (Wakefield and Salway, 2001). En résumé :

- **Effets aigus :** Les études *time series*, *case-crossover* et *de panel* sont adéquates pour détecter l'influence de variations journalières de la pollution.
- **Effets chroniques :** Les études de cohorte, plus coûteuses, permettent d'évaluer l'effet d'une exposition cumulative.
- **Modélisation de la tendance et de la saisonnalité :** Les *time series studies* intègrent ces facteurs via des fonctions de lissage dans les GLM/GAM. Les *case-crossover* les contrôlent par conception (périodes de référence).
- **Prise en compte des caractéristiques individuelles :** Les *case-crossover*, *études de panel* et *de cohorte* intègrent plus facilement des covariables individuelles.

Aucun plan d'étude n'est donc universellement supérieur : ils sont complémentaires en fonction des questions posées.

1.3 Comportement non linéaire de la relation dose-réponse

Selon Louis Anthony Cox Jr, la relation entre l'exposition à un polluant et le risque de développer une pathologie respiratoire ne se limite pas à un simple modèle linéaire (Cox Jr, 2021). Divers processus biologiques (activation de l'inflammasome, saturation des voies de clairance, etc.) induisent des réponses non linéaires, voire seuilaires. Dans de nombreuses études, les particules fines $PM_{2.5}$ et et les particules inhalables PM_{10}, les gaz et les fibres sont associées à des dommages respiratoires (Bang et al., 2015; Donaldson et al., 2010; Mazurek et al., 2017).

1.3.1 Limites des approches linéaires sans seuil (LNT)

Les modèles de type linéaire sans seuil (LNT) ont longtemps occupé une place centrale, en particulier pour leur simplicité et leur caractère "conservateur" (Belkebir et al., 2011; Jane Ellen Simmons and Boyes, 2005). Cependant, plusieurs réserves s'imposent :

- **Saturation des mécanismes de clairance :** Au-delà d'un certain niveau d'exposition, les macrophages alvéolaires peuvent être débordés, entraînant une inflammation chronique (Tran et al., 2001; Alisa DeStefano and Wallace, 2017).

— **Seuils biologiques :** L'activation de l'inflammasome NLRP3 survient dès qu'un niveau critique d'agression cellulaire est franchi (Groslambert and Py, 2018; Sayan and Mossman, 2015).
— **Rôle des pics d'exposition :** Les expositions brèves mais intenses peuvent avoir un effet plus marqué qu'une même dose cumulée plus étalée (Louis Anthony , Tony).

1.4 De l'épidémiologie à la pharmacocinétique

1.4.1 Modèles pharmacocinétiques et pharmacodynamiques (PBPK/PD)

Les modèles PBPK (Physiologically Based Pharmacokinetic) et PD (Pharmacodynamic) décrivent la distribution d'un composé et ses effets biologiques dans l'organisme. Appliqués à la toxicologie de la silice ou de l'amiante, ils soulignent le caractère crucial de la dynamique temporelle de l'exposition (Tran et al., 2001; Alisa DeStefano and Wallace, 2017) :

— **Silice cristalline (RCS) :** Tran et al. (2001) identifient un plateau lorsque la clairance atteint un équilibre avec l'accumulation. Au-delà, la réponse inflammatoire devient chronique.
— **Amiante :** Alisa DeStefano and Wallace (2017) suggèrent qu'à dose cumulée identique, des pics d'exposition plus prononcés provoquent des dépôts de fibres supérieurs à une exposition plus uniforme.

Ces résultats illustrent l'importance de modéliser la pharmacocinétique et la pharmacodynamique pour mieux évaluer le risque.

1.4.2 Modèles mécanistiques de l'inflammation

Le rôle de l'inflammasome NLRP3 dans l'inflammation persistante est de plus en plus documenté (Donaldson et al., 2010; Sayan and Mossman, 2015). Ce complexe protéique agit comme un "seuil biologique" :

— **Comportement tout-ou-rien :** L'assemblage du NLRP3 dépend de signaux multiples et d'une quantité critique de particules (Groslambert and Py, 2018).
— **Impact sur la réglementation :** Des pics d'exposition brefs mais intenses peuvent suffire à initier une inflammation auto-entretenue, imposant de repenser les normes sanitaires pour prendre en compte ces épisodes (Louis Anthony , Tony).

1.5 Applications à la pollution de l'air ambiant et enjeux de santé publique

Au-delà des milieux professionnels, la pollution urbaine (trafic routier, industrie, chauffage) contribue à l'augmentation des admissions hospitalières pour pathologies respiratoires (Cho et al., 2011; Gomes et al., 2014). Les premières approches statistiques étaient principalement

basées sur des modèles de séries temporelles liant les variations journalières de la pollution à la mortalité ou aux hospitalisations (Kim et al., 2003; Anthony Cox et al., 1996). Cependant, des recherches plus récentes insistent sur l'importance de capturer la dynamique fine de l'exposition et le rôle déterminant des pics (Gottschalk et al., 2016; Sayan and Mossman, 2015).

1.6 Critique de la modélisation par régression

Dans ses travaux, Louis Anthony Cox Jr souligne les limites des modèles de régression "réduits" (linéaires ou logistiques) pour établir un lien de causalité entre la pollution et les maladies respiratoires (Cox Jr, 2021). La présence de variables confondantes (facteurs socio-économiques, co-expositions, etc.) et de corrélations spurielles fragilise la validité des conclusions (Cui et al., 2003; Sneeringer, 2009; Selvin et al., 1984).

1.6.1 Modèles linéaires, logistiques et spatio-temporels

Si les modèles linéaires ou logistiques sont simples et largement utilisés (Cui et al., 2003), ils présentent plusieurs inconvénients (Fewell et al., 2007; JACOBS et al., 1979; Christenfeld et al., 2004; W-D., 2006) :

- **Manque de modélisation causale :** La chaîne complète (émission → transport → dose interne → effet) reste souvent implicite.
- **Dépendance au choix des covariables :** L'omission ou l'ajout de variables peut modifier la taille (voire le signe) de l'effet de la pollution (Sneeringer, 2009).
- **Risques de régression fallacieuse :** Des séries temporelles avec tendances communes peuvent produire des liens spurious.

Les modèles à effets mixtes (modèles hiérarchiques) (Wakefield, 2009) et les approches spatio-temporelles réduisent les biais liés à la structure des données, mais n'éliminent pas toutes les confusions causales ni les hypothèses fortes.

1.6.2 Modèles structuraux, réseaux bayésiens et QRA

Pour dépasser ces limites, diverses approches structurales (modèles d'équations structurelles, réseaux bayésiens) ont été proposées (Pearl, 2009; Shipley, 2016; Ellis and Wong, 2008). Elles formalisent la chaîne causale (émission, dispersion, exposome, dose-réponse) et intègrent des variables latentes. En parallèle, la *Quantitative Risk Assessment* (QRA) offre un cadre plus complet pour étudier chaque étape (émission, dispersion, exposome, dose-réponse, impact sanitaire) (Greenland and Brumback, 2002; Cox Jr., 2005; Jhih-Shyang Shih and Siikamäki, 2008). Leur principal inconvénient demeure la nécessité de données riches et d'expertise multidisciplinaire, mais elles renforcent la robustesse et la pertinence des conclusions (Greenland, 2001; Lucas, 1976).

1.7 Identifications des lacunes et pistes de recherche

Malgré des avancées considérables, plusieurs lacunes subsistent dans la littérature traitant de l'effet de la pollution de l'air sur les maladies respiratoires :

1. **Variabilité inter-individuelle :** Les différences génétiques, épigénétiques ou liées au mode de vie (tabagisme, comorbidités) sont souvent intégrées comme simples covariables, sans modéliser pleinement leur complexité.
2. **Rôle des pics et des durées d'exposition :** Peu d'études proposent une modélisation fine des épisodes aigus (forte pollution), alors qu'ils peuvent déclencher des réponses inflammatoires non linéaires.
3. **Couplage entre mécanismes biologiques et statistiques :** Les approches PBPK/PD sont prometteuses mais nécessitent des données toxicologiques précises, rarement disponibles à grande échelle. Les modèles purement statistiques risquent, eux, de négliger la saturation de certaines voies biologiques.

1.7.1 Potentiel des modèles de mélange gaussien

Une piste de recherche consiste à utiliser des **modèles de mélange gaussien** (GMM) pour mieux rendre compte de l'hétérogénéité de la population vis-à-vis de la pollution :

— **Identification de sous-groupes :** Les individus peuvent être classés selon leur sensibilité, certains étant plus vulnérables.

— **Modélisation d'effets non linéaires :** Les distributions de l'exposition ou de la réponse peuvent présenter plusieurs modes, correspondant à différents profils de réactivité.

— **Couplage avec des approches hiérarchiques ou mécanistiques :** Les GMM peuvent se combiner à des réseaux bayésiens ou à des modèles PBPK/PD pour capturer la diversité des mécanismes biologiques.

1.8 Conclusion

La modélisation de l'impact de la pollution de l'air sur les maladies respiratoires a considérablement évolué, depuis les premières approches linéaires sans seuil (LNT) jusqu'aux méthodologies plus avancées en épidémiologie, en toxicologie et en statistique. Les principaux enseignements issus de la littérature actuelle soulignent :

— Le **caractère non linéaire** et potentiellement **seuilaire** de la relation dose-réponse, lié à des processus biologiques (activation de l'inflammasome, saturation de la clairance).

— L'importance des **pics d'exposition**, qui peuvent exercer un impact disproportionné par rapport à la dose cumulée.

— La **nécessité de modèles causaux plus riches** (approches mécanistiques, réseaux bayésiens, QRA) pour identifier les leviers d'action et informer au mieux les politiques de santé publique.

De nombreuses pistes de recherche restent ouvertes : prise en compte de la variabilité inter-individuelle par des modèles de mélange, modélisation plus fine des épisodes aigus, intégration plus profonde des connaissances biologiques dans les approches statistiques, etc. L'enjeu final demeure d'offrir des outils de prédiction et d'aide à la décision fiables, afin de limiter l'impact sanitaire d'une pollution de l'air qui demeure un enjeu majeur de santé publique.

Bibliographie

Clyde F. Martin Alisa DeStefano and Dorothy I. Wallace. A dynamical model of the transport of asbestos fibres in the human body. *Journal of Biological Dynamics*, 11(1) :365–377, 2017. doi : 10.1080/17513758.2017.1355489. URL https://doi.org/10.1080/17513758.2017.1355489. PMID : 28770658.

Louis Anthony Cox, Michael G. Bird, and Larry Griffis. Isoprene cancer risk and the time pattern of dose administration. *Toxicology*, 113(1) :263–272, 1996. ISSN 0300-483X. doi : https://doi.org/10.1016/0300-483X(96)03455-5. URL https://www.sciencedirect.com/science/article/pii/0300483X96034555. Evaluation of Butadiene and Isoprene Health Risks.

HARLAND AUSTIN, W DANA FLANDERS, and KENNETH J ROTHMAN. Bias Arising in Case-Control Studies from Selection of Controls from Overlapping Groups. *International Journal of Epidemiology*, 18(3) :713–716, 09 1989. ISSN 0300-5771. doi : 10.1093/ije/18.3.713. URL https://doi.org/10.1093/ije/18.3.713.

Ki Moon Bang, Jacek M Mazurek, John M Wood, Gretchen E White, Scott A Hendricks, Ainsley Weston, Centers for Disease Control, Prevention (CDC), et al. Silicosis mortality trends and new exposures to respirable crystalline silica—United States, 2001–2010. *MMWR Morb Mortal Wkly Rep*, 64(5) :117–20, 2015.

Adrian G. Barnett, Gail M. Williams, Joel Schwartz, Anne H. Neller, Terri L. Best, Anton L. Petroeschevsky, and Rodney W. Simpson. Air pollution and child respiratory health : a case-crossover study in Australia and New Zealand. *American Journal of Respiratory and Critical Care Medicine*, 171(11) :1272–1278, June 2005. doi : 10.1164/rccm.200411-1586OC.

Adrian G. Barnett, Gail M. Williams, Joel Schwartz, Trudi L. Best, Anne H. Neller, Anna L. Petroeschevsky, and Rod W. Simpson. The effects of air pollution on hospitalizations for cardiovascular disease in elderly people in australian and new zealand cities. *Environmental Health Perspectives*, 114(7) :1018–1023, 2006. doi : 10.1289/ehp.8674. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.8674.

Thomas F. Bateson and Joel Schwartz. Control for Seasonal Variation and Time Trend in Case-Crossover Studies of Acute Effects of Environmental Exposures. *Epidemiology*, 10(5) : 539–544, 1999. ISSN 10443983. URL http://www.jstor.org/stable/3703342.

Thomas F. Bateson and Joel Schwartz. Selection Bias and Confounding in Case-Crossover Analyses of Environmental Time-Series Data. *Epidemiology*, 12(6) :654–661, November 2001.

Emel Belkebir, Christophe Rousselle, Cédric Duboudin, Laurent Bodin, and Nathalie Bonvallot. Haber's rule duration adjustments should not be used systematically for risk assessment in public health decision-making. *Toxicology Letters*, 204(2) :148–155, 2011. ISSN 0378-4274. doi : https://doi.org/10.1016/j.toxlet.2011.04.026. URL https://www.sciencedirect.com/science/article/pii/S0378427411001925.

Michelle L. Bell, Aidan McDermott, Scott L. Zeger, Jonathan M. Samet, and Francesca Dominici. Ozone and Short-term Mortality in 95 US Urban Communities, 1987-2000. *JAMA*, 292(19) :2372–2378, 11 2004a. ISSN 0098-7484. doi : 10.1001/jama.292.19.2372. URL https://doi.org/10.1001/jama.292.19.2372.

Michelle L Bell, Jonathan M Samet, and Francesca Dominici. Time-series studies of particulate matter. *Annu. Rev. Public Health*, 25(1) :247–280, 2004b.

N. E. Breslow, N. E. Day, and James J. Schlesselman. Statistical methods in cancer research. Volume 1 — the analysis of case-control studies. *Journal of Occupational Medicine*, 24(4) : 255–257, April 1982.

William C. S. Cho, Chung K. Kwan, Stephen Yau, Peter P. F. So, Patricia C. M. Poon, and Joseph S. K. Au. The role of inflammation in the pathogenesis of lung cancer. *Expert Opinion on Therapeutic Targets*, 15(9) :1127–1137, 2011. doi : 10.1517/14728222.2011.599801. URL https://doi.org/10.1517/14728222.2011.599801. PMID : 21751938.

Nicholas J. Christenfeld, Richard P. Sloan, Douglas Carroll, and Sander Greenland. Risk factors, confounding, and the illusion of statistical control. *Psychosomatic Medicine*, 66(6) : 868–875, Nov-Dec 2004. doi : 10.1097/01.psy.0000140008.70959.41.

Louis Anthony Cox Jr. *Quantitative risk analysis of air pollution health effects*. Springer, 2021.

Louis Anthony (Tony) Cox Jr. Some limitations of a proposed linear model for antimicrobial risk management. *Risk Analysis*, 25(6) :1327–1332, 2005. doi : https://doi.org/10.1111/j.1539-6924.2005.00703.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2005.00703.x.

Y. Cui, Z. F. Zhang, J. Froines, et al. Air pollution and case fatality of SARS in the People's Republic of China : an ecologic study. *Environmental Health*, 2 :15, 2003. doi : 10.1186/1476-069X-2-15. URL https://doi.org/10.1186/1476-069X-2-15.

Douglas W. Dockery, C. Arden Pope, Xiping Xu, John D. Spengler, James H. Ware, Martha E. Fay, Benjamin G. Ferris, and Frank E. Speizer. An association between air pollution and mortality in six u.s. cities. *New England Journal of Medicine*, 329(24) :1753–1759, 1993. doi : 10.1056/NEJM199312093292401. URL https://www.nejm.org/doi/full/10.1056/NEJM199312093292401.

F. Dominici, A. McDermott, M. Daniels, S. L. Zeger, and J. M. Samet. A Special Report to the Health Effects Institute on the Revised Analyses of the NMMAPS II Data., 2003.

Francesca Dominici, Jonathan M. Samet, and Scott L. Zeger. Combining Evidence on Air Pollution and Daily Mortality from the 20 Largest US Cities : A Hierarchical Modelling Strategy. *Journal of the Royal Statistical Society Series A : Statistics in Society*, 163 (3) :263–302, 01 2002. ISSN 0964-1998. doi : 10.1111/1467-985X.00170. URL https://doi.org/10.1111/1467-985X.00170.

Francesca Dominici, Roger D. Peng, Scott L. Zeger, Ronald H. White, and Jonathan M. Samet. Particulate Air Pollution and Mortality in the United States : Did the Risks Change from 1987 to 2000 ? *American Journal of Epidemiology*, 166(8) :880–888, 08 2007. ISSN 0002-9262. doi : 10.1093/aje/kwm222. URL https://doi.org/10.1093/aje/kwm222.

Ken Donaldson, Fiona A. Murphy, Rodger Duffin, et al. Asbestos, carbon nanotubes and the pleural mesothelium : a review of the hypothesis regarding the role of long fibre retention in the parietal pleura, inflammation and mesothelioma. *Particle and Fibre Toxicology*, 7 :5, 2010. doi : 10.1186/1743-8977-7-5. URL https://doi.org/10.1186/1743-8977-7-5.

Daniela D'Ippoliti, Francesco Forastiere, Carla Ancona, Nera Agabiti, Danilo Fusco, Paola Michelozzi, and Carlo A. Perucci. Air Pollution and Myocardial Infarction in Rome : A Case-Crossover Analysis. *Epidemiology*, 14(5) :528–535, September 2003. doi : 10.1097/01.ede.0000082046.22919.72.

Byron Ellis and Wing Hung Wong. Learning causal bayesian network structures from experimental data. *Journal of the American Statistical Association*, 103(482) :778–789, 2008. doi : 10.1198/016214508000000193. URL https://doi.org/10.1198/016214508000000193.

Zoe Fewell, George Davey Smith, and Jonathan A. C. Sterne. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies : A Simulation Study. *American Journal of Epidemiology*, 166(6) :646–655, 07 2007. ISSN 0002-9262. doi : 10.1093/aje/kwm165. URL https://doi.org/10.1093/aje/kwm165.

W. James Gauderman, Frank D. Gilliland, H. Vora, Edward Avol, Daniel Stram, Rob McConnell, Duncan Thomas, Fred Lurmann, Helene G. Margolis, Edward B. Rappaport, Kiros Berhane, and John M. Peters. Association between air pollution and lung function growth in southern california children : results from a second cohort. *American Journal of Respiratory and Critical Care Medicine*, 166(1) :76–84, Jul 2002. doi : 10.1164/rccm.2111021.

Mónica Gomes, Ana Luísa Teixeira, Ana Coelho, António Araújo, and Rui Medeiros. The role of inflammation in lung cancer. In Bharat B. Aggarwal, Bokyung Sung, and Subash Chandra Gupta, editors, *Inflammation and Cancer*, pages 1–23. Springer Basel, Basel, 2014. doi : 10.1007/978-3-0348-0837-8_1. URL https://doi.org/10.1007/978-3-0348-0837-8_1.

Rachel A. Gottschalk et al. Distinct nf-

κ

b and mapk activation thresholds uncouple steady-state microbe sensing from anti-pathogen inflammatory responses. *Cell Systems*, 2(6) :378–390, 2016.

Sander Greenland. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *International Journal of Epidemiology*, 30(6) :1343–1350, 12 2001. ISSN 0300-5771. doi : 10.1093/ije/30.6.1343. URL https://doi.org/10.1093/ije/30.6.1343.

Sander Greenland and Babette Brumback. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31(5) :1030–1037, 10 2002. ISSN 0300-5771. doi : 10.1093/ije/31.5.1030. URL https://doi.org/10.1093/ije/31.5.1030.

Marine Groslambert and Bénédicte F Py. Spotlight on the nlrp3 inflammasome pathway. *Journal of Inflammation Research*, 11 :359–374, 2018. doi : 10.2147/JIR.S141220. URL https://www.tandfonline.com/doi/abs/10.2147/JIR.S141220. PMID : 30288079.

T. J. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall/CRC, Boca Raton, FL, 1990.

JJK Jaakkola. Case-crossover design in air pollution epidemiology. *European Respiratory Journal*, 21(40 suppl) :81s–85s, 2003.

RODNEY L. JACOBS, EDWARD E. LEAMER, and MICHAEL P. WARD. Difficulties with testing for causation. *Economic Inquiry*, 17(3) :401–413, 1979. doi : https://doi.org/10.1111/j.1465-7295.1979.tb00538.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1465-7295.1979.tb00538.x.

Marina V. Evans Jane Ellen Simmons and William K. Boyes. Moving from external exposure concentration to internal dose : Duration extrapolation based on physiologically based pharmacokinetic derived estimates of internal dose. *Journal of Toxicology and Environmental Health, Part A*, 68(11-12) :927–950, 2005. doi : 10.1080/15287390590912586. URL https://doi.org/10.1080/15287390590912586. PMID : 16020185.

Holly Janes, Lianne Sheppard, and Thomas Lumley. Case–crossover analyses of air pollution exposure data : Referent selection strategies and their implications for bias. *Epidemiology*, 16(6) :717–726, November 2005a. doi : 10.1097/01.ede.0000181315.18836.9d.

Holly Janes, Lianne Sheppard, and Thomas Lumley. Overlap bias in the case-crossover design, with application to air pollution exposures. *Statistics in Medicine*, 24(2) :285–300, 2005b. doi : https://doi.org/10.1002/sim.1889. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.1889.

Karen Palmer Jhih-Shyang Shih, Dallas Burtraw and Juha Siikamäki. Air emissions of ammonia and methane from livestock operations : Valuation and policy options. *Journal of the Air & Waste Management Association*, 58(9) :1117–1129, 2008. doi : 10.3155/1047-3289.58.9.1117. URL https://doi.org/10.3155/1047-3289.58.9.1117.

Amy H Kim, Michael C Kohn, Abraham Nyska, and Nigel J Walker. Area under the curve as a dose metric for promotional responses following 2,3,7,8-tetrachlorodibenzo-p-dioxin exposure. *Toxicology and Applied Pharmacology*, 191(1) :12–21, 2003. ISSN 0041-008X. doi : https://doi.org/10.1016/S0041-008X(03)00225-4. URL https://www.sciencedirect.com/science/article/pii/S0041008X03002254.

Daniel Krewski, Michael Jerrett, Richard T Burnett, Renjun Ma, Edward Hughes, Yuanli Shi, Michelle C Turner, C Arden Pope III, George Thurston, Eugenia E Calle, et al. *Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality*, volume 140. Health Effects Institute Boston, MA, 2009.

Francine Laden, Joel Schwartz, Frank E. Speizer, and Douglas W. Dockery. Reduction in fine particulate air pollution and mortality : Extended follow-up of the Harvard Six Cities study. *American Journal of Respiratory and Critical Care Medicine*, 173(6) :667–672, March 2006. doi : 10.1164/rccm.200503-443OC. Epub 2006 Jan 19.

J T Lee and J Schwartz. Reanalysis of the effects of air pollution on daily mortality in seoul, korea : A case-crossover design. *Environmental Health Perspectives*, 107(8) :633–636, 1999. doi : 10.1289/ehp.99107633. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.99107633.

D. Levy, T. Lumley, L. Sheppard, J. Kaufman, and H. Checkoway. Referent selection in case–crossover analyses for acute health effects of air pollution. *Epidemiology*, 12 :186–192, March 2001a.

Drew Levy, Lianne Sheppard, Harvey Checkoway, Joel Kaufman, Thomas Lumley, Jane Koenig, and David Siscovick. A Case-Crossover Analysis of Particulate Matter Air Pollution and Out-of-Hospital Primary Cardiac Arrest. *Epidemiology*, 12(2) :193–199, March 2001b.

Jr Louis Anthony (Tony) Cox. Risk analysis implications of dose-response thresholds for nlrp3 inflammasome-mediated diseases : Respirable crystalline silica and lung cancer as an example. *Dose-Response*, 17(2) :1559325819836900, 2019. doi : 10.1177/1559325819836900. URL https://doi.org/10.1177/1559325819836900. PMID : 31168301.

Robert E. Lucas. Econometric policy evaluation : A critique. *Carnegie-Rochester Conference Series on Public Policy*, 1 :19–46, 1976. ISSN 0167-2231. doi : https://doi.org/10.1016/S0167-2231(76)80003-6. URL https://www.sciencedirect.com/science/article/pii/S0167223176800036.

Thomas Lumley and Drew Levy. Bias in the case – crossover design : implications for studies of air pollution. *Environmetrics*, 11(6) :689–704, 2000. doi : https://doi.org/10.1002/1099-095X(200011/12)11:6<689::AID-ENV439>3.0.CO;2-N. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1099-095X%28200011/12%2911%3A6%3C689%3A%3AAID-ENV439%3E3.0.CO%3B2-N.

M. Maclure and M. A. Mittleman. Should We Use a Case-Crossover Design ? *Annual Review of Public Health*, 21(1) :193–221, 2000. doi : https://doi.org/10.1146/annurev.publhealth.21.1.193.

Malcolm Maclure. The Case-Crossover Design : A Method for Studying Transient Effects on the Risk of Acute Events. *American Journal of Epidemiology*, 133(2) :144–153, 01 1991. ISSN 0002-9262. doi : 10.1093/oxfordjournals.aje.a115853. URL https://doi.org/10.1093/oxfordjournals.aje.a115853.

J. M. Mazurek, G. Syamlal, J. M. Wood, S. A. Hendricks, and A. Weston. Malignant Mesothelioma Mortality — United States, 1999–2015. *MMWR Morbidity and Mortality Weekly Report*, 66 :214–218, 2017. doi : 10.15585/mmwr.mm6608a3. URL http://dx.doi.org/10.15585/mmwr.mm6608a3.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition, 1989.

Mercedes Medina-Ramón, Antonella Zanobetti, and Joel Schwartz. The Effect of Ozone and PM10 on Hospital Admissions for Pneumonia and Chronic Obstructive Pulmonary Disease : A National Multicity Study. *American Journal of Epidemiology*, 163(6) :579–588, 01 2006. ISSN 0002-9262. doi : 10.1093/aje/kwj078. URL https://doi.org/10.1093/aje/kwj078.

William Navidi. Bidirectional Case-Crossover Designs for Exposures with Time Trends. *Biometrics*, 54(2) :596–605, 1998. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/3109766.

L M Neas, J Schwartz, and D Dockery. A case-crossover analysis of air pollution and mortality in philadelphia. *Environmental Health Perspectives*, 107(8) :629–631, 1999. doi : 10.1289/ehp.99107629. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.99107629.

Judea Pearl. Causal inference in statistics : An overview. *Statistics Surveys*, 3(none) :96 – 146, 2009. doi : 10.1214/09-SS057. URL https://doi.org/10.1214/09-SS057.

Roger D. Peng, Francesca Dominici, Roberto Pastor-Barriuso, Scott L. Zeger, and Jonathan M. Samet. Seasonal Analyses of Air Pollution and Mortality in 100 US Cities. *American Journal of Epidemiology*, 161(6) :585–594, 03 2005. ISSN 0002-9262. doi : 10.1093/aje/kwi075. URL https://doi.org/10.1093/aje/kwi075.

Annette Peters, Douglas W. Dockery, James E. Muller, and Murray A. Mittleman. Increased Particulate Air Pollution and the Triggering of Myocardial Infarction. *Circulation*, 103(23) : 2810–2815, 2001. doi : 10.1161/01.CIR.103.23.2810. URL https://www.ahajournals.org/doi/abs/10.1161/01.CIR.103.23.2810.

Annette Peters, Stephanie von Klot, Margit Heier, Ines Trentinaglia, Allmut Hörmann, H. Erich Wichmann, and Hannelore Löwel. Exposure to Traffic and the Onset of Myocardial Infarction. *New England Journal of Medicine*, 351(17) :1721–1730, 2004. doi : 10.1056/NEJMoa040203. URL https://www.nejm.org/doi/full/10.1056/NEJMoa040203.

J. M. Peters, E. Avol, W. Navidi, S. J. London, W. J. Gauderman, F. Lurmann, W. S. Linn, H. Margolis, E. Rappaport, H. Gong, and D. C. Thomas. A study of twelve Southern California communities with differing levels and types of air pollution. I. Prevalence of respiratory morbidity. *American Journal of Respiratory and Critical Care Medicine*, 159(3) : 760–767, March 1999a. doi : 10.1164/ajrccm.159.3.9804143.

John M. Peters, Edward Avol, W. James Gauderman, William S. Linn, William Navidi, Stephanie J. London, Helene Margolis, Edward Rappaport, H. Vora, Henry Jr. Gong, and Duncan C. Thomas. A study of twelve Southern California communities with differing levels and types of air pollution. II. Effects on pulmonary function. *American Journal of Respiratory and Critical Care Medicine*, 159(3) :768–775, March 1999b. doi : 10.1164/ajrccm.159.3.9804144.

C. Arden III Pope. Mortality effects of longer term exposures to fine particulate air pollution : Review of recent epidemiological evidence. *Inhalation Toxicology*, 19(sup1) :33–38, 2007. doi : 10.1080/08958370701492961. URL https://doi.org/10.1080/08958370701492961. PMID : 17886048.

C. Arden III Pope, Michael J. Thun, Mohan M. Namboodiri, Douglas W. Dockery, Joel S. Evans, Frank E. Speizer, and Clark W. Jr. Heath. Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory and Critical Care Medicine*, 151(3 Pt 1) :669–674, Mar 1995. doi : 10.1164/ajrccm/151.3_Pt_1.669.

J. M. Samet, F. Dominici, S. L. Zeger, J. Schwartz, and D. W. Dockery. The National Morbidity, Mortality, and Air Pollution Study. Part I : Methods and methodologic issues. *Research Report (Health Effects Institute)*, 94(Pt 1) :5–14 ; discussion 75–84, June 2000a.

J. M. Samet, S. L. Zeger, F. Dominici, F. Curriero, I. Coursac, D. W. Dockery, J. Schwartz, and A. Zanobetti. The National Morbidity, Mortality, and Air Pollution Study, Part II :

Morbidity and Mortality from Air Pollution in the United States. Health Effects Institute, Cambridge, MA., 2000b.

Jonathan M. Samet, Francesca Dominici, Frank C. Curriero, Ivan Coursac, and Scott L. Zeger. Fine Particulate Air Pollution and Mortality in 20 U.S. Cities, 1987–1994. *New England Journal of Medicine*, 343(24) :1742–1749, 2000c. doi : 10.1056/NEJM200012143432401. URL https://www.nejm.org/doi/full/10.1056/NEJM200012143432401.

Muriel Sayan and Brooke T. Mossman. The NLRP3 inflammasome in pathogenic particle and fibre-associated lung inflammation and diseases. *Particle and Fibre Toxicology*, 13 :51, 2015. doi : 10.1186/s12989-016-0162-4. URL https://doi.org/10.1186/s12989-016-0162-4.

James J Schlesselman. *Case Control Studies : Design, Conduct, Analysis*. Oxford University Press, New York, 1982.

S Selvin, D Merrill, L Wong, and S T Sacks. Ecologic regression analysis and the study of the influence of air quality on mortality. *Environmental Health Perspectives*, 54 :333–340, 1984. doi : 10.1289/ehp.8454333. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.8454333.

Bill Shipley. *Cause and correlation in biology : A user's guide to path analysis, structural equations and causal inference with R*. Cambridge university press, 2016.

Stacy Sneeringer. Does Animal Feeding Operation Pollution Hurt Public Health ? A National Longitudinal Study of Health Externalities Identified by Geographic Shifts in Livestock Production. *American Journal of Agricultural Economics*, 91(1) :124–137, 2009. doi : https://doi.org/10.1111/j.1467-8276.2008.01161.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8276.2008.01161.x.

J. M. Symons, L. Wang, E. Guallar, E. Howell, F. Dominici, M. Schwab, B. A. Ange, J. Samet, J. Ondov, D. Harrison, and A. Geyh. A Case-Crossover Study of Fine Particulate Matter Air Pollution and Onset of Congestive Heart Failure Symptom Exacerbation Leading to Hospitalization. *American Journal of Epidemiology*, 164(5) :421–433, 06 2006. ISSN 0002-9262. doi : 10.1093/aje/kwj206. URL https://doi.org/10.1093/aje/kwj206.

C. L. Tran, M. Graham, and D. Buchanan. A biomathematical model for rodent and human lung describing exposure, dose, and response to inhaled silica. Technical report, Institute of Occupational Medicine, 2001.

Sverre Vedal. Ambient particles and health : Lines that divide. *Journal of the Air & Waste Management Association*, 47(5) :551–581, 1997. doi : 10.1080/10473289.1997.10463922. URL https://doi.org/10.1080/10473289.1997.10463922.

Chen W-D. Testing for spurious regression in a panel data model with the individual number and time length growing. Journal of Applied Statistics, 33(8) :759–772, 2006.

Jon Wakefield. Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology*, 38(2) :330–336, 03 2009. ISSN 0300-5771. doi : 10.1093/ije/dyp179. URL https://doi.org/10.1093/ije/dyp179.

Jonathan Wakefield and Ruth Salway. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society : Series A (Statistics in Society)*, 164 (1) :119–137, 2001. doi : https://doi.org/10.1111/1467-985X.00191. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-985X.00191.

Antonella Zanobetti and Joel Schwartz. The Effect of Particulate Air Pollution on Emergency Admissions for Myocardial Infarction : A Multicity Case-Crossover Analysis. *Environmental Health Perspectives*, 113(8) :978–982, 2005. doi : 10.1289/ehp.7550. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.7550.

Chapitre 2

Répartition des maladies respiratoires chroniques par tranche d'âge

2.1 Introduction

Les maladies respiratoires chroniques constituent un enjeu majeur de santé publique. Afin d'étudier leur incidence à l'échelle des États-Unis, nous exploitons des données fournies par l'*Institute for Health Metrics and Evaluation* (IHME) dans le cadre du *Global Burden of Disease* (GBD) pour l'année 2021 (Institute for Health Metrics and Evaluation (IHME), 2024b).

Dans l'optique de limiter l'influence des valeurs atypiques, nous avons recouru à une régression linéaire robuste basée sur la norme de Huber (Andersen, 2008; Rousseeuw, 2019; Hampel et al., 1986; Rousseeuw and Leroy, 1987; Maronna et al., 2006; Hubert et al., 2008). Cette régression a été répétée plusieurs fois, dans le but de quantifier l'incertitude associée à l'estimation. Notre intérêt porte à la fois sur l'effet de l'âge (modélisé par un polynôme cubique) et sur l'effet global (incidence pour tous les groupes d'âge confondus), lequel est exprimé sur une échelle de *log-cotes*. Par souci de simplicité, le rôle du sexe est ignoré dans cette étude, mais nous considérons qu'il serait intéressant de l'inclure dans les recherches futures.

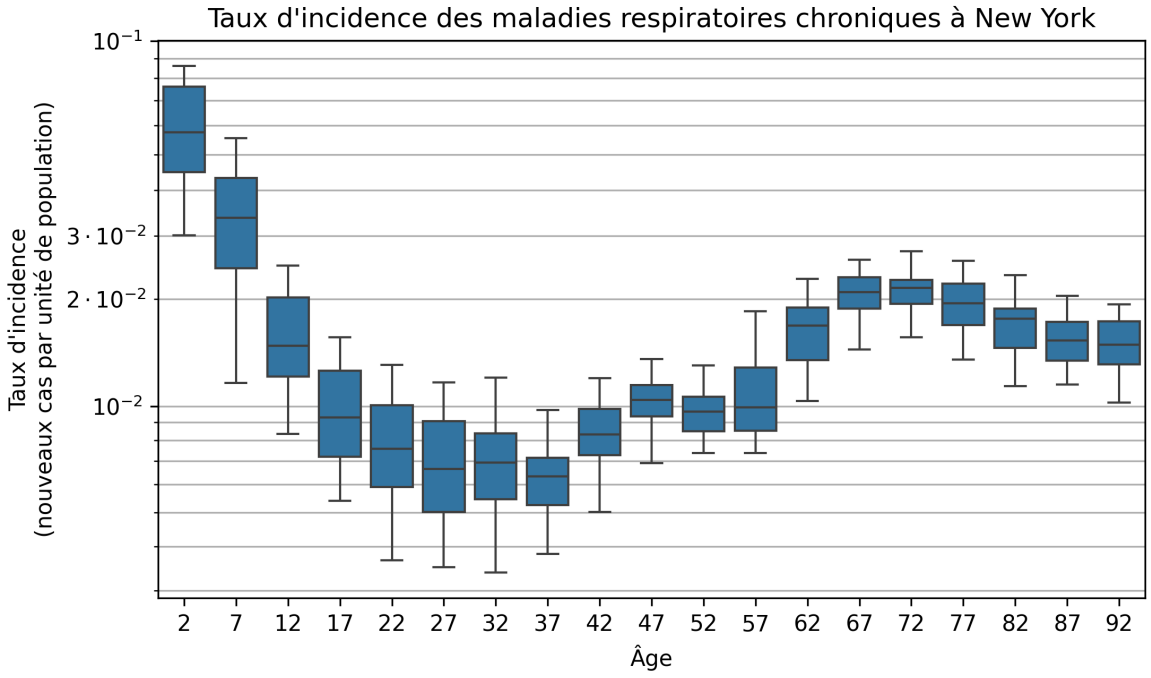
2.2 Statistiques descriptives

Les données portent sur la période allant de 1990 à 2021. Toute année postérieure à 2019 est exclue afin de prévenir les effets potentiels de la pandémie de COVID-19. Les données initiales incluent plusieurs groupes d'âge différenciés par des intervalles de quatre ans (< 5, 5–9, ..., 90–94), ainsi qu'un groupe couvrant l'ensemble des classes d'âge. Pour chaque État aux États-Unis, l'incidence (nouveaux cas pour 100 000 habitants) est associée à un intervalle d'incertitude $[\ell, u]$.

Selon l'aperçu des données et des outils GBD 2021 (Institute for Health Metrics and Evaluation

(IHME), 2024a), chaque estimation est recalculée 500 fois en échantillonnant des distributions plutôt que d'utiliser des valeurs ponctuelles pour les données d'entrée, les transformations de données et le choix du modèle. Pour obtenir l'intervalle d'incertitude à 95 %, on classe alors 1,000 valeurs obtenues (500 répétitions doublées afin de capturer les échantillonnages internes) de la plus petite à la plus grande, et on retient la 25e et la 975e valeurs.

Pour un groupe d'âge donné (proche de a ans), on note i_a l'incidence. Cette dernière est échantillonnée de façon aléatoire et uniforme dans l'intervalle $[\ell, u]$. Nous répétons le processus d'échantillonnage et d'ajustement 100 fois. Afin d'assurer la reproductibilité, nous avons fixé le générateur de nombres pseudo-aléatoires à 1729. La figure 2.1 montre l'incidence par âge pour l'État de New York : l'axe des ordonnées est en échelle logit, chaque point est placé au milieu de sa tranche d'âge et l'incidence est tirée au hasard entre les bornes inférieure et supérieure de l'intervalle d'incertitude selon la loi uniforme. Il y a une observation par tranche d'âge et par année d'étude (1990-2019).



Source : GBD 2021

FIGURE 2.1 – Taux d'incidence (par unité de population) à New York. L'échelle des ordonnées est logit.

La figure 2.1 suggère qu'une relation de forme cubique lie le taux d'incidence à l'âge. Le présent chapitre se propose de formaliser, vérifier puis généraliser cette observation à l'ensemble des États des États-Unis.

2.3 Méthodologie statistique

2.3.1 Modèle de régression robuste

l'objectif de cette section est de décrire comment nous relions l'incidence (par tranche d'âge) à l'incidence globale (tous âges confondus), tout en tenant compte de l'effet de l'âge de façon robuste aux observations atypiques (ou *outliers*).

Formulation du modèle

Pour chaque groupe d'âge (dont l'âge moyen est a), on dispose d'une valeur d'incidence i_a, ainsi que d'une incidence globale i_{globale} (tous âges confondus). Les deux valeurs sont exprimées pour 100 000 personnes. En notant

$$\text{logit}(x) \;=\; \log\Big(\frac{x}{100000 - x}\Big),$$

nous adoptons le modèle suivant à chaque itération :

$$\text{logit}(i_a) \;=\; \alpha \;+\; \beta_0 \,\text{logit}\big(i_{\text{globale}}\big) \;+\; \beta_1\, a \;+\; \beta_2\, a^2 \;+\; \beta_3\, a^3 \;+\; \varepsilon,$$

où $\alpha, \beta_0, \beta_1, \beta_2$ et β_3 sont les coefficients à estimer. Le choix de la fonction logit s'explique par la nature proportionnelle de l'incidence : en transformant l'incidence par $\log\big(\frac{i_a}{100000-i_a}\big)$, on évite le problème de valeurs extrêmes (taux extrêmement faibles ou très élevés) et on linearise mieux la relation avec l'âge.

Régression robuste avec la perte de Huber

Le terme aléatoire ε est supposé suivre un modèle « hubérisé », c'est-à-dire que la *fonction de perte* utilisée dans l'ajustement du modèle n'est pas simplement la somme des carrés des résidus (comme dans la régression linéaire classique), mais la *perte de Huber*.

— **Pourquoi la perte de Huber ?**

La fonction de perte de Huber est une combinaison de la perte quadratique (pour les petits résidus) et de la perte absolue (pour les grands résidus). Concrètement, lorsque l'écart entre la valeur prédite et la valeur observée est faible, le modèle se comporte comme une régression aux moindres carrés. En revanche, si cet écart est trop important (observations atypiques, erreurs de mesure importantes, etc.), la perte de Huber se comporte comme une perte absolue, limitant ainsi l'influence de ces valeurs extrêmes.

$$\rho_\delta(r) = \begin{cases} \frac{1}{2}\, r^2 & \text{si } |r| \le \delta, \\ \delta\,\big(|r| - \frac{\delta}{2}\big) & \text{sinon,} \end{cases}$$

où r désigne le résidu (écart entre la prédiction et l'observation) et δ est un paramètre de coupure. La régression robuste via `RLM(..., M=HuberT())` utilise par défaut le *tuning constant* $\delta = 1{,}345$.

— **Avantage principal : robustesse**

Grâce à cette approche, les valeurs potentiellement aberrantes sont « moins pénalisées », ce qui rend l'estimation des coefficients $(\alpha, \beta_0, \beta_1, \beta_2, \beta_3)$ plus stable et moins sensible aux données atypiques.

Échantillonnage pour refléter l'incertitude

l'incidence i_a et l'incidence globale i_{globale} ne sont pas des valeurs fixes, mais plutôt comprises dans un intervalle d'incertitude $[\ell, u]$. Pour chaque itération :

1. Nous échantillonnons aléatoirement i_a dans $[\ell_a, u_a]$ et i_{globale} dans $[\ell_{\text{all}}, u_{\text{all}}]$ suivant une loi uniforme.
2. Nous appliquons la transformation logit puis ajustons le modèle via la perte de Huber pour estimer $\alpha, \beta_0, \beta_1, \beta_2, \beta_3$.
3. Nous stockons les coefficients estimés.

En répétant cette opération un grand nombre de fois (typiquement plusieurs centaines ou milliers d'itérations), nous tenons compte de la variabilité due à l'estimation du taux d'incidence. Enfin, nous moyennons les coefficients estimés pour obtenir des valeurs centrales robustes, accompagnées d'intervalles de crédibilité (ou de confiance) reflétant la variabilité observée.

Cette procédure fournit ainsi une image plus complète de l'incertitude, contrairement à une seule estimation ponctuelle, tout en restant robuste aux observations extrêmes.

2.3.2 Tests de normalité des coefficients

À l'issue du processus d'itération décrit précédemment, nous obtenons, pour chaque État, un grand nombre d'estimations pour chacun des coefficients du modèle (α, β_0, β_1, β_2, β_3). Afin d'évaluer si ces distributions de coefficients (issues des multiples échantillonnages) peuvent être raisonnablement considérées comme normales, nous recourons au test de Shapiro–Wilk.

Le test de Shapiro–Wilk (souvent abrégé en *SW test*) vérifie la proximité d'un échantillon de données par rapport à une distribution normale de référence. Concrètement, il calcule un indice (*statistique* W) qui compare l'ordre des observations dans l'échantillon à l'ordre théorique de valeurs provenant d'une distribution normale.

— l'*hypothèse nulle* (H_0) stipule que les données sont issues d'une distribution normale.
— Une valeur de p-value élevée (généralement $p \geq 0{,}05$) indique qu'aucune preuve forte ne permet de rejeter l'hypothèse de normalité.
— À l'inverse, si la p-value est très faible ($p < 0{,}05$), on conclut que les données s'éloignent significativement d'une distribution normale.

Le test de Shapiro–Wilk est souvent privilégié pour des échantillons de taille faible à modérée. Pour des très grands échantillons, même de faibles déviations par rapport à la normalité peuvent conduire à un rejet de l'hypothèse nulle, ce qui doit être interprété avec prudence.

Application à nos coefficients

Pour chaque État, nous considérons le nuage de valeurs estimées pour un coefficient donné (par exemple, β_1). Nous appliquons le test de Shapiro–Wilk à cet échantillon :

1. **Estimation du coefficient.** Nous collectons les estimations issues des multiples itérations (après les tirages dans $[\ell,u]$ et l'ajustement du modèle robuste).
2. **Calcul de la statistique du test.** Le test de Shapiro–Wilk produit la statistique W et la *value* p associée.
3. **Interprétation.**
   — $p \geq 0{,}05$: aucune évidence pour rejeter l'hypothèse de normalité. Les estimations du coefficient peuvent être considérées comme approximativement normales.
   — $p < 0{,}05$: l'hypothèse de normalité est rejetée au seuil de 5 %. Les estimations présentent une ou plusieurs caractéristiques (asymétrie, queues épaisses, etc.) qui s'éloignent d'une distribution normale.

Ce diagnostic de normalité permet notamment de vérifier la validité d'hypothèses statistiques ultérieures (par exemple, la construction d'intervalles de confiance ou l'utilisation de tests paramétriques), et de justifier ou non l'emploi de méthodes complémentaires plus robustes.

2.4 Résultats

2.4.1 Distributions des valuers p

Le tableau 2.1 présente les valeurs p du test de Shapiro–Wilk appliqué aux distributions des coefficients robustes (α, β_0, β_1, β_2, β_3) pour chacun des 50 États et le District of Columbia. Chaque ligne du tableau correspond à un État et chaque colonne à l'un des coefficients estimés.

TABLE 2.1 – Résultats du test de Shapiro–Wilk (valeurs p) par État et par coefficient. Une valuer p supérieure à 0,05 indique qu'il n'existe pas de preuve suffisante pour rejeter l'hypothèse de normalité au seuil de 5 %. Source : GBD 2021.

État	Constante	Incidence (tous les âges)	Âge	Âge^2	Âge^3
Florida	0.900262	0.719677	0.962169	0.563169	0.482713
Indiana	0.046227	0.026316	0.259308	0.373499	0.295557
Arizona	0.266231	0.200417	0.644828	0.928012	0.821360
District of Columbia	0.192110	0.315677	0.630271	0.908799	0.589787
Virginia	0.459928	0.451890	0.933773	0.750506	0.775443
Minnesota	0.035898	0.039665	0.979166	0.985214	0.902774
Alabama	0.789109	0.930856	0.098207	0.100775	0.195099
Kentucky	0.399275	0.317714	0.368874	0.272306	0.461826
Wisconsin	0.578989	0.294828	0.665907	0.473481	0.608763
Louisiana	0.286008	0.391046	0.882237	0.869540	0.548751
Oregon	0.420299	0.295518	0.226837	0.439856	0.556087
Ohio	0.603272	0.872154	0.561491	0.186139	0.166579
Nevada	0.150145	0.349179	0.198798	0.036677	0.029377
Texas	0.379894	0.268752	0.559649	0.578125	0.659767
Mississippi	0.693148	0.823702	0.653253	0.994191	0.962616
Alaska	0.193820	0.306439	0.928482	0.827019	0.829079
Connecticut	0.097959	0.155841	0.929510	0.947488	0.971022
Georgia	0.857730	0.806463	0.595895	0.335236	0.297994
North Carolina	0.469925	0.581799	0.751414	0.751692	0.895083
Wyoming	0.004778	0.005386	0.847914	0.675758	0.684384
New Jersey	0.026770	0.037915	0.749857	0.773758	0.583323
Rhode Island	0.016973	0.012016	0.864882	0.423398	0.400641
Washington	0.120665	0.140095	0.533588	0.873975	0.887788
Missouri	0.232014	0.294303	0.640803	0.396093	0.177082
Vermont	0.593869	0.585375	0.731438	0.545405	0.350779
Massachusetts	0.550541	0.464981	0.556817	0.425361	0.412623
Maryland	0.333188	0.322741	0.817402	0.522865	0.593523
Utah	0.416914	0.233968	0.500383	0.599504	0.668291
South Carolina	0.205609	0.155599	0.740327	0.624069	0.717434
Michigan	0.082893	0.144240	0.486778	0.947819	0.976748
Kansas	0.135763	0.142002	0.965683	0.890911	0.923332
New Mexico	0.406371	0.332559	0.382242	0.204714	0.358130
Iowa	0.939622	0.951580	0.968355	0.829809	0.806361
New York	0.091702	0.153574	0.941865	0.989753	0.824668
Arkansas	0.736346	0.929694	0.544431	0.838450	0.780760
Delaware	0.003730	0.009623	0.501311	0.840542	0.642348
Pennsylvania	0.167352	0.156111	0.010462	0.057079	0.249097
Colorado	0.166951	0.337884	0.350400	0.225930	0.348391
Idaho	0.610961	0.651126	0.836059	0.759192	0.843739
West Virginia	0.153072	0.400901	0.245878	0.482385	0.503756
California	0.546507	0.537487	0.623897	0.528192	0.316758
Tennessee	0.826333	0.619585	0.625040	0.206343	0.232872
Oklahoma	0.056913	0.057975	0.922941	0.883862	0.748928
Montana	0.874656	0.766501	0.724507	0.810860	0.796844
Maine	0.823682	0.986323	0.996607	0.650141	0.438621
Illinois	0.643275	0.646493	0.286435	0.130874	0.108143
New Hampshire	0.003858	0.003582	0.132991	0.702489	0.824263
Nebraska	0.122026	0.063810	0.927237	0.808858	0.725433
North Dakota	0.435265	0.360600	0.069065	0.166695	0.055921
Hawaii	0.897619	0.935712	0.948668	0.592383	0.290948
South Dakota	0.251898	0.328966	0.964662	0.785235	0.691936

Bien que ce tableau fournisse le détail complet, il est parfois difficile d'appréhender la distribution globale des valeurs p à travers une simple liste de nombres. Afin de faciliter l'interprétation, nous avons recours à une représentation graphique sous forme de *diagramme en violon*, représentée à la figure 2.2.

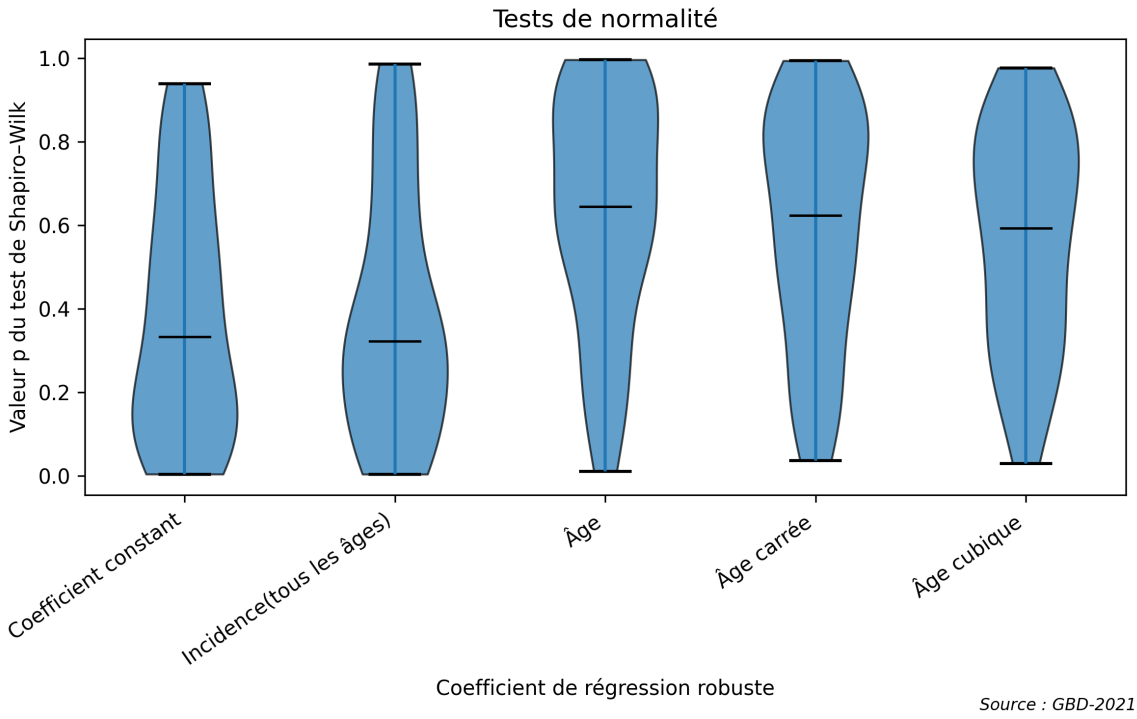


FIGURE 2.2 – Distribution en violon des valeurs p du test de Shapiro–Wilk, par coefficient (50 États et le District of Columbia). Chaque colonne illustre la *densité* (sur l'axe vertical) des valeurs p pour un coefficient donné. La largeur du violon reflète la proportion de valeurs p à chaque niveau.

Interprétation et bonnes pratiques de visualisation

— **Diagrammes en violon** :

une boîte à moustaches central (indiquant la médiane et les quartiles) avec une courbe de densité symétrique autour de l'axe vertical.

1. Ils permettent de repérer rapidement les zones où les valeurs p sont les plus concentrées et d'identifier d'éventuelles valeurs extrêmes.

— **Seuil critique de 5 % (valeur p = 0,05)** :

1. Si la plupart des valeurs p se situent *au-dessus* de 0,05, on ne rejette pas l'hypothèse de normalité pour la majorité des États.
2. Si, au contraire, de nombreuses valeurs p sont *inférieures* à 0,05, cela indique qu'on observe un écart significatif par rapport à la normalité dans plusieurs États.

— **Considérations pratiques** :

1. *Taille d'échantillon* : Pour des séries de coefficients obtenus après de multiples itérations, la puissance du test de Shapiro–Wilk peut être élevée et conduire à un rejet de l'hypothèse de normalité pour de légères déviations.
2. *Variabilité entre les coefficients* : Une différence notable dans la distribution des valeurs p entre deux coefficients peut suggérer que l'un est estimé de façon plus stable et plus conforme aux hypothèses de normalité que l'autre.

Dans l'ensemble, la combinaison du tableau 2.1 et du diagramme en violon (figure 2.2) offre une vision à la fois détaillée et synthétique de la normalité (ou non) des distributions de coefficients selon les 50 États et le District of Columbia. Cela permet ensuite de mieux apprécier la robustesse des inférences statistiques associées à chacun de ces coefficients.

2.4.2 Estimateurs moyens des coefficients

Au terme de nos multiples itérations, nous disposons, pour chaque État, d'un ensemble d'estimations pour les coefficients du modèle robuste : la constante α, l'effet global β_0 (liant l'incidence par tranche d'âge à l'incidence globale), ainsi que les composantes linéaire (β_1), quadratique (β_2) et cubique (β_3) de l'âge. Nous calculons ensuite la moyenne de ces estimations, État par État, afin d'obtenir cinq *estimateurs moyens* (un par coefficient).

Distribution globale des coefficients

La figure 2.3 illustre la distribution de ces coefficients moyens sur l'ensemble des 50 États et le District of Columbia, sous forme de diagrammes en violon. Chaque violon permet de repérer :

— **La densité des valeurs** : la largeur du violon renseigne sur la concentration des valeurs autour de certaines zones.

— **La médiane et les quartiles** : visibles grâce à un marqueur central (ligne horizontale) indiquant la médiane, et parfois des traits indiquant les quartiles.

— **Les éventuelles valeurs extrêmes ou asymétries** : si la forme du violon est nettement allongée ou décalée, cela peut révéler une plus grande variabilité ou une asymétrie dans les estimations d'un coefficient.

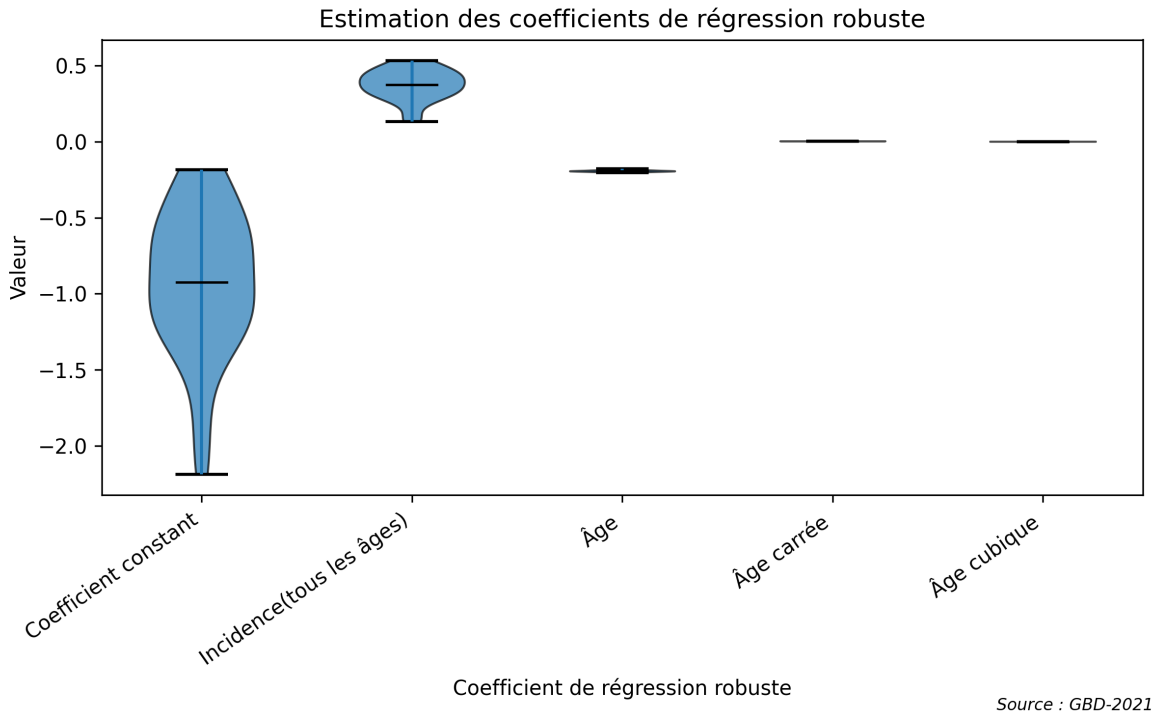


FIGURE 2.3 – Distribution en violon des coefficients moyens de la régression robuste, tous États confondus. La largeur de chaque violon indique la densité estimée des valeurs à chaque niveau.

Tableau récapitulatif par État

Pour une analyse plus détaillée, le tableau 2.2 ci-après répertorie ces coefficients moyens pour chaque État. Cette vue granulaire peut s'avérer pertinente, par exemple, si l'on souhaite comparer directement deux États voisins ou si l'on s'intéresse aux extrêmes (États pour lesquels les coefficients sont particulièrement élevés ou faibles).

TABLE 2.2 – Coefficients moyens du modèle robuste par État. Les valeurs présentées correspondent à la moyenne des estimations obtenues via les multiples itérations d'échantillonnage et d'ajustement (régression robuste). Source : GBD 2021.

État	Constante	Incidence (tous les âges)	Âge	Âge^2	Âge^3
Georgia	-0.671946	0.430326	-0.198000	0.004449	-0.000027
Idaho	-1.965648	0.158865	-0.190018	0.004262	-0.000026
Wisconsin	-1.222907	0.330755	-0.184288	0.004086	-0.000025
Kansas	-0.781263	0.406307	-0.196285	0.004359	-0.000027
Alaska	-1.884132	0.151040	-0.190711	0.004134	-0.000024
District of Columbia	-1.346955	0.269259	-0.179566	0.003815	-0.000023
New Hampshire	-0.838015	0.385153	-0.194291	0.004313	-0.000026
Maine	-0.545244	0.440159	-0.198476	0.004388	-0.000027
Nevada	-1.225596	0.304739	-0.190018	0.004194	-0.000025
Arkansas	-1.151253	0.351523	-0.185039	0.004144	-0.000025
New Jersey	-0.204345	0.514698	-0.200054	0.004410	-0.000027
Connecticut	-0.205225	0.513576	-0.198130	0.004352	-0.000026
Missouri	-0.544167	0.460457	-0.193552	0.004315	-0.000026
Colorado	-0.971400	0.343362	-0.199803	0.004422	-0.000027
Rhode Island	-0.418249	0.469570	-0.195052	0.004290	-0.000026
Nebraska	-1.120373	0.366173	-0.186783	0.004204	-0.000026
Florida	-0.492848	0.471961	-0.194814	0.004349	-0.000027
Mississippi	-1.106534	0.377411	-0.186475	0.004237	-0.000026
Vermont	-1.108899	0.326219	-0.192628	0.004250	-0.000026
Louisiana	-1.188533	0.331680	-0.192397	0.004349	-0.000027
Maryland	-0.623663	0.434829	-0.192420	0.004261	-0.000026
Oregon	-1.321628	0.297876	-0.186333	0.004138	-0.000025
Pennsylvania	-0.202719	0.529016	-0.193086	0.004259	-0.000026
Delaware	-0.614979	0.416641	-0.198489	0.004369	-0.000027
Illinois	-0.962975	0.374315	-0.187745	0.004173	-0.000025
Kentucky	-0.516919	0.468000	-0.193981	0.004348	-0.000027
West Virginia	-0.197746	0.525634	-0.195545	0.004340	-0.000026
Massachusetts	-0.896073	0.358072	-0.196900	0.004354	-0.000027
New York	-0.267371	0.513713	-0.186473	0.004108	-0.000025
Alabama	-0.566789	0.459257	-0.193196	0.004316	-0.000026
California	-1.249773	0.290689	-0.198156	0.004413	-0.000027
South Dakota	-1.408039	0.288593	-0.191311	0.004276	-0.000026
Montana	-1.275156	0.293896	-0.193819	0.004304	-0.000026
New Mexico	-0.969591	0.347537	-0.192991	0.004247	-0.000026
Ohio	-0.329737	0.515963	-0.190680	0.004256	-0.000026
Utah	-2.111485	0.149834	-0.177633	0.003913	-0.000023
Tennessee	-0.829140	0.420391	-0.188904	0.004265	-0.000026
Texas	-0.768053	0.393132	-0.204142	0.004559	-0.000028
Arizona	-0.912814	0.360985	-0.193649	0.004295	-0.000026
Indiana	-0.476412	0.469788	-0.195229	0.004352	-0.000027
Michigan	-0.482946	0.466191	-0.192444	0.004259	-0.000026
North Carolina	-0.785923	0.424856	-0.191566	0.004322	-0.000027
Minnesota	-1.238576	0.309178	-0.195062	0.004279	-0.000026
South Carolina	-0.572408	0.476576	-0.186667	0.004199	-0.000026
Virginia	-0.833369	0.390105	-0.193818	0.004322	-0.000026
Hawaii	-1.438490	0.238511	-0.187950	0.004071	-0.000024
North Dakota	-1.158724	0.323574	-0.194520	0.004304	-0.000026
Wyoming	-1.622545	0.218802	-0.191542	0.004249	-0.000026
Iowa	-1.186115	0.358169	-0.186266	0.004193	-0.000026
Washington	-1.344826	0.298453	-0.188853	0.004199	-0.000025
Oklahoma	-0.692055	0.415745	-0.197038	0.004394	-0.000027

2.4.3 Visualisation géographique des coefficients

La répartition spatiale des coefficients moyens est représentée sur les cartes choroplèthes ci-dessous, selon la palette de couleurs `Viridis`, qui est *color-blind friendly* et met en évidence les variations de valeur de façon progressive. Chaque carte met l'accent sur l'un des cinq coefficients.

Carte de la constante (α) :

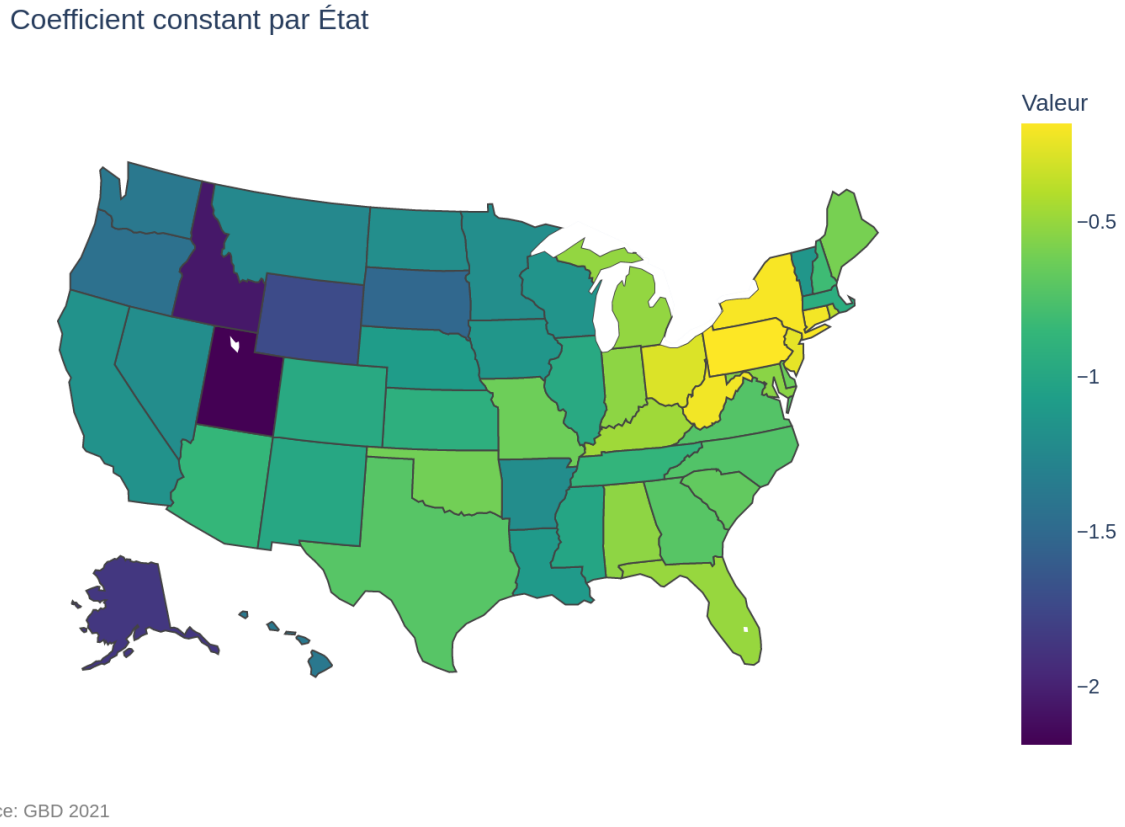


FIGURE 2.4 – Valeurs moyennes de α (coefficient constant) par État. Plus la couleur est foncée, plus la valeur estimée est élevée.

Carte du coefficient β_0 (incidence globale) :

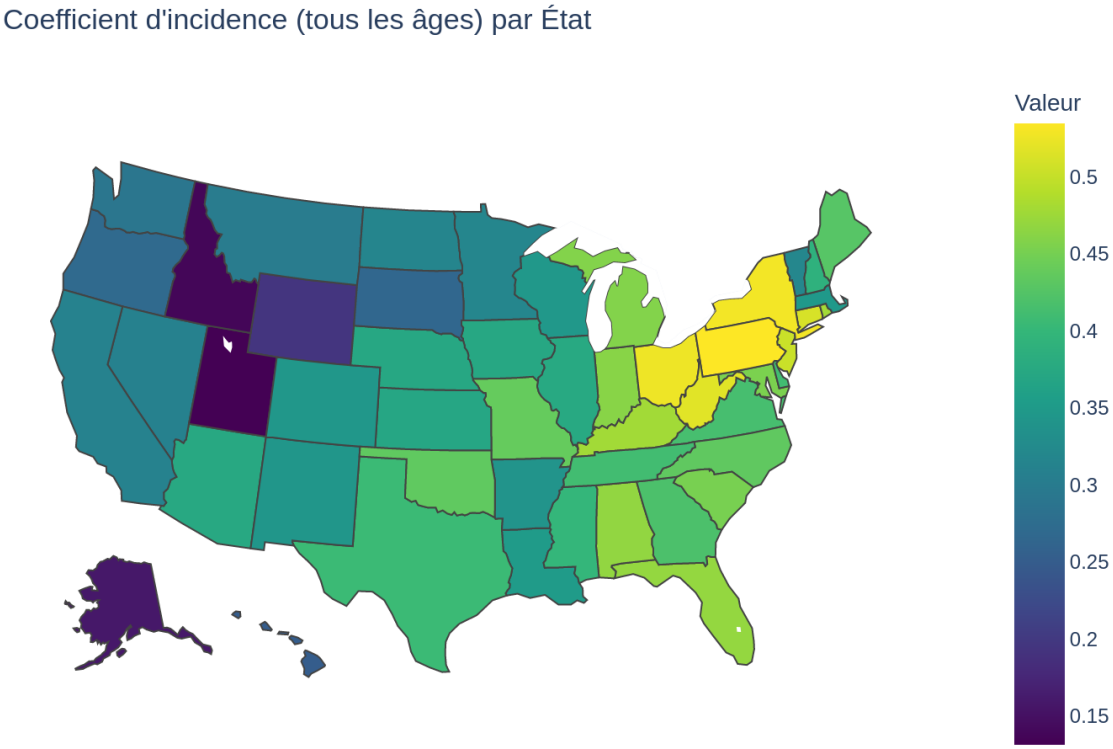


FIGURE 2.5 – Valeurs moyennes de β_0 (log-odds de l'incidence globale) par État. Les zones plus foncées indiquent un effet global plus prononcé de l'incidence.

Carte du coefficient β_1 (âge) :

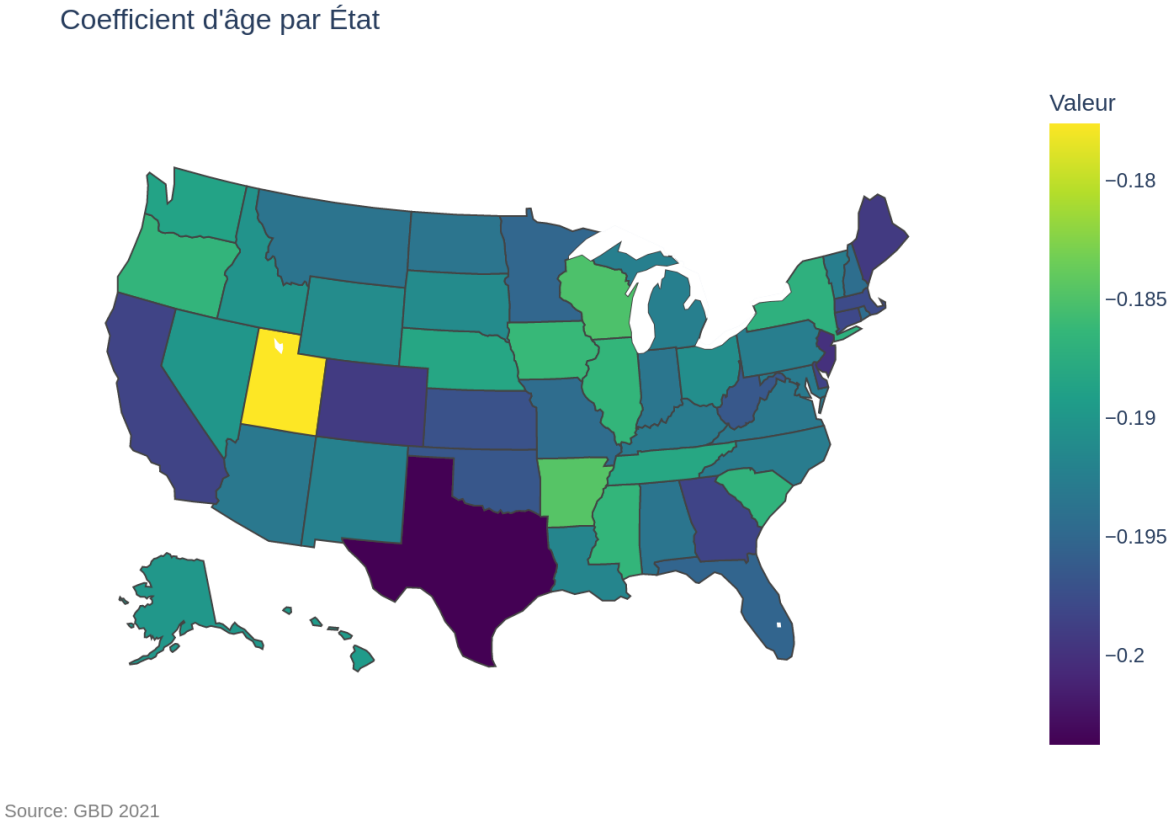


FIGURE 2.6 – Valeurs moyennes de β_1 (terme linéaire en âge) par État.

Carte du coefficient β_2 (âge au carré) :

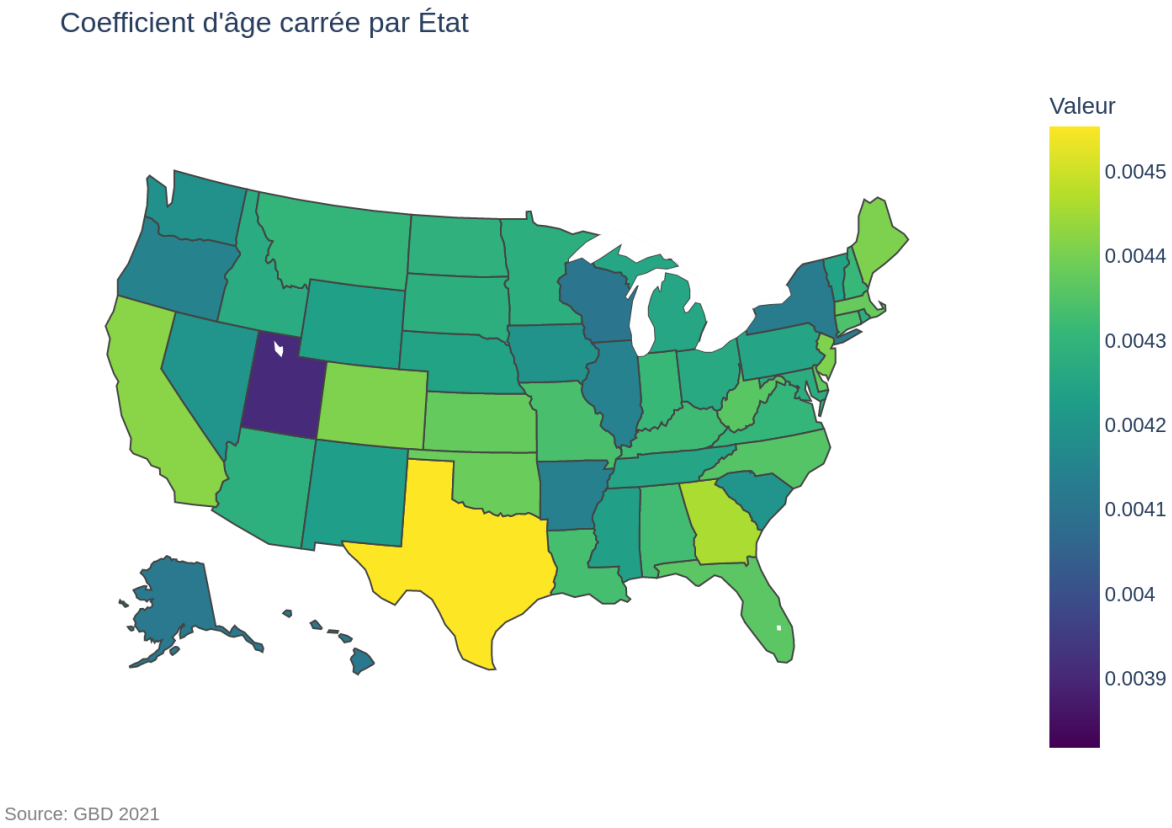


FIGURE 2.7 – Valeurs moyennes de β_2 (terme quadratique) par État.

Carte du coefficient β_3 (âge au cube) :

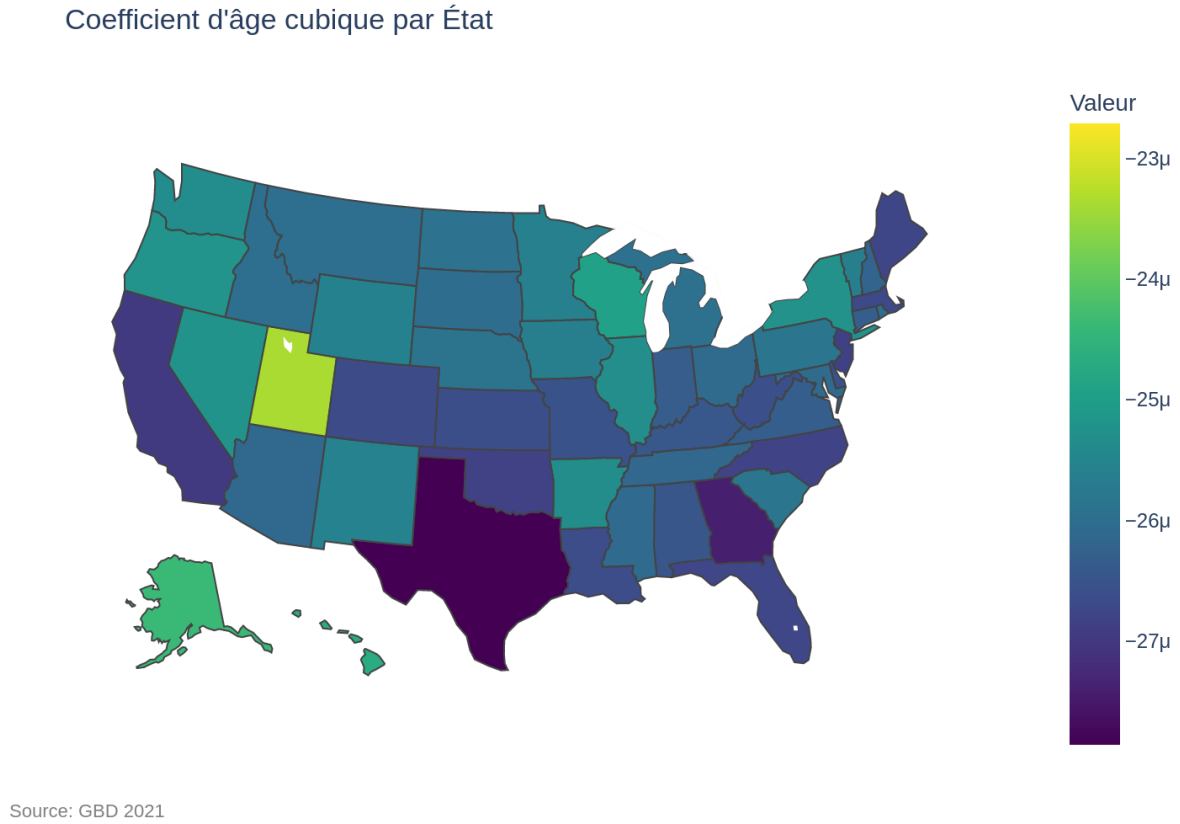


FIGURE 2.8 – Valeurs moyennes de β_3 (terme cubique) par État.

2.5 Qualité prédictive du modèle : synthèse par État

Paramètres de la validation croisée

La qualité de prédiction a été évaluée par **validation croisée à $k = 10$ plis** (*10-fold cross-validation*), avec permutation préalable des données (`shuffle = True`) et **graine aléatoire fixée à 42** afin de garantir la reproductibilité des résultats. [1] Le critère d'évaluation est la **sMAPE** (Symmetrical Mean Absolute Percentage Error), définie pour chaque observation $(y, \hat{y})$ par

$$\text{sMAPE} \;=\; \frac{|\,y - \hat{y}\,|}{\frac{|y|+|\hat{y}|}{2}} \;\times 100\ \%.$$

Un faible terme $\varepsilon = 10^{-8}$ est ajouté au dénominateur pour éviter toute division par zéro.

Cette métrique, bornée entre 0 et 100 %, est insensible aux valeurs nulles et pénalise de façon symétrique les sure- et sous-estimations. Pour chaque État et chaque pli, le modèle robuste à perte de Huber est ré-estimé puis évalué, produisant ainsi 10 valeurs de sMAPE par territoire.

1. En pratique, nous utilisons la classe `KFold` de SCIKIT-LEARN (`n_splits = 10, random_state = 42`).

Statistiques récapitulatives

Le tableau 2.3 rassemble, pour chacun des 50 États et le District of Columbia, trois indicateurs issus de ces 10 valeurs :

1. le *minimum* : meilleure performance observée ;
2. la *moyenne* : performance typique ;
3. le *maximum* : pire performance observée.

Dans l'ensemble, la **sMAPE moyenne se situe entre 4 % et 7 %**, ce qui traduit une excellente adéquation du modèle robuste ; même dans les cas les plus défavorables la sMAPE reste modérée (*maxima* rarement au-delà de 10 %).

TABLE 2.3 – Statistiques sMAPE (minimum, moyenne, maximum) par État. $k = 10$ plis, graine = 42, métrique = sMAPE. Source : GBD 2021.

État	smape (min)	smape (moyenne)	smape (max)
Mississippi	3.44	4.37	5.32
Arkansas	3.62	4.40	5.38
Florida	3.72	4.44	5.55
Missouri	3.93	4.48	5.00
South Dakota	3.53	4.56	5.77
Ohio	3.79	4.59	5.35
Indiana	4.12	4.60	5.03
Maryland	3.82	4.62	5.50
Tennessee	4.27	4.63	5.45
North Carolina	3.94	4.72	5.67
Alaska	4.06	4.75	5.47
Iowa	4.31	4.76	5.51
Wyoming	3.76	4.76	6.27
West Virginia	4.17	4.76	5.56
Nebraska	4.13	4.77	5.89
South Carolina	3.72	4.77	5.70
Georgia	3.90	4.77	5.37
Wisconsin	3.76	4.77	5.44
Kentucky	4.15	4.78	5.49
Louisiana	3.87	4.78	5.35
Nevada	4.15	4.78	6.20
Maine	4.42	4.80	5.45
Kansas	4.22	4.81	5.44
Washington	4.03	4.82	5.92
Alabama	4.09	4.84	5.44
Oklahoma	4.05	4.85	5.52
Michigan	4.25	4.86	5.45
Vermont	3.90	4.87	6.09
Utah	4.23	4.88	5.56
Rhode Island	4.28	4.88	5.67
New Hampshire	4.12	4.89	5.62
California	4.05	4.89	5.50
Illinois	3.77	4.91	5.71
Montana	4.05	4.92	5.73
Massachusetts	4.24	4.93	6.02
Delaware	4.02	4.95	5.85
Idaho	4.27	4.97	5.92
Arizona	4.15	4.99	5.65
Pennsylvania	4.42	4.99	5.84
Texas	4.36	5.02	6.27
Minnesota	4.17	5.04	6.31
District of Columbia	4.49	5.05	5.81
Virginia	4.21	5.06	5.58
Hawaii	4.43	5.06	6.39
North Dakota	4.28	5.07	6.68
Oregon	4.54	5.10	5.90
Connecticut	4.24	5.10	5.84
New Mexico	4.27	5.11	5.92
Colorado	4.29	5.13	5.87
New Jersey	4.55	5.25	5.89
New York	4.83	5.36	6.33

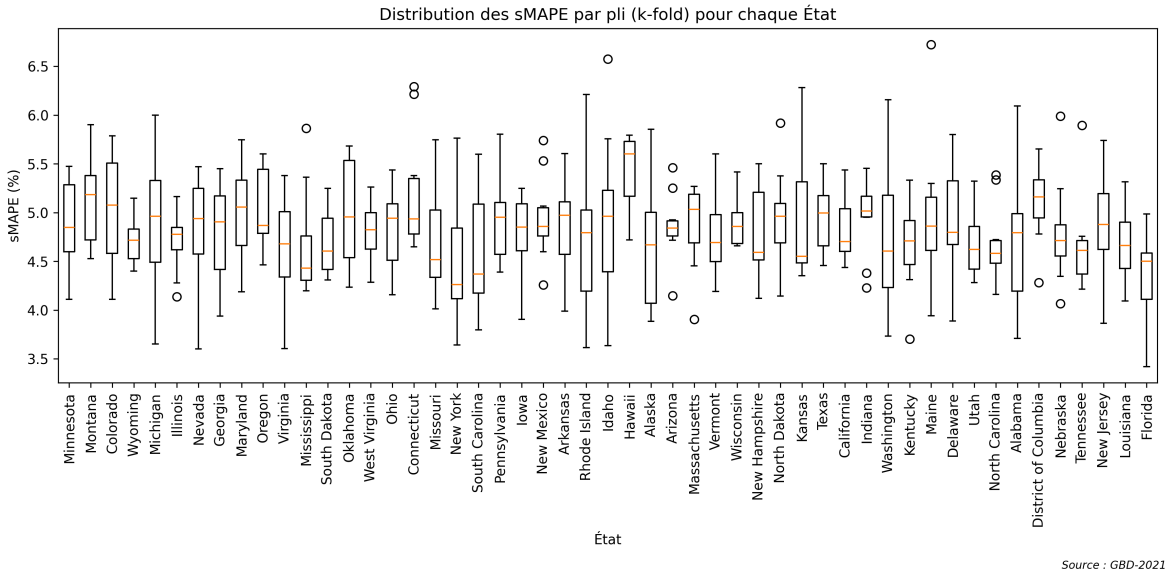


FIGURE 2.9 – Distribution des sMAPE par pli ($k = 10$) pour chaque État.

La figure 2.9 apporte une vision graphique de la dispersion des sMAPE :

— Les boîtes, centrées autour d'une médiane proche de 5 %, sont généralement compactes, signe d'une variabilité limitée entre plis.
— Quelques moustaches plus étendues (valeurs hors-quartile) indiquent, pour certains États, une sensibilité accrue au découpage aléatoire des données, sans toutefois atteindre des niveaux d'erreur critiques.

Avec un schéma de validation croisée rigoureux ($k = 10$, graine fixée) et une métrique relative stricte (sMAPE), la régression linéaire robuste à perte de Huber montre une **performance stable et élevée** sur l'ensemble des États. Les écarts territoriaux observés restent modestes et n'affectent pas les conclusions globales présentées dans les sections suivantes.

2.6 Discussion

Les résultats obtenus éclairent la relation entre l'incidence des maladies respiratoires chroniques et l'âge, tout en révélant des disparités géographiques au sein des États-Unis. Dans cette section, nous discutons de la pertinence de la méthode mise en œuvre, de l'interprétation possible des coefficients, des limites à prendre en compte et des pistes d'amélioration envisageables.

2.6.1 Robustesse de l'approche et adéquation du modèle

Le recours à une régression linéaire robuste fondée sur la perte de Huber s'avère approprié pour atténuer l'influence des valeurs atypiques (*outliers*). Contrairement à une régression aux moindres carrés ordinaires, qui peut être fortement affectée par des observations extrêmes, cette

méthode borne l'influence de ces points tout en conservant une bonne efficacité quand les résidus restent « normaux ». Les résultats de la validation croisée (sMAPE généralement inférieure à 10%) confirment la *bonne performance prédictive* du modèle robuste sur la quasi-totalité des États, malgré les variations inhérentes au découpage aléatoire des données.

En outre, l'échantillonnage répété de l'incidence dans ses intervalles d'incertitude permet de propager la variabilité des données de base jusqu'aux estimations finales, donnant ainsi une *vision réaliste de l'incertitude* des coefficients. Il en résulte une estimation plus nuancée qu'une simple régression ponctuelle.

2.6.2 Interprétation des coefficients et disparités régionales

Effet global (β_0) et constante (α)

Le coefficient β_0 capte l'effet d'*incidence globale* : en moyenne, à mesure que l'incidence globale (tous âges confondus) augmente, l'incidence par tranche d'âge suit une tendance ascendante. Il traduit donc un *niveau de base* (sur échelle logit) auquel viennent s'ajouter les variations spécifiques à l'âge. Les écarts de β_0 entre les États peuvent s'expliquer par des disparités en matière de pollution atmosphérique, de prévalence de facteurs de risque (p.ẽx. tabagisme, obésité), d'accès aux soins ou encore par des politiques de santé publique différenciées (programmes de dépistage, campagnes de vaccination, etc.).

Le terme α agit comme une constante d'ajustement sur l'échelle logit : certaines régions affichent une α relativement plus élevée, traduisant un taux de base majoré. Les raisons peuvent être multiples : structures démographiques particulières, densité urbaine, conditions socio-économiques, climat, etc.

Relation cubique avec l'âge β_1, β_2, β_3

Le choix d'un polynôme cubique pour modéliser l'effet de l'âge répond au constat visuel (figure 2.1) d'une courbe présentant plusieurs points d'inflexion. Les coefficients β_1, β_2 et β_3 indiquent que la relation entre l'âge et l'incidence n'est pas strictement linéaire : la probabilité de développer une maladie respiratoire chronique peut augmenter rapidement à certains âges, se stabiliser, puis évoluer différemment chez les personnes plus âgées. Ces phénomènes peuvent refléter :

- Des changements physiologiques liés au vieillissement (affaiblissement du système immunitaire, réduction de la capacité pulmonaire) ;
- Des expositions environnementales ou professionnelles cumulées au fil des ans (p.ẽx. tabagisme, inhalation de poussières) ;
- Des facteurs de comorbidités (maladies cardiovasculaires, diabète, etc.) plus fréquents avec l'âge.

Les *différences* dans les valeurs moyennes de $\beta_1, \beta_2, \beta_3$ entre États peuvent suggérer des contextes régionaux spécifiques (par exemple, dans certains États à forte altitude ou à climat sec, certaines pathologies respiratoires peuvent être moins fréquentes, ou encore la courbe d'incidence peut se décaler vers des âges plus avancés).

2.6.3 Qualité prédictive et validation croisée

La validation croisée à 10 plis (*k-fold cross-validation*), couplée à la sMAPE comme mesure d'erreur, montre une *dispersion relativement faible* des scores (5% à 7% en moyenne). Cette performance satisfaisante témoigne de la *bonne généralisation* du modèle, tout en indiquant une *relative stabilité* des estimations à travers divers sous-échantillons.

Il convient néanmoins de noter que la *taille des sous-échantillons*, dans la mesure où les données sont segmentées par tranches d'âge et par États, peut varier. La robustesse du modèle semble particulièrement bénéfique pour gérer des sous-ensembles de données plus réduits ou plus susceptibles de contenir des observations extrêmes (par exemple, certains âges très élevés pour lesquels l'incidence est moins bien estimée).

2.6.4 Limites et points d'attention

Malgré la solidité méthodologique, plusieurs limites méritent d'être soulignées :

1. **Exclusion des années postérieures à 2019 :** l'objectif était d'éviter l'effet confondant de la pandémie de COVID-19 sur les taux d'incidence des maladies respiratoires chroniques. Toutefois, cela implique que d'éventuelles *tendances récentes* (p.ẽx. meilleure prévention, détection précoce) sont absentes des données analysées.

   “‘

2. **Non prise en compte d'autres covariables :** le modèle se concentre sur la relation âge–incidence, assortie d'une correction via le taux d'incidence global, sans intégrer le sexe, les habitudes de vie (tabagisme, sédentarité) ou la pollution atmosphérique *in situ*. Dès lors, les coefficients captent *indirectement* des effets potentiellement confondants.
3. **Qualité et diversité des sources de données :** les estimations du *GBD* (*Global Burden of Disease*) sont elles-mêmes dérivées d'un ensemble complexe de sources hétérogènes (enquêtes, registres hospitaliers, recensements). Les incertitudes rapportées peuvent sous-estimer ou surestimer la variabilité réelle, suivant les États et les périodes considérées.
4. **Hypothèses distributionnelles :** malgré la robustesse face aux outliers, le modèle suppose un lien *logit*–linéaire (cubique) avec l'âge. Or, dans certaines régions (ou pour certains groupes d'âge), la réalité peut s'éloigner de ce schéma. Les tests de Shapiro–Wilk confirment globalement la normalité approximative des distributions de coefficients, mais de légères déviations peuvent exister dans les États possédant plus d'observations ou, à l'inverse, des échantillons plus restreints. “‘

2.6.5 Pistes d'amélioration et perspectives

Plusieurs développements pourraient affiner l'analyse et compléter la compréhension des phénomènes mis en évidence :

1. **Ajout de covariables pertinentes :** inclure des facteurs de risque (tabagisme, pollution locale, statut socio-économique) et la distinction *homme/femme* améliorerait la spécificité du modèle. Cette étape requiert toutefois des données complémentaires harmonisées à l'échelle de chaque État.

   “‘

2. **Modèles à effets mixtes ou spatio-temporels :** un *modèle hiérarchique* (ou *multilevel*) pourrait introduire des effets aléatoires par État, tout en tenant compte des interrelations spatiales (États voisins, zones urbaines versus rurales). Cela aiderait à mieux cerner la *structure de corrélation* entre territoires et à raffiner l'interprétation cartographique.
3. **Approches non paramétriques :** pour capturer plus librement la relation âge–incidence, on peut envisager des *splines*, des *GAM* (*Generalized Additive Models*) ou des méthodes d'apprentissage automatique. Bien que les modèles paramétriques, comme celui présenté, aient l'avantage d'être plus interprétables, des approches plus flexibles pourraient révéler des motifs plus complexes.
4. **Mise à jour post-pandémie :** il serait pertinent d'examiner la période à partir de 2020 pour évaluer l'impact d'une éventuelle *modification des comportements* (port du masque, distanciation sociale), ainsi que des éventuelles séquelles à long terme chez les patients ayant souffert de COVID-19 (*covid long*). “‘

2.6.6 Conclusion

Cette étude confirme qu'un *modèle cubique* en âge, appliqué à l'échelle *logit* du taux d'incidence, décrit de façon satisfaisante la relation entre l'incidence des maladies respiratoires chroniques et l'âge. L'emploi d'une **régression linéaire robuste** à la perte de Huber, combiné à l'échantillonnage aléatoire des intervalles d'incertitude, offre *une résistance accrue aux valeurs atypiques* et une *meilleure quantification de l'incertitude* par rapport à la régression classique.

Sur le plan *géographique*, les disparités mises en évidence (via les cartes choroplèthes) invitent à explorer les déterminants locaux (environnementaux, socio-économiques, politiques de santé) pour mieux comprendre pourquoi certains États affichent des coefficients plus marqués. Les analyses futures pourraient ainsi s'orienter vers un *modèle enrichi*, intégrant le sexe, la densité de population, la pollution, ou encore l'effet de politiques de santé ciblées, afin de préciser les mécanismes sous-jacents à la dynamique des maladies respiratoires chroniques aux États-Unis.

En somme, la *flexibilité* et la *robustesse* de l'approche présentée permettent de décrire et de comparer la relation âge–incidence à travers des contextes variés. Cette méthode pourrait

servir de base à des *analyses comparatives internationales*, ou s'étendre à d'autres pathologies chroniques où l'effet de l'âge est susceptible de présenter des non-linéarités marquées.

Bibliographie

Robert Andersen. *Modern methods for robust regression*. Number 152. Sage, 2008.

Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics : The Approach Based on Influence Functions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1986.

Mia Hubert, Peter J. Rousseeuw, and Stefan Van Aelst. High-Breakdown Robust Multivariate Methods. *Statistical Science*, 23 :92–119, 2008.

Institute for Health Metrics and Evaluation (IHME). *GBD 2021 Data and Tools Overview*. Institute for Health Metrics and Evaluation, Seattle, WA, May 2024a. URL http://www.healthdata.org/gbd. Updated May 2024.

Institute for Health Metrics and Evaluation (IHME). Global burden of disease 2021 : Findings from the gbd 2021 study, 2024b. URL https://www.healthdata.org/research-analysis/library/global-burden-disease-2021-findings-gbd-2021-study. Accessed : 2025-05-12.

Ricardo A. Maronna, Richard D. Martin, and Víctor J. Yohai. *Robust Statistics : Theory and Methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester, 2006.

Peter J. Rousseeuw. Robust Statistics, Part 1 : Introduction and Univariate Data. Lecture notes, LARS–IASC School, May 2019. Slides.

Peter J. Rousseeuw and Annick Leroy. *Robust Regression and Outlier Detection*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1987.