UCLouvain

epl

École polytechnique de Louvain

Évaluation des interventions non-pharmaceutiques contre le COVID-19 en Allemagne : comparaison de cadres d'inférence statistique et analyse de robustesse

Auteur: **Tessa Rouvez**
Promoteur: **Raphaël JUNGERS**
Lecteurs: **Frédéric Crevecoeur, Guillaume Berger**
Année académique 2025–2026
Master [120]: ingénieur civil en mathématiques appliquées

Résumé

Les politiques sanitaires mises en place pendant la pandémie de COVID-19 ont fait l'objet de nombreuses analyses statistiques. Pourtant, mesurer précisément leur efficacité reste complexe, notamment parce que l'épidémie évolue de manière très différente selon les régions et conserve une forte inertie dans le temps. Ce mémoire étudie ainsi l'influence du modèle statistique sur les conclusions obtenues concernant l'efficacité de ces mesures.

Il repose sur les données épidémiologiques de treize groupements régionaux allemands pendant l'année 2020. Trois approches sont comparées : un modèle mixte fréquentiste classique (`lme4`), un modèle bayésien hiérarchique (`brms/Stan`) et un modèle fréquentiste intégrant une correction autorégressive AR(1) (`glmmTMB`). Des simulations sur données synthétiques viennent compléter cette analyse afin de tester la robustesse des modèles face au problème de causalité inverse.

Les résultats montrent d'abord que les approches fréquentiste et bayésienne donnent pratiquement les mêmes estimations ($r \approx 0{,}9998$). Autrement dit, utiliser un modèle bayésien plus complexe ne corrige pas les problèmes liés à une mauvaise prise en compte de la dynamique temporelle. L'étude montre surtout que lorsque la dépendance entre les jours n'est pas correctement modélisée, les modèles ont tendance à attribuer aux mesures sanitaires une partie de l'évolution naturelle de l'épidémie.

Les simulations confirment ce problème : sans correction temporelle, les modèles classiques peuvent produire jusqu'à 100% de faux positifs. Une fois l'autocorrélation corrigée avec le filtre AR(1), les effets attribués aux mesures sanitaires ne restent plus statistiquement significatifs. Les simulations montrent cependant que cette correction ne réduit pas la capacité du modèle à détecter un véritable effet lorsqu'il existe, avec une puissance de détection proche de 99%.

Cette perte de significativité dans les données réelles ne signifie donc pas que les politiques sanitaires ont été inefficaces. Elle met surtout en évidence les limites des modèles de régression appliqués à des données observationnelles. Ce travail souligne ainsi l'importance de modéliser correctement la dynamique temporelle afin d'éviter des conclusions trompeuses sur l'effet réel des politiques de santé publique.

Déclaration IA

Je déclare sur l'honneur que ce mémoire est le résultat de mon travail personnel et de ma propre réflexion scientifique. Conformément aux principes de responsabilité, de transparence et d'authenticité instaurés par l'Université catholique de Louvain, j'atteste avoir utilisé de manière limitée et encadrée certains outils d'intelligence artificielle générative dans le cadre de ce travail.

Ces outils ont été utilisés pour m'accompagner dans la rédaction du code informatique nécessaire aux modélisations statistiques, à la création de graphiques, ainsi que pour réviser et améliorer la fluidité linguistique de mon texte. Je certifie que ces technologies n'ont agi que comme des outils d'implémentation. Elles ont uniquement servi d'outils d'assistance. Les choix méthodologiques, la construction des modèles, l'analyse des résultats et les conclusions présentées dans ce mémoire relèvent entièrement de ma responsabilité.

Je reconnais également avoir pris connaissance du fait qu'une utilisation non déclarée ou non conforme de ces outils constitue une irrégularité académique au regard du Règlement des études et des examens de l'Université catholique de Louvain (chapitre 4, section 7, articles 107 à 114).

Signature :

Tessa Rouvez

Remerciements

Je tiens tout d'abord à remercier mon promoteur, le Professeur Raphaël Jungers, pour son accompagnement tout au long de ce travail. Je lui suis particulièrement reconnaissante pour sa réactivité à chacune de mes questions, le temps qu'il m'a consacré, mais aussi pour ses encouragements constants, qui m'ont souvent aidée à avancer lorsque je me sentais un peu perdue.

Je remercie également mes lecteurs, Frédéric Crevecoeur et Guillaume Berger, pour le temps consacré à la lecture et à l'évaluation de ce mémoire.

Enfin, un grand merci à ma famille et à mes amis pour leur soutien et leurs encouragements.

Table des matières

Introduction

La pandémie de COVID-19 a été une crise sanitaire inédite qui a affecté la société et l'économie dans le monde entier. Cette crise a surtout montré à quel point la gestion d'une épidémie est complexe. Face à la propagation rapide du virus, les gouvernements ont dû agir, souvent dans l'urgence. Ils ont mis en place des mesures très variées appelées interventions non-pharmaceutiques (NPIs), allant des simples recommandations d'hygiène jusqu'aux confinements stricts.

Aujourd'hui, avec le recul et l'analyse de la littérature scientifique, deux grands éléments ressortent pour mieux comprendre la dynamique de cette pandémie.

D'abord, la pandémie n'a pas frappé partout de la même façon. De nombreuses études montrent que des caractéristiques locales comme la densité de population, la structure démographique ou le niveau d'urbanisation influencent directement la propagation du virus et la mortalité associée. Par exemple, à l'échelle européenne, l'étude de Goujon et al. (2022) [16] montre que les régions urbaines et denses présentent une mortalité plus élevée que les zones rurales. À l'inverse, l'analyse de la mortalité en Belgique par Bourguignon et al. (2024) [4] démontre à l'aide d'un modèle multiniveau que la densité de population a eu une influence très limitée.

D'autre part, toutes les mesures mises en place n'ont pas eu la même efficacité. Plusieurs études montrent que certaines mesures comme les restrictions de contact ou les limitations de rassemblement ont eu un impact beaucoup plus fort que les autres. Par exemple, Khazaei et al. (2023) [21], en utilisant un modèle bayésien hiérarchique sur les données allemandes, ont évalué l'impact de plusieurs mesures sanitaires. Leurs résultats montrent que l'efficacité de ces restrictions a beaucoup varié d'une région à l'autre. De même, García-García et al. (2022) [13] à partir de modèles additifs généralisés appliqués à l'Espagne montrent que certaines mesures ciblées (comme la limitation des rassemblements et les heures de fermeture obligatoires) ont un effet significatif sur la réduction du taux de reproduction du virus.

Problématique et Objectifs

Ces résultats montrent que la pandémie n'évolue pas de la même manière selon les régions et que les mesures sanitaires n'ont pas toutes le même effet dans un environnement qui évolue rapidement au fil du temps. Il devient alors difficile d'évaluer précisément l'impact réel des politiques publiques.

L'enjeu principal n'est donc pas uniquement d'identifier quelles mesures ont été efficaces, mais de comprendre dans quelle mesure ces conclusions dépendent de l'outil statistique utilisé. Les modèles statistiques classiques ne suffisent pas pour gérer toutes ces difficultés en même temps, ce qui justifie l'utilisation d'architectures mathématiques avancées comme les modèles multiniveaux et bayésiens.

Ce mémoire cherche à répondre à la question suivante : *Dans quelle mesure les résultats sur l'efficacité des politiques sanitaires sont-ils robustes face au choix du cadre d'inférence statistique et à la complexité des données réelles ?*

Pour y répondre, ce travail poursuit trois objectifs principaux :

1. **Comparer trois cadres d'inférence différents :** un modèle fréquentiste classique (`lme4`), un modèle bayésien hiérarchique (`brms/Stan`) et un modèle fréquentiste avec correction autorégressive explicite AR(1) (`glmmTMB`).
2. **Appliquer ces modèles à un cas empirique :** l'estimation de l'efficacité des mesures sanitaires en Allemagne durant la première année de la pandémie.
3. **Étudier la robustesse des modèles :** analyser leur stabilité face aux biais spatiaux, aux dépendances temporelles et aux retards administratifs.

Structure du mémoire

Pour répondre à cette problématique, ce mémoire est structuré de manière progressive. Le **chapitre 1** posera les fondements des modèles de régression et expliquera pourquoi la dépendance des données nous oblige à utiliser des modèles multiniveaux. Le **chapitre 2** détaillera la théorie derrière l'inférence bayésienne, indispensable pour intégrer l'incertitude dans nos calculs. Le **chapitre 3** présentera l'architecture des modèles, les données utilisées et le protocole de validation. Le **chapitre 4** sera consacré à l'analyse des résultats. Il abordera d'abord la validation technique des modèles, avant de comparer les différents cadres d'inférence. Ensuite, le **chapitre 5** mettra en évidence les limites méthodologiques fréquemment rencontrées dans la littérature existante. Le **chapitre 6** s'appuiera sur des données synthétiques pour valider notre démarche à l'aide de simulations. Le **chapitre 7** discutera plus largement des implications de nos résultats, de leurs limites et des perspectives pour les recherches futures. Enfin, le **chapitre 8** conclura ce travail.

Chapitre 1

Fondements des modèles de régression multiniveaux

1.1 Cadre général de la modélisation

Les données épidémiologiques présentent une structure complexe : les observations ne sont pas indépendantes, elles varient selon les régions, les périodes et les caractéristiques locales. Pour analyser correctement ce type de données, il est nécessaire d'utiliser des modèles statistiques capables de prendre en compte cette hétérogénéité.

Ce chapitre présente les principaux fondements mathématiques des modèles de régression utilisés dans ce mémoire, en particulier les modèles linéaires généralisés mixtes (GLMM) dans un cadre fréquentiste. Les concepts sont introduits progressivement en s'appuyant sur la méthodologie de Bourguignon et al. (2024) [4]. Nous commencerons par la régression logistique classique et ses limites face à des données structurées, avant d'introduire les modèles multiniveaux à effets aléatoires. Le chapitre suivant prolongera ensuite cette approche dans un cadre bayésien.

1.2 La modélisation des variables binaires

1.2.1 Limites de l'approche linéaire et contraintes de probabilité

Soit une variable aléatoire binaire $Y \in \{0,1\}$ où $Y = 1$ indique la survenue d'un évènement et $Y = 0$ son absence.
L'objectif de la modélisation est d'estimer la probabilité conditionnelle de succès

$p(x)$ en fonction de certaines caractéristiques X (âge, sexe, densité, etc.) :

$$p(x) = P(Y = 1 \mid X = x)$$

Intuitivement, une première approche consisterait à utiliser une régression linéaire classique (de type $Y = \beta X + \epsilon$) mais cette méthode ne convient pas ici : le modèle linéaire classique est conçu pour des variables continues et suppose une distribution normale des erreurs. Si nous l'appliquions ici, nous rencontrerions deux problèmes mathématiques [20] :

— premièrement, une droite de régression s'étend théoriquement à l'infini, elle pourrait donc ici prédire des valeurs hors de l'intervalle $[0, 1]$ (par exemple une probabilité négative) ;

— deuxièmement, le modèle linéaire suppose que l'erreur est constante partout (homoscédasticité). Or, pour une variable binaire, la variance dépend directement de la moyenne. Cette hypothèse fondamentale n'est donc pas respectée.

1.2.2 La régression logistique et la transformation Logit

Pour résoudre ces problèmes, l'approche standard consiste à utiliser la régression logistique. L'idée est de transformer la relation linéaire pour contraindre le résultat final à rester à l'intérieur de l'intervalle $[0, 1]$. Donc plutôt que de modéliser directement la probabilité $p(x)$, nous modélisons les **cotes** (ou *odds*) [20].

Ce changement de modèle entraîne aussi un changement de méthode d'estimation, ce qui permet de résoudre le problème de variance. Alors que les coefficients d'une régression linéaire classique sont estimés par la méthode des moindres carrés, la régression logistique utilise le maximum de vraisemblance. Avec cette approche, il n'est plus nécessaire de supposer que la variance reste constante. Le modèle tient directement compte du fait que la variabilité des données dépend de la probabilité estimée [20].

Définition 1.1 (Odds). Les cotes représentent le rapport entre la probabilité $p(x)$ qu'un événement se produise et la probabilité qu'il ne se produise pas :

$$\text{Odds}(x) = \frac{p(x)}{1 - p(x)} \tag{1.1}$$

L'intérêt de ce rapport est qu'il transforme une probabilité bornée [0,1] en une valeur comprise entre 0 et $+\infty$.
Pour permettre à notre modèle linéaire de couvrir l'ensemble des réels ($-\infty$ à $+\infty$), nous appliquons la fonction logarithme à ces cotes. C'est ce que l'on appelle la fonction logit [20] :

Définition 1.2 (Fonction logit).

$$\text{logit}(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right) \tag{1.2}$$

Le modèle logistique s'écrit alors :

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{1.3}$$

Cette transformation permet de lever la contrainte de bornage tout en conservant une structure linéaire dans les paramètres.

Pour revenir à la probabilité interprétable $p(x)$ à partir des coefficients estimés, l'idée est d'inverser la fonction logit pour obtenir la fonction **sigmoïde** (ou fonction logistique) :

$$p(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+...+\beta_k x_k)}} \tag{1.4}$$

Cette formulation mathématique garantit que, quelles que soient les valeurs prises par les variables explicatives X, la probabilité estimée restera toujours strictement comprise entre 0 et 1 [20].

Interprétation des coefficients : Odds Ratios

L'interprétation des résultats se fait généralement via les rapports de cotes (*Odds Ratios*, OR). Pour les obtenir, il suffit de calculer l'exponentielle des coefficients estimés par le modèle.

Définition 1.3 (Odds Ratio). Pour une variable x_j, l'Odds Ratio associé à son coefficient β_j est donné par :

$$\text{OR}_j = e^{\beta_j} \tag{1.5}$$

— $OR_j > 1$: augmentation du risque
— $OR_j < 1$: effet protecteur
— $OR_j = 1$: absence d'effet

1.3 Limites des modèles standards : la structure hiérarchique et la dépendance spatiale des données

La régression logistique permet donc de traiter correctement une variable binaire mais elle repose, comme la plupart des modèles statistiques standards, sur une hypothèse forte : l'indépendance des observations (i.i.d). Cela signifie que le résidu d'un individu i ne doit pas être corrélé avec celui d'un individu j.

1.3.1 Violation de l'hypothèse d'indépendance

Dans la réalité (encore plus lorsqu'on analyse des données spatiales), cette hypothèse est rarement vérifiée car les données sont naturellement structurées par groupes. Dans notre cas et donc dans le contexte épidémiologique étudié par Bourguignon et al. (2024) [4], les individus ne sont pas des entités isolées : ils habitent dans des communes spécifiques, qui appartiennent elles-mêmes à des provinces ou des régions.

Il existe une dépendance liée au contexte : deux personnes vivant dans la même commune partagent presque le même environnement (densité de population, accès aux soins,..). Elles auront donc tendance à avoir des risques de santé plus proches l'un de l'autre que deux personnes prises au hasard dans le pays. C'est cette ressemblance, créée par l'appartenance au même groupe, qui s'appelle la corrélation intra-classe.

1.3.2 Corrélation intra-classe

Définition 1.4 : Corrélation intra-classe (ICC). La corrélation intra-classe est un indicateur qui mesure à quel point les individus d'un même groupe se ressemblent. Elle calcule la part des différences observées (variance) qui s'explique uniquement par l'appartenance à ce groupe, plutôt que par les caractéristiques individuelles [10].

Si nous appliquons une régression logistique standard sur ces données, nous faisons comme si chaque habitant apportait une information totalement nouvelle et indépendante. Cependant, ce n'est pas correct : deux voisins apportent une information redondante. En ignorant cette structure, le modèle sous-estime l'incertitude (les erreurs-types sont trop faibles). Cela augmente artificiellement le risque de conclure à la significativité statistique d'un effet qui n'existe pas en réalité. Ce phénomène est appelé l'inflation du risque d'erreur de première espèce [20, 37, 22].

1.4 La solution méthodologique : les modèles multiniveaux

Pour corriger ce biais et modéliser correctement l'hétérogénéité territoriale, il est nécessaire d'étendre le modèle logistique à un Modèle Linéaire Généralisé Mixte (GLMM), aussi appelé modèle multiniveau. L'idée centrale de ces modèles est de ne plus voir la variance comme un bloc unique, mais de la décomposer en deux sources distinctes.

1.4.1 Les effets fixes et aléatoires

— **Les effets fixes (β) :** ils correspondent à la partie « moyenne » du modèle. Ils mesurent l'impact des variables explicatives (comme l'âge, le sexe, ou la densité de population) sur la variable réponse, en supposant que cet impact est constant sur l'ensemble de la population étudiée.

— **Les effets aléatoires (u_j) :** ils capturent la variabilité spécifique aux groupes (ici, les communes). Contrairement aux effets fixes, nous ne cherchons pas à estimer la valeur exacte pour chaque commune isolément, mais nous considérons que les communes observées sont un échantillon aléatoire tiré d'une population plus large de communes possibles [37].

1.4.2 Formulation mathématique

Considérons un individu i appartenant à une commune j. En s'inspirant de l'approche de Bourguignon et al. (2024) [4] le modèle logistique multiniveau à ordonnée à l'origine aléatoire s'écrit :

$$\text{logit}(p_{ij}) = \underbrace{\beta_0 + \beta_1 x_{ij} + \beta_2 z_j}_{\text{Partie Fixe}} + \underbrace{u_j}_{\text{Partie Aléatoire}} \tag{1.6}$$

où :

— x_{ij} représente les caractéristiques propres à l'individu (ex : son âge).

— z_j représente les caractéristiques contextuelles connues de sa commune (ex : la densité de population).

— u_j est l'effet aléatoire propre à la commune j.

1.4.3 Propriétés et interprétation de l'effet aléatoire u_j

Ce terme u_j peut être interprété comme une variable latente au niveau du groupe, dans le sens où il capture des facteurs non observés spécifiques à chaque commune j que l'on ne mesure pas directement (par exemple, le fait que les habitants respectent vraiment les mesures ou pas).

Mathématiquement, nous supposons que ces effets aléatoires suivent une distribution normale centrée en zéro :

$$u_j \sim \mathcal{N}(0, \sigma_u^2) \tag{1.7}$$

C'est cette variance σ_u^2 qui détermine si le territoire a une réelle influence et qui quantifie les différences entre les communes.

— Si σ_u^2 est proche de 0, cela signifie que l'endroit où l'on vit n'a pas d'importance : les différences de mortalité s'expliquent uniquement par les caractéristiques individuelles des personnes.

— Si σ_u^2 est élevé, cela indique une forte disparité géographique non expliquée par les variables du modèle (par exemple : à âge et profil égaux, nous avons plus de risques de mourir dans certaines communes que dans d'autres à cause de facteurs locaux inconnus).

Ce terme u_j sert à isoler les caractéristiques locales de chaque commune pour éviter qu'elles ne viennent fausser les résultats généraux (β). En tenant compte de ces particularités locales, le modèle devient plus rigoureux et ajuste sa précision. Nous pouvons donc mesurer à quel point le risque dépend vraiment du lieu de vie, par rapport aux caractéristiques personnelles.

1.5 Transition vers notre cas d'étude : de la variable binaire aux données de comptage

Ce cadre théorique basé sur la régression logistique a permis d'illustrer intuitivement la mécanique des modèles multiniveaux, en prenant pour exemple l'étude de Bourguignon et al. (2024) [4]. Mais dans le cadre de notre application empirique aux données allemandes, la variable d'intérêt n'est plus une probabilité binaire (être infecté ou non) mais un comptage journalier de nouveaux cas.

Le modèle final utilisé dans ce mémoire conservera exactement la même structure d'effets fixes et aléatoires expliquée ci-dessus, mais la distribution mathématique sous-jacente sera adaptée. Nous passerons d'une distribution binomiale à une distribution binomiale négative, conçue pour modéliser des données de comptage présentant une surdispersion (comme le nombre de cas d'infections), ce qui sera détaillé au chapitre 3.

Chapitre 2

Fondements des modèles bayésiens hiérarchiques

Dans le chapitre précédent, nous avons vu comment les modèles multiniveaux permettent de gérer l'hétérogénéité spatiale en traitant les territoires comme des groupes distincts (par exemple, les communes dans le cas de la Belgique). Ils reposent sur une hypothèse classique : les paramètres du modèle (par exemple, l'impact de la densité) sont des valeurs fixes et uniques qu'il faut estimer.

Cependant, cette méthode atteint ses limites face à la complexité et l'incertitude de certaines données épidémiologiques. C'est notamment ce qui justifie l'approche adoptée dans l'étude de Khazaei et al. (2023)[21].

L'approche bayésienne permet de dépasser ces limites en traitant les paramètres non plus comme des constantes, mais comme des variables aléatoires. Elle représente alors l'incertitude sous forme de distributions de probabilité [37].

Dans l'étude de Khazaei et al. (2023) [21] sur la propagation du COVID-19 en Allemagne, cette approche est exploitée via un modèle bayésien hiérarchique permettant d'estimer simultanément la dynamique épidémique, l'effet des interventions non-pharmaceutiques (NPIs) et la variabilité entre les 16 Länder.

Ce chapitre présente les bases de cette méthode, avant de détailler la structure hiérarchique de ces modèles, leur application à l'étude des NPIs, et enfin les méthodes numériques utilisées pour leur estimation.

2.1 Fondements de l'inférence bayésienne

2.1.1 De la valeur fixe à la variable aléatoire

Dans le cadre fréquentiste classique, un paramètre θ est considéré comme une quantité fixe mais inconnue. L'incertitude provient uniquement du fait que l'échantillon observé est aléatoire. L'objectif est alors de calculer une estimation ponctuelle $\hat{\theta}$ (un seul chiffre) qui se rapproche le plus possible de cette « vraie » valeur unique et, éventuellement, un intervalle de confiance basé sur la distribution d'échantillonnage [37, 20].

L'inférence bayésienne inverse cette logique : le paramètre θ est modélisé comme une variable aléatoire. L'objectif n'est plus de trouver un chiffre unique, mais de déterminer sa distribution de probabilité en fonction des données que l'on observe. Cette méthode permet donc de représenter l'incertitude de manière explicite [37].

2.1.2 Théorème de Bayes

Soient y les données observées (par exemple des séries temporelles de cas, hospitalisations, admissions en soins intensifs et décès) et θ les paramètres inconnus que nous cherchons à estimer. L'inférence bayésienne repose sur le théorème de Bayes [14] :

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)} \tag{2.1}$$

où :

— $p(\theta)$ est la **distribution a priori** (*prior*), représentant l'état de connaissance ou d'incertitude sur θ avant l'observation des données ;

— $p(y \mid \theta)$ est la **vraisemblance** (*likelihood*). Elle décrit la probabilité d'observer les données y si les paramètres valaient θ ;

— $p(\theta \mid y)$ est la **distribution a posteriori** (*posterior*), qui combine l'information a priori et l'information apportée par les données. C'est notre connaissance mise à jour (l'objectif final de l'analyse bayésienne) [37] ;

— $p(y)$ est la **vraisemblance marginale** (*marginal likelihood*), donnée par :

$$p(y) = \int p(y \mid \theta)\, p(\theta)\, d\theta \tag{2.2}$$

C'est une constante de normalisation qui assure que l'intégrale de la distribution a posteriori soit bien égale à 1.

En pratique, comme le dénominateur est constant par rapport à θ, nous travaillons souvent à une constante multiplicative près, ce qui conduit à l'expression proportionnelle :

$$p(\theta \mid y) \propto p(y \mid \theta)\, p(\theta) \tag{2.3}$$

2.1.3 Interprétation probabiliste et intervalles de crédibilité

Une fois que nous avons calculé la distribution a posteriori $p(\theta \mid y)$, nous disposons de toute l'information disponible. Si nous avons ensuite besoin d'une valeur unique pour résumer cette distribution, nous pouvons utiliser par exemple la moyenne a posteriori $\mathbb{E}[\theta \mid y]$ ou la médiane [37].

Grâce à l'approche bayésienne, il est aussi facile de quantifier l'incertitude via les intervalles de crédibilité. Un intervalle de crédibilité à 95% est un intervalle $[a, b]$ tel que :

$$P(\theta \in [a, b] \mid y) = 0{,}95 \tag{2.4}$$

Cela signifie qu'une fois les données observées, il y a 95% de probabilité que le vrai paramètre θ se trouve dans cet intervalle [37].

2.2 Modèles hiérarchiques bayésiens

Comme indiqué au chapitre précédent, les données épidémiologiques suivent souvent une structure hiérarchique. Nous n'avons pas des observations isolées mais des mesures répétées dans le temps, regroupées par territoires (communes, provinces) qui sont eux-mêmes dans des ensembles géographiques plus larges (régions, pays). Pour analyser correctement ces données, il faut un modèle qui respecte cette structure emboîtée. C'est ce que permettent les modèles hiérarchiques bayésiens [21].

L'idée est de trouver un bon équilibre : ne pas supposer que toutes les communes sont identiques, mais ne pas non plus les traiter comme des unités totalement indépendantes. Ce partage d'information entre groupes (*partial pooling*) est précisément ce qui rend l'approche de Khazaei et al. (2023) robuste, même pour des régions où les données sont plus rares ou bruitées [37].

2.2.1 Structure hiérarchique à trois niveaux

Un modèle hiérarchique bayésien est généralement construit en trois niveaux :

Niveau 1 : Ce qui est observé (le modèle des données). Pour chaque unité i appartenant au groupe j, l'observation y_{ij} est générée par une loi de probabilité qui dépend d'un paramètre local θ_j. Dans leur étude, Khazaei et al. (2023) utilisent des distributions adaptées (telles que la loi de Poisson ou la loi binomiale négative) pour modéliser le nombre de cas ou de décès. Mathématiquement, ce premier niveau s'écrit :

$$y_{ij} \mid \theta_j \sim p(y_{ij} \mid \theta_j) \tag{2.5}$$

Niveau 2 : Paramètres de groupe (*Prior* hiérarchique). Nous considérons que les paramètres locaux θ_j ne sont pas indépendants, mais qu'ils sont issus d'une même distribution commune, elle-même dirigée par des hyperparamètres ϕ :

$$\theta_j \mid \phi \sim p(\theta_j \mid \phi), \quad j = 1, \ldots, J \tag{2.6}$$

Cette hypothèse formalise l'**échangeabilité** des groupes conditionnellement à ϕ. Elle contraint les estimations locales à rester cohérentes avec la moyenne globale, ce qui empêche les groupes ayant peu de données de produire des résultats extrêmes ou aberrants [37, 14].

Niveau 3 : Hyperparamètres (*Hyperprior*). Les hyperparamètres ϕ (qui contrôlent la moyenne et la variance générale entre les groupes) ne sont pas fixés arbitrairement. Ils possèdent eux-mêmes une distribution de probabilité a priori :

$$\phi \sim p(\phi) \tag{2.7}$$

Ce dernier niveau capture l'incertitude sur la structure d'ensemble. Il permet au modèle d'apprendre si les régions sont globalement très hétérogènes ou au contraire très similaires [14, 37].

2.2.2 Distribution a posteriori conjointe et partial pooling

En combinant les trois niveaux (données, groupes, hyperparamètres) avec le théorème de Bayes, nous obtenons la distribution a posteriori conjointe. Elle rassemble l'ensemble de l'information disponible sur tous les paramètres en même temps.

Mathématiquement, nous avons :

$$p(\theta_1, \ldots, \theta_J, \phi|y) \propto \underbrace{\left[\prod_{j=1}^{J} \prod_{i} p(y_{ij}|\theta_j)\right]}_{\text{Vraisemblance (Données)}} \times \underbrace{\left[\prod_{j=1}^{J} p(\theta_j|\phi)\right]}_{\text{Distribution des groupes (\textit{prior} hiérarchique)}} \times \underbrace{p(\phi)}_{\text{Hyperprior}} \tag{2.8}$$

Cette formule permet au modèle de trouver un juste milieu entre :

— *No pooling* : aucune mise en commun. Chaque région est estimée séparément, ce qui peut produire des résultats très instables pour les petits groupes.

— *Complete pooling* : mise en commun totale. Toutes les régions sont considérées comme identiques, ce qui ignore l'hétérogénéité locale.

Le modèle hiérarchique se situe entre ces deux extrêmes grâce au *partial pooling* (mise en commun partielle). Si une commune dispose de peu de données, le modèle évite de se baser uniquement sur ces observations limitées et va combler ce manque

d'information en s'appuyant sur la moyenne nationale (se rapprochant du *complete pooling*). À l'inverse, si une commune a beaucoup de données, le modèle lui fait confiance et respecte ses spécificités locales (se rapprochant du *no pooling*) [37, 21, 14].

2.3 Application aux NPIs : ordonnées et pentes aléatoires

Pour bien modéliser la pandémie, il est nécessaire de définir mathématiquement comment les mesures sanitaires s'appliquent sur le territoire.

Dans notre modèle principal, nous utilisons une approche avec **ordonnée à l'origine aléatoire** (*Random Intercept*). Concrètement, cela signifie que le modèle permet à chaque Land d'avoir son propre point de départ (une incidence de base différente, capturée par l'effet spatial). En revanche, l'effet propre de chaque loi sanitaire (noté β_k) est considéré comme un effet fixe et global. Nous partons de l'hypothèse qu'une politique fédérale, comme la fermeture des écoles, a la même efficacité théorique partout en Allemagne.

Toutefois, comme le soulignent Khazaei et al.(2023) [21], l'adhésion de la population à une règle peut varier selon la culture ou la politique locale. Pour en tenir compte, il est possible de complexifier le modèle en ajoutant des **pentes aléatoires** (*Random Slopes*). Dans ce cas, nous supposons que l'efficacité $\alpha_{k,j}$ de la mesure k dans le Land j n'est plus fixe, mais suit une distribution normale :

$$\alpha_{k,j} \sim \mathcal{N}(\mu_k, \sigma_k^2) \tag{2.9}$$

où :

— μ_k est l'efficacité moyenne de la loi au niveau national.

— σ_k^2 mesure à quel point cette efficacité varie d'une région à l'autre.

Grâce à cette formule, le modèle calcule un coefficient ajusté pour chaque région, tout en utilisant le mécanisme de lissage spatial (*partial pooling*) pour éviter de tirer des conclusions extrêmes basées sur trop peu de données [21].

Dans le chapitre 4, nous utiliserons cette approche complexe à pentes aléatoires comme un test de robustesse. Cela nous permettra de vérifier mathématiquement notre hypothèse initiale : l'efficacité des restrictions a-t-elle vraiment été homogène sur tout le territoire, ou y a-t-il eu des déviations régionales significatives ?

2.4 Variables latentes et choix de modélisation

L'une des principales difficultés dans la modélisation du COVID-19 est que le nombre réel d'infections n'est jamais observé directement. Les cas rapportés ne

correspondent qu'à une partie des infections réelles, car ils dépendent notamment du nombre de tests réalisés, des stratégies de dépistage et des délais de déclaration.

Face à cette difficulté, certains travaux, comme celui de Khazaei et al. (2023) [21], introduisent une variable latente $I_{j,t}$ représentant le nombre réel d'infections. Cette variable est estimée à partir de plusieurs sources de données (cas, hospitalisations, admissions en soins intensifs et décès), en intégrant explicitement les délais épidémiologiques via des modèles probabilistes dynamiques.

Cependant, l'intégration de ce type de modèle implique une complexité computationnelle importante et repose sur des hypothèses structurelles fortes concernant la dynamique de transmission.

Dans le cadre de ce mémoire, nous adoptons une approche plus parcimonieuse en modélisant directement la variable des cas observés. Afin de tenir compte, de manière indirecte, des délais entre l'infection réelle et sa détection, nous introduisons un décalage temporel strict (*lag* de 7 jours) sur les variables explicatives. Comme le justifient Hunter et al. (2021) [19], ce délai permet d'approximer à la fois le temps d'incubation biologique et le retard administratif des tests. Par ailleurs, la dynamique globale de l'épidémie est capturée à l'aide de fonctions de lissage temporel (splines naturelles), permettant d'absorber les grandes tendances non observées des vagues épidémiques.

Si cette simplification ne permet pas d'interprétation directe en termes de dynamique épidémiologique sous-jacente (comme le nombre exact d'infections réelles ou l'estimation du taux de reproduction R_t), elle offre un compromis entre interprétation, robustesse et simplicité. Notre modèle se concentre ainsi uniquement sur l'association statistique directe entre les politiques publiques et les cas observés.

2.5 Estimation numérique : méthodes MCMC

2.5.1 Problème de solutions analytiques

En statistique bayésienne, l'estimation des paramètres repose sur le calcul de la distribution *a posteriori*. Cependant, l'évaluation exacte de cette distribution demande le calcul de la vraisemblance marginale $p(y)$. Ce terme implique la résolution d'une intégrale multidimensionnelle sur l'ensemble de l'espace des paramètres Θ :

$$p(y) = \int_{\Theta} p(y \mid \theta)\, p(\theta)\, d\theta \tag{2.10}$$

Dans la plupart des applications, cette intégrale est impossible à résoudre avec des méthodes classiques [37, 14].

2.5.2 Méthodes MCMC et estimation bayésienne

Comme la distribution *a posteriori* ne peut pas être calculée directement dans la plupart des modèles complexes, il faut passer par des méthodes d'approximation numérique. Pour cela, les approches bayésiennes utilisent généralement les méthodes de Monte-Carlo par chaînes de Markov (MCMC). Le principe est de générer progressivement une suite de valeurs $(\theta^{(1)}, \theta^{(2)}, \dots)$ qui finit par reproduire la distribution recherchée $p(\theta \mid y)$ [14].

Le principe fondamental de ces algorithmes dérive de la méthode de Metropolis-Hastings. À chaque itération t, l'algorithme propose une nouvelle valeur candidate θ^* à partir de la valeur courante $\theta^{(t)}$, selon une loi de proposition $q(\theta^* \mid \theta^{(t)})$. Cette nouvelle valeur est ensuite acceptée ou rejetée avec une probabilité α, calculée à partir du rapport des densités [8] :

$$\alpha = \min\left(\frac{p(\theta^*|y)q(\theta^{(t)}|\theta^*)}{p(\theta^{(t)}|y)q(\theta^*|\theta^{(t)})}, 1\right) \tag{2.11}$$

Si ces premières approches bayésiennes reposaient sur des propositions aléatoires locales (à l'image de la marche aléatoire de Metropolis-Hastings), les modélisations contemporaines exigent des outils plus performants.

L'approche utilisée dans notre étude (via le package `brms` [6]) repose sur le moteur de calcul probabiliste `Stan`. Il utilise l'algorithme **Hamiltonian Monte Carlo (HMC)**, et plus précisément sa variante adaptative **NUTS (No-U-Turn Sampler)**. Cette méthode est aujourd'hui utilisée dans la littérature récente en modélisation épidémiologique [3, 33].

Contrairement aux méthodes d'exploration locale purement aléatoires, HMC s'appuie sur la forme de la distribution pour orienter les propositions de nouvelles valeurs vers les zones les plus probables. Cela permet d'explorer l'espace des paramètres de manière plus efficace, en particulier dans des modèles complexes avec de nombreuses dimensions [17].

Comme l'algorithme démarre à partir d'un point arbitraire, les premières itérations servent à atteindre une zone stable de la distribution. Cette étape correspond à la phase de chauffe (*burn-in*) et les échantillons générés pendant cette période ne sont pas conservés pour l'analyse finale. Une fois la chaîne stabilisée, elle peut alors explorer correctement la distribution *a posteriori* [14].

Afin de consolider la qualité de l'inférence, Khazaei et al. (2023) [21] ont exécuté 8 chaînes de Markov parallèles comportant chacune 50 000 itérations. Cette redondance limite fortement le risque que l'algorithme reste piégé dans un optimum local et renforce la confiance dans la convergence et la stabilité des estimations.

Dans notre implémentation, nous avons adapté cette puissance de calcul à notre modèle en utilisant 4 chaînes parallèles de 2 000 itérations (dont 1 000 de *burn-in*). Cette approche est plus simple à faire tourner mais elle reste largement suffisante et donne des résultats fiables, comme nous le verrons avec les diagnostics de convergence au chapitre 4.

Chapitre 3

Méthodologie

3.1 Données et variables

3.1.1 Source des données et période étudiée

Les données utilisées dans ce travail proviennent du dépôt public mis à disposition par Rehms et Khazaei (2023) [29], qui regroupe les sources utilisées par l'étude de Khazaei et al. [21]. Ce jeu de données correspond à une série temporelle journalière décrivant l'évolution de l'épidémie ainsi que les différentes mesures mises en place dans les seize Länder allemands. Les données épidémiologiques sont initialement issues de l'Institut Robert Koch [30]. Pour garantir un nombre suffisant d'observations, certaines régions ont été regroupées (ex : Berlin-Brandenburg), ce qui fait au final une analyse de 13 groupements régionaux. L'étude porte uniquement sur la période allant du 19 février 2020 au 31 décembre 2020. Ce choix permet de se concentrer sur les deux premières vagues de l'épidémie, juste avant l'arrivée de la vaccination et des nouveaux variants, afin d'isoler au mieux l'effet des interventions non-pharmaceutiques.

3.1.2 Traitement des variables spatiales et administratives

Pour prendre en compte la dimension géographique de l'épidémie, la densité de population de chaque région a été ajoutée comme variable de contrôle. En vérifiant le code source de l'étude originale, nous avons repéré une erreur dans les données démographiques : la population attribuée à Thuringe (2 910 875) correspondait en réalité à une duplication des données du Schleswig-Holstein. Cette valeur a donc été corrigée et remplacée par la valeur officielle de 2 120 237 habitants, issue des registres officiels Destatis de 2020 [35]. Pour vérifier l'impact de cette correction, le modèle a aussi été estimé avec la valeur erronée initiale. Les résultats restent quasiment inchangés, avec des écarts inférieurs à 0,001 sur les coefficients des NPIs.

Cela montre que cette erreur n'a pas influencé les conclusions. La correction a toutefois été conservée pour garantir la cohérence et la qualité des données.

La variable de densité de population a ensuite été standardisée (centrée et réduite). Comme cette variable possède une échelle très différente des indicateurs binaires des mesures sanitaires, cette transformation permet d'optimiser la convergence des algorithmes d'estimation et de garantir la stabilité numérique du modèle.

En plus de la dimension spatiale, les données épidémiologiques journalières présentent des biais liés à l'organisation administrative. Les capacités variables des laboratoires et des autorités sanitaires locales engendrent des délais de déclaration fréquents [31]. En pratique, cela se traduit souvent par une sous-déclaration des cas durant le week-end, suivie d'un rattrapage en début de semaine. Afin d'éviter que le modèle n'interprète ces variations administratives comme des évolutions réelles de l'épidémie, une variable catégorielle indiquant le jour de la semaine (*weekday*) a été ajoutée aux modèles, conformément à l'approche de Khazaei et al. (2023) [21].

3.1.3 Mesures sanitaires (NPIs)

Les mesures sanitaires étudiées sont codées sous forme binaire : 0 si la mesure est inactive, et 1 si elle est active. Huit interventions différentes sont prises en compte :

1. **SchoolClosure** : fermeture partielle ou totale des établissements scolaires (primaires et secondaires).
2. **Curfew** : couvre-feu et restriction de sortie (interdiction de quitter son domicile sans motif valable).
3. **MaskShoppingcenters** : obligation du port du masque dans les centres commerciaux et les points de vente.
4. **RestaurantClosure** : fermeture des restaurants (à l'exception de la vente à emporter et de la livraison).
5. **RestaurantTestRequired** : accès aux restaurants autorisé uniquement sur présentation d'un test négatif.
6. **EventsUpTo100** : limitation des événements publics en intérieur à un maximum de 100 personnes.
7. **ContactrestrictionLessThan5** : restriction des contacts à un maximum de 5 personnes (à l'exception des personnes d'un même foyer).
8. **ContactrestrictionStrict** : restriction stricte des contacts aux seules personnes d'un même foyer.

En plus de ces mesures, une variable indicatrice appelée **generalBehavioralChanges** a été ajoutée. Elle permet de prendre en compte les changements de

comportement spontanés de la population, comme la distanciation sociale, qui ne sont pas directement liés aux décisions politiques. Cette variable s'active (prend la valeur 1) dans chaque Land dès l'entrée en vigueur de la toute première intervention non-pharmaceutique (majoritairement à la mi-mars 2020) et garde sa valeur jusqu'à la fin de la période d'observation.

Une attention particulière doit aussi être portée au nombre de jours pendant lesquels chaque mesure est active. La plupart des NPIs sont en place sur de longues périodes : le port du masque dans les centres commerciaux cumule 1855 Länder-jours, la fermeture des écoles 1661, et la fermeture des restaurants 1165. En revanche, la restriction stricte des contacts (**ContactrestrictionStrict**) est beaucoup moins fréquente. Elle n'a été activée que pendant 98 observations (Länder-jours), ce qui ne représente qu'environ 2,4% des 4 022 observations totales de la période étudiée. Cela signifie qu'une quantité d'information limitée est disponible pour estimer précisément son effet. C'est ce qui explique en partie l'instabilité de son coefficient lors des analyses de robustesse.

3.1.4 Le décalage temporel (lag)

Lorsqu'une mesure sanitaire est mise en place, son effet n'apparaît pas immédiatement dans les données épidémiologiques. Les cas observés à une date donnée correspondent en réalité à des infections survenues plusieurs jours auparavant. D'après Hunter et al. (2021) [19], la période d'incubation du COVID-19 est d'environ 5 jours en moyenne. À cela s'ajoutent encore le délai d'apparition des symptômes, la réalisation du test PCR et l'enregistrement administratif du résultat. Au total, le délai observé entre l'infection et son apparition dans les statistiques est d'environ une semaine.

Pour tenir compte de cette réalité, nous avons appliqué un décalage temporel (*lag*) de **7 jours** à toutes les variables liées aux mesures sanitaires. Autrement dit, le modèle explique le nombre de cas observés à un jour donné en fonction des mesures qui étaient en place une semaine plus tôt.

3.2 Architecture des modèles

3.2.1 Distribution binomiale négative et surdispersion

La variable d'intérêt, notée $Y_{i,t}$, correspond au nombre de nouveaux cas dans le Land i au jour t. Comme il s'agit d'un nombre de cas, nous sommes dans un cadre de données de comptage. Le modèle le plus simple dans ce cas est la régression de Poisson. Mais ce modèle suppose que la moyenne et la variance sont égales, ce qui

est rarement le cas en pratique.

Dans nos données, les cas de COVID-19 varient énormément d'un jour à l'autre, notamment à cause des événements de super-propagation (*clusters*) ou des campagnes de dépistage irrégulières. Pour le vérifier, un test préliminaire a été effectué sur les données et montre une forte surdispersion (ratio de 73,51), ce qui indique une variance beaucoup plus grande que la moyenne. Le modèle de Poisson n'est donc pas adapté ici. Nous utilisons par conséquent une distribution Binomiale Négative, qui permet de mieux gérer cette variabilité grâce à un paramètre de dispersion ϕ :

$$Y_{i,t} \sim \text{NB}(\mu_{i,t}, \phi) \tag{3.1}$$

3.2.2 Effet spatial et modèle mixte (Partial Pooling)

Chaque Land a ses propres spécificités locales (politique de dépistage, infrastructures, etc.) qui influencent le niveau de base des infections. Pour prendre en compte ces différences tout en gardant un modèle parcimonieux, nous avons utilisé un modèle à effets mixtes.

Concrètement, nous avons ajouté une ordonnée à l'origine aléatoire (*Random Intercept*, codée `(1 | country)`) pour chaque Land. L'avantage principal de cette structure est le lissage partiel (*Partial pooling*). L'algorithme se sert des informations issues des grands Länder ayant beaucoup de données (comme la Bavière) pour stabiliser les estimations des Länder plus petits (comme la Sarre). Cela rend les résultats spatiaux beaucoup plus fiables et robustes.

3.2.3 Contrôle temporel (splines naturelles)

Sur le plan temporel, l'épidémie évolue de manière progressive sous forme de vagues. Au départ, des variables liées aux saisons (hiver, printemps, été, automne) avaient été envisagées pour modéliser le temps, mais cette approche forçait l'algorithme à créer des sauts brusques lors des changements de saison.

Pour éviter ce problème, nous avons utilisé une spline naturelle (`ns(date_num, df=4)`). Cette fonction permet de tracer une courbe lisse, avec quatre degrés de liberté, qui suit l'évolution globale de l'épidémie dans le temps. Grâce à cela, le modèle capture la tendance générale sans perturber l'estimation de l'effet propre des mesures sanitaires.

Une analyse de sensibilité sur ce paramètre a été réalisée en faisant varier le nombre de degrés de liberté (df = 4, 6, 8, 10). Les résultats montrent que l'autocorrélation résiduelle ne dépend pas de ce choix. Ce point sera discuté plus en détail au chapitre 4.

3.2.4 L'équation globale

L'équation finale du modèle relie l'incidence moyenne $\mu_{i,t}$ aux différentes variables explicatives :

$$\log(\mu_{i,t}) = \log(P_i) + \beta_0 + u_i + \sum_{k=1}^{8} \beta_k X_{k,i,t-7} + \gamma B_{i,t} + \delta D_i + \sum_{m=1}^{6} \theta_m W_{m,t} + f(t) \quad (3.2)$$

Dans cette formule :

- $\log(P_i)$ est le logarithme de la population du Land, utilisé comme *offset* pour modéliser un taux d'incidence par habitant plutôt qu'un volume brut de cas ;
- β_0 est l'ordonnée à l'origine globale ;
- $u_i \sim \mathcal{N}(0, \sigma_u^2)$ représente l'effet aléatoire propre au Land i (*random intercept*) ;
- β_k sont les coefficients des huit NPIs évaluées, avec le décalage de 7 jours (dans le modèle glmmTMB AR(1), **RestaurantTestRequired** étant exclue en raison d'un nombre insuffisant de données) ;
- γ et δ mesurent respectivement l'impact du changement de comportement citoyen ($B_{i,t}$, dont le moment d'activation dépend du Land) et de la densité de population standardisée (D_i) ;
- les paramètres θ_m sont les effets fixes associés aux variables indicatrices des jours de la semaine ($W_{m,t}$) afin de contrôler le biais de déclaration administratif (effet week-end) ;
- $f(t)$ est la fonction non linéaire modélisant la dynamique temporelle globale de l'épidémie grâce à une spline naturelle.

3.2.5 Cadres d'inférence statistique : fréquentiste, bayésien et autorégressif

Définition 3.1 : cadres d'inférence statistique. L'inférence statistique regroupe l'ensemble des méthodes permettant de tirer des conclusions sur des paramètres inconnus à partir de données observées, tout en quantifiant l'incertitude associée à ces estimations, notamment par la construction d'intervalles de confiance ou de crédibilité [37].

Dans le contexte de ce travail, l'objectif n'est pas d'établir une inférence causale pure, mais plutôt de vérifier si nos estimations de l'efficacité des mesures sanitaires (ainsi que leur incertitude) restent stables lorsque nous changeons de méthode statistique. Notre équation globale a donc été évaluée avec trois approches d'inférence

différentes.

La première approche repose sur un cadre fréquentiste classique via un Modèle Linéaire Généralisé Mixte (GLMM). Les paramètres de ce modèle ont été estimés par la méthode du maximum de vraisemblance, en utilisant la fonction `glmer.nb` du package `lme4`.

Ensuite, une approche bayésienne hiérarchique a été implémentée à l'aide du package `brms`, basé sur le langage probabiliste `Stan`. Au sein de ce modèle, les paramètres sont estimés par échantillonnage de Monte-Carlo par chaînes de Markov (MCMC). Contrairement à l'approche fréquentiste, cette méthode permet d'intégrer des connaissances préalables sous forme de distributions *a priori* (*priors*).

Pour limiter le surapprentissage sans imposer de contraintes trop fortes au modèle, nous avons utilisé des *priors* faiblement informatifs. Une distribution normale centrée réduite a ainsi été appliquée aux coefficients fixes :

$$\beta_k, \gamma, \delta, \theta_m \sim \mathcal{N}(0, 1)$$

Pour l'ordonnée à l'origine, nous avons conservé le *prior* par défaut du package `brms`, qui correspond à la loi de Student-$t(3,\ 0,\ 2.5)$ [6]. Comme l'expliquent Veenman et al. [37], cette distribution possède en effet des queues plus épaisses qu'une loi normale classique, ce qui offre au modèle une plus grande flexibilité face aux éventuelles valeurs extrêmes ou atypiques.

Enfin, une approche fréquentiste a été testée en ajoutant **une structure autorégressive de premier ordre, noté AR(1)**. Les données épidémiologiques présentent en effet une inertie importante : le nombre de cas au jour t dépend fortement du nombre de cas au jour $t-1$. Même si la spline naturelle capte la dynamique globale des vagues, elle ne suffit pas toujours à corriger cette dépendance journalière. Nous avons donc implémenté un troisième modèle via le package `glmmTMB` qui intègre un terme AR(1) indépendant à chaque Land. Ce modèle permet ainsi de tester si les effets des NPIs restent significatifs même après correction de cette autocorrélation.

3.2.6 Diagnostic et validation

Pour garantir que les résultats ne sont pas faussés par des problèmes techniques, plusieurs tests ont été réalisés.

La suffisance des données. Avant de valider les modèles, il était important de vérifier qu'il y avait assez de données pour estimer correctement les paramètres.

Dans ce travail, 13 régions sont observées quotidiennement sur près d'un an. Cela correspond à environ 191 observations par paramètre pour le Modèle A et 183 pour le Modèle C (cette légère différence s'explique par le fait que le Modèle C inclut un paramètre supplémentaire ρ pour l'autocorrélation temporelle AR(1)). Ces valeurs sont largement au-dessus des recommandations de la littérature, qui suggèrent un minimum d'environ 10 observations par paramètre pour garantir des estimations fiables et éviter le surapprentissage [28].

Le surapprentissage bayésien. Dans la même logique, pour notre modèle bayésien (`brms`), une validation croisée de type Leave-One-Out (PSIS-LOO) a été utilisée. Cette méthode estime la capacité prédictive du modèle face à de nouvelles données. La fiabilité de cette estimation est évaluée par le paramètre de Pareto k, qui mesure l'influence de chaque observation sur les résultats [38]. Obtenir des valeurs $k < 0{,}7$ garantit qu'aucune donnée isolée n'a une influence disproportionnée sur les résultats, confirmant l'absence de surapprentissage et la stabilité du modèle [20].

Le pouvoir explicatif du modèle (R^2). Pour évaluer la qualité globale du modèle, nous utilisons le coefficient de détermination (R^2). Cet indicateur permet de mesurer la proportion de la variabilité des données qui est capturée par le modèle [20].

Dans le cas des modèles à effets mixtes (GLMM), le calcul classique du R^2 n'est cependant pas directement applicable. Nous utiliserons donc la méthode de calcul du R^2 conditionnel développée par Nakagawa et Schielzeth (2013) [27]. Cette approche permet de quantifier la variance expliquée par l'architecture globale en combinant à la fois les effets fixes (l'impact des mesures sanitaires) et les effets aléatoires (l'hétérogénéité géographique entre les Länder).

Le test de colinéarité (VIF). Étant donné que plusieurs mesures sont souvent mises en place en même temps, il existe un risque de multicolinéarité qui pourrait masquer l'effet spécifique de chaque intervention. Le test du Facteur d'Inflation de la Variance (VIF) permet de vérifier l'absence de dépendance trop forte entre ces variables explicatives. Un VIF égal à 1 indique l'absence totale de colinéarité. Des valeurs inférieures à 10 sont généralement considérées comme acceptables et garantissent l'absence de problème de colinéarité susceptible de fausser les estimations [20, 19].

L'autocorrélation spatiale (indice de Moran). L'indice de Moran permet d'évaluer dans quelle mesure les valeurs d'une variable sont similaires entre des zones géographiquement voisines [26]. Plus précisément, il mesure la corrélation entre

la valeur observée dans une zone et la moyenne (pondérée) des valeurs observées dans les zones proches. Appliqué aux résidus du modèle via le package `DHARMa`, ce test vérifie que les erreurs de prédiction ne sont pas regroupées géographiquement (absence de clusters d'erreurs). Le respect de cette condition confirme que l'ordonnée à l'origine aléatoire par Land suffit à capturer l'hétérogénéité spatiale de l'épidémie.

L'inflation de zéros et la dispersion. Toujours à l'aide du package `DHARMa` (`testDispersion` et `testZeroInflation`), nous avons évalué la qualité de la distribution binomiale négative via des tests de dispersion et d'inflation de zéros. Cela permet de vérifier que le modèle gère correctement les irrégularités liées aux données administratives (comme les sous-déclarations des week-ends) [6].

L'autocorrélation temporelle (ACF). Les modèles de régression classiques supposent que les erreurs sont indépendantes les unes des autres. Avec des données temporelles, cette hypothèse peut facilement être violée : les erreurs d'un jour peuvent rester liées à celles des jours précédents. Lorsque cette dépendance n'est pas prise en compte, le modèle a tendance à sous-estimer les erreurs types. Les intervalles de confiance deviennent alors artificiellement trop étroits, ce qui augmente le risque de conclure à tort qu'un effet est significatif [20].

Pour vérifier cette hypothèse d'indépendance, l'autocorrélation temporelle des résidus a été analysée à l'aide de corrélogrammes (ACF) [20]. Un corrélogramme montre la corrélation entre les résidus d'un jour donné et ceux des jours précédents (appelés *lags*). Le graphique affiche aussi des bandes de confiance à 95 %, représentées par des lignes pointillées. Lorsque les barres restent à l'intérieur de ces limites, les résidus peuvent être considérés comme indépendants. À l'inverse, si certaines barres dépassent ces seuils, cela indique la présence d'une dépendance temporelle significative.

Diagnostics de convergence MCMC. La convergence du modèle bayésien est évaluée à l'aide de trace plots, qui représentent l'évolution des chaînes MCMC au fil des itérations. Ils permettent de vérifier que les chaînes explorent correctement l'espace des paramètres et atteignent une distribution stationnaire. Une bonne convergence se traduit par des chaînes qui se mélangent bien, sans tendance apparente [37]. La convergence est aussi testée par le critère $\hat{R}$, qui compare la variance entre les différentes chaînes MCMC à la variance au sein de chaque chaîne. Selon les recommandations récentes de Vehtari et al. (2021) [39], une valeur $\hat{R} < 1{,}01$ est exigée pour garantir que toutes les chaînes ont correctement convergé vers la même distribution stationnaire.

Analyse de sensibilité. Enfin, une analyse de sensibilité a été réalisée sur le décalage temporel. Plutôt que de figer le modèle sur un délai unique de 7 jours, le modèle a été intégralement recalculé sous quatre scénarios de latence (4, 7, 10 et 14 jours). Cela permet de vérifier si les résultats sont robustes ou s'ils dépendent du choix du délai.

Chapitre 4

Résultats

Ce chapitre présente les résultats de nos trois modèles. Nous commencerons par vérifier que ces derniers sont fiables et bien construits, avant d'analyser les estimations de l'efficacité des politiques sanitaires et leur stabilité entre les cadres d'inférence.

4.1 Diagnostics et validation des modèles

Avant d'interpréter les résultats, il faut s'assurer que les modèles sont techniquement fiables. Cette section présente l'ensemble des diagnostics réalisés.

Convergence MCMC (Trace Plot). Contrairement à l'approche fréquentiste, qui cherche une seule solution optimale, l'approche bayésienne reconstruit la distribution complète des paramètres en explorant l'espace des solutions via les chaînes de Markov. Il est donc important de vérifier que cet algorithme a correctement convergé.

La figure 4.1 montre l'évolution des quatre chaînes MCMC pour deux variables clés : la restriction stricte des contacts et la fermeture des écoles. La première a été choisie car elle est au cœur du phénomène de causalité inverse, tandis que la seconde correspond à une mesure très discutée politiquement.

Les quatre chaînes se mélangent bien, sans dérive ni comportement particulier visible. Cela indique que l'algorithme a bien convergé et qu'il explore de manière stable les différentes valeurs possibles des paramètres. Les valeurs de $\hat{R} = 1$ pour l'ensemble des paramètres confirment cette bonne convergence.

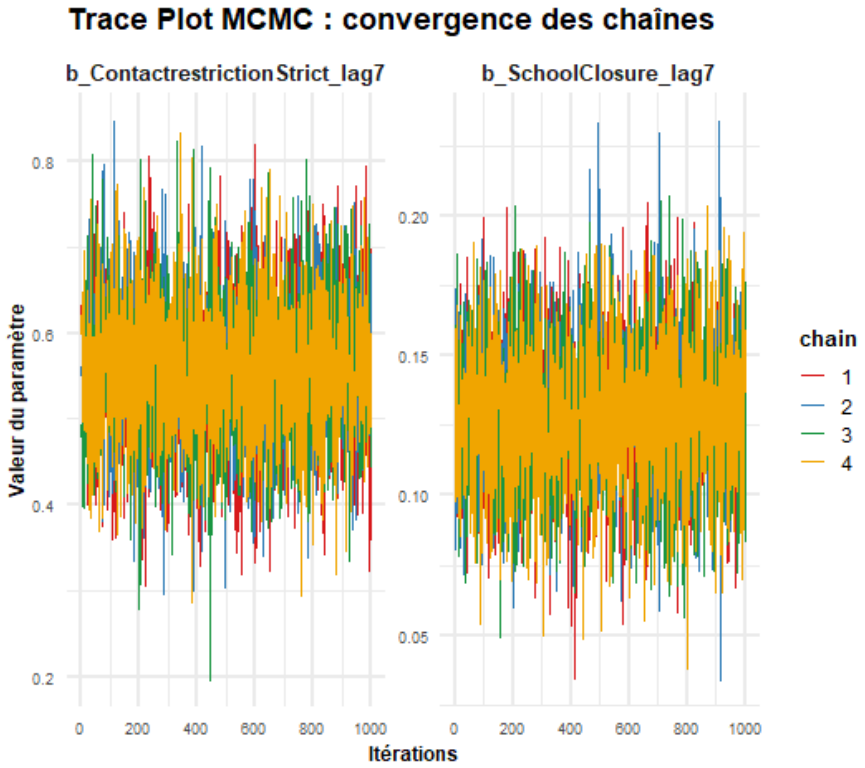
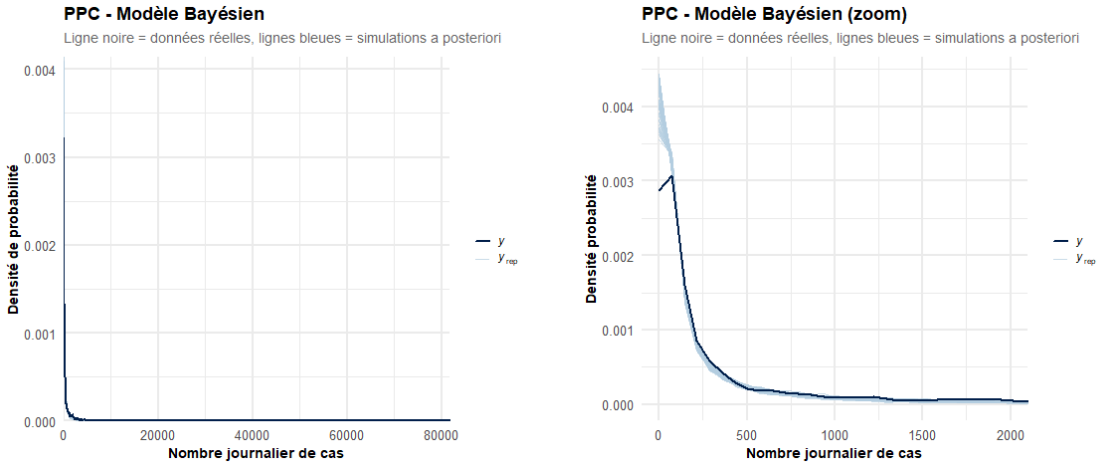


FIGURE 4.1 – **Trace Plot des chaînes MCMC pour l'estimation de la restriction stricte des contacts et de la fermeture des écoles.** Les quatre chaînes de Markov se mélangent parfaitement pour former une bande horizontale dense et homogène, sans aucune tendance visible. Cela confirme que l'algorithme a correctement convergé vers une distribution stationnaire stable.

Ajustement de la distribution (Posterior Predictive Check). Pour valider le choix de la loi binomiale négative et vérifier si notre modèle est capable de reproduire les données réelles, nous avons réalisé un contrôle prédictif a posteriori (PPC), présenté à la figure 4.2. L'objectif est de vérifier si le modèle est capable de générer des données (lignes bleues) dont la forme correspond aux données réelles observées (ligne noire). Dans notre cas, les distributions sont très proches, ce qui montre que le modèle capture bien la forme globale des données. En zoomant sur les valeurs proches de zéro, nous pouvons voir que la concentration de jours à « zéro cas » est globalement absorbée par la loi de surdispersion, même si la sous-déclaration des cas durant le week-end génère encore un léger écart résiduel.



(a) Vue large (b) Vue zoomée sur les zéros

FIGURE 4.2 – **Contrôle prédictif *a posteriori* (PPC) du modèle bayésien.** La vue zoomée (b) illustre l'ajustement du modèle face à la sous-déclaration administrative des week-ends.

Absence de surapprentissage (PSIS LOO-CV). Étant donné la complexité du modèle (plusieurs variables, effets aléatoires, spline), il existe un risque de surapprentissage (*overfitting*). Pour vérifier cela, une validation croisée *Leave-One-Out* (approximée par PSIS-LOO) a été utilisée. La figure 4.3 montre les valeurs du paramètre k de Pareto pour l'ensemble des 4 022 observations de notre jeu de données. Tous les points se situent largement en dessous du seuil critique de 0,7 (la grande majorité oscillant entre -0,3 et 0,3). Cela signifie qu'aucune observation particulière (par exemple un pic de cas isolé) ne domine les résultats. Le modèle reste donc stable et généralise correctement. De plus, le modèle estime le nombre de paramètres effectifs à $p\_loo = 32.4$, une valeur cohérente avec la structure de notre équation.

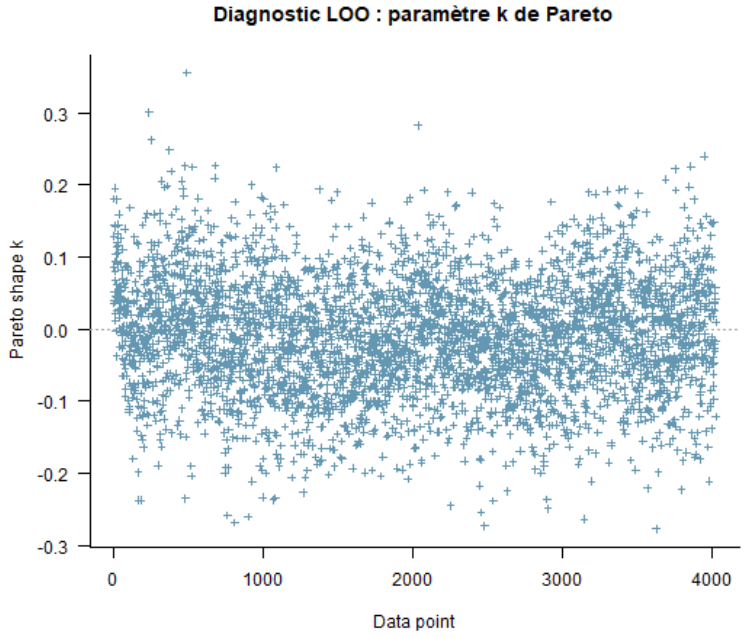


FIGURE 4.3 – **Diagnostic de validation croisée (PSIS-LOO) du modèle bayésien.** Le graphique évalue l'influence de chaque observation via le paramètre k de Pareto. Tous les points se situent en dessous du seuil critique de 0,7. Ce résultat confirme qu'aucune donnée individuelle n'a une influence disproportionnée sur l'estimation, garantissant ainsi la stabilité du modèle et l'absence de surapprentissage.

Colinéarité et stabilité du modèle (VIF). Étant donné que plusieurs mesures sanitaires (NPIs) ont souvent été mises en place en même temps, il existait un risque élevé de multicolinéarité. La stabilité de l'équation globale a été vérifiée par le calcul du Facteur d'Inflation de la Variance (VIF). Les valeurs du VIF pour les huit NPIs et la variable de comportement restent toutes inférieures à 2,5 (entre 1,0 et 2,31 pour le Modèle A, entre 1,0 et 2,29 pour le Modèle B). La spline temporelle introduit logiquement un peu plus de corrélation, mais cela reste dans des niveaux acceptables (VIF = 7.16 pour le Modèle A, 4.48 pour le Modèle B), ce qui est mathématiquement attendu. En pratique, le modèle est donc capable d'isoler l'effet propre de chaque loi, malgré leurs activations simultanées.

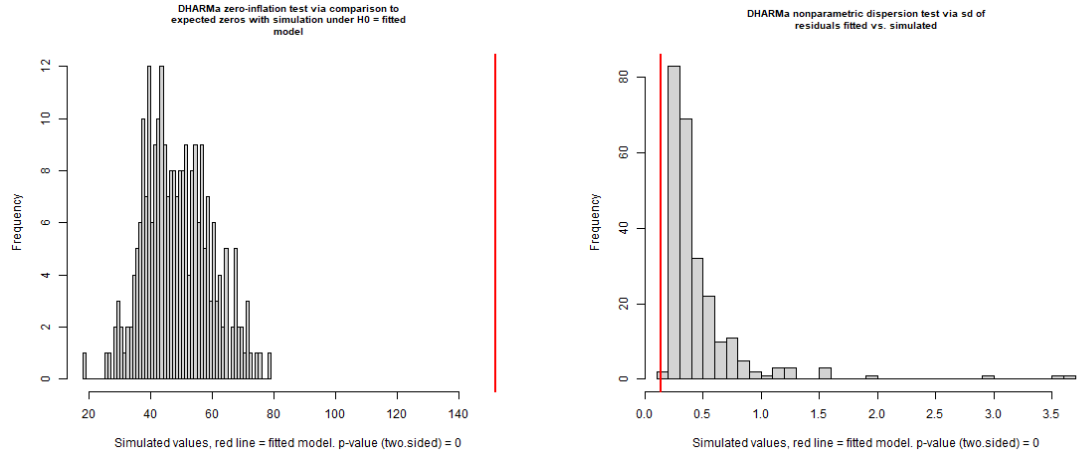
Pouvoir explicatif (R^2). Pour quantifier la part de variance expliquée, nous avons calculé le R^2 conditionnel de Nakagawa. Le Modèle B (bayésien) atteint un score de 79,7 % (CrI 95 % [77,3 ; 82,1]). L'équation parvient ainsi à expliquer près de 80 % de la variance d'un phénomène épidémiologique pourtant très instable.

Indépendance spatiale (indice de Moran). Nous avons aussi vérifié que les erreurs du modèle ne sont pas spatialement corrélées par l'Indice de Moran (via le package `DHARMa`). L'indice de Moran calculé sur les résidus des treize groupements régionaux retourne une valeur de -0,045 avec une p-value non significative de 0,575. L'hypothèse d'autocorrélation spatiale est donc rejetée. L'introduction de la densité de population et de l'ordonnée à l'origine aléatoire (`1 | country`) par Land a donc bien neutralisé les effets géographiques. La justification de ce choix de modélisation est détaillée à la section 4.2, qui illustre concrètement pourquoi chaque Land nécessite son propre niveau de base.

Limites liées aux données (DHARMa). Même si le modèle capture bien les différences entre régions, l'analyse des résidus met en évidence certaines limites liées aux données elles-mêmes.

Premièrement, le test d'inflation de zéros (4.4a) indique que le nombre de jours sans cas déclarés est environ trois fois plus élevé que ce que le modèle prédit sous la loi binomiale négative (ratio = 3,09, p-value $< 0{,}001$). Comme l'illustrait déjà le graphique PPC, cette anomalie s'explique par les biais administratifs, notamment les week-ends où les tests sont moins nombreux (la sous-déclaration systématique du dimanche et du lundi).

Deuxièmement, le test de dispersion (4.4b) indique que la variance résiduelle n'est pas parfaitement capturée (`dispersion` = 0,30, p-value $< 0{,}001$). En cherchant à absorber l'hétérogénéité des données épidémiologiques, le modèle s'attend à plus de variation qu'il n'en reste une fois celui-ci ajusté. Bien que notre variable de contrôle *weekday* lisse ce phénomène en moyenne, ces deux diagnostics combinés suggèrent qu'un modèle encore plus complexe (comme un modèle avec inflation de zéros) pourrait améliorer l'ajustement.



(a) **Test d'inflation de zéros** (b) **Test de dispersion résiduelle**

FIGURE 4.4 – La ligne rouge observée se situe en dehors des distributions simulées (histogrammes) pour les deux tests ($p < 0{,}001$)

Autocorrélation temporelle (ACF). Il s'agit du diagnostic le plus important de cette section. Jusqu'ici, tous les autres diagnostics valident l'architecture du modèle. Cependant, l'analyse des corrélogrammes sur les résidus des Modèles A et B révèle un problème structurel majeur sur les treize groupements régionaux.

Malgré l'ajout de la spline naturelle, qui capture correctement les grandes vagues de l'épidémie, une autocorrélation temporelle importante reste présente. Les erreurs de prédiction du jour t restent fortement liées à celles du jour précédent, $t - 1$.

Cela montre que le modèle ne parvient pas à capturer complètement l'inertie quotidienne de l'épidémie. Or, les mesures sanitaires sont souvent mises en place pendant les périodes de forte circulation du virus. Comme le modèle ne prend pas explicitement en compte cette inertie, il risque d'attribuer une grande partie de cette baisse naturelle de la transmission aux seules interventions gouvernementales [3, 9]. Les effets des mesures sanitaires peuvent donc paraître artificiellement plus importants qu'ils ne le sont réellement.

Pour corriger ce point, nous avons introduit un troisième modèle avec une structure autorégressive (Modèle C), présenté dans la section 4.5.

Ce problème ne concerne pas uniquement nos données. Les études de référence traitent cette dépendance temporelle de différentes façons. Khazaei et al. (2023) [21] l'intègrent à travers une variable latente dynamique. García-García et al. (2022) [13], de leur côté, utilisent des splines temporelles dans un modèle additif généralisé (GAM), une approche assez proche de celle utilisée ici, mais qui ne

corrige pas explicitement la dépendance résiduelle entre jours consécutifs. D'autres études importantes, comme Hunter et al. (2021) [19] ou Liu et al. (2021) [24], reconnaissent ne pas avoir modélisé cette inertie, ce que nous discutons plus en détail au chapitre 5.

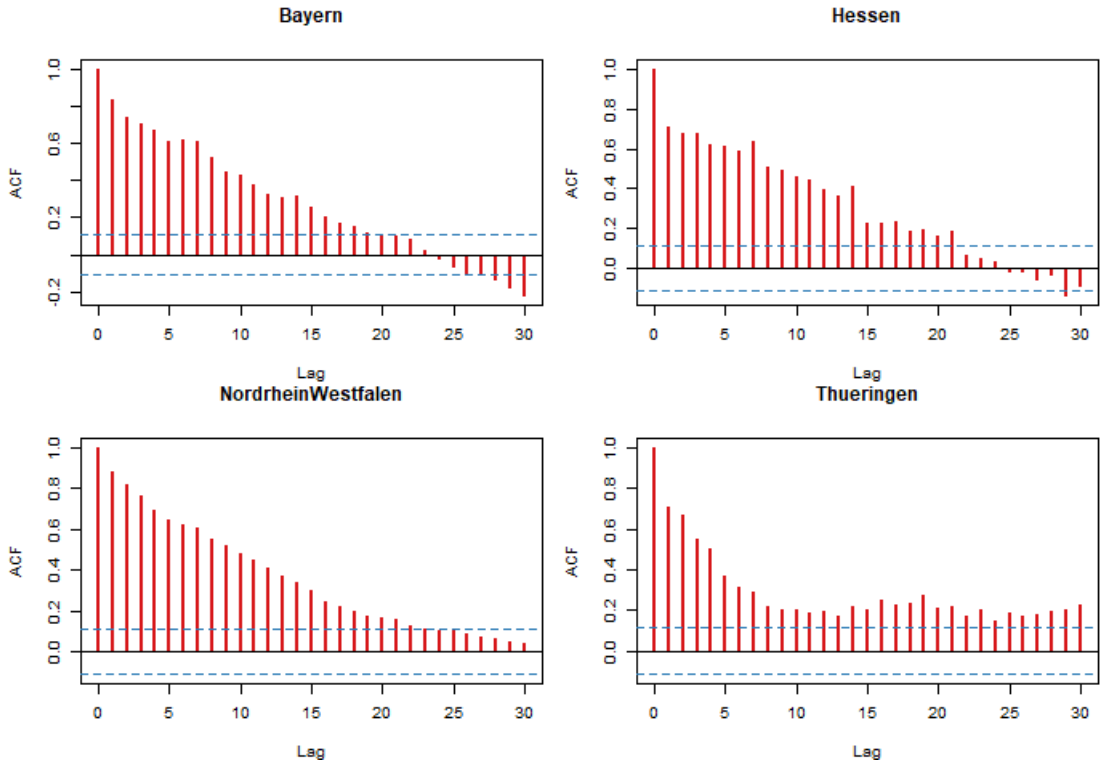


FIGURE 4.5 – **Corrélogrammes (ACF) des résidus du Modèle A pour quatre régions représentatives.** (Les graphiques pour l'ensemble des 13 régions sont disponibles en annexe A.4). Les barres d'autocorrélation dépassent partout les seuils de significativité (lignes pointillées bleues) sur de nombreux jours consécutifs. Cela démontre que dans le modèle standard l'erreur de prédiction d'un jour reste fortement corrélée à celle de la veille.

Sensibilité aux degrés de liberté de la spline. Le choix initial de (df=4) repose sur l'idée de capturer les grandes variations saisonnières observées au cours de l'année. Afin de vérifier que ce choix n'influence pas les conclusions, nous avons répété l'analyse avec des splines plus flexibles (df=6, 8 et 10).

La figure 4.6 montre que les corrélogrammes des résidus restent pratiquement identiques dans les quatre scénarios. Quelle que soit la flexibilité de la spline, une forte autocorrélation résiduelle persiste. Ces résultats indiquent que le problème ne provient pas du choix des degrés de liberté, mais de l'absence d'une correction

explicite de la dépendance temporelle.

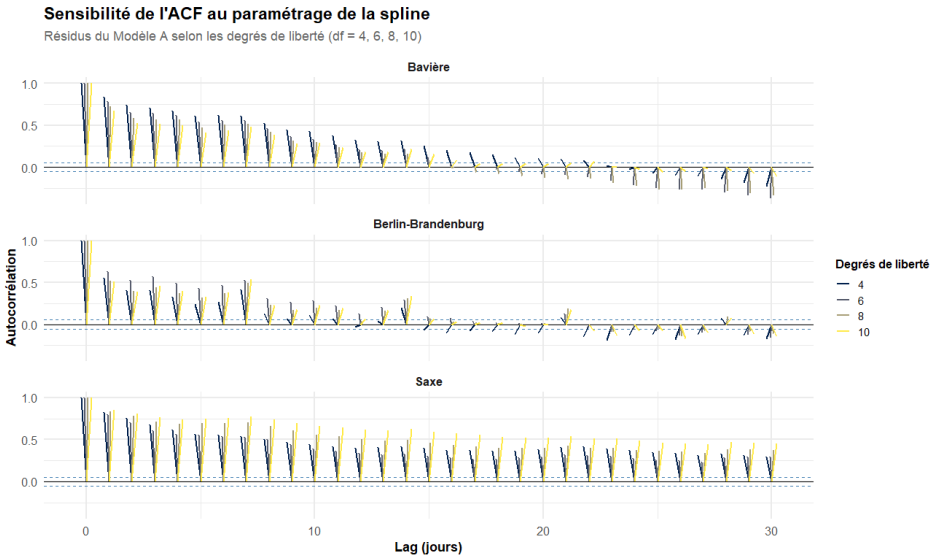


FIGURE 4.6 – **Analyse de sensibilité de l'autocorrélation des résidus (ACF) selon les degrés de liberté (*df*) de la spline.** Dans les quatre scénarios, l'autocorrélation persiste au-delà des bandes de confiance (lignes pointillées bleues). Cela montre que modifier les degrés de liberté de la spline ne permet pas de corriger ce problème.

4.2 Analyse spatiale et hétérogénéité régionale

Bien que l'analyse de l'autocorrélation temporelle ait montré les limites des Modèles A et B pour tirer des conclusions causales globales, ces architectures restent utiles pour étudier la dimension spatiale de manière isolée. En effet, l'intégration simultanée d'une correction temporelle AR(1) et d'une analyse spatiale détaillée dans un seul modèle pose des problèmes importants de convergence , comme l'illustrera l'échec du Modèle D bayésien (voir section 4.5). Il est donc plus rigoureux de procéder par étapes méthodologiques successives.

Analyser d'abord les différences entre régions, avant d'introduire la correction temporelle (qui sera faite avec le Modèle C), permet de vérifier que les hypothèses spatiales sont cohérentes et de justifier une éventuelle simplification du modèle.

4.2.1 Cas concret : la Sarre

Pour comprendre pourquoi un effet par Land a été introduit, il est utile d'examiner le cas de la Sarre (figure 4.7). Ce graphique met en évidence que la Sarre (en jaune) suit la même évolution que le reste de l'Allemagne (en noir) : une première vague au printemps, puis une seconde beaucoup plus forte à l'automne et en hiver.

Par contre, les niveaux d'incidence ne sont pas du tout les mêmes. Le reste du pays dépasse largement les 2000 cas en moyenne journalière, alors que la Sarre reste en dessous de 500.

Si nous avions utilisé un modèle classique sans effet aléatoire, le modèle aurait simplement estimé une moyenne globale pour l'ensemble de l'Allemagne. Cela aurait conduit à surestimer le niveau de cas dans les petits Länder comme la Sarre et à sous-estimer celui des grandes régions.

Avec l'ajout d'une ordonnée à l'origine aléatoire, chaque Land conserve son propre niveau de base. Le modèle peut alors mieux distinguer les différences régionales et isoler plus correctement l'effet des mesures sanitaires.

Ce choix est validé statistiquement par le test de Moran présenté en section 4.1, qui confirme que l'ordonnée à l'origine aléatoire suffit à neutraliser les effets géographiques.

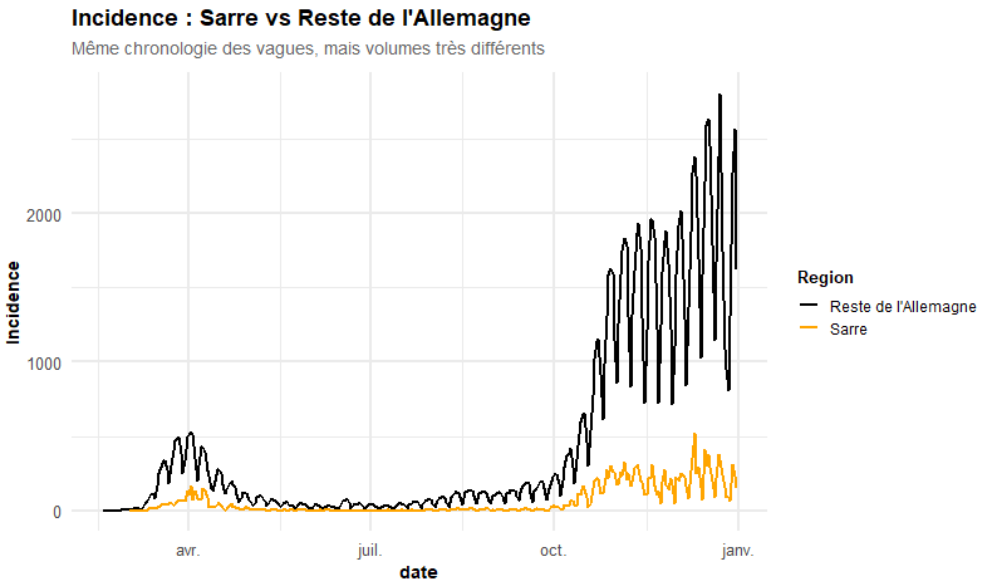


FIGURE 4.7 – **Comparaison de l'incidence journalière entre Sarre et le reste de l'Allemagne.** Les deux courbes suivent une chronologie épidémique identique mais présentent des volumes de cas très différents. Cela justifie le fait d'utiliser un modèle mixte avec une ordonnée à l'origine aléatoire par Land.

4.2.2 Homogénéité de l'efficacité des mesures (Pentes aléatoires)

Ensuite, nous avons voulu vérifier si les mesures sanitaires avaient le même effet partout en Allemagne, ou si leur efficacité variait selon les régions. Pour tester cette hypothèse, nous avons ajouté des pentes aléatoires (*Random Slopes*) sur la restriction stricte des contacts (l'une des mesures les plus importantes et débattues). Cette spécification permet à l'algorithme de calculer un score d'efficacité différent pour chaque région.

Le résultat est présenté à la figure 4.8. Les points bleus montrent l'écart d'efficacité de chaque Land par rapport à la moyenne nationale. Nous pouvons constater que tous les points sont très proches de zéro et cela suggère que l'efficacité de cette restriction a été pratiquement identique sur tout le territoire allemand. Ce résultat est important, car il confirme que notre choix initial (utiliser des effets fixes globaux pour les mesures) est cohérent [1].

1. Sur le plan algorithmique, l'ajout de pentes aléatoires rend le modèle beaucoup plus complexe. Lors de l'échantillonnage bayésien, cela a entraîné quelques transitions divergentes (9 au total). Même si ce nombre reste marginal, cela indique que le modèle atteint ses limites de complexité et

La conclusion principale reste toutefois très claire : l'effet moyen varie peu d'une région à l'autre et reste proche de zéro. Cela justifie l'utilisation de coefficients globaux dans le modèle principal.

Pour vérifier que ce résultat ne concernait pas uniquement cette mesure, le même test a également été appliqué à d'autres politiques importantes disposant de davantage de données, notamment la fermeture des écoles et l'obligation du port du masque. Les résultats, présentés en annexe A.3, montrent que même avec une incertitude beaucoup plus faible, les différences d'efficacité entre régions restent très limitées. Cela renforce l'idée qu'un modèle à effets fixes constitue une approximation pertinente pour l'ensemble des mesures sanitaires.

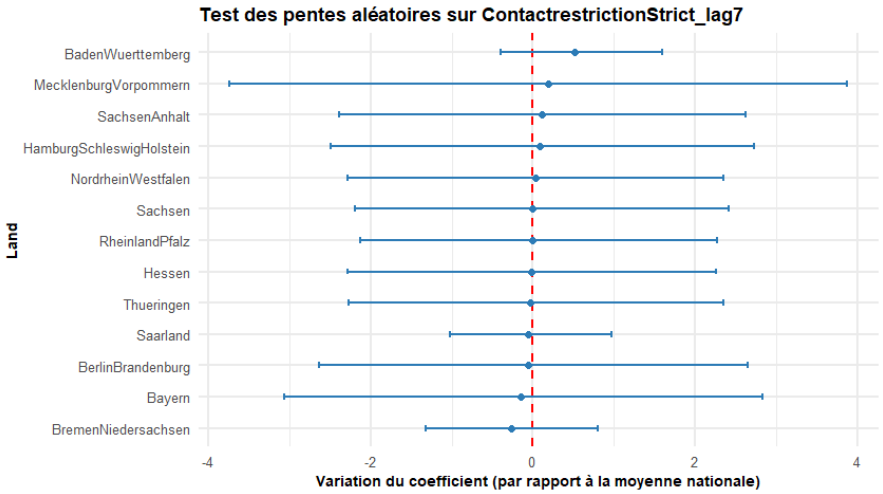


FIGURE 4.8 – **Test des pentes aléatoires : variation régionale de l'efficacité de la restriction stricte des contacts.** Le graphique montre l'écart d'efficacité de la mesure pour chaque Land par rapport à la moyenne nationale (ligne pointillée rouge). Toutes les barres d'incertitude croisent le zéro, cela confirme que l'impact de la loi est similaire dans toutes les régions, ce qui justifie d'utiliser un effet fixe pour toute l'Allemagne.

que les intervalles de crédibilité doivent être interprétés avec prudence.

Bilan de l'architecture spatiale : Ces analyses confirment l'architecture de notre modèle. D'une part, les fortes disparités des niveaux d'incidence de base entre les régions justifient l'introduction d'une ordonnée à l'origine aléatoire. D'autre part, l'analyse des pentes aléatoires démontre que l'effet des mesures sanitaires reste globalement homogène sur l'ensemble du territoire. L'ajout de pentes aléatoires n'est par conséquent pas nécessaire.

4.3 Stabilité algorithmique : fréquentiste vs bayésien

Cette section répond à une question importante du mémoire : le choix de la méthode statistique modifie-t-il les conclusions sur l'efficacité des mesures sanitaires ? Pour y répondre, nous avons comparé les estimations du Modèle A (`lme4`) et du Modèle B (`brms`).

Le tableau 4.1 met côte à côte les coefficients estimés pour les différentes mesures sanitaires. Il apparaît que les valeurs obtenues sont quasiment identiques. Les écarts sont très faibles, toujours inférieurs à 0,01. La corrélation de Pearson entre ces deux séries d'estimations atteint d'ailleurs $r = 0{,}9998$ [2].

Globalement, les deux modèles produisent donc des résultats équivalents, que ce soit sur le signe des effets, leur intensité ou leur classement.
La figure 4.9 illustre également cette similitude. Chaque point, représentant une NPI, est parfaitement aligné sur la diagonale. Cela confirme que les deux approches arrivent exactement aux mêmes conclusions.

2. La mesure `RestaurantTestRequired` a été retirée de cette comparaison car elle est trop peu présente dans les données, ce qui rend son estimation très instable dans le cadre bayésien.

Mesure Sanitaire (NPI)	Modèle A (Fréquentiste)	Modèle B (Bayésien)	Écart absolu
Restriction contacts < 5	-0,119	-0,122	0,003
Restriction contacts stricte	0,554	0,558	0,004
Couvre-feu	0,073	0,076	0,003
Événements ≤ 100 personnes	0,072	0,071	0,001
Masque (centres commerciaux)	-0,102	-0,109	0,007
Fermeture restaurants	-0,048	-0,042	0,006
Fermeture écoles	0,132	0,129	0,003
Test requis restaurants	*Exclu — insuffisance de données*		

TABLE 4.1 – **Comparaison des coefficients estimés par les cadres fréquentiste (Modèle A) et bayésien (Modèle B).** Les écarts absolus entre les deux approches ont une différence maximale de 0,007 sur l'échelle logarithmique. Cette quasi-équivalence démontre que les deux cadres d'inférence produisent des résultats identiques.

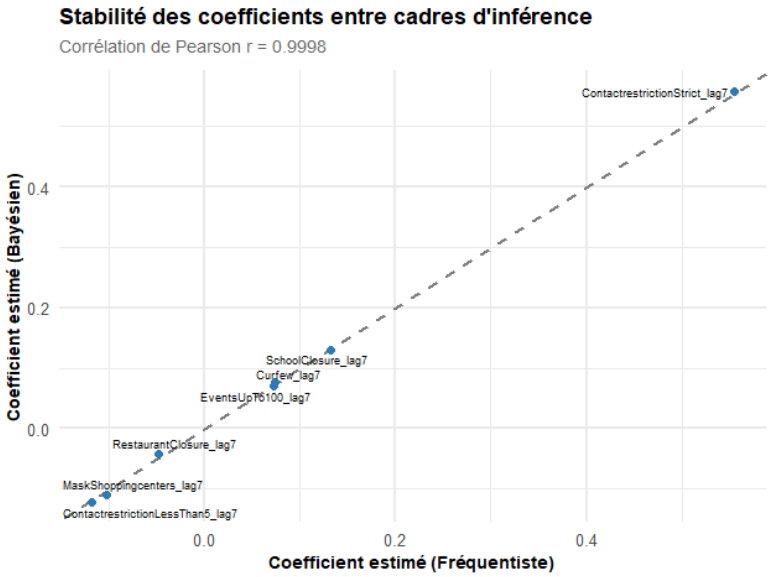


FIGURE 4.9 – **Scatter Plot de concordance entre les estimations fréquentistes et bayésiennes.** Tous les points s'alignent parfaitement sur la diagonale (ligne pointillée), illustrant une corrélation de Pearson quasi-absolue ($r = 0{,}9998$).

4.3.1 Comparaison de l'incertitude

Une autre question importante était de savoir si le modèle fréquentiste sous-estimait l'incertitude par rapport au modèle bayésien. Les intervalles de confiance fréquentistes et les intervalles de crédibilité bayésiens ne s'interprètent pas de la même manière [3].

Malgré cette différence théorique, comparer la largeur de ces intervalles reste utile pour voir si une méthode apparaît artificiellement plus précise que l'autre.

Les résultats montrent que ce n'est pas le cas. La largeur des intervalles est presque identique entre les deux approches. Le ratio moyen entre la largeur des intervalles bayésiens (CrI 95 %) et fréquentistes (CI 95 %) est de 1,001, soit une différence d'à peine 0,1 % en moyenne.

La figure 4.10 confirme ce constat. Pour chaque politique sanitaire, les barres d'incertitude bayésiennes (en bleu) et fréquentistes (en rouge) sont presque de la même taille. Aucun modèle ne produit des intervalles systématiquement plus larges ou plus étroits que l'autre.

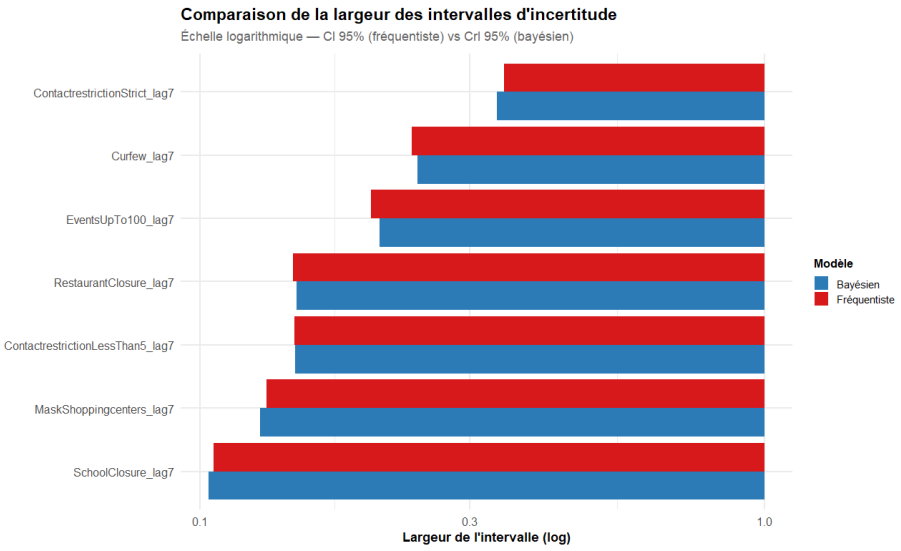


FIGURE 4.10 – **Comparaison de l'incertitude (largeur des intervalles à 95 %) entre les cadres fréquentiste et bayésien, sur une échelle logarithmique.** Pour chaque mesure sanitaire, les deux approches génèrent des barres d'une longueur quasi identique.

3. Pour rappel, l'intervalle de crédibilité bayésien (défini à la section 2.1.3) signifie directement qu'il y a 95 % de probabilité que le paramètre se situe dans l'intervalle compte tenu des données [37]. À l'inverse, un intervalle de confiance fréquentiste indique que si l'expérience était répétée un grand nombre de fois, 95 % des intervalles construits contiendraient la vraie valeur du paramètre.

Bilan algorithmique : L'analyse démontre que le choix entre une approche fréquentiste et bayésienne ne modifie ni les estimations, ni leur incertitude. Dans le cadre de ces données, la complexité calculatoire de l'échantillonnage MCMC et l'utilisation de distributions *a priori* n'apportent aucune plus-value par rapport à un modèle mixte classique résolu par maximum de vraisemblance.

4.4 Efficacité des mesures sanitaires et causalité inverse

Pour interpréter concrètement les résultats, les coefficients du modèle ont été convertis en pourcentages de variation de l'incidence. Cela permet de mieux comprendre l'effet estimé des mesures sanitaires.

Nous comparons ensuite un modèle exploratoire simple, sans spline et basé sur des variables saisonnières, au modèle final intégrant le lissage temporel. Cette démarche permet d'illustrer à quel point la manière de modéliser la dynamique de l'épidémie influence l'interprétation des effets des mesures sanitaires.

4.4.1 Le paradoxe de la restriction stricte (causalité inverse)

Sur le premier graphique exploratoire (figure 4.11), la restriction stricte des contacts semblait très efficace, affichant une réduction des cas d'environ 36 %. Cependant, sur notre modèle final (figure 4.12), ce résultat change complètement : la mesure bascule à gauche de l'axe zéro, ce qui correspond à une augmentation des cas. [4]

Ce résultat inattendu suggère un biais statistique bien connu en modélisation épidémiologique : la **causalité inverse** (ou endogénéité). En modélisant correctement la dynamique de l'épidémie avec les splines, le modèle capture mieux la tendance temporelle de fond, ce qui révèle le biais de causalité inverse.

En effet, les autorités ont généralement instauré ces restrictions strictes lorsque la situation sanitaire était déjà très critique. Le modèle capte ainsi une forte corrélation positive entre l'activation de la mesure et un niveau d'incidence élevé, confondant ainsi la cause et la conséquence.

Autrement dit, les décisions politiques réagissent à l'épidémie tout autant que l'épidémie réagit à ces décisions. Dans ce cas, il devient très difficile de distinguer

4. Sur l'échelle logarithmique, un coefficient positif $\beta > 0$ implique $e^{\beta} > 1$, soit une réduction $(1 - e^{\beta}) \times 100 < 0$, c'est-à-dire une augmentation de l'incidence.

ce qui relève réellement de l'effet de la mesure.

Cette observation met en évidence la limite principale de cette approche. Le modèle permet bien d'identifier des associations globales et de révéler des biais, comme la causalité inverse. En revanche, avec ce type de données, il reste difficile d'isoler clairement l'effet propre des mesures sans prendre en compte plus finement la dynamique temporelle de l'épidémie.

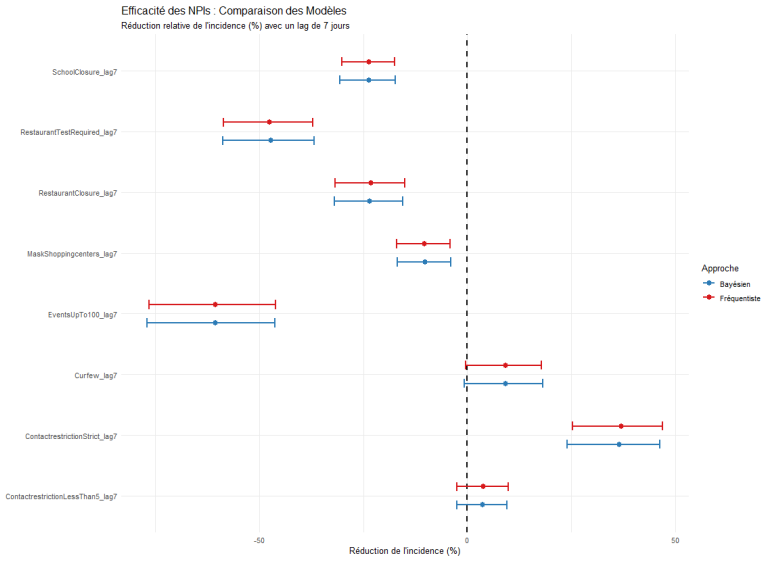


FIGURE 4.11 – **Modèle sans splines.** Sans filtre temporel, le Modèle A tendance à attribuer l'évolution de la courbe épidémique aux seules mesures sanitaires. Les coefficients risquent alors d'absorber l'effet d'autres facteurs non mesurés.

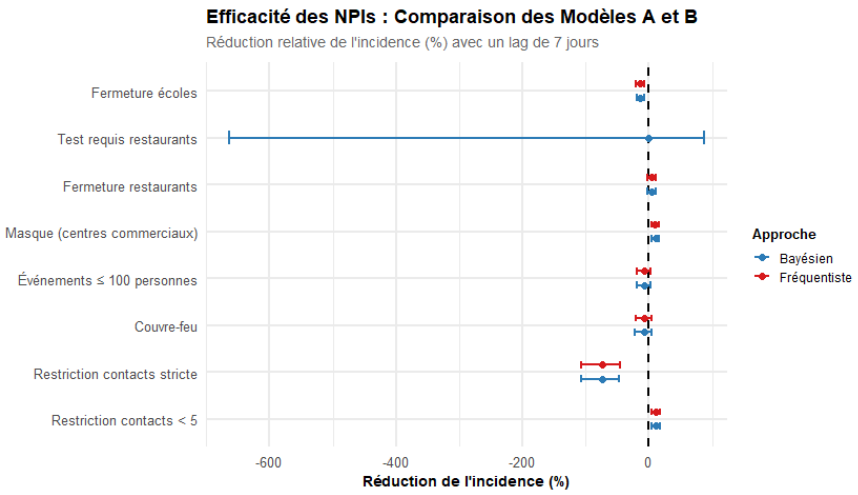


FIGURE 4.12 – **Modèle final (avec splines temporelles).** L'intégration des splines permet de lisser et de capter la dynamique épidémique de fond. Les coefficients des NPIs reflètent ainsi plus fidèlement l'impact spécifique de chaque mesure.

4.4.2 Analyse des autres mesures

L'observation du modèle final (figure 4.12) met en évidence plusieurs points intéressants. D'abord, les résultats fréquentistes (points rouges) et bayésiens (points bleus) confirment visuellement la stabilité des estimations évoquée précédemment.

Ensuite, nous pouvons remarquer que la variable `RestaurantTestRequired` possède une barre d'incertitude très élevée. Ceci s'explique par le fait que cette mesure a été trop peu utilisée pour que le modèle puisse estimer son effet de manière fiable.

Enfin, pour la majorité des autres lois (fermeture des écoles, couvre-feu, restrictions de rassemblement), les effets estimés sont proches de zéro. Cela suggère que leur effet propre, une fois la dynamique globale de l'épidémie prise en compte, est difficile à isoler statistiquement.

4.4.3 Analyse de sensibilité temporelle

Dans toutes les sections précédentes, nous avons supposé un délai d'incubation et de déclaration de 7 jours. Cependant, la biologie du virus et les délais administratifs des tests varient dans la réalité. Pour s'assurer que nos conclusions et notamment le phénomène de causalité inverse observé à la section précédente ne dépendent pas trop de ce choix, les modèles ont été ré-estimés avec plusieurs décalages : 4, 7, 10 et 14 jours.

La figure 4.13 présente les résultats obtenus pour plusieurs délais. L'objectif n'est pas ici de déterminer un délai "optimal", mais plutôt de vérifier si les estimations restent stables lorsque ce paramètre varie.

Le graphe montre que les coefficients changent peu d'un scénario à l'autre : pour chaque mesure, les estimations restent regroupées dans une zone assez proche. Par exemple, l'effet de la fermeture des restaurants reste proche de zéro quel que soit le délai choisi.

La restriction stricte des contacts se distingue néanmoins : bien que son effet reste orienté dans le même sens, son amplitude varie davantage selon le décalage, ce qui traduit une sensibilité plus forte à ce paramètre.

Cette stabilité doit toutefois être interprétée avec prudence. D'un côté, elle suggère que les résultats ne dépendent pas fortement du choix du délai. Mais elle reflète aussi une limite du modèle : les mesures sanitaires restent souvent en place pendant de longues périodes. Décaler leur effet de quelques jours modifie donc assez peu l'information utilisée par le modèle.

Ces résultats ne remettent pas en cause l'existence d'un délai d'incubation. Ils montrent surtout que le modèle capture principalement des tendances globales. Avec une autocorrélation aussi forte dans les données, il devient difficile d'identifier précisément des effets à l'échelle journalière, ce qui renforce l'intérêt d'ajouter une correction temporelle explicite comme celle du Modèle C.

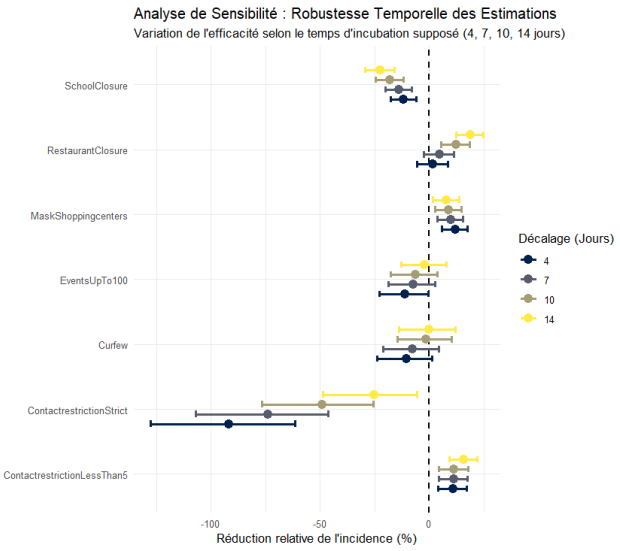


FIGURE 4.13 – **Analyse de sensibilité temporelle : variation du délai d'incubation sur les estimations.** Les coefficients varient peu d'un scénario à l'autre.

4.5 Le filtre temporel : la correction autorégressive (Modèle C)

Dans la section 4.1, l'analyse des corrélogrammes (ACF) avait mis en évidence un problème important : les résidus de nos modèles de base restaient fortement autocorrélés dans le temps.

Concrètement, cela signifie que le nombre de cas observés un jour donné t reste encore fortement lié à celui de la veille $(t-1)$, même après modélisation. Une partie de cette inertie risquait donc d'être attribuée à tort aux mesures sanitaires.

Pour corriger ce biais, un terme autorégressif d'ordre 1 (AR(1)) a été ajouté au modèle. Une estimation bayésienne avec cette structure temporelle a d'abord été testée (Modèle D). Cependant, ce modèle n'a pas convergé de manière satisfaisante. Les diagnostics montrent des valeurs de $\hat{R}$ allant jusqu'à 1,20 pour le terme autorégressif et 1,40 pour le paramètre de dispersion, ainsi que des tailles d'échantillon efficaces très faibles (Bulk ESS inférieurs à 30 pour certains paramètres, alors que la littérature recommande un minimum d'environ 100 échantillons efficaces par chaîne de Markov (ici 4 chaînes) pour obtenir une estimation stable [39]). Cela s'explique par la complexité du modèle, qui combine une distribution binomiale négative (fortement surdispersée), une dépendance temporelle explicite et une structure hiérarchique. Dans ce cadre, l'estimation bayésienne devient difficile et ne permet pas d'obtenir des résultats fiables.

L'analyse de la dépendance temporelle s'est donc appuyée sur une approche fréquentiste (Modèle C). Estimé via le solveur `glmmTMB`, ce modèle reprend la même structure que le modèle mixte initial (Modèle A), tout en intégrant un terme autorégressif AR(1) spécifique à chaque Land, et converge de manière satisfaisante.

La figure 4.14 montre les corrélogrammes des résidus du Modèle C pour quatre régions représentatives (les graphiques pour l'ensemble des treize régions sont disponibles en annexe A.4). Il y a une grande différence avec le modèle initial (figure 4.5). Presque tous les retards (*lags*) se situent maintenant à l'intérieur des bandes de confiance bleues, ce qui signifie que la forte inertie journalière a été correctement neutralisée.

Toutefois, l'analyse des résidus montre encore quelques petits pics aux retards 7, 14 et 21 jours dans certaines régions, comme la Hesse. Cela reflète probablement la persistance du cycle hebdomadaire lié aux biais des déclarations administratives. Même si la variable *weekday* et le filtre AR(1) corrigent l'essentiel de la dynamique temporelle, ils ne suffisent pas à éliminer complètement ces effets hebdomadaires.

L'essentiel de l'autocorrélation qui pouvait biaiser l'estimation des NPIs a bien

été corrigé par le modèle.

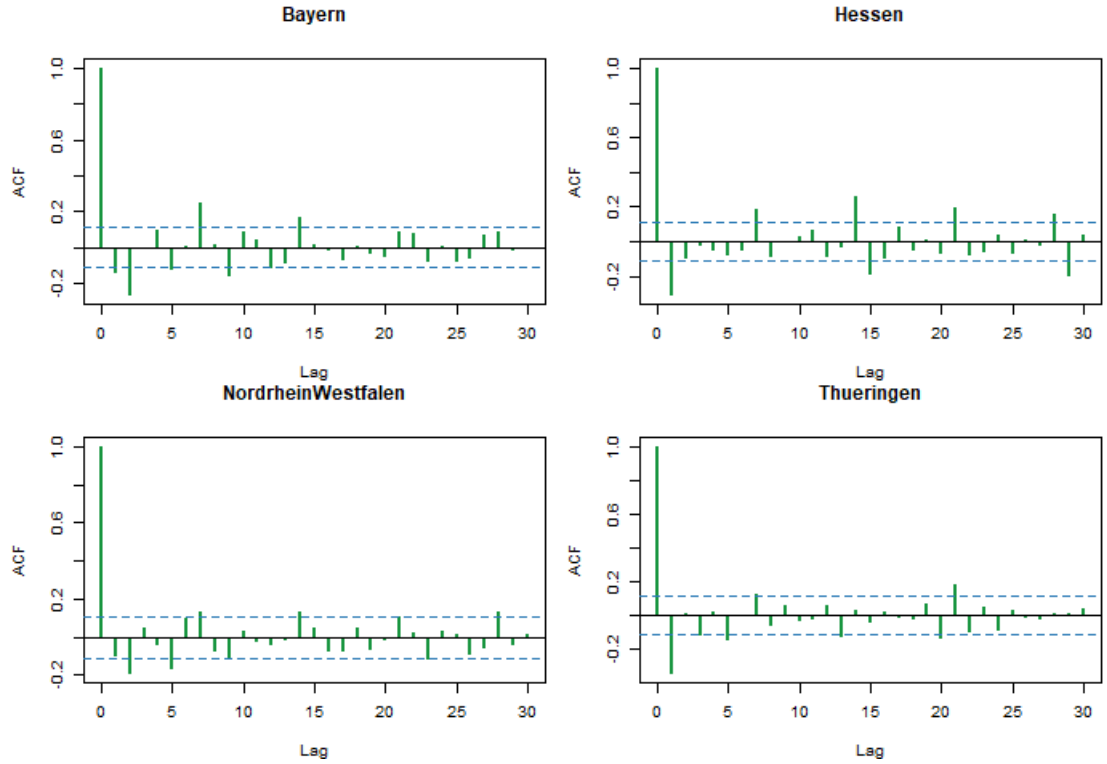


FIGURE 4.14 – **Corrélogrammes (ACF) des résidus du Modèle C.** La plupart des barres redescendent sous le seuil de significativité (lignes bleues pointillées).

4.5.1 Amélioration du modèle

L'introduction du terme AR(1) change fortement les performances du modèle. Le paramètre d'autocorrélation est estimé à $\rho = 0{,}98$, ce qui confirme une très forte dépendance d'un jour à l'autre dans les données épidémiologiques. Cette correction améliore aussi nettement la qualité globale du modèle : le critère d'information d'Akaike (AIC) passe de 44 334 (Modèle A) à 38 989 (Modèle C), soit une diminution de plus de 5 300 points, ce qui est une amélioration importante.

De plus, cette correction temporelle permet au modèle de s'adapter indirectement aux irrégularités administratives. Le diagnostic des résidus via `DHARMa` (figure 4.15) montre que l'inflation de zéros, qui était problématique dans les modèles précédents, n'est désormais plus significative (*ratio* = 1,38, $p = 0{,}208$). Mathématiquement, le Modèle C est donc le seul à valider l'intégralité de nos hypothèses de modélisation.

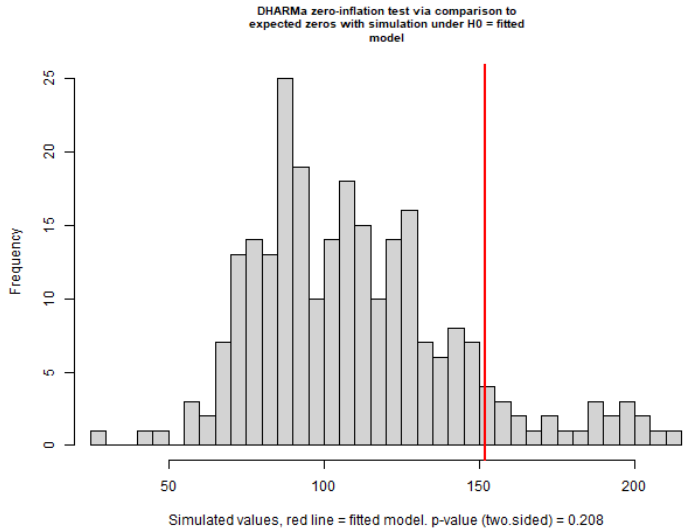


FIGURE 4.15 – **Test d'inflation de zéros sur le Modèle C (AR(1)).** L'histogramme montre le nombre de zéros attendu selon les simulations du modèle. La ligne rouge indique le nombre de zéros réellement observés dans les données empiriques (152 jours sans nouveaux cas). Même si cette valeur se situe dans la partie droite de la distribution simulée, elle reste compatible avec ce que le modèle peut produire ($p = 0{,}208$). Cela indique que le Modèle C reproduit correctement la fréquence des jours sans nouveaux cas, sans qu'il soit nécessaire d'ajouter un mécanisme d'inflation de zéros.

4.5.2 Disparition de la significativité des NPIs

Cependant, cette amélioration a une conséquence épidémiologique importante. La figure 4.16 compare les estimations des trois modèles. Une fois l'autocorrélation journalière prise en compte par le terme AR(1) (en vert), les intervalles de confiance de l'ensemble des mesures sanitaires s'élargissent ou se décalent pour inclure la valeur de zéro. Autrement dit, les effets des mesures ne sont plus statistiquement significatifs.

4.5.3 Interprétation du résultat

Cette observation constitue un résultat central de ce mémoire. Il montre que ce que les modèles A et B interprétaient comme un effet des politiques sanitaires semble en partie lié à la dynamique naturelle de l'épidémie.

La figure 4.16 illustre bien cette dynamique en comparant les estimations obtenues pour les sept NPIs conservées. Les modèles fréquentiste (en rouge) et

bayésien (en bleu) donnent des résultats presque identiques. Mais, dès que le filtre AR(1) est ajouté dans le Modèle C, tous les coefficients se rapprochent fortement de zéro.

Le cas de **ContactrestrictionStrict** est également intéressant. Dans les modèles A et B, cette mesure semblait avoir un effet très marqué. Avec le Modèle C, cet effet devient beaucoup plus incertain et se rapproche de zéro. Cela suggère qu'une partie du signal observé auparavant ne traduisait pas une réelle efficacité, mais provenait principalement de l'autocorrélation temporelle qui n'avait pas été prise en compte.

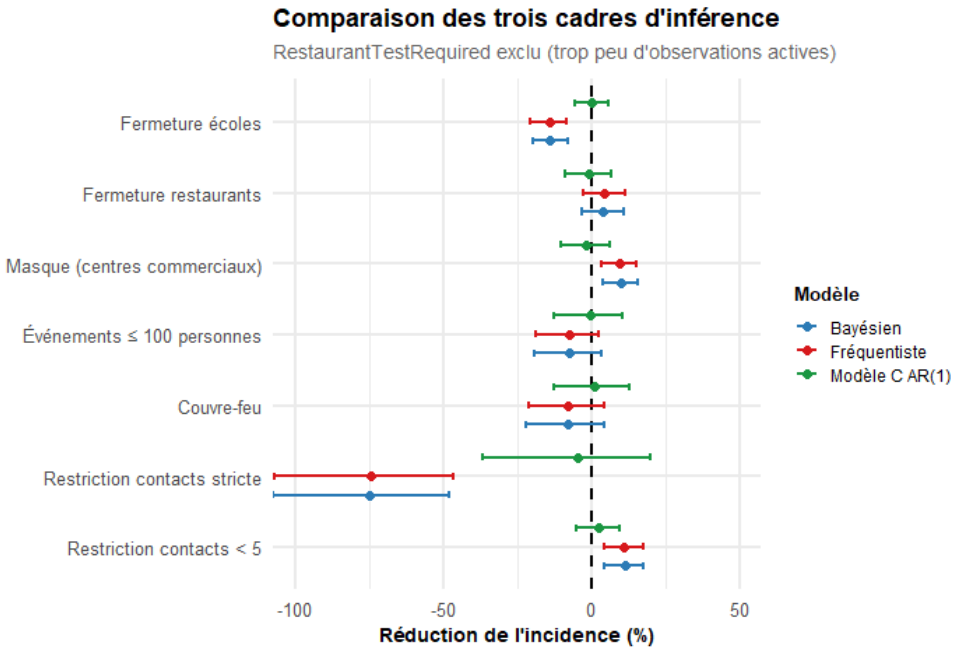


FIGURE 4.16 – **Comparaison finale des coefficients des NPIs selon les trois cadres d'inférence.** Les estimations des modèles de base fréquentiste (rouge) et bayésien (bleu) sont presque identiques, confirmant la similitude des deux approches. En revanche, l'introduction de la structure autorégressive AR(1) dans le Modèle C (vert) déplace la quasi-totalité des effets vers zéro.

Bilan de la modélisation : En conclusion, le choix entre approche fréquentiste et bayésienne ne modifie quasiment pas les résultats obtenus. En revanche, la prise en compte explicite de la dépendance temporelle a un impact majeur sur les estimations.

Une fois la dépendance temporelle correctement prise en compte, les effets des mesures sanitaires ne peuvent plus être distingués statistiquement de zéro.

Ce résultat ne veut pas dire que ces mesures ne sont pas efficaces. Il montre plutôt que, avec nos données, il est difficile de séparer clairement leur effet de l'évolution naturelle de l'épidémie, qui dépend beaucoup de ce qu'il s'est passé les jours précédents.

En pratique, cela montre que la manière dont les cas évoluent dans le temps influence fortement ce que le modèle attribue aux politiques sanitaires. Cette limite sera discutée plus en détail dans le chapitre suivant.

Pour faciliter la lecture, le tableau 4.2 résume l'ensemble des diagnostics réalisés. Il présente, pour chaque test, son objectif, les seuils de référence, les valeurs obtenues et les conclusions associées.

Test / Diagnostic	Objectif	Seuil admis	Valeur obtenue	Conclusion
Ratio obs/paramètres (Modèles A et C)	Vérifier la suffisance des données.	> 10 obs/param	191 (Modèle A) 183 (Modèle C)	**Validé.** Données largement suffisantes.
Fréquence d'activation (Länder-jours)	Vérifier si chaque NPI est estimable.	> 100 observations actives	6/7 NPIs évaluées : OK sauf Restr. stricte : 98 obs.	**Attention.** Une NPI en limite du seuil.
Trace Plot & $\hat{R}$ (Modèle B)	Vérifier la convergence de l'algorithme MCMC.	$\hat{R} < 1.01$ Chaînes mélangées	$\hat{R} = 1$ pour tous les paramètres	**Validé.** Distribution stable, bonne exploration de l'espace.
PSIS-LOO (Pareto k) (Modèle B)	Détecter un éventuel surapprentissage (overfitting).	$k \leq 0{,}7$	Tous $< 0{,}7$	**Validé.** Modèle stable, aucune observation aberrante ne domine.
Colinéarité (VIF) (Modèles A et B)	Vérifier que les NPIs ne sont pas trop corrélées entre elles.	VIF < 10	NPIs $< 2{,}5$ Spline(df=4) 7,16(A), 4,48(B)	**Validé.** Le modèle isole correctement l'effet de chaque loi.
Indice de Moran (Modèle A)	Tester l'autocorrélation spatiale des résidus.	p-value $> 0{,}05$ (absence de corrélation)	Indice $= -0{,}045$ $p = 0{,}575$	**Validé.** L'effet aléatoire par Land neutralise bien les biais spatiaux.
Inflation de zéros (Modèle A)	Vérifier si la distribution binomiale négative suffit.	p-value $> 0{,}05$	Ratio $= 3{,}09$ $p < 0{,}001$	**Échec initial.** Sous-déclaration des weekends. *Corrigé dans le Modèle C ($p = 0{,}208$).*
Dispersion (DHARMa) (Modèle A)	Vérifier l'adéquation de la variance résiduelle.	p-value $> 0{,}05$	Disp. $= 0{,}30$ $p < 0{,}001$	**Avertissement.** Le modèle prédit plus de variabilité qu'il n'en reste dans les résidus.
Corrélogramme (ACF) (Tous les modèles)	Détecter l'autocorrélation temporelle (inertie).	Lags dans la bande de confiance	Forte corrélation dans A et B Bruit blanc dans C	**Non validé.** Révèle une dépendance temporelle non modélisée dans A et B, justifiant l'introduction du Modèle C (AR(1)).

TABLE 4.2 – Synthèse des diagnostics statistiques

Chapitre 5

Limites méthodologiques dans la littérature existante

Pour comprendre l'intérêt d'introduire un modèle avec une structure autorégressive explicite (notre Modèle C), il est utile de comparer notre approche avec celle de travaux qui ont marqué l'évaluation des politiques sanitaires. Deux études en particulier, celles de Liu et al. (2021) [24] et de Hunter et al. (2021) [19], offrent un point de comparaison direct avec nos Modèles A et B.

Dans l'étude de Liu et al. (2021) [24], les auteurs cherchent à estimer l'impact de treize catégories de mesures sanitaires sur le taux de reproduction R_t dans 130 pays. Leur approche repose sur un modèle linéaire à effets fixes qui s'apparente à notre Modèle A :

$$R_{it} = \alpha_i + \sum \beta X_{it} + u_{it} \tag{5.1}$$

Ici, α_i représente les effets fixes par pays et X_{it} les mesures publiques. Bien que cette méthode permette de capter des tendances globales, elle présente une limite importante que nous cherchons justement à examiner : la dépendance temporelle des données n'est pas modélisée de manière explicite dans le terme d'erreur u_{it}.

Dans le cas du COVID-19, les cas observés un jour donné restent fortement liés à ceux de la veille. Ignorer cette dépendance peut alors conduire le modèle à confondre la dynamique naturelle de l'épidémie avec l'effet propre des mesures sanitaires.

Leurs résultats illustrent parfaitement ce problème. Certaines mesures, telles que l'obligation de rester à domicile ou la fermeture des transports publics, présentent un coefficient positif, ce qui suggère, selon leur modèle, qu'elles sont associées à une augmentation de R_t. Les auteurs eux-mêmes reconnaissent que ces résultats contre-intuitifs s'expliquent par le fait que les restrictions les plus strictes ont souvent été

introduites simultanément, lors de l'accélération rapide de l'épidémie. Le modèle capte alors la dynamique de la vague plutôt que l'effet propre des interventions. C'est exactement ce que notre Modèle C cherche à éviter.

L'article de Hunter et al. (2021) [19] adopte une architecture beaucoup plus sophistiquée, se rapprochant de notre Modèle B. Ils utilisent un cadre bayésien avec des Modèles Additifs Mixtes (GAMM) associés à une distribution binomiale négative pour gérer la surdispersion et comme nous, des fonctions de lissage (*splines*) pour capturer l'évolution du temps. Ils intègrent également une structure spatiale (modèle de Besag-York-Mollié) pour prendre en compte les interactions entre pays voisins.

Pourtant malgré cette complexité, la dépendance temporelle à court terme n'est pas explicitement modélisée dans les résidus. Les auteurs reconnaissent d'ailleurs que les données présentent une dépendance sérielle inévitable, sans pour autant chercher à l'intégrer directement dans leurs équations.

Cette limite peut influencer l'interprétation des résultats. Par exemple, dans le Tableau 3 de leur étude, les auteurs observent que les mesures de confinement strict (*stay-at-home orders*) sont associées à une augmentation du nombre de cas, avec un IRR atteignant 2,55 (IC 95% : 1,94–3,35) entre 22 et 28 jours après leur mise en place. Les auteurs reconnaissent eux-mêmes que ce résultat est surprenant.

Une explication possible est liée à la causalité inverse. Les mesures les plus strictes sont généralement introduites lorsque l'épidémie est déjà en forte accélération. Si le modèle ne prend pas correctement en compte cette dynamique temporelle, il peut alors attribuer une partie de la hausse naturelle des cas aux mesures sanitaires elles-mêmes car elles apparaissent au même moment.

Les travaux de Banholzer et al. (2021) [3] illustrent une autre limite des modèles bayésiens hiérarchiques semi-mécanistes. Dans cette architecture, la composante mécaniste repose sur l'équation de renouvellement : un principe épidémiologique simple modélisant le fait que les infections d'un jour donné dépendent directement du nombre d'infections des jours précédents. Mais, le taux de transmission (R_t) qui multiplie ces infections est modélisé de manière déterministe, uniquement en fonction des mesures sanitaires en place. Pour gérer l'effet du temps, les auteurs intègrent une « fonction de réponse temporelle retardée » (*time-delayed response function*), modélisée par une spline linéaire qui fait progressivement passer l'efficacité de la mesure de 0 à 100% sur une durée de trois jours.

La principale limite de cette approche est qu'elle n'intègre pas de mécanisme pour absorber les variations non observées. Autrement dit, toute baisse du taux de transmission R_t est, par construction, attribuée aux mesures sanitaires. Si les cas diminuent naturellement en raison de changements de comportements spontanés

(par exemple, une réduction volontaire des contacts ou un plus grand respect de l'hygiène), le modèle ne peut pas en tenir compte et attribue directement cette évolution aux politiques publiques. Cela peut conduire à une surestimation de leur effet, en particulier lorsque plusieurs mesures sont mises en place à des dates proches.

Ce problème structurel est d'ailleurs mis en évidence dans les travaux de Soltesz et al. (2020) [34]. Ils proposent une analyse critique du modèle de Flaxman et al. (2020) [11], très utilisé au début de la pandémie. Ce modèle concluait que les confinements stricts avaient été la principale source de réduction de la transmission en Europe. Or, Soltesz et al. montrent que ces résultats reposent fortement sur les hypothèses du modèle. Comme plusieurs interventions ont été introduites presque simultanément, il devient très difficile d'en distinguer les effets respectifs de manière robuste. En modifiant légèrement les hypothèses de ce modèle, les auteurs démontrent que l'algorithme attribue systématiquement la baisse des cas à la dernière mesure introduite, quelle qu'elle soit. Cela met en évidence un problème d'identifiabilité important, lié à la structure même du modèle.

L'évolution récente de la littérature confirme que cette nécessité de modéliser le temps est désormais partagée par d'autres chercheurs. Les travaux de Sharma et al. (2021) [33] illustrent bien cette évolution. En analysant l'efficacité des NPIs lors de la seconde vague en Europe (à partir de données régionales couvrant 114 régions dans sept pays européens), les auteurs ont introduit une innovation majeure dans leur cadre bayésien : une « marche aléatoire latente » (*latent random walk*) spécifique à chaque région, destinée à absorber les tendances non observées et les corrélations structurelles résiduelles.

D'un point de vue mathématique, cette évolution est intéressante. Une marche aléatoire peut être vue comme un cas particulier de processus très fortement persistant, correspondant en effet à un cas extrême de dépendance temporelle, où chaque valeur dépend très fortement de la précédente. Dans notre Modèle C (fréquentiste avec filtre AR(1)), nous estimons un coefficient très élevé ($\rho \approx 0{,}98$), ce qui montre que les données sont fortement liées dans le temps. La structure introduite par Sharma et al. (2021) [33] permet aussi de capter cette dépendance entre les jours et a donc un rôle très proche de celui de notre filtre AR(1). En pratique, cela donne des résultats plus fiables en évitant d'attribuer à tort l'évolution naturelle de l'épidémie aux mesures sanitaires.

Cette idée est également mise en avant par Lison et al. (2023) [23] dans leur revue méthodologique publiée dans *The Lancet Public Health*. Les auteurs rappellent que des facteurs non observés (comme des changements spontanés de comportement) peuvent facilement être confondus avec l'effet des politiques si la dynamique

temporelle est mal modélisée. Ils insistent donc sur l'importance de valider la robustesse des modèles, notamment à l'aide de simulations.

Bilan de la littérature : L'ensemble de ces études montre que le problème de dépendance temporelle ne concerne pas uniquement nos données allemandes. Cette difficulté revient régulièrement dans la littérature et constitue une limite plus générale des modèles appliqués aux données épidémiologiques.

La principale contribution de notre approche est d'intégrer explicitement cette dépendance via le terme AR(1) du Modèle C. C'est pour vérifier si cette correction permet réellement de réduire les biais mis en évidence dans la littérature que des simulations sur données synthétiques seront réalisées dans le chapitre suivant.

Chapitre 6

Validation par données synthétiques

6.1 Objectifs de la simulation

Les chapitres précédents ont mis en évidence les difficultés liées à l'analyse des données réelles de la pandémie. Une question fondamentale reste alors posée : comment isoler l'effet d'une mesure sanitaire de la dynamique naturelle de l'épidémie ?

D'un côté, l'épidémie présente une forte inertie temporelle : le nombre de cas d'un jour dépend fortement de celui des jours précédents. De l'autre, les gouvernements mettent généralement en place les restrictions lorsque la situation sanitaire se détériore rapidement.

C'est pour contourner ce problème, et pour suivre les recommandations de Lison et al. (2023) [23] discutées au chapitre précédent, que nous passons ici par une simulation basée sur des données synthétiques. L'idée est de créer artificiellement une épidémie dont nous maîtrisons exactement le fonctionnement. Contrairement aux données réelles, la vérité terrain (*ground truth*) est ici connue : les paramètres de la dynamique épidémique et l'effet causal des mesures sont fixés par construction. Cela permet de vérifier mathématiquement si les modèles parviennent à estimer correctement ces effets.

Cette démarche poursuit deux objectifs. Le premier est de vérifier si l'inertie naturelle de l'épidémie suffit, à elle seule, à générer des faux positifs dans un modèle classique (c'est-à-dire détecter un effet statistiquement significatif qui n'existe pas en réalité). Le second est de vérifier que le Modèle C, qui tient compte de cette dépendance temporelle, reste capable de détecter un véritable effet lorsqu'il est présent.

Concrètement, les simulations reposent sur un jeu de données synthétiques regroupant treize régions observées quotidiennement pendant 300 jours, soit 3900 observations par itération. Chaque expérience est répétée 100 fois pour les modèles A et C afin d'obtenir des résultats suffisamment stables. Le cas du Modèle B, estimé dans un cadre bayésien, est analysé séparément sur un nombre plus limité d'itérations en raison du coût computationnel élevé de l'échantillonnage MCMC.

6.2 Construction des données simulées

Le nombre de cas observés pour une région i au temps t, noté $Y_{i,t}$, est simulé à partir d'une distribution binomiale négative afin de reproduire la forte variabilité typique des données épidémiologiques :

$$Y_{i,t} \sim \text{NegBin}(\mu_{i,t}, \theta)$$

L'espérance $\mu_{i,t}$ est définie par :

$$\log(\mu_{i,t}) = \alpha + u_i + \epsilon_{i,t}$$

où α représente un niveau moyen global, $u_i \sim \mathcal{N}(0,\sigma^2)$ un effet aléatoire spécifique à chaque région, et $\epsilon_{i,t}$ une composante temporelle. Pour reproduire l'inertie naturelle de l'épidémie, cette composante temporelle suit un processus autorégressif de premier ordre :

$$\epsilon_{i,t} = \rho\,\epsilon_{i,t-1} + \nu_{i,t}, \qquad \nu_{i,t} \sim \mathcal{N}(0,\sigma_\nu^2)$$

Le paramètre ρ contrôle le niveau de dépendance dans le temps. Lorsque $\rho = 0$, les observations sont indépendantes d'un jour à l'autre. À l'inverse, lorsque ρ est élevé, les valeurs évoluent de manière plus progressive et dépendent fortement des jours précédents, ce qui correspond mieux à une situation épidémique telle qu'observée dans les analyses des chapitres précédents.

Le niveau moyen global en échelle logarithmique a été fixé à $\alpha = 2{,}8$, c'est-à-dire à environ 16 cas par jour en moyenne ($\exp(2{,}8)$) lorsqu'il n'y a pas de variation particulière. L'effet aléatoire u_i introduit une hétérogénéité entre les régions : certaines auront une incidence plus élevée que d'autres, ce qui permet par exemple de reproduire des écarts liés à la densité de population. Cet effet est tiré aléatoirement en début de simulation avec un écart-type $\sigma = 0{,}35$, puis reste fixe dans le temps.

Le paramètre de dispersion $\theta = 15$ permet à la loi binomiale négative de produire des fluctuations importantes autour de la moyenne, comme les pics de cas observés dans les données réelles.

L'objectif de ces choix n'est pas de reproduire parfaitement une vraie épidémie, mais de générer un processus de données suffisamment réalistes pour comparer les modèles dans un cadre contrôlé et identique pour tous.

6.2.1 Construction d'une mesure placebo

Tout d'abord, une mesure sanitaire fictive, notée $NPI_{i,t}$, est introduite. Cette variable est construite de manière volontairement **endogène** : sa probabilité d'activation dépend du nombre de cas observés la veille.

Concrètement, plus le nombre de cas augmente, plus la probabilité que la mesure soit activée devient élevée. Cette probabilité est générée à l'aide d'une fonction logistique inverse selon l'équation suivante :

$$\mathbb{P}(NPI_{i,t} = 1) = \text{logit}^{-1}\Big(\gamma_0 + \gamma_1 \log(Y_{i,t-1} + 1)\Big)$$

Le terme $Y_{i,t-1}$ représente le nombre de cas observés la veille. Un logarithme est appliqué à cette valeur, avec l'ajout d'une constante $+1$ pour éviter les problèmes lorsque le nombre de cas est nul. La fonction logit^{-1} transforme ensuite ce résultat en une probabilité comprise entre 0 et 1.

Une fois cette probabilité calculée, la variable $NPI_{i,t}$ est tirée aléatoirement selon une loi binomiale. La mesure peut donc apparaître de manière aléatoire, mais a plus de chances d'être activée lorsque l'épidémie s'aggrave. Ce mécanisme reproduit une situation réaliste : dans la pratique, les gouvernements renforcent généralement les restrictions lorsque les cas augmentent.

En revanche, cette mesure placebo n'intervient jamais dans l'équation qui génère les cas simulés. Elle n'a donc aucun effet réel sur l'épidémie simulée. Toute significativité détectée par les modèles correspond alors à un faux positif : le modèle conclut à un effet alors qu'il n'en existe aucun par construction.

Dans les simulations, les paramètres sont fixés à $\gamma_0 = -1$ et $\gamma_1 = 1{,}2$. Ces valeurs ne proviennent pas d'une estimation empirique, mais servent à reproduire un comportement gouvernemental plausible. Avec $\gamma_0 = -1$, la probabilité d'activation de base est d'environ 27 %, même lorsque le nombre de cas est faible. Le paramètre $\gamma_1 = 1{,}2$ contrôle ensuite la sensibilité aux variations des cas : plus l'incidence augmente, plus la probabilité d'activer la mesure devient élevée, sans devenir systématiquement égale à 1.

6.3 Résultats des simulations

6.3.1 Impact de l'autocorrélation

La première expérience consiste à faire varier le paramètre ρ, afin d'évaluer l'impact de l'autocorrélation temporelle sur les performances des modèles. Les

valeurs testées sont :

$$\rho \in \{0; 0{,}15; 0{,}3; 0{,}4; 0{,}5; 0{,}7; 0{,}85; 0{,}95; 0{,}98\}$$

Pour chaque valeur de ρ, 100 simulations indépendantes ont été réalisées. À chaque itération, le Modèle A (sans correction temporelle) et le Modèle C (avec une structure AR(1)) ont été estimés puis leur taux de faux positifs a été calculé.

Les résultats sont présentés sur la figure 6.1. Lorsque $\rho = 0$, c'est-à-dire en l'absence d'autocorrélation, le Modèle A présente un taux de faux positifs de 2 % et le Modèle C de 4 %, des valeurs proches du seuil théorique de 5 %. En revanche, dès que l'autocorrélation augmente, les performances du Modèle A se dégradent rapidement. Son taux de faux positifs atteint 19 % pour $\rho = 0{,}15$, 40 % pour $\rho = 0{,}3$, puis 61 % pour $\rho = 0{,}4$. À partir de $\rho = 0{,}7$, il atteint 100 % : dans toutes les simulations, le modèle détecte à tort un effet significatif pour une mesure qui n'en a en réalité aucun.

À l'inverse, le Modèle C reste très stable sur toutes les valeurs testées. Son taux de faux positifs varie entre 3 % et 8 %, ce qui reste cohérent avec le seuil théorique de 5 %, compte tenu de la variabilité liée aux simulations.

Comme le montre le graphique 6.1, la courbe du Modèle A augmente fortement avec ρ, alors que celle du Modèle C reste quasiment plate.

Autocorrélation (ρ)	**Faux positifs Modèle A(%)**	**Faux positifs Modèle C (%)**
0,00	2	4
0,15	19	4
0,30	40	8
0,40	61	8
0,50	91	3
0,70	**100**	7
0,85	100	4
0,95	100	4
0,98 (Allemagne)	**100**	**4**

TABLE 6.1 – **Taux de faux positifs selon le niveau d'autocorrélation (ρ).** Résultats basés sur 100 simulations par scénario. Sans correction temporelle (Modèle A), la proportion de faux positifs augmente rapidement jusqu'à atteindre 100 % lorsque l'autocorrélation est élevée. Le Modèle C corrige ce biais et maintient le taux d'erreur autour des 5 % attendus.

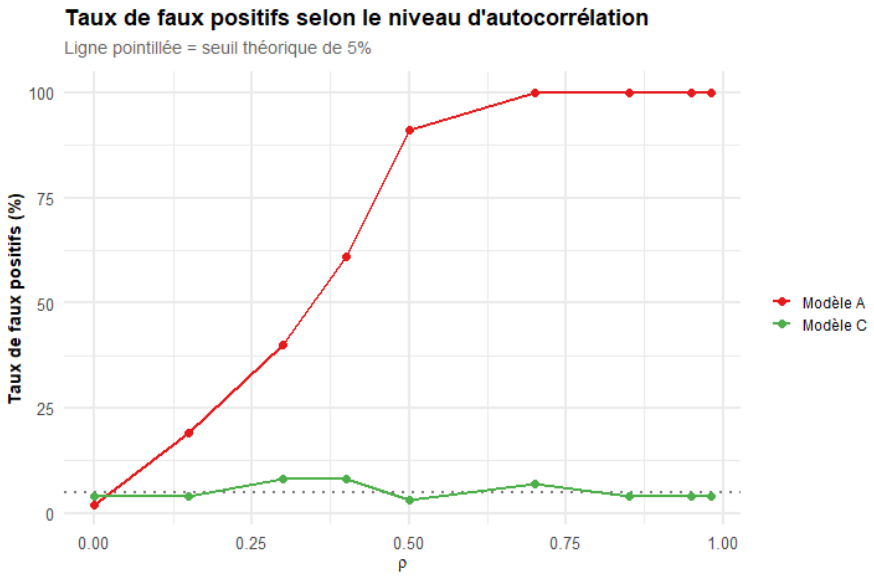


FIGURE 6.1 – **Évolution du taux de faux positifs selon l'autocorrélation temporelle ρ.** Ce graphique montre l'importance de corriger la dépendance temporelle. Sans correction (Modèle A, en rouge), le taux de fausses découvertes grimpe rapidement pour atteindre 100 % à partir de $\rho \geq 0{,}70$. À l'inverse, le Modèle C (en vert) parvient à contrôler ce risque en restant stable et proche du seuil théorique de 5 % (ligne pointillée), confirmant sa robustesse statistique face à l'inertie des données.

Ces résultats montrent que le Modèle A fonctionne correctement lorsque les observations sont indépendantes, mais devient rapidement trompeur dès que la dépendance temporelle devient plus réaliste. Le Modèle C, lui, reste stable et fiable dans tous les scénarios testés.

6.3.2 Le cas du modèle bayésien

Une deuxième analyse a été réalisée avec le Modèle B pour vérifier si le passage à un cadre bayésien permettait de corriger le problème observé précédemment.

Le protocole reste globalement le même, mais en raison du coût computationnel élevé lié à l'échantillonnage MCMC sous `Stan`, l'analyse a été limitée à un seul scénario comprenant 10 simulations (avec 2 chaînes et 1000 itérations chacune). Le niveau d'autocorrélation a été fixé à $\rho = 0{,}85$. Cette valeur correspond à une forte dépendance temporelle, cohérente avec une dynamique épidémique, tout en restant suffisamment éloignée de la limite de non-stationnarité ($\rho = 1$) pour assurer

la convergence du modèle.

Dans 9 simulations sur 10, l'intervalle de crédibilité à 95 % exclut zéro. Autrement dit, le modèle détecte presque systématiquement un effet significatif alors que la mesure simulée n'a en réalité aucun effet. Le taux de faux positifs observé atteint donc 90 %.

Même si ce résultat repose sur un nombre de simulations plus limité, il va clairement dans le même sens que ce qui a été observé précédemment. Le problème ne vient pas du choix entre approche fréquentiste ou bayésienne, mais du fait que la dépendance temporelle n'est pas correctement prise en compte.

En pratique, cela signifie qu'un modèle bayésien hiérarchique simple, avec seulement une ordonnée à l'origine aléatoire par région, peut détecter un effet significatif qui n'existe pas réellement s'il ne tient pas compte explicitement de la dépendance temporelle.

6.4 Capacité à détecter un effet réel

Une question importante se pose alors : en corrigeant les faux positifs, le Modèle C ne risque-t-il pas aussi de perdre sa capacité à détecter de vrais effets ? Autrement dit, le filtre autorégressif pourrait-il absorber non seulement le bruit, mais aussi une partie du signal réel ?

Pour répondre à cette question, une seconde simulation a été réalisée. Cette fois, une véritable mesure sanitaire a été introduite, avec un effet causal négatif de 20 % sur le nombre de cas attendus. Dans le cadre du modèle log-linéaire, cela correspond à :

$$\beta = \ln(0{,}8) \approx -0{,}223$$

Le paramètre d'autocorrélation est fixé à $\rho = 0{,}85$, pour conserver une dynamique temporelle réaliste. Contrairement à la simulation précédente, la variable $NPI_{i,t}$ est ici construite de manière **exogène** : ses dates d'activation sont tirées aléatoirement et ne dépendent pas de l'évolution des cas. Cela permet de tester directement la capacité des modèles à détecter un effet réel, sans biais de causalité inverse.

Comme précédemment, l'expérience est répétée 100 fois pour les Modèles A et C. Les résultats sont très clairs : les deux modèles identifient l'effet avec une puissance d'environ **99 %**. Le coefficient moyen estimé est de $-0{,}236$ pour le Modèle A et de $-0{,}226$ pour le Modèle C, ce dernier étant légèrement plus proche de la vraie valeur $-0{,}223$. Le biais moyen est ainsi de $-0{,}013$ pour le Modèle A, contre seulement

−0,002 pour le Modèle C.

La figure 6.2 montre la distribution des estimations sur l'ensemble des simulations. Comme nous pouvons le voir, les deux distributions sont bien centrées autour de la vraie valeur (ligne pointillée). Les estimations du Modèle C sont légèrement plus resserrées, ce qui indique une meilleure précision.

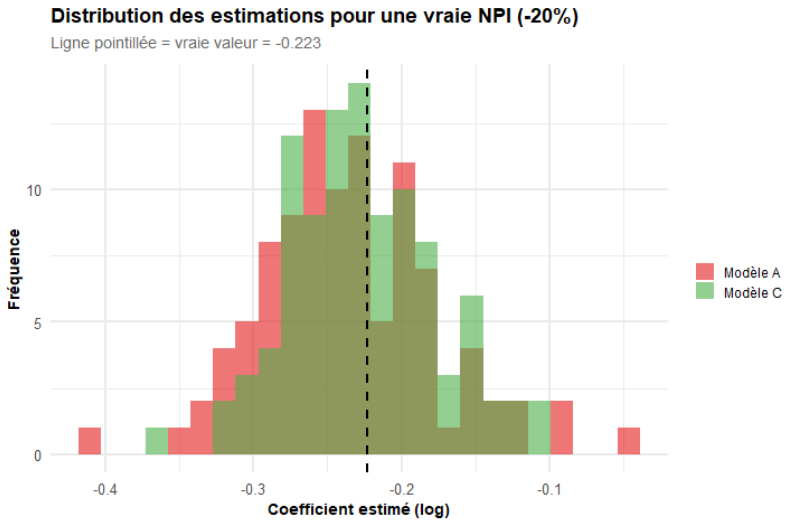


FIGURE 6.2 – **Distribution des estimations pour une NPI ayant un effet réel de −20 %.** Ce graphique compare le Modèle A (rouge) et le Modèle C (vert) face à un effet théorique connu ($\beta \approx -0{,}223$, ligne pointillée). Les deux distributions se superposent et se centrent sur cette valeur. Ceci démontre que la correction de l'autocorrélation dans le Modèle C permet de neutraliser le risque de faux positifs sans pour autant perdre en sensibilité pour détecter les effets réels.

Ces résultats montrent aussi que le Modèle A fonctionne correctement lorsque les mesures sont prises indépendamment de l'évolution des cas. Le problème apparaît avec les données réelles, où les gouvernements réagissent justement à la dégradation de la situation sanitaire. Ici, le Modèle C se montre plus robuste. Il reste capable de détecter un véritable effet lorsqu'il existe, tout en évitant d'attribuer à tort la dynamique naturelle de l'épidémie aux mesures sanitaires.

6.5 Conclusion

L'ensemble de ces simulations met en avant que les difficultés mentionnées dans les chapitres précédents ne viennent pas uniquement des données empiriques mais relèvent d'un problème plus général de modélisation.

Lorsque les variables explicatives sont endogènes et que la dynamique temporelle est forte, un modèle standard sans correction adaptée peut produire un grand nombre de faux positifs. Dans nos simulations, ce phénomène apparaît dès que l'autocorrélation devient un peu élevée. À l'inverse, l'introduction d'une structure autorégressive permet de corriger ce biais tout en conservant une très bonne capacité à détecter les effets réels. Le Modèle C limite ainsi les faux positifs sans perdre en précision.

Bilan des simulations : Ces résultats tendent à prouver que dans l'évaluation des politiques sanitaires, le simple choix de l'algorithme statistique (fréquentiste ou bayésien) ne suffit pas à garantir la validité des résultats. La manière dont la dynamique temporelle et l'inertie de l'épidémie sont intégrées dans l'équation mathématique joue un rôle central pour éviter de conclure à l'efficacité d'une mesure qui n'en a en réalité aucune.

Chapitre 7

Discussion

Les analyses des chapitres précédents ont mis en évidence des résultats parfois contre-intuitifs. Nous avons notamment constaté qu'ignorer la dynamique temporelle ne créait pas seulement de l'imprécision statistique, mais pouvait inverser complètement les conclusions sur l'efficacité d'une mesure sanitaire. Ces observations soulèvent plusieurs questions importantes : dans quelle mesure est-il réellement possible d'interpréter les effets estimés comme des effets causaux ? Et jusqu'à quel point ces conclusions dépendent-elles du modèle utilisé ?

L'objectif de ce chapitre est de prendre du recul sur ces résultats, d'en comprendre les implications épidémiologiques et statistiques, et d'en définir les limites afin d'ouvrir de nouvelles perspectives de recherche.

7.1 Équivalence des cadres d'inférence

Ce travail a comparé trois cadres d'inférence statistique appliqués aux mêmes données épidémiologiques allemandes : un modèle fréquentiste standard (Modèle A), une approche bayésienne hiérarchique (Modèle B) et un modèle fréquentiste avec correction autorégressive (Modèle C).

Le premier constat majeur de ce travail concerne la comparaison entre les cadres d'inférence. L'application du modèle fréquentiste standard (Modèle A) et du modèle bayésien hiérarchique (Modèle B) a conduit à des estimations quasiment identiques. La corrélation de Pearson entre les coefficients de ces deux modèles est de $r \approx 0{,}9998$, et les largeurs de leurs intervalles d'incertitude sont équivalentes.

Ces résultats démontrent que, sur ce jeu de données, la complexité de l'échantillonnage MCMC et l'utilisation de probabilités a priori (priors) n'apportent pas de gain d'information significatif par rapport à une approche par maximum de vraisemblance. Cela signifie que lorsque le modèle est fondamentalement mal spécifié

par rapport à la dynamique temporelle, changer simplement de cadre d'inférence (de fréquentiste à bayésien) ne suffit pas à corriger les biais. L'enjeu de la robustesse n'est donc pas le choix de l'algorithme d'estimation, mais bien la structure mathématique de l'équation.

7.2 Inertie épidémique et causalité inverse

C'est l'introduction de la structure autorégressive (Modèle C) qui modifie les résultats. L'analyse des corrélogrammes révèle une autocorrélation résiduelle forte dans les Modèles A et B. Une fois cette dépendance corrigée par le terme AR(1) dans le Modèle C, les effets de toutes les mesures sanitaires ne sont plus statistiquement significatifs et les intervalles de confiance sont en moyenne 17% plus larges. Le Modèle A identifiait pourtant l'effet de quatre de ces mesures comme significatif.

Ce résultat ne signifie pas que les interventions non-pharmaceutiques étaient inefficaces. Il met surtout en évidence une limite importante de ce type d'approche appliquée à des données observationnelles : le problème de causalité inverse.

En réalité, les mesures sanitaires sont généralement instaurées lorsque la situation se dégrade rapidement. Les modèles standards ont donc tendance à associer ces mesures à la hausse des cas simplement parce qu'elles apparaissent au même moment.

C'est ce que démontrent nos simulations du chapitre 6. Elles prouvent que, même en l'absence d'effet réel, une dynamique temporelle forte combinée à des mesures réactives suffit à tromper les modèles classiques et à générer des faux positifs. Le Modèle C permet d'éviter cette erreur. Cependant, il nous enseigne aussi qu'avec ce type d'approche, il devient très difficile d'isoler proprement l'effet causal des lois sanitaires de la dynamique naturelle de l'épidémie.

Un troisième résultat clé concerne le changement de signe de la restriction stricte des contacts. Dans le modèle exploratoire sans correction temporelle, cette mesure semblait réduire les cas d'environ 36 % mais cet effet s'inverse et devient positif dans le Modèle A (avec spline), illustrant le biais de causalité inverse.

Pour vérifier si ce résultat était uniquement dû à une région en particulier, une analyse de sensibilité spatiale a été réalisée en retirant chaque Land un par un (*leave-one-region-out*). Les résultats montrent que ce phénomène persiste. Le signe de l'effet ne change jamais, quelle que soit la région retirée et reste identique sur l'ensemble des délais temporels testés (4, 7, 10 et 14 jours).

En revanche, même si le signe est stable, l'amplitude du coefficient varie selon la région retirée (allant de 0,19 sans le Bade-Wurtemberg à 0,81 sans la Sarre). Cette instabilité reflète surtout la faible fréquence d'activation de cette mesure

(98 observations, soit 2,4 % du panel total) plutôt qu'une véritable hétérogénéité régionale. Les résultats détaillés de cette analyse sont présentés en annexe A.2.

L'inertie naturelle de l'épidémie et la causalité inverse expliquent pourquoi il est difficile d'isoler l'effet propre des mesures avec ce type de données. L'absence de significativité dans le Modèle C ne signifie pas que les mesures étaient inutiles. Elle indique surtout que ces outils ne permettent pas de démontrer clairement leur effet causal.

7.3 Limites de l'approche

Les résultats obtenus dans ce travail doivent être interprétés en tenant compte de plusieurs limites importantes.

La nature des données. Comme évoqué précédemment, les politiques sanitaires ne sont pas appliquées de manière aléatoire. Elles sont généralement introduites en réaction à une dégradation de la situation sanitaire. Même avec une correction autorégressive, il reste donc très difficile d'éliminer complètement les biais de causalité inverse. Les coefficients estimés doivent ainsi être interprétés comme des associations statistiques et non comme des preuves causales.

L'agrégation spatiale. L'analyse repose sur treize groupements régionaux, ce qui limite la capacité du modèle à capter des dynamiques locales plus fines. L'analyse de sensibilité spatiale (*leave-one-region-out*) montre d'ailleurs que certains Länder moins peuplés, comme la Sarre, peuvent influencer davantage l'estimation de certaines mesures.

Le signe des coefficients reste néanmoins stable quelle que soit la région retirée. En revanche, l'amplitude de certains effets (en particulier celui de la restriction stricte des contacts) varie selon le Land exclu. Cette instabilité s'explique surtout par le faible nombre d'observations disponibles pour cette mesure, avec seulement 98 Länder-jours d'activation sur l'ensemble de la période étudiée.

Les choix de modélisation temporelle. L'utilisation d'un délai fixe de 7 jours pour toutes les mesures est une simplification. L'analyse de sensibilité montre que les résultats ne changent pas beaucoup entre 4 et 14 jours, mais cela implique également que le modèle capte surtout des tendances longues plutôt que la dynamique journalière de l'incubation. Même si le terme AR(1) corrige le principal problème d'autocorrélation, il suppose que cette dépendance est la même pour tous les Länder et constante dans le temps, ce qui reste une approximation.

La qualité des données de déclaration. Les données de déclaration présentent des biais administratifs bien connus, comme la sous-déclaration durant les week-ends. L'ajout de la variable *weekday* permet de corriger une grande partie de ce problème. De plus, le test du modèle *ZINB* montre que les conclusions restent globalement inchangées malgré cette correction supplémentaire (gain de 369 points d'AIC sans changement sur les coefficients des NPIs).

Dans l'ensemble, ces limites ne remettent pas en cause les résultats obtenus, mais elles demandent de les interpréter avec prudence. Elles montrent aussi que ce type d'approche purement statistique a ses limites. Pour aller plus loin, il faudrait utiliser des méthodes mieux adaptées à l'inférence causale (comme l'approche par contrôle synthétique) ou des modèles épidémiologiques plus détaillés (de type SEIR), comme nous le détaillerons dans la section suivante.

7.4 Perspectives de recherche

Les limites identifiées dans ce travail ouvrent plusieurs pistes pour la suite.

Les méthodes causales. Les modèles utilisés ici permettent surtout de mettre en évidence des associations, mais pas d'établir un lien causal clair entre les mesures et l'évolution des cas. Certaines méthodes d'inférence causale, comme les variables instrumentales ou les approches quasi-expérimentales, sont souvent utilisées pour traiter ce problème. Cependant, dans le contexte des épidémies, ces méthodes présentent également certaines limites.

Les méthodes de type différences-en-différences (DID) ou à effets fixes bidirectionnels (TWFE), comme celles utilisées par Fowler et al. (2020) [12], tentent de contourner les biais en comparant des régions traitées et non traitées. Cependant, Callaway et Li (2023) [7] démontrent via simulation que ces méthodes peuvent être structurellement inadaptées à l'évaluation des politiques sanitaires. Leur argument central est que la transmission virale est un processus non linéaire dont l'évolution dépend de l'état courant de l'épidémie, ce qui viole systématiquement l'hypothèse de tendances parallèles indispensable au DID. Les régions qui adoptent des mesures strictes sont souvent celles où l'épidémie progresse le plus rapidement. L'algorithme risque alors d'inverser la cause et la conséquence : il attribue l'augmentation des cas à la politique sanitaire, alors qu'en réalité, c'est l'accélération de l'incidence qui a motivé son instauration.

Une piste intéressante pour contourner le problème de causalité inverse est la méthode du contrôle synthétique, recommandée par Lison et al. (2023) [23]. Au lieu

d'estimer l'effet des mesures avec une régression classique, cette approche construit un scénario alternatif. Pour chaque région qui applique une mesure, le principe consiste à modéliser une région jumelle synthétique. Cette région de comparaison est calculée en combinant les données de plusieurs autres régions qui n'ont pas encore adopté la restriction.

Il devient alors possible de comparer l'évolution réelle de l'épidémie avec celle de cette région synthétique. Cela donne une estimation mathématique de ce qui se serait passé si la mesure n'avait pas été prise. En comparant des trajectoires complètes plutôt qu'un lien direct entre les cas et l'activation des lois, cette méthode limite fortement le problème de causalité inverse.

L'étude de Mitze et al. (2020) [25] applique cette méthode avec des données allemandes. Les chercheurs ont évalué l'impact du port du masque obligatoire, imposé en avance au début du mois d'avril 2020 dans la ville d'Iéna. Pour isoler l'effet de cette décision, ils ont construit une « Iéna synthétique » à partir d'autres villes qui n'avaient pas encore imposé le masque à ce moment-là.

Grâce à cette comparaison, ils ont pu démontrer que l'introduction du masque avait entraîné une réduction de près de 25 % du nombre cumulé de cas en vingt jours. Cependant, l'application de cette méthode à nos données présente de nouveau une limite pratique. En Allemagne, presque tous les Länder ont mis en place des mesures similaires à quelques jours d'intervalle. Il est donc très difficile de trouver des régions « sans restriction » pour servir de point de comparaison. Malgré cette difficulté, le recours à ce type d'approche reste une piste indispensable pour dépasser les limites de nos modèles actuels et mesurer le véritable impact des décisions politiques.

Une meilleure modélisation temporelle. Le terme AR(1) corrige une grande partie de l'autocorrélation résiduelle, mais il reste une approximation relativement simple. Une piste plus ambitieuse pourrait être de combiner ces modèles statistiques avec des modèles épidémiologiques mécanistes de type SIR ou SEIR. Ces approches représentent directement les mécanismes de transmission du virus et permettraient d'intégrer des éléments comme les infections non détectées ou la variabilité des délais d'incubation.

Des données plus fines. Une autre piste serait de travailler à l'échelle des districts allemands (*Kreise*) plutôt qu'à celle des Länder. Avec plus de groupes spatiaux, le modèle pourrait mieux capturer les dynamiques locales de l'épidémie et réduire le biais lié à l'agrégation géographique.

Travailler à ce niveau de détail pose cependant un problème pratique : les districts ont des volumes de cas journaliers bien plus faibles, ce qui rend les données plus instables. Certains auraient probablement trop peu d'observations pour que le modèle puisse en tirer quelque chose de fiable, comme le notent Sharma et al. (2021) [33].

Des variables complémentaires. Enfin, intégrer des données de mobilité (telles que la géolocalisation des téléphones) ou des indicateurs mesurant le respect réel des mesures permettrait de combler l'écart entre l'annonce d'une loi et le comportement effectif des citoyens. Cette approche a déjà fait ses preuves. Les travaux de Sears et al. (2023) [32] aux États-Unis ou de Gibbs et al. (2022) [15] au Ghana démontrent par exemple que les données de téléphonie mobile captent très bien les changements de comportements spontanés de la population. Ces variables permettraient de mieux distinguer l'effet des politiques officielles des changements spontanés de comportement, qui jouent probablement un rôle important dans la dynamique épidémique.

Chapitre 8

Conclusion

Ce mémoire vise à répondre à une question méthodologique centrale : *les conclusions sur l'efficacité des politiques sanitaires dépendent-elles de la méthode statistique utilisée ?* Les résultats obtenus montrent que la réponse doit être plus nuancée et qu'une grande prudence est nécessaire dans l'interprétation de ce type de modèles.

Trois constats principaux ressortent de cette analyse :

— **La complexité d'un modèle ne garantit pas de meilleurs résultats.** Dans notre étude, le passage d'une approche fréquentiste à une approche bayésienne ne change pratiquement rien aux estimations obtenues ($r \approx 0{,}9998$). Cette similarité montre surtout qu'un algorithme plus sophistiqué ne corrige pas automatiquement un problème lié à la structure du modèle lui-même.

— **La dépendance temporelle joue un rôle essentiel.** Les modèles standards supposent que les observations journalières sont indépendantes, alors que les cas d'un jour restent fortement liés à ceux des jours précédents. En ignorant cette inertie, une partie de la dynamique naturelle de l'épidémie est attribuée à tort aux mesures sanitaires. Une fois cette dépendance intégrée avec le filtre AR(1) du Modèle C, les effets statistiquement significatifs disparaissent presque entièrement.

— **Le problème de causalité inverse est majeur.** Les simulations sur données synthétiques ont montré que, sans correction temporelle, les modèles peuvent produire jusqu'à 100 % de faux positifs. En pratique, les restrictions sont généralement mises en place lorsque la situation sanitaire se dégrade déjà fortement. Il devient alors très difficile, avec des modèles de régression classiques, de distinguer ce qui relève réellement de l'effet des mesures et ce qui correspond simplement à l'évolution naturelle de l'épidémie.

Cette perte de la significativité statistique ne signifie pas que les mesures sanitaires étaient inefficaces. Elle met surtout en évidence les limites des modèles appliqués à des données observationnelles. Finalement, l'objectif de ce mémoire n'est pas d'identifier quelle mesure est la plus efficace, mais plutôt de montrer qu'un résultat statistiquement convaincant ne correspond pas nécessairement à une véritable relation causale.

Ce travail m'a aussi fait prendre conscience de la responsabilité qui accompagne la modélisation statistique. Pendant une crise comme celle du COVID-19, il existe une forte attente pour obtenir rapidement des réponses claires et des chiffres interprétables. Pourtant, des modèles très complexes peuvent donner une impression de certitude alors que certaines hypothèses importantes ne sont pas forcément respectées.

Ce mémoire m'a donc appris à prendre davantage de recul face aux résultats statistiques, même lorsqu'ils paraissent solides. Il m'a surtout montré qu'il est essentiel de savoir ce que le modèle mesure réellement, ce qu'il ne prend pas en compte et jusqu'où les données permettent d'interpréter les résultats.

Bibliographie

[1] Alfano, V., & Ercolano, S. (2020). The efficacy of lockdown against COVID-19 : A cross-country panel analysis. *Applied Health Economics and Health Policy, 18*(4), 509–517. https://doi.org/10.1007/s40258-020-00596-3

[2] Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics : An empiricist's companion*. Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72

[3] Banholzer, N., van Weenen, E., Lison, A., Cenedese, A., Seeliger, A., Kratzwald, B., Tschernutter, D., Salles, J. P., Bottrighi, P., Lehtinen, S., Feuerriegel, S., & Vach, W. (2021). Estimating the effects of non-pharmaceutical interventions on the number of new infections with COVID-19 during the first epidemic wave. *PLOS ONE, 16*(6), e0252827. https://doi.org/10.1371/journal.pone.0252827

[4] Bourguignon, M., Damiens, J., Doignon, Y., Eggerickx, T., Plavsic, A., Sanderson, J.-P., & Bertrand, A. (2024). *Déterminants individuels et spatiaux de la mortalité pendant la pandémie de Covid-19 : le cas de la Belgique en 2020*. Démographie et sociétés, Document de travail 36, UCLouvain.

[5] Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A. B., Leech, G., Altman, G., Mikulik, V., Norman, A. J., Monrad, J. T., Besiroglu, T., Ge, H., Hartwick, M. A., Teh, Y. W., Chindelevitch, L., Gal, Y., & Kulveit, J. (2021). Inferring the effectiveness of government interventions against COVID-19. *Science, 371*(6531), eabd9338. https://doi.org/10.1126/science.abd9338

[6] Bürkner, P.-C. (2017). brms : An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*, 1–28. https://doi.org/10.18637/jss.v080.i01

[7] Callaway, B., & Li, T. (2023). Policy evaluation during a pandemic. *Journal of Econometrics, 236*(1), 105454. https://doi.org/10.1016/j.jeconom.2023.03.009

[8] Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician, 49*(4), 327–335. https://doi.org/10.2307/2684568

[9] Chin, V., Ioannidis, J. P. A., Tanner, M. A., & Cripps, S. (2021). Effect estimates of COVID-19 non-pharmaceutical interventions are non-robust and highly model-dependent. *Journal of Clinical Epidemiology, 136*, 96–132. https://doi.org/10.1016/j.jclinepi.2021.03.014

[10] *Coefficient de corrélation Intraclasse (ICC) pour évaluer la fiabilité.* (2025, 15 septembre). STAT Inférentielle. https://statinferentielle.fr/coefficient-correlation-intraclasse-icc/

[11] Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Ghani, A. C., Donnelly, C. A., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C., & Bhatt, S. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature, 584*(7820), 257–261. https://doi.org/10.1038/s41586-020-2405-7

[12] Fowler, J., Hill, S., Levin, R., & Obradovich, N. (2020). *The effect of stay-at-home orders on COVID-19 infections in the United States*. medRxiv. https://doi.org/10.1101/2020.04.13.20063628

[13] García-García, D., Herranz-Hernández, R., Rojas-Benedicto, A., León-Gómez, I., Larrauri, A., Peñuelas, M., Guerrero-Vadillo, M., Ramis, R., & Gómez-Barroso, D. (2022). Assessing the effect of non-pharmaceutical interventions on COVID-19 transmission in Spain, 30 August 2020 to 31 January 2021. *Euro Surveillance, 27*(19), 2100869. https://doi.org/10.2807/1560-7917.ES.2022.27.19.2100869

[14] Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3e éd.). CRC Press. http://www.stat.columbia.edu/~gelman/book/

[15] Gibbs, H., Liu, Y., Abbott, S., Baffoe-Nyarko, I., Laryea, D. O., Akyereko, E., Kuma-Aboagye, P., Asante, I. A., Mitjà, O., LSHTM CMMID COVID-19 Working Group, Ampofo, W., Asiedu-Bekoe, F., Marks, M., & Eggo, R. M. (2022). Association between mobility, non-pharmaceutical interventions, and COVID-19 transmission in Ghana : A modelling study using mobile phone data. *PLOS Global Public Health, 2*(9), e0000502. https://doi.org/10.1371/journal.pgph.0000502

[16] Goujon, A., Natale, F., Ghio, D., & Conte, A. (2022). Demographic and territorial characteristics of COVID-19 cases and excess mortality in the European Union during the first wave. *Journal of Population Research, 39*(4), 533–556. https://doi.org/10.1007/s12546-021-09263-3

[17] Hoffman, M. D., & Gelman, A. (2014). *The No-U-Turn Sampler : Adaptively setting path lengths in Hamiltonian Monte Carlo*. Journal of Machine Learning Research, 15(1), 1593–1623.

[18] Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., Druckenmiller, H., Huang, L. Y., Hultgren, A., Krasovich, E., Lau, P., Lee, J., Rolf, E., Tseng, J., & Wu, T. (2020). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature, 584*(7820), 262–267. https://doi.org/10.1038/s41586-020-2404-8

[19] Hunter, P. R., Colón-González, F. J., Brainard, J., & Rushton, S. (2021). Impact of non-pharmaceutical interventions against COVID-19 in Europe in 2020 : A quasi-experimental non-equivalent group and time series design study. *Eurosurveillance, 26*(28), 2001401. https://doi.org/10.2807/1560-7917.ES.2021.26.28.2001401

[20] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning : with applications in R* (2nd ed.). Springer. https://www.statlearning.com

[21] Khazaei, Y., Küchenhoff, H., Hoffmann, S., Syliqi, D., & Rehms, R. (2023). Using a Bayesian hierarchical approach to study the association between non-pharmaceutical interventions and the spread of Covid-19 in Germany. *Scientific Reports, 13*(1), 18900. https://doi.org/10.1038/s41598-023-45950-2

[22] Larmarange, J., Barnier, J., Biaudet, J., Briatte, F., Bouchet-Valat, M., Gallic, E., Giraud, F., Gombin, J., Kauffmann, M., Lalanne, C., & Robette, N. (2019). *Analyse-R. Introduction à l'analyse d'enquêtes avec R et RStudio*. Zenodo. https://doi.org/10.5281/zenodo.2669067

[23] Lison, A., Banholzer, N., Sharma, M., Mindermann, S., Unwin, H. J. T., Mishra, S., Stadler, T., Bhatt, S., Ferguson, N. M., Brauner, J., & Vach, W. (2023). Effectiveness assessment of non-pharmaceutical interventions : Lessons learned from the COVID-19 pandemic. *The Lancet Public Health, 8*(4), e311–e317. https://doi.org/10.1016/S2468-2667(23)00046-4

[24] Liu, Y., Morgenstern, C., Kelly, J., Lowe, R., Munday, J., Villabona-Arenas, C. J., Gibbs, H., Pearson, C. A. B., Prem, K., Leclerc, Q. J., Meakin, S. R.,

Edmunds, W. J., Jarvis, C. I., Gimma, A., Funk, S., Quaife, M., Russell, T. W., Emory, J. C., Abbott, S., ...CMMID COVID-19 Working Group. (2021). The impact of non-pharmaceutical interventions on SARS-CoV-2 transmission across 130 countries and territories. *BMC Medicine, 19*(1), 40. https://doi.org/10.1186/s12916-020-01872-8

[25] Mitze, T., Kosfeld, R., Rode, J., & Wälde, K. (2020). Face masks considerably reduce COVID-19 cases in Germany. *Proceedings of the National Academy of Sciences, 117*(51), 32293–32301. https://doi.org/10.1073/pnas.2015954117

[26] Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17–23. https://doi.org/10.2307/2332142

[27] Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

[28] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*(12), 1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3

[29] Rehms, R., & Khazaei, Y. (2023). *RaphaelRe/COVID_NPIs_Germany* [Code source Python]. GitHub. https://github.com/RaphaelRe/COVID_NPIs_Germany

[30] Robert Koch-Institut. (2020). *SARS-CoV-2 Infektionen in Deutschland* [Jeu de données]. Zenodo. https://doi.org/10.5281/zenodo.20250679

[31] Schilling, J., Tolksdorf, K., Marquis, A., Faber, M., Pfoch, T., Buda, S., Haas, W., Schuler, E., Altmann, D., Grote, U., Diercke, M., & RKI COVID-19 Study Group. (2021). Die verschiedenen Phasen der COVID-19-Pandemie in Deutschland : Eine deskriptive Analyse von Januar 2020 bis Februar 2021. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz, 64*(9), 1093–1106. https://doi.org/10.1007/s00103-021-03394-x

[32] Sears, J., Villas-Boas, J. M., Villas-Boas, S. B., & Villas-Boas, V. (2023). Are we #Stayinghome to flatten the curve ? *American Journal of Health Economics, 9*(1), 71–95. https://doi.org/10.1086/721705

[33] Sharma, M., Mindermann, S., Rogers-Smith, C., Leech, G., Snodin, B., Ahuja, J., Sandbrink, J. B., Monrad, J. T., Altman, G., Dhaliwal, G., Finnveden, L., Norman, A. J., Oehm, S. B., Sandkühler, J. F., Aitchison, L., Gavenčiak, T., Mellan, T., Kulveit, J., Chindelevitch, L., ...Brauner, J. M. (2021). Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. *Nature Communications, 12*(1), 5820. https://doi.org/10.1038/s41467-021-26013-4

[34] Soltesz, K., Gustafsson, F., Timpka, T., Jaldén, J., Jidling, C., Heimerson, A., Schön, T. B., Spreco, A., Ekberg, J., Dahlström, Ö., Bagge Carlson, F., Jöud, A., & Bernhardsson, B. (2020). The effect of interventions on COVID-19. *Nature, 588*(7839), E26–E28. https://doi.org/10.1038/s41586-020-3025-y

[35] Statistisches Bundesamt. (2021). *Bevölkerung nach Nationalität und Geschlecht*. Consulté le 2 janvier 2026 sur https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Bevoelkerung/Bevoelkerungsstand/Tabellen/zensus-geschlecht-staatsangehoerigkeit-2020.html

[36] Taboga, M. (2021). *Markov Chain Monte Carlo (MCMC) diagnostics*. Consulté le 22 décembre 2025, à l'adresse https://www.statlect.com/fundamentals-of-statistics/Markov-Chain-Monte-Carlo-diagnostics

[37] Veenman, M., Stefan, A. M., & Haaf, J. M. (2024). Bayesian hierarchical modeling : An introduction and reassessment. *Behavior Research Methods, 56*(5), 4600–4631. https://doi.org/10.3758/s13428-023-02204-3

[38] Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

[39] Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization : An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis, 16*(2), 667–718. https://doi.org/10.1214/20-BA1221

Annexe A

Annexes

A.1 Comparaison des estimations des NPIs entre les modèles fréquentiste et bayésien

TABLE A.1 – Comparaison des estimations des NPIs : Modèle A(Fréquentiste) vs Modèle B (Bayésien)

Mesure (NPI)	**Modèle**	**Coefficient (β)**	**IC / CrI à 95%**	**p-value**	**Variation estimée (%)**
ContactrestrictionLessThan5	Fréquentiste (A)	$-0{,}119$	$[-0{,}192; -0{,}045]$	0,002	+11,2%
	Bayésien (B)	$-0{,}122$	$[-0{,}190; -0{,}050]$	-	+11,4%
ContactrestrictionStrict	Fréquentiste (A)	0,554	$[0{,}381; 0{,}726]$	$< 0{,}001$	−74,0%
	Bayésien (B)	0,558	$[0{,}390; 0{,}730]$	-	−74,7%
Curfew (Couvre-feu)	Fréquentiste (A)	0,073	$[-0{,}045; 0{,}191]$	0,226	−7,6%
	Bayésien (B)	0,076	$[-0{,}040; 0{,}200]$	-	−7,9%
EventsUpTo100	Fréquentiste (A)	0,072	$[-0{,}028; 0{,}172]$	0,159	−7,4%
	Bayésien (B)	0,071	$[-0{,}030; 0{,}170]$	-	−7,3%
MaskShoppingcenters	Fréquentiste (A)	$-0{,}102$	$[-0{,}167; -0{,}036]$	0,002	+9,7%
	Bayésien (B)	$-0{,}109$	$[-0{,}170; -0{,}040]$	-	+10,3%
RestaurantClosure	Fréquentiste (A)	$-0{,}048$	$[-0{,}121; 0{,}024]$	0,197	+4,7%
	Bayésien (B)	$-0{,}042$	$[-0{,}120; 0{,}030]$	-	+4,1%
SchoolClosure	Fréquentiste (A)	0,132	$[0{,}078; 0{,}184]$	$< 0{,}001$	−14,1%
	Bayésien (B)	0,129	$[0{,}080; 0{,}180]$	-	−13,7%

Note : IC = Intervalle de Confiance (Fréquentiste). CrI = Intervalle de Crédibilité (Bayésien).
Une réduction négative indique une augmentation estimée des cas (effet de causalité inverse non corrigé).

A.2 Analyse spatiale : Leave-one-region-out

La figure A.1 présente l'évolution des coefficients estimés pour chaque mesure sanitaire lorsque les Länder sont retirés itérativement de la base de données.

Le point rouge représente l'estimation du modèle complet et les points bleus représentent les estimations partielles. Cette analyse démontre visuellement que la plupart des mesures sont très stables spatialement. L'exception est la mesure **ContactrestrictionStrict**, dont l'amplitude varie fortement selon le Land retiré (de 0,19 à 0,81), ce qui s'explique par son très faible nombre de jours d'activation.

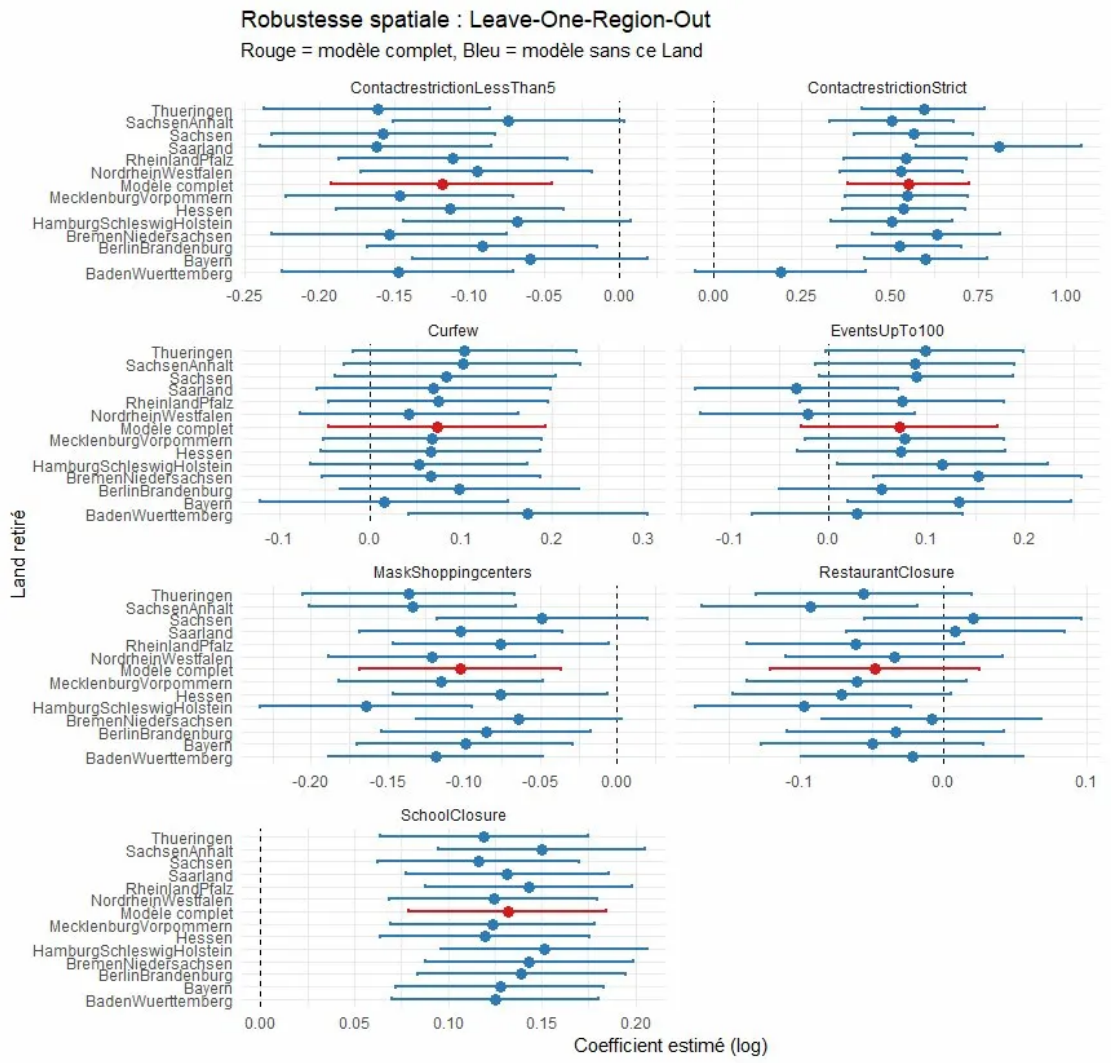


FIGURE A.1 – Stabilité spatiale des coefficients des NPIs par exclusion itérative (Leave-one-region-out).

A.3 Analyse de l'homogénéité spatiale des autres mesures (pentes aléatoires)

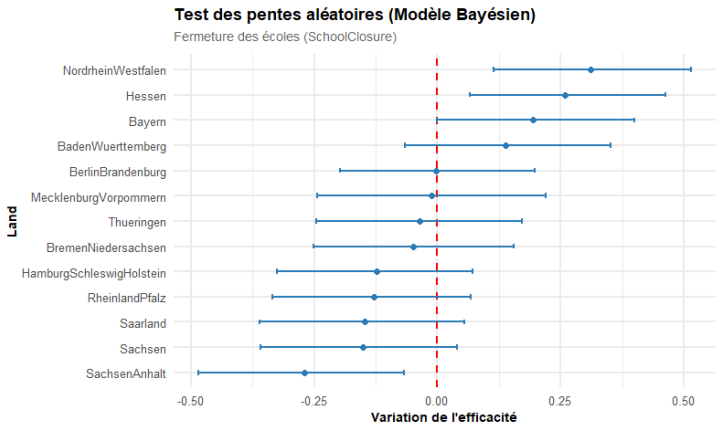


FIGURE A.2 – **Variation régionale de l'efficacité de la fermeture des écoles.** L'écart par rapport à la moyenne nationale (ligne rouge) est minime et les intervalles croisent le zéro, confirmant un effet homogène.

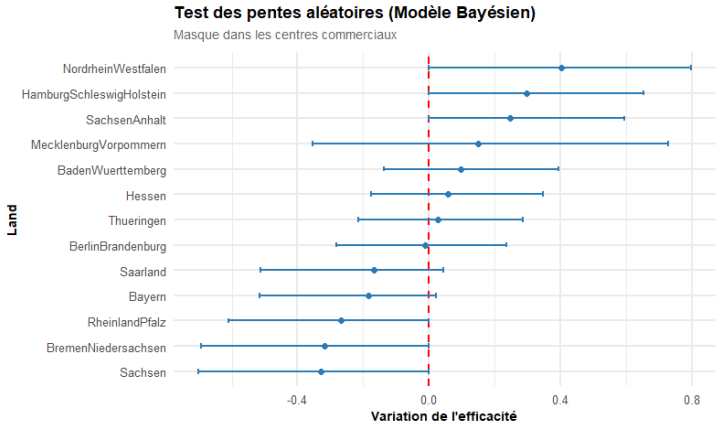


FIGURE A.3 – **Variation régionale de l'efficacité du port du masque.** Tout comme pour les écoles, l'impact de la mesure ne présente pas de déviation régionale significative.

A.4 Corrélogrammes des résidus pour l'ensemble des régions

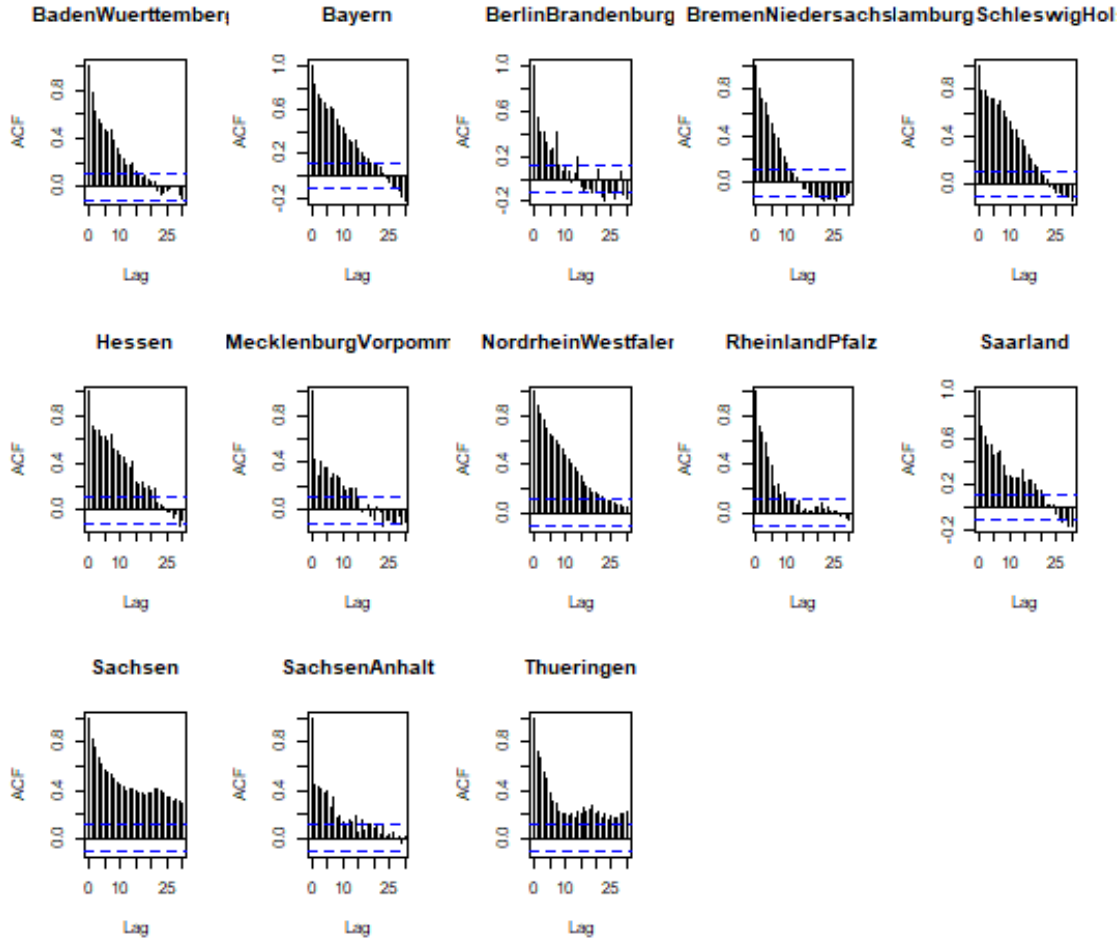


FIGURE A.4 – **Corrélogrammes (ACF) des résidus du Modèle A pour les treize groupements régionaux.** Les barres d'autocorrélation dépassent partout les seuils de significativité (lignes pointillées bleues) sur de nombreux jours consécutifs. Cela démontre que dans le modèle standard l'erreur de prédiction d'un jour reste fortement corrélée à celle de la veille.

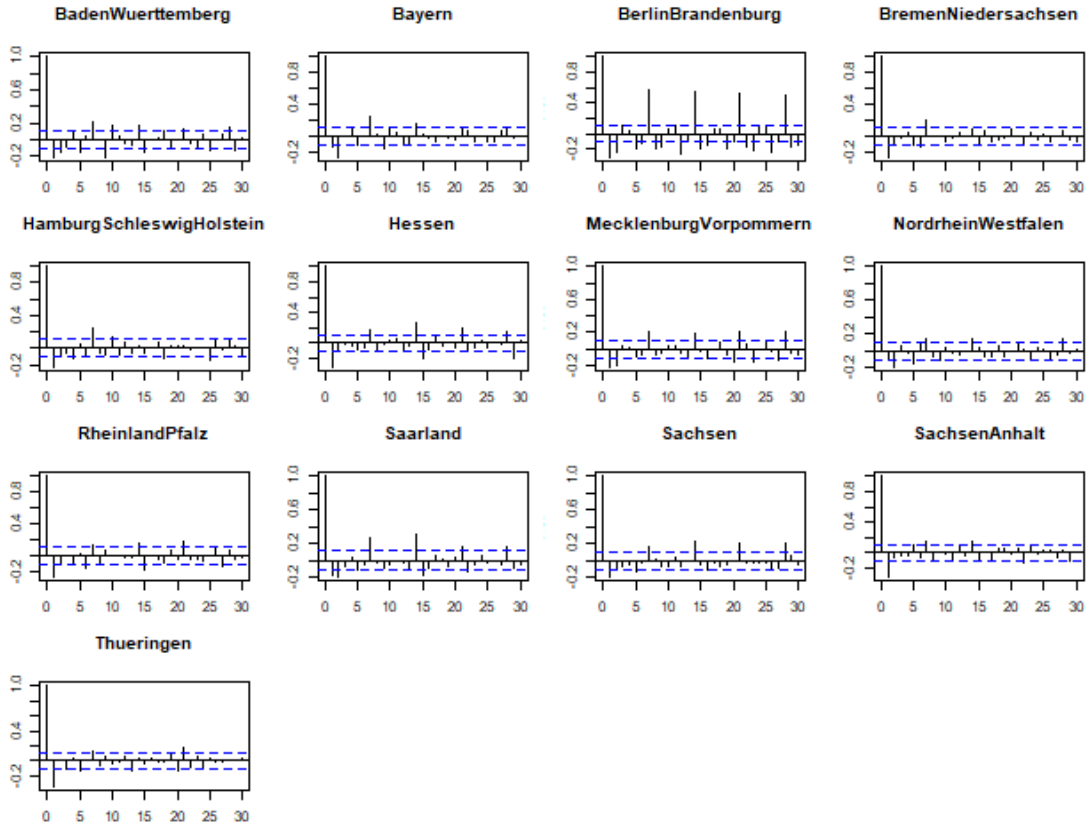


FIGURE A.5 – **Corrélogrammes (ACF) des résidus du Modèle C pour les treize régions.** La plupart des barres redescendent sous le seuil de significativité (lignes pointillées bleues), confirmant la validité du filtre AR(1) sur l'ensemble du territoire.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**