Maladies infectieuses et données agrégées : estimation de la fraction attribuable et prise en compte de biais

Felix Cheysson

Thèse de doctorat

NNT : 2020UPASR012

Maladies infectieuses et données agrégées : estimation de la fraction attribuable et prise en compte de biais

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 570, santé publique (EDSP)
Spécialité de doctorat : Biostatistiques
Unité de recherche : Université Paris-Saclay, UVSQ, Inserm, CESP, 94807, Villejuif, France
Référent : Faculté de médecine

Thèse présentée et soutenue en visioconférence totale, le 17 novembre 2020, par

Felix Cheysson

Composition du jury :

Bruno Falissard PU-PH, Inserm, Université Paris-Saclay	Président
Patricia Reynaud-Bouret Professeur, Université Côte-d'Azur	Rapporteure & Examinatrice
Jean-Christophe Thalabard Professeur émérite, Université de Paris	Rapporteur & Examinateur
Jacques Bénichou PU-PH, Université de Rouen	Examinateur
Karen Leffondré Maître de conférences, Université de Bordeaux	Examinatrice
François Roueff Professeur, Télécom Paris, Université Paris Saclay	Examinateur
Laurence Watier Chercheur, Inserm, Université Paris Saclay	Directrice
Gabriel Lang IGPEF, AgroParisTech, Université Paris Saclay	Codirecteur

Remerciements

Je remercie mes deux directeurs de thèse, Laurence et Gabriel, qui m'ont encadré ces quatre dernières années. Tout d'abord, cette thèse n'aurait pas existé sans le soutien de Laurence, qui m'a accompagné dans mes choix professionnels depuis le stage de master, et conseillé au cours de mes virages académiques. Merci pour ton écoute et ta gentillesse, sans oublier les verres au Général Beuret ! Je remercie également Gabriel, avec qui j'ai pris grand plaisir à travailler tout au long de ma thèse. Merci pour ton accueil, ta disponibilité et ta bienveillance. Avant tout, merci à tous les deux pour la confiance que vous m'avez accordée tout au long de cette thèse, mais aussi d'avoir su me rassurer quand je doutais de mon travail.

Je tiens ensuite à remercier tous les membres du jury de la soutenance : Patricia Reynaud-Bouret et Jean-Christophe Thalabard, qui m'ont fait l'honneur d'être rapporteurs de ce manuscrit ; Bruno Falissard, d'avoir accepté d'être président du jury ; et Jacques Bénichou, Karen Leffondré et François Roueff, qui ont bien voulu être examinateurs de la soutenance.

Je tiens à remercier François Roueff une seconde fois pour son coup de pouce qui nous a permis de débloquer le travail sur les processus de Hawkes agrégés. Je remercie également Isabelle Pontais et Céline Casério-Schönemann de Santé publique France pour avoir mis à notre disposition les données du réseau Oscour, mais aussi Yann Le Strat et Daniel Lévy Bruhl pour les échanges pertinents que nous avons eus. Par ailleurs, un grand merci à Sophie Donnet pour son cours sur l'algorithme EM, qui nous a fait gagner un temps précieux.

Je souhaite remercier toutes les personnes qui m'ont accompagné tout au long de ma thèse au sein de mes deux labos d'accueil. D'abord tous les doctorants avec qui j'ai partagé l'expérience de la thèse : ceux qui étaient là avant moi, Anna et Anna, Marie, Yann, Mathieu, Rana, Clothilde, Hélène ; ceux qui ont commencé avec moi : Timothée, Marie, Marion, Mehdi, Mélanie, Jeanne, Audrey ; et ceux qui m'ont rejoint : Raphaëlle, Martina, Annarosa, Saint-Clair, Claire, Tâm, Jonathan, Lison, David, Salam. Mais aussi tous ceux qui m'ont accueilli à bras ouverts : Annick, Bich-Tram, Elisabeth, Lulla, Solen, Isabelle, Pierre et aussi Pierre, Sylvain, Julien, Sophie, Christophe, Christelle, Joon, Céline, Tristan, Sarah, Laure, Stéphane,

Colette, Séverine, Lénaig, Marie-Laure, Eric, Jessica, Julie, Jade, Erica, Paul, Gaspard, Marie-Pierre, Christian. Enfin Liliane et Didier, qui m'ont permis d'effectuer ma thèse dans les meilleures conditions. C'est grâce à vous tous que ces quatre années de thèse ont été si agréables et si heureuses, et c'est aussi grâce à vous que je souhaite continuer dans la recherche académique.

Plus particulièrement, je voudrais remercier mes compagnons de thèse. Marie, toujours joyeuse, jamais fatiguée, et généreuse comme tout ! Tu as été l'un des piliers de l'Agro, ta bonne humeur est contagieuse, tu la partages avec tous ceux que tu rencontres. Timothée, pour ta gentillesse et ta compassion. Ce fut un grand plaisir de t'avoir comme cobureau. J'espère que tu es aussi heureux dans tes nouvelles études que je l'ai été à ton contact. Martina, Raphaëlle, Annarosa et Saint-Clair, pour tous ces moments heureux que nous avons partagés autour des pauses thés, pour votre soutien pendant cette dernière année, pour votre joie et votre bonne humeur. Anna et Lénaig, pour nos conversations, nos délires, nos déjeuners. Vous m'accompagnez depuis bien avant ma thèse, et je vous remercie de votre amitié. À Mélanie, Audrey, Marion, Salam, Mehdi, Lison, David, Jonathan et Paul, pour avoir apporté la vie dans le laboratoire et supporté mes bêtises autour du café dans le couloir. Vous allez tous me manquer.

Merci aussi à tous mes amis d'enfance, vous qui m'accompagnez depuis tant d'années. François, Marc, Thibaut, et leurs compagnes, Claire et Sophie. Pour tous ces moments passés ensemble, les vacances à Saint-Gilles, les jeux de société et jeux de rôle, les parties pendant le confinement. Merci aussi à Marine et Harold (même si tu es aussi de ma famille), pour vos talents acérés de détective, qui nous ont permis de garder cent pour cent de réussite à nos escape games. Merci pour votre amitié indéfectible.

Je remercie bien sûr ma famille, qui a toujours été là pour moi, quand je n'allais pas bien et quand j'allais bien aussi. Mes frères, Anatole et Arthur, qui êtes dans mon cœur, même si je ne vous vois pas suffisamment. Maman et Papa, pour votre soutien permanent, votre amour et de m'avoir toujours encouragé dans mes projets. Merci pour toute votre affection.

Je remercie aussi toutes mes belles familles, Bruno, Elyse et Claire, Isabelle et Maxime, mais aussi Christian, Agnès, Colette, Maxime, Bénédicte et Rénald. Merci pour votre accueil chaleureux.

Enfin, je tiens à remercier Annie, ma compagne, qui m'a soutenu tout au long

de ma thèse. Merci pour ta présence, ta compréhension, et surtout ta patience : je sais que je n'ai pas été facile à supporter... Tu es gentille, attentionnée, et je suis heureux que tu aies été à mes côtés durant ces années. Je t'aime.

Et merci à Touille, surtout si tu ne montes pas sur le clavier pendant la soutenance !

Liste des productions scientifiques

Publications scientifiques

- **F. Cheysson**, M. A. Vibet, D. Guillemot, et L. Watier. Estimation of exposure-attributable fractions from time series : A simulation study. *Statistics in medicine*, 37(24) :3437-3454, 2018.
- **F. Cheysson**, C. Brun-Buisson, L. Opatowski, L. Le Fouler, C. Caserio-Schönemann, I. Pontais, D. Guillemot, L. Watier. Outpatient antibiotic use attributable to viral acute lower respiratory tract infections during the cold season. Soumis.
- **F. Cheysson**, G. Lang. Strong mixing condition for Hawkes processes and application to Whittle estimation from count data. En révision.

Communications orales

Congrès nationaux

- **F. Cheysson**, D. Guillemot, L. Watier. Estimating Attributable Fractions using Bayesian Inference : a Simulation Study with Epidemic Exposure. Présentée aux $7^{ème}$ *Rencontres des Jeunes Statisticiens*, SFdS, Porquerolles, Avril 2017.
- **F. Cheysson**, G. Lang, L. Watier. Estimation of attributable fractions from time series. Présentée aux *Sommets de Rochebrune*, MIA-Paris, Rochebrune, Mars 2018.
- **F. Cheysson**, M.A. Vibet, D. Guillemot, L. Watier. Estimation of attributable fractions from time series. Présentée à la *Journée des Jeunes Chercheurs*, Société Française de Biométrie, Paris, Juin 2018.
- **F. Cheysson**, M.A. Vibet, D. Guillemot, L. Watier. Estimation of attributable fractions from time series. Présentée aux $50^{ème}$ *Journées de Statistique*, SFdS, Paris Saclay, Juin 2018.
- **F. Cheysson**, G. Lang, L. Watier. Whittle estimator for the discrete Hawkes process. Présentée aux $8^{ème}$ *Rencontres des Jeunes Statisticiens*, SFdS, Porque-

rolles, Avril 2019.

- **F. Cheysson**, G. Lang, L. Watier. Spectral estimation of Hawkes count data in discrete time. Présentée aux *51ème Journées de Statistique*, SFdS, Nancy, Juin 2019.
- **F. Cheysson**, G. Lang. A strong mixing condition for Hawkes processes and its application to Whittle estimation from count data. Présentée au séminaire *Workshop : Statistical methods for Hawkes processes*, Sorbonne Université, Paris, Mars 2020.
- **F. Cheysson**, G. Lang. Estimation of Hawkes processes from binned observations using Whittle likelihood. Présentée à la conférence *Séries chronologiques : nouveaux résultats et applications statistiques*, CIRM, Marseille, Sept. 2020.

Congrès internationaux

- **F. Cheysson**, G. Lang, L. Watier. Spectral estimation of Hawkes count data in discrete time. Présentée à *7th Channel Network Conference*, International Biometric Society, Rothamsted Research, UK, Juil. 2019 — Prix de la *Meilleure présentation orale étudiante*.

Autre production

- **F. Cheysson**. *hawkesbow* : R package for the estimation of HAWKES processes from Binned Observations using Whittle likelihood. `https://github.com/fcheysson/hawkesbow`

Table des matières

Introduction

Les maladies infectieuses sont causées par des organismes, comme des bactéries, virus, champignons ou parasites qui sont transmis directement ou indirectement d'une personne à une autre. Elles constituent une des principales causes de décès dans le monde, en particulier dans les pays à faible revenu et chez les jeunes enfants. [1] Trois maladies infectieuses (les infections des voies respiratoires inférieures, les maladies diarrhéiques et la tuberculose) ont été classées parmi les dix premières causes de décès dans le monde en 2016 par l'Organisation mondiale de la santé. Au premier rang, les infections des voies respiratoires inférieures seraient responsables de 3,0 millions de décès, soit 5,2% des décès dans le monde en 2016. [2]

Pour surveiller et étudier les maladies infectieuses et de façon plus générale l'état de santé d'une population, les agences sanitaires collectent de façon continue et systématique des données de santé essentielles aux actions de santé publique (Thacker et Berkelman, 1988). Afin de pouvoir analyser et interpréter ces données en vue d'actions de prévention, elles produisent de nombreux indicateurs de santé, comme par exemple le taux d'incidence ou de prévalence d'une maladie ou le nombre annuel de décès. Ces indicateurs sont souvent agrégés, soit construits à partir en agrégeant des données individuelles à l'échelle d'une population donnée (classe d'âge, hommes / femmes, ...) ou d'un territoire géographique (pays, région, commune, ...) pour une période de temps donnée (année, semaine, ...), soit directement recueillis à l'échelle d'un territoire ou d'un établissement de santé. En conséquence, une partie de l'épidémiologie analytique consiste à étudier ces indicateurs agrégés et en particulier à développer des modèles spatiaux et/ou temporels permettant de travailler à partir de ces indicateurs (Elliott et Wartenberg, 2004; Auchincloss *et al.*, 2012; Bhaskaran *et al.*, 2013).

Lorsqu'une association entre un facteur de risque et un événement d'intérêt (la survenue d'une maladie, ou le nombre de décès par exemple) est supposée causale, un indicateur d'intérêt est la fraction attribuable car il reflète l'impact du facteur de

1. `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`, dernière visite le 03/09/20.

2. `https://www.who.int/healthinfo/global_burden_disease/GHE2016_Deaths_WBInc_2000_2016.xls`, dernière visite le 03/09/20.

risque sur l'incidence de l'événement à l'échelle d'une population. Elle correspond au nombre d'événements qui seraient évités si le facteur de risque était supprimé et peut être utilisée pour prévoir l'effet d'une élimination du facteur de risque. Malgré de nombreux travaux, l'estimation de la fraction attribuable à partir de données individuelles est un problème toujours d'actualité, puisque l'attribution des sources peut être une question difficile selon la qualité des données et la présence de facteurs de confusion (Benichou, 2001; Bard *et al.*, 2005). Lorsque son estimation est réalisée à partir d'indicateurs agrégés, elle nécessite d'abord une étape de modélisation pour déterminer la force de l'association entre le facteur de risque et l'événement d'intérêt. Dans ce cadre, la définition mathématique explicite de la fraction attribuable n'est pas simple puisqu'elle dépend du modèle choisi, et que son estimation, notamment celle de sa variance, peut présenter des difficultés techniques relatives au modèle.

Il a aussi été montré que l'agrégation de données individuelles, outre la perte d'information inhérente à l'agrégation, peut être source de biais difficiles à évaluer. Lorsque cette agrégation est effectuée à l'échelle du territoire, on parle du problème des unités spatiales modifiables (*Modifiable Areal Unit Problem*, MAUP). Ce problème a été constaté pour la première fois en 1934 (Gehlke et Biehl, 1934), puis ensuite exploré par Openshaw et Taylor (1979), et décrit en détails par Openshaw (1984), qui a montré que les coefficients de régression estimés et leurs variances peuvent varier considérablement en fonction des frontières ou des échelles choisies pour l'agrégation et qu'il n'existe pas de choix naturel optimal. Le MAUP a été largement étudié en géographie physique (Dark et Bram, 2007), mais aussi dans les disciplines qui utilisent des données agrégées à l'échelle de zones administratives, comme l'économie (Briant *et al.*, 2010; Kitchin et Mcardle, 2015) et l'épidémiologie (Araujo Navas *et al.*, 2020; Parenteau et Sawada, 2011).

De manière similaire, il a été montré plus récemment que le choix du découpage temporel pouvait avoir, lui aussi, un impact sur les résultats (Cheng et Adepeju, 2014). Des biais peuvent être introduits par le choix de l'échelle de temps (par exemple la semaine plutôt que le jour), de la définition des unités de temps (par exemple commencer la semaine le mardi plutôt que le lundi) et les effets de bord liés à la durée d'observation. Par analogie au MAUP, Cheng et Adepeju définissent ces effets sous le terme de problème des unités temporelles modifiables (*Modifiable Temporal Unit Problem*, MTUP).

Lorsqu'elles sont disponibles, travailler directement avec des données indivi-

duelles permet de s'affranchir de ces difficultés. Lorsque les indicateurs étudiés sont directement les cas individuels d'une pathologie, les processus ponctuels sont des modèles adaptés. Ils définissent un cadre de calcul explicite pour l'étude statistique d'événements localisés précisément dans l'espace et/ou le temps (Gatrell *et al.*, 1996; Benes *et al.*, 2005; Diggle *et al.*, 2013). Lorsque l'on s'intéresse plus précisément à des pathologies infectieuses contagieuses, c'est-à-dire les maladies infectieuses qui se transmettent facilement par contact avec une personne malade ou ses sécrétions, les processus de Hawkes sont une famille de processus ponctuels appropriés car ils permettent de prendre en compte les interactions entre individus. En effet, ils sont définis de sorte que l'occurrence d'un événement augmente la probabilité d'occurrence des événements futurs. En outre, leurs paramètres sont directement interprétables dans un contexte épidémiologique. Afin de comprendre les biais résultants des MAUP et MTUP dans le cadre de la modélisation de maladies contagieuses, il parait intéressant d'étudier les effets de l'agrégation de données sur le processus de Hawkes.

Objectifs de la thèse

Ce travail de thèse cherche à résoudre les difficultés méthodologique décrites précédemment lorsque l'on travaille sur des données agrégées dans un contexte de santé publique. Nous nous intéressons d'une part à l'estimation des fractions attribuables et d'autre part à l'impact de l'agrégation des données sur le processus de Hawkes.

Les premiers travaux présentent des méthodes d'estimation de la fraction attribuable pour des séries temporelles lorsque le facteur de risque est épidémique et les événements d'intérêt saisonniers. Pour les modèles statistiques les plus fréquemment rencontrés dans la littérature épidémiologique, nous établissons les estimateurs de la fraction attribuable et de leurs variances.

Les estimateurs proposés sont ensuite considérés pour étudier l'impact des pathologies infectieuses hivernales sur l'usage des antibiotiques dans la communauté. Deux études de cas portant sur cet usage sont développées, pour déterminer la contribution des syndromes grippaux d'une part, des infections des voies respiratoires basses d'autre part.

Enfin, nous nous intéressons au processus de Hawkes comme modèle d'étude des maladies contagieuses et étudions l'impact de l'agrégation de données sur leur estimation. Nous expliquons leur intérêt potentiel dans le contexte épidémiologique et proposons une méthode d'estimation adaptée lorsque les événements de contagion ne sont pas observés individuellement, mais agrégés sur des unités de temps régulières. Pour évaluer l'avantage des processus de Hawkes par rapport aux modèles statistiques classiquement utilisés pour les maladies contagieuses, nous proposons de comparer ces modèles *via* la fraction attribuable, qui est un indicateur de santé dont l'interprétation ne dépend pas du modèle.

Chapitre 1

Fraction attribuable et données agrégées

1.1. Contexte

Un facteur de risque est défini par l'Organisation mondiale de la santé comme "tout attribut, caractéristique ou exposition d'un sujet qui augmente la probabilité de développer une maladie". [1] De par sa définition, c'est un facteur dont l'exposition est associée à la survenue de la maladie. Le plus souvent, il ne s'agit pas d'une simple association statistique entre l'exposition au facteur et la survenue de la maladie et il est important d'identifier les éléments en faveur d'une relation causale. Cette étape complexe, étayée par des arguments issus le plus souvent d'études épidémiologiques de nature observationnelle, s'appuie sur un ensemble de critères appelés "critères de Hill" dont la confrontation permet d'apprécier le degré de plausibilité de la nature causale de la relation (Hill, 1965; Rothman et Greenland, 2005).

Ces critères, initialement au nombre de neuf, se regroupent en deux catégories. D'une part, les critères qui caractérisent la nature de l'association, avec comme critères forts de causalité, sa *force* et la *reproductibilité des résultats*, mais aussi sa *temporalité* : la cause doit précéder l'effet. Et d'autre part, les critères de nature contextuelle qui mettent en perspective les résultats des études épidémiologiques par rapport aux connaissances biologiques pertinentes en lien avec l'association considérée, parmi lesquels on trouve la *plausibilité biologique de l'association* comme critère fort dans l'établissement d'une relation de causalité.

Pour quantifier la force de l'association entre une maladie et une exposition, plusieurs mesures, dites d'association, ont été définies (Dicker *et al.*, 2006). Dans le cas d'une exposition binaire (exposés *versus* non exposés), on distingue : **l'excès de risque**, égal à la différence de risques de maladie entre les sujets exposés et non

1. `https://www.who.int/topics/risk_factors/fr/` (dernier accès 06/05/2020)

exposés ; et **le risque relatif (RR)**, égal au rapport des risques ou d'incidences de la maladie entre sujets exposés et non exposés. Ces mesures d'association sont facilement estimées à partir des études de cohorte.

Dans les études cas-témoins où l'incidence de la maladie n'est pas connue, on compare cette fois la fréquence d'exposition chez les cas et les témoins. Dans ce cadre, la mesure d'association est **l'odds ratio**, défini comme le rapport des côtes d'exposition chez les cas et les témoins. L'interprétation de cette mesure est difficile, c'est pourquoi des auteurs ont proposé des approximations du risque relatif à partir de l'odds ratio (Robbins *et al.*, 2002; Viera, 2008; Grant, 2014). Il a ainsi été montré que lorsque la maladie est rare (prévalence inférieure à 10%), l'odds ratio était une bonne approximation du risque relatif (Zhang et Yu, 1998; Viera, 2008).

Néanmoins, ces mesures d'association ne renseignent pas sur la proportion de cas liés à l'exposition, ni à l'échelle des sujets exposés, ni à celle d'une population cible. Elles ne permettent donc pas de quantifier l'impact de l'exposition sur la prévalence de la maladie et donc la proportion des cas sur laquelle les politiques de prévention pourraient agir. Sous l'hypothèse d'une relation causale, **le risque attribuable (RA)** est une mesure qui permet de mesurer cet impact. À l'échelle d'une population cible, le risque attribuable est défini comme la proportion, parmi tous les cas, de ceux attribuables à l'exposition. Cette proportion représente la proportion des cas qui seraient évités si l'exposition pouvait être éliminée.

En notant $\mathbb{P}(D)$ la probabilité de survenue de la maladie dans la population, composée de sujets exposés au facteur de risque E et non-exposés $\bar{E}$, et $\mathbb{P}(D|\bar{E})$ la probabilité de survenue de la maladie chez les non-exposés uniquement, le risque attribuable s'écrit :

$$\mathrm{RA} = \frac{\mathbb{P}(D) - \mathbb{P}(D|\bar{E})}{\mathbb{P}(D)}. \tag{1.1}$$

À la différence du risque relatif, qui mesure uniquement l'association entre le facteur de risque et la survenue de la maladie, la formulation qui suit montre que le risque attribuable dépend à la fois de la force de l'association et de la prévalence de l'exposition (Miettinen, 1974) :

$$\mathrm{RA} = p_{E|D}\frac{\mathrm{RR} - 1}{\mathrm{RR}}, \tag{1.2}$$

où $p_{E|D}$ désigne la prévalence de l'exposition parmi les sujets malades, et $(\mathrm{RR} - 1)/\mathrm{RR}$ correspond à la part attribuable uniquement chez les sujets exposés.

Puisqu'il dépend de la prévalence de l'exposition dans la population cible, une

valeur élevée du risque relatif ne correspond donc pas systématiquement à une valeur élevée du risque attribuable.[2] Par ailleurs, la prévalence de l'exposition pouvant varier fortement selon les populations, dans le temps et dans l'espace, le risque attribuable n'est, le plus souvent, pas transposable d'une population à une autre.

Comme pour les mesures d'association, la présence de facteurs de confusion peut conduire à des estimations du risque attribuable non valides (Walter, 1980, 1983). Des méthodes permettant de prendre en compte des facteurs de confusion dans l'estimation du risque attribuable ont aussi été développées (Benichou, 2001; Bard *et al.*, 2005).

Le risque attribuable tel que défini ci-dessus, nécessite de disposer de données à l'échelle individuelle, principalement issues d'enquêtes épidémiologiques (voir l'Annexe 1.A pour des détails concernant l'estimation du risque attribuable). Néanmoins, la notion de risque attribuable est aussi retrouvée dans la littérature en l'absence de données individuelles. Par exemple, en s'appuyant sur des indicateurs agrégés, de nombreux travaux cherchent à identifier la part de la mortalité attribuable à la pollution (Schwartz et Marcus, 1990), à la grippe (Nunes *et al.*, 2011; Muscatello *et al.*, 2013), au tabac (Fenelon et Preston, 2012), aux canicules (Fouillet *et al.*, 2006; Aboubakri *et al.*, 2019), *etc.* Sans être exhaustif, d'autres exemples portent sur le nombre d'hospitalisations attribuables aux virus respiratoires hivernaux (Zhou *et al.*, 2012), ou bien le nombre de prescriptions d'antiémétiques anti-dopaminergiques attribuables aux épidémies de gastro-entérite aiguës (Roussel *et al.*, 2013). Notons que le terme de risque attribuable est privilégié pour les études portant sur la survenue d'une maladie ou la mortalité ; dans les autres contextes, on parlera plus volontiers de **fraction attribuable (FA)**.

Dans ces travaux, les estimations s'appuient le plus souvent sur des données agrégées au cours du temps, ou séries temporelles, construites à partir d'indicateurs de santé issus de systèmes de surveillance. Les données étant agrégées, il est impossible de distinguer les cas exposés des non exposés et, pour estimer le nombre de cas associés à l'exposition, un modèle incluant l'exposition doit être construit. Ces modèles sont le plus souvent issus de la classe des modèles linéaires (mixtes) généralisés. On retrouve fréquemment les régressions linéaire, de Poisson et négative-binomiale (Gilca *et al.*, 2009; Thompson *et al.*, 2009; Perrin *et al.*, 2010; Zhou *et al.*, 2012; Bernier *et al.*, 2014), les modèles ARIMA, permettant de prendre en compte une

2. En particulier, l'équation (1.2) implique RA $< p_{E|D}$.

dépendance au cours du temps (Carrat et Valleron, 1995; Gasparrini et Leone, 2014) mais aussi la régression de Serfling lorsque l'exposition présente un profil épidémique (Charu *et al.*, 2011; Nunes *et al.*, 2011).

1.2. Objectifs

Nous présentons les méthodes d'estimation de la fraction attribuable lorsque l'on travaille avec des données agrégées sous forme de séries temporelles, dans le cadre de l'analyse de l'association entre une exposition épidémique et un indicateur de santé saisonnier. Pour les modèles statistiques fréquemment rencontrés dans la littérature épidémiologique, nous établissons les distributions asymptotiques des estimateurs de la fraction attribuable. À partir d'une étude de simulation, nous étudions ensuite l'impact du modèle sur l'estimation de la fraction attribuable, puis l'impact de la non prise en compte d'un décalage temporel dans l'association entre l'exposition et l'événement de santé.

1.3. Estimateurs de la fraction attribuable

Nous considérons les modèles linéaires généralisés pour l'estimation de la fraction attribuable. En notant $\{y(t)\}$ la série du nombre de cas observés et $\{x(t)\}$ celle reflétant le niveau de l'exposition, où $t = t_1, t_2, \ldots$ désigne des dates d'observations régulières (par exemple la semaine), ces modèles sont caractérisés par trois éléments :

- La composante déterministe du modèle, appelée *prédicteur linéaire*, qui modélise l'espérance $\mu(t)$ de $y(t)$. Elle est combinaison linéaire des variables prédictives du modèle, $\beta x(t) + \eta(t)$, où $\eta(t)$ désigne l'ensemble des autres termes du prédicteur linéaire : $\eta(t) = Z(t)\gamma$, avec $Z(t)$ la matrice de design correspondant aux autres variables explicatives que le facteur d'exposition, et γ un vecteur de paramètres.
- La composante aléatoire, appelée aussi *structure d'erreur*, est définie par la loi de probabilité $\mathcal{L}_\theta$ de $y(t)$, où θ désigne le vecteur des paramètres du modèle. Elle détermine en particulier la relation entre l'espérance et la variance du modèle.
- La relation fonctionnelle g entre ces deux composantes, appelée *fonction de lien*, qui relie l'espérance $\mu(t)$ de $y(t)$ au prédicteur linéaire $\beta x(t) + \eta(t)$.

Ainsi, le modèle s'écrit :

$$y(t) \sim \mathcal{L}_\theta\Big(\mu(t)\Big),$$
$$g\Big(\mu(t)\Big) = \beta x(t) + \eta(t).$$

Dans le cadre des modèles linéaires généralisés,[3] les probabilités bien définies en (1.1) ne peuvent pas être définies à partir des paramètres. La définition de la fraction attribuable est alors fonction du modèle et, sous l'hypothèse d'un lien de causalité entre l'exposition et l'indicateur d'intérêt, correspond à la proportion des cas qui seraient évités si l'exposition était éliminée.

En notant $\{y^*(t)\}$ la série du nombre de cas en absence de l'exposition, telle que

$$y^*(t) \sim \mathcal{L}_\theta\Big(\mu^*(t)\Big),$$
$$g\Big(\mu^*(t)\Big) = \eta(t),$$

la fraction attribuable sur une période de temps $\mathcal{T} = \{t_1, t_2, \ldots\}$ s'écrit

$$\mathrm{FA}(\mathcal{T}) = \frac{\sum_{t\in\mathcal{T}} \Big(y(t) - y^*(t)\Big)}{\sum_{t\in\mathcal{T}} y(t)}. \tag{1.3}$$

Pour estimer la fraction attribuable, il faut donc estimer la valeur de $y^*(t)$, connaissant la valeur de $y(t)$. Toutefois, selon le modèle retenu, l'estimateur intuitif $\widehat{\mu}^*(t)$ de l'espérance de $y^*(t)$ n'est en général pas un bon estimateur, car il peut mener à des estimations négatives de la fraction attribuable. En effet, il n'est pas certain que $y(t) > \mu^*(t)$.[4]

Pour corriger ce point, l'estimateur de $y^*(t)$ choisi doit intégrer la composante aléatoire de $y(t)$, *i.e.* la valeur prise par la structure d'erreur. Dans ce qui suit, les estimateurs de la fraction attribuable pour les modèles couramment utilisés dans ce contexte sont présentés. On notera $\chi(t) = y(t) - y^*(t)$ la variable représentant le nombre de cas attribuables à l'exposition, $\{\varepsilon(t)\}$ un bruit blanc Gaussien et n le nombre d'observations.

Régression linéaire Lorsque la structure d'erreur est Gaussienne et que la fonction de lien est l'identité, le modèle linéaire généralisé

$$y_t \sim \mathcal{N}(\mu_t, \sigma^2),$$
$$\mu_t = \beta x(t) + \eta(t)$$

3. À l'exception du modèle binomial, bien entendu.

4. Par exemple, pour une régression de Poisson telle que $\mu(t) = 100$ et $\mu^*(t) = 90$, on a $\mathbb{P}(y(t) < 90) = 0{,}17$.

correspond au modèle de régression linéaire simple :

$$y(t) = \beta x(t) + \eta(t) + \varepsilon(t),$$

où la variance de $\varepsilon(t)$ est égale à σ^2.

Dans ce cadre, le calcul de la fraction attribuable ne pose pas de problème particulier. La structure d'erreur étant additive, le nombre de cas attribuables à l'exposition vaut simplement $\chi(t) = \beta x(t)$ et la fraction attribuable est définie par :

$$\mathrm{FA}(\mathcal{T}) = \beta \frac{\sum_{t\in\mathcal{T}} x(t)}{\sum_{t\in\mathcal{T}} y(t)}.$$

Lorsqu'il existe une dépendance temporelle entre les observations, le bruit $\{\varepsilon(t)\}$ peut être modélisé par un modèle ARMA [5] par exemple, (Shumway et Stoffer, 2011) ce qui n'impacte pas la formule ci-dessus, mais peut avoir une influence sur l'estimation de β.

Soit $\widehat{\beta}_n$ l'estimateur du maximum de vraisemblance de $\beta \in \theta$, l'estimateur de la fraction attribuable est donné par :

$$\widehat{\mathrm{FA}}_n(\mathcal{T}) = \widehat{\beta}_n \frac{\sum_{t\in\mathcal{T}} x(t)}{\sum_{t\in\mathcal{T}} y(t)}$$

et sa loi asymptotique peut être obtenue par la méthode delta (Doob, 1935; Oehlert, 1992) :

$$\sqrt{n}\left(\widehat{\mathrm{FA}}_n(\mathcal{T}) - \mathrm{FA}(\mathcal{T})\right) \to \mathcal{N}\left(0, \left(\frac{\sum_{t\in\mathcal{T}} x(t)}{\sum_{t\in\mathcal{T}} y(t)}\right)^2 I^{-1}(\beta)\right),$$

où $I(\beta)$ désigne l'information de Fisher de β.

Régression de Poisson La régression de Poisson s'écrit :

$$y(t) \sim \mathcal{P}\Big(\mu(t)\Big),$$
$$g\Big(\mu(t)\Big) = \beta x(t) + \eta(t).$$

La fonction de lien usuelle est la fonction logarithme, mais l'identité peut aussi être considérée.

À la différence de la régression linéaire, la structure d'erreur du modèle n'est ici pas additive, ce qui rend plus difficile le calcul du nombre de cas attribuables à l'exposition, $\chi(t)$. Afin d'introduire la structure d'erreur de $y(t)$ dans l'estimateur

5. AutoRegressive Moving Average model.

de la fraction attribuable, on suppose que $\chi(t)$ est indépendante de $y^*(t)$. Alors, étant donné la relation $y(t) = y^*(t) + \chi(t)$, il s'avère que $\chi(t)$ suit également une loi de Poisson, d'espérance $\mu(t) - \mu^*(t)$. [6] Dans ce cadre, la distribution conditionnelle de $\chi(t)$ sachant $y(t)$ est une loi binomiale.

En effet, la distribution jointe de $y(t)$, $y^*(t)$ et $\chi(t)$ est donnée par

$$\mathbb{P}(y(t) = n, y^*(t) = k, \chi(t) = l) = \begin{cases} \mathbb{P}(y^*(t) = k, \chi(t) = l), & \text{si } n = k + l, \\ 0, & \text{sinon.} \end{cases}$$

Posons $n = k + l$, on a alors :

$$\begin{aligned} &\mathbb{P}(y^*(t) = k, \chi(t) = l \mid y(t) = n) \\ &\qquad = \frac{\mathbb{P}(y(t) = n, y^*(t) = k, \chi(t) = l)}{\mathbb{P}(y(t) = n)} \\ &\qquad = \frac{\mathbb{P}(y^*(t) = k)\mathbb{P}(\chi(t) = l)}{\mathbb{P}(y(t) = n)} \\ &\qquad = \frac{e^{-\mu^*(t)}(\mu^*(t))^k}{k!} \frac{e^{-(\mu(t)-\mu^*(t))}(\mu(t) - \mu^*(t))^l}{l!} \frac{n!}{e^{-\mu(t)}(\mu(t))^n} \\ &\qquad = \binom{n}{k} \left(\frac{\mu^*(t)}{\mu(t)}\right)^k \left(\frac{\mu(t) - \mu^*(t)}{\mu(t)}\right)^l. \end{aligned}$$

On reconnait que la distribution conditionnelle de $\big(y^*(t), \chi(t)\big)$ sachant le nombre de cas $y(t)$ est une loi multinomiale dont on peut déduire que :

$$\chi(t)|y(t) \sim B(y(t), p_t) \tag{1.4}$$

où

$$p_t = \frac{\mu(t) - \mu^*(t)}{\mu(t)}. \tag{1.5}$$

6. Omettant la dépendance en t pour alléger les notations, comme $y \sim \mathcal{P}(\mu)$, sa fonction génératrice des probabilités est donnée par

$$G_y(z) = e^{\mu(z-1)}.$$

Similairement, celle de $y^* + \chi$ s'écrit (par indépendance)

$$G_{y^*+\chi}(z) = G_{y^*}(z)G_\chi(z) = e^{\mu^*(z-1)}G_\chi(z).$$

Étant donné que $y = y^* + \chi$, on obtient l'égalité des fonctions génératrices des probabilités, soit

$$e^{\mu(z-1)} = e^{\mu^*(z-1)}G_\chi(z).$$

En conséquence, $G_\chi(z) = e^{(\mu-\mu^*)(z-1)}$, c'est-à-dire que χ suit une distribution de Poisson d'espérance $\mu - \mu^*$.

Intuitivement, chaque cas de $y(t)$ est généré soit par le niveau de base $y^*(t)$, soit par la contribution de l'exposition $\chi(t)$.

En conséquence, la fraction attribuable s'écrit :

$$\mathrm{FA}(\mathcal{T}) = \frac{\sum_{t\in\mathcal{T}} p_t y(t)}{\sum_{t\in\mathcal{T}} y(t)}.$$

Comme pour la régression linéaire, en utilisant les propriétés de l'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ du vecteur de paramètres θ, l'estimateur de la fraction attribuable est donné par :

$$\widehat{\mathrm{FA}}_n(\mathcal{T}) = \frac{\sum_{t\in\mathcal{T}} \widehat{p}_t y(t)}{\sum_{t\in\mathcal{T}} y(t)}, \tag{1.6}$$

où $\widehat{p}_t = (\widehat{\mu}(t) - \widehat{\mu}^*(t))/\widehat{\mu}(t)$ désigne l'estimateur de p_t. Sa loi asymptotique peut également être obtenue par la méthode delta :

$$\sqrt{n}\left(\widehat{\mathrm{FA}}_n(\mathcal{T}) - \mathrm{FA}(\mathcal{T})\right) \to \mathcal{N}\Big(0, [\nabla g(\theta)]^\top I^{-1}(\theta)[\nabla g(\theta)]\Big),$$

où $g(\theta) = \mathrm{FA}(\mathcal{T})$.

Lorsque les données présentent une surdispersion (la variance est supérieure à l'espérance), on peut introduire une correction pour ajuster l'estimation de la variance des paramètres du modèle : on parle alors d'un modèle de quasi-Poisson. Les résultats établis ci-dessus persistent, et le calcul de la fraction attribuable est identique.

Une alternative à cette correction est d'utiliser un modèle négatif-binomial, généralisant la régression de Poisson, et plus flexible dans la paramétrisation de la variance que le modèle de quasi-Poisson. Pour plus de détails sur l'analyse et la modélisation de données de comptage par régressions de Poisson, quasi-Poisson et négative-binomiale, le lecteur peut se référer à Cameron et Trivedi (1998). Dans le cadre d'un modèle négatif-binomial, il n'existe pas de loi simple pour $\chi(t)$ telle que $y(t) = y^*(t) + \chi(t)$, même sous l'hypothèse d'indépendance avec $y^*(t)$. On ne peut donc déterminer explicitement la distribution de $\chi(t)$ connaissant $y(t)$. On posera donc, par défaut, le même estimateur (1.6) de la fraction attribuable que pour la régression de Poisson.

Régression de Serfling Lorsque l'exposition $\{x(t)\}$ présente un profil épidémique, c'est-à-dire que $x(t) \approx 0$ sauf pendant une période épidémique $\mathcal{T}_E = \{t_{i_1}, \ldots, t_{i_k}\}$, la régression de Serfling est souvent utilisée pour modéliser l'impact de $\{x(t)\}$ sur

le nombre de cas $\{y(t)\}$ (Serfling, 1963). Dans ce cadre, $y(t)$ est modélisé par une régression linéaire périodique en dehors de la période épidémique :

$$y(t) = \eta(t) + \varepsilon(t), \qquad \forall t \in \mathcal{T}_E^c,$$

où $\mathcal{T}_E^c$ désigne les périodes non épidémiques, et $\eta(t)$ est généralement un prédicteur périodique, par exemple :

$$\eta(t) = m + \alpha t + \sum_{k=1}^{K} \left[\gamma_k \cos\left(\frac{2k\pi}{T} t\right) + \delta_k \sin\left(\frac{2k\pi}{T} t\right) \right],$$

avec m une constante, α un paramètre correspondant au terme linéaire et γ_k et δ_k des paramètres correspondant aux harmoniques de période T/k de la saisonnalité.

En pratique, Serfling (1963) propose d'ajuster le modèle sur les périodes endémiques uniquement (sans épidémie). Si l'exposition présente des épidémies saisonnières (par exemple la grippe), les périodes épidémiques de la série d'exposition $\{x(t)\}$ sont déterminées par le dépassement d'une valeur seuil choisie (Costagliola *et al.*, 1991).

Pendant l'épidémie, le nombre de cas n'étant pas modélisé, les résidus du modèle $\widehat{\varepsilon}(t)$ $(t \in \mathcal{T}_E)$ ne sont pas estimés. Il est donc impossible d'intégrer la structure des erreurs de $y(t)$ sur les estimations de $y^*(t)$. C'est pourquoi on définit l'estimateur de la fraction attribuable à partir de l'espérance de $y^*(t)$, même si il peut mener à des estimations négatives de la fraction attribuable :

$$\widehat{\mathrm{FA}}(\mathcal{T}_E) = \frac{\sum_{t \in \mathcal{T}_E} \Big(y(t) - \widehat{\mu}^*(t) \Big)}{\sum_{t \in \mathcal{T}_E} y(t)}.$$

Prise en compte du décalage Un point important à considérer lorsque l'on s'intéresse à l'association entre deux séries temporelles concerne la temporalité de l'effet. Pour l'étudier, il est possible d'introduire dans les modèles des décalages temporels, de la forme :

$$y(t) \sim \mathcal{L}_\theta\Big(\mu(t)\Big),$$
$$g\Big(\mu(t)\Big) = \sum_{k=l}^{L} \beta_k x(t-k) + \eta(t),$$

où l et L représentent les décalages minimal et maximal de l'association (éventuellement $l < 0$, signifiant que $y(t)$ dépend des valeurs futures de $x(t)$), et β_k les paramètres associés à chaque décalage.

En pratique, l'exploration du décalage entre les deux séries passe par l'étude des corrélations croisées. Ces corrélations décrivent la force de l'association linéaire entre les deux séries pour tous les décalages k possibles, et permettent en particulier d'identifier le délai minimum l et maximum L de l'association. Pour que les corrélations croisées soient interprétables, il est nécessaire d'effectuer un blanchiment (retrait de la tendance, de la saisonnalité et des autocorrélations résiduelles) d'au moins une des deux séries (Bowie et Prothero, 1981; Shumway et Stoffer, 2011). Il n'existe pas de consensus sur la série à blanchir, et certains auteurs préfèrent blanchir la série à expliquer $\{y(t)\}$ (Hubert *et al.*, 1992), d'autres la série explicative $\{x(t)\}$ (Opatowski *et al.*, 2013), et d'autres encore les deux (Bowie et Prothero, 1981).

Dans le cadre de l'étude d'un facteur de risque, l'analyse de la temporalité de l'effet est essentielle. Comme la cause doit précéder l'effet, l'unidirectionnalité de l'association est un argument majeur en faveur d'une relation de causalité. En revanche, si la cause et l'effet sont simultanés, cela pourrait simplement être le reflet d'un ou plusieurs facteurs de confusion qui agissent simultanément sur les deux phénomènes d'intérêt. Les valeurs des décalages minimum l et maximum L sont donc importantes pour pouvoir étayer l'interprétation de l'association, et en particulier, pour une relation causale, il est nécessaire que $l \geq 0$.

S'il existe des décalages temporels, la fraction attribuable doit alors être corrigée en conséquence. Dans le cadre de la régression linéaire avec erreurs Gaussiennes, elle devient :

$$\mathrm{FA}(\mathcal{T}) = \frac{\sum_{t\in\mathcal{T}} \sum_{k=l}^{L} \beta_k x(t-k)}{\sum_{t\in\mathcal{T}} y(t)}.$$

Pour la régression de Poisson, c'est la valeur de p_t qui change, selon le même raisonnement. Notons toutefois qu'il n'est pas possible de prendre en compte ces décalages temporels avec la régression de Serfling, puisqu'elle ne modélise pas d'association.

L'interprétation usuelle du décalage dans la fraction attribuable signifie que l'exposition a un effet retardé ou prolongé, et que le nombre de cas attribuables au temps t dépend de l'exposition aux temps $t-l, t-l-1, \ldots, t-L$. Gasparrini et Leone (2014) discutent d'une approche alternative, prospective plutôt que rétrospective, et propose un estimateur de la fraction attribuable qui calcule l'impact que l'exposition à un temps t a sur le nombre de cas aux temps $t+l, t+l+1, \ldots, t+L$. Ces deux approches répartissent donc différemment le nombre de cas attribuables dans le temps, et le choix de l'une d'entre elles doit être motivé par la question posée.

Ici, nous avons choisi l'approche rétrospective, plus classique et plus fréquemment rencontrée, bien que les estimateurs définis précédemment peuvent aussi s'adapter à l'approche prospective.

1.4. Etude de simulation

La notion de fraction attribuable est souvent retrouvée dans l'étude de phénomènes infectieux qui présentent une saisonnalité (Simonsen *et al.*, 2005; Rousel *et al.*, 2013; Opatowski *et al.*, 2013). Ces saisonnalités, parfois non expliquées, peuvent cependant être corrélées à des phénomènes épidémiques, par exemple les épidémies de grippe (Muscatello *et al.*, 2013; Matias *et al.*, 2017). On cherche alors à déterminer si l'association est causale, et le cas échéant, à calculer la fraction attribuable pour en déterminer l'impact.

L'estimation de la fraction attribuable étant fonction du modèle retenu (voir Section 1.3), les estimations peuvent être très variables (Gilca *et al.*, 2009; Thompson *et al.*, 2009; Lemaitre *et al.*, 2012). Le choix du modèle le plus adapté à la question posée se fonde, le plus souvent, sur la nature des données, mais il peut être difficile de choisir *a priori* entre plusieurs possibilités théoriquement valides.

L'étude de simulation qui suit a pour objectif d'étudier l'impact du modèle et de la valeur de la fraction attribuable sur son estimation, dans le cadre de l'analyse de l'association entre une exposition épidémique et un phénomène saisonnier. Elle permettra aussi de valider les estimateurs définis en Section 1.3, ainsi que leur variance. Un deuxième jeu de simulation s'intéressera aux performances des estimations de la fraction attribuable dans le cas où le décalage de l'association entre l'exposition et le nombre de cas est mal identifié.

1.4.1. Processus de simulation

Des séries hebdomadaires ont été simulées sur une période de $N = 10$ ans.

Simulation de l'exposition Nous avons choisi de simuler l'exposition de sorte qu'elle présente des profils semblables à ceux des épidémies de grippe. Pour cela, la série temporelle du niveau d'exposition $\{x(t)\}$ est construite par la superposition de N profils épidémiques $\{l_k(t)\}$ bruités par des termes d'erreur $\{w_k(t)\}$ $(1 \leq k \leq N)$: $x(t) = \sum_{k=1}^{N} \Big(l_k(t) + w_k(t)\Big)$.

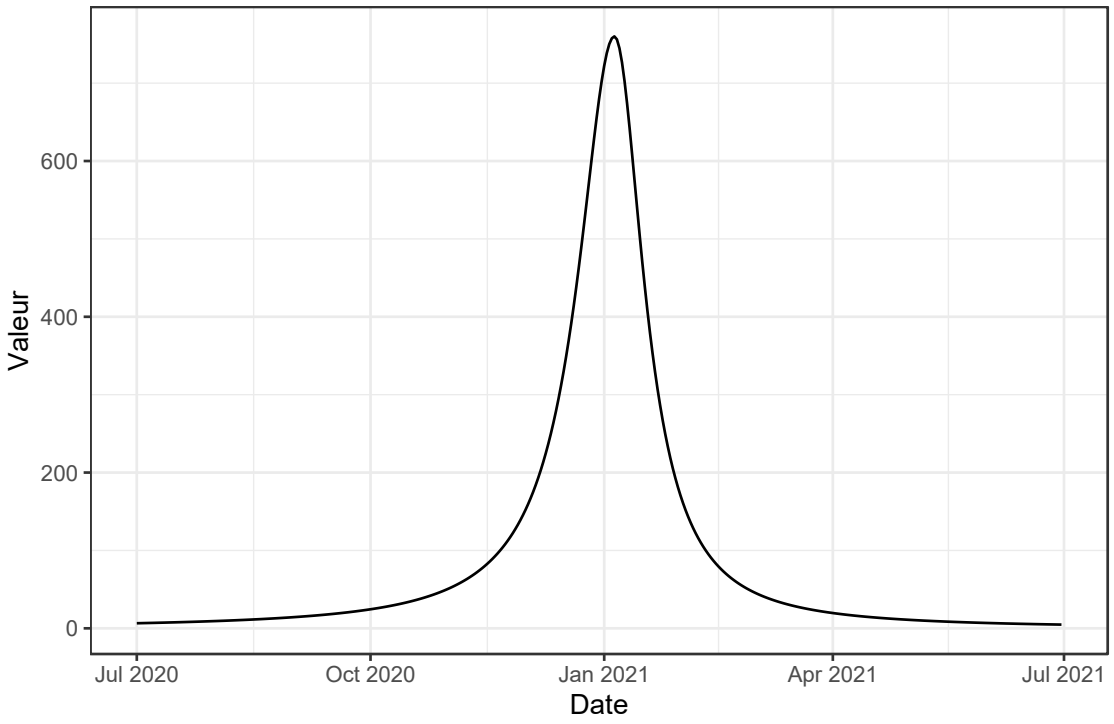


FIGURE 1.1 – Fonction de Lorentz asymétrique avec $h_k = 760$ pour $100000, \tau_k =$ 5 janvier, $r_k = 35$ jours et $s_k = 28$ jours.

Le modèle choisi pour générer les profils épidémiques $\{l_k(t)\}$ chaque année k est une fonction lorentzienne asymétrique (voir Figure 1.1), qui présente l'avantage de pouvoir être paramétrée directement à partir des caractéristiques des épidémies :

$$l_k(t) = \begin{cases} h_k \dfrac{(r_k/2)^2}{(r_k/2)^2 + (t-\tau_k)^2}, & t < \tau_k, \\ h_k \dfrac{(s_k/2)^2}{(s_k/2)^2 + (t-\tau_k)^2}, & t \geq \tau_k, \end{cases}$$

où les paramètres (h_k, τ_k, r_k, s_k) désignent respectivement le taux d'incidence au pic, la date du pic, la durée entre le début de l'épidémie et le pic, et la durée entre le pic et la fin de l'épidémie. Les termes d'erreur $\{w_k(t)\}$ suivent une distribution Gaussienne centrée, de variance proportionnelle au niveau épidémique : $w_k(t) \sim_{iid} \mathcal{N}\Big(0, \alpha \cdot l_k(t)\Big)$.

Pour que les simulations ressemblent aux épidémies de grippe en durée, date du pic et incidence au pic, les paramètres (h_k, τ_k, r_k, s_k) sont tirés aléatoirement selon les distributions données en Table 1.1. Ces distributions sont déterminées à partir des caractéristiques des épidémies de syndromes grippaux rapportées par le réseau Sentinelles (voir Annexe 1.B). Par ailleurs, les périodes épidémiques $\mathcal{T}_E$ de l'exposition, qui seront utilisées pour le calcul de la fraction attribuable, sont

déterminées en fixant un seuil de dépassement égal à 279 cas pour 100 000 personnes, conformément aux outils de détection du réseau Sentinelles pour la surveillance des syndromes grippaux en France. [7]

TABLEAU 1.1 – Distributions choisies des paramètres de simulation de l'exposition.

Paramètre	**Distribution**	**Espérance**	**Écart-type**
h_k : taux d'incidence au pic	Log-normale	760 pour 100 000	300 pour 100 000
τ_k : date du pic	Normale	5 janv.	30 jours
r_k : durée avant le pic	Log-normale	35 jours	7 jours
s_k : durée après le pic	Log-normale	28 jours	7 jours
α : facteur d'erreur	Valeur fixée	4	—

Simulation du nombre de cas Nous avons choisi de simuler la série du nombre de cas de sorte qu'elle présente une forte saisonnalité. Pour cela, la partie du prédicteur linéaire indépendante de l'exposition est donnée par le terme périodique suivant :

$$\eta(t) = \mu + \sum_{k\in\{1,2,4\}} \left[\gamma_k \cos\left(\frac{2k\pi}{52}t\right) + \delta_k \sin\left(\frac{2k\pi}{52}t\right)\right]. \tag{1.7}$$

Trois modèles sont considérés pour générer trois versions du nombre de cas— deux additifs : la régression linéaire avec erreurs autocorrélées et la régression de Poisson avec fonction de lien identité ; et un multiplicatif : la régression de Poisson avec fonction de lien logarithme.

À partir du modèle choisi et de la partie indépendante de l'exposition, la série du nombre de cas en absence de l'exposition, $\{y^*(t)\}$, est simulée. La contribution de l'exposition, $\{\chi(t)\}$, est ensuite simulée — par le processus déterministe $\chi(t) = \beta x(t)$ pour la régression linéaire, ou selon une loi de Poisson d'espérance $\mu(t) - \mu^*(t)$ pour les régressions de Poisson —, en fixant le paramètre β de telle sorte que la valeur de la fraction attribuable soit fixée à une valeur donnée. Cette fraction, calculée sur l'ensemble des N années, est donnée par

$$\overline{\mathrm{FA}} = \frac{1}{N}\sum_{k=1}^{N} \mathrm{FA}(\mathcal{T}_{E_k}),$$

7. `http://www.sentiweb.fr/document/methode_detection_epidemies` (dernier accès 06/06/2020).

Tableau 1.2 – Paramètres de simulation de la série du nombre de cas.

Paramètre	**Modèles additifs**	**Modèle multiplicatif**
Terme périodique $\eta(t)$		
μ : niveau moyen	1250	7
γ_1 : harmonique de période 52	-400	-0,25
γ_2 : harmonique de période 26	0	-0,05
γ_4 : harmonique de période 13	100	0,1
δ_1 : harmonique de période 52	0	0
δ_2 : harmonique de période 26	-100	-0,05
δ_4 : harmonique de période 13	-100	-0,05
Termes spécifiques à la régression linéaire avec erreurs autocorrélées		
Polynôme autorégressif saisonnier SAR(1)$(1)_{52}$	$(1-\phi_1 z)(1-\varphi_1 z^{52})$	
ϕ_1 : coefficient autorégressif d'ordre 1	0,4	
φ_1 : coefficient autorégressif d'ordre 52	0,7	
σ^2 : Variance du terme d'erreur	2500	

où $\mathcal{T}_{E_k}$ désigne la période épidémique de l'année k.

Les valeurs choisies pour les paramètres du terme périodique et des paramètres spécifiques à la régression linéaire avec erreurs autocorrélées sont présentées dans la Table 1.2. Ces valeurs ont été déterminées à partir de la série du taux hebdomadaire de prescriptions antibiotiques pour 100 000 en ville, présentée dans le Chapitre suivant. Les valeurs considérées pour $\overline{\text{FA}}$ vont de 10% à 50%, avec un pas de 10%.

Pour étudier l'impact d'un décalage temporel mal ou non identifié, un deuxième jeu de simulation est généré de manière similaire en simulant la contribution de l'exposition, $\{\chi(t)\}$, à partir de l'exposition $\{x(t-1)\}$.

Les simulations ont été effectuées avec R (R Core Team, 2019), à partir des fonctions `rnorm`, `rlnorm` et `rpois`, qui s'appuient sur le générateur de nombres aléatoires Mersenne Twister.

1.4.2. Analyse des simulations

Pour chacune des simulations, la fraction attribuable est estimée avec six modèles : trois modèles additifs, deux multiplicatifs et la régression de Serfling. Il s'agit,

pour les modèles additifs (notés “a-”) de la régression linéaire avec erreurs autocorrélées, “a-Gaussien”, de la régression de Poisson, “a-Poisson”, et de la régression négative-binomiale, “a-NégBin”, avec fonction de lien identité. Pour les modèles multiplicatifs (notés “m-”), ce sont la régression de Poisson, “m-Poisson”, et la régression négative-binomiale, “m-NégBin”, avec fonction de lien logarithme. La régression de Serfling ne modélisant pas directement la contribution de l'exposition, elle n'est pas considérée comme les autres modèles additifs.

Critères d'évaluation Notons s le nombre de simulations effectuées pour chaque modèle et chaque jeu de paramètres et $\widetilde{\mathrm{FA}}$ l'estimation de la fraction attribuable moyenne $\overline{\mathrm{FA}}$. Ces estimations sont évaluées et comparées par les critères suivants (Burton *et al.*, 2006) :

- Le *biais relatif* (BR) mesure l'écart relatif entre l'estimation $\widetilde{\mathrm{FA}}$ et la vraie valeur :

$$\mathrm{BR} = \frac{1}{s}\sum_{i=1}^{s}\frac{\widetilde{\mathrm{FA}}_i - \overline{\mathrm{FA}}}{\overline{\mathrm{FA}}}.$$

  Le modèle préféré est celui avec le biais relatif le plus petit, en valeur absolue.

- La racine de l'*erreur quadratique moyenne* (REQM) est une mesure de l'écart-type de l'estimateur :

$$\mathrm{REQM} = \sqrt{\frac{1}{s}\sum_{i=1}^{n}\left(\widetilde{\mathrm{FA}}_i - \overline{\mathrm{FA}}\right)^2}.$$

  Pour déterminer si la REQM est acceptable, on peut faire le rapport de la REQM sur la valeur de la fraction attribuable. Bien qu'il n'existe pas de recommandations, le modèle peut être considéré bon si ce rapport est inférieur à 10% et mauvais s'il est supérieur à 40%.

- La probabilité de recouvrement (PR) correspond à la proportion des simulations pour lesquelles l'intervalle de confiance de niveau 95% contient la vraie valeur $\overline{\mathrm{FA}}$, c'est-à-dire telles que :

$$\overline{\mathrm{FA}} \in \left[\widetilde{\mathrm{FA}} - 1{,}96\sqrt{\mathrm{Var}\left(\widetilde{\mathrm{FA}}\right)}; \widetilde{\mathrm{FA}} + 1{,}96\sqrt{\mathrm{Var}\left(\widetilde{\mathrm{FA}}\right)}\right],$$

  où le calcul de l'intervalle de confiance fait appel à la normalité asymptotique de l'estimateur de la fraction attribuable. La valeur nominale de la probabilité de recouvrement est de 95%.

1.4.3. Résultats

Sans décalage temporel de l'association Pour chaque valeur de la fraction attribuable moyenne, et chacun des trois modèles considérés pour simuler le nombre de cas, $s = 1000$ simulations ont été générées. Des exemples de simulation de l'exposition et du nombre d'événements pour chacun des modèles sont donnés en Figure 1.2, ainsi qu'en Annexe 1.C (Figures 1.C.1 et 1.C.2). Les expositions simulées présentent des profils épidémiques très semblables aux épidémies de grippe reportées par le réseau Sentinelles, et les simulations du nombre de cas présentent un schéma saisonnier marqué, avec peu de fluctuations pendant les périodes estivales mais de fortes fluctuations pendant les périodes hivernales. Les simulations issues de la régression linéaire "a-Gaussien" présentent davantage de variabilité résiduelle, en lien avec la structure d'erreur imposée au modèle. Pour des valeurs élevées de la fraction attribuable ($\geq 30\%$), les simulations issues du modèle "m-Poisson" présentent des écarts beaucoup plus larges entre les pics hivernaux et les creux estivaux, et les nombres de cas sont concentrés sur la période épidémique. Les critères d'évaluation de l'estimation de la fraction attribuable pour chacun des modèles de simulation sont reportés en Annexe 1.D (Tableaux 1.D.1, 1.D.2 et 1.D.3).

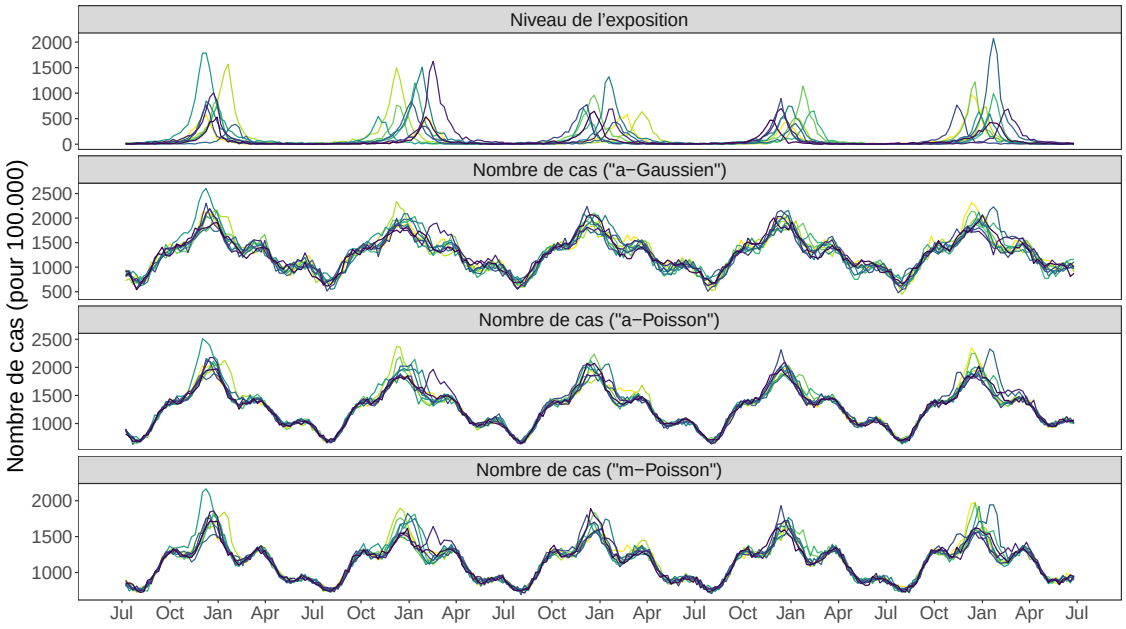


FIGURE 1.2 – Dix jeux de simulation pour les séries d'exposition et du nombre de cas, avec $\overline{\text{FA}} = 10\%$. Seules les 5 premières années simulées ont été représentées.

Comme attendu, lorsque le modèle d'estimation est identique à celui qui a servi à

simuler le nombre de cas (voir Figure 1.3), la fraction attribuable est correctement estimée quelque soit sa valeur, avec un biais relatif inférieur à 0,2%, une erreur quadratique moyenne proche de 0 et un taux de recouvrement proche de 95%, à l'exception du modèle "a-Gaussien", pour lequel le taux de recouvrement est légèrement plus faible et se situe autour de 93%. La taille de la fraction attribuable ne semble pas avoir d'impact sur le biais relatif et le taux de recouvrement, alors que l'erreur quadratique moyenne diminue légèrement lorsque la fraction attribuable augmente pour les modèles additifs "a-Gaussien" et "a-Poisson" (de moitié et d'un tiers respectivement, entre les valeurs de 10% et 50% de la fraction attribuable). Comme la fraction attribuable est bien retrouvée lorsque le modèle d'estimation est identique au modèle de simulation, ces modèles seront pris comme référence pour comparer les autres modèles d'estimation (Figure 1.4).

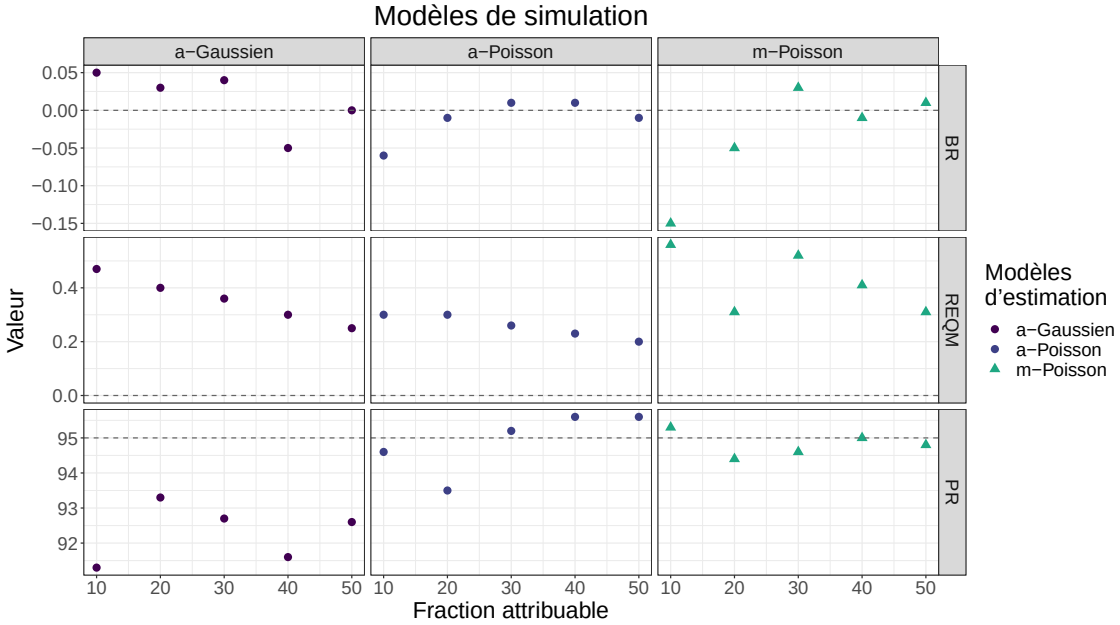


FIGURE 1.3 – Critères d'évaluation pour l'estimation de la fraction attribuable, lorsque le modèle d'estimation est identique au modèle de simulation. BR = biais relatif, REQM = racine de l'erreur quadratique moyenne, PR = probabilité de recouvrement. Les lignes pointillées correspondent aux valeurs nominales des critères d'évaluation.

Quelque soit le modèle de simulation, les performances de **la régression de Serfling** sont médiocres. En effet, l'estimation de la fraction attribuable est toujours sous-estimée avec un biais relatif qui varie entre -10% et -20%, une racine de l'erreur quadratique moyenne qui augmente linéairement avec la taille de la fraction

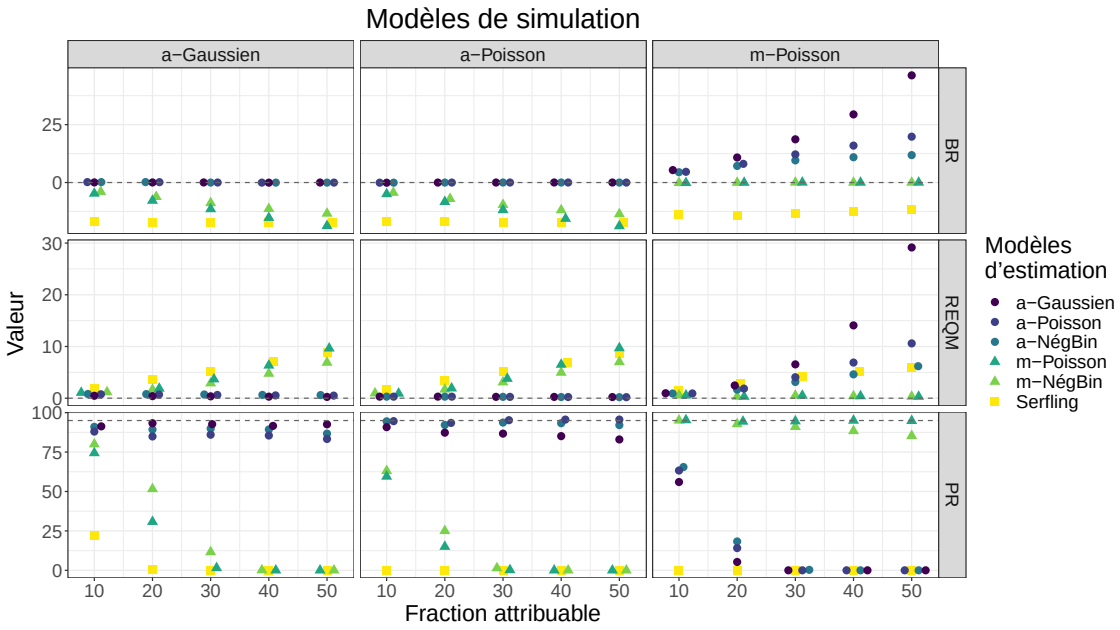


FIGURE 1.4 – Critères d'évaluation pour l'estimation de la fraction attribuable, selon le modèle de simulation (en colonne). BR = biais relatif, REQM = racine de l'erreur quadratique moyenne, PR = probabilité de recouvrement. Les lignes pointillées correspondent aux valeurs nominales des critères d'évaluation.

attribuable (de 1,5 à 9), et enfin un taux de recouvrement presque toujours égal à 0%.

Les estimations des **modèles additifs "a-Gaussien", "a-Poisson" et "a-NégBin"** sont similaires. En effet, lorsque la simulation est générée par un modèle additif, les biais relatifs et les erreurs quadratiques moyennes de ces modèles sont presque identiques aux modèles de référence. Néanmoins, on observe un taux de recouvrement légèrement plus faible pour les modèles "a-Poisson" et "a-NégBin" lorsque le modèle de simulation est "a-Gaussien" (jusqu'à 9,3 points de pourcentage de moins lorsque $\overline{\text{FA}} = 50\%$), et pour les modèles "a-Gaussien" et "a-NégBin" lorsque la simulation est générée par le modèle "a-Poisson" (jusqu'à 12,6 points).

Lorsque la simulation est généré par le modèle multiplicatif "m-Poisson", les performances des modèles additifs sont peu satisfaisantes. Pour une fraction attribuable de 10%, bien que le biais relatif et l'erreur quadratique moyenne sont toutes les deux petites, le taux de recouvrement est faible, de l'ordre de 60%. Lorsque la fraction attribuable augmente, le biais relatif et la racine de l'erreur quadratique moyenne augmentent, de façon quasi-linéaire pour les modèles "a-Poisson" et "a-NégBin" et atteignent, pour une fraction attribuable de 50%, 19,80% et 11,84%

respectivement pour le biais relatif, et 10,6 et 6,19 respectivement pour la racine de l'erreur quadratique moyenne. Pour le modèle "a-Gaussien", le biais relatif et la racine de l'erreur quadratique moyenne augmentent de façon quadratique en fonction de la fraction attribuable, atteignant 46,28% et 29,15 pour une fraction attribuable de 50%. Les taux de recouvrement des trois modèles diminuent suivant les valeurs de la fraction attribuable, et atteignent 0% pour des valeurs de la fraction attribuable supérieures ou égales à 30%.

De manière analogue, les estimations des **modèles multiplicatifs "m-Poisson" et "m-NégBin"** sont similaires. Lorsque la simulation est générée par un modèle additif, les biais relatifs et erreurs quadratiques moyennes de ces modèles sont acceptables lorsque la fraction attribuable vaut 10%, mais augmentent linéairement (en valeur absolue), pour atteindre environ -18% et 10 respectivement pour "m-Poisson", et -13,5% et 7 respectivement pour "m-NégBin". Les taux de recouvrement sont médiocres lorsque la fraction attribuable vaut 10% (entre 60 et 80 %), et tombent à 0% lorsqu'elle est supérieure ou égale à 40%.

Dans le cas où la simulation est générée par le modèle multiplicatif "m-Poisson", les performances du modèle "m-NégBin" sont très similaires au modèle de référence. On remarque une légère différence entre les taux de recouvrement des deux modèles : pour des valeurs de la fraction attribuable supérieures ou égales à 30%, les taux de recouvrement pour "m-NégBin" s'écartent de 4 à 10 points de pourcentage de celles de "m-Poisson".

Avec décalage temporel inconnu de l'association Le deuxième jeu de simulation porte sur les conséquences de l'estimation de la fraction attribuable lorsqu'il existe un décalage temporel d'une unité de temps non pris en compte dans l'association entre les deux séries. Les critères d'évaluation du jeu de simulation utilisé pour étudier l'impact d'un décalage temporel inconnu sont reportés en Annexe 1.D (Tableaux 1.D.4, 1.D.5 et 1.D.6). Lorsque le modèle d'estimation est identique à celui qui a servi à simuler le nombre de cas (voir Figure 1.5), la fraction attribuable n'est pas correctement estimée, même pour des valeurs faibles de la fraction attribuable. Les biais relatifs sont élevés (entre -17% et -29% pour "a-Gaussien", autour de -7% pour "a-Poisson" et "m-Poisson"), la racine de l'erreur quadratique augmente avec la valeur de la fraction attribuable (de 1,82 à 24,26 pour "a-Gaussien", de 0,83 à 3,08 pour "a-Poisson" et de 0,9 à 4,06 pour "m-Poisson"). Les taux de recouvrement sont

médiocres pour $\overline{\text{FA}} = 10\%$ (respectivement 12,4%, 42,7% et 37,7%) et s'approchent de 0% pour des valeurs de la fraction attribuable plus élevées. En particulier, les performances du modèle "a-Gaussien", sensible à la structure d'autocorrélation des données, se détériorent considérablement lorsque la fraction attribuable augmente.

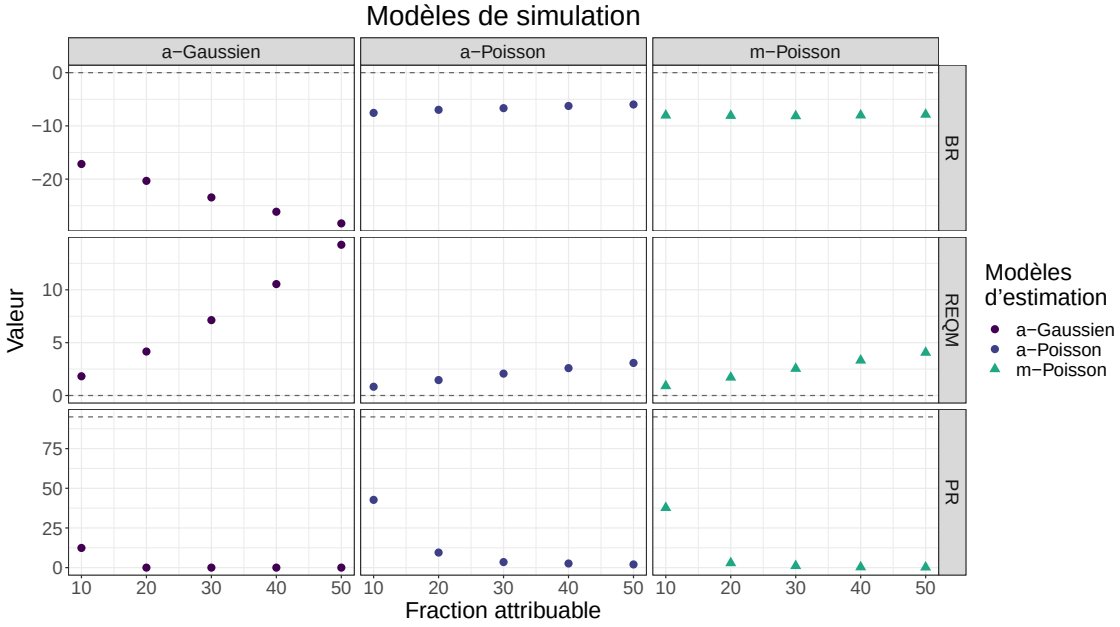


FIGURE 1.5 – Critères d'évaluation pour l'estimation de la fraction attribuable, lorsque le modèle d'estimation est identique au modèle de simulation, mais que le décalage temporel est mal identifié. BR = biais relatif, REQM = racine de l'erreur quadratique moyenne, PR = probabilité de recouvrement. Les lignes pointillées correspondent aux valeurs nominales des critères d'évaluation.

Les performances des autres modèles d'estimation sont représentées Figure 1.6. De manière générale, aucun des modèles n'est performant pour l'estimation de la fraction attribuable. Pour des valeurs de la fraction attribuable supérieures à 30%, les biais relatifs sont supérieurs à 5%, et les taux de recouvrement ne dépassent pas 40%.

Pour des valeurs de la fraction attribuable inférieures à 30%, on note que le modèle "a-NégBin" est le meilleur des six modèles d'estimation pour tous les modèles de simulation considérés (même pour le modèle de simulation "m-Poisson"), avec les biais relatifs et les racines de l'erreur quadratique moyenne les plus faibles, et les taux de recouvrement les plus élevés. Cependant, ces performances restent médiocres, avec des biais relatifs entre -7% et +4%, et des taux de recouvrement inférieurs à 80%.

Les performances de la régression de Serfling sont identiques aux jeux de simulations précédents, c'est-à-dire qu'aucun des critères d'évaluation n'est bon : le biais relatif est autour de -20%, la racine de l'erreur quadratique moyenne augmente linéairement avec la fraction attribuable et le taux de recouvrement est presque toujours égal à 0%.

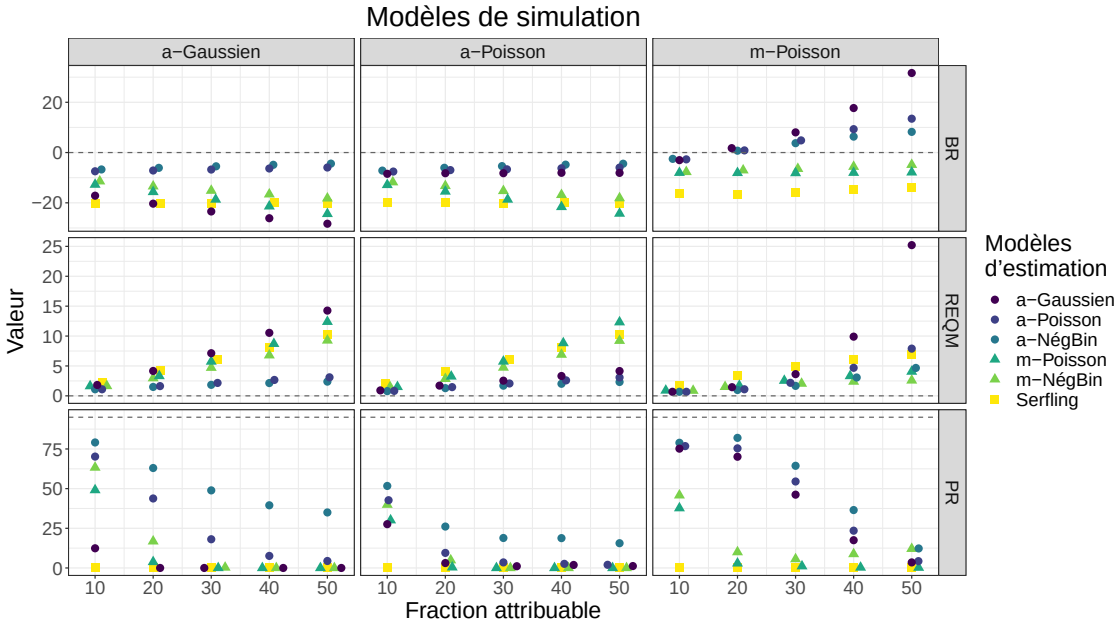


FIGURE 1.6 – Critères d'évaluation pour l'estimation de la fraction attribuable, selon le modèle de simulation (en colonne), lorsque le décalage temporel est mal identifié. BR = biais relatif, REQM = racine de l'erreur quadratique moyenne, PR = probabilité de recouvrement. Les lignes pointillées correspondent aux valeurs nominales des critères d'évaluation.

Résumé des résultats Les critères d'évaluation de l'étude de simulation ont révélé que : *(i)* lorsque le modèle d'estimation est identique au modèle de simulation, la fraction attribuable est correctement estimée ; *(ii)* les modèles ne sont pas robustes à une erreur sur la fonction de lien choisie : les performances de l'estimation sont mauvaises lorsque l'identité est utilisée à la place du logarithme et *vice versa* ; *(iii)* la régression de Serfling sous-estime systématiquement la fraction attribuable, quelque soit la fonction de lien du modèle de simulation ; *(iv)* les modèles ne sont pas robustes à la non prise en compte ou une erreur dans la prise en compte du décalage temporel d'une unité de temps de l'association entre les deux séries : l'estimateur de la fraction attribuable est alors biaisé.

1.5. Discussion

Dans ce travail, nous avons montré comment définir des estimateurs de la fraction attribuable dans le cadre de l'analyse de l'association entre une exposition épidémique et un phénomène saisonnier à partir de données agrégées au cours du temps. Pour les modèles fréquemment utilisés en épidémiologie, nous avons établi que ces estimateurs sont consistants et asymptotiquement Gaussiens, permettant de calculer des intervalles de confiance. Ces estimateurs ont été validés par une étude de simulation qui a aussi montré que les modèles ne sont ni robustes à une erreur de spécification de la fonction de lien de l'association, ni à une erreur dans la prise en compte d'un décalage temporel dans l'association entre la série du nombre d'événements et celle de l'exposition. Les estimateurs sont par ailleurs d'autant plus sensibles à ces erreurs que la fraction attribuable est élevée.

La fraction attribuable dépend d'une variable aléatoire non observée, le niveau de base du nombre de cas, au contraire du risque attribuable, qui est une constante définie directement à partir des probabilités de l'exposition et de la maladie. Alors qu'une stratégie consiste à estimer le niveau de base du nombre de cas par son espérance (Gilca *et al.*, 2009; Fenelon et Preston, 2012), nous avons présenté un estimateur de la fraction attribuable qui estime ce niveau en intégrant la composante aléatoire observée sur le nombre d'événements. Il s'agit d'un estimateur dit *contrefactuel*, qui s'intéresse à l'impact réel plutôt que l'impact moyen qu'aurait eu une intervention sur l'exposition. Cette idée a également été introduite sous une forme différente par Gasparrini et Leone (2014), qui présentent un estimateur contrefactuel en étendant la notion de risque aux modèles à retards échelonnés (Gasparrini *et al.*, 2010), qui forment une famille spécifique de modèles linéaires généralisés. Depuis, l'utilisation d'estimateurs contrefactuels de la fraction attribuable a rencontré un certain intérêt, notamment lorsque les facteurs d'exposition sont les températures extrêmes (Deng *et al.*, 2019; Achebak *et al.*, 2019). Nos travaux ont détaillé l'estimateur contrefactuel de la fraction attribuable dans le cadre de deux modèles linéaires généralisés particuliers. Un cadre plus général pour les modèles linéaires généralisés est présenté en Annexe 1.E, en faisant appel aux travaux sur la causalité de Pearl (2009).

L'étude de simulation a permis de comparer l'impact du choix du modèle sur l'estimation de la fraction attribuable pour six modèles : la régression linéaire avec

erreurs autocorrélées, les régressions de Poisson et négative-binomiale avec fonctions de lien identité et logarithme, et la régression de Serfling. Les résultats montrent une grande similarité entre les estimations issues des modèles additifs (fonction de lien identité), et de même entre les modèles multiplicatifs (fonction de lien logarithme). En particulier, nous observons des résultats très similaires entre les simulations issues des régressions de Poisson et négatives-binomiales. Enfin, la régression de Serfling a sous-estimé systématiquement la fraction attribuable.

Les similarités entre la régression linéaire et le modèle de Poisson additif étaient attendues, puisque la loi de Poisson peut être approchée par une distribution gaussienne lorsque l'espérance est élevée. Une différence entre la régression linéaire et le modèle de Poisson additif est cependant observée lorsque les simulations sont générées par un modèle de Poisson multiplicatif : lorsque la valeur de la fraction attribuable augmente, la racine de l'erreur quadratique moyenne augmente de façon linéaire pour le modèle de Poisson additif, mais de façon quadratique pour la régression linéaire. Ceci pourrait être dû à la différence entre les structures de variance des deux modèles : constante pour la régression linéaire et proportionnelle à l'espérance pour la régression de Poisson. En conséquence, la régression linéaire surestime la variance résiduelle pour tenir compte de la dispersion du nombre de cas. Les ressemblances entre les estimations des régressions de Poisson et négatives-binomiales pourraient être dues au fait qu'il existe peu de différence d'ordre de grandeur entre les valeurs minimales et maximales de la série du nombre d'événements (entre 500 et 2500 pour une fraction attribuable de 10%). Il est intéressant de noter que les régressions de quasi-Poisson et négatives-binomiales pourraient conduire à des estimations très différentes, car la forme du paramètre de dispersion pondère différemment les observations de la série (Ver Hoef et Boveng, 2007). Enfin, les mauvaises performances de la régression de Serfling proviennent du fait que l'exposition n'est pas tout à fait nulle autour des périodes épidémiques. En conséquence, l'ajustement du niveau de base du nombre d'événements est sur-évalué, et la fraction attribuable sous-estimée.

En résumé, à l'exception du modèle de Serfling, dont les performances ne sont pas bonnes dans tous les cas considérés, et que je ne recommande donc pas pour l'estimation de la fraction attribuable lorsque l'exposition est épidémique, les performances des autres modèles dépendent principalement de la nature additive ou multiplicative de la série du nombre d'événements. En particulier, même pour des

valeurs faibles de la fraction attribuable, les critères d'évaluation des modèles sont insuffisants lorsque la fonction de lien est mal identifiée, avec des taux de recouvrement bien inférieurs au taux nominal de 95%. Le choix de la fonction de lien est donc essentiel et doit, en pratique, être motivé par l'observation de la série du nombre d'événements, et la connaissance des phénomènes d'exposition. Par exemple, une série dont l'amplitude des fluctuations saisonnières varie de façon proportionnelle avec le niveau de la série invite à utiliser la fonction de lien logarithme. Dans le cas où aucun élément n'indique spécifiquement la fonction de lien logarithme, il peut être préférable de choisir la fonction de lien l'identité, qui mènera à une paramétrisation du modèle plus simple à interpréter. Ces éléments sont un premier pas vers le choix du modèle, permettant de déterminer la fonction de lien qui doit être utilisée. Il est ensuite nécessaire de choisir la distribution des erreurs, à partir des variations résiduelles observées : de variance constante pour la régression linéaire, proportionnelle à l'espérance pour la régression de Poisson, surdispersée pour la régression négative-binomiale. En particulier, dans le cas où les données de comptage sont rares, les régressions de type Poisson devraient être préférées puisque l'approximation par une loi gaussienne n'est pas valide. Si une structure résiduelle persiste après la modélisation, il faut alors intégrer une structure de dépendance temporelle : par exemple avec un modèle ARMA ou, si l'approximation par une loi gaussienne n'est pas adaptée, un modèle de Poisson autorégressif (Brandt et Williams, 2000).

Aucun des modèles n'est robuste lorsque le décalage dans l'association entre les deux séries est mal identifié ou non pris en compte. En effet, puisque l'exposition est ici épidémique, un décalage, même d'une semaine, réduit fortement la corrélation entre le niveau de l'exposition et le nombre de cas, et conduit à une sous-estimation de l'association. Il est donc essentiel d'identifier et de prendre en compte le décalage temporel qui peut exister dans l'association entre les deux séries. On remarquera que, de façon prévisible, les résultats issus du modèle de Serfling n'ont pas été fortement affectés par le décalage temporel, car il ne modélise pas l'association entre l'exposition et le nombre de cas.

Une problématique importante que nous n'avons pas abordée dans cette étude de simulation concerne le cas de plusieurs expositions simultanées. Une approche classique consiste à inclure un paramètre pour chaque facteur d'exposition dans le modèle afin d'obtenir des estimateurs de la fraction attribuable pour chacun (Benichou, 2001). Toutefois, l'interprétation des fractions attribuables peut être compli-

quée si le modèle n'est pas additif, car la fraction attribuable globale peut ne pas être égale à la somme des fractions attribuables des différents facteurs d'exposition, sauf si les séries des facteurs d'exposition ne sont jamais positives simultanément (Walter, 1983). En particulier, l'oubli d'un facteur d'exposition risque de biaiser les estimations de la fraction attribuable des autres facteurs considérés ; s'il est fortement corrélé avec ces derniers, leurs fractions attribuables pourraient être surestimées. Notons que, parmi les modèles considérés, la régression de Serfling n'est pas adaptée pour l'étude de plusieurs facteurs d'exposition, puisqu'elle est ajustée en dehors des périodes épidémiques. Pour les autres modèles, il serait pertinent de réaliser une nouvelle étude de simulation qui s'intéresse à au moins deux facteurs d'exposition et d'étudier l'impact de la corrélation de ces facteurs sur l'estimation de la fraction attribuable.

Ce travail a défini un cadre pour l'estimation de la fraction attribuable à partir de données agrégées, et présente des résultats de simulation lorsque l'exposition est épidémique et sous l'hypothèse d'un lien de causalité avec le phénomène saisonnier. Il s'inscrit dans une littérature où la définition de la fraction attribuable à partir de séries temporelles n'est souvent pas explicite et présente également des problèmes d'estimation. En particulier, l'estimation de la variance et des intervalles de confiance de la fraction attribuable que nous proposons offre une alternative aux approximations parfois sommaires (Perrin *et al.*, 2010; Roussel *et al.*, 2013), ou à l'utilisation d'algorithmes de Monte-Carlo plus longs à mettre en place (Deng *et al.*, 2019; Achebak *et al.*, 2019).

Deux applications épidémiologiques de l'estimation de la fraction attribuable aux pathologies hivernales en lien avec la consommation d'antibiotiques sont présentées dans le chapitre qui suit.

Chapitre 2

Applications de la fraction attribuable

2.1. Contexte

L'antibiorésistance est le phénomène qui consiste pour une bactérie à être résistante aux antibiotiques, c'est-à-dire pour laquelle certaines classes d'antibiotiques ne provoquent aucun effet thérapeutique. Elle peut être naturelle, caractéristique d'une espèce bactérienne spécifique (comme par exemple la résistance à l'amoxicilline de *Klebsiella pneumoniae*, ou la résistance à la colistine des bactéries à Gram +). Elle peut également être acquise, par un processus classiquement considéré comme le résultat d'une sélection naturelle, dirigé par l'utilisation d'antibiotiques et propagé par transmission verticale (d'une bactérie mère à ses bactéries filles) mais aussi horizontale (d'une bactérie à une autre), par transfert de gènes *via* des plasmides (Kaplan, 2014; Sun *et al.*, 2019). On rencontre aujourd'hui de nombreuses bactéries résistantes, faisant craindre une augmentation des situations d'impasse thérapeutique (Barbier et Wolff, 2010; Grall *et al.*, 2011).

Il s'agit d'un grave problème de santé publique (French, 2010), qui ne diminue pas en intensité (European Centre for Disease Prevention and Control, 2019), malgré les initiatives internationales (Rex, 2014; World Health Organization, 2015) et nationales [1] destinées à le combattre. En 2015 en Europe, l'antibiorésistance fut responsable de 33 000 décès et 670 000 infections à des bactéries résistantes (Cassini *et al.*, 2019). En France la même année, ce furent 5 500 décès liés à l'antibiorésistance et 125 000 patients développèrent une infection liée à une bactérie résistante (Cassini *et al.*, 2019). L'année suivante, Opatowski *et al.* (2019) rapportent 140 000 infections liées à une bactérie résistante.

Le mésusage, défini comme "une utilisation intentionnelle et inappropriée d'un

1. `https://solidarites-sante.gouv.fr/IMG/pdf/feuille_de_route_antibioresistance_nov_2016.pdf`, dernière visite le 01/09/20.

médicament ou d'un produit, non conforme à l'autorisation de mise sur le marché ou à l'enregistrement, ainsi qu'aux recommandations de bonnes pratiques",[2] et la surconsommation des antibiotiques contribuent au développement et à la diffusion de l'antibiorésistance (Guillemot *et al.*, 1998; van de Sande-Bruinsma *et al.*, 2008; Costelloe *et al.*, 2010; Bell *et al.*, 2014). En médecine humaine, la consommation des antibiotiques provient principalement des soins ambulatoires (>90% en France, ANSM, 2017), majoritairement comme traitement des infections respiratoires aiguës (McCaig, 1995; Alves Galvão *et al.*, 2016). Ils sont souvent prescrits sans qu'un motif clinique ne soit justifié (Dolk *et al.*, 2018; Gulliford *et al.*, 2014; Nadeem Ahmed *et al.*, 2010). En conséquence, une grande part du mésusage et de la surconsommation a lieu au cours des épidémies hivernales d'infections respiratoires.

Si aucun expert ne s'oppose à l'utilisation d'antibiotiques en cas de suspicion d'infection bactérienne, en revanche leur utilisation ne fait pas consensus pour les infections respiratoires virales. Certains experts considèrent que l'utilisation des antibiotiques est toujours inappropriée dans le cas d'infections virales (Fleming-Dutra *et al.*, 2016), s'appuyant sur des recommandations d'experts et des preuves qu'ils ne réduisent ni la durée de la maladie, ni la fréquence des surinfections, ni la mortalité (Alves Galvão *et al.*, 2016; Gonzales *et al.*, 1997; Gulliford *et al.*, 2016). D'autres experts et certaines recommandations suggèrent leur utilisation dans des cas spécifiques (Smith *et al.*, 2018; Pouwels *et al.*, 2018), surtout dans le cas de présentation clinique ambiguë. Par exemple, Smith *et al.* (2018) estiment que les antibiotiques peuvent être utilisés chez 10 à 20% des patients présentant un tableau clinique d'infection respiratoire aiguë sans comorbidité. Dans le cas des infections respiratoires basses, des auteurs ont montré que jusqu'à 50% des prescriptions antibiotiques étaient inappropriées, et souvent associées aux patients adultes (18-65 ans) pour lesquels les médecins ont ressenti qu'ils désiraient une prescription d'antibiotique (Dekker *et al.*, 2015; Dolk *et al.*, 2018; McKay *et al.*, 2016).

La majorité des estimations de la fraction des prescriptions évitables au cours des épidémies d'infections respiratoires aiguës proviennent d'enquêtes portant sur des groupes d'âge et des conditions spécifiques (par exemple, infections de l'oreille ou bronchite aiguë), souvent à partir de base d'enregistrements médicaux électroniques. Par exemple, Silverman *et al.* (2017) ont montré que, dans l'Ontario (Canada), chez

2. `https://solidarites-sante.gouv.fr/soins-et-maladies/medicaments/glossaire/article/mesusage`, dernière visite le 10/08/2020.

les personnes âgées (> 65 ans) sans comorbidité grave, 46% des personnes présentant une infection respiratoire haute aiguë d'étiologie supposée non bactérienne se sont vus prescrire des antibiotiques après avoir consulté leur médecin généraliste. Une étude de Fleming-Dutra *et al.* (2016) portant sur le mésusage des antibiotiques aux Etats-Unis a montré, à partir de bases de données échantillonnant les visites aux médecins libéraux, aux services d'urgence, aux services ambulatoires et de court-séjour des hôpitaux, que les infections respiratoires aiguës contribuent à hauteur de 110 prescriptions inappropriées pour 1 000 habitants par an, soit plus de 20% des prescriptions totales.

Au début des années 2000, on observait en France une augmentation des pneumocoques résistants à la pénicilline comme cause bactérienne des infections invasives d'origine communautaire et des pneumonies (Schuchat *et al.*, 1997; Marston *et al.*, 1997). Plusieurs études démontraient que la consommation d'antibiotiques étaient un facteur clef dans le taux de pneumocoques résistants aux bêta-lactamines et dans la diffusion des pneumocoques résistants à la pénicilline (Lipsitch, 2001; Harris *et al.*, 2002; Samore *et al.*, 2006; van de Sande-Bruinsma *et al.*, 2008). En outre, la France était le pays européen le plus consommateur d'antibiotiques, avec 32,2 doses quotidiennes pour 1 000 habitants (Goossens *et al.*, 2005). Les jeunes enfants ($\leq$ 5 ans) et les personnes âgées ($\geq$ 75 ans) étaient les classes d'âge les plus exposées à l'utilisation d'antibiotiques, et représentaient plus de 40% des prescriptions (Akkerman *et al.*, 2004; Sharland, 2007; Sommet *et al.*, 2004).

Pour lutter contre les pneumocoques résistants à la pénicilline, et plus généralement contre le mésusage et la surconsommation, le gouvernement français a lancé en 2001 un plan national visant préserver l'efficacité des antibiotiques (Carlet et Le Coz, 2015). En 2002, dans le cadre de ce plan, la caisse nationale de l'assurance maladie a mis en place une vaste campagne de sensibilisation autour du slogan "Les antibiotiques, c'est pas automatique !", pour amener les usagers du système de santé (le personnel de santé mais aussi les patients) à remettre en question la prescription systématique d'antibiotiques, en particulier au cours des épidémies saisonnières d'infections respiratoires virales. Cette campagne, reconduite chaque hiver jusqu'en 2005, fut un véritable succès, avec une diminution de la consommation d'antibiotiques de 25% pendant la période hivernale au bout de 5 ans (Sabuncu *et al.*, 2009).

Une autre campagne de communication, qui s'inscrit dans un deuxième plan national de 2007 à 2010, a permis de poursuivre les actions commencées par le plan

2001-2005. Intitulée “Les antibiotiques, utilisés à tort, ils deviendront moins forts”, cette campagne lancée en 2009, a davantage porté sur l'explication du phénomène de résistance. Elle a rencontré moins de succès que la précédente campagne, mais a néanmoins permis de stabiliser la diminution de la consommation d'antibiotiques (Trinh *et al.*, 2018), qui s'accompagne toutefois d'un remplacement thérapeutique : le nombre de prescriptions de macrolides a continué de baisser, tandis que le nombre de prescriptions de pénicillines est reparti à la hausse (Bernier *et al.*, 2014; Watier *et al.*, 2017). Au final, entre 2002 et 2012, ces dix années de communication ont tout de même permis d'éviter plus de 40 millions de prescriptions (Carlet et Le Coz, 2015). En particulier, le nombre de prescriptions associées aux infections respiratoires basses a diminué de 39,9% entre entre 2001 et 2009 (Chahwakilian *et al.*, 2011).

Le plan national d'alerte sur les antibiotiques qui a suivi, de 2011 à 2016, s'inscrit dans la continuité des actions précédentes, tout en s'attachant à mieux encadrer la dispensation des antibiotiques, à œuvrer en plus étroite coordination avec le monde de la santé animale et à promouvoir une utilisation plus adaptée des antibiotiques (Ministère chargé de la Santé, 2011).

Enfin, en 2016, en accord avec la démarche “One Health” recommandée par l'Organisation Mondiale de la Santé, une action coordonnée visant à maitriser l'antibiorésistance a été lancée, s'appuyant sur une feuille de route interministérielle construite par un groupe de travail spécial. [3] Les recommandations de ce groupe se focalisent sur quatre objectifs majeurs : favoriser et approfondir les recherches en matière d'antibiorésistance ; renforcer la surveillance à travers des indicateurs partagés entre les secteurs d'activité ; améliorer l'usage des antibiotiques ; et accroitre la sensibilisation des populations au risque de l'antibiorésistance et au bon usage des antibiotiques.

En 2017, la France était le troisième pays le plus consommateur d'antibiotiques en Europe, avec 31,3 doses quotidiennes pour 1 000 habitants, dont 90% provenant des soins ambulatoires (ANSM, 2017). Les infections respiratoires étaient à l'origine de deux prescriptions d'antibiotiques sur trois, et quatre prescriptions sur cinq appartenaient aux classes thérapeutiques des bêta-lactamines ou des macrolides (ANSM, 2017).

3. `https://solidarites-sante.gouv.fr/IMG/pdf/feuille_de_route_antibioresistance_nov_2016.pdf`, dernière visite le 01/09/20.

Nous considérons ici comme mésusage les prescriptions d'antibiotiques potentiellement inappropriées, c'est-à-dire celles qui ne sont pas justifiées cliniquement (à la différence d'une mauvaise dose ou d'une mauvaise durée de traitement). Réduire le mésusage est essentiel pour diminuer à la fois la résistance aux antibiotiques et les événements indésirables. Cependant, la fraction d'antibiotiques inappropriée et susceptible de réduction est inconnue (Metlay, 2015). Il est important de préciser les causes du mésusage et d'estimer la part des prescriptions qui leur sont attribuables pour guider les politiques de santé publique visant à combattre la dissémination de l'antibiorésistance. Comme une part importante du mésusage correspond à des prescriptions pour des infections respiratoires virales aiguës, l'estimation de la fraction attribuable des prescriptions antibiotiques associées à ces infections donnerait un indicateur du mésusage.

2.2. Objectifs

Nous étudions l'impact des maladies infectieuses hivernales sur la consommation des antibiotiques dans la communauté à travers deux études de cas. La première s'intéresse à l'impact de la campagne de prévention "Les antibiotiques, c'est pas automatique !" sur la diminution du mésusage attribuable aux épidémies de grippes saisonnières et compare les différents estimateurs établis dans la Section 1.3. Sur la période 2010-2017, nous étudions ensuite plus largement les contributions des infections des voies respiratoires inférieures à l'usage des antibiotiques pendant la période hivernale, et détaillons ces fractions attribuables pour les deux classes d'âge les plus exposées aux antibiotiques : les enfants et les personnes âgées.

2.3. Sources de données

Caisse Nationale de l'Assurance Maladie (CNAM)

En 2002, une collaboration de l'Institut Pasteur avec la CNAM et le Régime Social des Indépendants (RSI) a permis de collecter, de juillet 2000 à juin 2010, toutes les données de remboursement des assurés du Régime Général, de la Section Locale Mutualiste et du RSI pour les antibiotiques systémiques (classe J01 de la classification anatomique, thérapeutique et chimique). En 2013, une base de rembour-

sements de médicaments thérapeutiques ou préventifs des infections bactériennes communautaires (MIBAC) a été construite pour prolonger le suivi au-delà de 2010, d'enrichir la base existante par les données de remboursements de la Mutualité Sociale Agricole, et de chainer les individus (Commission Nationale de l'Informatique et des Libertés - CNIL, accord DR-2018-311). Le taux de couverture de cette base est de plus de 95% de la population métropolitaine française.

La population de la base concerne tous les bénéficiaires ayant reçu au moins un remboursement, sur la période étudiée, pour une prescription d'antibiotiques (classe J01 de la classification ATC). À chaque remboursement est associé, entre autres, la date de prescription et de soins, le code identifiant de présentation, l'année de naissance, le sexe, le département. Ainsi, chaque année, on dénombre environ 60 millions de remboursements de prescriptions antibiotiques pour 30 millions de bénéficiaires.

Insee

L'**Insee** (pour *Institut national de la statistique et des études économiques*) est une direction générale du ministère de l'Économie et des Finances. Il a pour mission de collecter, analyser et diffuser des informations sur l'économie et la société française sur l'ensemble de son territoire. Il est en particulier responsable du recensement de la population, afin de connaître la diversité et l'évolution de la population de la France.

Depuis 2004, le recensement repose sur une collecte d'information annuelle, concernant successivement tous les territoires communaux au cours d'une période de cinq ans. Les communes de moins de 10 000 habitants réalisent une enquête de recensement portant sur toute la population, à raison d'une commune sur cinq chaque année. Les communes de 10 000 habitants ou plus, réalisent tous les ans une enquête par sondage auprès d'un échantillon d'adresses représentant 8% de leurs logements.

Les données démographiques au 1er janvier de chaque année pour tous les territoires administratifs de France sont disponibles en ligne. [4] L'Insee fournit ainsi des statistiques sur les habitants et les logements, leur nombre et leurs caractéristiques : répartition par sexe et âge, professions, conditions de logement, etc.

4. `https://www.insee.fr/`, dernière visite le 16/08/2020.

Réseau Sentinelles

Le **réseau Sentinelles** (Valleron *et al.*, 1986) est un réseau de recherche et de veille en soins de premiers recours (médecine générale et pédiatrie) en France métropolitaine. [5] Il a pour mission, entre autres, de constituer des bases de données en médecine générale et en pédiatrie, à des fins de veille sanitaire, de prévision épidémique et de recherche. Aujourd'hui, il est coordonné par l'équipe "Surveillance et Modélisation des maladies transmissibles" de l'Institut Pierre Louis d'Épidémiologie et de Santé Publique, de l'Inserm et de Sorbonne Université, en collaboration avec l'agence nationale de Santé publique, Santé publique France.

Ce réseau est composé de médecins généralistes libéraux (1 314 au 1er janvier 2018, soit 2,1% des médecins généralistes libéraux) et pédiatres libéraux (116, soit 4,3% des pédiatres libéraux), volontaires, répartis sur le territoire métropolitain français. Chaque semaine, ces médecins transmettent les données de leurs patients vus en consultation pour les indicateurs suivis *via* une connexion Internet sécurisée. À partir de ces données, le réseau estime le taux d'incidence hebdomadaire pour chaque indicateur, permettant de suivre son évolution dans le temps et dans l'espace. En 2018, le réseau Sentinelles surveillait de façon continue des informations sur dix indicateurs de santé (dont neuf maladies infectieuses, comme les diarrhées aiguës ou les oreillons).

Réseau Oscour®

Santé publique France, l'agence nationale de santé publique française, est chargée de surveiller le bien-être de la population et d'alerter sur les menaces épidémiologiques, à travers plusieurs systèmes de surveillance, des enquêtes épidémiologiques ou comportementales et des bases de données médico-administratives. **Oscour®** (pour *organisation de la surveillance coordonnée des urgences*) est l'un de ces systèmes, créé en 2004, basé sur un réseau de services d'urgence hospitaliers qui collecte quotidiennement les données individuelles des patients pour la surveillance syndromique. En août 2014, environ 600 services d'urgence étaient inclus dans le réseau, représentant 80% de la fréquentation des services d'urgence nationaux. Plusieurs rapports détaillent ce réseau, ainsi que l'évaluation du système de surveillance syndromique (Josseran *et al.*, 2010; Fouillet *et al.*, 2015; Pelat *et al.*, 2017).

5. `https://www.sentiweb.fr/`, dernière visite le 15/07/2020.

Pour chaque visite dans un service d'urgence du réseau, les variables démographiques (âge, sexe), administratives (région de résidence) et médicales du patient sont enregistrées, ainsi que les diagnostics médicaux principaux et associés, encodés selon la dixième édition de la classification internationale des maladies (CIM-10). À l'aide des codes CIM-10, des indicateurs pour différents groupes syndromiques sont construits, avec des détails tels que, notamment, l'étiologie bactérienne ou virale présumée de l'infection.

Le réseau Oscour® collecte de façon continue les informations sur un certain nombre d'indicateurs de santé (décès, maladies infectieuses, événements sanitaires, etc.) (Bousquet *et al.*, 2013). L'exhaustivité du codage des diagnostics dans les services, de l'ordre de 60% en 2004, s'est amélioré et est stable depuis 2009, avec près de 80% des visites pour lesquelles le diagnostic est renseigné (Caillère *et al.*, 2011). En particulier, il existe une très bonne concordance entre les données du surveillance des syndromes grippaux issues du réseau Oscour® et du réseau Sentinelles (Josseran *et al.*, 2006), qui est le système de surveillance de référence pour la grippe en France (Flahault *et al.*, 1997).

2.4. Prescriptions antibiotiques attribuables aux syndromes grippaux

Nous nous intéressons à déterminer le nombre de prescriptions antibiotiques associées aux syndromes grippaux, sur la période 2002-2010 qui correspond à la campagne de sensibilisation à l'utilisation des antibiotiques, "Les antibiotiques, c'est pas automatique !", reconduite chaque hiver de 2002 à 2005, et aux cinq années qui l'ont succédée. Cette campagne a été un succès, puisqu'il a été montré que le nombre de prescriptions d'antibiotiques a diminué d'un quart entre 2002 et 2007 (Sabuncu *et al.*, 2009). Cette réduction est en particulier en faveur d'un mésusage avant la campagne : si les prescriptions ont pu être diminuées, c'est qu'elles étaient inappropriées ou tout du moins évitables.

Dans ce cadre, nous étudions ici plus précisément la contribution des épidémies saisonnières de syndromes grippaux à l'usage des antibiotiques et calculons la fraction des prescriptions qui leur sont attribuables. Ce cadre d'application a aussi été considéré pour estimer et comparer, dans un cas réel, l'ensemble des estimateurs développés en Section 1.3.

2.4.1. Matériel

Pour la période de juillet 2002 à juin 2010, les données sont issues pour les prescriptions antibiotiques de la CNAM et du RSI, et pour les syndromes grippaux du réseau Sentinelles. Les classes d'antibiotiques retenues sont les bêta-lactamines et macrolides (classe chimique anatomique thérapeutique J01C, J01D, J01F), car ce sont les classes les plus représentées pour combattre les syndromes grippaux. À partir des données individuelles d'antibiotiques prescrits et remboursés aux patients en ville, et des données de l'Insee, le taux hebdomadaire de remboursements d'antibiotiques pour 100 000 habitants a été construit.

L'incidence des cas de grippe a été approchée par celle des syndromes grippaux. Les syndromes grippaux sont définis par le réseau Sentinelles comme la combinaison d'une fièvre supérieure à 39°C, d'apparition brutale, accompagnée de myalgies et de signes respiratoires. Cet indicateur est suivi depuis la création du réseau en 1984, et sa surveillance a été terminée en 2020, suite au passage de la pandémie de Covid-19 au stade 3 en France, et remplacée par la surveillance plus large des infections respiratoires aiguës. Les caractéristiques des différentes épidémies saisonnières de grippe, dont les périodes épidémiques nécessaires pour le calcul des fractions attribuables, ont été déterminées par le réseau Sentinelles et sont présentées en Annexe 1.B.

2.4.2. Méthodes

Le nombre de prescriptions d'antibiotiques est modélisé en fonction de l'incidence des syndromes grippaux. Six modèles ont été utilisés : la régression linéaire avec erreurs autocorrélées ("a-Gaussien"), les régressions de Poisson et négative-binomiale avec fonctions de lien identité ("a- Poisson" et "a-NegBin") et logarithme ("m-Poisson" et "m-NegBin") et la régression de Serfling. Un terme périodique, identique à celui introduit à la relation (1.7), est inclus dans la régression pour tenir compte des saisonnalités de la série.

Pour prendre en compte d'éventuelles variabilités annuelles dans l'association entre les épidémies de grippe et le taux de remboursements d'antibiotiques, le paramètre β_k associé à la série d'incidence des syndromes grippaux est différent pour chaque épidémie saisonnière k. Le prédicteur linéaire des modèles (sauf de Serfling) s'écrit :

$$g\Big(\mu(t)\Big) = \eta(t) + \sum_{k=1}^{8} \beta_k x_k(t), \tag{2.1}$$

où g désigne la fonction de lien choisie, $\{x_k(t)\}$ la série d'incidence des syndromes grippaux de la saison k.

Pour déterminer s'il existe un décalage dans l'association entre la série d'incidence des syndromes grippaux et celles du taux de remboursements d'antibiotiques, plusieurs décalages temporels ont été testés et comparés à partir du BIC (Bayesian Information Criterion). À partir des estimations des modèles, les fractions de prescriptions antibiotiques attribuables aux épidémies grippales et leurs intervalles de confiance à 95% ont été estimés (Section 1.3).

2.4.3. Résultats

Les taux hebdomadaires de remboursements d'antibiotiques et d'incidence de syndromes grippaux pour 100 000 personnes sont présentés Figure 2.1. La série du nombre de remboursements antibiotiques présente une saisonnalité marquée, atteint généralement son maximum en décembre ou en janvier, et son minimum au mois d'août. La série d'incidence des syndromes grippaux présente des profils épidémiques avec des pics épidémiques qui se produisent entre octobre et mars. Visuellement, on observe une forte corrélation positive entre les pics épidémiques des syndromes grippaux et les maxima du taux de remboursements antibiotiques.

L'analyse préalable des données n'a pas montré de décalage entre les deux séries étudiées (résultats non montrés). Les modèles retenus n'incluent donc pas de décalage temporel. Les estimations des paramètres et de la fraction attribuable pour chacun des modèles sont données dans les Tableaux 2.1 et 2.2. Les fractions attribuables et leurs intervalles de confiance à 95% sont également représentées Figure 2.2.

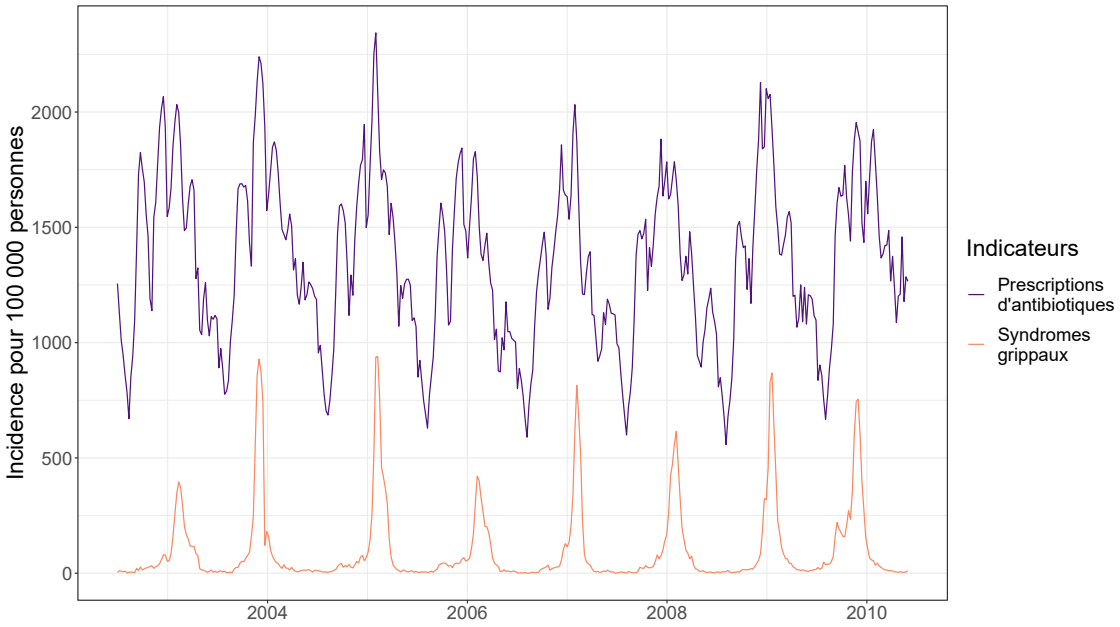


FIGURE 2.1 – Taux de remboursements d'antibiotiques (bêta-lactamines et macrolides) et d'incidence des syndromes grippaux, par semaine pour 100 000, en France métropolitaine.

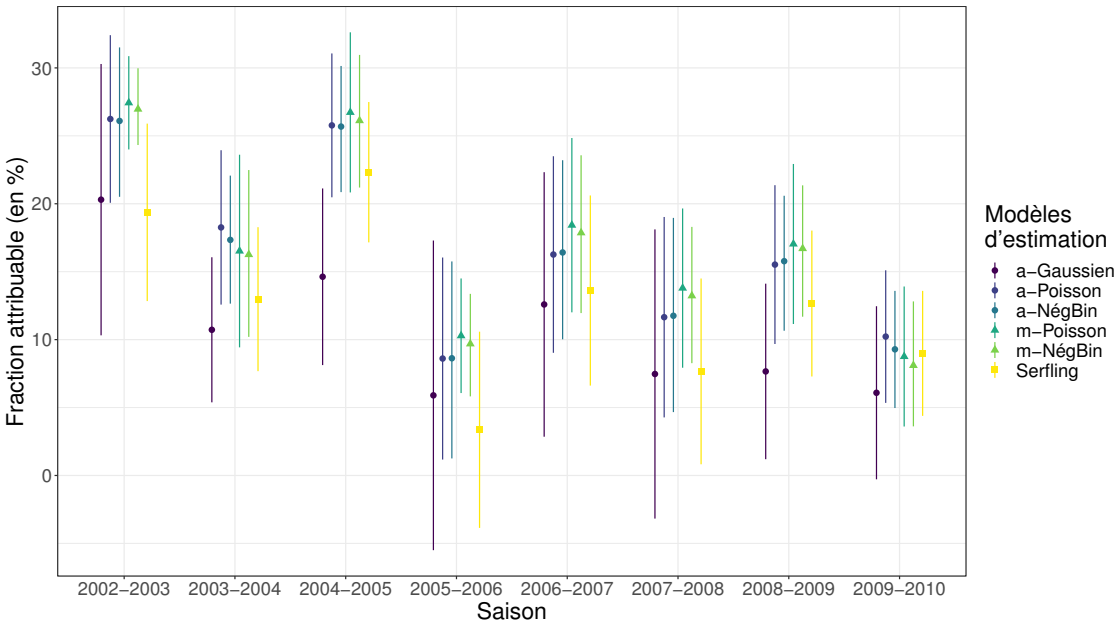


FIGURE 2.2 – Estimations et intervalles de confiance à 95% de la fraction des remboursements d'antibiotiques attribuables à chaque épidémie saisonnière de grippe de 2002 à 2010.

TABLEAU 2.1 – Estimation (et écarts-type) des paramètres pour chacun des six modèles.

	Additifs : **a-**			Multiplicatifs† : **m-**		
Paramètres	**Gaussien**	**Poisson**	**NégBin**	**Poisson**	**NégBin**	**Serfling**
μ	1 320 (30)	1 285 (10)	1 286 (10)	7 138 (8)	7 135 (8)	1 315 (9)
γ_1	-353 (39)	-336 (14)	-342 (14)	-258 (11)	-265 (11)	-359 (13)
γ_2	-10 (34)	-14 (11)	-11 (11)	-34 (9)	-34 (9)	-14 (12)
γ_4	98 (25)	107 (11)	111 (10)	80 (9)	87 (9)	108 (12)
δ_1	4 (39)	2 (12)	-5 (12)	1 (10)	-7 (9)	-12 (12)
δ_2	-80 (35)	-103 (12)	-112 (11)	-76 (10)	-90 (10)	-85 (13)
δ_4	-84 (25)	-102 (11)	-110 (10)	-72 (9)	-84 (9)	-121 (12)
$\beta_{2002-03}$	1,24 (0,31)	1,63 (0,23)	1,64 (0,26)	1,07 (0,14)	1,09 (0,16)	
$\beta_{2003-04}$	0,40 (0,10)	0,67 (0,11)	0,72 (0,14)	0,33 (0,06)	0,33 (0,07)	
$\beta_{2004-05}$	0,54 (0,12)	0,94 (0,11)	0,95 (0,13)	0,59 (0,06)	0,60 (0,08)	
$\beta_{2005-06}$	0,31 (0,30)	0,44 (0,20)	0,44 (0,21)	0,35 (0,15)	0,37 (0,15)	
$\beta_{2006-07}$	0,43 (0,17)	0,55 (0,13)	0,54 (0,14)	0,40 (0,09)	0,41 (0,10)	
$\beta_{2007-08}$	0,29 (0,21)	0,44 (0,15)	0,43 (0,16)	0,34 (0,10)	0,35 (0,11)	
$\beta_{2008-09}$	0,31 (0,13)	0,64 (0,13)	0,63 (0,14)	0,41 (0,07)	0,42 (0,08)	
$\beta_{2009-10}$	0,29 (0,16)	0,45 (0,12)	0,51 (0,14)	0,24 (0,07)	0,26 (0,08)	
ϕ_1	0,68 (0,04)					
φ_1	0,53 (0,05)					
σ^2	9 573					24 586

† $\times 10^{-3}$.

La comparaison des paramètres d'association montre une très grande différence entre l'épidémie de grippe de 2002-03 et celles qui ont suivi, mais des différences plus faibles entre les épidémies de 2003 à 2010. Par exemple, pour la régression linéaire avec erreurs autocorrélées, la différence entre les épidémies de grippe de 2002-03 et de 2003-04 est statistiquement significative (p-value = 0.005) tandis qu'il n'existe pas de différence significative entre les épidémies après 2003 (p-value > 0,11).

Pour tous les modèles considérés, on observe une diminution de la fraction attribuable entre les épidémies de 2002-03 et de 2003-04, suivie d'un retour au niveau initial pour l'épidémie de 2004-05 sauf pour la régression linéaire avec erreurs autocorrélées. L'année suivante, la fraction attribuable a atteint son minimum sur la période considérée (à l'exception des modèles multiplicatifs). Entre l'épidémie de

TABLEAU 2.2 – Estimation (et écarts-type) des fractions attribuables pour chacun des six modèles.

	Additifs : **a-**			Multiplicatifs : **m-**		
Paramètres	**Gaussien**	**Poisson**	**NégBin**	**Poisson**	**NégBin**	**Serfling**
$FA_{2002-03}$	20.3 (5.1)	26.1 (2.8)	26.2 (3.1)	27.0 (1.6)	27.4 (1.8)	19.4 (3.3)
$FA_{2003-04}$	10.7 (2.7)	17.3 (2.4)	18.3 (2.9)	16.3 (3.1)	16.5 (3.6)	13.0 (2.7)
$FA_{2004-05}$	14.6 (3.3)	25.7 (2.3)	25.8 (2.7)	26.1 (2.7)	26.7 (3.0)	22.3 (2.6)
$FA_{2005-06}$	5.9 (5.8)	8.6 (3.6)	8.6 (3.8)	9.7 (2.0)	10.3 (2.2)	3.4 (3.7)
$FA_{2006-07}$	12.6 (5.0)	16.4 (3.4)	16.3 (3.7)	17.9 (2.9)	18.4 (3.3)	13.6 (3.6)
$FA_{2007-08}$	7.5 (5.4)	11.8 (3.5)	11.7 (3.8)	13.2 (2.7)	13.8 (3.0)	7.7 (3.5)
$FA_{2008-09}$	7.7 (3.3)	15.8 (2.7)	15.5 (3.0)	16.7 (2.6)	17.0 (3.0)	12.7 (2.7)
$FA_{2009-10}$	6.1 (3.3)	9.3 (2.3)	10.2 (2.5)	8.1 (2.3)	8.8 (2.6)	9.0 (2.3)

2002-03 et celle de 2005-06, la fraction attribuable a diminué de 16 points de pourcentage, soit presque de deux tiers (cinq sixièmes pour la régression de Serfling). Ensuite, les fractions attribuables se stabilisent à un niveau un peu plus élevé jusqu'à la fin de la période.

Alors que les estimations issues des régressions de Poisson et négatives-binomiales (".-Poisson" et ".-NegBin") sont similaires, celles estimées par des régressions linéaires avec erreurs autocorrélées ("a-Gaussien") et de Serfling sont systématiquement inférieures (sauf pour la saison 2009-10 pour le modèle de Serfling). Pour la régression de Serfling, ceci est en accord avec l'étude de simulation, qui indiquent des estimations entre 10% et 20% inférieures en moyenne. Pour la régression linéaire, cette différence s'explique par le fait que le modèle intègre une structure d'autocorrélation saisonnière (paramètre autorégressif d'ordre 52), qui tend à propager les valeurs élevées de la série et donc à diminuer les valeurs des paramètres d'association.

2.4.4. Discussion

Dans ce travail, nous avons estimé la part des prescriptions antibiotiques attribuables aux syndromes grippaux au cours des épidémies hivernales, entre juillet 2002 et juin 2010, en France métropolitaine. Nous avons considéré six modèles pour l'estimation de la fraction attribuable : la régression linéaire avec erreurs autocorré-

lées, les régressions de Poisson et négative-binomiale avec fonctions de lien identité et logarithme, et la régression de Serfling. Tous les modèles montrent une diminution significative de la fraction attribuable sur la période 2002-2006, d'environ 16 points de pourcentage, soit près d'un tiers, suivie d'une stabilisation jusqu'en 2010. Cette diminution coïncide avec la mise en place de la campagne de sensibilisation "Les antibiotiques, c'est pas automatique !".

Il existe peu de différences entre les estimations issues des régressions de Poisson et négatives-binomiales, additives et multiplicatives, tandis que celles issues de la régression de Serfling sont presque toujours inférieures. Ce constat est en accord avec les résultats de l'étude de simulation du Chapitre 1 pour des valeurs de la fraction attribuable peu élevées ($\leq 20\%$). En revanche, les estimations issues de la régression linéaire avec erreurs autocorrélées sont différentes des autres modèles additifs, suggérant qu'il existe une structure d'autocorrélation avec un fort impact sur l'estimation de la fraction attribuable.

Bien que les épidémies de grippe ne soient pas identiques chaque saison, et que certaines des valeurs de la fraction attribuable puissent être expliquées par des épidémies peu contagieuses — par exemple, l'épidémie de l'hiver 2005-2006 a infecté deux fois moins de personnes que celle de l'hiver précédent sur une période similaire de l'année (Tableau 1.B.1 en Annexe 1.B.1) —, cette différence seule ne peut expliquer les variations de fractions attribuables. En effet, l'épidémie de 2002-2003 est comparable avec celle de 2005-2006, alors que la fraction attribuable de la seconde est quatre fois inférieure à la première. Cette diminution pourrait donc être interprétée comme une diminution du mésusage associé aux épidémies grippales, même s'il est difficile de déterminer avec certitude si ces prescriptions étaient inappropriées sans les données individuelles. On pourrait l'expliquer soit par la diminution du nombre de prescriptions des médecins, soit par la diminution de la pression imposée aux médecins par les patients pour la prescription d'antibiotique, retrouvée comme facteur de risque de sur-prescription (Dekker *et al.*, 2015).

Ainsi, les estimations de la fraction attribuable obtenues soutiennent les effets positifs de la campagne "Les antibiotiques, c'est pas automatique !", dirigée vers les patients et le personnel de santé, puisqu'elle a conduit à une diminution des prescriptions d'antibiotiques associées aux symptômes grippaux. Ces résultats sont d'ailleurs comparables aux travaux de Sabuncu *et al.* (2009), qui ont estimé une diminution de presque moitié du paramètre d'association après la mise en place de

la campagne, comparable avec la diminution progressive que nous avons observée. Ces résultats concordent également avec l'enquête rétrospective de Chahwakilian *et al.* (2011) qui rapportent une diminution de la proportion des consultations pour syndrome grippal résultant en une prescription d'antibiotique, d'environ 30% en 2002 à presque 10% sur la période 2006-2009. Par ailleurs, la stabilisation de la fraction attribuable entre les années 2006 et 2010 est étayée par plusieurs autres publications (Bernier *et al.*, 2014; Carlet et Le Coz, 2015).

Notons toutefois que les intervalles de confiance estimés sont vraisemblablement trop petits, car ils ne tiennent pas compte de l'incertitude sur la série d'incidence des syndromes grippaux. En effet, puisque cette série est construite à partir d'un échantillon de médecins généralistes et pédiatres libéraux représentant moins de 5% des praticiens libéraux (voir Section 2.3), le réseau Sentinelles fournit les intervalles de confiance de l'incidence des syndromes grippaux. Il faudrait les prendre en compte pour corriger les intervalles de confiance des fraction attribuables.

À partir des données agrégées de bases médico-administratives, nous sommes parvenus à estimer la fraction des remboursements de prescriptions antibiotiques attribuables aux différentes épidémies de syndromes grippaux, et à quantifier la diminution du mésusage associé aux syndromes grippaux. Toutefois, cette analyse statistique n'a pas pris en compte d'autres pathologies infectieuses hivernales, ce qui peut engendrer des effets de confusion : il est possible que les fractions attribuables aient été sur-estimées. Par la suite, nous considérons également les infections des voies respiratoires basses comme facteurs d'exposition pour le calcul de la fraction attribuable.

2.5. Prescriptions antibiotiques attribuables aux infections respiratoires basses

Nous nous intéressons à déterminer le nombre de prescriptions antibiotiques associées aux différentes infections respiratoires basses virales, sur la période 2010-2017. Nous détaillons en particulier la contribution de ces infections pour les deux classes d'âge les plus exposées aux antibiotiques : les enfants et les personnes âgées. En regard des résultats de l'analyse précédente et de la présence d'autocorrélation dans les séries de taux de remboursements d'antibiotiques, seule la régression linéaire avec erreurs autocorrélées a été considérée.

2.5.1. Matériel

Pour la période de janvier 2010 à décembre 2017, les données sont issues pour les prescriptions antibiotiques de la base de données MIBAC et pour les infections respiratoires basses du réseau Oscour®. Les classes d'antibiotiques retenues sont les bêta-lactamines et les macrolides (classe chimique anatomique thérapeutique J01C, J01D, J01F), car ce sont les classes les plus représentées pour combattre les infections respiratoires basses. À partir des données individuelles d'antibiotiques prescrits et remboursés aux patients en ville, et des données de l'Insee, le taux hebdomadaire de remboursements d'antibiotiques pour 100 000 habitants a été construit.

Pour l'incidence des infections respiratoires basses, cinq groupes syndromiques ont été retenus dans le cadre de ce travail. Il s'agit de quatre groupes spécifiques : bronchiolite, bronchite, syndrome grippal et pneumonie ; et d'un groupe global, les infections respiratoires basses, qui comprend les quatre premiers groupes, plus quelques autres infections respiratoires (détails dans l'Annexe 2.A). Pour chacun des groupes, deux séries chronologiques, selon l'étiologie microbienne (bactérienne ou virale), ont été construites par le réseau Oscour® en agrégeant le nombre de visites associées chaque semaine. Comme l'inclusion des services d'urgence dans le réseau Oscour® a continué d'augmenter pendant la période d'étude, le nombre de visites par groupe syndromique a été rapporté au nombre total de visites toute cause confondue, et exprimé comme proportion.

D'après les codes CIM-10 déclarés dans le système de surveillance Oscour®, 99% des bronchites reportées au cours des visites aux services d'urgence du réseau entre janvier 2010 et décembre 2017 étaient d'étiologie virale, tandis que 98% des pneumopathies reportées durant cette période étaient d'étiologie non virales. Par conséquent, nous n'avons pas utilisé les fréquences de diagnostic des bronchites non virales, ni des pneumopathies virales, dans ce qui suit.

Trois classes d'âges ont été considérées : la population générale (toutes classes d'âges confondues), les enfants de 5 ans et moins, et les personnes âgées de 75 ans et plus. Pour chaque classe d'âge, le taux de remboursements hebdomadaire pour 100 000 personnes, et la proportion de visites associées à chaque groupe syndromique sont représentés Figure 2.3.

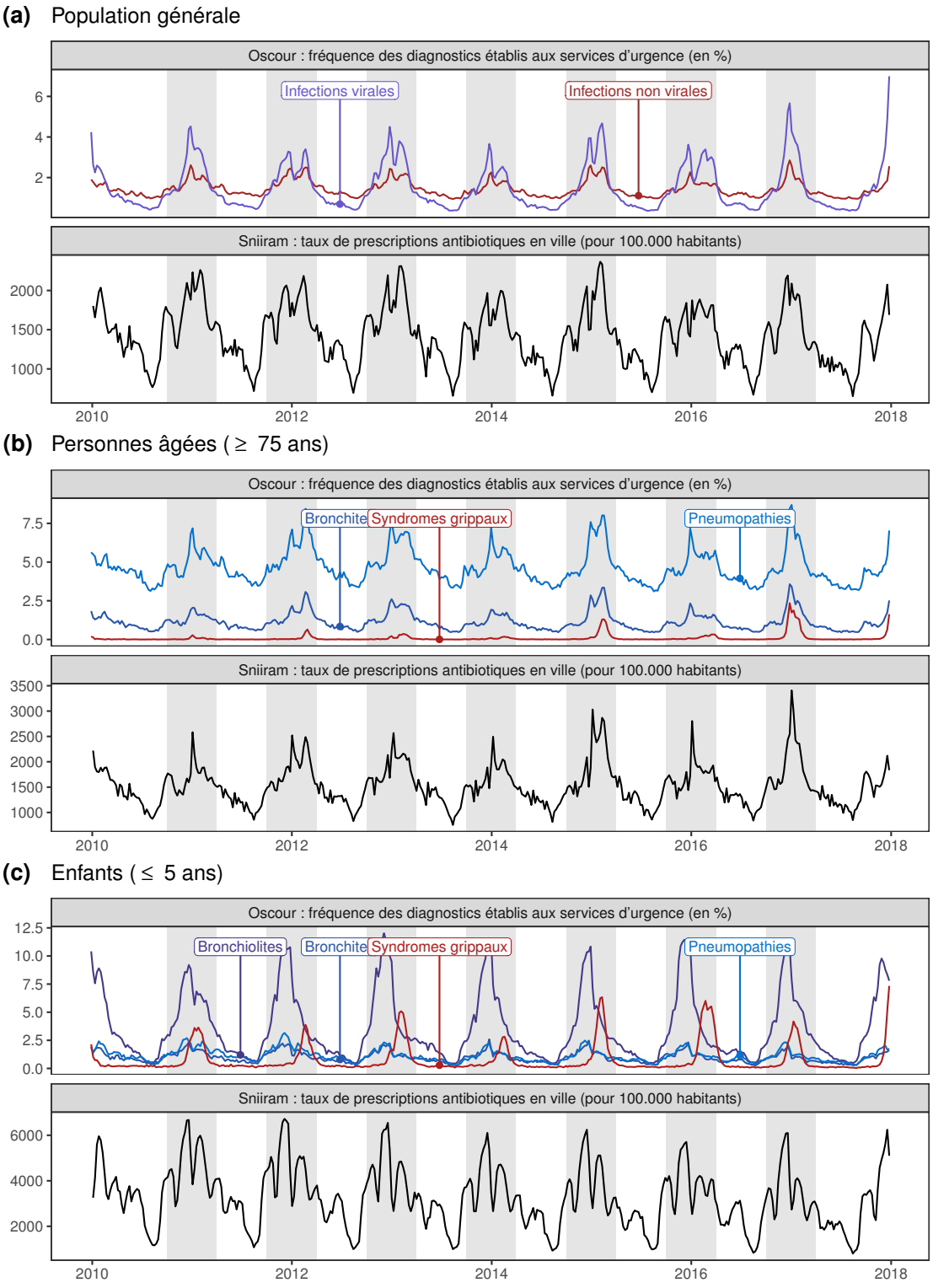


FIGURE 2.3 – Evolution hebdomadaire du taux de remboursements d'antibiotiques (bêta-lactamines et macrolides) pour 100 000 en ville (bas de chaque sous-figure) et de la proportion de visites aux services d'urgence associées à un diagnostic d'infection respiratoire basse (haut), en France métropolitaine. Les zones grises correspondent aux périodes hivernales (d'octobre à mars inclus). (a) À l'échelle de la population générale ; (b) chez les personnes âgées de 75 ans et plus ; (c) chez les enfants de 5 ans et moins.

2.5.2. Méthodes

Pour décrire les séries de taux de remboursements d'antibiotiques et d'incidence des infections respiratoires basses, nous avons estimé les moyennes des séries lors des périodes hivernales. Afin de tenir compte de la présence d'autocorrélation dans les séries, les écart-types des moyennes ont été ajustés à l'aide des estimateurs de Newey-West (aussi parfois appelés estimateurs sandwich). Les estimateurs de Newey-West peuvent être calculés à partir de la librairie `sandwich` (Zeileis, 2004) de `R` (R Core Team, 2020).

Dans un premier temps, pour identifier les contributions globales des infections respiratoires basses à l'usage des antibiotiques, nous avons modélisé le nombre de prescriptions en fonction du groupe syndromique global uniquement, en distinguant toutefois les infections virales des non virales. Le modèle choisi, noté M1, est une régression linéaire avec erreurs autocorrélées, nécessaire pour tenir compte de la présence d'autocorrélation dans l'évolution du nombre de prescriptions antibiotiques. Ce modèle s'écrit de la forme suivante :

$$\begin{aligned}[\#\text{Tx prescriptions}]_t = \mu &+ \beta_{\text{Inf. non-virales}} \times [\text{Infections non virales}]_t \\ &+ \beta_{\text{Inf. virales}} \times [\text{Infections virales}]_t \\ &+ \nu(t),\end{aligned}$$

où μ désigne la moyenne de la série en l'absence d'exposition, et $\{\nu(t)\}$ est un processus ARMA [6] (Shumway et Stoffer, 2011). La variance des erreurs du processus ARMA est dénotée σ^2.

Dans un second temps, afin de pouvoir comparer les contributions des différentes infections respiratoires, un deuxième modèle est ajusté, en considérant cette fois les groupes syndromiques restreints (pneumopathies, bronchiolites, bronchites et syndromes grippaux), et uniquement les enfants de 5 ans et moins, et les personnes

6. AutoRegressive Moving Average model.

âgées de 75 et plus. Ce modèle, noté M2, s'écrit :

$$\begin{aligned}[\#\text{Tx prescriptions}]_t = \mu &+ \beta_{\text{Pneumopathies}} \times [\text{Pneumopathies}]_t \\ &+ \beta_{\text{Bronchiolites}} \times [\text{Bronchiolites}]_t \\ &+ \beta_{\text{Bronchites}} \times [\text{Bronchites}]_t \\ &+ \beta_{\text{Synd. grippaux}} \times [\text{Syndromes grippaux}]_t \\ &+ \nu(t),\end{aligned}$$

où le terme $\beta_{\text{Bronchiolites}} \times [\text{Bronchiolites}]_t$ n'est pas inclus pour les personnes âgées, chez qui les bronchiolites sont peu souvent rapportées. Les modèles M1 et M2 ont été estimés par maximum de vraisemblance grâce à la fonction `arima` de `R`, selon la démarche usuelle de Box et Jenkins (Box *et al.*, 2015). Pour chaque modèle, l'identification de la structure d'autocorrélation a été validée par l'étude des fonctions d'autocorrélation des résidus et des tests de corrélation de Ljung-Box jusqu'à l'ordre 52, et la normalité des résidus a été validée par l'étude du graphique quantile-quantile des résidus et des tests de normalité de Kolmogorov-Smirnov. Pour ces tests, nous avons choisi un risque de première espèce de 0,01, pour éviter de sur-paramétrer les modèles.

À partir des estimations des modèles, les fractions de prescriptions antibiotiques attribuables aux différents groupes syndromiques et leurs intervalles de confiance à 95% ont été estimés (Section 1.3). Afin de pouvoir comparer les contributions des différentes infections respiratoires basses, dont les périodes épidémiques sont hivernales mais ne coïncident pas nécessairement, les fractions attribuables ont été calculées sur toute la période hivernale (d'octobre à mars inclus).

2.5.3. Résultats

Analyse descriptive Les trois séries des taux hebdomadaires de remboursements d'antibiotiques en ville sur la période de janvier 2010 à décembre 2017 considérées pour les trois populations étudiées sont présentées Figure 2.3. Pour chacune des séries, on observe une forte saisonnalité avec des taux de prescriptions plus élevés au cours des périodes hivernales. En particulier, pour la population générale et les enfants de 5 ans et moins, cinq pics de consommation par an se distinguent, avec une très forte régularité chez les enfants principalement : en octobre, décembre, fin janvier/début février, fin mars/début avril, et début juin. En revanche, pour les

personnes âgées de 75 ans et plus, un seul pic saisonnier par an peut être systématiquement observé début janvier.

À l'échelle de la population générale, le taux de remboursements d'antibiotiques en ville est en moyenne de 1 415 prescriptions pour 100 000 personnes (noté par la suite pcmp) par semaine entre 2010 et 2017. Pour les enfants et les personnes âgées, les moyennes sont de 3 268 et 1 524 pcmp par semaine respectivement. Au cours des périodes hivernales, le nombre de prescriptions monte en moyenne à 1 701 pcmp par semaine à l'échelle de la population entière, 4 178 pcmp par semaine pour les enfants et 1 822 pcmp par semaine pour les personnes âgées (Tableau 2.3). Il n'y a pas de variation significative du taux de remboursements entre les différentes périodes hivernales à l'échelle de la population et des personnes âgées (p-value $>$ 0,10), mais une diminution significative chez les enfants de 5 ans et moins, avec 1 029 pcmp par semaine de moins pendant l'hiver 2016-17 par rapport à l'hiver 2010-11 (p-value = 0,01) (résultats non fournis).

Les différents groupes syndromiques (en global ou spécifiques) présentent tous des saisonnalités marquées, avec une proportion de visites nettement plus élevées durant les périodes hivernales (Figure 2.3). Au sein de chaque classe d'âge considérée et à l'exception des syndromes grippaux, les séries d'incidence des différents groupes syndromiques sont très fortement corrélées entre elles (Annexe 2.B), avec des coefficients de corrélation élevés ($\geq$ 0,83).

Chez les enfants de 5 ans et moins, les pics d'incidence associés à l'apparition des bronchiolites ont systématiquement précédé ceux des syndromes grippaux au cours des périodes hivernales de 2010 à 2017. En particulier, le deuxième pic annuel d'utilisation d'antibiotiques chez les enfants était fortement associé aux pics d'incidence de bronchiolites, avec un coefficient de corrélation égal à 0,79 (IC : 0,75-0,82) entre les deux séries. En revanche, bien que visuellement liés, les pics d'incidence des syndromes grippaux n'étaient que faiblement associés au troisième pic d'utilisation d'antibiotiques, avec un coefficient de corrélation de 0,36 (IC : 0,27-0,44).

À l'échelle de la population entière, les infections respiratoires basses représentent 4,07% des visites aux services d'urgence pendant les périodes hivernales, réparties en 41% de pneumopathies, 22% de bronchiolites, 21% de bronchites et 15% de syndromes grippaux (Tableau 2.3). Chez les enfants, la proportion de visites avec un diagnostic d'infections virales est 3,6 fois plus élevée, dont 68% de bronchiolites. Chez les personnes âgées, ce sont les infections non virales qui sont

Tableau 2.3 – Taux de remboursements d'antibiotiques en ville (pour 100 000 habitants par semaine) et proportion de visites avec un diagnostic d'infection respiratoire basse aux services d'urgence (en % par semaine) : moyenne et écarts-type des moyennes sur les périodes hivernales de janvier 2010 à décembre 2017.

	Tous âges	75 ans et plus	5 ans et moins
Taux de remboursements d'antibiotiques (pour 100 000 par semaine)			
	1 701 (35,8)	1 822 (54,2)	4 178 (115,4)
Proportions de visites aux services d'urgences (en % par semaine)			
Infections respiratoires basses non virales			
Toutes infections	1,74 (0,03)	5,84 (0,14)	1,54 (0,03)
Pneumopathies	1,66 (0,03)	5,45 (0,14)	1,53 (0,03)
Infections respiratoires basses virales			
Toutes infections	2,33 (0,07)	1,79 (0,10)	8,37 (0,16)
Bronchiolites	0,88 (0,04)		5,71 (0,20)
Bronchites	0,86 (0,02)	1,55 (0,07)	1,33 (0,03)
Syndromes grippaux	0,60 (0,06)	0,18 (0,03)	1,39 (0,15)

3,4 fois plus fréquentes que dans la population entière.

Ajustement des modèles Pour chacun des modèles, les critères d'adéquation sont donnés en Annexe 2.C et les estimations des paramètres dans le Tableau 2.4. La structure d'autocorrélation des modèles correspondant aux enfants de 5 ans et moins est différente de celle des modèles correspondant à la population générale et aux personnes âgées de 75 ans et plus. Pour ces derniers, seule une structure autorégressive (AR) est retrouvée, alors qu'une structure autorégressive et moyenne mobile (ARMA) est retrouvée pour les enfants de 5 ans et moins. Par ailleurs, ni la structure d'autocorrélation ni l'estimation des paramètres de la structure ne sont fortement modifiées selon les variables explicatives prises en compte (globale ou spécifiques).

TABLEAU 2.4 – Estimation (et écarts-type) des paramètres des régressions linéaires avec erreurs autocorrélées[1] pour chacun des modèles.

	Tous âges	75 ans et plus		5 ans et moins	
Paramètres	M1	M1	M2	M1	M2
μ	893 (69)	697 (78)	693 (75)	1 432 (147)	1 377 (144)
$\beta_{\text{Inf. non-virales}}$	0,23 (0,05)	0,11 (0,02)		0,71 (0,09)	
$\beta_{\text{Pneumopathies}}$			0,13 (0,02)		0,70 (0,09)
$\beta_{\text{Inf. virales}}$	0,13 (0,02)	0,21 (0,02)		0,19 (0,02)	
$\beta_{\text{Bronchiolites}}$					0,14 (0,02)
$\beta_{\text{Bronchites}}$			0,20 (0,04)		0,44 (0,11)
$\beta_{\text{Synd. grippaux}}$			0,18 (0,04)		0,21 (0,03)
ϕ_1	0,29 (0,06)	0,14 (0,05)	0,14 (0,05)	0,69 (0,04)	0,66 (0,05)
ϕ_2	0,23 (0,05)	0,25 (0,05)	0,24 (0,05)		
φ_1	0,82 (0,03)	0,78 (0,03)	0,77 (0,03)	0,90 (0,03)	0,91 (0,03)
θ_3				-0,21 (0,06)	-0,20 (0,06)
θ_6				-0,13 (0,06)	-0,11 (0,06)
ϑ_1				-0,39 (0,07)	-0,40 (0,07)
σ^2	9 244	12 012	12 074	66 780	64 725

[1] Les structures d'autocorrélation identifiées sont un SAR(2)$(1)_{52}$, de polynôme autorégressif $\Phi(z) = (1-\phi_1 z-\phi_2 z^2)(1-\varphi_1 z^{52})$ pour la population générale et les personnes âgées de 75 ans et plus ; et un SARMA(1, 6)$(1, 1)_{52}$, de polynômes autorégressif $\Phi(z) = (1-\phi_1 z)(1-\varphi_1 z^{52})$ et moyenne mobile $\Theta(z) = (1-\theta_3 z^3-\theta_6 z^6)(1-\vartheta_1 z^{52})$ pour les enfants de 5 ans et moins.

Fractions attribuables Les moyennes des fractions attribuables sur les périodes hivernales entre janvier 2010 et décembre 2017 sont présentées dans le Tableau 2.5. La somme des fractions attribuables des groupes syndromiques spécifiques, estimées à partir des modèles M2, est cohérente avec la fraction attribuable de l'indicateur global, estimée à l'aide des modèles M1.

Au cours des périodes hivernales et à l'échelle de la population générale, 40% (intervalle de confiance à 95% : 29-52%) des prescriptions antibiotiques sont associées aux infections respiratoires basses, dont 23% (IC : 13-33%) et 17% (IC :

Tableau 2.5 – Fraction des remboursements d'antibiotiques en ville attribuables aux infections respiratoires basses : moyenne [et intervalles de confiance à 95%] sur les périodes hivernales entre 2010 et 2017.

	Tous âges	75 ans et plus	5 ans et moins
Infections respiratoires basses non virales			
Toutes infections	23 [13-33]	36 [24–47]	26 [20–32]
Pneumonies		38 [26–50]	25 [19–32]
Infections respiratoires basses virales			
Toutes infections	17 [13-22]	20 [16–25]	38 [31–46]
Bronchiolites			19 [13–26]
Bronchites		17 [10–24]	14 [07–21]
Syndromes grippaux		02 [01–03]	07 [05–09]

13-22%) attribuables aux infections non virales et virales respectivement.

Chez les personnes âgées, la contribution des infections virales était comparable à celle de l'ensemble de la population, mais les infections non virales ont contribué à hauteur de 36% (IC : 24-47%) des prescriptions antibiotiques. En revanche pour les enfants de 5 ans et moins, c'est la fraction attribuable des infections non virales qui était comparable (26%, IC : 20-32%), alors que celle des infections virales ont représenté 38% (IC : 31-46%) des prescriptions, les bronchiolites étant les causes principales, responsables de la moitié de ces prescriptions (19%, IC : 13-26%).

Au cours des différentes périodes hivernales sur la période étudiée, les fractions attribuables des différents syndromes respiratoires ont peu évolué (Figure 2.4). En effet, seule la contribution des syndromes grippaux, entre 5 et 10% chez les enfants, varie selon les années, mais cette contribution est négligeable à l'échelle de la population.

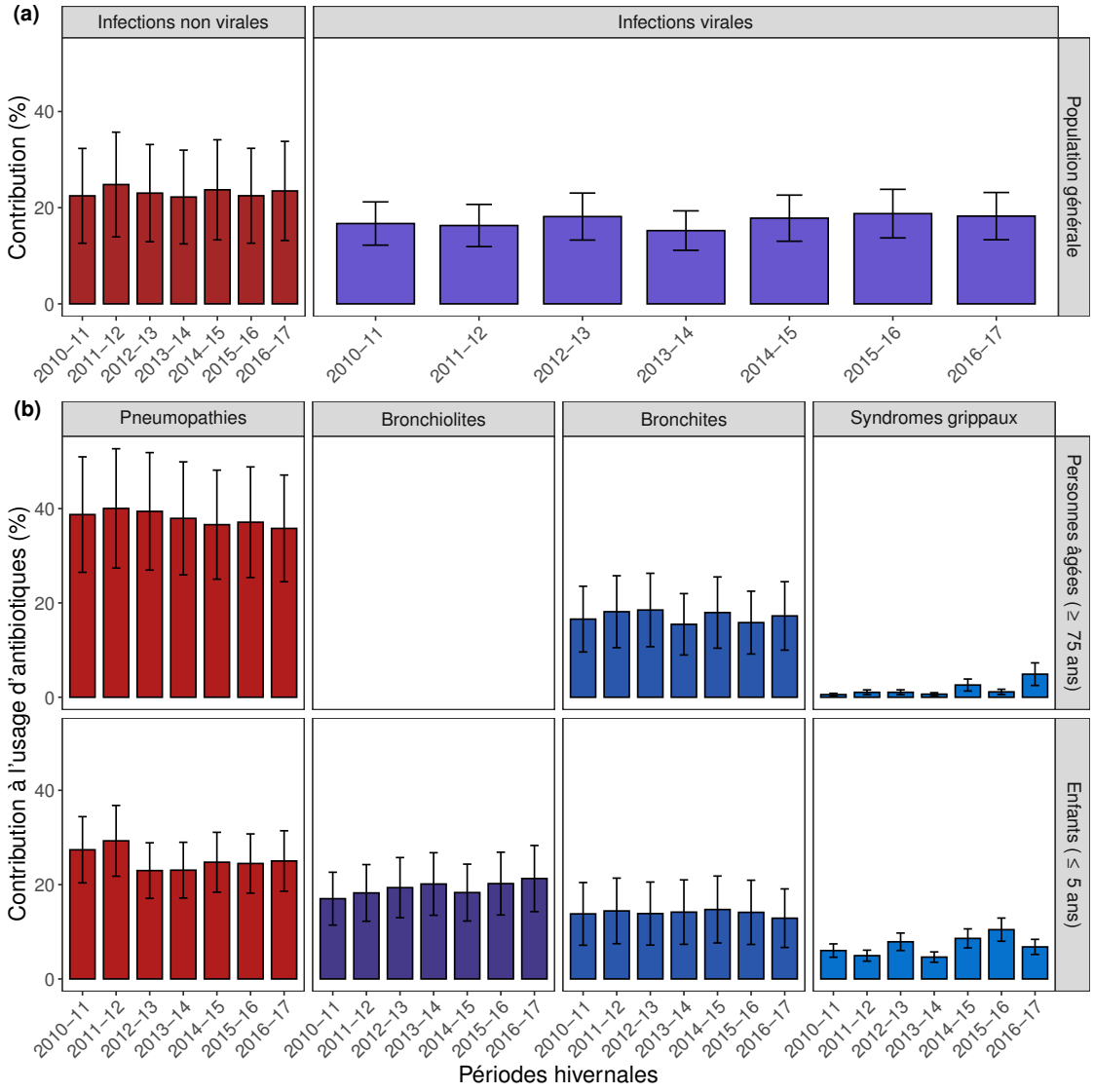


FIGURE 2.4 – Fractions des prescriptions antibiotiques en ville attribuables aux différents syndromes respiratoires au cours des périodes hivernales. (a) À l'échelle de la population générale. (b) Chez les personnes âgées de 75 ans et plus (haut) et les enfants de 5 ans et moins (bas).

2.5.4. Discussion

Ce travail estime que, durant les périodes hivernales, les infections respiratoires basses virales contribueraient pour 17% des prescriptions antibiotiques en ville, et jusqu'à 20% et 38% pour les personnes âgées et les enfants respectivement. En particulier chez les enfants, les bronchiolites seraient responsables de 19% des prescriptions d'antibiotiques.

Déterminer si ces prescriptions antibiotiques sont appropriées ou non est difficile, puisqu'il s'agit de données agrégées. Comme proposé dans la littérature, on supposera que la proportion des prescriptions d'antibiotiques pour les infections respiratoires basses virales qui sont inappropriées est évaluée entre 50% et 100% (Dekker *et al.*, 2015; Fleming-Dutra *et al.*, 2016). En particulier pour la bronchiolite du nourrisson, il n'a pas été prouvé que l'utilisation des antibiotiques réduiraient la sévérité de la maladie, ni sa durée (Farley *et al.*, 2014). Dans la situation la plus optimiste, le mésusage représenterait donc 8,5% des prescriptions antibiotiques, soit 145 prescriptions pour 100 000 personnes par semaine, à l'échelle de la population, et 19%, soit 794 prescriptions pour 100 000 personnes par semaine, chez les enfants. Dans la situation pessimiste, ce mésusage atteindrait 17%, soit 289 prescriptions pour 100 000 personnes par semaine, à l'échelle de la population générale, et 38%, soit 1 588 prescriptions pour 100 000 personnes par semaine, chez les enfants.

Pour savoir si une prescription antibiotique est justifiée, il est nécessaire de connaitre précisément la nature de l'infection ainsi que son étiologie microbienne. Dans ce travail, ces éléments sont déterminés à partir du code CIM-10 documenté lors de la visite aux services d'urgences. Concernant la nature de l'infection, Henriksen *et al.* (2014) ont montré que la sensibilité et valeur prédictive positive des codes CIM-10 reportés à la sortie du patient du service d'urgence sont faibles pour les infections respiratoires basses (70,6% et 71,1% respectivement), en particulier pour les diagnostics de pneumonie. À notre connaissance, il n'existe pas d'étude ayant déterminé la sensibilité et valeur prédictive positive des codes CIM-10 pour caractériser l'étiologie microbienne en services d'urgence. En particulier, il est vraisemblable que les codes CIM-10 de pneumopathies recueillis par le réseau Oscour®, qui rapportent pour 98% d'entre elles une étiologie non virale, ne soient pas représentatifs des pneumopathies rencontrées dans le cadre de soins de ville, pour lesquelles une étiologie virale est de plus en plus souvent rencontrée (Ieven *et al.*, 2018; Jain *et al.*, 2015; Ruuskanen *et al.*, 2011). Cela signifierait que la proportion

de prescriptions inappropriées associées aux diagnostics de pneumopathies est non nulle, et d'autant plus importante que l'étiologie virale est fréquemment retrouvée. Une autre considération importante est de savoir si les visites aux services d'urgence reflètent correctement l'épidémiologie des infections dans la population générale. Puisque nous avons modélisé le nombre de prescriptions ambulatoires d'antibiotiques en utilisant la fréquentation des services d'urgence, celle-ci doit être un bon indicateur des événements sanitaires dans la communauté. Des études (Josseran *et al.*, 2006, 2010; Pelat *et al.*, 2017) ont montré que, bien que construit à l'origine pour répondre à la nécessité de détecter les menaces de santé publique, le réseau Oscour® était suffisamment sensible pour évaluer l'impact sanitaire tant des conditions environnementales que des maladies infectieuses, notamment en le comparant aux données de syndromes grippaux du réseau Sentinelles, qui est le système de surveillance de référence pour la grippe en France (Flahault *et al.*, 1997).

Même si elles contribuent aussi à l'utilisation ambulatoire d'antibiotiques, les infections des voies respiratoires supérieures n'ont pas été prises en compte dans l'étude. En effet, si la fréquentation des urgences semble refléter avec précision les épidémies d'infections respiratoires basses aiguës dans la communauté, ce n'est probablement pas le cas pour les infections des voies respiratoires supérieures, qui sont le plus souvent prises en charge en ville. Néanmoins, cette non prise en compte ne modifie pas les conclusions de cette étude comme le montre la grande quantité de prescriptions inappropriées attribuables aux infections des voies respiratoires inférieures. En effet, les estimations des fractions attribuables seraient encore plus élevées en incluant dans le modèle les infections respiratoires hautes.

Pour pouvoir faire des comparaisons entre elles, les fractions attribuables ont été estimées sur les périodes hivernales (26 semaines, d'octobre à mars inclus) plutôt que sur les périodes épidémiques respectives de chaque groupe syndromique. Elles peuvent donc paraître sous-estimées par rapport aux estimations qui ne s'intéressent qu'aux périodes épidémiques. En particulier, les fractions attribuables estimées en Section 2.4.3 sont plus élevées car calculées sur les périodes épidémiques plus courtes (entre 5 et 16 semaines, Tableau 1.B.1). En revanche, les résultats sont comparables à l'étude de Fleming-Dutra *et al.* (2016) qui estiment 110 prescriptions inappropriées pour 1 000 habitants par an, soit 212 prescriptions pour 100 000 personnes par semaine, attribuables aux infections respiratoires aux Etats-Unis.

Les corrélations élevées entre les séries d'incidence des différents groupes syn-

dromiques pourraient suggérer une circulation simultanée des différentes infections respiratoires ou bien des mauvais classements lors de la codification des visites aux services d'urgence. En particulier, nous avons supposé l'étiologie microbienne de certains des codes CIM-10 dont l'étiologie n'est pas spécifiée, comme les bronchites aiguës codées J20, que nous avons considérées comme virales. Néanmoins, comme les groupes syndromiques spécifiques sont exclusifs et que les modèles sont additifs, la somme des contributions des syndromes spécifiques est cohérente avec la contribution estimée à partir de l'indicateur global des infections respiratoires basses. Cette observation montre non seulement la robustesse des indicateurs des groupes syndromiques reportés par le réseau Oscour®, mais renforce les résultats de l'analyse et leur interprétation.

2.6. Conclusions

Nous avons mis en évidence l'impact de la campagne de sensibilisation "Les antibiotiques, c'est pas automatique !" sur la baisse du nombre de prescriptions d'antibiotiques associé aux épidémies de grippe saisonnières. Nous avons estimé que le nombre de prescriptions associé à la grippe a diminué de plus de moitié au cours de cette période, ce qui pourrait être interprété comme une diminution du mésusage des antibiotiques, témoignant du succès de la campagne de sensibilisation.

Nous avons ensuite montré que les infections virales aiguës des voies respiratoires inférieures étaient des facteurs importants de sur-prescription d'antibiotiques en ville. Plus précisément, elles représenteraient pendant les périodes hivernales jusqu'à 17% de l'utilisation globale d'antibiotiques en ville à l'échelle de la population générale, soit 289 prescriptions pour 100 000 habitants par semaine, et jusqu'à 38% chez les enfants de 5 ans et moins, soit 1 588 prescriptions pour 100 000 habitants par semaine, dont la moitié attribuables aux bronchiolites.

Chapitre 3

Processus de Hawkes et données agrégées

3.1. Contexte

Les maladies contagieuses sont des maladies infectieuses qui se transmettent facilement par contact avec une personne malade ou ses sécrétions. Leur modélisation présente une difficulté fondamentale par rapport aux maladies peu ou non transmissibles : les cas ne sont pas indépendants, car les individus infectés sont des facteurs de diffusion de la maladie. Si cette structure de dépendance n'est pas prise en compte lorsque l'on s'intéresse à identifier un facteur de risque, elle peut conduire à des biais dans l'estimation des paramètres, et donc dans l'interprétation des résultats (Paynter, 2016; Mishra et Baral, 2020). De plus, si le niveau d'agrégation est grossier, une partie des informations disponibles sur la structure de dépendance, et donc la transmission de la maladie, est perdue, puisque les interactions à des échelles plus faibles que le niveau d'agrégation ne sont plus observées.

Lorsque l'on travaille avec des séries temporelles construites en agrégeant des données, les processus autorégressifs, et plus généralement ARMA (*autoregressive moving average*) ou ARIMAX (*autoregressive integrated moving average with exogenous variables*), sont une famille de modèles adaptés pour prendre en compte des dépendances temporelle. En effet, le théorème de décomposition de Wold établit que toute série temporelle stationnaire peut être représentée, et donc modélisée, par un processus ARMA. Toutefois, l'estimation des paramètres du processus peut en pratique poser problème, car elle nécessite d'identifier manuellement les ordres des parties autorégressives et moyenne mobile. De plus, même s'il est bien identifié, les paramètres de la structure de dépendance d'un processus ARMA peuvent être difficiles à interpréter dans un contexte épidémiologique.

À la place, nous considérons les processus de Hawkes qui sont une famille de modèles présentant des propriétés d'auto-excitation. Le processus de Hawkes dé-

crit une succession d'événements en insistant sur la dépendance des observations actuelles par rapport aux événements passés : une série d'événements dans le passé influence positivement le nombre de nouveaux cas. Cette dépendance est modélisée sous la forme d'une intensité conditionnelle qui représente l'espérance attendue du nombre de nouveaux cas sur un pas de temps en fonction des événements passés. Ces processus sont donc adaptés pour modéliser des maladies contagieuses à l'échelle individuelle et renseigner sur la structure de dépendance temporelle et donc la transmission de la maladie. En particulier, les paramètres de ce modèle sont interprétables directement : ils décrivent la durée de contagion de la maladie, mais aussi le taux de reproduction et la fréquence d'occurrence de *clusters*.

Cependant, dans le cas où les données sont agrégées, c'est-à-dire que les cas de la maladie sont dénombrés sur des intervalles de temps réguliers, les paramètres du processus de Hawkes ne peuvent être estimés avec des méthodes classiques : les estimateurs usuels du maximum de vraisemblance ou des moments ne sont pas explicites. Une approche alternative de leur estimation peut être développée en s'appuyant sur des méthodes spectrales, qui consistent à déplacer le problème dans le cadre fréquentiel plutôt que temporel. Pour prouver que ces méthodes fournissent des bons estimateurs, il est nécessaire de montrer que le processus de Hawkes vérifie des hypothèses plus fortes que dans le cas des méthodes classiques.

3.2. Objectifs

Nous montrons que les processus de Hawkes sont adaptés pour modéliser des maladies contagieuses, et étudions l'impact de l'agrégation de données sur son estimation. Dans un premier temps, nous rappelons les définitions et notions essentielles pour l'étude des processus de Hawkes. Ensuite, nous développons une méthode d'estimation dans le cas où les événements du processus ne sont pas observés, mais seulement dénombrés sur des unités de temps régulières. Pour montrer que cette méthode d'estimation produit des estimateurs consistants et asymptotiquement normaux des paramètres, nous prouvons que les processus de Hawkes sont fortement mélangeants. Enfin, nous illustrons cette méthode par des études de simulation et une étude de cas réelle.

3.3. Présentation des outils statistiques

3.3.1. Processus ponctuel

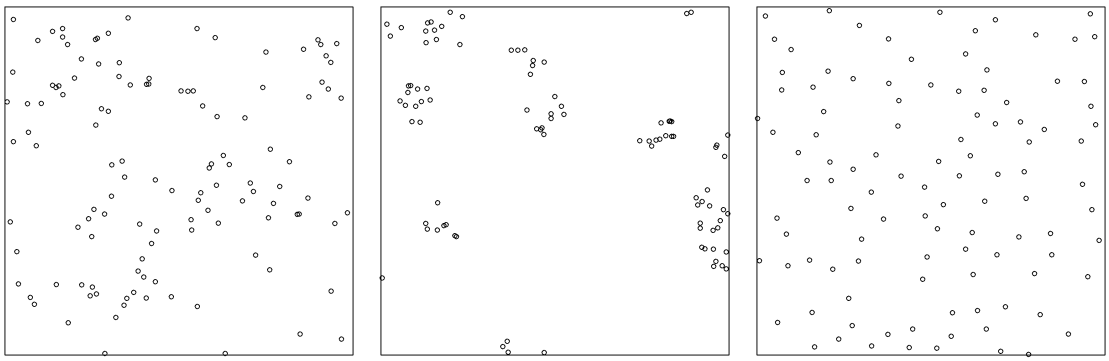
Les processus ponctuels sont des outils de modélisation permettant de définir un cadre formel de calcul pour l'étude statistique d'événements ou d'objets localisés précisément dans l'espace (Figure 3.1a). On retrouve leur utilisation dans des contextes très différents, par exemple pour modéliser et analyser les positions des arbres ou des nids d'oiseaux (en écologie statistique, Ludwig et Reynolds, 1988), celles des étoiles et des galaxies (en astrostatistique, Babu et Feigelson, 1996), ou encore celles des défauts dans un *wafer* de silicium (en science des matériaux, Ohser et Mücklich, 2000). Les processus ponctuels sont également utilisés pour analyser des séquences d'événements aléatoires, observés dans le temps (Figure 3.1b), par exemple le partage des tweets sur la plateforme Twitter (Farajtabar *et al.*, 2014; Zhao *et al.*, 2015), ou les accidents majeurs dans les mines de charbon au Royaume-Uni (Cox et Lewis, 1966).

En épidémiologie, les processus ponctuels spatiaux ont été utilisés pour étudier des phénomènes localisés précisément, comme, par exemple, l'incidence de la leucémie infantile à North Humberside, au Royaume-Uni (Diggle et Chetwynd, 1991), le cancer du poumon et du larynx dans le Lancashire, au Royaume-Uni (Diggle, 1990; Kelsall et Diggle, 1995) ou l'incidence d'encéphalite à tiques en République Tchèque (Benes *et al.*, 2005). Les processus ponctuels temporels sont également utilisés, par exemple pour la prédiction du parcours de soins des patients aux Etats-Unis (Xu *et al.*, 2016).

Les ouvrages de référence pour l'étude théorique des processus ponctuels sont Daley et Vere-Jones (2003); Møller et Waagepetersen (2004). Pour une approche plus accessible, je recommande vivement Baddeley (2007) et Floch *et al.* (2018), dont je me suis fortement inspiré pour introduire le formalisme des processus ponctuels. Pour une approche tournée vers le domaine épidémiologique, voir Elliott *et al.* (2000).

Les processus ponctuels peuvent être définis formellement, de façon équivalente, selon deux approches complémentaires.

Définition par la position des points Intuitivement, un processus ponctuel peut être considéré comme un tirage aléatoire d'un ensemble de points de $\mathbb{R}^d$ ($d =$



(a) Gauche : un processus de Poisson ; centre : un processus agrégé (Matérn) ; droite : un processus régulier (Matérn II).

(b) Haut : un processus de Poisson ; centre : un processus agrégé (Matérn) ; bas : un processus régulier (Matérn II).

FIGURE 3.1 – Réalisations de trois processus ponctuels. (a) Processus ponctuels spatiaux, $d = 2$. (b) Processus ponctuels temporels, $d = 1$.

$1, 2, 3$ ou 4 en pratique). Plus formellement, un processus ponctuel X peut être défini comme une variable aléatoire, c'est-à-dire une application $\omega \mapsto X(\omega)$ de Ω, l'univers des possibles, à valeurs dans l'ensemble des semis de $\mathbb{R}^d$.

Définition. On appelle **semis de points** (ou configuration de points) tout ensemble au plus dénombrable de points $x = \{x_1, x_2, \ldots\}$ de $\mathbb{R}^d$.

Chaque éventualité $\omega \in \Omega$ détermine donc l'ensemble des points $X(\omega) = \{X_1(\omega), X_2(\omega), \ldots\}$ du processus ponctuel.

Définition par le comptage des points Il est possible d'aller plus loin en identifiant X avec la famille d'applications

$$\omega \mapsto N(\omega, B) \coloneqq \mathrm{card}(X(\omega) \cap B), \qquad B \subset \mathbb{R}^d,$$

qui comptent le nombre de points du processus ponctuel qui sont dans l'ensemble B. (card(A) désigne le nombre d'éléments de l'ensemble A.) (Figure 3.2) Pour tout $\omega \in \Omega$, l'application $N(\omega, \cdot)$, notée juste $N(\cdot)$ s'il n'existe pas d'ambiguïté, est une mesure, appelée la mesure de comptage associée à $X(\omega)$.

Définition. On appelle **mesure de comptage** une application N qui vérifie les propriétés suivantes :

- N est à valeurs entières non négatives : pour tout ensemble B,

$$N(B) \in \mathbb{N}.$$

- N est additive : pour tous ensembles disjoints A, B,

$$N(A \cup B) = N(A) + N(B).$$

- L'ensemble vide est de mesure nulle :

$$N(\varnothing) = 0.$$

La définition par les mesures de comptages semble moins directe, mais permet de puiser dans la théorie des mesures aléatoires (Kallenberg, 1983; Baccelli *et al.*, 2020) pour étudier les processus ponctuels. L'espace des réalisations d'un processus ponctuel dans $\mathbb{R}^d$ est alors $\mathfrak{N}$, l'ensemble de toutes les mesures de comptage sur $\mathbb{R}^d$, et un processus ponctuel est donc une variable aléatoire à valeurs dans $\mathfrak{N}$.

Équivalence entre les définitions On peut vérifier que la connaissance des valeurs $N(B)$ pour tout compact B donne suffisamment d'information pour reconstruire complètement le semis : les points du semis correspondent aux positions x telles que $N(\{x\}) > 0$. Inversement, on peut construire la mesure de comptage en associant une masse de Dirac δ_x en tout point x du semis. Ces points x sont appelés les *atomes* de la mesure de comptage N.

Définition. Une **masse de Dirac** (ou mesure de Dirac) au point x est une mesure supportée par x et de masse unitaire :

$$\forall A \subset \mathbb{R}^d, \quad \delta_x(A) = \mathbb{1}_A(x) = \begin{cases} 1, & \text{si } x \in A, \\ 0, & \text{sinon,} \end{cases}$$

où $\mathbb{1}_A$ est appelée la fonction indicatrice de A.

La masse de Dirac au point x associe la mesure 1 aux ensembles qui contiennent x et 0 à ceux qui ne le contiennent pas. La mesure de comptage N est alors donnée par :

$$N = \sum_{x \in X} \delta_x, \tag{3.1}$$

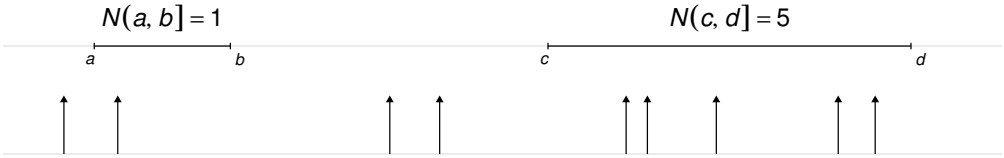
et vaut, pour tout $B \in \mathbb{R}^d$ compact,

$$N(B) = \sum_{x \in X} \mathbb{1}_B(x).$$

N compte bien le nombre de point du semis X qui sont dans l'ensemble B.

En raison de l'équivalence entre les définitions des processus ponctuels, nous alternerons souvent entre la notation X pour désigner le processus lorsqu'il est considéré comme un semis aléatoire, et la notation N pour la mesure de comptage associée (Figure 3.2).

(a) Comme semis aléatoire : la réalisation du processus est un ensemble de points.



(b) Comme mesure de comptage aléatoire : la réalisation du processus est la mesure N qui compte le nombre de points dans chaque ensemble.

FIGURE 3.2 – Deux formalismes pour définir les processus ponctuels.

Définition du tirage aléatoire Un processus ponctuel étant un phénomène aléatoire, il doit être décrit en définissant l'espace des résultats possibles, puis en définissant l'ensemble des événements possibles et leurs probabilités. Rappelons que dès qu'un tirage aléatoire est effectué dans un ensemble de réalisations non dénombrable, il n'est plus possible d'affecter une probabilité à chaque réalisation. Dans le

cas d'une variable normale par exemple, on définit la probabilité sur des intervalles et non sur chacune des valeurs réelles. La probabilité sera définie pour chaque intervalle. On l'étend ensuite à tout ensemble que l'on peut décrite comme une union ou intersection infinie d'intervalles ; un tel ensemble est appelé ensemble borélien. L'ensemble de ces boréliens est appelé tribu borélienne.

Dans le cas de notre processus ponctuel, nous n'affectons pas de probabilité à chaque semis. À la place, on procède en construisant les ensembles de semis qui vont être les événements pour le processus ponctuel. L'ensemble de ces événements est appelé la tribu du processus. Définissons maintenant notre choix d'événements pour le processus ponctuel. Un événement élémentaire pour le processus ponctuel est un événement du type

$$\mathcal{A}_{B,m} = \{N \in \mathfrak{N} : N(B) = m\}, \tag{3.2}$$

c'est-à-dire l'événement tel qu'il y a exactement m points dans la région B, où $B \subset \mathbb{R}^d$ est compact et m est un entier naturel. Remarquons que cet événement contient une infinité de semis, puisque seul le nombre de points dans l'ensemble B fixé est connu, mais pas leur position, et rien n'est dit des points à l'extérieur de B.

Soit $\mathcal{N}$ la tribu engendrée par tous les événements de la forme $\mathcal{A}_{B,m}$, c'est-à-dire le plus petit ensemble contenant tous ces événements, leurs intersections finies ou infinies et leurs complémentaires. $\mathcal{N}$ est l'ensemble des événements pour le processus ponctuel. L'ensemble $\mathfrak{N}$, ainsi muni de la tribu $\mathcal{N}$, est l'espace des résultats pour un processus ponctuel sur $\mathbb{R}^d$.

Un processus ponctuel X peut maintenant être défini formellement, à partir de la mesure de comptage N associée à la réalisation du semis.

Définition. Un **processus ponctuel** X sur $\mathbb{R}^d$ est une application mesurable d'un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ dans l'espace des mesures de comptage localement finies $(\mathfrak{N}, \mathcal{N})$:

$$\begin{aligned} N : (\Omega, \mathcal{F}, \mathbb{P}) &\to (\mathfrak{N}, \mathcal{N}), \\ \omega &\mapsto N(\omega, \cdot). \end{aligned}$$

On supposera également par le suite que les processus ponctuels sont localement finis et simples, c'est-à-dire que toute région bornée possède une mesure finie : pour tout compact $B \subset \mathbb{R}^d$, $N(B) < \infty$; et que deux points du processus ne sont jamais exactement superposés : pour tout $x \in \mathbb{R}^d$, $N(\{x\}) \leq 1$. Ceci permet de réduire considérablement la complexité de l'ensemble des événements d'intérêts $\mathcal{N}$.

La mesurabilité du processus, *i.e.* la certitude que chaque événement de $\mathcal{N}$ peut être mesuré en probabilité, est assurée par la construction de la tribu $\mathcal{N}$ qui garantit que les variables $N(B)$, avec $B \subset \mathbb{R}^d$ compact, sont des variables aléatoires définies également sur $(\Omega, \mathcal{F}, \mathbb{P})$. Il s'agit d'ailleurs de la plus petite tribu garantissant cette propriété.

Processus de Poisson et processus d'agrégat Le processus de Poisson est sans aucun doute le plus emblématique des processus ponctuels, et forme la base à partir de laquelle la plupart des autres modèles sont construits.

Définition. Le **processus de Poisson** homogène, d'*intensité* $\beta > 0$, est un processus ponctuel N tel que

- pour tout compact $B \subset \mathbb{R}^d$, le nombre de points $N(B)$ suit une loi de Poisson de paramètre $\beta \cdot |B|$;
- si $B_1, \ldots, B_m$ sont des régions disjointes, alors $N(B_1), \ldots, N(B_m)$ sont indépendants.

Ici, $|\cdot|$ représente la mesure de Lebesgue : il s'agit de la longueur en une dimension, de la surface en deux dimensions, du volume en trois dimensions, etc.

Grâce à ces deux propriétés, appelées propriétés d'*homogénéité* et d'*indépendance* respectivement, le processus de Poisson est le seul processus permettant de générer des semis de points complètement aléatoires (Daley et Vere-Jones, 2003, Chapitre 2.2). En particulier, l'homogénéité assure l'absence de préférence entre région de l'espace, et entraine que le nombre de points attendus dans chaque région $B \subset \mathbb{R}^d$ soit proportionnel à son "volume" (au sens de la mesure de Lebesgue) : $\mathbb{E}[N(B)] = \beta \cdot |B|$. β désigne donc le nombre moyen de points par unité de volume. Par ailleurs, la propriété d'indépendance assure qu'il n'existe pas d'interaction entre les points du processus.

Il est possible de relâcher la condition d'homogénéité : on considère une *fonction d'intensité* $\beta : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, de sorte que le nombre moyen de points dans un compact $B \subset \mathbb{R}^d$ vaut

$$\mathbb{E}[N(B)] = \int_B \beta(s) \mathrm{d}s.$$

Dans ce cas, le processus de Poisson est dit *inhomogène* : les points vont préférer les régions d'intensité élevée.

Un processus dérivé du processus de Poisson est le **processus d'agrégat de Poisson**, qui permet de générer des semis présentant des structures d'agrégation des

points. À partir d'un premier processus ponctuel de Poisson X, appelé processus des centres, on remplace chaque point $x \in X$ du processus par un ensemble fini de points G_x, appelé l'agrégat de centre x. La superposition de tous les agrégats forment le processus d'agrégat $Y = \cup_{x \in X} G_x$.

On suppose généralement que les différents agrégats G_x sont des processus indépendants. Un exemple simple de processus d'agrégat est le processus de Matérn, pour lequel le processus des centres X est un processus de Poisson homogène, et chaque agrégat G_x est composé d'un nombre aléatoire de points indépendamment et identiquement distribués autour de x (voir Figure 3.1(a, b) centre).

Somme, intensité et stationnarité Nous rappelons brièvement certaines notations et propriétés essentielles pour traiter les processus ponctuels.

Pour toute fonction f à valeurs dans $\mathbb{R}^d$, on notera

$$N(f) \coloneqq \int_{\mathbb{R}^d} f(s) N(\mathrm{d}s) = \sum_i f(X_i)$$

la somme des $f(X_i)$ sur tous les points X_i du processus. L'écriture sous forme d'intégrale fait implicitement appel à la représentation du processus comme somme de masses de Dirac, donnée par la relation (3.1) : $N(\mathrm{d}s) = \sum_{x \in X} \delta_x(s)\mathrm{d}s$.

Les lois des processus ponctuels sont complexes. En pratique, on étudie donc préférentiellement les moments du processus.

Définition. La **mesure d'intensité** d'un processus ponctuel N est la mesure ν_1 définie par

$$\nu_1(B) = \mathbb{E}[N(B)],$$

pour tout compact $B \subset \mathbb{R}^d$.

Si elle existe, la **fonction d'intensité** du processus est la fonction $\lambda : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ telle que

$$\nu_1(B) = \int_B \lambda(s)\mathrm{d}s.$$

La mesure d'intensité correspond au moment d'ordre 1 du processus et informe sur le nombre moyen (espérance) de points dans toute région. La fonction d'intensité (ou parfois juste intensité) correspond au nombre attendu de points dans une région infinitésimale $\mathrm{d}s$:

$$\lambda(x) = \lim_{|\mathrm{d}s| \to 0} \frac{\mathbb{E}[N(x + \mathrm{d}s)]}{|\mathrm{d}s|}.$$

Par exemple, pour le processus de Poisson homogène, on a toujours $\mathbb{E}[N(B)] = \beta \cdot |B|$: l'intensité du processus est donc bien la fonction constante égale à β.

Enfin, comme une seule observation d'un semis de points ne permet pas de distinguer les propriétés statistiques du processus ponctuel en absence d'information supplémentaire, une hypothèse souvent supposée est celle de la stationnarité du processus.

Définition. Un processus ponctuel X est dit **stationnaire** si ses propriétés statistiques sont invariantes par translation, c'est-à-dire si, pour tout vecteur $v \in \mathbb{R}^d$, la distribution du processus translaté $X + v$ (obtenue en changeant chaque point $x \in X$ en $x + v$) est identique à la distribution de X.

L'hypothèse de stationnarité permet d'assurer que les observations dans des ensembles translatés ont la même loi de probabilité. En considérant de tels sous-ensembles d'observation, on obtient une répétition d'expérience qui rend possible l'estimation. Pour exemple, le processus de Poisson homogène est stationnaire, tandis que le processus de Poisson inhomogène ne l'est pas.

Processus ponctuel temporel Un processus ponctuel temporel peut être traité différemment des processus en plus grandes dimensions, puisque le temps possède un ordre naturel. Il est alors possible d'étudier les points du semis, appelés aussi temps d'arrivée ou d'occurrence dans ce cadre, et que nous noterons désormais T_i au lieu de X_i, selon leur ordre $\ldots < T_{-1} < T_0 < T_1 < \ldots$, où T_i représente le temps d'arrivée du i-ème événement (Figure 3.3a). On peut également s'intéresser aux temps inter-arrivée $V_i = T_{i+1} - T_i$ (Figure 3.3b), l'avantage étant que, pour certains processus (par exemple le processus de Poisson, ou les processus de renouvellement), les variables $V_1, V_2, \ldots$ sont indépendantes.

Un outil statistique particulièrement bien adapté à l'ordre naturel des processus ponctuels temporels est l'intensité conditionnelle, qui permet d'étudier comment les temps d'arrivée passés influencent la probabilité d'occurrence des temps d'arrivée futurs.

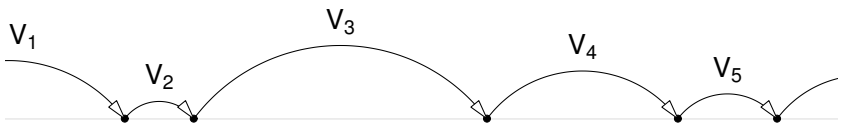
Définition. Si elle existe, la fonction d'**intensité conditionnelle** d'un processus ponctuel temporel N est la fonction $\lambda^* : \mathbb{R} \to \mathbb{R}_{\geq 0}$ telle que

$$\lambda^*(t) = \lim_{h \to 0} \frac{\mathbb{E}\left[N\Big((t, t+h]\Big) \mid \mathcal{H}(t)\right]}{h},$$

où $\mathcal{H}(t)$ désigne la tribu engendrée par les événements du type

$$\{N \in \mathfrak{N} : N(B) = m, B \subset (-\infty, t]\},$$

(a) Temps d'occurrence $\{T_i\}$.



(b) Temps inter-arrivée $\{V_i\}$.

FIGURE 3.3 – Représentation d'un processus ponctuel temporel.

et correspond à l'histoire du processus jusqu'au temps t.

Intuitivement, $\lambda^*(t)\mathrm{d}t$ est la probabilité qu'il existe un point de N dans un intervalle $(t, t+\mathrm{d}t]$ de taille infinitésimale $\mathrm{d}t$, conditionnellement aux temps d'arrivée dans le passé du processus. $\lambda^*(t)$ est donc une variable aléatoire d'espérance

$$\mathbb{E}[\lambda^*(t)] = \lambda(t).$$

3.3.2. Processus de Hawkes

Les processus de Hawkes linéaires sont une famille de processus stochastiques pour lesquels l'occurrence d'un événement augmente la probabilité d'occurrence des événements futurs. Ils ont été introduits par Hawkes (1971a,b), qui proposa la première définition d'un processus ponctuel auto-excitant, en s'appuyant sur les processus éponymes de Cox (1955) et les travaux de Bartlett (1963) sur l'analyse spectrale des processus ponctuels. En raison des propriétés d'auto-excitation des processus de Hawkes, ils sont bien adaptés pour modéliser des processus ponctuels présentant des agrégats de points. Alors que les premières applications concernaient presque exclusivement la sismologie pour analyser les répliques générées après la survenue d'un tremblement de terre (Adamopoulos, 1976; Ogata, 1988), leur utilisation s'est rapidement étendue à de nombreuses autres disciplines, notamment la neurophysiologie (Chornoboy *et al.*, 1988), la finance (Bacry *et al.*, 2015), la génomique

(Reynaud-Bouret et Schbath, 2010) et plus récemment l'épidémiologie (Meyer *et al.*, 2012).

Définition Le processus de Hawkes est un processus ponctuel dont la fonction d'intensité conditionnelle est stochastique et dépend des temps d'arrivée du passé.

Définition. Un **processus de Hawkes** linéaire sur $\mathbb{R}$ est un processus ponctuel N avec fonction d'intensité conditionnelle

$$\begin{aligned}\lambda^*(t) &= \eta + \int_0^t h(t-u)N(\mathrm{d}u) \\ &= \eta + \sum_{T_i<t} h(t-T_i).\end{aligned}$$

La constante $\eta > 0$ est appelée *intensité d'immigration* et la fonction mesurable $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ *fonction de reproduction.*

La fonction d'intensité conditionnelle du processus de Hawkes possède donc deux composantes distinctes : le premier terme, l'intensité d'immigration, est un terme déterministe et correspond à une probabilité constante de trouver un point du processus dans un intervalle ; en revanche, le deuxième terme dépend des réalisations passées du processus et détermine comment elles influencent la probabilité d'occurrence des points futurs. Ainsi, chaque nouveau point du processus va faire augmenter la probabilité d'occurrence des points selon la fonction de reproduction. La Figure 3.4 illustre la notion d'intensité conditionnelle pour un processus de Hawkes *exponentiel*, c'est-à-dire dont la fonction de reproduction est de la forme $h(t) = \alpha e^{-\beta t}$.

Équivalence avec les processus de branchement Le processus de Hawkes linéaire peut être représenté par un processus d'agrégat de Poisson (Hawkes et Oakes, 1974). En effet, le processus consiste en un flux d'*immigrants*, qui arrivent selon un processus de Poisson X_c d'intensité η, et sont les centres des agrégats. Ensuite, un immigrant arrivé au temps T_i génère des *enfants* selon un processus de Poisson inhomogène G_{T_i} de fonction d'intensité $h(\cdot - T_i)$. Ceux-ci à leur tour sont les centres d'autres processus d'agrégat et génèrent indépendamment d'autres enfants selon la même loi, et ainsi de suite *ad infinitum*. Les processus B_{T_i} formé d'un immigrant arrivé au temps T_i et de tous ses *descendants* — c'est-à-dire ses enfants, les enfants de ses enfants, *etc.* —, sont appelés *processus de branchement*,

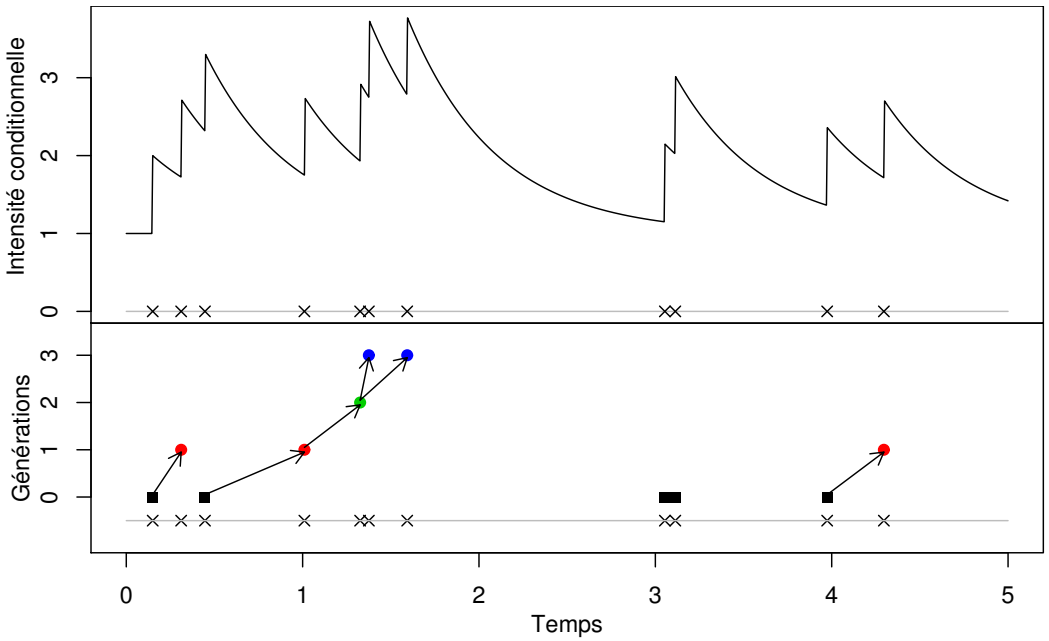


FIGURE 3.4 – Réalisation d'un processus de Hawkes exponentiel, avec $\eta = 1$ et $h(t) = e^{-2t}$: les croix correspondent aux temps d'arrivée du processus. (Haut) Représentation de l'intensité conditionnelle du processus : chaque point du processus augmente la probabilité d'occurrence des points futurs, selon un terme de forme exponentielle. (Bas) Représentation du processus sous forme de branchements : chaque immigrant (carrés noirs, de génération 0) peut générer des enfants (points rouges, de génération 1), qui peuvent générer à leur tour des enfants, et ainsi de suite *ad infinitum*.

et sont indépendants. Enfin, le processus de Hawkes X peut être défini comme la superposition de tous ces processus de branchement :

$$X = \bigcup_{t \in X_c} B_t.$$

La Figure 3.4 illustre l'équivalence du processus de Hawkes avec sa représentation sous forme de processus de branchement.

Cette représentation sous forme de processus d'agrégat permet de se rattacher à la théorie habituelle de Galton–Watson. Sans perte de généralité, considérons un processus de branchement dont l'immigrant arrive au temps 0. Définissons Z_k comme le nombre de points de génération k, *i.e.* $Z_0 = 1$ pour l'immigrant, puis Z_1 indique le nombre d'enfants que l'immigrant génère, Z_2 le nombre d'enfants que les enfants de l'immigrant génèrent, *etc.* Alors $(Z_k)_{k \in \mathbb{N}}$ est un processus de Galton–Watson.

En particulier, $(Z_{k+1} \mid Z_k = z)$ $(k, z \in \mathbb{N})$ suit une distribution de Poisson de paramètre $z\mu$, où $\mu \coloneqq \int_{\mathbb{R}} h(t)\mathrm{d}t$. Alors, selon les propriétés classiques de Galton–Watson, une condition suffisante pour l'existence du processus de Hawkes est $\mu < 1$ qui assure que le nombre total de descendants de tout immigrant est fini avec une probabilité 1 et a une espérance finie. Cette condition garantit également que le processus est strictement stationnaire.

Par la suite, nous écrirons fréquemment la fonction de reproduction de la forme

$$h = \mu h^*,$$

où $\int_{\mathbb{R}} h^*(t)\mathrm{d}t = 1$. Cette écriture permet de faire apparaitre explicitement le *taux de reproduction* μ d'une part, qui représente le nombre moyen d'enfants par individu, et h^*, que nous appelons le *noyau de reproduction*, et qui est une fonction de densité.

Intérêt du processus en épidémiologie Un des principaux attraits des processus de Hawkes pour la modélisation des processus ponctuels réside dans cette représentation sous forme de processus de branchement. En épidémiologie, cette représentation est particulièrement bien adaptée pour modéliser et prédire la diffusion de maladies contagieuses. Les temps d'arrivée représentent dans ce cas les nouveaux individus infectés, qui peuvent à leur tour transmettre la maladie à d'autres personnes. Les immigrants du modèle décrivent les cas pour lesquels aucun individu n'a été identifié comme source de transmission, par exemple des cas importés. Le processus de branchement généré par un immigrant du modèle représente alors un cluster de la maladie.

L'estimation des paramètres du processus de Hawkes permettrait alors de renseigner des informations importantes concernant la contagiosité d'une maladie. Ainsi, l'intensité d'immigration décrirait la fréquence d'apparition des clusters, tandis que la forme de la fonction de reproduction informerait sur la durée de contagiosité de la maladie. Enfin, le taux de reproduction μ pourrait être rapproché du R_0 épidémiologique qui représente le nombre moyen de personnes qu'un individu contagieux peut infecter.

Pourtant, l'utilisation du processus de Hawkes pour modéliser des maladies contagieuses est récente. Britton (2010) n'en fait pas mention dans sa revue de littérature des modèles épidémiques stochastiques, et les premières publications utilisant le processus de Hawkes dans le domaine de l'épidémiologie apparaissent en 2009. Ainsi, Meyer *et al.* (2012) ont proposé un modèle auto-excitant pour modéliser la

transmission des infections invasives à méningocoques en Allemagne qui a permis de montrer l'influence de l'âge des individus et du sérogroupe dans la dynamique de cette transmission. Rizoiu *et al.* (2018) ont proposé une extension du modèle, pour tenir compte de la taille finie de la population et donc de la réduction du nombre d'individus susceptibles au cours de la transmission, en s'inspirant du modèle classique SIR. Récemment, Chiang *et al.* (2020) ont montré que la modélisation de la pandémie de Covid-19 par processus de Hawkes est performante pour suivre et prédire l'évolution de la pandémie.

Estimation L'estimation des paramètres des processus de Hawkes a été étudiée de manière approfondie lorsque les temps d'arrivée sont observés, en s'appuyant principalement sur les méthodes du maximum de vraisemblance (Ogata, 1978; Ozaki et Ogata, 1979; Ogata, 1988). La fonction de vraisemblance d'une réalisation $\{t_1, \dots, t_p\}$ d'un processus de Hawkes observé sur l'intervalle $[0, T]$ est donnée par

$$\mathcal{V}_n(\theta) = \left[\prod_{i=1}^{p} \lambda^*(t_i)\right] \exp\left(-\int_0^T \lambda^*(u)\mathrm{d}u\right).$$

Il s'agit d'une vraisemblance conditionnelle, puisque les temps d'arrivée avant le temps 0 ne sont pas observés. Les estimateurs du maximum de vraisemblance pour les processus de Hawkes ont des bonnes propriétés asymptotiques : ils sont consistants et asymptotiquement gaussiens (Ogata, 1978).

3.4. Estimation du processus agrégé

Nous considérons ici que les temps d'arrivée ne sont pas observés ; au lieu de cela, le temps est découpé en intervalles réguliers correspondant par exemple à des jours ou des semaines et le nombre de temps d'arrivée dans chaque intervalle est compté. Les méthodes du maximum de vraisemblance ne sont plus applicables à ces données de comptage par intervalle, puisque la vraisemblance de la série de comptage ne peut pas être calculée explicitement.

L'estimation des paramètres du processus de Hawkes, en particulier de la fonction de reproduction, est alors un défi majeur. En effet, le phénomène d'agrégation des données, *i.e.* de compter les temps d'arrivée par intervalle au lieu de les observer directement, fait perdre une partie de l'information sur l'interaction entre les temps d'arrivée. Des méthodes naïves, qui supposeraient par exemple une distribution uni-

forme des temps d'arrivée à l'intérieur des intervalles, ignoreraient les interactions existantes et biaiseraient les résultats de l'estimation.

Kirchner (2016) a proposé une estimation non-paramétrique, en approximant la distribution de la série de comptage par un processus INAR(∞), et a montré que l'estimation conditionnelle par moindres carrés donne des estimateurs consistants et asymptotiquement normaux pour le processus de Hawkes sous-jacent lorsque la taille des intervalles tend vers zéro (Kirchner, 2017). Malheureusement, bien que ces estimateurs soient adaptés à la plupart des jeux de données pour lesquels la taille d'intervalle peut être choisie arbitrairement petite, ces estimations sont biaisées pour les jeux de données qui présentent des intervalles de taille importante.

Comme la vraisemblance de la série de comptage n'est pas calculable, on pourrait tenter un algorithme d'*Expectation Maximisation*, comme cela se fait pour les processus multivariés (Olson et Carley, 2013) ou lorsque l'intensité de l'immigration est un processus de renouvellement (Wheatley *et al.*, 2016), en considérant la structure de branchement — quels points sont des immigrants, et qui est le parent de chaque enfant — comme une donnée manquante. Pour un processus observé en temps discret, une approche analogue qui considérerait les temps d'arrivée comme des variables latentes n'est malheureusement pas adaptée, car il n'existe pas de forme explicite pour la distribution conditionnelle des temps d'arrivée compte tenu du nombre de temps d'arrivée par intervalle. Les algorithmes d'*Expectation Maximisation* stochastiques (Celeux *et al.*, 1995), qui permettent d'approcher cette distribution conditionnelle, ne résolvent pas le problème puisque les résultats de convergence habituels sont basés sur des vraisemblances de familles exponentielles (Delyon *et al.*, 1999), ce qui exclut les processus de Hawkes.

Après avoir introduit la notion de série de comptage pour désigner le processus agrégé, nous proposons une approche spectrale pour l'estimation des processus de Hawkes à partir de leur série de comptage en temps discret. Alors que la vraisemblance de la série de comptage n'est pas explicitement calculable, sa décomposition spectrale peut être directement reliée à celle du processus de Hawkes. Pour une introduction accessible et intuitive à l'analyse spectrale des séries temporelles, je recommande vivement Percival et Walden (2009).

Nous définissons ensuite un estimateur des paramètres du processus de Hawkes à partir de la série de comptage, en s'appuyant sur les travaux de Whittle (1952). Pour assurer que l'estimation est correcte, nous établissons d'abord une condition

de faible dépendance pour les processus de Hawkes linéaires. Cette condition, qui affirme que la covariance entre les temps d'arrivée décroit suffisamment vite lorsque l'intervalle de temps qui les sépare augmente, assure de bonnes propriétés asymptotiques pour l'estimateur de Whittle.

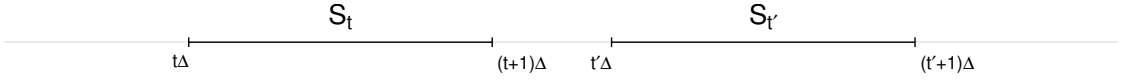
Enfin, nous illustrons cette approche par des expériences numériques et une étude de cas réel.

3.4.1. Série de comptage

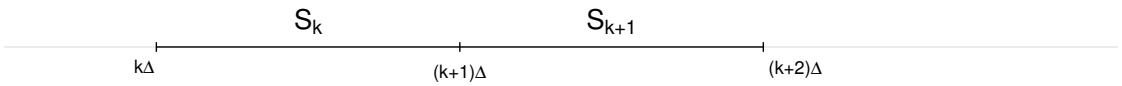
Nous nous intéressons aux séries temporelles générées par le comptage des temps d'arrivée du processus Hawkes, c'est-à-dire les séries temporelles obtenues en comptant le nombre de temps d'arrivée du processus sur des intervalles de longueur fixe. Nous donnons deux définitions de ces séries temporelles, selon que les extrémités des intervalles puissent prendre n'importe quelle valeur réelle, ou qu'elles soient limitées à une grille régulière (voir Figure 3.5) :

Définition. La série de comptage pour des intervalles de taille Δ associée à un processus ponctuel N est la série temporelle $(S_t)_{t\in\mathbb{R}} = \Big\{N\Big((t\Delta,(t+1)\Delta]\Big)\Big\}_{t\in\mathbb{R}}$ ou $(S_k)_{k\in\mathbb{Z}} = \Big\{N\Big((k\Delta,(k+1)\Delta]\Big)\Big\}_{k\in\mathbb{Z}}$, générée par la mesure de comptage sur des intervalles de taille Δ.

Notons que la connaissance de la série de comptage en temps continu $(S_t)_{t\in\mathbb{R}}$ est équivalente à la connaissance du processus ponctuel N. En effet, les temps d'arrivée du processus sont les temps s tels que $(S_{s/\Delta})$ est discontinu et $\lim_{t\to s^+} S_{t/\Delta} - \lim_{t\to s^-} S_{t/\Delta} = 1$. Au contraire, la connaissance de la série de comptage en temps discret $(S_k)_{k\in\mathbb{Z}}$ ne donne pas d'information sur les temps d'arrivée à l'intérieur de chaque intervalle $(k\Delta,(k+1)\Delta]$.



(a) En temps continu : $(S_t)_{t\in\mathbb{R}}$



(b) En temps discret : $(S_k)_{k\in\mathbb{Z}}$

FIGURE 3.5 – Série de comptage pour des intervalles de taille Δ.

3.4.2. Analyse spectrale des séries de comptage

Nous établissons la fonction de densité spectrale pour les séries de comptage de Hawkes à partir du spectre de Bartlett du processus. Le *spectre de Bartlett* d'un processus ponctuel stationnaire de second ordre N sur $\mathbb{R}$ est défini comme la mesure unique, positive, symétrique Γ telle que, pour toute fonction à décroissance rapide φ et ψ sur $\mathbb{R}$ (voir Daley et Vere-Jones, 2003, Proposition 8.2.I)

$$\mathrm{Cov}\Big(N(\varphi), N(\psi)\Big) = \int_{\mathbb{R}} \widetilde{\varphi}(\omega)\widetilde{\psi^*}(\omega)\Gamma(\mathrm{d}\omega), \tag{3.3}$$

où $\psi^*(u) = \psi(-u)$, et $\widetilde{\cdot}$ désigne la transformée de Fourier :

$$\widetilde{\varphi}(\omega) = \int_{\mathbb{R}} e^{-i\omega s}\varphi(s)\mathrm{d}s.$$

Intuitivement, le spectre de Bartlett est l'équivalent, pour un processus ponctuel, de la fonction de densité spectrale d'une série temporelle. En particulier, il s'agit de la mesure qui permet de décomposer la covariance du processus dans le domaine de Fourier, selon ses différentes composantes périodiques.

Pour le processus de Hawkes linéaire, le spectre de Bartlett admet une densité donnée par (Daley et Vere-Jones, 2003, Example 8.2(e))

$$\gamma(\omega) = \frac{m}{2\pi}\left|1 - \widetilde{h}(\omega)\right|^{-2} \tag{3.4}$$

où $m = \mathbb{E}\Big[N(0,1]\Big] = \eta\,(1 - \int_{\mathbb{R}} h(t)dt)^{-1}$.

Cela nous permet de calculer la fonction de densité spectrale de la série de comptage en temps continu (Figure 3.6a) :

Proposition 1. *Soit N un processus de Hawkes linéaire sur $\mathbb{R}$, et $\{S_t\}_{t\in\mathbb{R}} = \{N(t\Delta, (t+1)\Delta]\}_{t\in\mathbb{R}}$ sa série de comptage en temps continu associée. Alors S_t possède une densité spectrale donnée par la fonction*

$$f_S(\omega) = m\,\Delta\,\mathrm{sinc}^2\left(\frac{\omega}{2}\right)\left|1 - \widetilde{h}\left(\frac{\omega}{\Delta}\right)\right|^{-2}. \tag{3.5}$$

Démonstration. Soit $\varphi = \mathbb{1}_{(0,\Delta]}$ et $\psi = \mathbb{1}_{(\Delta u, \Delta(u+1)]}$. On a

$$\widetilde{\varphi}(\omega) = \int_0^{\Delta} e^{-i\omega s}\mathrm{d}s = \frac{i}{\omega}\left[e^{-i\omega\Delta} - 1\right],$$
$$\widetilde{\psi^*}(\omega) = \int_{-\Delta(u+1)}^{-\Delta u} e^{-i\omega s}\mathrm{d}s = \frac{i}{\omega}e^{i\omega\Delta u}\left[1 - e^{i\omega\Delta}\right].$$

Alors, en utilisant (3.3) et (3.4), la fonction d'autocovariance de S_t est donnée par

$$\begin{aligned}\gamma_S(u) &= \mathrm{Cov}(S_0, S_u)\\ &= \mathrm{Cov}\Big(N(\varphi), N(\psi)\Big)\\ &= \int_{\mathbb{R}} \frac{1}{\omega^2}\, e^{i\omega\Delta u} \left|e^{i\omega\Delta} - 1\right|^2 \Gamma(\mathrm{d}\omega)\\ &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{i\omega u} m\Delta\, \mathrm{sinc}^2\left(\frac{\omega}{2}\right)\left|1 - \widetilde{h}\left(\frac{\omega}{\Delta}\right)\right|^{-2} \mathrm{d}\omega.\end{aligned}$$

□

Pour la série de comptage en temps discret, il est nécessaire de prendre en compte les effets de recouvrement spectral (*aliasing* en anglais), qui replie les composantes de haute fréquence sur les basses fréquences (Figure 3.6b) :

Corollaire 1. *Soit N un processus de Hawkes linéaire sur $\mathbb{R}$, et $(S_k)_{k\in\mathbb{Z}} = \Big\{N\Big((k\Delta, (k+1)\Delta]\Big)\Big\}_{k\in\mathbb{Z}}$ sa série de comptage en temps discret. Alors S_k possède une densité spectrale donnée par la fonction*

$$f(\omega) = \sum_{k\in\mathbb{Z}} f_S(\omega + 2k\pi)$$

où $f_S(\cdot)$ est la fonction définie par (3.5).

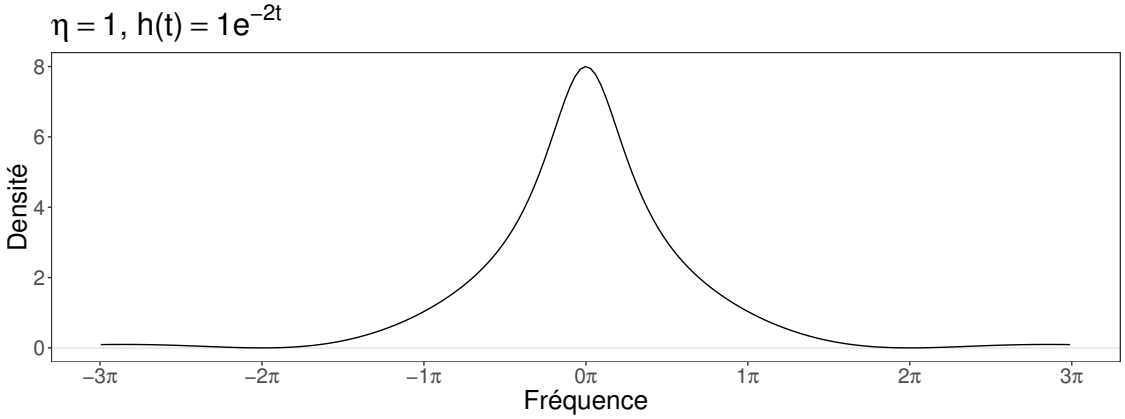
Démonstration. La densité spectrale et l'autocovariance de la série en temps discret forment une paire de Fourier, donnée par

$$f(\omega) = \sum_{k\in\mathbb{Z}} \gamma(k) e^{-i\omega k}, \qquad \gamma(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\omega) e^{i\omega k}\mathrm{d}\omega.$$

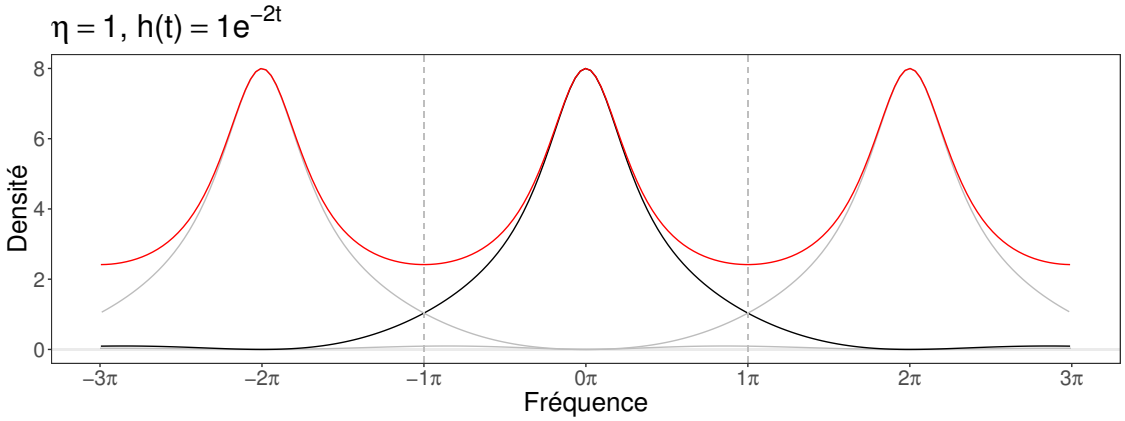
Pour tout $u \in \mathbb{Z}$, on a :

$$\begin{aligned}\mathrm{Cov}(S_0, S_u) &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{i\omega u} m\Delta\, \mathrm{sinc}^2\left(\frac{\omega}{2}\right)\left|1 - \widetilde{h}\left(\frac{\omega}{\Delta}\right)\right|^{-2} \mathrm{d}\omega\\ &= \frac{1}{2\pi}\sum_{k\in\mathbb{Z}}\int_{(2k-1)\pi}^{(2k+1)\pi} e^{i\omega u} m\Delta\, \mathrm{sinc}^2\left(\frac{\omega}{2}\right)\left|1 - \widetilde{h}\left(\frac{\omega}{\Delta}\right)\right|^{-2} \mathrm{d}\omega\\ &= \frac{1}{2\pi}\sum_{k\in\mathbb{Z}}\int_{-\pi}^{\pi} \underbrace{e^{i2k\pi u}}_{=1 \text{ since } u\in\mathbb{Z}} e^{i\omega u} m\Delta\, \mathrm{sinc}^2\left(\frac{\omega + 2k\pi}{2}\right)\left|1 - \widetilde{h}\left(\frac{\omega + 2k\pi}{\Delta}\right)\right|^{-2} \mathrm{d}\omega\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega u}\sum_{k\in\mathbb{Z}} f_S(\omega + 2k\pi)\mathrm{d}\omega.\end{aligned}$$

□



(a) Fonction de densité spectrale de la série de comptage en temps continue.



(b) Fonction de densité spectrale de la série de comptage en temps continu (courbe noire) et en temps discret (courbe rouge). Le recouvrement spectral replie les hautes fréquences sur les basses : la densité spectrale de la série de comptage en temps discret est égale à la superposition des densités spectrales de la série de comptage en temps continu décalées de $2k\pi$ (courbes grises).

FIGURE 3.6 – Fonction de densité spectrale pour les séries de comptage par intervalle de taille $\Delta = 1$ associées au processus de Hawkes exponentiel avec intensité d'immigration $\eta = 1$ et fonction de reproduction $h(t) = e^{-2t}$.

3.4.3. Propriétés de mélange fort

Rosenblatt (1956) a introduit le coefficient de mélange fort pour mesurer la dépendance entre des tribus, ce qui a suscité des décennies d'intérêt pour la théorie de la faible dépendance pour les séries temporelles et les champs aléatoires (voir Bradley (2005) pour un examen des conditions de mélange). Les conditions de mélange fournissent de très fortes inégalités de covariance et des méthodes de couplage

(Doukhan, 1994; Rio, 2017) pour obtenir des preuves de propriétés asymptotiques pour l'estimation de paramètres, à condition que les coefficients de mélange diminuent suffisamment vite. Cependant, ces coefficients sont formulés par rapport à des tribus riches et donc difficiles à contrôler même pour des modèles très simples. Pour cette raison pratique, les coefficients de mélange de régularité absolue sont souvent préférés car ils peuvent être facilement calculés pour les processus de Markov et les fonctions de processus de Markov (Davydov, 1974).

Westcott (1972) a étendu la définition du mélange aux processus ponctuels, prouvant par exemple que les processus d'agrégats de Poisson sont mélangeants au sens ergodique (Westcott, 1971). Cependant, en l'absence d'informations précises sur les coefficients de mélange fort, le cadre de dépendance faible n'a pas conduit à un développement statistique important dans la modélisation des processus ponctuels. Des travaux récents ont porté sur le calcul de coefficients de mélange forts pour certaines classes de processus ponctuels (Heinrich et Pawlas, 2013; Poinas *et al.*, 2019), en s'appuyant sur les résultats des séries temporelles et des champs aléatoires et en utilisant le fait que les tribus générées par des ensembles dénombrables sont plus pauvres que celles générées par des ensembles continus.

Dans ce contexte, nous considérons les propriétés de mélange pour les processus de Hawkes et établissons comme résultat une condition de mélange forte avec un taux de décroissance polynomial. Comme les processus de Hawkes avec fonction de reproduction exponentielle sont des processus markoviens déterministes par morceaux (Oakes, 1975), on pourrait espérer calculer des coefficients de mélange de régularité absolue. Cependant, comme cela ne s'étendrait pas aux autres fonctions de reproduction, nous établissons à la place une condition de mélange fort qui s'applique à toute classe de fonction, à condition qu'ils aient un moment fini d'ordre $1+\delta$, $\delta > 0$.

On rappelle que, pour un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$ et $\mathcal{A}, \mathcal{B}$ deux sous-tribus de $\mathcal{F}$, le coefficient de mélange fort de Rosenblatt est défini comme la mesure de dépendance entre $\mathcal{A}$ et $\mathcal{B}$ (Rosenblatt, 1956) :

$$\alpha(\mathcal{A}, \mathcal{B}) \coloneqq \sup\Big\{ |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{A}, B \in \mathcal{B} \Big\}.$$

Cette définition peut être adaptée à un processus ponctuel N sur $\mathbb{R}$, en définissant (voir Poinas *et al.*, 2019)

$$\alpha_N(r) \coloneqq \sup_{t \in \mathbb{R}} \alpha\Big(\mathcal{E}^t_{-\infty}, \mathcal{E}^\infty_{t+r}\Big),$$

où $\mathcal{E}_a^b$ représente la tribu générée par les événements cylindriques sur l'intervalle $(a, b]$, c'est-à-dire les événements $\mathcal{A}_{B,m}$ donnés par la relation (3.2), où $B \subset (a, b]$. Pour la suite correspondante $(S_k)_{k\in\mathbb{Z}}$, le coefficient de mélange fort prend la forme

$$\alpha_X(r) \coloneqq \sup_{n\in\mathbb{Z}} \alpha\Big(\mathcal{F}_{-\infty}^n, \mathcal{F}_{n+r}^\infty\Big),$$

où $\mathcal{F}_a^b$ représente la tribu générée par $(S_k)_{a\le k\le b}$.

On dit que le processus ponctuel N (resp. la séquence (S_k)) est fortement mélangeant si $\alpha_N(r)$ (resp. $\alpha_X(r)$) $\to 0$ lorsque $r \to \infty$. Intuitivement, la condition de mélange fort signifie que la dépendance entre les événements passés et futurs diminue uniformément jusqu'à zéro à mesure que l'écart de temps entre eux augmente. Notons que, puisque $\mathcal{F}_a^b \subset \mathcal{E}\Big((a,b]\Big)$, $\alpha_X(r) \le \alpha_N(r)$ pour tout r.

Théorème 1. *Soit N un processus de Hawkes linéaire sur $\mathbb{R}$ avec fonction de reproduction $h = \mu h^*$, où $\mu = \int_{\mathbb{R}} h < 1$ et $\int_{\mathbb{R}} h^* = 1$. Supposons qu'il existe $\delta > 0$ tel que le moment d'ordre $1 + \delta$ du noyau de reproduction h^* est fini :*

$$\nu_{1+\delta} \coloneqq \int_{\mathbb{R}} t^{1+\delta} h^*(t)\mathrm{d}t < \infty.$$

Alors N est fortement mélangeant et

$$\alpha_N(r) = \mathcal{O}\left(r^{-\delta}\right).$$

En résumé, la preuve comporte deux parties : premièrement, nous ramenons le problème à un seul arbre de Galton-Watson à temps continu en utilisant la représentation du processus de Hawkes sous forme de processus de branchement ; deuxièmement, nous établissons une borne supérieure pour les coefficients de mélange fort de l'arbre. Pour établir cette borne supérieure, on utilise le fait que le processus de Galton–Watson s'éteint presque sûrement et que le noyau de reproduction h^* a un moment fini ; la probabilité qu'il existe un descendant de génération k à une grande distance de l'immigrant tend rapidement vers 0 lorsque k augmente. La preuve est détaillée dans l'Annexe 3.B.

Enfin, comme conséquence immédiate du Théorème 1, nous obtenons le corollaire suivant pour les séries de comptage de Hawkes :

Corollaire 2. *Soit N un processus de Hawkes comme dans le Théorème 1, et $(S_k)_{k\in\mathbb{Z}} = \Big\{N\Big((k\Delta, (k+1)\Delta]\Big)\Big\}_{k\in\mathbb{Z}}$ sa série de comptage associée. Alors, (S_k) est fortement mélangeant et*

$$\alpha_X(r) = \mathcal{O}\left(r^{-\delta}\right).$$

3.4.4. Estimation paramétrique des séries de comptage

Pour une série temporelle linéaire stationnaire $(S_k)_{k\in\mathbb{Z}}$ avec fonction de densité spectrale $f_\theta(\cdot)$, et θ un vecteur de paramètres inconnu, Hosoya (1974) et Dzhaparidze (1974), s'appuyant sur les travaux de Whittle (1952), proposèrent comme estimateur de θ la statistique

$$\widehat{\theta}_n = \arg\min_{\theta\in\Theta} \mathcal{L}_n(\theta) \tag{3.6}$$

où

$$\mathcal{L}_n(\theta) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left(\log f_\theta(\omega) + \frac{I_n(\omega)}{f_\theta(\omega)}\right)\mathrm{d}\omega \tag{3.7}$$

est la log-vraisemblance spectrale de la série temporelle, et $I_n(\omega) = (2\pi n)^{-1}\left|\sum_{k=1}^n S_k\, e^{-ik\omega}\right|^2$ est le périodogramme calculé à partir des réalisations $(S_k)_{1\leq k\leq n}$. Ils établirent également les propriétés asymptotiques de l'estimateur sous des conditions de régularité appropriées.

Dzhaparidze (1986) étendit ces résultats à des cas plus généraux, et en particulier à des processus stationnaires vérifiant les conditions de mélange de Rosenblatt. Les conditions et théorèmes suivants sont donc des adaptations de ceux trouvés dans Dzhaparidze (1986, Theorem II.7.1 et II.7.2) pour les séries de comptage associées aux processus de Hawkes.

Théorème 2. *Soit N un processus de Hawkes sur $\mathbb{R}$ avec fonction de reproduction $h = \mu h^*$, où $\mu = \int_{\mathbb{R}} h < 1$ et $\int_{\mathbb{R}} h^* = 1$, et $(S_k)_{k\in\mathbb{Z}} = \big(N(k, k+1]\big)_{k\in\mathbb{Z}}$ sa série de comptage en temps discret associée, avec fonction de densité spectrale f_θ. Supposons que les conditions de régularité suivantes sont vérifiée :*

(A1) Le vrai paramètre θ_0 appartient à un ensemble compact Θ de $\mathbb{R}^p$.

(A2) Pour tous $\theta_1 \neq \theta_2$ dans Θ, alors $f_{\theta_1} \neq f_{\theta_2}$ pour tout ω.

(A3) La fonction f_θ^{-1} est dérivable par rapport à θ et ses dérivées partielles $(\partial/\partial\theta_k)f_\theta^{-1}$ sont continues en $\theta \in \Theta$ et $-\pi \leq \omega \leq \pi$.

De plus, supposons qu'il existe un $\delta > 0$ tel que le noyau de reproduction h^* a un moment fini d'ordre $2+\delta$. Alors l'estimateur $\widehat{\theta}_n$ défini par (3.6) *(avec $\mathcal{L}_n(\theta)$ donné par* (3.7)*), est consistant, i.e. $\widehat{\theta}_n \to \theta_0$ en probabilité.*

Démonstration. La seule condition de Dzhaparidze (1986, Théorème II.7.1) que nous devons vérifier est qu'il existe un $\gamma > 2$ tel que $\mathbb{E}[|S_k|^{2\gamma}]$ soit fini et que l'inégalité suivante soit vérifiée :

$$\sum_{r=1}^{\infty}\big(\alpha_X(r)\big)^{1-2/\gamma} < \infty. \tag{3.8}$$

Puisque le processus de Hawkes linéaire admet des moments exponentiels finis si h^* a un moment d'ordre $\delta \in (0,1]$ (Roueff *et al.*, 2016, Théorème 4), $\mathbb{E}[|S_k|^{2\gamma}]$ est fini pour tout γ. Alors, en utilisant le corollaire 2, il existe toujours un $\gamma > 2$ qui satisfait (3.8). □

Soit Γ_θ la matrice définie par la relation :

$$\Gamma_\theta = \left(\frac{1}{4\pi} \int_{-\pi}^{\pi} \frac{\partial}{\partial \theta_k} \log f_\theta(\omega) \, \frac{\partial}{\partial \theta_l} \log f_\theta(\omega) \, \mathrm{d}\omega \right)_{1 \le k,l \le p}.$$

Si la série temporelle (S_k) était gaussienne, cette matrice serait en fait la limite, lorsque $n \to \infty$ de la matrice d'information de Fisher (Dzhaparidze, 1986, Section II.2.2). Ici, comme (S_k) n'est pas gaussien, les propriétés asymptotiques de l'estimateur de Whittle dépendent des statistiques du quatrième ordre de la série de comptage et nous définissons donc la matrice suivante :

$$C_{4,\theta} = \left(\frac{1}{8\pi} \int \int_{-\pi}^{\pi} f_{4,\theta}(\omega_1, -\omega_1, -\omega_2) \, \frac{\partial}{\partial \theta_k} \frac{1}{f_\theta(\omega_1)} \, \frac{\partial}{\partial \theta_l} \frac{1}{f_\theta(\omega_2)} \, \mathrm{d}\omega_1 \mathrm{d}\omega_2 \right)_{1 \le k,l \le p}$$

où $f_{4,\theta}(\cdot,\cdot,\cdot)$ est la densité du cumulant spectral de quatrième ordre de la série de comptage. Nous avons le résultat suivant :

Théorème 3. *Soit N un processus de Hawkes linéaire comme dans le Théorème 2, et $(S_k)_{k\in\mathbb{Z}} = \big(N(k\Delta,(k+1)\Delta]\big)_{k\in\mathbb{Z}}$ sa série de comptage en temps discret associée, avec fonction de densité spectrale f_θ. Supposons que les conditions (A1), (A2), (A3) sont vérifiées, et que :*

(A4) La fonction f_θ est deux fois dérivable par rapport à θ et ses dérivées partielles du second degré $(\partial^2/\partial\theta_k\partial\theta_l) f_\theta$ sont continues en $\theta \in \Theta$ et $-\pi \le \omega \le \pi$.

Alors l'estimateur $\widehat{\theta}_n$ est asymptotiquement normal et

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \underset{n\to\infty}{\sim} \mathcal{N}\left(0, \Gamma_{\theta_0}^{-1} + \Gamma_{\theta_0}^{-1} C_{4,\theta_0} \Gamma_{\theta_0}^{-1}\right).$$

Remarque. Le calcul de l'intégrale du cumulant spectral d'ordre quatre dans C_{4,θ_0} n'est pas trivial. Nous renvoyons aux travaux de Shao (2010) pour une méthode élégante permettant de calculer un estimateur de cette intégrale.

Remarque. Une extension directe des résultats prouvés dans cette section concerne les processus de Hawkes non causaux, pour lesquels le noyau de reproduction h^* peut prendre des valeurs positives sur $\mathbb{R}_{\le 0}$. Ces processus ne peuvent être définis en fonction de leur intensité conditionnelle, puisque les temps d'arrivée dépendent du

futur. On les définira plutôt à partir de la représentation sous forme de processus de branchement présentée dans la Section 3.3.2, représentation qui n'est pas invalidée par le choix d'un noyau de reproduction non causal : les immigrants et leurs descendants peuvent alors générer des points dans le futur, mais aussi dans le passé. Dans ce cadre, tous les résultats établis sont directement applicables aux processus de Hawkes non causaux, ou peuvent être étendus sans effort.

3.5. Étude de simulation

Nous illustrons la procédure d'estimation et les propriétés asymptotiques de l'approche spectrale pour les séries de comptage des processus de Hawkes. Pour mettre en évidence les différents théorèmes des sections précédentes, nous considérons deux noyaux h^* pour la fonction de reproduction : le noyau exponentiel pour lequel tous les moments existent et le noyau de Pareto dont les moments supérieurs ne sont pas finis.

Les simulations et estimations suivantes ont été réalisées avec notre librairie *hawkesbow*, disponible librement (`https://github.com/fcheysson/hawkesbow`), écrite à la fois en R (R Core Team, 2019) et en C++ en utilisant Rcpp (Eddelbuettel et François, 2011).

3.5.1. Procédure de simulation

Noyau exponentiel Nous considérons d'abord un processus de Hawkes linéaire avec une fonction de reproduction exponentielle :

$$\lambda(t) = \eta + \mu \int \beta e^{-\beta(t-u)} \mathrm{d}N(u),$$

i.e. avec noyau de reproduction $h^*(t) = \beta e^{-\beta t}$ pour $t \geq 0$. Notons que le processus vérifie les conditions des deux Théorèmes 2 et 3.

En utilisant la représentation de branchement du processus de Hawkes, nous avons simulé 1000 réalisations du processus sur l'intervalle $[0, T]$ avec les valeurs des paramètres $\eta = 1$, $\mu = 0{,}5$ et $\beta = 1$. Pour chacune des simulations, nous avons créé quatre séries temporelles en comptant le nombre de temps d'arrivée dans des intervalles de taille $\Delta = 0{,}25$, $0{,}5$, 1 ou 2 respectivement. Nous avons ensuite estimé les paramètres η, μ et β comme dans la Section 3.4.4 pour chacune des quatre séries temporelles. Nous avons comparé ces estimations aux estimations habituelles

du maximum de vraisemblance (Figure 3.7). Comme ces dernières utilisent toute l'information disponible sur les temps d'arrivée, elles sont meilleures que toute estimation basée sur les séries de comptage par intervalle, et fournissent donc un scénario optimal pour les estimations de Whittle lorsque la taille des intervalles tend vers 0. Pour le noyau exponentiel, un ensemble de 1000 simulations et les estimations de Whittle associées, avec $T = 1000$ et $\Delta = 1$, prend environ 4 minutes sur un ordinateur portable équipé d'un processeur Intel i5.

Noyau de Pareto Nous considérons maintenant un processus de Hawkes linéaire avec un noyau de reproduction de Pareto : $h^*_\gamma(t) = \gamma a^\gamma t^{-\gamma-1}$ pour $t \geq a$. Nous rappelons que les moments d'une distribution de Pareto sont tous finis jusqu'à, mais sans inclure, l'ordre γ. Nous illustrons les théorèmes des sections précédentes en considérant trois cas pour le paramètre de forme γ, chacun satisfaisant davantage les hypothèses nécessaires : *(i)* $\gamma = 1$, l'espérance de h^* est infinie et le processus ne satisfait pas la condition du théorème 1 ; *(ii)* $\gamma = 2$, le processus est fortement mélangeant, mais la variance est infinie et le processus ne satisfait pas les hypothèses du Théorème 2 ; *(iii)* $\gamma = 3$, le processus est fortement mélangeant et satisfait les hypothèses des Théorèmes 2 et 3, mais les moments d'ordre 3 et plus n'existent pas.

Comme pour le noyau exponentiel, nous avons simulé 1000 simulations du processus de Hawkes pour chaque $\gamma \in \{1, 2, 3\}$, avec des valeurs de paramètres $\eta = 1$, $\mu = 0{,}5$, et $a_3 = 2/3$ pour $\gamma = 3$, $a_2 = 1/2$ pour $\gamma = 2$, de sorte que les noyaux de Pareto h^*_3 et h^*_2 et le noyau exponentiel possèdent la même espérance. Pour le noyau de Pareto h^*_1, nous avons choisi $a_1 = 1/3$ arbitrairement. Nous n'avons pas pu comparer les estimations de Whittle à celles du maximum de vraisemblance, car ces dernières étaient trop gourmande d'un point de vue computationnel, la fonction de vraisemblance contenant un grand nombre de points de discontinuité par rapport au paramètre de position du noyau a : $p(p-1)/2$ points de discontinuité, avec p le nombre de temps d'arrivée du processus.Les estimations pour le noyau de Pareto sont situées dans l'Annexe 3.A. Pour le noyau de Pareto, un ensemble de 1000 simulations et les estimations de Whittle associées, avec $T = 1000$ et $\Delta = 1$, prend environ 14 minutes sur un ordinateur portable équipé d'un processeur Intel i5.

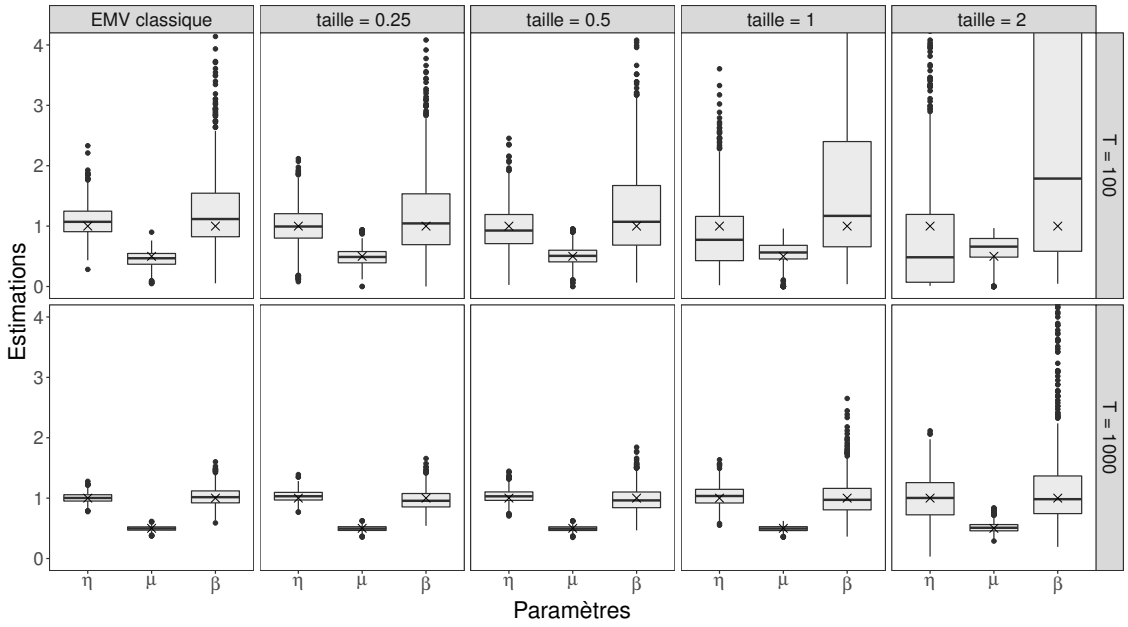


FIGURE 3.7 – Estimations des paramètres η, μ et β pour 1000 simulations du processus de Hawkes linéaire avec noyau de reproduction $h^*(t) = \beta e^{-\beta t}$ sur l'intervalle $[0, T]$. Les vraies valeurs des paramètres (croix) sont : $\eta = 1$, $\mu = 0{,}5$, $\beta = 1$. La colonne de gauche correspond aux estimations du maximum de vraisemblance. Les autres colonnes correspondent aux estimations de Whittle, selon les différentes tailles d'intervalle.

3.5.2. Résultats et interprétation

Noyau exponentiel Pour $T = 100$ et Δ petit, l'estimateur de Whittle est presque aussi bon que l'estimateur du maximum de vraisemblance. En revanche, l'estimation se détériore massivement pour des tailles d'intervalle plus élevées, notamment pour le paramètre d'intensité du noyau exponentiel β. Cela s'explique par le fait que des tailles d'intervalle élevées par rapport à l'échelle du noyau de reproduction rendent difficile la détection des interactions entre les temps d'arrivée. Cette difficulté peut être évaluée par la probabilité qu'un temps d'arrivée dans un intervalle ait un enfant dans le même intervalle : en supposant que le processus est stationnaire, cette probabilité est égale à $\Delta^{-1} \int_0^\Delta \int_u^\Delta \beta e^{-\beta(t-u)} \mathrm{d}t\mathrm{d}u = 1 - (\beta\Delta)^{-1}(1 - e^{-\beta\Delta})$. Par exemple, avec $\beta = 1$ et $\Delta = 2$, on obtient une probabilité de 0,57, *i.e.* 57% de l'information concernant l'interaction entre les temps d'arrivée du processus de Hawkes est localisée à l'intérieur des intervalles et donc inaccessible à l'observation. Heureusement, en augmentant T, les propriétés asymptotiques garantissent que les

estimations de Whittle s'améliorent, même pour des intervalles de taille élevée.

Pour illustrer davantage les propriétés asymptotiques de l'estimation, notamment sa vitesse de convergence, nous calculons l'erreur quadratique moyenne, définie par EQM $= K^{-1}\sum(\widehat{\theta}_n - \theta_0)^2$, pour les estimations de chaque ensemble de $K = 1000$ simulations pour T et Δ donnés (Figure 3.8). Lorsque T est grand, la pente de l'erreur quadratique moyenne par rapport à T atteint -1 (en échelle log-log) pour tous les paramètres et presque toutes les tailles d'intervalle, ce qui illustre le taux de convergence $\mathcal{O}(n^{-1})$ énoncé dans le Théorème 3. Lorsque T est petit, les estimations de l'intensité d'immigration η et du taux de reproduction μ semblent déjà avoir atteint le taux de convergence optimal, tandis que l'erreur quadratique moyenne pour le paramètre d'intensité β du noyau exponentiel est jusqu'à un ordre de grandeur et demi plus élevé que ce à quoi on pourrait s'attendre en extrapolant l'erreur quadratique moyenne pour T grand. Enfin, il convient de noter que, pour des tailles d'intervalle raisonnables ($\Delta \leq 1$), les estimations de Whittle du taux de reproduction μ ont une erreur quadratique moyenne similaire à celles du maximum de vraisemblance.

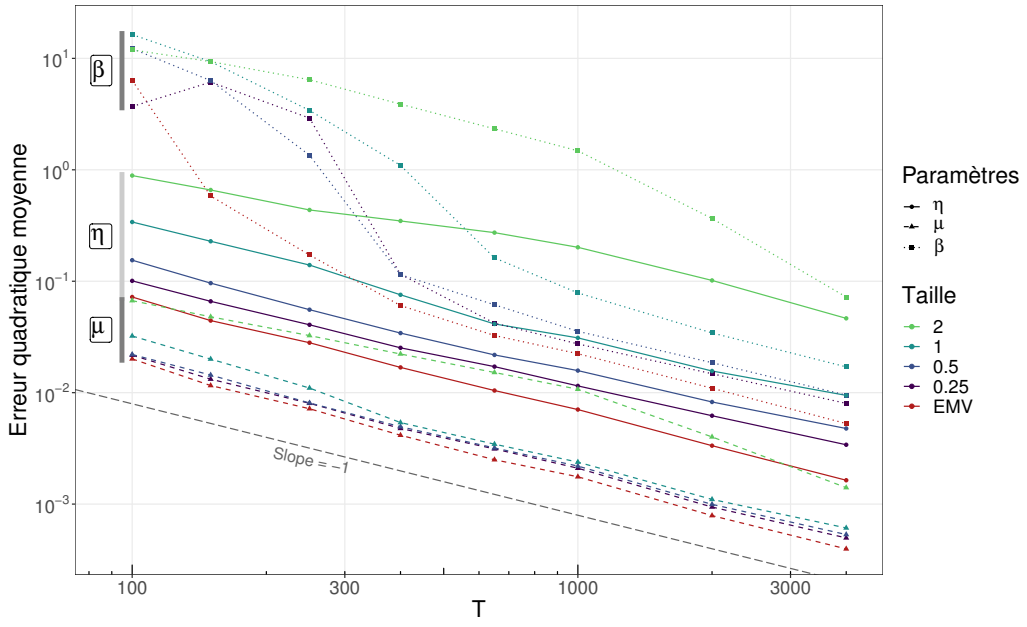


FIGURE 3.8 – Erreur quadratique moyenne pour les estimations des paramètres η, μ et β pour 1000 simulations du processus de Hawkes linéaire avec noyau de reproduction $h^*(t) = \beta e^{-\beta t}$ sur l'intervalle $[0, T]$, à l'échelle log-log. La droite grise pointillée représente la pente idéale de -1, *i.e.* une vitesse de convergence en $\mathcal{O}(n^{-1})$.

Noyau de Pareto Les performances des estimations ponctuelles sont étonnamment similaires pour toutes les valeurs de γ. Tant l'intensité d'immigration η que le taux de reproduction μ présentent la vitesse optimale de convergence $\mathcal{O}(n^{-1})$ pour tous les T considérés et presque toutes les tailles d'intervalle. D'autre part, les estimations pour le paramètre de position a du noyau de Pareto montrent un comportement singulier. Alors que pour les tailles d'intervalles 0,5, 1 et 2, l'erreur quadratique moyenne par rapport à T atteint asymptotiquement la pente idéale de -1 (bien qu'avec un ordre de grandeur entre $\Delta = 0{,}5$ et $\Delta = 1$, et un autre entre $\Delta = 1$ et $\Delta = 2$), elle ne semble pas avoir atteint un régime asymptotique similaire pour la taille d'intervalle 0,25, qui présente une erreur quadratique moyenne presque similaire à $\Delta = 2$. Nous ne sommes pas en mesure d'expliquer ce comportement.

Il est intéressant de noter que les estimations ponctuelles présentent de bons comportements asymptotiques pour toutes les valeurs de γ, sauf pour la taille d'intervalle $\Delta = 0{,}25$, même si les noyaux de Pareto h_2^* et h_1^* ne vérifient pas les hypothèses des Théorèmes 2 et 3, ce qui suggère que la condition sur les moments du noyau de reproduction dans le Théorème 1 est trop restrictive. Néanmoins, elle est suffisamment faible pour que l'approche spectrale que nous avons développée puisse être utile pour des applications dans de nombreuses disciplines.

3.6. Étude de cas : transmission de la rougeole à Tokyo

La rougeole est une maladie virale très contagieuse, principalement transmise par des gouttelettes et caractérisée par une éruption cutanée accompagnée d'une fièvre. Malgré les efforts déployés dans le monde entier pour éradiquer la maladie, elle est réapparue dans les pays développés, principalement par le biais de cas importés et de personnes non vaccinées, générant des foyers mineurs d'infection. Au Japon, la rougeole est une maladie à déclaration obligatoire : tous les cas diagnostiqués doivent être signalés au gouvernement, puis faire l'objet d'une enquête afin de contenir les foyers potentiels.

L'Institut National Japonais des Maladies Infectieuses publie des rapports hebdomadaires ainsi que des tableaux de données de surveillance pour toutes les maladies à déclaration obligatoire. [1] Nous considérons ici le nombre de cas de rougeole

1. `https://www.niid.go.jp/niid/en/survaillance-data-table-english.html`, dernière

dans la préfecture de Tokyo, d'août 2012 à février 2020 (Figure 3.9). Nous modélisons les données de comptage hebdomadaires en utilisant un processus de Hawkes linéaire avec un noyau de reproduction gaussien :

$$h^*(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\nu)^2}{2\sigma^2}\right),$$

puis estimons les paramètres η, μ, ν et σ comme dans la Section 3.4.4. Nous considérons le processus comme stationnaire car l'impact de la saisonnalité est faible par rapport à la variabilité temporelle observée.

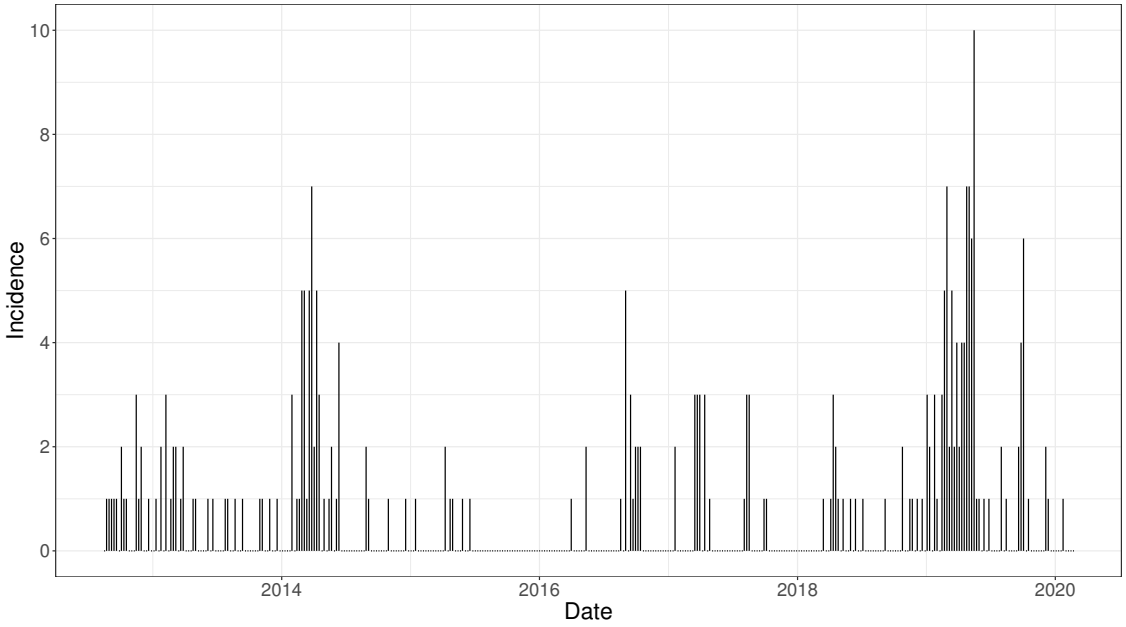


FIGURE 3.9 – Nombre hebdomadaire des cas de rougeole à Tokyo. Entre la troisième semaine d'août 2012 et la troisième semaine de février 2020, 264 cas de rougeole ont été déclarés dans la préfecture de Tokyo.

Pour le noyau de reproduction gaussien, on trouve $\widehat{\nu} = 9{,}8$ jours et $\widehat{\sigma} = 5{,}9$ jours, ce qui correspond à un écart interquartile de 7,9 jours. Ces estimations peuvent être reliées aux caractéristiques cliniques du virus : la période d'incubation de la rougeole est en moyenne de 10 à 12 jours, et la transmission se produit généralement de 4 jours avant à 4 jours après l'apparition de l'éruption cutanée (Centers for Disease Control and Prevention, 2015). Pour l'intensité d'immigration et le taux de reproduction, on trouve $\widehat{\eta} = 0{,}040$ jour^{-1} et $\widehat{\mu} = 0{,}72$. On constate que les cas dont la source de transmission est inconnue (*i.e.* les immigrants du modèle) représentent

visite le 01/09/2020.

$1 - \widehat{\mu} = 28\%$ de tous les cas de rougeole, une valeur proche des données trouvées dans (Nishiura *et al.*, 2017, Figure 3), qui rapporte 23 cas importés parmi 106 événements contagieux au Japon, 2016.

3.7. Discussion

Dans ce travail, nous avons établi une condition de mélange fort avec un taux de décroissance polynomial pour les processus de Hawkes linéaires, puis nous avons proposé une procédure d'estimation de Whittle à partir de leurs données de comptage. À notre connaissance, il s'agit du premier travail qui étudie les conditions de mélange fort pour l'estimation des processus de Hawkes. Cette approche présente des caractéristiques intéressantes : *(i)* elle a de bonnes propriétés asymptotiques, similaires à la méthode du maximum de vraisemblance ; *(ii)* elle est facile à mettre en œuvre et flexible, puisque la seule valeur spécifiée par l'utilisateur est la transformée de Fourier $\widetilde{h}$ du noyau de reproduction h^* ; *(iii)* elle est efficace d'un point de vue computationnel, avec une complexité en $\mathcal{O}(n \log n)$, n le nombre d'intervalles considérés, due au calcul du périodogramme par une transformée de Fourier rapide, par rapport à $\mathcal{O}(p^2)$, p le nombre de temps d'arrivée, pour la méthode du maximum de vraisemblance (sauf lorsque le noyau est exponentiel, auquel cas la complexité est réduite à $\mathcal{O}(p)$ avec un minimum d'efforts (Ozaki et Ogata, 1979), ce qui la rend plus efficace que notre approche) ; *(iv)* elle est particulièrement bien adaptée aux applications où la taille des intervalles ne peut être choisie arbitrairement.

Nous nous attendons à ce que les résultats soient également vrais pour les processus de Hawkes multivariés, avec des modifications minimales. En effet, les propriétés de mélange fort ont été obtenues en utilisant certaines propriétés de l'arbre Galton–Watson qui s'étendent au cas multitype. En outre, l'analyse spectrale des processus de comptage de Hawkes peut être directement étendue au cas multivarié, en utilisant les résultats de Daley et Vere-Jones (2003, Exemple 8.3(c)) sur le spectre multivarié de Bartlett des processus ponctuels mutuellement excitants. Néanmoins, nous avons décidé de nous concentrer sur le cas univarié par souci de concision et de clarté.

Un intérêt majeur du cas multivarié pour la modélisation des maladies contagieuses est de pouvoir faire la distinction entre plusieurs groupes à risque. Dans ce cadre, les événements d'infection de chaque groupe est modélisé par un processus

de Hawkes, et ces processus peuvent interagir *via* des fonctions de reproduction croisées. L'estimation de ces fonctions de reproduction croisées donne alors une information sur la transmission de la maladie, et donc sur les contacts, entre les différents groupes. Les processus de Hawkes seraient donc des outils utiles dans le cas où les contacts entre groupes sont mal connus.

Les résultats asymptotiques établis dans ce travail s'inscrivent dans une démarche classique de preuves pour les séries temporelles. En général, les résultats sont d'abord établis dans un cadre d'indépendance (ici, un processus pour lequel il n'existe pas d'interaction entre les points est le processus de Poisson), puis étendus dans un cadre de faible dépendance pour les processus stationnaires. À cet égard, la méthode d'estimation proposée pour les séries de comptage est au processus de Hawkes ce que la régression de Poisson est au processus de Poisson. Une troisième étape de la démarche statistique concerne l'établissement de résultats asymptotiques en relâchant l'hypothèse de stationnarité : il faut alors définir une classe plus large de processus quasi-stationnaires.

Le développement des outils statistiques présentés a pris plus de temps que prévu. Des pistes d'extension des résultats asymptotiques pour les processus de Hawkes-non stationnaires, essentiels pour pouvoir être plus largement applicables aux jeux de données rencontrés en épidémiologie, sont proposées en conclusion de ce manuscrit. Ces extensions sont nécessaires pour offrir des applications pratiques satisfaisantes pour la recherche en santé publique, puisque l'hypothèse de stationnarité n'est pas vérifiée pour l'incidence de la plupart des maladies contagieuses.

Conclusions et perspectives

Dans ce travail de thèse, nous nous sommes intéressés à certaines difficultés méthodologiques inhérentes à l'utilisation de données agrégées. Nous avons précisé un cadre pour la modélisation et l'estimation de la fraction d'événements de santé attribuables à un facteur de risque à partir de séries temporelles. Nous avons également établi les lois asymptotiques des estimateurs pour les modèles souvent rencontrés dans la littérature de santé publique. À travers une étude de simulation, nous avons montré que l'estimation de la fraction attribuable est sensible au choix du modèle : il doit être motivé par la connaissance épidémiologique et l'observation statistique des données. En particulier, il est nécessaire d'évaluer s'il existe un décalage temporel entre le facteur de risque et l'événement de santé. En outre, j'ai montré que la régression de Serfling n'est pas un bon modèle pour l'estimation de la fraction attribuable.

Les estimateurs proposés ont permis d'étudier l'impact de pathologies infectieuses hivernales sur la consommation d'antibiotiques. La campagne de sensibilisation "Les antibiotiques, c'est pas automatique !" a permis de réduire le nombre de remboursement d'antibiotiques attribuables aux épidémies de syndromes grippaux de plus de moitié entre les années 2002 et 2010. Les infections respiratoires basses virales contribueraient pour 17% de l'utilisation globale d'antibiotiques en ville à l'échelle de la population générale, soit 289 prescriptions pour 100 000 habitants par semaine, et jusqu'à 38% chez les enfants de 5 ans et moins, soit 1588 prescriptions pour 100 000 habitants par semaine dont la moitié attribuables aux bronchiolites.

Enfin, nous avons proposé une approche spectrale de l'estimation des processus de Hawkes à partir de données agrégées, et montré son intérêt potentiel pour l'étude des maladies contagieuses. Nous avons montré des propriétés de faible dépendance pour le processus, permettant d'établir la consistance et la normalité asymptotique des estimateurs proposés.

Toutefois, pour la recherche en santé publique, il est nécessaire de développer davantage ces outils statistiques, dans la mesure où les paramètres descriptifs du processus de Hawkes stationnaire sont constants au cours du temps, ce qui n'est pas vérifié dans la plupart des cadres d'étude concernant les maladies contagieuses. Par

exemple, la bronchiolite, maladie contagieuse dont le taux d'incidence en France métropolitaine entre 2010 et 2017 est présenté dans le Chapitre 2, présente une saisonnalité marquée. En faisant l'hypothèse que les paramètres varient suffisamment lentement, voire sont constants sur des périodes fixées, ils pourraient être estimés sur chacune de ces périodes à partir des résultats établis dans le cas stationnaire (Kumazawa et Ogata, 2014). Par exemple, dans le cas des maladies contagieuses, si l'on considère que les paramètres ne diffèrent qu'entre la période estivale et la période hivernale, il serait possible d'ajuster deux processus de Hawkes selon la période.

Cette méthode peut donner des résultats heuristiques intéressants, mais ne permet pas de garantir la précision de l'approximation asymptotique des résultats obtenus. De plus, dans l'optique d'identifier des associations entre la diffusion de maladies contagieuses et des facteurs de risque (p. ex. liens environnementaux, ou avec d'autres maladies infectieuses), inclure des variables explicatives dans la paramétrisation du modèle va à l'encontre de l'hypothèse de stationnarité. Pour ces raisons, un développement méthodologique supplémentaire est nécessaire pour définir de manière univoque un cadre de quasi-stationnarité dans lequel les résultats de consistance et de normalité asymptotique de l'estimation seront préservés. Une version non-stationnaire du processus de Hawkes a été introduite par Chen et Hall (2013) en laissant l'intensité d'immigration dépendre du temps. Roueff *et al.* (2016) ont proposé une version plus générale, pour laquelle la fonction de reproduction est aussi autorisée à dépendre du temps. Pour inclure des variables explicatives dans cette formulation de l'intensité conditionnelle, il est possible d'en faire dépendre les paramètres du processus de Hawkes (voir p. ex. les travaux de Meyer *et al.*, 2012). Ainsi, la paramétrisation de l'intensité d'immigration permet de faire évoluer avec le temps la fréquence d'apparition des clusters, celle du taux de reproduction le nombre moyen de personnes qu'un individu contagieux peut infecter, et celle du noyau de reproduction la durée de contagiosité de la maladie.

L'utilisation de processus de Hawkes non-stationnaire pose cependant la question de l'estimation des paramètres dans un cadre asymptotique inhabituel. En effet, pour les processus stationnaires, les propriétés des estimateurs sont généralement établies dans le cas asymptotique où la fenêtre d'observation du processus (ou le nombre d'intervalles de temps observés dans le cas des séries de comptage) tend vers l'infini. Ce cadre n'est pas adapté aux processus non-stationnaires, puisque

des observations futures ne renseignent pas sur les valeurs des paramètres générant les observations présentes. Il est donc nécessaire de redéfinir le cadre asymptotique utilisé, d'adapter la méthode d'estimation et de démontrer de nouveaux résultats de consistance et de normalité asymptotique pour les estimateurs des paramètres. Pour les processus de Hawkes, deux cadres asymptotiques ont été récemment considérés : Chen et Hall (2013) proposent une situation où l'intensité d'immigration tend vers l'infini, tandis que la fenêtre d'observation du processus reste fixée ; un autre cadre asymptotique a été introduit par Dahlhaus (1997) pour l'étude des séries temporelles non-stationnaires et repris dans les travaux de Roueff *et al.* (2016) pour l'étude des processus de Hawkes non-stationnaires.

L'analyse spectrale des séries temporelles non-stationnaires a déjà été étudiée, et des estimateurs de Whittle adaptés ont été proposés (Dahlhaus, 1997). Il faudrait vérifier que les séries de comptage issues des processus de Hawkes satisfont aux conditions énoncées dans ces travaux, au prix d'un certain effort. Alternativement, il serait possible d'étendre directement les résultats établis dans le Chapitre 3 au cas non-stationnaire : il est raisonnable de penser que les propriétés de faible dépendance établies pour le processus de Hawkes stationnaire à partir de sa représentation sous forme d'arbres de Galton–Watson peuvent être également démontrées dans le cas non-stationnaire, pour lequel le processus de Hawkes garde cette représentation. Il faudrait ensuite étendre les résultats asymptotiques établis par Dzhaparidze (1986) pour des séries temporelles non-stationnaires. L'estimateur de Whittle étant connu pour être robuste à un certain nombre d'écarts aux hypothèses (Taqqu et Teverovsky, 1997; Giraitis et Taqqu, 1999; Sousa-Vieira, 2016), il est probable qu'il le soit aussi lorsque l'hypothèse de stationnarité des processus de Hawkes est relâchée.

Un complément à apporter aux travaux portant sur le processus de Hawkes est d'étendre les résultats considérés à l'échelle spatiale, qui est indispensable pour prendre en compte la diffusion des maladies contagieuses. Il est donc nécessaire d'étendre les résultats en plus grande dimension pour comprendre l'effet de l'agrégation spatiale des données. Ceci est d'autant plus important que la transmission des maladies contagieuses dans l'espace peut se produire à des échelles géographiques petites — des échelles que des données à la résolution de zones administratives habituelles (commune, département, région, ...) ne permettraient pas d'analyser (Meyer *et al.*, 2012). L'extension des résultats établis en plus grande dimension ne devrait pas poser de problèmes majeurs, bien que le passage à des dimensions plus élevées

nécessite, pour le Théorème 1, que les moments du noyau de reproduction existent jusqu'à un ordre plus élevé — jusqu'à l'ordre $d + \delta$, $\delta > 0$, où d est la dimension de l'espace considéré, à cause de la dernière intégration de la preuve du théorème. La difficulté de l'extension en plus grande dimension réside en réalité dans les outils d'estimation du processus : au contraire du cas temporel pour lequel la fonction de densité spectrale peut être explicitée pour n'importe quel choix de taille d'intervalle, dans le cas spatial la géométrie des régions d'étude impacte fortement la forme de la densité spectrale. Bien qu'il soit facile de calculer la densité spectrale pour une grille spatiale régulière, cela ne tiendrait pas compte des géométries beaucoup plus complexes des zones géographiques ou administratives à l'échelle desquelles les données sont agrégées. Il est donc nécessaire de développer davantage les outils d'estimation de la densité spectrale dans le cas de géométries arbitraires.

Enfin, pour déterminer l'apport des processus de Hawkes par rapport aux modèles classiquement utilisés pour la modélisation des maladies contagieuses, il est nécessaire de pouvoir comparer ces modèles dans des situations bien définies par le contexte épidémiologique. Au contraire des paramètres d'association des processus de Hawkes et des autres modèles qui ne sont pas directement comparables, la fraction attribuable est une statistique adaptée puisque son interprétation ne dépend pas du modèle. De plus, elle est facilement estimable pour les processus de Hawkes, puisqu'ils sont définis à partir de l'intensité conditionnelle qui peut être vue comme une fonction de risque (Daley et Vere-Jones, 2003). Ainsi, la probabilité d'occurrence d'un cas de maladie peut être décomposée selon les différentes sources de risque, et il est possible de déterminer la probabilité que le cas ait été généré par le facteur de risque. Une étape supplémentaire est néanmoins requise lorsque l'on travaille avec des données agrégées, puisque la fonction d'intensité conditionnelle nécessite les temps d'arrivée exacts pour être calculée. Il faudrait alors soit pouvoir prédire les temps d'arrivée du processus à partir des données agrégées et des estimations des paramètres, soit établir la loi de l'estimateur de la fraction attribuable pour la série de comptage du processus.

Par suite, la comparaison des estimations de la fraction attribuable pour le processus de Hawkes et pour les autres modèles sur des données simulées dans un contexte épidémiologique choisi permettrait d'évaluer l'apport de ces processus pour la recherche en santé publique.

Annexes

1.A. Risque attribuable

En notant D la survenue de la maladie et E l'exposition au facteur de risque, le risque attribuable s'écrit :

$$\mathrm{RA} = \frac{\mathbb{P}(D) - \mathbb{P}(D|\bar{E})}{\mathbb{P}(D)}. \tag{9}$$

Le risque attribuable mesure, *à l'échelle de la population étudiée*, l'excès relatif de risque de la maladie attribuable à l'exposition, c'est-à-dire la proportion parmi tous les sujets malades de ceux attribuables à l'exposition.

À la différence du risque relatif (RR), qui mesure le rapport du taux d'incidence de la maladie entre les sujets exposés et non exposés,

$$\mathbb{P}(D|E) = \mathrm{RR} \times \mathbb{P}(D|\bar{E}),$$

le risque attribuable dépend à la fois de la force de l'association entre le facteur d'exposition et la maladie et de la prévalence de ce facteur dans la population, p_E. Cette relation est mise en évidence en décomposant dans (9) la probabilité de la maladie $\mathbb{P}(D)$ par $\mathbb{P}(D|E)p_E + \mathbb{P}(D|\bar{E})(1 - p_E)$ pour obtenir (Cole et MacMahon, 1971; Miettinen, 1974) :

$$\mathrm{RA} = \frac{p_E(\mathrm{RR} - 1)}{1 + p_E(\mathrm{RR} - 1)}. \tag{10}$$

Une troisième formulation du risque attribuable permet de faire ressortir la prévalence de l'exposition parmi les sujets malades $p_{E|D} = p_E \times \mathbb{P}(D|E)/\mathbb{P}(D)$ (Miettinen, 1974) :

$$\mathrm{RA} = p_{E|D}\frac{\mathrm{RR} - 1}{\mathrm{RR}}. \tag{11}$$

Ces deux formulations alternatives permettent de rattacher le risque attribuable à la prévalence de l'exposition, et à la force de son association avec la maladie. Ainsi, par exemple, en l'absence d'exposition dans la population, $p_E \approx 0$, ou d'association, $\mathrm{RR} = 1$, le risque attribuable est nul. Inversement, lorsque la population entière est exposée, $p_E \approx 1$, le risque attribuable est équivalent au risque des personnes exposées. Enfin, on notera que lorsque le facteur d'exposition provoque surement la

maladie, RR $\to \infty$, le risque attribuable est alors égal à la prévalence de l'exposition parmi les sujets malades.

Une valeur élevée du risque relatif ne correspond donc pas systématiquement à une valeur élevée du risque attribuable et dépend de la prévalence de l'exposition dans la population cible. En général, le risque attribuable n'est pas transposable d'une population à une autre, car la prévalence (mais aussi parfois le risque relatif dans l'étude de facteurs génétiques par exemple) peut varier fortement selon les populations, mais aussi dans le temps et dans l'espace.

Dans le cadre des études transversales ou de cohorte pour lesquelles l'échantillonnage de la population ne dépend pas de l'exposition ni de la maladie des individus, le risque attribuable peut être estimé en modélisant le processus d'échantillonnage par une loi multinomiale à cinq paramètres (la taille de l'échantillon et les quatre probabilités d'état malades / exposés). Tous les termes intervenant dans les relations (9), (10) et (11) peuvent alors être estimés, et l'estimateur du risque attribuable est

$$\widehat{\mathrm{RA}} = \frac{ad - bc}{(a+b)(b+d)}, \tag{12}$$

où a, b, c et d sont respectivement les nombres de cas exposés, de cas non exposés, de témoins exposés et de témoins non exposés.

Dans le cadre des études cas-témoins ou de cohorte pour lesquelles l'échantillonnage dépend de l'exposition ou de la maladie, le processus d'échantillonnage ne suit plus une loi multinomiale mais deux lois binomiales indépendantes (une pour les cas, l'autre pour les témoins). Les prévalences de la maladie et de l'exposition ne peuvent plus être estimées à l'échelle de la population. Il faut alors utiliser la formulation (10), approcher la prévalence de l'exposition dans la population à partir de la proportion des témoins exposés en faisant l'hypothèse d'une maladie rare, et approcher l'estimateur du risque relatif par celui de l'odds ratio. Alternativement, la formulation (11) peut être utilisée puisque la prévalence de l'exposition parmi les cas est directement estimable. Ces deux formulations mènent alors à l'estimateur suivant pour le risque attribuable :

$$\widehat{\mathrm{RA}} = \frac{ad - bc}{d(a+b)}. \tag{13}$$

Des estimateurs asymptotiques de la variance et des intervalles de confiance Gaussiens du risque attribuable peuvent être calculés à partir des relations (12) et (13) en appliquant la méthode delta aux estimateurs du maximum de vraisemblance

des variables a, b, c et d. D'autres méthodes ont également été proposées pour obtenir des intervalles de confiance plus précis à partir de transformations du risque attribuable (Leung et Kupper, 1981; Llorca et Delgado-Rodríguez, 2000).

Dans le cas où plusieurs expositions E_1 et E_2 sont considérées, l'estimateur brut du risque attribuable est biaisé, excepté les cas où les expositions sont indépendantes ou lorsque l'exposition à E_2 seule n'augmente pas le risque de survenue de la maladie (Walter, 1980). Il convient donc d'ajuster l'estimateur du risque attribuable pour les facteurs de confusions connus ou suspectés. Par ailleurs, sauf indépendance des expositions, l'élimination d'une exposition modifie généralement la distribution de l'autre. Il faut alors prendre en compte l'effet du changement de distribution dans l'estimation du risque attribuable, en corrigeant la valeur de référence (c'est-à-dire la probabilité de la maladie chez les non-exposés) ; par exemple, l'introduction d'exercices physiques quotidiens comme prévention de maladie cardiaque peut aussi agir sur l'hypertension, également facteur de risque des maladies cardiaques (Walter, 1976). Des revues détaillées des méthodes ajustées d'estimation sont disponibles dans la littérature (Benichou, 2001; Bard *et al.*, 2005).

1.B. Historique des épidémies de syndromes grippaux en France

TABLEAU 1.B.1 – Historique des épidémies de syndromes grippaux en France, données issues du réseau Sentinelles (`https://www.sentiweb.fr/france/fr/?page=epidemies`).

Saison	Début (Semaine)	Fin (Semaine)	Durée (Semaines)	Pic (Semaine)	Incidence au pic (taux pour 100.000 habitants)	Incidence totale (taux pour 100.000 habitants)
2018 / 2019	2019-03	2019-08	6	2019-06	599	2 457
2017 / 2018	2017-50	2018-10	13	2017-52	459	3 382
2016 / 2017	2016-50	2017-05	8	2017-03	410	2 720
2015 / 2016	2016-04	2016-14	11	2016-11	467	3 465
2014 / 2015	2015-03	2015-11	9	2015-06	827	4 413
2013 / 2014	2014-05	2014-09	5	2014-07	325	1 284
2012 / 2013	2012-51	2013-11	13	2013-05	770	5 531
2011 / 2012	2012-05	2012-12	8	2012-08	452	2 276
2010 / 2011	2010-51	2011-07	9	2011-01	490	3 491
2009 / 2010	2009-37	2009-52	16	2009-49	754	5 515
2008 / 2009	2008-51	2009-08	10	2009-04	868	4 459
2007 / 2008	2008-02	2008-10	9	2008-06	615	3 468
2006 / 2007	2007-03	2007-09	7	2007-06	815	3 398
2005 / 2006	2006-04	2006-13	10	2006-06	421	2 598
2004 / 2005	2005-03	2005-12	10	2005-06	939	5 106
2003 / 2004	2003-45	2004-01	9	2003-49	928	4 667
2002 / 2003	2003-05	2003-15	11	2003-07	396	2 533
2001 / 2002	2002-01	2002-08	8	2002-04	848	3 893
2000 / 2001	2000-50	2001-07	10	2001-05	471	2 629
1999 / 2000	1999-49	2000-06	10	2000-01	922	5 593
1998 / 1999	1998-53	1999-11	12	1999-07	896	5 581
1997 / 1998	1998-06	1998-17	12	1998-14	547	4 178

Suite à la page suivante

Tableau 1.B.1 – suite de la page précédente

Saison	Début (Semaine)	Fin (Semaine)	Durée (Semaines)	Pic (Semaine)	Incidence au pic (taux pour 100.000 habitants)	Incidence totale (taux pour 100.000 habitants)
1996 / 1997	1996-48	1997-05	10	1996-51	1 106	5 175
1995 / 1996	1995-47	1996-02	8	1995-51	1 299	4 818
1994 / 1995	1995-11	1995-18	8	1995-14	431	1 925
1993 / 1994	1993-46	1994-01	8	1993-49	1 565	5 402
1992 / 1993	1993-03	1993-13	11	1993-06	500	3 268
1991 / 1992	1991-49	1992-06	10	1991-51	666	3 542
1990 / 1991	1991-06	1991-10	5	1991-08	381	1 386
1989 / 1990	1989-48	1990-06	11	1989-51	1 463	8 207
1988 / 1989	1988-46	1989-02	9	1988-50	1 793	8 227
1987 / 1988	1988-09	1988-15	7	1988-11	566	2 432
1986 / 1987	1987-04	1987-09	6	1987-06	533	2 415
1985 / 1986	1986-02	1986-12	11	1986-07	886	6 160
1984 / 1985	1985-03	1985-14	12	1985-05	1 155	7 758

1.C. Exemples de jeux de données simulés

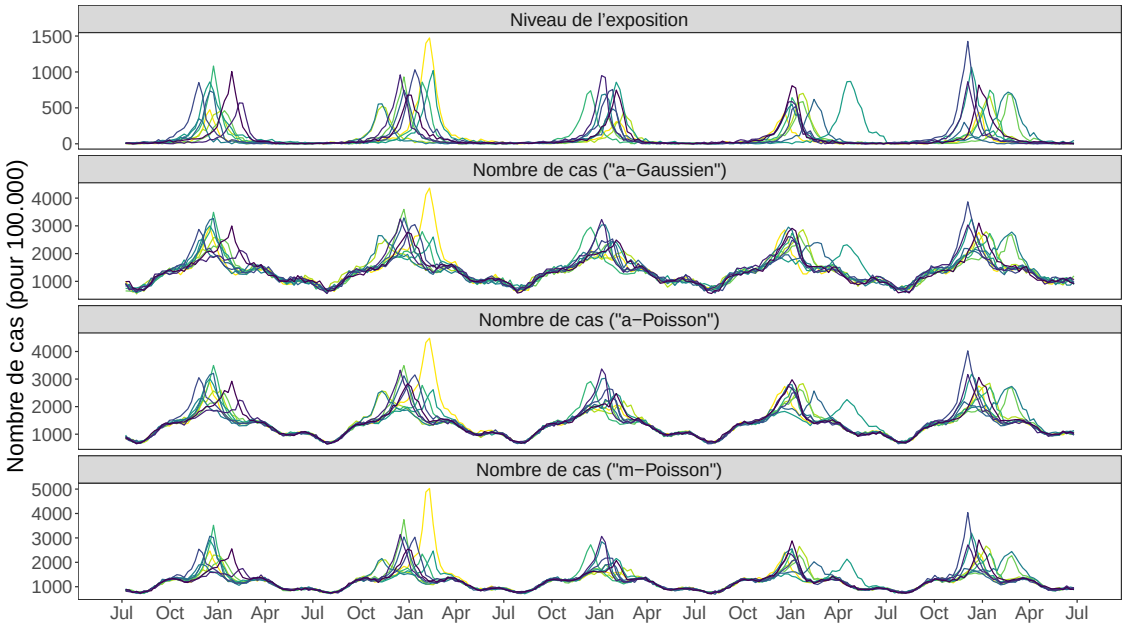


FIGURE 1.C.1 – Dix jeux de simulation pour les séries d'exposition et du nombre de cas, avec $\overline{\text{FA}} = 30\%$. Seules les 5 premières années simulées ont été représentées.

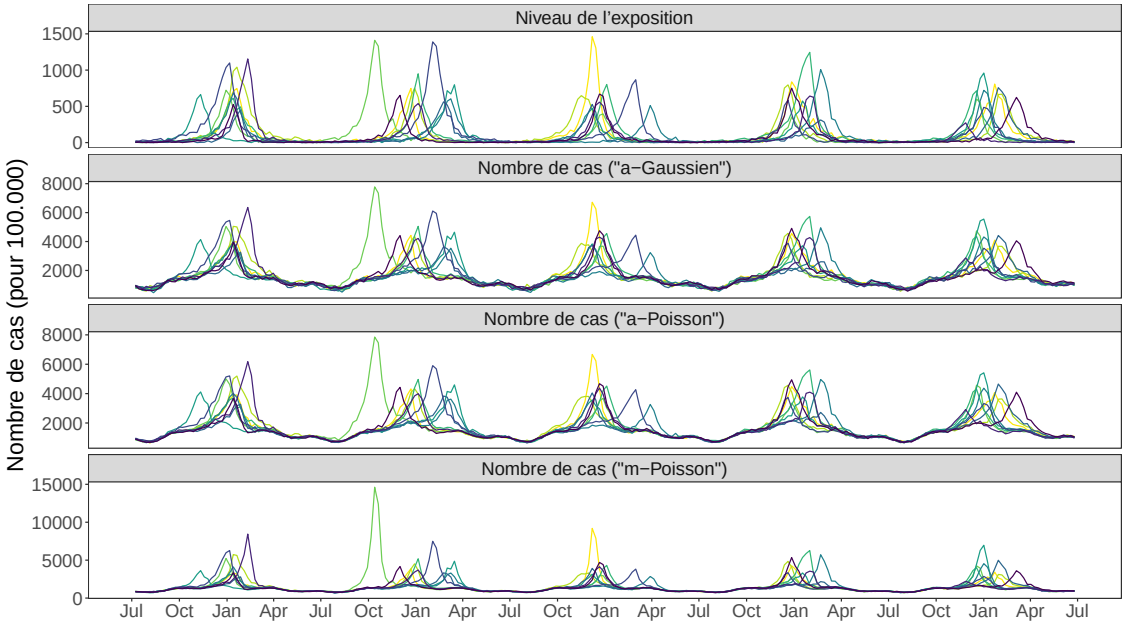


FIGURE 1.C.2 – Dix jeux de simulation pour les séries d'exposition et du nombre de cas, avec $\overline{\text{FA}} = 50\%$. Seules les 5 premières années simulées ont été représentées.

1.D. Critères d'évaluation de l'étude de simulation

TABLEAU 1.D.1 – Critères d'évaluation pour les estimations de la fraction attribuable, lorsque le nombre de cas a été simulé par le modèle "a-Gaussien" (indiqué par la colonne dont les valeurs sont en gras).

		Additifs : **a-**			Multiplicatifs : **m-**		
$\overline{FA}$	**Critère**	**Gaussien**	**Poisson**	**NégBin**	**Poisson**	**NégBin**	**Serfling**
10%	BR (%)	**0,05**	0,19	0,18	-4,74	-3,99	-16,95
	REQM	**0,47**	0,76	0,82	1,07	1,18	1,97
	PR (%)	**91,3**	88,0	91,0	74,4	80,0	22,0
20%	BR (%)	**0,03**	0,16	0,19	-7,81	-6,17	-17,12
	REQM	**0,40**	0,71	0,79	1,87	1,66	3,57
	PR (%)	**93,3**	84,9	89,3	30,8	51,6	0,3
30%	BR (%)	**0,04**	0,01	-0,01	-11,45	-8,81	-17,09
	REQM	**0,36**	0,63	0,72	3,67	2,90	5,23
	PR (%)	**92,7**	86,0	89,9	1,5	11,6	0,0
40%	BR (%)	**-0,05**	-0,03	-0,03	-15,21	-11,36	-17,26
	REQM	**0,30**	0,54	0,64	6,34	4,76	7,01
	PR (%)	**91,6**	85,5	89,3	0,0	0,1	0,0
50%	BR (%)	**0**	-0,03	-0,03	-18,72	-13,37	-17,42
	REQM	**0,25**	0,50	0,60	9,64	6,89	8,81
	PR (%)	**92,6**	83,3	86,7	0,0	0,0	0,0

TABLEAU 1.D.2 – Critères d'évaluation pour les estimations de la fraction attribuable, lorsque le nombre de cas a été simulé par le modèle "a-Poisson" (indiqué par la colonne dont les valeurs sont en gras).

		Additifs : **a-**			Multiplicatifs : **m-**		
$\overline{FA}$	**Critère**	**Gaussien**	**Poisson**	**NégBin**	**Poisson**	**NégBin**	**Serfling**
10%	BR (%)	-0,08	**-0,06**	-0,06	-4,95	-4,39	-16,78
	REQM	0,30	**0,30**	0,30	0,91	0,95	1,74
	PR (%)	90,8	**94,6**	94,5	59,5	63,1	0,0
20%	BR (%)	-0,01	**-0,01**	-0,01	-8,32	-7,03	-16,96
	REQM	0,30	**0,30**	0,30	1,92	1,70	3,45
	PR (%)	87,3	**93,5**	92,2	15,0	25,0	0,0
30%	BR (%)	0,01	**0,01**	0,01	-11,86	-9,59	-17,06
	REQM	0,26	**0,26**	0,26	3,78	3,09	5,18
	PR (%)	86,7	**95,2**	93,7	0,2	1,4	0,0
40%	BR (%)	0,02	**0,01**	0,01	-15,61	-11,91	-17,13
	REQM	0,25	**0,23**	0,23	6,50	4,96	6,93
	PR (%)	85,1	**95,6**	93,3	0,0	0,0	0,0
50%	BR (%)	0,01	**-0,01**	-0,01	-18,81	-13,61	-17,23
	REQM	0,22	**0,20**	0,20	9,70	7,01	8,71
	PR (%)	83,0	**95,6**	92,1	0,0	0,0	0,0

TABLEAU 1.D.3 – Critères d'évaluation pour les estimations de la fraction attribuable, lorsque le nombre de cas a été simulé par le modèle "m-Poisson" (indiqué par la colonne dont les valeurs sont en gras).

		Additifs : **a-**			Multiplicatifs : **m-**		
$\overline{\text{FA}}$	**Critère**	**Gaussien**	**Poisson**	**NégBin**	**Poisson**	**NégBin**	**Serfling**
10%	BR (%)	5,37	4,63	4,45	**-0,15**	-0,15	-13,80
	REQM	0,96	0,92	0,91	**0,56**	0,56	1,58
	PR (%)	56,0	63,3	65,5	**95,3**	95,1	0,1
20%	BR (%)	10,81	8,07	7,16	**-0,05**	-0,05	-14,18
	REQM	2,47	1,86	1,65	**0,31**	0,31	2,92
	PR (%)	5,3	14,1	18,3	**94,4**	92,8	0,0
30%	BR (%)	18,66	12,15	9,57	**0,03**	0,03	-13,58
	REQM	6,54	4,06	3,14	**0,52**	0,52	4,17
	PR (%)	0,0	0,0	0,3	**94,6**	91,0	0,0
40%	BR (%)	29,42	15,94	10,92	**-0,01**	-0,01	-12,68
	REQM	14,08	6,89	4,61	**0,41**	0,41	5,15
	PR (%)	0,0	0,0	0,0	**95,0**	88,4	0,0
50%	BR (%)	46,28	19,80	11,84	**0,01**	0,01	-11,60
	REQM	29,15	10,60	6,19	**0,31**	0,31	5,87
	PR (%)	0,0	0,0	0,0	**94,8**	85,3	0,0

TABLEAU 1.D.4 – Critères d'évaluation pour les estimations de la fraction attribuable, lorsque le nombre de cas a été simulé par le modèle "a-Gaussien" (indiqué par la colonne dont les valeurs sont en gras) et que le décalage temporel est mal identifié.

		Additifs : **a-**			Multiplicatifs : **m-**		
$\overline{\text{FA}}$	**Critère**	**Gaussien**	**Poisson**	**NégBin**	**Poisson**	**NégBin**	**Serfling**
10%	BR (%)	**-17,16**	-7,45	-6,72	-12,75	-11,42	-20,29
	REQM	**1,82**	1,10	1,10	1,63	1,65	2,27
	PR (%)	**12,4**	70,2	79,1	49,1	63,3	0,2
20%	BR (%)	**-20,33**	-7,14	-6,09	-15,67	-13,42	-20,32
	REQM	**4,16**	1,63	1,49	3,32	2,96	4,20
	PR (%)	**0,0**	43,8	63,0	3,8	16,8	0,0
30%	BR (%)	**-23,44**	-6,76	-5,47	-18,68	-15,15	-20,07
	REQM	**7,13**	2,16	1,84	5,74	4,71	6,13
	PR (%)	**0,0**	18,1	48,9	0,0	0,3	0,0
40%	BR (%)	**-26,13**	-6,34	-4,87	-21,39	-16,60	-20,06
	REQM	**10,54**	2,66	2,14	8,71	6,79	8,14
	PR (%)	**0,0**	7,6	39,5	0,0	0,0	0,0
50%	BR (%)	**-28,32**	-5,95	-4,40	-24,40	-18,21	-20,37
	REQM	**14,26**	3,10	2,38	12,39	9,27	10,32
	PR (%)	**0,0**	4,4	35,0	0,0	0,0	0,0

TABLEAU 1.D.5 – Critères d'évaluation pour les estimations de la fraction attribuable, lorsque le nombre de cas a été simulé par le modèle "a-Poisson" (indiqué par la colonne dont les valeurs sont en gras) et que le décalage temporel est mal identifié.

		Additifs : **a-**			Multiplicatifs : **m-**		
$\overline{\text{FA}}$	**Critère**	**Gaussien**	**Poisson**	**NégBin**	**Poisson**	**NégBin**	**Serfling**
10%	BR (%)	-8,46	**-7,56**	-7,18	-12,87	-11,79	-20,04
	REQM	0,91	**0,83**	0,79	1,50	1,48	2,06
	PR (%)	27,6	**42,7**	51,7	30,1	39,9	0,0
20%	BR (%)	-8,23	**-6,98**	-6,06	-15,48	-13,27	-19,77
	REQM	1,71	**1,46**	1,29	3,25	2,85	4,02
	PR (%)	3,1	**9,5**	26,1	0,4	4,9	0,0
30%	BR (%)	-8,22	**-6,67**	-5,40	-18,68	-15,26	-20,11
	REQM	2,54	**2,07**	1,70	5,75	4,74	6,11
	PR (%)	1,1	**3,5**	18,9	0,0	0,0	0,0
40%	BR (%)	-8,06	**-6,26**	-4,81	-21,63	-16,83	-19,99
	REQM	3,32	**2,59**	2,02	8,83	6,90	8,10
	PR (%)	1,9	**2,6**	18,8	0,0	0,0	0,0
50%	BR (%)	-8,09	**-5,98**	-4,44	-24,27	-18,12	-20,12
	REQM	4,15	**3,08**	2,32	12,32	9,22	10,19
	PR (%)	1,2	**2,0**	15,6	0,0	0,0	0,0

TABLEAU 1.D.6 – Critères d'évaluation pour les estimations de la fraction attribuable, lorsque le nombre de cas a été simulé par le modèle "m-Poisson" (indiqué par la colonne dont les valeurs sont en gras) et que le décalage temporel est mal identifié.

		Additifs : **a-**			Multiplicatifs : **m-**		
$\overline{\text{FA}}$	**Critère**	**Gaussien**	**Poisson**	**NégBin**	**Poisson**	**NégBin**	**Serfling**
10%	BR (%)	-3,02	-2,71	-2,51	**-8,03**	-7,70	-16,34
	REQM	0,68	0,68	0,68	**0,90**	0,87	1,78
	PR (%)	75,2	76,8	78,9	**37,7**	45,8	0,1
20%	BR (%)	1,75	0,86	0,74	**-8,09**	-7,04	-16,62
	REQM	1,45	1,10	0,98	**1,71**	1,51	3,42
	PR (%)	70,1	75,4	82,0	**2,9**	10,0	0,0
30%	BR (%)	8,03	4,84	3,73	**-8,15**	-6,45	-15,99
	REQM	3,63	2,18	1,67	**2,54**	2,05	4,88
	PR (%)	46,2	54,5	64,4	**1,1**	5,6	0,0
40%	BR (%)	17,71	9,31	6,34	**-8,01**	-5,62	-14,89
	REQM	9,90	4,69	3,07	**3,32**	2,39	6,04
	PR (%)	17,5	23,5	36,5	**0,3**	8,8	0,0
50%	BR (%)	31,65	13,47	8,23	**-7,87**	-4,85	-13,76
	REQM	25,21	7,90	4,67	**4,06**	2,58	6,99
	PR (%)	3,5	4,3	12,3	**0,2**	12,1	0,0

1.E. Modèles causaux fonctionnels

Pour alléger les notations, nous omettons la dépendance en t dans cette annexe. Mathématiquement, la difficulté de la définition d'un estimateur contrefactuel de la fraction attribuable provient du fait qu'il faut garantir que le tirage aléatoire qui produit y^* soit en relation avec le tirage qui a produit y, et donc que l'observation de y informe sur la réalisation de y^*. Nous avons proposé que cette relation passe par un terme d'erreur et avons montré, pour la régression linéaire et celle de Poisson, que cela permet de définir des estimateurs contrefactuels de la fraction attribuable, mais que ces méthodes ne peuvent pas être étendues facilement, comme par exemple pour la régression négative-binomiale, même sous des hypothèses fortes d'indépendance entre y^* et χ, qui représente la contribution du facteur d'exposition x.

Nous proposons une approche alternative basée sur une comparaison de quantile, à partir des travaux de Pearl (2009) qui a posé les fondations d'une approche moderne de la causalité en statistique. Au lieu de s'intéresser à la distribution conditionnelle $p(\cdot|x)$ de y sachant x, nous souhaitons définir une distribution *interventionnelle*, $p(\cdot|do(x))$, qui représenterait la distribution de y qui aurait été observée si une intervention avait pu fixer la valeur du facteur d'exposition sans toucher aux autres variables. L'opérateur *do* (anglais du verbe "faire") a été introduit par Pearl pour formaliser l'impact des interventions en statistique. Des introductions pédagogiques à l'utilisation de l'opérateur *do* sont disponibles (Pearl, 2010; Tucci, 2013; Huszár, 2018, 2019a,b).

Pour définir la distribution interventionnelle, il est nécessaire d'introduire **un modèle causal fonctionnel** qui contient les liens de causalité entre les variables (Figure 1.E.1a) :

$$\begin{aligned} x &= u_x, \\ y &= f_\theta(x, u_y), \end{aligned}$$

où u_x et u_y, appelées **variables d'erreur**, sont des variables aléatoires caractérisant les lois de x et y. L'introduction des variables d'erreur est l'étape clef permettant de dissocier le caractère aléatoire et la relation de causalité des variables x et y. La question contrefactuelle posée, à savoir "combien de cas auraient pu être évités si l'exposition avait été fixée à x_0 ?", peut alors être traitée de la manière suivante :

- À partir des données, déterminer les valeurs des variables d'erreur u_x et u_y

(Figure 1.E.1b). Il s'agit des valeurs qui ont généré l'exposition x et le nombre de cas y observés, et qui ne seront pas modifiées par l'intervention.

- Modifier le modèle par l'intervention $do(x = x_0)$:

$$x = x_0.$$

La valeur de x est remplacé par x_0 sans toucher aux autres variables. En particulier, le lien entre u_x et x est rompu par l'intervention : ce modèle est dit *mutilé* (Figure 1.E.1c).

- Déterminer la valeur contrefactuelle y^* de y, qui correspond au nombre de cas post-intervention, à partir du modèle mutilé :

$$y^* = f_\theta(x_0, u_y).$$

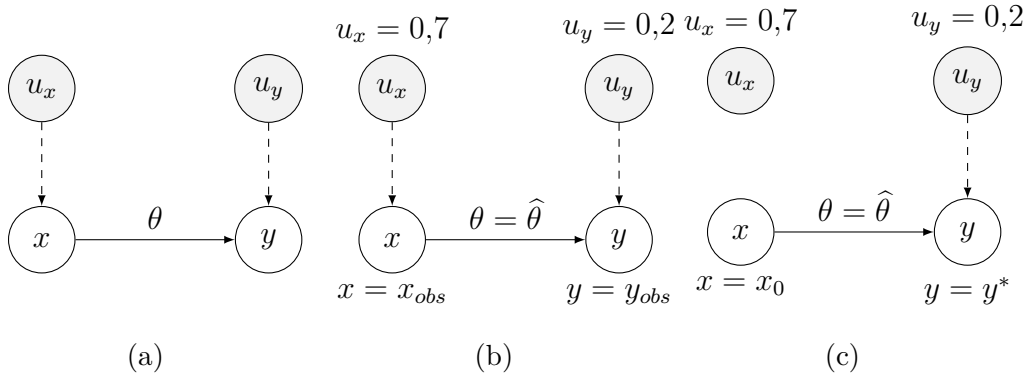


FIGURE 1.E.1 – Modèle causal pour répondre à une question contrefactuelle à propos d'une observation $u = (u_x, u_y)$. (a) Le modèle général, (b) le modèle spécifique à l'observation $u = (0{,}7; 0{,}2)$, (c) le modèle mutilé pour lequel l'exposition est fixée à $x = x_0$ conformément à la question contrefactuelle.

La fraction attribuable est finalement calculée à partir du modèle causal et donnée par :

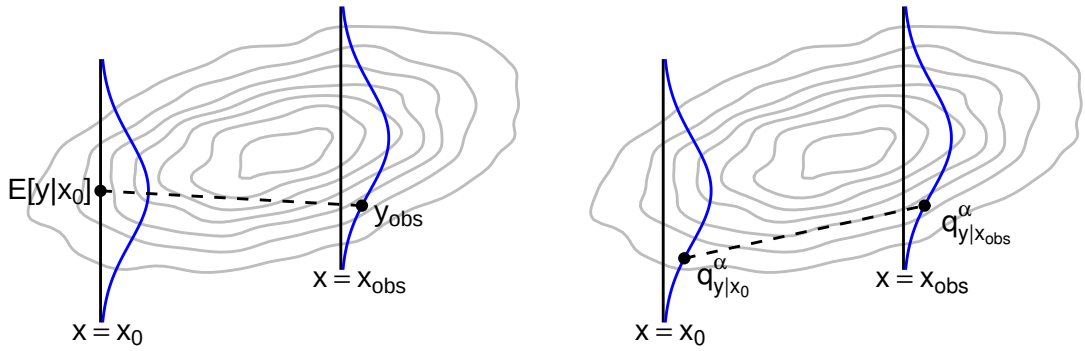
$$\mathrm{FA} = \frac{f_\theta(x, u_y) - f_\theta(x_0, u_y)}{f_\theta(x, u_y)}. \tag{14}$$

En pratique, sa loi peut être approchée par un algorithme de Monte-Carlo, en générant un échantillon de (u_x, u_y). L'estimateur contrefactuel, *i.e.* spécifique à l'observation, est défini par :

$$\widehat{\mathrm{FA}} = \frac{y_{obs} - f_{\widehat{\theta}}(x_0, \widehat{u}_y)}{y_{obs}}.$$

Un exemple important de modèle fonctionnel causal est celui pour lequel $u_y \sim \mathcal{U}(0,1)$ et $f_\theta(x,\cdot) = F^{-1}_{y|x}(\cdot)$, où $F_{y|x}$ est la fonction de répartition de la loi conditionnelle $p(y|x)$. Alors, pour tout $u_y \in [0,1]$, $f_\theta(x,u_y) = F^{-1}_{y|x}(u_y) = q^{u_y}_{y|x}$ est le quantile de niveau u_y de la loi $p(y|x)$. L'intervention $do(x = x_0)$ donne alors la variable contrefactuelle $y^* = f_\theta(x_0,u_y)$ qui est le quantile de niveau u_y de la loi conditionnelle $p(y|x_0)$. En d'autres termes, on attribue à la variable contrefactuelle y^* le quantile de $p(y|x_0)$ correspondant au quantile observé de $p(y|x)$ (Figure 1.E.2b). Ce choix de modèle fonctionnel causal est intéressant puisqu'il peut s'appliquer à n'importe quel modèle linéaire généralisé.

Remarquons que la variable d'erreur u_x n'est en pratique pas nécessaire, puisque le facteur d'exposition x est directement observé. Toutefois, son introduction permet de mettre en valeur la dissociation entre le caractère aléatoire et la relation de causalité des variables x et y.



(a) (b)

FIGURE 1.E.2 – Distribution jointe $p(x,y)$ et distributions conditionnelles $p(y|x_0)$ et $p(y|x_{obs})$. (a) Si l'on ne définit pas y^* sur le même espace de probabilité que y, cela peut conduire à des estimations négatives de la fraction attribuable. (b) On peut attribuer à la variable contrefactuelle y^* le quantile de $p(y|x_0)$ correspondant au quantile observé de $p(y|x)$.

2.A. Codes CIM-10 pour les groupes syndromiques considérés

TABLEAU 2.A.1 – Composition des groupes syndromiques considérés, en fonction des codes CIM-10 recueillis. Le groupe syndromique "Inf. resp. basse" désigne les diagnostics non inclus dans les groupes spécifiques, mais inclus dans le groupe global des infections respiratoires basses.

Code CIM10	Intitulé	Classification syndromique	
		Syndrome	Virale ?
J09	Grippe, à virus grippal zoonotique ou pandémique identifié	Syndrome grippal	Oui
J10*	Grippe, à virus grippal saisonnier identifié	Syndrome grippal	Oui
J11*	Grippe, virus non identifié	Syndrome grippal	Oui
J12	Pneumopathies virales, non classées ailleurs	Pneumopathie	Oui
J120	Pneumopathie adénovirale	Pneumopathie	Oui
J121	Pneumopathie due au virus respiratoire syncytial [VRS]	Pneumopathie	Oui
J122	Pneumopathie due au virus paragrippaux	Pneumopathie	Oui
J128	Autres pneumopathies virales	Pneumopathie	Oui
J129	Pneumopathie virale, sans précision	Pneumopathie	Oui
J13	Pneumonie due à Streptococcus Pneumopathiee	Pneumopathie	Non
J14	Pneumopathie due à Haemophilus influenzae	Pneumopathie	Non
J15*	Pneumopathies bactériennes, non classées ailleurs	Pneumopathie	Non
J16	Pneumopathie due à d'autres micro-organismes infectieux, non classée ailleurs	Pneumopathie	Oui
J160	Pneumopathie due à Chlamydia	Pneumopathie	Non

Suite à la page suivante

Tableau 2.A.1 – suite de la page précédente

Code CIM10	Intitulé	Classification syndromique	
		Syndrome	Virale ?
J168	Pneumopathie due à d'autres micro-organismes infectieux	Pneumopathie	Oui
J17	Pneumopathie au cours de maladies classées ailleurs	Pneumopathie	Oui
J170	Pneumopathie au cours de maladies bactériennes classées ailleurs	Pneumopathie	Non
J171	Pneumopathie au cours de maladies virales classées ailleurs	Pneumopathie	Oui
J172	Pneumopathie au cours de mycoses	Inf. resp. basse	Oui
J173	Pneumopathie au cours de maladies parasitaires	Inf. resp. basse	Oui
J178	Pneumopathie au cours d'autres maladies classées ailleurs	Pneumopathie	Oui
J18*	Pneumopathie à micro-organisme non précisé	Pneumopathie	Non
J20	Bronchite aiguë	Bronchite	Oui
J200	Bronchite aiguë due à Mycoplasma pneumoniae	Bronchite	Non
J201	Bronchite aiguë due à Haemophilus influenzae	Bronchite	Non
J202	Bronchite aiguë due à des steptocoques	Bronchite	Non
J203	Bronchite aiguë due au virus Coxsackie	Bronchite	Oui
J204	Bronchite aiguë due aux virus paragrippaux	Bronchite	Oui
J205	Bronchite aiguë due au virus respiratoire syncytial [VRS]	Bronchite	Oui
J206	Bronchite aiguë due à des rhinovirus	Bronchite	Oui
J207	Bronchite aiguë due à des virus ECHO	Bronchite	Oui

Suite à la page suivante

TABLEAU 2.A.1 – suite de la page précédente

Code CIM10	Intitulé	Classification syndromique	
		Syndrome	Virale ?
J208	Bronchite aiguë due à d'autres micro-organismes précisés	Bronchite	Oui
J209	Bronchite aiguë, sans précision	Bronchite	Oui
J21	Bronchiolite aiguë	Bronchiolite	Oui
J210	Bronchiolite aiguë due au virus respiratoire syncytial [VRS]	Bronchiolite	Oui
J218	Bronchiolite aiguë due à d'autres micro-organismes précisés	Inf. resp. basse	Oui
J219	Bronchiolite, sans précision	Inf. resp. basse	Oui
J40	Bronchite	Bronchite	Oui
J69	Pneumopathie due à des substances solides et liquides	Inf. resp. basse	Non
J690	Pneumopathie due à des aliments et des vomissements	Inf. resp. basse	Non
J691	Pneumopathie due à des huiles et des essences	Inf. resp. basse	Non
J698	Pneumopathie due à d'autres substances solides et liquides	Inf. resp. basse	Non
J84	Autres affections pulmonaires interstitielles	Inf. resp. basse	Oui

2.B. Corrélation entre les groupes syndromiques

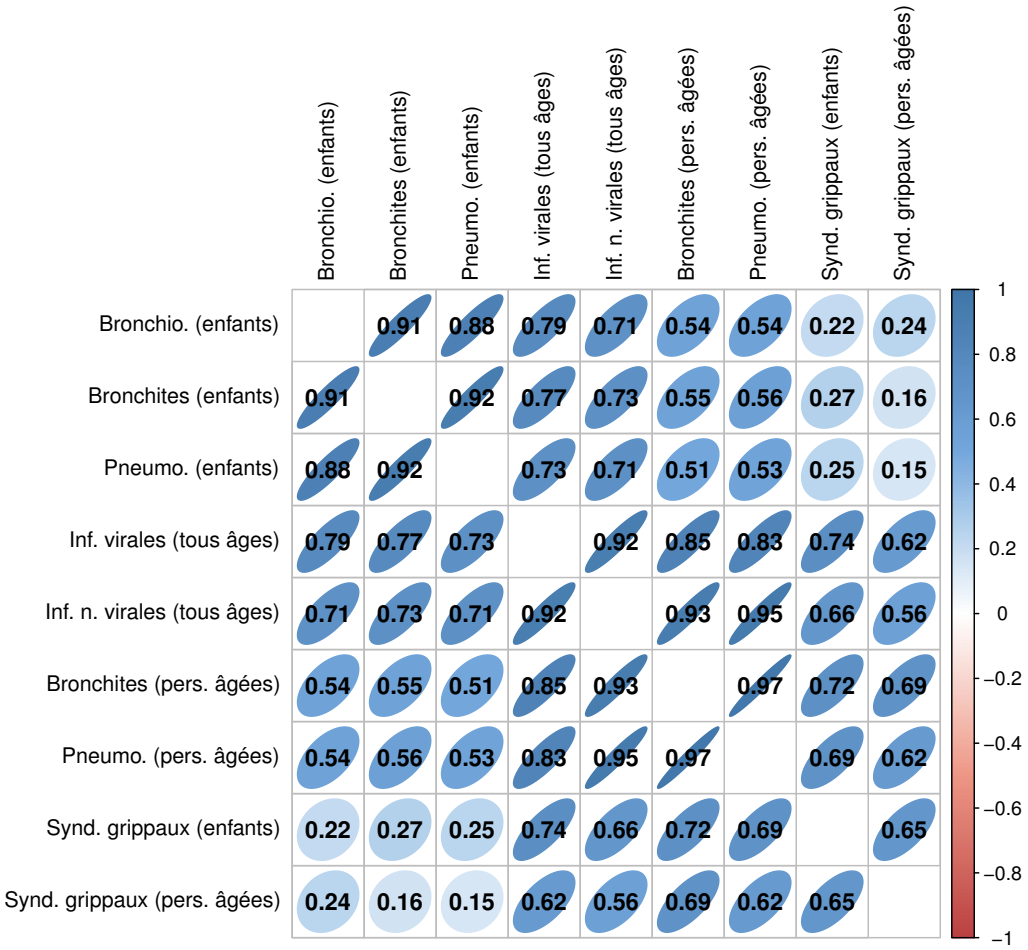


FIGURE 2.B.1 – Graphiques de corrélation entre les séries d'incidence des différents syndromes respiratoires pour chaque classe d'âge considérée.

2.C. Validation des hypothèses des modèles de la Section 2.5.2

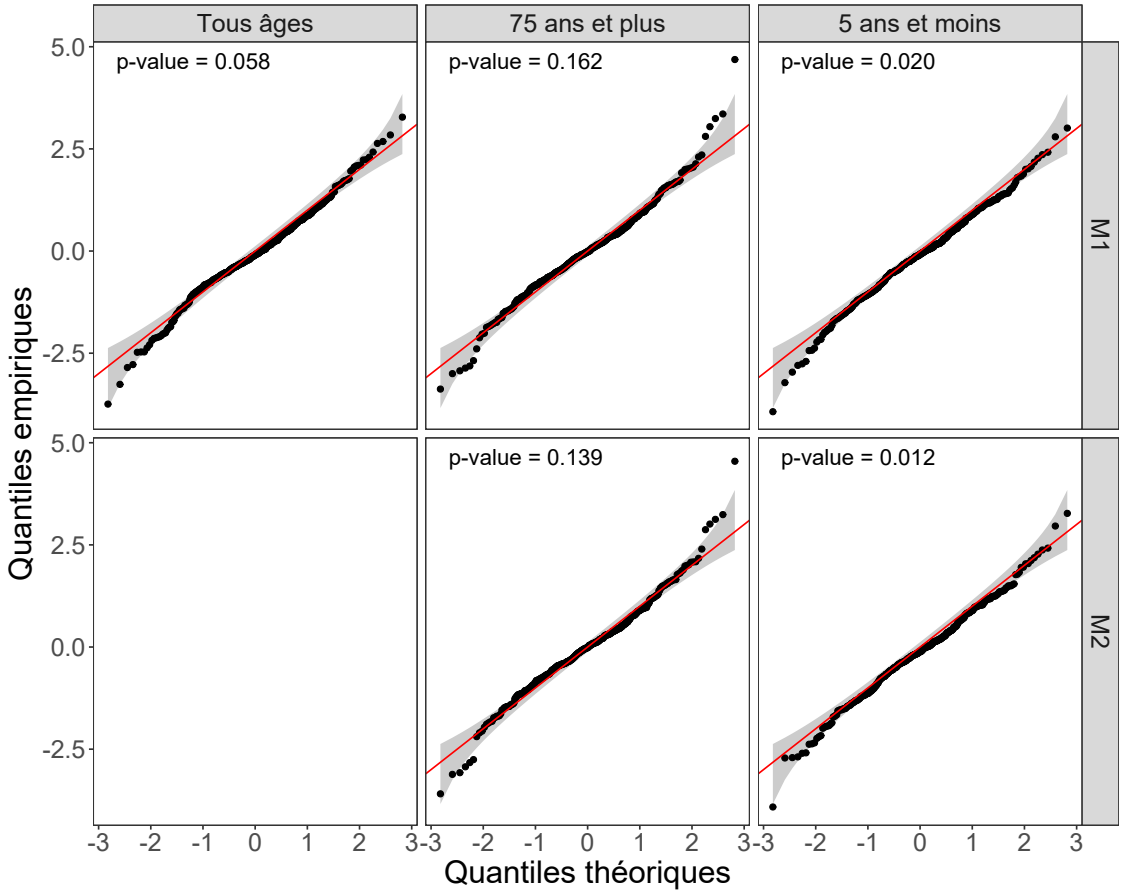


FIGURE 2.C.1 – Graphiques quantile-quantile des résidus standardisés pour chacun des cinq modèles définis en Section 2.5.2. La droite rouge est la première bissectrice. La bande grise correspond aux intervalles de confiance point par point à 95% des quantiles de la loi normale centrée réduite. Les p-values se rapportent à des tests de normalité de Kolmogorov-Smirnov.

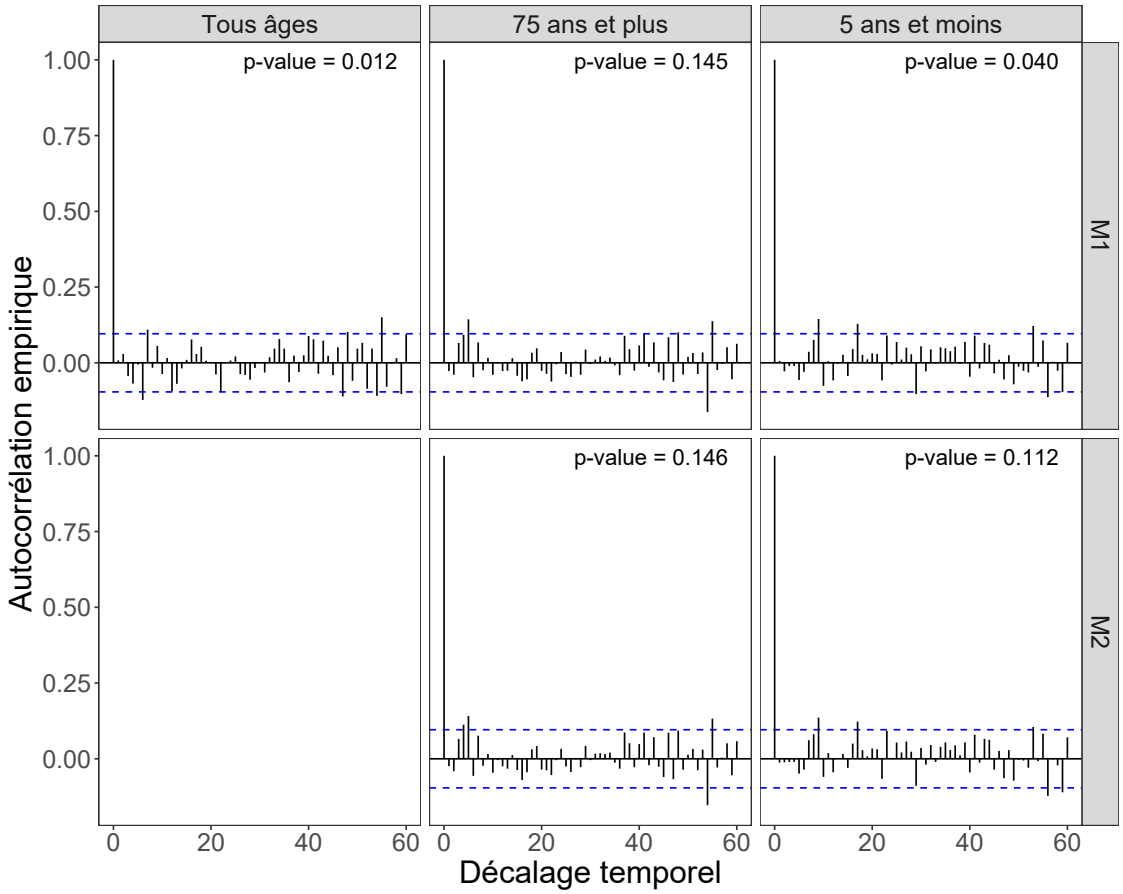


FIGURE 2.C.2 – Fonctions d'autocorrélation empiriques des résidus pour chacun des cinq modèles définis en Section 2.5.2. Les intervalles délimités par les droites bleues hachurées correspondent aux intervalles de confiance asymptotiques à 95% des fonctions d'autocorrélation pour la loi normale. Les p-values se rapportent à des tests d'autocorrélation de Ljung-Box jusqu'à l'ordre 52.

3.A. Figures de la Section 3.5

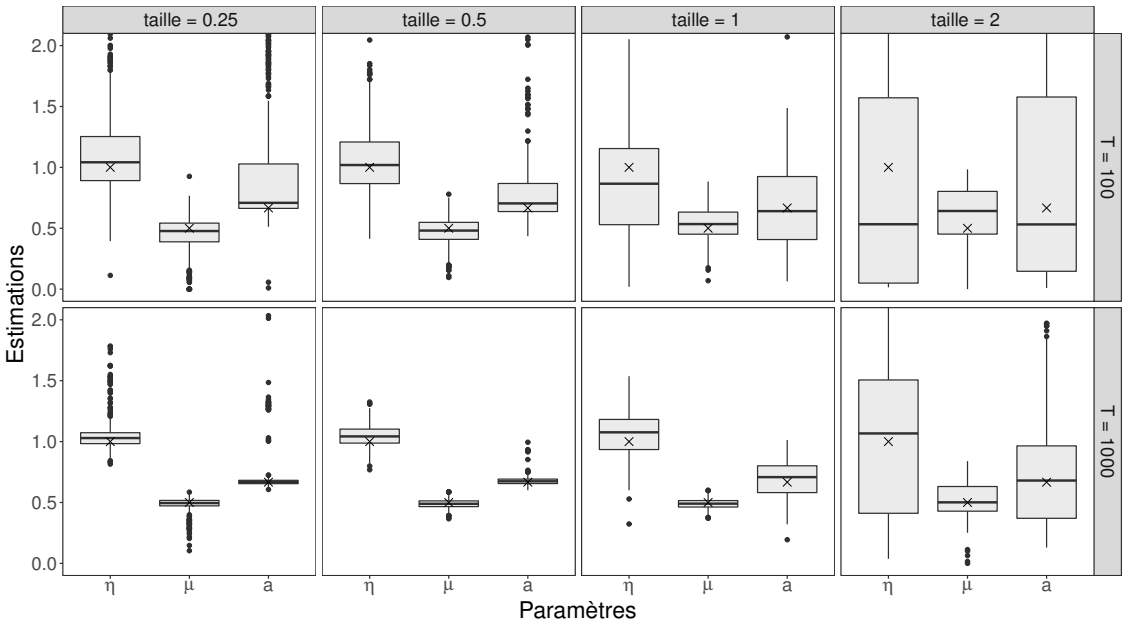


FIGURE 3.A.1 – Estimations des paramètres η, μ et a pour 1000 simulations du processus de Hawkes linéaire avec noyau de reproduction $h_3^*(t) = 3a^3t^{-4}$ sur l'intervalle $[0, T]$. Les vraies valeurs des paramètres (croix) sont : $\eta = 1$, $\mu = 0.5$, $a = 2/3$.

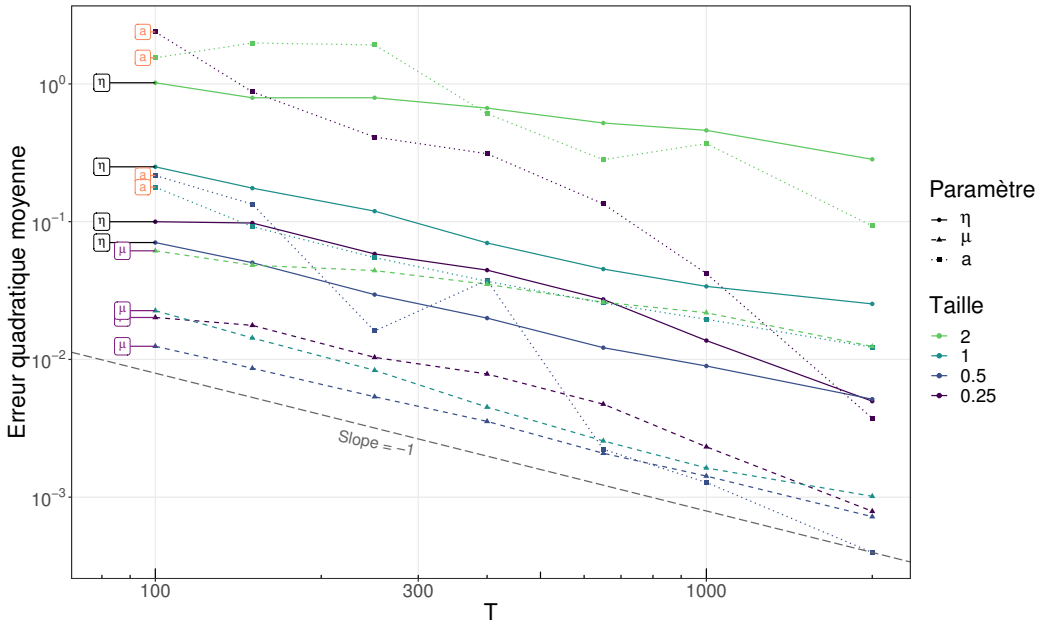


FIGURE 3.A.2 – Erreur quadratique moyenne pour les estimations des paramètres η, μ et a pour 1000 simulations du processus de Hawkes linéaire avec noyau de reproduction $h_3^*(t) = 3a^3t^{-4}$ sur l'intervalle $[0, T]$, à l'échelle log-log. La droite grise pointillée représente la pente idéale de -1, *i.e.* une vitesse de convergence en $\mathcal{O}(n^{-1})$.

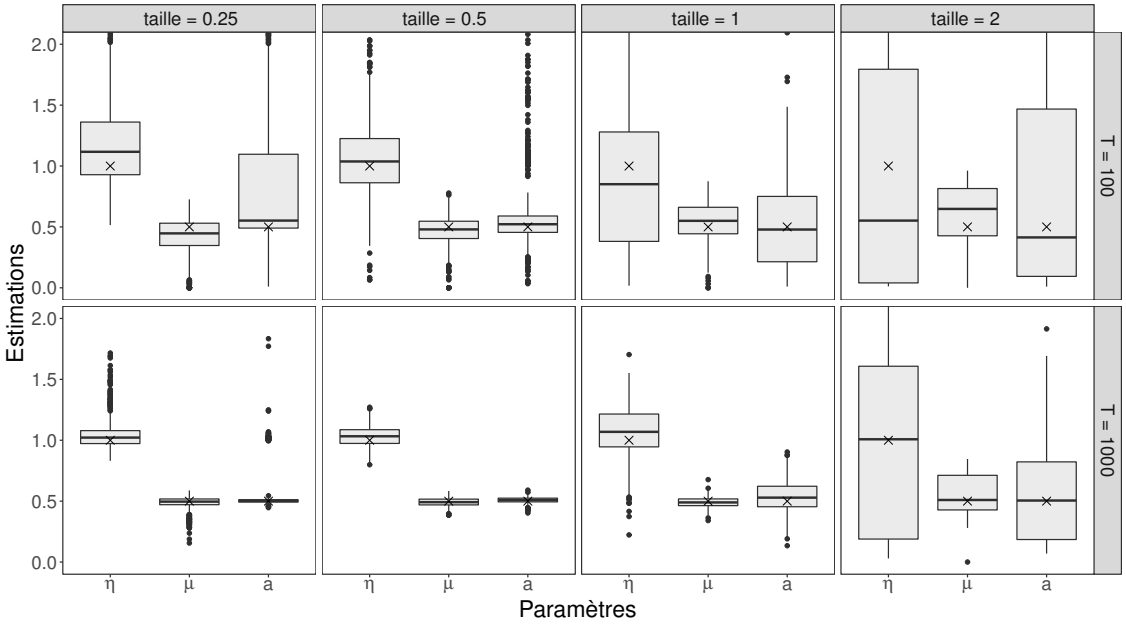


FIGURE 3.A.3 – Estimations des paramètres η, μ et a pour 1000 simulations du processus de Hawkes linéaire avec noyau de reproduction $h_2^*(t) = 2a^2t^{-3}$ sur l'intervalle $[0, T]$. Les vraies valeurs des paramètres (croix) sont : $\eta = 1$, $\mu = 0.5$, $a = 1/2$.

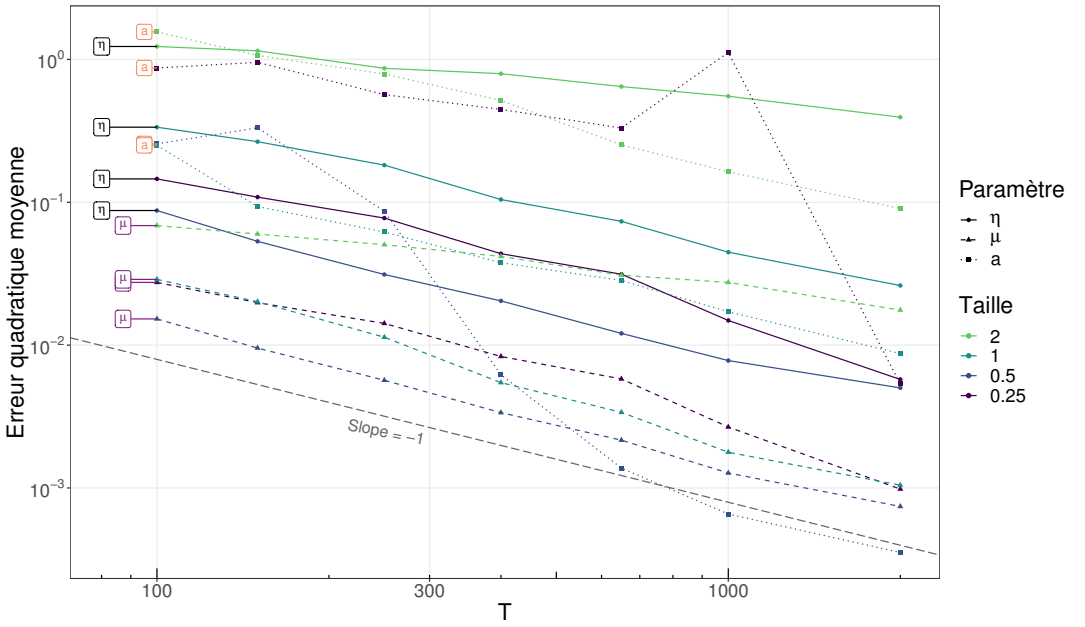


FIGURE 3.A.4 – Erreur quadratique moyenne pour les estimations des paramètres η, μ et a pour 1000 simulations du processus de Hawkes linéaire avec noyau de reproduction $h_2^*(t) = 2a^2t^{-3}$ sur l'intervalle $[0, T]$, à l'échelle log-log. La droite grise pointillée représente la pente idéale de -1, *i.e.* une vitesse de convergence en $\mathcal{O}(n^{-1})$.

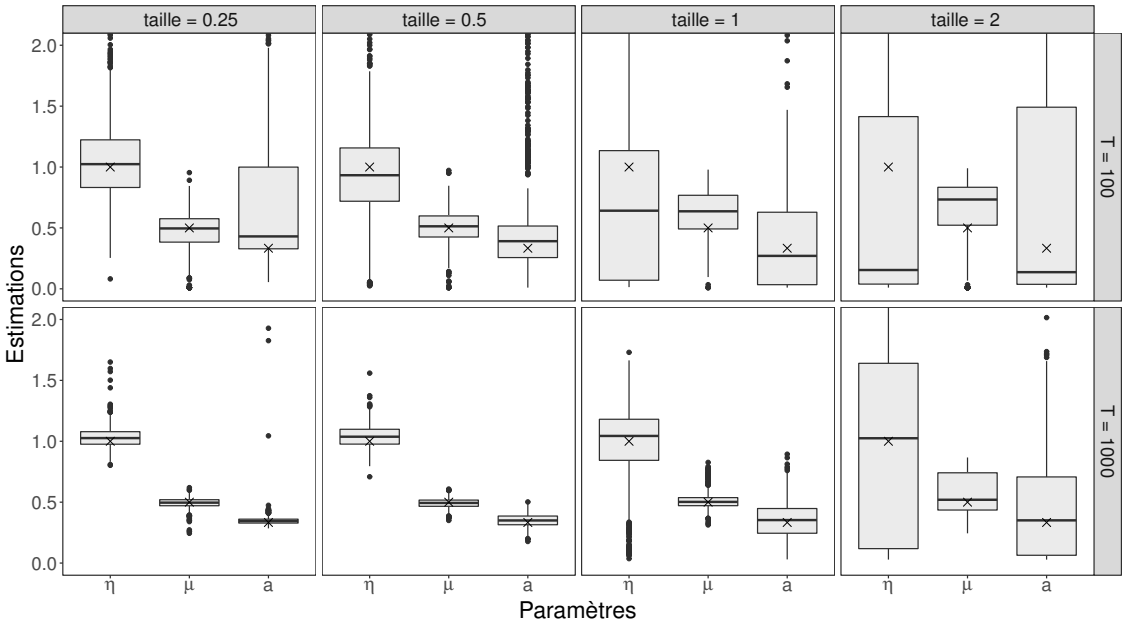


FIGURE 3.A.5 – Estimations des paramètres η, μ et a pour 1000 simulations du processus de Hawkes linéaire avec noyau de reproduction $h_1^*(t) = a^1t^{-2}$ sur l'intervalle $[0, T]$. Les vraies valeurs des paramètres (croix) sont : $\eta = 1$, $\mu = 0.5$, $a = 1/3$.

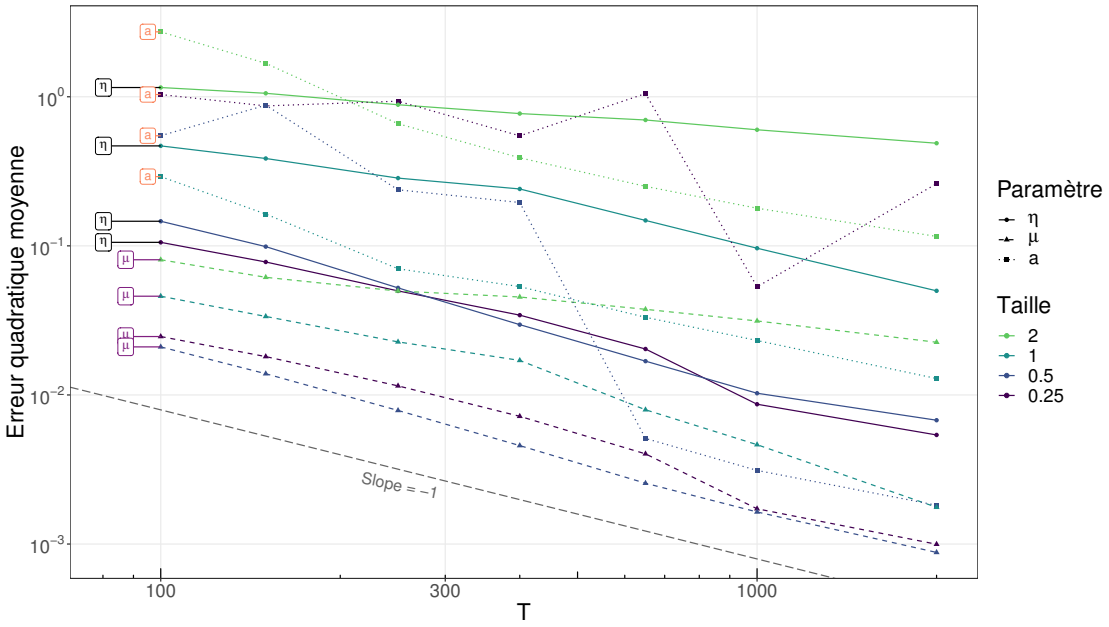


FIGURE 3.A.6 – Erreur quadratique moyenne pour les estimations des paramètres η, μ et a pour 1000 simulations du processus de Hawkes linéaire avec noyau de reproduction $h_1^*(t) = a^1 t^{-2}$ sur l'intervalle $[0, T]$, à l'échelle log-log. La droite grise pointillée représente la pente idéale de -1, *i.e.* une vitesse de convergence en $\mathcal{O}(n^{-1})$.

3.B. Preuve du Théorème 1

Cette annexe est directement issue de l'article *Strong mixing condition for Hawkes processes and application to Whittle estimation from count data*, que nous avons soumis pour publication. Elle est conservée en anglais. Dans un premier temps, nous rappelons les notations relatives à la représentation sous forme de processus de branchement du processus de Hawkes, et introduisons celles relatives aux mesures de comptage du processus de branchement. Ensuite, nous prouvons le Théorème 1.

Notations

The linear Hawkes process is a specific case of the Poisson *cluster process* (Hawkes et Oakes, 1974). Briefly, the process consists of a stream of *immigrants*, the cluster centres, which arrive according to a Poisson process N_c with intensity measure η. Then, an immigrant at time T_i generates *offsprings* according to an inhomogenous Poisson process $N_1(\cdot|T_i)$ with intensity measure $h(\cdot - T_i)$. These in turn independently generate further offsprings according to the same law, and so on *ad infinitum*. The *branching processes* $N(\cdot|T_i)$, consisting of an immigrant at time T_i and all their *descendants*, are therefore independent. Finally, the Hawkes process N is defined as the superposition of all branching processes :

$$\forall A \in \mathcal{B}(\mathbb{R}), N(A) = N_c\Big(N(A|\cdot)\Big).$$

This cluster representation links to the usual Galton–Watson theory. Without loss of generality, consider one branching process whose immigrant has time 0. Define Z_k as the number of points of generation k, *i.e.* $Z_0 = 1$ for the immigrant, then Z_1 denotes the number of offsprings that the immigrant generates, Z_2 the number of offsprings that the offsprings of the immigrants generate, *etc.* Then $(Z_k)_{k\in\mathbb{N}}$ is a Galton–Watson process.

Preuve du Théorème 1

By definition, for a given Hawkes process N, we have

$$\alpha_N(r) \coloneqq \sup_{t\in\mathbb{R}} \alpha\Big(\mathcal{E}_{-\infty}^{t}, \mathcal{E}_{t+r}^{\infty}\Big) = \sup_{t\in\mathbb{R}} \sup_{\substack{A\in\mathcal{E}_{-\infty}^{t} \\ B\in\mathcal{E}_{t+r}^{\infty}}} \Big|\mathrm{Cov}\Big(\mathbb{1}_A(N), \mathbb{1}_B(N)\Big)\Big|,$$

where $\mathbb{1}_{\mathcal{A}}(N)$ is the indicator function of the cylinder set $\mathcal{A}$, *i.e.* for an elementary cylinder set $\mathcal{A}_{B,m} = \{N \in \mathfrak{N} : N(B) = m\}$, $\mathbb{1}_{\mathcal{A}_{B,m}}(N) = 1$ if $N(B) = m$ and 0 otherwise.

We recall that a point process N is said to be positively associated if, for all families of pairwise disjoint Borel sets $(A_i)_{1\leq i\leq k}$ and $(B_j)_{1\leq j\leq l}$, and for all coordinate-wise increasing functions $F : \mathbb{N}^k \to \mathbb{R}$ and $G : \mathbb{N}^l \to \mathbb{R}$, it satisfies

$$\mathrm{Cov}\Big(F\Big(N(A_1),\dots,N(A_k)\Big), G\Big(N(B_1),\dots,N(B_l)\Big)\Big) \geq 0.$$

We start by stating a useful property (see Gao et Zhu, 2018, Section 2.1, key property (e)), which follows from Hawkes processes being infinitely divisible processes :

Proposition 2. *The stationary Hawkes process is positively associated.*

Using this proposition and Poinas *et al.*'s work on associated point processes (Poinas *et al.*, 2019), the following lemma controls the covariance of the indicator functions by the covariance of the count measure of the process, then rescale the problem to a single branching process, thanks to the independence between clusters of a Hawkes process.

Lemme 1. *Let s,t,$u \in \mathbb{R}$ and $r > 0$ such that $s < t < t + r < u$, and let $\mathcal{A} \in \mathcal{E}_s^t, \mathcal{B} \in \mathcal{E}_{t+r}^u$. Then,*

$$\Big|\mathrm{Cov}\Big(\mathbb{1}_{\mathcal{A}}(N), \mathbb{1}_{\mathcal{B}}(N)\Big)\Big| \leq \int \Big|\mathrm{Cov}\Big(N\Big((s,t]\Big|y\Big), N\Big((t+r,u]\Big|y\Big)\Big)\Big| M_c(\mathrm{d}y)$$

where $M_c(\cdot)$ refers to the first-order moment of the centre process N_c.

Démonstration. Using Proposition 2 and (Poinas *et al.*, 2019, Theorem 2.5), we have

$$\Big|\mathrm{Cov}\Big(\mathbb{1}_{\mathcal{A}}(N), \mathbb{1}_{\mathcal{B}}(N)\Big)\Big| \leq \Big|\mathrm{Cov}\Big(N\Big((s,t]\Big), N\Big((t+r,u]\Big)\Big)\Big|.$$

Then, conditioning by the cluster centre process N_c (see for example Daley et Vere-Jones, 2003, Exercise 6.3.4) :

$$\begin{aligned}\mathrm{Cov}\Big(N\Big((s,t]\Big), N\Big((t+r,u]\Big)\Big) &= \int \mathrm{Cov}\Big(N\Big((s,t]\Big|y\Big), N\Big((t+r,u]\Big|y\Big)\Big) M_c(\mathrm{d}y) \\ &\quad + \int \mathbb{E}\Big[N\Big((s,t]\Big|x\Big)\Big]\mathbb{E}\Big[N\Big((t+r,u]\Big|y\Big)\Big] C_c(\mathrm{d}x\times\mathrm{d}y),\end{aligned}$$

where $M_c(\cdot)$ and $C_c(\cdot)$ refer to the first-order moment measure and the covariance measure of the centre process N_c respectively. Since the centre process is Poisson, $C_c \equiv 0$ and the second term is zero. □

We are now interested in deriving an upper bound for the covariance of counts of a single branching process. Without loss of generality, we consider a cluster whose immigrant is located at time 0. Let Z_k denote the number of points of generation k, and by $Z_k^{(s,t]}$ those that are located in the interval $(s,t]$. By definition, we have

$$N\Big((s,t]\Big|0\Big) = \sum_{k=0}^{+\infty} Z_k^{(s,t]}.$$

Then, the covariance between two intervals for a branching process is

$$\mathrm{Cov}\Big(N\Big((s,t]\Big|0\Big), N\Big((t+r,u]\Big|0\Big)\Big) = \sum_{k=0}^{+\infty}\sum_{l=0}^{+\infty} \mathrm{Cov}\left(Z_k^{(s,t]}, Z_l^{(t+r,u]}\right).$$

Before continuing further, we will need a few results on the Galton–Watson process $(Z_k)_{k\in\mathbb{N}}$:

Lemme 2. *The expectation, variance and second-order moment of Z_k are*

$$\begin{aligned}
\mathbb{E}[Z_k] &= \mu^k, \\
\mathrm{Var}(Z_k) &= \mu^k \sum_{j=0}^{k-1} \mu^j = \mu^k \frac{1-\mu^k}{1-\mu}, \\
\mathbb{E}[Z_k^2] &= \mu^k \sum_{j=0}^{k} \mu^j = \mu^k \frac{1-\mu^{k+1}}{1-\mu}.
\end{aligned}$$

Démonstration. Call ϕ_k the probability-generating function of Z_k :

$$\forall s \in [0,1], \phi_k(s) = \mathbb{E}[s^{Z_k}].$$

It is well-known, for a Galton–Watson process, that $(\phi_k)_{k\in\mathbb{N}}$ verifies

$$\forall k \in \mathbb{N}, \phi_{k+1} = \phi_k \circ \phi_1$$

where in our case ϕ_1 is the probability-generating function of a Poisson process with parameter μ. Differentiating the recurrence relation up to order 2 then evaluating it in $s = 1$ gives the following relations :

$$\begin{aligned}
\phi'_{k+1}(1) &= \phi'_1(1)\phi'_k(1), \\
\phi''_{k+1}(1) &= \phi''_1(1)\phi'_k(1) + (\phi'_1(1))^2\phi''_k(1),
\end{aligned}$$

where $\phi'_k(1)$ and $\phi''_k(1)$ are related to the moments of the process by

$$\mathbb{E}[Z_k] = \phi'_k(1), \qquad \mathrm{Var}(Z_k) = \phi''_k(1) + \phi'_k(1) - (\phi'_k(1))^2.$$

Finally plugging in the initial conditions for the Poisson variable Z_1, $\phi'_1(1) = \mu$ and $\phi''_1(1) = \mu^2$, yields the expected result. □

Lemme 3. *The covariance and second-order product moment of* (Z_k) *are*

$$\mathrm{Cov}(Z_k, Z_l) = \mu^{k\vee l} \sum_{j=0}^{k\wedge l-1} \mu^j = \mu^{k\vee l} \frac{1-\mu^{k\wedge l}}{1-\mu},$$
$$\mathbb{E}[Z_k Z_l] = \mu^{k\vee l} \sum_{j=0}^{k\wedge l} \mu^j = \mu^{k\vee l} \frac{1-\mu^{k\wedge l+1}}{1-\mu},$$

where $k \vee l = \max(k,l)$ *and* $k \wedge l = \min(k,l)$.

Démonstration. This is a straightforward recurrence, noting that

$$\begin{aligned}\mathrm{Cov}(Z_k, Z_{k+h}) &= \mathrm{Cov}\left(Z_k, \sum_{i=1}^{+\infty} \mathbb{1}_{\{Z_{k+h-1}\geq i\}} Z_{1,i}\right) \\ &= \mathbb{E}\left[Z_{1,1}\right] \mathrm{Cov}\left(Z_k, \sum_{i=1}^{+\infty} \mathbb{1}_{\{Z_{k+h-1}\geq i\}}\right) \\ &= \mu\, \mathrm{Cov}(Z_k, Z_{k+h-1}).\end{aligned}$$

wherein $Z_{1,i}$ denotes the number of offsprings of the point i of generation $k+h-1$, is independent of $Z_{1,j}$ $(i \neq j)$, of Z_{k+h-1} and of Z_k, and has the same distribution as Z_1. □

Let T_i^k denote the time of arrival of the i-th point of generation k. It has a parent T_j^{k-1} (when $k > 0$). Let Δ_i^k be the associated inter-arrival time, *i.e.* $\Delta_i^k = T_i^k - T_j^{k-1}$. Then, for each point i of generation k, there exists a sequence $(\alpha_{i,k}^{(j)})_{1\leq j\leq k}$, with $\alpha_{i,k}^{(k)} = i$, denoting the indices of the ancestors of T_i^k, such that

$$T_i^k = \sum_{j=1}^{k} \Delta^j_{\alpha_{i,k}^{(j)}}.$$

For the stationary Hawkes process, the Δ_i^k are independent of all other Δ_j^l, and identically distributed according to the measure h^*. As a consequence, we get the following lemma :

Lemme 4. *For* $k \in \mathbb{N}$ *and* $1 \leq i, j \leq Z_k$,

(i) T_i^k *and* T_j^k *are identically distributed, with distribution measure equal to the* k*-multiple convolution of* h^* *with itself,*

(ii) For $\delta > 0$*, there is a upper bound on the* m*-th moment of* T_1^k *:*

$$\mathbb{E}\Big[(T_1^k)^{1+\delta}\Big] \leq k^{1+\delta}\ \mathbb{E}\Big[(\Delta_1^1)^{1+\delta}\Big] = k^{1+\delta}\, \nu_{1+\delta}$$

where $\nu_{1+\delta} \coloneqq \int_{\mathbb{R}} t^{1+\delta} h^*(t)\mathrm{d}t$.

Démonstration. (ii) Using Hölder's inequality :

$$\begin{aligned} T_1^k &= \sum_{j=1}^{k} 1 \cdot \Delta^j_{\alpha^{(j)}_{1,k}} \\ &\leq \left(\sum_{j=1}^{k} 1^{\frac{1+\delta}{\delta}} \right)^{\frac{\delta}{1+\delta}} \cdot \left(\sum_{j=1}^{k} (\Delta^j_{\alpha^{(j)}_{1,k}})^{1+\delta} \right)^{\frac{1}{1+\delta}} \\ &= k^{\frac{\delta}{1+\delta}} \left(\sum_{j=1}^{k} (\Delta^j_{\alpha^{(j)}_{1,k}})^{1+\delta} \right)^{\frac{1}{1+\delta}}. \end{aligned}$$

□

Additionally, since for any point of the branching process offsprings are generated by a Poisson process, the arrival times, say Δ_i^k, are independent from the number of offsprings generated at the current or past generations. Conversely, since the reproduction mean μ does not depend on the time, the number of offsprings generated at any generation, say Z_l, are independent from the past arrival times. Consequently, we have the following lemma :

Lemme 5. *For k,$l \in \mathbb{N}$ and $1 \leq i \leq Z_k$, T_i^k and Z_l are independent.*

Remarque. This lemma separates the genealogy of the Galton–Watson process (Z_k) from the arrival times (T_i^k) of the branching process, analogously to how the Poisson process is a binomial process with Poisson-distributed number of points. Then, a cluster in a stationary Hawkes process is equivalent to a Galton–Watson process (Z_k), upon which the ancestors $(\alpha_{i,k}^{(k-1)})$ are drawn equiprobably from the Z_{k-1} possible ancestors and the (Δ_i^k) independently with distribution function h^*. Intuitively, since each point j of generation $k-1$ generates offsprings according to the same intensity measure, then each point of generation k has ancestor j with equiprobability. This is analogous to the backwards simulation of a Wright-Fisher process without the constant population size restriction.

We state a useful lemma for the covariance of the product of independent random variables.

Lemme 6. *Let $(X_i^k)_{i,k\in\mathbb{N}}$ and $(Y_j^l)_{j,l\in\mathbb{N}}$ be two collections of random variables such that, for all $i, j, k, l \in \mathbb{N}$, the variables X_i^k and Y_j^l are independent. Then*

$$\operatorname{Cov}(X_i^k Y_i^k, X_j^l Y_j^l) = \mathbb{E}[X_i^k X_j^l] \operatorname{Cov}(Y_i^k, Y_j^l) + \mathbb{E}[Y_i^k] \mathbb{E}[Y_j^l] \operatorname{Cov}(X_i^k, X_j^l).$$

Démonstration. Writing the expression of the covariance then adding and substracting the term $\mathbb{E}[X_i^k X_j^l]\,\mathbb{E}[Y_i^k]\,\mathbb{E}[Y_j^l]$ yields the relation. □

We can now derive an upper bound for $\mathrm{Cov}\left(Z_k^{(s,t]}, Z_l^{(t+r,u]}\right)$:

Lemme 7. *Suppose that there exists $\delta > 0$ such that $\nu_{1+\delta} < \infty$, and $l \geq 0$. Then*

$$\left|\mathrm{Cov}\left(Z_k^{(s,t]}, Z_l^{(t+r,u]}\right)\right| \leq 2\,\frac{l^{1+\delta}\,\nu_{1+\delta}}{(t+r)^{1+\delta}}\mu^{k\vee l}\frac{1-\mu^{k\wedge l+1}}{1-\mu}.$$

Démonstration. We have

$$\begin{aligned}\mathrm{Cov}\left(Z_k^{(s,t]}, Z_l^{(t+r,u]}\right) &= \mathrm{Cov}\left(\sum_{i=1}^{Z_k}\mathbb{1}_{\{T_i^k\in(s,t]\}}, \sum_{j=1}^{Z_l}\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right)\\ &= \sum_{i=1}^{+\infty}\sum_{j=1}^{+\infty}\mathrm{Cov}\left(\mathbb{1}_{\{Z_k\geq i\}}\mathbb{1}_{\{T_i^k\in(s,t]\}}, \mathbb{1}_{\{Z_l\geq j\}}\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right).\end{aligned}$$

Then, by Lemmas 5 and 6,

$$\begin{aligned}&\mathrm{Cov}\left(\mathbb{1}_{\{Z_k\geq i\}}\mathbb{1}_{\{T_i^k\in(s,t]\}}, \mathbb{1}_{\{Z_l\geq j\}}\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right)\\ &\quad= \mathbb{E}\left[\mathbb{1}_{\{Z_k\geq i\}}\mathbb{1}_{\{Z_l\geq j\}}\right]\mathrm{Cov}\left(\mathbb{1}_{\{T_i^k\in(s,t]\}}, \mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right)\\ &\quad+ \mathbb{E}\left[\mathbb{1}_{\{T_i^k\in(s,t]\}}\right]\mathbb{E}\left[\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right]\mathrm{Cov}\left(\mathbb{1}_{\{Z_k\geq i\}}, \mathbb{1}_{\{Z_l\geq j\}}\right).\end{aligned}$$

For the first term,

$$\begin{aligned}\mathrm{Cov}\left(\mathbb{1}_{\{T_i^k\in(s,t]\}}, \mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right) &= \mathbb{E}\left[\mathbb{1}_{\{T_i^k\in(s,t]\}}\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{T_i^k\in(s,t]\}}\right]\mathbb{E}\left[\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right]\\ &\leq \mathbb{E}\left[\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right]\\ &\leq \mathbb{P}\left(T_j^l \geq t+r\right)\\ &\leq \frac{\mathbb{E}\left[(T_1^l)^{1+\delta}\right]}{(t+r)^{1+\delta}}\\ &\leq \frac{l^{1+\delta}\,\nu_{1+\delta}}{(t+r)^{1+\delta}},\end{aligned}$$

using Markov's inequality for the second to last inequality, and Lemma 4 for the last one. Similarly,

$$\begin{aligned}\mathrm{Cov}\left(\mathbb{1}_{\{T_i^k\in(s,t]\}}, \mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right) &= \mathbb{E}\left[\mathbb{1}_{\{T_i^k\in(s,t]\}}\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right] - \mathbb{E}\left[\mathbb{1}_{\{T_i^k\in(s,t]\}}\right]\mathbb{E}\left[\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right]\\ &\geq -\mathbb{E}\left[\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right]\\ &\geq -\frac{l^{1+\delta}\,\nu_{1+\delta}}{(t+r)^{1+\delta}},\end{aligned}$$

The second term is straightforward,

$$\left|\mathbb{E}\left[\mathbb{1}_{\{T_i^k\in(s,t]\}}\right]\mathbb{E}\left[\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right]\right| \leq \mathbb{E}\left[\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right] \leq \frac{l^{1+\delta}\,\nu_{1+\delta}}{(t+r)^{1+\delta}}.$$

Then :

$$\begin{aligned}
&\left|\sum_{i=1}^{+\infty}\sum_{j=1}^{+\infty}\mathrm{Cov}\left(\mathbb{1}_{\{Z_k\geq i\}}\mathbb{1}_{\{T_i^k\in(s,t]\}},\mathbb{1}_{\{Z_l\geq j\}}\mathbb{1}_{\{T_j^l\in(t+r,u]\}}\right)\right| \\
&\quad\leq \frac{l^{1+\delta}\,\nu_{1+\delta}}{(t+r)^{1+\delta}}\left|\sum_{i=1}^{+\infty}\sum_{j=1}^{+\infty}\mathbb{E}\left[\mathbb{1}_{\{Z_k\geq i\}}\mathbb{1}_{\{Z_l\geq j\}}\right]+\sum_{i=1}^{+\infty}\sum_{j=1}^{+\infty}\mathrm{Cov}\left(\mathbb{1}_{\{Z_k\geq i\}},\mathbb{1}_{\{Z_l\geq j\}}\right)\right| \\
&\quad= \frac{l^{1+\delta}\,\nu_{1+\delta}}{(t+r)^{1+\delta}}\left|\mathbb{E}\left[Z_kZ_l\right]+\mathrm{Cov}\left(Z_k,Z_l\right)\right| \\
&\quad\leq 2\,\frac{l^{1+\delta}\,\nu_{1+\delta}}{(t+r)^{1+\delta}}\,\mu^{k\vee l}\,\frac{1-\mu^{k\wedge l+1}}{1-\mu},
\end{aligned}$$

using Lemma 3 for the last inequality. □

Straightforwardly, since $\sum\mu^k$ and $\sum l^{1+\delta}\mu^l$ are summable for $\delta>0$, we get the following lemma :

Lemme 8. *Suppose that there exists $\delta>0$ such that $\nu_{1+\delta}<\infty$. Then,*

$$\left|\mathrm{Cov}\Big(N\big((s,t]\big|0\big),N\big((t+r,u]\big|0\big)\Big)\right| = \mathcal{O}\left(\frac{1}{(t+r)^{1+\delta}}\right).$$

All that is left to prove Theorem 1 is to integrate the upper bound with respect to the first-moment measure of the centre process. Using the notations of Lemmas 1 and 8, and with $M_c(\cdot)=\eta\ell(\cdot)$ where $\ell(\cdot)$ is the Lebesgue measure,

$$\begin{aligned}
\left|\mathrm{Cov}\big(\mathbb{1}_{\mathcal{A}}(N),\mathbb{1}_{\mathcal{B}}(N)\big)\right| &\leq \int_{\mathbb{R}}\left|\mathrm{Cov}\Big(N\big((s,t]\big|y\big),N\big((t+r,u]\big|y\big)\Big)\right|M_c(\mathrm{d}y) \\
&= \int_{-\infty}^{t}\left|\mathrm{Cov}\Big(N\big((s,t]\big|y\big),N\big((t+r,u]\big|y\big)\Big)\right|M_c(\mathrm{d}y) \\
&= \mathcal{O}\left(\int_{-\infty}^{t}\frac{1}{(t+r-y)^{1+\delta}}\mathrm{d}y\right) \\
&= \mathcal{O}\left(r^{-\delta}\right).
\end{aligned}$$

This upper bound is valid for any $s,u\in\mathbb{R}$, therefore holds for $\mathcal{A}\in\mathcal{E}_{-\infty}^{t}, \mathcal{B}\in\mathcal{E}_{t+r}^{\infty}$.

■

Bibliographie

O. Aboubakri, N. Khanjani, Y. Jahani, et B. Bakhtiari. Attributable risk of mortality associated with heat and heat waves : A time-series study in Kerman, Iran during 2005–2017. *J. Therm. Biol.*, 82(February) :76–82, 2019.

H. Achebak, D. Devolder, et J. Ballester. Trends in temperature-related age-specific and sex-specific mortality from cardiovascular diseases in Spain : a national time-series analysis. *Lancet Planet. Heal.*, 3(7) :e297–e306, jul 2019.

L. Adamopoulos. Cluster models for earthquakes : Regional comparisons. *J. Int. Assoc. Math. Geol.*, 8(4) :463–475, aug 1976.

A. E. Akkerman, J. C. van der Wouden, M. M. Kuyvenhoven, J. P. Dieleman, et T. J. Verheij. Antibiotic prescribing for respiratory tract infections in Dutch primary care in relation to patient age clinical entities. *J. Antimicrob. Chemother.*, 54(6) : 1116–1121, 2004.

M. G. Alves Galvão, M. A. Rocha Crispino Santos, et A. J. Alves da Cunha. Antibiotics for preventing suppurative complications from undifferentiated acute respiratory infections in children under five years of age. *Cochrane Database Syst. Rev.*, 29(2), feb 2016.

ANSM. Rapport annuel : La consommation d'antibiotiques en France en 2016. Technical report, Agence nationale de sécurité du médicament et des produits de santé, 2017.

A. L. Araujo Navas, F. Osei, R. J. Soares Magalhães, L. R. Leonardo, et A. Stein. Modelling the impact of MAUP on environmental drivers for Schistosoma japonicum prevalence. *Parasites and Vectors*, 13(1) :1–18, 2020.

A. H. Auchincloss, S. Y. Gebreab, C. Mair, et A. V. Diez Roux. A Review of Spatial Methods in Epidemiology, 2000–2010. *Annu. Rev. Public Health*, 33(1) :107–122, apr 2012.

G. Babu et E. Feigelson. *Astrostatistics*. Chapman & Hall, London, 1996.

F. Baccelli, B. Blaszczyszyn, et M. Karray. *Random Measures, Point Processes, and Stochastic Geometry*. 2020.

E. Bacry, I. Mastromatteo, et J.-F. Muzy. Hawkes Processes in Finance. *Mark. Microstruct. Liq.*, 1(1) :1550005, jun 2015.

A. Baddeley. Spatial Point Processes and their Applications. In W. Weil, editor, *Stoch. Geom.*, pages 1–75. Springer, Berlin, Heidelberg, 2007.

F. Barbier et M. Wolff. Multirésistance chez Pseudomonas aeruginosa . *Médecine/Sciences*, 26(11) :960–968, 2010.

D. Bard, R. Barouki, S. Benhamou, J. Bénichou, J. Clavel, E. Jougla, et G. Launoy. Risque attribuable. In *Cancer - Approch. méthodologique du lien avec l'environnement*, pages 69–92. Les éditions Inserm, 2005.

M. S. Bartlett. The Spectral Analysis of Point Processes. *J. R. Stat. Soc. Ser. B*, 25(2) :264–296, 1963.

B. G. Bell, F. Schellevis, E. Stobberingh, H. Goossens, et M. Pringle. A systematic review and meta-analysis of the effects of antibiotic consumption on antibiotic resistance. *BMC Infect. Dis.*, 14(1) :13, dec 2014.

V. Benes, K. Bodlak, J. Moller, et R. Waagepetersen. A case study on point process modelling in disease mapping. *Image Anal. Stereol.*, 24(Copyright 2006, IEE) : 159–168, 2005.

J. Benichou. A review of adjusted estimators of attributable risk. *Stat. Methods Med. Res.*, 10(3) :195–216, 2001.

A. Bernier, E. Delarocque-Astagneau, C. Ligier, M.-A. Vibet, D. Guillemot, et L. Watier. Outpatient antibiotic use in France between 2000 and 2010 : after the nationwide campaign, it is time to focus on the elderly. *Antimicrob. Agents Chemother.*, 58(1) :71–7, jan 2014.

K. Bhaskaran, A. Gasparrini, S. Hajat, L. Smeeth, et B. Armstrong. Time series regression studies in environmental epidemiology. *Int. J. Epidemiol.*, 42(4) : 1187–1195, 2013.

V. Bousquet, C. Caserio-Schönemann, et Comité de pilotage OSCOUR. La surveillance des urgences par le réseau OSCOUR. Technical report, 2013.

C. Bowie et D. Prothero. Finding causes of seasonal diseases using time series analysis. *Int. J. Epidemiol.*, 10(1) :87–92, 1981.

G. E. Box, G. M. Jenkins, G. C. Reinsel, et G. M. Ljung. *Time series analysis : forecasting and control*. John Wiley & Sons, 2015.

R. C. Bradley. Basic properties of strong mixing conditions. A survey and some open questions. *Probab. Surv.*, 2(1) :107–144, 2005.

P. T. Brandt et J. T. Williams. A Linear Poisson Autoregressive Model. *Polit. Anal.*, 9(1997) :164–184, 2000.

A. Briant, P. P. Combes, et M. Lafourcade. Dots to boxes : Do the size and shape of spatial units jeopardize economic geography estimations ? *J. Urban Econ.*, 67 (3) :287–302, 2010.

T. Britton. Stochastic epidemic models : A survey. *Math. Biosci.*, 225(1) :24–35, may 2010.

A. Burton, D. G. Altman, P. Royston, et R. L. Holder. The design of simulation studies in medical statistics. *Stat. Med.*, 25(24) :4279–4292, dec 2006.

N. Caillère, C. Caserio-Schönemann, N. Fournet, A. Fouillet, D. Pateron, C. Leroy, et L. Josseran. Surveillance des urgences - Réseau OSCOUR (Organisation de la surveillance coordonnée des urgences) - Résultats nationaux 2004/2011. Technical report, 2011.

A. Cameron et P. K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press, 1998.

J. Carlet et P. Le Coz. Report from the Special Working Group for Keeping Antibiotics Effective. A report for the French Health Minister. Technical report, Ministère des affaires sociales, de la santé et des droits des femmes, 2015.

F. Carrat et A. J. Valleron. Influenza mortality among the elderly in France, 1980-90 : how many deaths may have been avoided through vaccination ? *J. Epidemiol. Community Health*, 49(4) :419–25, aug 1995.

A. Cassini, L. D. Högberg, D. Plachouras, A. Quattrocchi, A. Hoxha, G. S. Simonsen, M. Colomb-Cotinat, M. E. Kretzschmar, B. Devleesschauwer, M. Cecchini, D. A. Ouakrim, T. C. Oliveira, M. J. Struelens, C. Suetens, D. L. Monnet, R. Strauss, K. Mertens, T. Struyf, B. Catry, K. Latour, I. N. Ivanov, E. G. Dobreva, A. Tambic Andraševic, S. Soprek, A. Budimir, N. Paphitou, H. Žemlicková, S. Schytte Olsen, U. Wolff Sönksen, P. Märtin, M. Ivanova, O. Lyytikäinen, J. Jalava, B. Coignard, T. Eckmanns, M. Abu Sin, S. Haller, G. L. Daikos, A. Gikas, S. Tsiodras, F. Kontopidou, Á. Tóth, Á. Hajdu, Ó. Guólaugsson, K. G. Kristinsson, S. Murchan, K. Burns, P. Pezzotti, C. Gagliotti, U. Dumpis, A. Liuimiene, M. Perrin, M. A. Borg, S. C. de Greeff, J. C. Monen, M. B. Koek, P. Elstrøm, D. Zabicka, A. Deptula, W. Hryniewicz, M. Caniça, P. J. Nogueira, P. A. Fernandes, V. Manageiro, G. A. Popescu, R. I. Serban, E. Schréterová, S. Litvová, M. Štefkovicová, J. Kolman, I. Klavs, A. Korošec, B. Aracil, A. Asensio, M. Pérez-Vázquez, H. Billström, S. Larsson, J. S. Reilly, A. Johnson, et S. Hopkins. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European Economic Area in 2015 : a population-level modelling analysis. *Lancet Infect. Dis.*, 19(1) :56–66, jan 2019.

G. Celeux, D. Chauveau, et J. Diebolt. On Stochastic Versions of the EM Algorithm. Technical Report RR-2514, INRIA, 1995.

Centers for Disease Control and Prevention. *Epidemiology and Prevention of Vaccine-Preventable Diseases*. Public Health Foundation, Washington D.C., 13 edition, 2015.

P. Chahwakilian, B. Huttner, B. Schlemmer, et S. Harbarth. Impact of the French campaign to reduce inappropriate ambulatory antibiotic use on the prescription and consultation rates for respiratory tract infections. *J. Antimicrob. Chemother.*, 66(12) :2872–2879, 2011.

V. Charu, G. Chowell, L. S. Palacio Mejia, S. Echevarría-Zuno, V. H. Borja-Aburto, L. Simonsen, M. A. Miller, et C. Viboud. Mortality burden of the A/H1N1 pandemic in Mexico : A comparison of deaths and years of life lost to seasonal influenza. *Clin. Infect. Dis.*, 53(10) :985–993, 2011.

F. Chen et P. Hall. Inference for a Nonstationary Self-Exciting Point Process with an

Application in Ultra-High Frequency Financial Data Modeling. *J. Appl. Probab.*, 50(04) :1006–1024, dec 2013.

T. Cheng et M. Adepeju. Modifiable temporal unit problem (MTUP) and its effect on space-time cluster detection. *PLoS One*, 9(6) :1–10, 2014.

W.-h. Chiang, X. Liu, et G. Mohler. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. pages 1–19, 2020.

E. S. Chornoboy, L. P. Schramm, et A. F. Karr. Maximum likelihood identification of neural point process systems. *Biol. Cybern.*, 59(4-5) :265–275, 1988.

P. Cole et B. MacMahon. Attributable risk percent in case-control studies. *Br. J. Prev. Soc. Med.*, 25(4) :242–244, nov 1971.

D. Costagliola, A. Flahault, D. Galinec, P. Carmerin, J. Menares, et A.-J. Valleron. A routine tool for detection and assessment of epidemice of influenza-like syndromes in France. *Am. J. Public Health*, 81(11) :97–99, 1991.

C. Costelloe, C. Metcalfe, A. Lovering, D. Mant, et A. D. Hay. Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients : systematic review and meta-analysis. *BMJ*, 340(may18 2) :c2096–c2096, jun 2010.

D. R. Cox. Some Statistical Methods Connected with Series of Events. *J. R. Stat. Soc. Ser. B*, 17(2) :129–157, 1955.

D. R. Cox et P. A. W. Lewis. *The statistical analysis of series of events*. Methuen's Monographs on Applied Probability and Statistics. Springer, Netherlands, 1966.

R. Dahlhaus. Fitting time series models to nonstationary processes. *Ann. Stat.*, 25 (1) :1–37, 1997.

D. J. Daley et D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Probability and its Applications. Springer-Verlag, New York, 2003.

S. J. Dark et D. Bram. The modifiable areal unit problem (MAUP) in physical geography. *Prog. Phys. Geogr.*, 31(5) :471–479, 2007.

Y. A. Davydov. Mixing Conditions for Markov Chains. *Theory Probab. Its Appl.*, 18(2) :312–328, mar 1974.

A. R. J. Dekker, T. J. M. Verheij, et A. W. van der Velden. Inappropriate antibiotic prescription for respiratory tract indications : most prominent in adult patients. *Fam. Pract.*, 32(4) :cmv019, apr 2015.

B. Delyon, M. Lavielle, et E. Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Ann. Stat.*, 27(1) :94–128, 1999.

C. Deng, Z. Ding, L. Li, Y. Wang, P. Guo, S. Yang, J. Liu, Y. Wang, et Q. Zhang. Burden of non-accidental mortality attributable to ambient temperatures : a time series study in a high plateau area of southwest China. *BMJ Open*, 9(2) :e024708, feb 2019.

R. C. Dicker, F. Coronado, D. Koo, et R. G. Parrish. *Principles of Epidemiology in Public Health Practice ; An Introduction to Applied Epidemiology and Biostatistics*. Centers for Disease Control and Prevention, 2006.

P. J. Diggle. A Point Process Modelling Approach to Raised Incidence of a Rare Phenomenon in the Vicinity of a Prespecified Point. *J. R. Stat. Soc. Ser. A (Statistics Soc.*, 153(3) :349–362, 1990.

P. J. Diggle et A. G. Chetwynd. Second-Order Analysis of Spatial Clustering for Inhomogeneous Populations. *Biometrics*, 47(3) :1155–1163, sep 1991.

P. J. Diggle, P. Moraga, B. Rowlingson, et B. M. Taylor. Spatial and Spatio-Temporal Log-Gaussian Cox Processes : Extending the Geostatistical Paradigm. *Stat. Sci.*, 28(4) :542–563, 2013.

F. C. K. Dolk, K. B. Pouwels, D. R. M. Smith, J. V. Robotham, et T. Smieszek. Antibiotics in primary care in England : which antibiotics are prescribed and for which conditions ? *J. Antimicrob. Chemother.*, 73(Suppl 2) :ii2–ii10, feb 2018.

J. L. Doob. The Limiting Distributions of Certain Statistics. *Ann. Math. Stat.*, 6 (3) :160–169, sep 1935.

P. Doukhan. *Mixing : Properties and Examples*. Springer-Verlag, New York, 1994.

K. Dzhaparidze. *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. Springer Series in Statistics. Springer New York, New York, NY, 1986.

K. O. Dzhaparidze. A New Method for Estimating Spectral Parameters of a Stationary Regular Time Series. *Theory Probab. Its Appl.*, 19(1) :122–132, dec 1974.

D. Eddelbuettel et R. François. Rcpp : Seamless R and C++ Integration. *J. Stat. Softw.*, 40(8) :1–18, 2011.

P. Elliott et D. Wartenberg. Spatial epidemiology : Current approaches and future challenges. *Environ. Health Perspect.*, 112(9) :998–1006, 2004.

P. Elliott, J. Wakefield, N. Best, et D. Briggs. *Spatial Epidemiology : Methods and Applications*. Oxford University Press, Oxford, 2000.

European Centre for Disease Prevention and Control. Surveillance of antimicrobial resistance in Europe 2018. Technical report, ECDC, Stockholm, 2019.

M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, et L. Song. Shaping social activity by incentivizing users. *Adv. Neural Inf. Process. Syst.*, 3(January) : 2474–2482, 2014.

R. Farley, G. K. Spurling, L. Eriksson, et C. B. Del Mar. Antibiotics for bronchiolitis in children under two years of age. *Cochrane Database Syst. Rev.*, (10), oct 2014.

A. Fenelon et S. H. Preston. Estimating Smoking-Attributable Mortality in the United States. *Demography*, 49(3) :797–818, aug 2012.

A. Flahault, E. Boussard, J.-F. Vibert, et A.-J. Valleron. Sentiweb remains efficient tool for nationwide surveillance of disease. *BMJ*, 314(7091) :1418–1418, may 1997.

K. E. Fleming-Dutra, A. L. Hersh, D. J. Shapiro, M. Bartoces, E. A. Enns, T. M. File, J. A. Finkelstein, J. S. Gerber, D. Y. Hyun, J. A. Linder, R. Lynfield, D. J. Margolis, L. S. May, D. Merenstein, J. P. Metlay, J. G. Newland, J. F. Piccirillo, R. M. Roberts, G. V. Sanchez, K. J. Suda, A. Thomas, T. M. Woo, R. M. Zetts, et L. A. Hicks. Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011. *JAMA*, 315(17) :1864–1873, may 2016.

J.-M. Floch, E. Marcon, et F. Puech. Les configurations de points. In V. Loonis et M.-P. de Bellefon, editors, *Man. d'analyse Spat. Théorie mise en œuvre Prat. avec R*, pages 73–114. Insee, 2018.

A. Fouillet, G. Rey, F. Laurent, G. Pavillon, S. Bellec, C. Guihenneuc-Jouyaux, J. Clavel, E. Jougla, et D. Hémon. Excess mortality related to the August 2003 heat wave in France. *Int. Arch. Occup. Environ. Health*, 80(1) :16–24, sep 2006.

A. Fouillet, V. Bousquet, I. Pontais, A. Gallay, et C. Caserio-Schönemann. The French Emergency Department OSCOUR Network : Evaluation After a 10-year Existence. *Online J. Public Health Inform.*, 7(1) :11984, feb 2015.

G. French. The continuing crisis in antibiotic resistance. *Int. J. Antimicrob. Agents*, 36 :S3–S7, nov 2010.

X. Gao et L. Zhu. Functional central limit theorems for stationary Hawkes processes and application to infinite-server queues. *Queueing Syst.*, 90(1-2) :161–206, 2018.

A. Gasparrini et M. Leone. Attributable risk from distributed lag models. *BMC Med. Res. Methodol.*, 14(1) :1–8, 2014.

A. Gasparrini, B. Armstrong, et M. G. Kenward. Distributed lag non-linear models. *Stat. Med.*, 29(21) :2224–2234, sep 2010.

A. C. Gatrell, T. C. Bailey, P. J. Diggle, B. S. Rowlingson, et B. S. Rowlingsont. Point Spatial application pattern analysis geographical epidemiology. *Trans. Inst. Br. Geogr.*, 21(1) :256–274, 1996.

C. E. Gehlke et K. Biehl. Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material. *J. Am. Stat. Assoc.*, 29(185) :169, mar 1934.

R. Gilca, G. De Serres, D. Skowronski, G. Boivin, et D. L. Buckeridge. The Need for Validation of Statistical Methods for Estimating Respiratory Virus-Attributable Hospitalization. *Am. J. Epidemiol.*, 170(7) :925–936, oct 2009.

L. Giraitis et M. S. Taqqu. Whittle estimator for finite-variance non-Gaussian time series with long memory. *Ann. Stat.*, 27(1) :178–203, 1999.

R. Gonzales, J. F. Steiner, et M. A. Sande. Antibiotic prescribing for adults with colds, upper respiratory tract infections, and bronchitis by ambulatory care physicians. *JAMA*, 278(11) :901–4, 1997.

H. Goossens, M. Ferech, R. Vander Stichele, et M. Elseviers. Outpatient antibiotic use in Europe and association with resistance : a cross-national database study. *Lancet*, 365(9459) :579–587, feb 2005.

N. Grall, A. Andremont, et L. Armand-Lefèvre. Résistance aux carbapénèmes : vers une nouvelle impasse ? *J. des Anti-Infectieux*, 13(2) :87–102, 2011.

R. L. Grant. Converting an odds ratio to a range of plausible relative risks for better communication of research findings. *BMJ*, 348(jan24 1) :f7450–f7450, jan 2014.

D. Guillemot, C. Carbon, B. Balkau, P. Geslin, H. Lecoeur, F. Vauzelle-Kervroëdan, G. Bouvenot, et E. Eschwége. Low Dosage and Long Treatment Duration of β-Lactam. *JAMA*, 279(5) :365, feb 1998.

M. C. Gulliford, A. Dregan, M. V. Moore, M. Ashworth, T. van Staa, G. McCann, J. Charlton, L. Yardley, P. Little, et L. McDermott. Continued high rates of antibiotic prescribing to adults with respiratory tract infection : survey of 568 UK general practices. *BMJ Open*, 4(10) :e006245, oct 2014.

M. C. Gulliford, M. V. Moore, P. Little, A. D. Hay, R. Fox, A. T. Prevost, D. Juszczyk, J. Charlton, et M. Ashworth. Safety of reduced antibiotic prescribing for self limiting respiratory tract infections in primary care : cohort study using electronic health records. *BMJ*, 354 :i3410, jul 2016.

A. D. Harris, M. H. Samore, M. Lipsitch, K. S. Kaye, E. Perencevich, et Y. Carmeli. Control-Group Selection Importance in Studies of Antimicrobial Resistance : Examples Applied to Pseudomonas aeruginosa, Enterococci, and Escherichia coli . *Clin. Infect. Dis.*, 34(12) :1558–1563, 2002.

A. G. Hawkes. Spectra of Some Self-Exciting and Mutually Exciting Point Processes. *Biometrika*, 58(1) :83–90, 1971a.

A. G. Hawkes. Point Spectra of Some Mutually Exciting Point Processes. *J. R. Stat. Soc. Ser. B*, 33(3) :438–443, 1971b.

A. G. Hawkes et D. Oakes. A cluster process representation of a self-exciting process. *J. Appl. Probab.*, 11(03) :493–503, sep 1974.

L. Heinrich et Z. Pawlas. Absolute regularity and Brillinger-mixing of stationary point processes. *Lith. Math. J.*, 53(3) :293–310, 2013.

D. P. Henriksen, S. L. Nielsen, C. B. Laursen, J. Hallas, C. Pedersen, et A. T. Lassen. How Well Do Discharge Diagnoses Identify Hospitalised Patients with Community-Acquired Infections ? – A Validation Study. *PLoS One*, 9(3) :e92891, mar 2014.

A. B. Hill. The Environment and Disease : Association or Causation ? *Proc. R. Soc. Med.*, 58(5) :295–300, 1965.

Y. Hosoya. *Estimation problems on stationary time series models.* Ph.d. dissertation, Yale University, 1974.

B. Hubert, L. Watier, P. Garnerin, et S. Richardson. Meningococcal Disease and Influenza-like Syndrome : A New Approach to an Old Question. *J. Infect. Dis.*, 166(3) :542–545, sep 1992.

F. Huszár. ML beyond Curve Fitting : An Intro to Causal Inference and do-Calculus, 2018.

F. Huszár. Causal Inference 3 : Counterfactuals, 2019a.

F. Huszár. Causal Inference 2 : Illustrating Interventions via a Toy Example, 2019b.

M. Ieven, S. Coenen, K. Loens, C. Lammens, F. Coenjaerts, A. Vanderstraeten, B. Henriques-Normark, D. Crook, K. Huygen, C. C. Butler, T. J. Verheij, P. Little, K. Zlateva, A. van Loon, E. C. Claas, et H. Goossens. Aetiology of lower respiratory tract infection in adults in primary care : a prospective study in 11 European countries. *Clin. Microbiol. Infect.*, 24(11) :1158–1163, 2018.

S. Jain, W. H. Self, R. G. Wunderink, S. Fakhran, R. Balk, A. M. Bramley, C. Reed, C. G. Grijalva, E. J. Anderson, D. M. Courtney, J. D. Chappell, C. Qi, E. M. Hart, F. Carroll, C. Trabue, H. K. Donnelly, D. J. Williams, Y. Zhu, S. R. Arnold, K. Ampofo, G. W. Waterer, M. Levine, S. Lindstrom, J. M. Winchell, J. M. Katz, D. Erdman, E. Schneider, L. A. Hicks, J. A. McCullers, A. T. Pavia, K. M. Edwards, et L. Finelli. Community-acquired pneumonia requiring hospitalization among U.S. adults. *N. Engl. J. Med.*, 373(5) :415–427, 2015.

L. Josseran, J. Nicolau, N. Caillère, P. Astagneau, et G. Brücker. Syndromic surveillance based on emergency department activity and crude mortality : two examples. *Eurosurveillance*, 11(12) :225–9, 2006.

L. Josseran, A. Fouillet, N. Caillère, D. Brun-Ney, D. Ilef, G. Brucker, H. Medeiros, et P. Astagneau. Assessment of a Syndromic Surveillance System Based on Morbidity Data : Results from the Oscour® Network during a Heat Wave. *PLoS One*, 5(8) :e11984, aug 2010.

O. Kallenberg. *Random Measures*. Academic Press, 1983.

T. Kaplan. The Role of Horizontal Gene Transfer in Antibiotic Resistance. *Eukaryon*, 10(March) :80–81, 2014.

J. E. Kelsall et P. J. Diggle. Non-parametric estimation of spatial variation in relative risk. *Stat. Med.*, 14(21-22) :2335–2342, nov 1995.

M. Kirchner. Hawkes and INAR(∞) processes. *Stoch. Process. their Appl.*, 126(8) : 2494–2525, aug 2016.

M. Kirchner. An estimation procedure for the Hawkes process. *Quant. Financ.*, 17 (4) :571–595, apr 2017.

R. Kitchin et G. Mcardle. Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards. (January 2019), 2015.

T. Kumazawa et Y. Ogata. Nonstationary etas models for nonstandard earthquakes. *Ann. Appl. Stat.*, 8(3) :1825–1852, 2014.

M. Lemaitre, F. Carrat, G. Rey, M. Miller, L. Simonsen, et C. Viboud. Mortality Burden of the 2009 A/H1N1 Influenza Pandemic in France : Comparison to Seasonal Influenza and the A/H3N2 Pandemic. *PLoS One*, 7(9) :1–11, 2012.

H. M. Leung et L. L. Kupper. Comparisons of Confidence Intervals for Attributable Risk. *Biometrics*, 37(2) :293, jun 1981.

M. Lipsitch. Measuring and Interpreting Associations between Antibiotic Use and Penicillin Resistance in Streptococcus pneumoniae. *Clin. Infect. Dis.*, 32(7) : 1044–1054, 2001.

J. Llorca et M. Delgado-Rodríguez. A comparison of several procedures to estimate the confidence interval for attributable risk in case-control studies. *Stat. Med.*, 19(8) :1089–1099, 2000.

J. Ludwig et J. Reynolds. *Statistical Ecology : A Primer on Methods and Computing*. John Wiley & Sons, New York, 1988.

B. J. Marston, J. F. Plouffe, T. M. File, B. A. Hackman, S. J. Salstrom, H. B. Lipman, M. S. Kolczak, et R. F. Breiman. Incidence of community-acquired pneumonia requiring hospitalization : Results of a population-based active surveillance study in Ohio. *Arch. Intern. Med.*, 157(15) :1709–1718, 1997.

G. Matias, R. Taylor, F. Haguinet, C. Schuck-Paim, R. Lustig, et V. Shinde. Estimates of hospitalization attributable to influenza and RSV in the US during 1997-2009, by age and risk status. *BMC Public Health*, 17(1) :1–14, 2017.

L. F. McCaig. Trends in Antimicrobial Drug Prescribing Among Office-Based Physicians in the United States. *JAMA*, 273(3) :214, jan 1995.

R. McKay, A. Mah, M. R. Law, K. McGrail, et D. M. Patrick. Systematic Review of Factors Associated with Antibiotic Prescribing for Respiratory Tract Infections. *Antimicrob. Agents Chemother.*, 60(7) :4106–4118, jul 2016.

J. P. Metlay. Editorial commentary : Setting national targets for antibiotic use. *Clin. Infect. Dis.*, 60(9) :1317–1318, 2015.

S. Meyer, J. Elias, et M. Höhle. A Space-Time Conditional Intensity Model for Invasive Meningococcal Disease Occurrence. *Biometrics*, 68(2) :607–616, 2012.

O. S. Miettinen. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am. J. Epidemiol.*, 99(5) :325–332, 1974.

Ministère chargé de la Santé. Plan national d'alerte sur les antibiotiques 2011-2016. Technical report, 2011.

S. Mishra et S. D. Baral. Rethinking the population attributable fraction for infectious diseases. *Lancet Infect. Dis.*, 20(2) :155–157, 2020.

J. Møller et R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes*. Chapman & Hall/CRC, 2004.

D. J. Muscatello, A. T. Newall, D. E. Dwyer, et C. R. MacIntyre. Mortality Attributable to Seasonal and Pandemic Influenza, Australia, 2003 to 2009, Using a Novel Time Series Smoothing Approach. *PLoS One*, 8(6) :e64734, jun 2013.

M. Nadeem Ahmed, M. M. Muyot, S. Begum, P. Smith, C. Little, et F. J. Windemuller. Antibiotic Prescription Pattern for Viral Respiratory Illness in Emergency Room and Ambulatory Care Settings. *Clin. Pediatr. (Phila).*, 49(6) :542–547, jun 2010.

H. Nishiura, K. Mizumoto, et Y. Asai. Assessing the transmission dynamics of measles in Japan, 2016. *Epidemics*, 20(May 2010) :67–72, 2017.

B. Nunes, C. Viboud, A. Machado, C. Ringholz, H. Rebelo-de Andrade, P. Nogueira, et M. Miller. Excess mortality associated with influenza epidemics in Portugal, 1980 to 2004. *PLoS One*, 6(6) :e20661, 2011.

D. Oakes. The Markovian self-exciting process. *J. Appl. Probab.*, 12(01) :69–77, mar 1975.

G. W. Oehlert. A Note on the Delta Method. *Am. Stat.*, 46(1) :27–29, feb 1992.

Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann. Inst. Stat. Math.*, 30(1) :243–261, 1978.

Y. Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Assoc.*, 83(401) :9–27, 1988.

J. Ohser et F. Mücklich. *Statistical Analysis of Microstructures in Materials Science*. John Wiley & Sons, Chichester, 2000.

J. F. Olson et K. M. Carley. Exact and approximate EM estimation of mutually exciting hawkes processes. *Stat. Inference Stoch. Process.*, 16(1) :63–80, 2013.

L. Opatowski, E. Varon, C. Dupont, L. Temime, S. van der Werf, L. Gutmann, P.-Y. Boëlle, L. Watier, et D. Guillemot. Assessing pneumococcal meningitis association with viral respiratory infections and antibiotics : insights from statistical and mathematical models. *Proc. Biol. Sci.*, 280(1764) :20130519, 2013.

M. Opatowski, P. Tuppin, K. Cosker, M. Touat, G. De Lagasnerie, D. Guillemot, J. Salomon, C. Brun-Buisson, et L. Watier. Hospitalisations with infections related to antimicrobial-resistant bacteria from the French nationwide hospital discharge database, 2016. *Epidemiol. Infect.*, 147 :1–9, 2019.

S. Openshaw. The modifiable areal unit problem. *Concepts Tech. Mod. Geogr.*, 38 : 1–41, 1984.

S. Openshaw et P. J. Taylor. A million or so correlation coefficients : three experiments on the modifiable areal unit problem. In Wrigley, editor, *Stat. Appl. Spat. Sci.*, pages 127–44. London, 1979.

T. Ozaki et Y. Ogata. Maximum likelihood estimation of Hawkes' self-exciting point processes. *Ann. Inst. Stat. Math.*, 31(1) :145–155, dec 1979.

M. P. Parenteau et M. C. Sawada. The modifiable areal unit problem (MAUP) in the relationship between exposure to NO2 and respiratory health. *Int. J. Health Geogr.*, 10(1) :58, 2011.

S. Paynter. Incorporating Transmission into Causal Models of Infectious Diseases for Improved Understanding of the Effect and Impact of Risk Factors. *Am. J. Epidemiol.*, 183(6) :574–582, 2016.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl. The Foundations of Causal Inference. *Sociol. Methodol.*, 40(1) :75–149, aug 2010.

C. Pelat, I. Bonmarin, M. Ruello, A. Fouillet, C. Caserio-Schönemann, D. Levy-Bruhl, Y. Le Strat, O. Retel, B. Hubert, F. Golliot, L. King, I. M. Njoya, C. Saura, et L. Filleul. Improving regional influenza surveillance through a combination of automated outbreak detection methods : The 2015/16 season in France. *Eurosurveillance*, 22(32) :1–10, 2017.

D. B. Percival et A. T. Walden. Introduction to Spectral Analysis. In *Spectr. Anal. Univariate Time Ser.*, pages 1–29. 2009.

J. B. Perrin, C. Ducrot, J. L. Vinard, E. Morignat, A. Gauffier, D. Calavas, et P. Hendrikx. Using the National Cattle Register to estimate the excess morta-

lity during an epidemic : Application to an outbreak of Bluetongue serotype 8. *Epidemics*, 2(4) :207–214, 2010.

A. Poinas, B. Delyon, et F. Lavancier. Mixing properties and central limit theorem for associated point processes. *Bernoulli*, 25(3) :1724–1754, aug 2019.

K. B. Pouwels, F. C. K. Dolk, D. R. M. Smith, J. V. Robotham, et T. Smieszek. Actual versus 'ideal' antibiotic prescribing for common conditions in English primary care. *J. Antimicrob. Chemother.*, 73(Suppl 2) :ii19–ii26, feb 2018.

R Core Team. R : A Language and Environment for Statistical Computing, 2019.

R Core Team. *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

J. H. Rex. ND4BB : addressing the antimicrobial resistance crisis. *Nat. Rev. Microbiol.*, 12(4) :231–232, apr 2014.

P. Reynaud-Bouret et S. Schbath. Adaptive estimation for hawkes processes ; Application to genome analysis. *Ann. Stat.*, 38(5) :2781–2822, 2010.

E. Rio. *Asymptotic Theory of Weakly Dependent Random Processes*, volume 80 of *Probability Theory and Stochastic Modelling*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2017.

M.-A. Rizoiu, S. Mishra, Q. Kong, M. Carman, et L. Xie. SIR-Hawkes. In *Proc. 2018 World Wide Web Conf.*, pages 419–428, New York, New York, USA, 2018. ACM Press.

A. S. Robbins, S. Y. Chao, et V. P. Fonseca. What's the Relative Risk ? A Method to Directly Estimate Risk Ratios in Cohort Studies of Common Outcomes. *Ann. Epidemiol.*, 12(7) :452–454, 2002.

M. Rosenblatt. A Central Limit Theorem and a Strong Mixing Condition. *Proc. Natl. Acad. Sci.*, 42(1) :43–47, jan 1956.

K. J. Rothman et S. Greenland. Causation and Causal Inference in Epidemiology. *Am. J. Public Health*, 95(Suppl 1) :144–150, 2005.

F. Roueff, R. von Sachs, et L. Sansonnet. Locally stationary Hawkes processes. *Stoch. Process. their Appl.*, 126(6) :1710–1743, jun 2016.

V. Roussel, T. Tritz, C. Souty, C. Turbelin, C. Arena, B. Lambert, A. Lillo-LeLouët, S. Kernéis, T. Blanchon, et T. Hanslik. Estimating the excess of inappropriate prescriptions of anti-dopaminergic anti-emetics during acute gastroenteritis epidemics in France. *Pharmacoepidemiol. Drug Saf.*, 22(10) :1080–1085, aug 2013.

O. Ruuskanen, E. Lahti, L. C. Jennings, et D. R. Murdoch. Viral pneumonia. *Lancet*, 377(9773) :1264–1275, apr 2011.

E. Sabuncu, J. David, C. Bernède-Bauduin, S. Pépin, M. Leroy, P.-Y. Boëlle, L. Watier, et D. Guillemot. Significant reduction of antibiotic use in the community after a nationwide campaign in France, 2002-2007. *PLoS Med.*, 6(6) :e1000084, jun 2009.

M. H. Samore, M. Lipsitch, S. C. Alder, B. Haddadin, G. Stoddard, J. Williamson, K. Sebastian, K. Carroll, O. Ergonul, Y. Carmeli, et M. A. Sande. Mechanisms by which antibiotics promote dissemination of resistant pneumococci in human populations. *Am. J. Epidemiol.*, 163(2) :160–170, 2006.

A. Schuchat, K. Robinson, J. D. Wenger, L. H. Harrison, M. Farley, A. L. Reingold, L. Lefkowitz, et B. A. Perkins. Bacterial Meningitis in the United States in 1995. *N. Engl. J. Med.*, 337(14) :970–976, oct 1997.

J. Schwartz et A. Marcus. Mortality and air pollution in London : a time series analysis. *Am. J. Epidemiol.*, 131(1) :185–94, jan 1990.

R. E. Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public Health Rep.*, 78(6) :494–506, 1963.

X. Shao. A self-normalized approach to confidence interval construction in time series. *J. R. Stat. Soc. Ser. B*, 72(3) :343–366, jun 2010.

M. Sharland. The use of antibacterials in children : A report of the specialist advisory committee on antimicrobial resistance (SACAR) paediatric subgroup. *J. Antimicrob. Chemother.*, 60(SUPPL. 1) :15–26, 2007.

R. H. Shumway et D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer, 4 edition, 2011.

M. Silverman, M. Povitz, J. M. Sontrop, L. Li, L. Richard, S. Cejic, et S. Z. Shariff. Antibiotic Prescribing for Nonbacterial Acute Upper Respiratory Infections in Elderly Persons. *Ann. Intern. Med.*, 166(11) :765, jun 2017.

L. Simonsen, T. A. Reichert, C. Viboud, W. C. Blackwelder, R. J. Taylor, et M. A. Miller. Impact of Influenza Vaccination on Seasonal Mortality in the US Elderly Population. *Arch. Intern. Med.*, 165 :265–272, 2005.

D. R. M. Smith, F. C. K. Dolk, K. B. Pouwels, M. Christie, J. V. Robotham, et T. Smieszek. Defining the appropriateness and inappropriateness of antibiotic prescribing in primary care. *J. Antimicrob. Chemother.*, 73(Suppl 2) :ii11–ii18, feb 2018.

A. Sommet, C. Sermet, P. Y. Boëlle, M. Tafflet, C. Bernède, et D. Guillemot. No significant decrease in antibiotic use from 1992 to 2000, in the French community. *J. Antimicrob. Chemother.*, 54(2) :524–528, 2004.

M. E. Sousa-Vieira. Applicability of the Whittle estimator to non-stationary and non-linear long-memory processes. *J. Simul.*, 10(3) :182–192, aug 2016.

D. Sun, K. Jeannot, Y. Xiao, et C. W. Knapp. Editorial : Horizontal Gene Transfer Mediated Bacterial Antibiotic Resistance. *Front. Microbiol.*, 10(3) :565–591, aug 2019.

M. S. Taqqu et V. Teverovsky. Robustness of Whittle-type estimators for time series with long-range dependence. *Commun. Stat. Part C Stoch. Model.*, 13(4) : 723–757, 1997.

S. B. Thacker et R. L. Berkelman. Public health surveillance in the United States. *Epidemiol. Rev.*, 10(1) :164–190, 1988.

W. W. Thompson, E. Weintraub, P. Dhankhar, P. Y. Cheng, L. Brammer, M. I. M. I. Meltzer, J. S. Bresee, et D. K. Shay. Estimates of US influenza-associated deaths made using four different methods. *Influenza Other Respi. Viruses*, 3(1) : 37–49, 2009.

N. T. H. Trinh, P. Chahwakilian, T. A. Bruckner, S. Sclison, C. Levy, M. Chalumeau, D. Milic, R. Cohen, et J. F. Cohen. Discrepancies in national time trends of

outpatient antibiotic utilization using different measures : a population-based study in France. *J. Antimicrob. Chemother.*, 73(5) :1395–1401, may 2018.

R. R. Tucci. Introduction to Judea Pearl's Do-Calculus. *arXiv e-prints*, page arXiv :1305.5506, apr 2013.

A. J. Valleron, E. Bouvet, P. Garnerin, J. Ménarès, I. Heard, S. Letrait, et J. Lefaucheux. A computer network for the surveillance of communicable diseases : The French experiment. *Am. J. Public Health*, 76(11) :1289–1292, 1986.

N. van de Sande-Bruinsma, H. Grundmann, D. Verloo, E. Tiemersma, J. Monen, H. Goossens, et M. Ferech. Antimicrobial Drug Use and Resistance in Europe. *Emerg. Infect. Dis.*, 14(11) :1722–1730, nov 2008.

J. M. Ver Hoef et P. L. Boveng. Quasi-Poisson vs. Negative Binomial Regression : how should we model overdispersed count data ? *Ecology*, 88(11) :2766–2772, nov 2007.

A. J. Viera. Odds Ratios and Risk Ratios : What's the Difference and Why Does It Matter ? *South. Med. J.*, 101(7) :730–734, jul 2008.

S. D. Walter. The Estimation and Interpretation of Attributable Risk in Health Research. *Biometrics*, 32(4) :829, 1976.

S. D. Walter. Prevention for Multifactorial Diseases. *Am. J. Epidemiol.*, 112(3) : 409–416, sep 1980.

S. D. Walter. Effects of interaction, confounding and observational error on attributable risk estimation. *Am. J. Epidemiol.*, 117(5) :598–604, may 1983.

L. Watier, P. Cavalié, B. Coignard, et C. Brun-Buisson. Comparing antibiotic consumption between two European countries : Are packages an adequate surrogate for prescriptions ? *Eurosurveillance*, 22(46) :1–6, 2017.

M. Westcott. On Existence and Mixing Results for Cluster Point Processes. *J. R. Stat. Soc. Ser. B*, 33(2) :290–300, 1971.

M. Westcott. The probability generating functional. *J. Aust. Math. Soc.*, 14(4) : 448–466, 1972.

S. Wheatley, V. Filimonov, et D. Sornette. The Hawkes process with renewal immigration & its estimation with an EM algorithm. *Comput. Stat. Data Anal.*, 94 : 120–135, feb 2016.

P. Whittle. Some results in time series analysis. *Scand. Actuar. J.*, 1952(1-2) : 48–60, jan 1952.

World Health Organization. Global Action Plan on Antimicrobial Resistance. Technical report, WHO, Geneva, 2015.

H. Xu, W. Wu, S. Nemati, et H. Zha. Patient flow prediction via discriminative learning of mutually-correcting processes. *IEEE Trans. Knowl. Data Eng.*, 29 (1) :157–171, apr 2016.

A. Zeileis. Econometric Computing with HC and HAC Covariance Matrix Estimators. *J. Stat. Softw.*, 11(10) :128–129, 2004.

J. Zhang et K. F. Yu. What's the Relative Risk ? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes. *JAMA*, 280(19) :1690, nov 1998.

Q. Zhao, M. A. Erdogdu, H. Y. He, A. Rajaraman, et J. Leskovec. SEISMIC : A self-exciting point process model for predicting tweet popularity. *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2015-Augus :1513–1522, 2015.

H. Zhou, W. W. Thompson, C. G. Viboud, C. M. Ringholz, P. Y. Cheng, C. Steiner, G. R. Abedi, L. J. Anderson, L. Brammer, et D. K. Shay. Hospitalizations associated with influenza and respiratory syncytial virus in the United States, 1993-2008. *Clin. Infect. Dis.*, 54(10) :1427–1436, 2012.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique (EDSP)

Titre : Maladies infectieuses et données agrégées : estimation de la fraction attribuable et prise en compte de biais

Mots clés : Maladies infectieuses, Données agrégées, Fraction attribuable, Série temporelle, Processus de Hawkes

Résumé : La surveillance épidémiologique repose le plus souvent sur l'analyse d'indicateurs de santé agrégés. Nous étudions les problèmes méthodologiques rencontrés lorsque l'on travaille sur ce type de données dans un contexte de santé publique. Dans un premier temps, nous nous intéressons au calcul de la fraction attribuable lorsque l'exposition est épidémique et le nombre d'événements de santé saisonnier. Pour les modèles statistiques de séries temporelles les plus souvent utilisés, nous présentons une méthode d'estimation de cette fraction et de ses intervalles de confiance. Ce travail nous a permis de montrer que la campagne de sensibilisation "Les antibiotiques, c'est pas automatique !" avait conduit à une diminution de plus de moitié des prescriptions antibiotiques associées aux épidémies de syndromes grippaux dès 2005. Par ailleurs, récemment 17% des prescriptions seraient attribuables aux infections virales des voies respiratoires basses pendant la période hivernale, et près de 38% chez les enfants, dont la moitié attribuables aux bronchiolites. Dans un second temps, nous proposons les processus de Hawkes comme modèles pour les maladies contagieuses et étudions l'impact de l'agrégation des données sur leur estimation. Dans ce contexte, nous développons une méthode d'estimation des paramètres du processus et prouvons que les estimateurs ont de bonnes propriétés asymptotiques. Ces travaux fournissent des outils statistiques pour éviter certains biais dus à l'agrégation de données individuelles pour l'étude de fractions attribuables et de maladies contagieuses.

Title : Infectious diseases and aggregate data : estimating attributable fractions and controlling for bias

Keywords : Infectious diseases, Aggregate data, Attributable fraction, Time series, Hawkes process

Abstract : Epidemiological surveillance is most often based on the analysis of aggregate health indicators. We study the methodological problems encountered when working with this type of data in a public health context. First, we focus on calculating the attributable fraction when the exposure is epidemic and the number of health events exhibits a seasonality. For the most frequently used time series models, we present a method for estimating this fraction and its confidence intervals. This work enabled us to show that the awareness campaign "Antibiotics are not automatic !" led to a reduction of more than half of the antibiotic prescriptions associated with influenza epidemics as early as 2005. Moreover, recently 17% of prescriptions are thought to be attributable to viral infections of the lower respiratory tract during the cold period, and nearly 38% in children, half of which attributable to bronchiolitis. In a second step, we propose Hawkes processes as models for contagious diseases and study the impact of data aggregation on their estimation. In this context, we develop a method for estimating the process parameters and prove that the estimators have good asymptotic properties. This work provides statistical tools to avoid some biases due to the use of aggregate data for the study of attributable fractions and contagious diseases.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France