Modelling the Risks of Measles Outbreaks Near Elimination

Alexis Robert

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy
of the
University of London

May 2021

Department of Infectious Disease Epidemiology

Faculty of Epidemiology and Population Health

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by the UK Medical Research Council (Grant number: MR/N013638/1)

Research group affiliation(s):
Centre for the Mathematical Modelling of Infectious Diseases

Declaration of own work

I Alexis Robert, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Alexis Robert, 31st May 2021

Abstract

Although the global burden of measles has been substantially reduced since the introduction of the first measles vaccine in the 1960s, large outbreaks continue to affect populations in every WHO region. Even in countries with a high national vaccine uptake, social and spatial heterogeneity in coverage lead to under-immunised populations, where the importation of cases can cause large transmission clusters. This thesis explores how local transmission risk can be identified using different data sources: i) routinely collected individual-level case surveillance data, and ii) population-level factors such as vaccination coverage and recent outbreaks.

In the absence of regular sub-national serological surveys, transmission trees from previous outbreaks can be used to identify areas repeatedly associated with transmission events. I developed the R package o2geosocial to reconstruct who infected whom from routinely collected surveillance data, and to compute the number of cases per transmission cluster, i.e. the cluster size distribution. This method infers the infector, infection date and importation status of each case using their onset date, location, age, and genotype. In the first chapter of the thesis, I outlined the methodology implemented in the package and applied it to simulated local outbreaks. The method was able to reconstruct the simulated transmission dynamics and highlighted regions repeatedly associated with secondary transmission. In the second chapter, I applied o2geosocial to data from the national measles database in the United States, which lists cases reported between 2001 and 2016. Both studies illustrated the ability of this method to reconstruct transmission history from widely collected epidemiological information in a variety of contexts and geographical scales.

Countries become eligible for certification of World Health Organisation's measles elimination status after national transmission is interrupted for three years in the presence of high national vaccine coverage. Recent major outbreaks in countries where measles had been declared eliminated (e.g., United Kingdom, Brazil, Greece) illustrate that current indicators of elimination may be imperfect predictors of outbreak risk. In the third and fourth chapters of this thesis, I studied the impact of recent levels of local incidence and vaccine uptake on the risks of importation, cross-regional and local transmissions by implementing a time-series model using the R package surveillance. I applied this model using the daily number of cases reported in France between 2009 and 2017 and discussed how local indicators can inform the risks of national outbreaks.

Acknowledgements

First and foremost, I would like to thank my supervisors, Dr Adam Kucharski and Pr Sebastian Funk, for the help and support they have given me throughout the past four years, from writing up the application to the submission of this PhD. I'm deeply grateful for their advice, and for offering me a chance to work on such a fascinating topic. Their insights and encouragement have constantly helped me become a better researcher and these projects would not have been possible without their guidance.

I would also like to thank the people I have collaborated with during the various projects of this thesis. To Dr Paul Gastañaduy for welcoming me for two weeks in the Centers for Disease Control and Prevention, and being so generous with his time to discuss and set up the first project of this PhD. To Dr Helen Johnson, for inviting me to work together at the European Centre for Disease Prevention and Control, and giving her time and resources to develop this research project. Finally, to Pr Jacco Wallinga for agreeing to supervise me during my three-month research internship, and whose passion and enthusiasm have been an incredible example for me.

This research would also not have been possible without the generous funding support of the Medical Research Council.

I first joined the Center for Mathematical Modelling of Infectious Diseases as an MSc intern in 2015, and many people have been essential in giving me the passion and interest for this field. Firstly, I would like to thank Dr Anton Camacho, who was my first supervisor here, and was patient and kind enough to give me an opportunity to work on mathematical modelling projects. His unshakeable dedication to keeping a healthy balance of work and fun by maintaining regular pub trips have set the best possible example for me to try and follow. Secondly, I thank Dr Rosalind Eggo for being such an incredible mentor when I was a research assistant, this PhD project would not have been the same without the multitude of advice and the guidance she gave me during this first year. Finally, I owe a great deal of gratitude to Pr John Edmunds, whose support and mentorship during the last five years have been invaluable. I would also like to thank all the colleagues I had the chance to sit with in Room 107 at a time where I had just arrived in London, and who made my integration in this team seem so easy, despite my, back then, somewhat limited understanding of both the field and the language.

The last three years would not have been the same without the support and solidarity of the many PhD students I met at LSHTM. To Naomi, for being the best co-lead an ECR coordinator could wish for; and to Ali, for being the best ECR coordinator an ECR could wish for. To Richard for our always insightful discussion and his encouragement. To Charlotte, Isabel, Sophie, Quentin, James, Chris, and Akira for

making LG22 such a supportive and fun place to work in. To Orlagh, Kelly, and Emily, for always being up for a pumphandle pub trip.

Bringing this PhD to fruition would also have been impossible without the support of many friends here and overseas, so a massive thank you to Annalisa, Arnaud, Camille, Catherine, Damien, Irene, Katie, Katarina, Margaux, Michaela, Minh-Thành, Nicolas, and Romain, for many amazing memories that kept me going in the tougher times. Of course, I could not state how grateful I am to Mathias, for being both the best flatmate and friend anyone could ask for, and to Maëva, for bearing with me through so many years and interminable phone calls.

None of this would have been possible without the constant support of my family. Merci à Hélène, Véronique, Valérie, Didier, Martial, Estelle, Eulalie, et Lucas pour leur soutien depuis tant d'années. Merci à Jérémy, pour avoir toujours été à mes côtés, et dont l'enthousiasme et la volonté de découverte sont une inspiration permanente. Merci à Mamie pour être une présence positive et un soutien infaillible depuis le début, et pour avoir égayé mes dimanches de confinement. Maman, Papa, je ne pense jamais pouvoir vous remercier assez pour l'aide permanente que vous m'avez donné toutes ces années. Rien de tout cela n'aurait été possible sans vos conseils et votre amour. Merci.

Finally, to Alyce, for making everything so much better when she is around, and for being the most wonderful person to be with, especially during a lockdown.

List of abbreviations

CDC: Centers for Disease Control and Prevention

CFR: Case Fatality Ratio

CI: Confidence Interval

ECDC: European Centre for Disease Prevention and Control

EU: European Union

IgM: Immunoglobulin M

LIC: Low Income Countries

MCMC: Monte Carlo Markov Chains

MCV: Measles Containing Vaccines

MMR: Measles-Mumps-Rubella

MMR-V: Measles-Mumps-Rubella-Varicella

MR: Measles-Rubella

NPI: Non-Pharmaceutical Interventions

RNA: Ribonucleic acid

RT-PCR: Reverse transcription polymerase chain reaction

SIA: Supplementary Immunisation Activities

UMIC: Upper Middle-Income Country

UK: United Kingdom

USA: United States of America

WHO: World Health Organisation

Contents

Figures

Tables

Outline of thesis

Aim and objectives

The aim of this PhD is to develop methods to improve the identification of groups and local areas most vulnerable to measles outbreaks using routinely collected surveillance data, and better understand their impact on the dynamics of measles transmission in countries with low national incidence and high national vaccine uptake.

The specific objectives of the thesis are:

> Objective 1: Develop an inference method to use routinely collected epidemiological data in order to reconstruct who-infected-whom.
>
> Objective 2: Highlight the areas where importations of measles was most likely to cause secondary cases in the United States using the inferred transmission trees.
>
> Objective 3: Estimate whether recent incidence and local vaccine coverage were associated with a lower level of local and cross-regional transmission during the past ten years in France.
>
> Objective 4: Explore the impact of variation in vaccine coverage on the number of cases and the spatial spread of measles.
>
> Objective 5: Adapt currently suitable Epidemic-Endemic transmission models to daily case counts in order to assess the sensitivity of the results to the use of aggregated surveillance data.

Layout

This thesis follows a 'research paper' style, meaning that some of the chapters are publications in peer reviewed academic journals. I have published one first-author paper (Chapter 3), one other is currently under review (Chapter 2). Chapter 4 has been submitted to an academic journal, and Chapter 5 is not currently written as an independent paper. These chapters are preceded by an introduction and followed by a discussion. This thesis therefore contains six chapters in total:

1. **Background and introduction to the research questions**
2. **o2geosocial: Reconstructing who-infected-whom from routinely collected surveillance data**: This paper presents the R package *o2geosocial*, developed during this PhD, that aims to infer the probability of connection between cases from routinely collected surveillance data.
3. **Probabilistic reconstruction of measles transmission clusters from routinely collected surveillance data**: This paper shows an application of *o2geosocial* to routinely collected surveillance data

collected during measles outbreaks in the United States between 2001 and 2016, and assesses whether the inferred transmission clusters match the contact tracing investigations.

4. **The impact of local vaccine coverage and recent incidence on measles transmission in France**: This paper presents an estimation of the association between local vaccine coverage and the recent level of incidence, and the daily case counts per department in France between 2009 and 2018. This analysis is carried out using the Epidemic-Endemic framework, implemented in the R package *surveillance*. It also shows an evaluation of the impact of variation in coverage and incidence on future outbreaks.
5. **Comparison of aggregated and non-aggregated models in the Epidemic-Endemic framework**: The aim of this chapter is to analyse the impact of aggregated data in the Epidemic-Endemic framework on parameter estimations and calibration of the models, using simulated outbreaks. It shows the added value of using non-aggregated data when available.
6. **Discussion and conclusions**.

Chapter 1. Background

1.1. Measles virus and the impact of measles vaccines

1.1.1. Symptoms and case definition of measles

Measles is an infectious viral disease spreading among humans, who are its only known reservoir [1]. It is caused by the measles virus, an Ribonucleic acid (RNA) virus member of the Paramyxoviridae family of viruses, and is transmitted via respiratory droplets. Symptoms commonly associated with measles include fever, cough, coryza and conjunctivitis, followed by a characteristic rash, red eyes, sensitivity to light, and Koplick's spots inside the mouth. Measles complications are most common in young infants, adults older than 20 years, pregnant women, and immunocompromised or malnourished individuals [2]. The most frequent complications include ear infections, diarrhoea, and pneumonia. Despite the availability of a safe and highly effective vaccine, measles remains a leading cause of morbidity and mortality in young children, causing 140,000 deaths in 2018, mostly among children under the age of 5, according to the World Health Organization's (WHO) estimations [3]. WHO's guidelines detail the case definition and classification of measles cases is as follows [4]:

- A suspected case is defined by fever and maculopapular rash, or if a health-care worker suspect measles.
- A laboratory-confirmed case is a suspected case that was confirmed positive by testing in a WHO-accredited laboratory. Common methods for confirmation of measles infection include detection of measles-specific Immunoglobulin M (IgM) antibody (on blood specimen) and measles RNA by real-time Reverse Transcription Polymerase Chain Reaction (RT-PCR) (throat, or nasopharyngeal swabs).
- An epidemiologically linked measles case is a suspected case that was not lab-positive, but was geographically and temporally related to a lab-confirmed case, or another epidemiologically linked case.
- A clinically compatible case is a suspected case that was not lab-confirmed, nor epidemiologically linked, but had fever, a maculopapular rash, and at least one of cough, coryza and conjunctivitis.

This classification is illustrated in Figure 1.1.

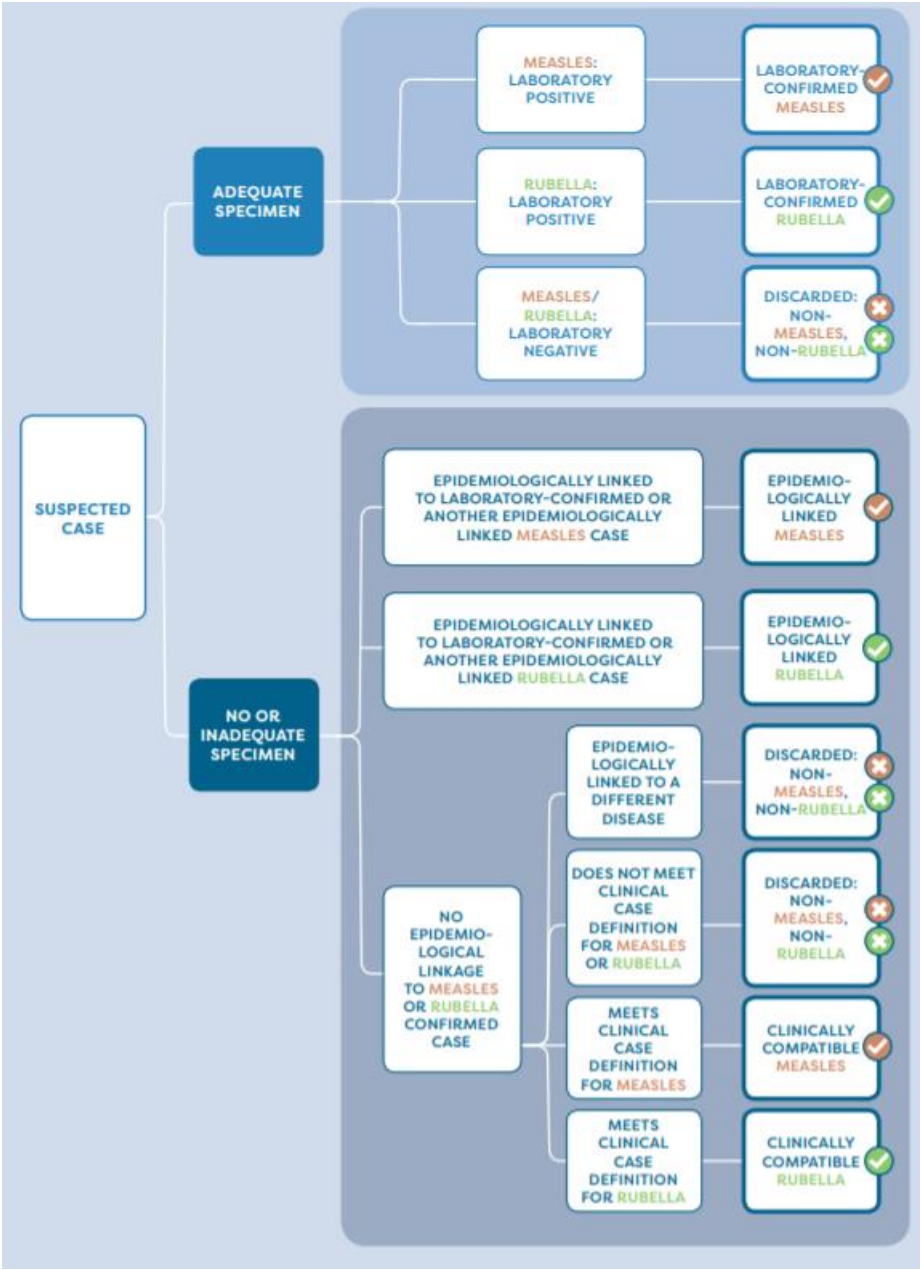

Figure 1.1: Classification of suspected measles and rubella cases from [4].

On the other hand, the case classification used in the United States is defined by the Centers for Disease Control and Prevention (CDC) as follows: a clinically compatible case (no laboratory confirmation, no epidemiologic linkage to a confirmed case) is defined as “probable”; a laboratory-confirmed, or epidemiologically linked case is “confirmed” [5]. In Europe, the European Centre for Disease Prevention and Control (ECDC) defines a case as “possible” if they are clinically compatible, “probable” if they are clinically compatible and epidemiologically linked to a laboratory-confirmed case, and “confirmed” if they are clinically compatible and lab confirmed, and have not been recently vaccinated [6].

Upon exposure, the incubation period ranges from 7 to 18 days, and a rash appears after 14 days [7]. The immunity gained after infection or vaccination lasts for at least several decades [8].

Measles is highly infectious. The estimated mean number of secondary transmissions caused by an infected individual across the course of the disease among an otherwise completely susceptible population ranges from 9 to 18 individuals [1,9]. This value, defined as the basic reproduction number, is one of the highest amongst directly transmitted pathogens [10]. Therefore, measles virus can spread very quickly in populations with low immunity to the virus, and cause a large number of severe infections, which can ultimately lead to deaths. The Case Fatality Ratio (CFR) is context-specific, it was estimated to be around 1.6% in Low Income Countries (LIC) and 0.7% in Upper Middle-Income Country (UMIC) [11]. Before the introduction of measles containing vaccines (MCV), measles was responsible for more than 2 million deaths annually, mostly among children [12,13].

1.1.2. The development of measles-containing vaccines and immunisation programs

Prior to the development of vaccines, measles mortality declined steadily in most industrialized countries in the first half of the 20th Century, in part thanks to the use of antibiotics and improvements in living conditions [12,14,15]. Nevertheless, almost all children were still contracting measles before adolescence [16]. Measles became a vaccine-preventable disease in 1963, with the first introduction of a measles vaccine in the population [17]. Measles vaccines are commonly administered as combined vaccines with those for rubella (MR), mumps (MMR), or varicella (MMR-V). The MMR vaccine efficacy is high: two doses is roughly 97% effective at preventing measles; one dose is about 93% effective [18]. Optimal protection is therefore reached after 2 doses of vaccine.

Determining the optimal age of vaccination in routine immunisation programs is a matter of balancing various context-specific factors: In countries with active transmission, WHO recommends that the first dose be administered at nine months after birth in routine immunisation programs, and the second dose should be given at age 15-18 months, with at least four week between the two doses [19]. In these settings, delaying the administration of vaccine doses would lead to children risking infection prior to their vaccination. These vaccine doses would then fail to prevent infection, and potential severe symptoms, and would be given to children who had already acquired immunity through infection. On the other hand, in countries with lower levels of transmission, WHO recommends that the first dose should be administered at 12 months of age. The second dose should then be administered from two years of age. This is justified because the average age of infection will increase if the level of transmission is lower, therefore the risks of infection before vaccination should be reduced [19]. Furthermore, later vaccination can benefit from higher seroconversion rates in the vaccinated children, which means that measles antibodies in the patient are higher than the protective threshold, and that the individuals should therefore be protected against infection. Indeed, seroconversion rates were shown to increase between 4 and 11 months of age [20,21]. Earlier age at first dose may also lead to an increase in the number of vaccine failures after two doses [22].

Finally, another factor that must be accounted for in calculations of the routine vaccination schedule is reducing the susceptibility gap between the decay of maternal antibody and the administration of the first dose. Indeed, most children are born with immunity to measles thanks to maternal antibodies, due to previous infection or vaccination of the mother. The maternal antibodies decay over time, and can interfere with the ability of the vaccine to induce an immune response at younger ages. The amount of antibody transferred and the time of protection depend on various factors: vaccinated mothers transfer fewer antibodies to their infant than those who acquired immunity through infection [23,24], and in vaccinated mothers, the duration of protection may be inversely correlated with childbearing age [25]. In countries that have maintained high levels of vaccine coverage and low levels of transmission for decades, studies suggest that a proportion of infants could be susceptible at birth [26–28]. The recommendations for the age of vaccination are therefore informed by several factors and may need to be periodically re-evaluated.

On top of routine vaccinations, supplementary immunisation activities (SIA), or mass vaccination campaigns, can be implemented to strengthen the level of protection in the population. An SIA is defined as the administration of a supplementary dose of vaccine to a group, that can be spatially or socially defined, during a short period, regardless of the recipients' previous vaccination histories. Countries that reach low levels of coverage with routine immunisation programs can use periodic SIAs to close the gaps in target coverage [29,30]. In other countries, reactive mass vaccination campaigns can be part of the public health response to increases in cases, especially when certain age groups or populations are known to be under-vaccinated [31].

1.1.3. The impact of vaccination campaigns on measles transmission

The development of national routine vaccination programs and SIAs around the world increased the coverage of MCV. This led to a substantial decrease in the burden of measles in every WHO region (Africa, Americas, South East Asia, Europe, Eastern Mediterranean, and Western Pacific), both in terms of incidence and mortality [14] (Figure 1.2). Indeed, a historic low incidence was reached in 2016, with 18 cases per million population, down from 145 per million in 2000 [32], and 13.8 million estimated deaths were prevented by measles vaccination between 2000–2012 [33]. In September 2013, all six WHO regions accepted the target defined by the Global Vaccine Action Plan for 2012-2020 to eliminate measles in every region [34]. Elimination of measles in a region is defined by three verification criteria [35]:

- Documented interruption of endemic measles virus transmission for a period of at least 36 months from the last known endemic case.
- The presence of high quality (“verification standard”) surveillance.

- Genotyping evidence that supports interrupted transmission (i.e., no indigenous chains of transmission persisting for at least one year). Indigenous chains of transmission refer to local endemic transmission that does not correspond to an importation.

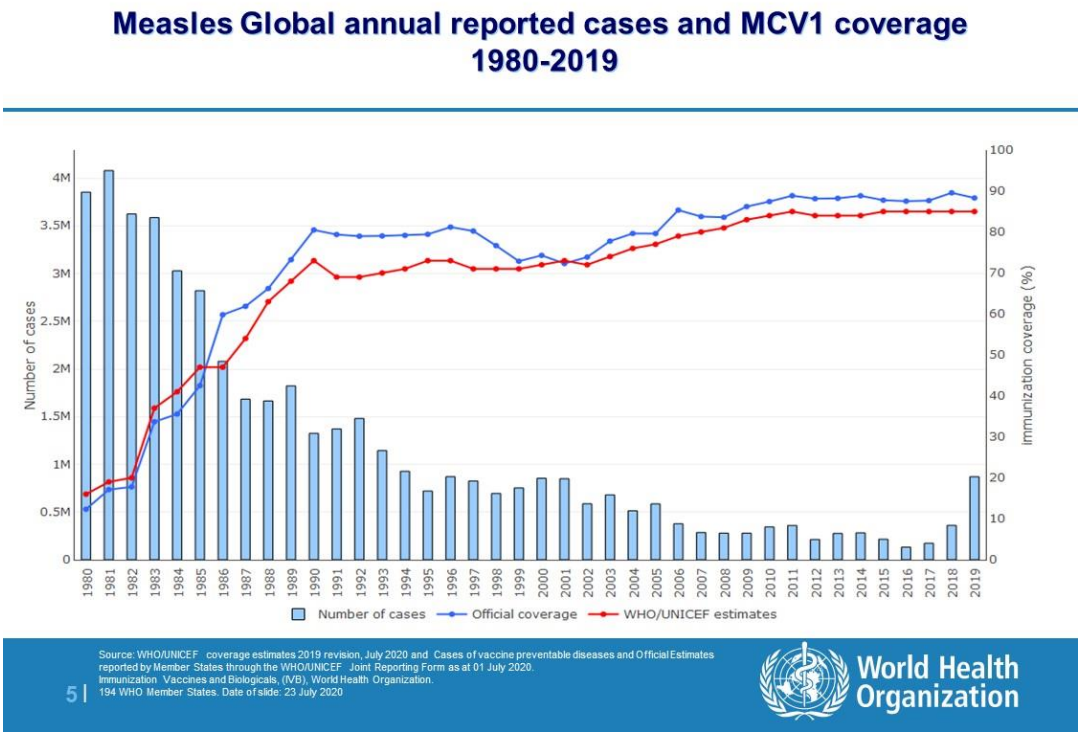

Figure 1.2: The number of measles cases and vaccine coverage reported worldwide to WHO between 1980 and 2019, taken from [36]

Thus far, the only WHO region to ever declare measles eliminated was the Americas in September 2016 [37], following the declaration of elimination in the United States in 2000 [38]. The success of the Pan American Health Organization in interrupting measles transmission relied upon combined strategies applied uniformly across the Americas, including high vaccine coverage reached through routine immunisation programs, catch-up campaigns during periods of low transmission, and follow-up campaigns to ensure high levels of immunity at the age of school entry [39].

Moreover, although the regional number of cases was stable, some European countries regularly reported outbreaks between 2012 and 2017, while 43 countries (91%) had interrupted endemic measles transmission for more than 12 months, including 37 (70%) that had sustained interruption for more than 36 months [40].

However, in 2017, the global incidence started increasing again, and major outbreaks were reported in countries that had not reported endemic transmission in years. The global incidence in 2018 was at its highest since 2011, and the number of cases in 2018 corresponded to an 167% increase compared to 2016. At a regional level, this corresponded to increases of 246% in the African Region, 16,732% in the

Americas, 931% in the Eastern Mediterranean Region, 1,791% in Europe, and 26% in South East Asia [32].

This recrudescence in incidence was especially acute in regions that had reached the elimination status or where deemed "near elimination". In Europe, more than 80,000 cases were reported in 2018, the highest figure in 20 years [39,40]. Numerous countries reported endemic transmission that started in 2017 and lasted throughout 2018 (e.g. Romania [41], Greece [42], France[43], Ukraine [44]). Large measles outbreaks were also reported in South America in 2018 [45]. The United States reported 375 cases in 2018, and 1,282 in 2019, its highest figure since 1992. As a consequence of the recent increase of cases, the elimination status was revoked in several countries only a few years after they had reached it, for example in Albania, Czechia, Greece, and the United Kingdom in Europe, or Venezuela and Brazil in the Americas [34,46,47]. Given that the Americas and Europe were thought to be in best position to eliminate measles, it is key for public health organisations to understand the factors that led to the resurgence of measles transmission in near elimination settings.

1.2. Measles burden in high-income high-coverage settings: the need for indicators

1.2.1. The effect of immunity gaps on transmission risks in countries near elimination

As measles is highly infectious, high levels of vaccine coverage are needed to control the spread of the virus in a population. In theory, routine vaccination programs would give immunity to every child by administering two doses of vaccine. In practice, because the efficacy of the vaccine is not perfect, and factors such as age at vaccination can impact the seroconversion rate, not every administration of a vaccine confers immunity. Furthermore, protection from a vaccine can wane over time [48], and not every individual can be vaccinated, for instance because of severe allergies or a weakened immune system. Failure to administer vaccines to eligible children due to vaccine hesitancy or other factors further increases the fraction of the population that is susceptible to measles [49].

Because of the high transmissibility of the measles virus, high levels of population immunity are required to interrupt transmission. Assuming random mixing of individuals and lifelong immunity upon infection, one can use the basic reproduction number R_0 to compute the minimum vaccine uptake needed to interrupt endemic transmission and reach a "herd immunity threshold", which refers to a population where the vaccine coverage is sufficient to reduce the risks of transmission among susceptible individuals thanks to the presence of immune individuals [50]. First estimations of this threshold have showed that $1 - \frac{1}{R_0} = 89 - 94\%$ of the population is required to be immune to potentially achieve measles elimination [51]. However, this theoretical estimate ignores non-homogeneous population mixing, imperfect vaccine efficacy and spatial heterogeneity in vaccine coverage [52,53].

Since lifelong immunity is gained after infection, the level of immunity in the population also depends on previous levels of transmission. Therefore, the proportion of children that received two doses of vaccine shows an incomplete picture of the risks of future outbreaks. Given that in the absence of immunisation activities, almost all children would be infected by measles [16], the level of immunity in adults that were born before the implementation of large-scale immunisation programs is high. Upon implementation of the first immunisation activities (both routine vaccination programs and SIAs), the number of susceptibles in the population will be minimal due to all adults and teenagers having gained immunity through infection, and a proportion of the children being vaccinated. The transmission expected after the implementation of vaccine campaigns is therefore very low. This period has been coined as the "honeymoon period", and was repeatedly observed both in real data and epidemiological models [54–56]. Individuals that were missed by vaccination campaigns, and had not been infected given the drop in measles transmission during this period therefore remain susceptible and accumulate over time. After several years, the number of susceptibles in the population can become sufficient to trigger new outbreaks in an older population than during the endemic phase, thus ending the so-called "honeymoon period".

Minimising the number of individuals missed by routine immunisation is therefore central for bringing measles transmission under control. The target immunity levels per age group was first computed by the WHO European Region in the 1990s using age-stratified contact data to compute age-specific transmission rates [57]. They recommended that at least 85% of 1–4-year-olds, 90% of 5–9-year-olds and 95% of 10-year-olds and older possess immunity against measles to achieve elimination. Gaps in immunity that could result from incomplete routine programs should be closed by SIAs. Nevertheless, these levels were based on social contact patterns from pre-vaccination measles epidemiology in England and Wales. Subsequent studies, using recently observed age-stratified contact patterns, have shown that the target immunity level of 5-9 year olds may not be sufficient to remove endemic transmission, and would need to be raised to 95% [58].

Gaps in immunity can have long-lasting consequences, with cohorts maintaining a high proportion of susceptibles decades after a drop in routine vaccine uptake. A clear example was observed in the United Kingdom: In 1998, an article published in Lancet falsely linked the MMR vaccine to autism in children and gained media attention, causing a drop in the MMR vaccine coverage in the United Kingdom [59–61]. The study was later debunked [62–64], but the vaccination uptake of the first dose at 2 years old in England and Wales, exceeding 90% before the controversy, dropped to below 80% in 2003. Vaccination rates have been increasing ever since, reaching 89.3% for the second dose at 5 years old in 2015 [65]. Despite this sustained annual increase and several catch-up campaigns aiming to reduce this susceptibility gap [66–68], individuals born in the late 1990s – early 2000s may remain the most

susceptible to measles virus in the country. Indeed, this cohort was linked to several large transmission events since 2010, where the overwhelming majority of infected teenagers had never been vaccinated [69,70].

Immunisation strategies must be designed to reach high vaccine coverage homogenously in the population, and must be maintained consistently to avoid the emergence of gaps in immunisation. In Europe, the age distribution of the cases in recent outbreaks highlighted the shift in immunisation profile: adults older than 20 years old accounted for 30% of the cases in Europe in 2015 [71], and 37% in 2018 [40], whereas in 2006 and 2007, this proportion was below 20% [72]. The proportion of adolescents and adults infected even reached 50% in some outbreaks [73]. These cases are too old to be eligible for national routine immunisation programmes and can only be targeted by SIAs. Furthermore, the shift in the age distribution of cases also triggers changes in the routes and settings associated with measles transmission: recent outbreaks were reported during art and music festivals and other mass gatherings [70,74,75]. The large outbreaks triggered after the accumulation of susceptible individuals can subsequently affect infants in the period between the decay of maternal antibodies and the immunisation programmes [71,73].

Gaps in immunity can also arise from spatially or socially heterogeneous vaccine coverage across the country, and can create pockets of susceptibles although the overall uptake is high. In these social or spatial subgroups with lower levels of immunity, susceptible individuals are more likely to be in contact with each other than would be expected if the vaccine coverage was homogeneous [76]. These disparities in coverage can then lead to localised outbreaks [77]. For instance, in Europe, various outbreaks were reported within populations with lower probability of vaccination such as: Roma communities in Bulgaria, Anthroposophist communities in Germany, the Netherlands and Switzerland, Jewish ultra-orthodox communities in Belgium, United Kingdom and orthodox Protestants in the Netherlands [78,79]. Spatial variations in coverage have also been identified as the main driver of measles outbreaks in Germany [80]. Similarly, in the United States, recent outbreaks affected under-immunised Amish and Jewish communities [77,81].

Various causes have been identified to explain the struggles in many countries near elimination to reach a uniformly high vaccination coverage. Firstly, the role of socio-economic factors in causing variation and inequalities in vaccine coverage has been repeatedly highlighted. Toffoluti et al linked the adoption of austerity measures and local decreases in public health expenditures in Italy to drops in regional vaccination coverage [82], and Danis et al highlighted that in Greece, socio-economic factors explained under immunization better than parental beliefs and attitudes towards immunization [83]. Muscat also showed that lack of information and poor access to health care were major factors for explaining low

vaccination coverages in certain community [84]. These factors can lead to entire sub-populations lacking vaccination and being at highest risks of large outbreaks. Muscat identifies Roma and Sinti, Traveller, Anthroposophist, Ultra-orthodox Jewish communities, and recent immigrants as populations particularly at risk of measles outbreaks.

Secondly, vaccine scepticism and refusal of parents to immunise their children can lead to increasingly large pockets of susceptibles. Certain religious beliefs have been known to influence the decision to vaccinate children for a long time, and under-immunised religious communities have been at risk of outbreaks in many low-incidence settings [85]. Furthermore, the rising influence of the antivaccine movement is an area of increasing concern to public health bodies, and can lead to decreases in vaccination coverage in many near elimination settings [86,87]. Vaccine hesitancy can be caused by vaccine-related controversy, such as the drop in coverage that was observed in England and Wales following the media attention given to the debunked study linking MMR vaccination and autism. Following this type of controversy, SIAs targeting the age groups affected by the decrease in vaccine coverage are needed to close the immunity gaps. Vaccine confidence must also be restored through direct discussions to address parental concerns about vaccines [88], or communication to fight the antivaccine misinformation. Indeed, Hotez et al identified three major elements boosting the influence of the antivaccine movement in the United States: a media empire comprised of hundreds of disinformation websites and active social media accounts, a political arm with political action committees in various Western states, and a predatory behaviour to target insular groups [87]. Strategies must be designed to counter these elements, and limit the spread of mistrust in vaccine.

Finally, the example of Venezuela shows the impact of an unprecedented humanitarian and political crisis on the spread of vaccine preventable diseases: measles was re-established as endemic in the country, and led to importations in neighbouring countries, threatening the elimination status reached by the Americas in 2016 [37].

Spatial and social immunity gaps caused by the unequal vaccination coverage in near elimination settings have had a clear impact on measles transmission in the WHO European region in the last ten years, and have been obvious during recent outbreaks in the United States. It is therefore key to understand how epidemiological and vaccination data can be used to solve the challenges raised by the current state of immunisation in countries near elimination.

1.2.2. Identification of sub-national regions and groups with lower immunity

The resurgence of measles observed in various countries that had previously reached elimination, or were approaching it, highlights the existence of previously unidentified immunity gaps. Locating and identifying gaps in immunity is therefore key for two reasons:

- Firstly, to strengthen the immunisation programs by understanding how these immunity gaps came to exist and make sure these groups can be reached by routine vaccination in near future, and design catch-up campaigns to close the immunity gaps,
- Secondly, to locate the areas and groups that are currently most at-risk of outbreak upon importations of measles. Knowing the profile of the groups with low immunisation also brings information into the settings where large outbreaks are more likely, for instance schools, universities, or large gatherings.

Nevertheless, identifying the pockets of susceptibles distributed in a country is challenging. The most direct way to figure out the immunity profile of the population is through serosurveys, whereby specimens from a defined population are collected and tested to determine the presence of antibodies as a measure of immunity in the population [56,89]. Therefore, serosurveys provide a measure of immunity that contains both vaccine-induced and infection-induced immunity. They bring a punctual set of evidence of how close a given population is from herd immunity, and therefore are key to assessing the risk of outbreaks and identifying high-risk population subgroups. Finally, routinely conducted serosurveys (referred to as serosurveillance) also helps monitor immunity over time, thus highlighting groups that could be targeted by SIAs. Nevertheless, serosurveys come with various limitations: they can be costly, and logistically difficult to conduct, they also require a substantial time commitment. Consequently, they are rarely carried out in countries where the burden of measles is relatively low. Therefore, little information on the state of measles susceptibility in near elimination settings can be drawn from recent serosurveys.

The identification of pockets of susceptibles can also be informed by using vaccine coverage data. The nationwide coverage information is collected by most of countries with well-performing surveillance data (which includes near elimination countries). Information on the national uptake of first and second vaccine doses among given age groups can help identify drops in vaccination following surges in levels of vaccine hesitancy or declining access to the routine immunisation programs. The birth years associated with temporal drops in immunisation can be targeted by SIAs once they are too old to be captured by routine immunisation programs. Coverage data provides more accurate descriptions of the national immunity profile in countries where the level of measles transmission has been low for years, since most of the immunity is achieved by coverage rather than previous infection. Nevertheless, distinguishing spatial sub-national heterogeneity in immunity using coverage data is much more challenging. Indeed, historical data on vaccine coverage at a local level are rarely available. Local estimations of vaccine coverage also require a reliable population denominator at the same spatial scale, which is often not available. Furthermore, they can be unreliable because of movements within the country since the population may not correspond to the local vaccine uptake twenty years before.

Finally, vaccine coverage does not account for the imperfect vaccine efficacy, which means that vaccine coverage could be an overestimation of the level of vaccine-induced immunity in the population, especially in countries that deliver the first dose of vaccine early.

Routinely collected disease surveillance data can be used to identify areas with higher susceptibility. As argued in the previous subsection, the age distribution of cases can shed light on the intensity of transmission in the country, and on the existence of immunity gaps. Epidemiological investigations can also highlight routes of transmission and super-spreading events, hence improving the understanding of the transmission dynamics and improving future surveillance. Local effective reproduction numbers, which describe the average number of individuals infected by a case in an area, can also be computed from surveillance data and show regions repeatedly associated with increased transmission. However, this type of analysis is subject to various limitations. Firstly, regions with low incidence can report several consecutive years with very low numbers of cases, during which time the number of susceptibles increases (this phenomenon is referred to as 'replenishment of susceptibles'). During these periods, surveillance data will display little information about the risk of outbreaks. The amount of information brought by surveillance data therefore depends on the amount of transmission in the country. Secondly, measles outbreaks can be widely underreported, especially among communities that are less in contact with public health authorities [90]. This can lead to entire transmission chains not being reported, and ultimately to an underestimation of the number of secondary infections per case. Finally, extensive case investigations are needed to assess the importation status and the clustering of cases, in order to identify the different concurrent transmission chains and independent importations. Indeed, the raw local number of cases is not sufficient to evaluate the susceptibility of a region. For instance, a region that repeatedly reports independent importations without secondary transmissions could be deemed protected against large outbreaks, whereas another region with a similar number of cases all grouped in the same transmission cluster would be especially at risk of large transmission events.

Routinely collected surveillance data can be enhanced through information on genetic sequences of measles. Indeed, molecular surveillance of measles is defined as an essential aspect of elimination programs, but is hampered by measles virus' genetic stability [91,92]. Molecular surveillance aims to identify whether sequenced cases come from repeated introductions (i.e. their sequences are too different to be linked in the same transmission cluster), or if they were caused by endemic transmission (i.e. their sequences are similar enough to be connected). However, measles virus shows very limited sequence variability both in outbreaks settings and in laboratory [93,94]. Furthermore, although WHO recognises 24 different measles virus genotypes, only four of these have recently been detected, with the B3 and D8 genotypes accounting for an overwhelming majority of the sequences reported in recent years [95]. Comparing whole-genome sequences would exhibit more variability between the sequences,

but collecting whole genomes is expensive and labour-intensive [96]. Therefore, molecular surveillance of measles virus is currently mostly based on 450 nucleotides of the C-terminal region of the nucleoprotein (N450), but the genetic variation of this sequence of nucleotides is limited in the B3 and D8 genotypes [95] . Novel approaches are now being designed to integrate more regions of the measles genetic sequences in molecular surveillance, without increasing the technical and financial cost of sequence collection [97,98].

In conclusion, the identification of the areas and groups most susceptible to measles transmission can be informed by multiple data sources, each with its own strengths and limitations. Developing tools in order to be able to combine these data sources to maximise the information they provide is therefore crucial to assess the risks of measles outbreaks in near elimination settings.

1.3. Mathematical modelling to identify and study heterogeneous transmission risks

1.3.1. Introduction to mathematical modelling of measles outbreaks

Thanks to improvements in the electronic surveillance of infectious diseases, combined with the development of modern methods able to study the transmission dynamics of infectious diseases, mathematical analysis and models are now playing a central part in public health strategies. They are used to propose and test theories, as well as compare, plan, implement and evaluate detection, prevention, and control programs [1,99,100].

Mathematical models are limited by a balance between accuracy (ability to reflect data with correct predictions), transparency (clarity of the role and impact of the model and its components), and flexibility (ability to be adapted to new situations) [101]. Models can improve our understanding of infectious disease dynamics by estimating parameters quantifying the intensity of transmission.

Model parameters can be estimated from data using a wide range of techniques and analyses. Bayesian statistics allows for the inclusion of external information from previous studies using the prior distributions. It relies on Bayes' theorem, which is used to compute the posterior probability by multiplying the prior probability distribution and the likelihood function. Monte Carlo Markov Chains (MCMC), where proposal distributions are used to update the set of parameters depending on its previous values, is an efficient and commonly used method to sample from a posterior distribution when an analytical form of that distribution is unavailable [102,103]. One of the most popular methods is the Metropolis-Hastings algorithm, that simulates MCMC where an update is proposed and is then accepted or rejected depending on its likelihood. The efficient convergence of the MCMC chain depends on the movement functions implemented in the Metropolis-Hastings algorithm.

Characterised by its high infectiousness over a relatively short period and long-lasting immunity, measles has served as a model of an acute, immunising infectious disease for studies of infectious disease dynamics [54,104,105]. Numerous studies have focused on the pre-vaccination era and were used to design vaccination catch-up campaigns, or to evaluate or modify vaccination routines, notably in England and Wales [13,106–109]. Historically, since measles almost exclusively affected children, age-structured compartmental models have been widely used for measles modelling, to understand regional transmission, or to estimate the transmission rates before the implementation of vaccination campaigns [110–112].

Mathematical models have also been used to study the spread of measles in highly vaccinated areas, and give insight into the transmission dynamics in settings near elimination [52,113,114]. These studies aimed to improve the detection of potential losses in herd immunity, and the impact of future variations in the vaccine coverage. Models also showed the potential for local spread when national vaccine coverage is above the elimination threshold, highlighting the impact of super-spreading events [115,116]. “Super-spreading events” refer to transmission events where an unusually high number of secondary transmissions are linked to a single case or settings, for instance because of a high number of potentially infectious contacts, and their impacts have been documented for various infectious diseases [117–119].

This PhD focuses on two specific aspects of the mathematical modelling of measles outbreaks: firstly, the use of Bayesian statistics and MCMC to reconstruct transmission history and infer who-infected-whom; and secondly the so-called Epidemic-Endemic framework, used to disentangle the association between various covariates and the incidence at a sub-national level. Both frameworks use a variety of data sources, one of the most informative being the regions of residency (or notification) of the cases. Therefore, models are needed in order to estimate the likelihood of connection between two cases depending on their location.

1.3.2. Mathematical models to estimate the connectivity between regions

The connectivity between regions in a given area of interest quantifies the movements of populations between the different regions. Empirical measures of movements can be generated from commuting data [120], mobile phones’ GPS, or social network data [121]. These data cannot directly be applied to infectious disease modelling, specifically measles virus, since the spatial spread of the virus often focuses on specific age groups with different mobility, for instance school-aged children, who do not commute, are less likely to have a phone, and are often excluded from movement data because of sensitivity. Furthermore, the behaviour of infected individuals may differ from typical movement behaviour, i.e., symptomatic cases may not commute to work, or control measures may change the

mobility. Mobility models can therefore be integrated in the modelling of the virus spread to account for the specificities of disease-related mobility.

The set of models most commonly used to describe connectivity between regions are the gravity models, where the connectivity only depends on the number of inhabitants and the distance between the regions [122]. In the simplest form of the gravity approach, the probability of commuting between two regions i and j is proportional to the product of the origin population size m_i, the destination population size m_j, and a function of the distance between i and j: $f(d_{ij})$: $p_{ij} \propto m_i^{\alpha} m_j^{\beta} f(d_{ij}, \gamma)$, with α, β, and γ adjusting for the impact of each variable. The gravity approach has been widely used to describe the spread of infectious diseases [110,123,124]. Various formulations have been developed, for instance the function of distance between the two regions has commonly been described as a power-law decay ($f(d_{ij}, \gamma) = d_{ij}^{-\gamma}$), but can also be set as an exponential decay ($f(d_{ij}, \gamma) = e^{-\gamma d_{ij}}$) [120]. The spread of infectious diseases can also be represented as piecewise gravity models, where different functions are used to compute the connectivity between regions depending on the absolute distance between them [123]. Such framing supposes that the connectivity between nearby regions cannot be described by the same function and parameters as those used to describe the connectivity between regions that are further away from each other.

Although the gravity models' appeal relies on their simplicity and their flexibility, one major shortcoming of this framework is that the connectivity between two regions is influenced by the size of neighbouring regions. In other words, human movement between two urban regions i and j with only rural areas between the two regions may be different from the connectivity if another urban region is located between i and j. Stouffer's "law of intervening opportunity" posits that the connectivity between two regions directly depends on the number of "intervening opportunities" between the two regions [125]. In other words, the "distance" between i and j would be better quantified as the number of inhabitants living between i and j, than by the absolute geographic distance. This corresponds to the Stouffer's rank model [125,126]. Several alternative models describing human movements are related to Stouffer's law, such as the radiation model [127], that also does not take absolute distance into account, but proposes an alternative reduction of connectivity due to competing regions; or Fotheringham's competing destinations model [128], that integrates both absolute distance and competing populations, and allows for the fact that competing regions may also boost the connectivity between two regions (Figure 1.3).

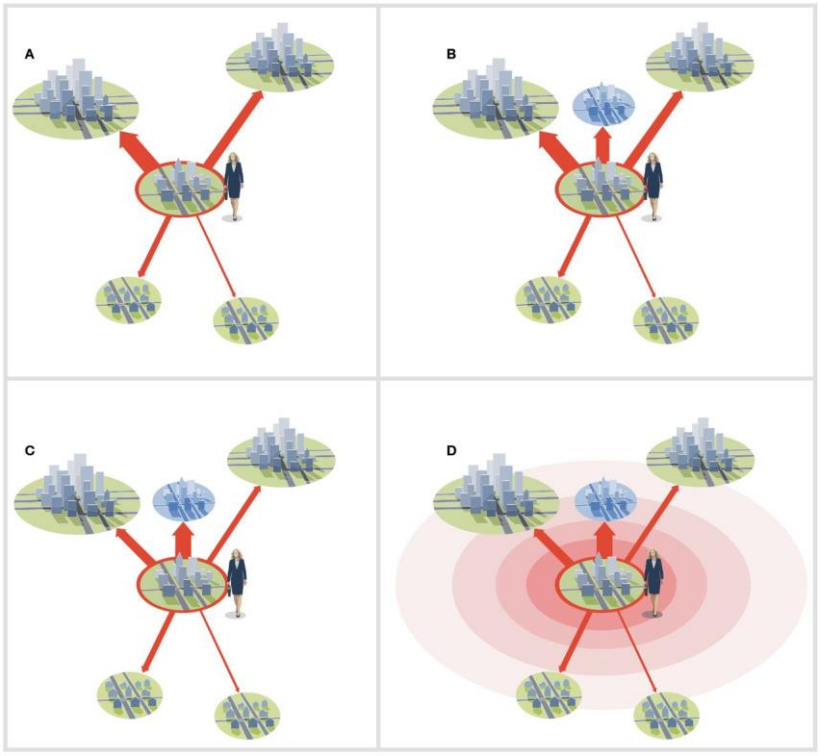

Figure 1.3: Taken from Bjørnstad et al [129]: Spatial interaction models predict the flux of human movements between population centres (cities, towns, villages) as a function of the distribution of the population. In this diagram, the relative magnitude of the fluxes from a focal town to other population centres are represented by the widths of the arrows. In the widely-employed gravity models (A), interactions among cities are strictly pairwise. Thus, the addition of a new town (B) has no effect on the movement to other towns. In Fotheringham's competing destinations model (C), however, competition or synergy among nearby communities can reduce or augment fluxes. Stouffer's model of intervening opportunities and the radiation model (D) posit that movement from one city to another is diminished by the presence of opportunities in communities more proximal to the source city.

Various publications aimed to compare these models and evaluate which one was best adapted to the reconstruction of movements between regions. Results were dependent on the data sources and the distance of connection. Bjørnstad et al. used weekly pre-vaccination (1944-1965) measles cases counts in England and Wales and highlighted that Stouffer's rank models outperformed the gravity model, although this dataset may not represent modern movements between regions [129]. On the other hand, Lenormand et al. showed that the exponential gravity model was best able to fit commuting flows in different European and American countries. This result was more marked for shorter distance trips [120]. Therefore, different mathematical models can be used to analyse the connectivity between sub-national regions, and applied to the spatial spread of measles.

1.3.3. Mathematical models to reconstruct who infected whom

A key parameter to understand and analyse the spread of a virus in a community is the effective reproduction number, which describes the average number of individuals infected by a case in a community. When the reproduction number is above 1, the number of cases increases over time, possibly leading to a large epidemic. To eventually stop the spread of a virus in a population, the

reproduction number needs to be brought below 1. The individual reproduction number can be computed by studying the history of individual transmissions, i.e. “who-infected-whom”. Indeed, from the transmission trees, one can compute the number of secondary transmissions per reported case, and study the factors associated with major transmission events. However, transmission trees have a number of limitations: since they are dependent on the proportion of cases that get reported to the surveillance system, they may under-estimate the number of secondary transmissions caused by a given case. They also do not account for opportunities of transmission that did not happen. Therefore, there is no way to distinguish between individuals who did not cause further transmissions because they had no contact (i.e. they do not give any information on the level of immunity in the population), and those who had many contacts, but did not cause further transmission because all their contacts were immune (i.e. this hints that the immunity level in the population is high). These points are especially important for pathogens where subclinical cases are frequent. Nevertheless, transmission trees can be used to highlight areas, or individual variables, repeatedly associated with secondary transmissions within the reported cases.

Transmission trees can be reconstructed through patient interviews, whereby cases are asked about their movements and contacts during their infectious period. This involves the identification of people who may have come into contact with cases, based on epidemiological data directly collected from the interviewed individuals (e.g. age, gender, when and where contact with a confirmed case occurred, onset date). This approach has been applied to recent epidemic situations, such as the 2014-16 Ebola virus (EBOV) outbreak in Western Africa [130] or the SARS coronavirus 2 epidemic (SARS-CoV-2) [131]. However, contact tracing has several limitations: it is sensitive to reporting, ascertainment or recall biases. It is also difficult to implement when only a subset of the cases is reported to the surveillance system, because of asymptomatic transmission or imperfect surveillance. Furthermore, contract tracing is expensive and time-consuming. Therefore, it is not always carried out as a standard measles surveillance procedure.

Statistical methods have been developed to infer transmission trees using epidemiological or genetic data [132–134]: Wallinga and Teunis first developed a likelihood-based estimation procedure taking the onset date of cases as inputs to reconstruct likely transmission trees [133]. In the Wallinga-Teunis algorithm, the distribution of the generation time, which describes the typical number of days between an infector and an infectee, is used to compute the likelihood of transmission between an infector and an infectee. The individual reproduction number is then calculated from the probability of infection between each case in the dataset.

This method has been the bedrock of most transmission tree inference methods because it requires a limited amount of routinely collected information in many epidemic settings. It was modified to integrate the estimation of the date of infection, which provides more information on the time when a case was infectious using the latent period of the disease. It can also be used to estimate the probability of a case being an importation, meaning that the case has no plausible ancestor reported in the dataset. Other methods have also integrated the probability of unreported cases between connected individuals to account for a situation where case A infected case B, who infected case C, but only A and C were reported. Case A and C are therefore linked, with one missing generation, and the probability of missing generations between cases can be added as a parameter estimated by the model. The Wallinga-Teunis algorithm has been adapted as a Bayesian framework, which allows for the inclusion of priors of the parameters of the model [135]. The Bayesian inference methods commonly use the Metropolis–Hastings algorithm with MCMC to sample from the posterior distribution of the parameters and transmission trees.

Nevertheless, in the event of concurrent transmission chains reported in a given community, the onset date may not be enough to disentangle the transmission history, and the Wallinga-Teunis method has been supplemented with other data sources to improve the accuracy of the inference procedure. For instance, genetic sequences are now of special interest when identifying chains of transmission, as patients can be easily sampled, and genetic distance between samples can be computed. When the genetic distance between two samples is low, the cases are more likely to belong to the same part of a given transmission tree. Using sequence data alone to reconstruct the history of transmission during outbreaks is increasingly frequent as sampling of cases is becoming more common [136–139]. Methods such as the R package *outbreaker2*, have been developed to unify genetic distances and epidemiological data in the same likelihood [140–143]. The inclusion of such data can substantially increase the accuracy of the reconstructed transmission trees [140]. However, as described in the previous section, measles virus is known to evolve slowly [92]. Direct transmission links cannot be inferred using genetic sequences as samples from unrelated individuals can be very close genetically [96,144]. On the other hand, because there is little variability in measles sequences, genotypes can be used to identify importations, since different genotypes cannot be found in the same transmission cluster.

The age of the cases, combined with social contact data, can also bring information on the chances of a connection between two reported cases. Indeed, mixing between different age groups is not homogenous, and incorporating the risks of infectious contacts between age groups in methods to reconstruct transmission trees can improve their accuracy. For instance, Edmunds et al. highlighted the uneven distribution of infectious contacts in measles and meningococcal outbreaks [145]. Data on social mixing patterns are usually collected using prospective surveys where participants complete regular

diary entries listing the number of people they have been in contact with. Mixing matrices can be computed using demographic information on the age of the participants and their contacts. Assuming that infectious contacts are analogous to social contacts, especially for respiratory viruses, one can use these mixing matrices to describe the spread of viruses among and between different age groups [146]. The Polymod contact survey carried out in 2006 in eight European countries further demonstrated the higher contact rate within age groups compared to between age groups [147]. The survey highlighted that the number of contacts was highest among schoolchildren, whereas adults over the age of 65 had the lowest number. The data collected during the Polymod survey has since been used as a reference for age mixing in mathematical models.

The onset date, age-group and location of cases are part of the variables routinely collected during measles outbreaks in most countries near elimination. Therefore, it is crucial to develop methods that combine these variables to maximise the information that routine surveillance can bring to the reconstruction of measles transmission trees, and identify the settings where epidemiological investigations or further variables are needed to disentangle the history of transmission. The transmission clusters and importation status inferred from these reconstruction methods can then yield insight into the regions where importations of cases led to large transmission clusters, or the features associated with the cases involved in large transmission events.

1.3.4. Mathematical models to analyse heterogeneous transmission risks

One of the challenges caused by the heterogeneity in the level of immunity within countries near elimination is to identify the local surveillance or coverage data that can be indicative of higher risks of outbreak. Indeed, as explained in the second section, local values of coverage can be scarce, or may not accurately describe the current level of vaccination in a given area. Moreover, WHO uses the level of incidence in the past three years to assess the eligibility of a country for the elimination status [35], but periods of low incidence can be associated with the replenishment of susceptible individuals, which would increase the risks of imminent outbreaks in a population. Therefore, there is a need to design models able to evaluate the association between local covariates and levels of incidence.

The Epidemic-Endemic modelling framework corresponds to a class of models developed to analyse sub-national case count data and disentangle the separate impact of various covariates on the number of local and cross-regional transmissions, and importations. It was first developed by Held et al., and has since been adapted to various settings and pathogens [80,148–150]. This framework is implemented as part of the R package *surveillance*¸ and is often referred to as the *hhh4* class of models.

In the Epidemic-Endemic framework, the expected number of cases ($\mu_{i,t}$) reported in the region i at time t depends on three sources of transmission (called “components”):

i. The *autoregressive* component ($\lambda_{i,t}$) represents the impact of $Y_{i,t-1}$, the number of cases in i at the previous time step, on the number of cases in i at t. The predictor $\lambda_{i,t}$ indicates the average number of new cases expected at in i at t per case in i at $t-1$.

ii. The *neighbourhood* component ($\phi_{i,t}$) represents the impact of $Y_{j,t-1}$, the number of cases reported for each region j around i at the previous time step, on the number of cases in i at t. The exact impact of cases in these regions on cases in i is determined by a distance matrix ω which quantifies the connectivity between the different regions. If $\phi_{i,t}$ is high, cases in regions around i are more likely to cause new cases in i, whereas a low value of $\phi_{i,t}$ indicates that cross regional transmissions towards i are less likely.

iii. The *endemic* component ($\nu_{i,t}$) represents the background number of new cases occurring in region i, regardless of the current number of cases in i, or in the regions around i. If $\nu_{i,t}$ is high, new cases in i are common, regardless of the number of cases in or around i at the previous time step. Since the endemic component does not depend on Y_{t-1}, it represents the background importations that cannot be linked to the other two components. Therefore, these cases either correspond to importations from outside the modelled area, or cases that are not otherwise predicted by the other two components.

The neighbouring and autoregressive component represent the 'Epidemic' part of the *hhh4* model, i.e. the section that depends on current levels of transmission in the dataset. The number of observed cases at t in i $Y_{i,t}$, usually follows a negative binomial distribution to allow for overdispersion [151], with the overdispersion parameter ψ being estimated by the model. The *hhh4* model can also be parametrised as a Poisson regression. The full equation for the mean number of cases in region i at time t is:

$$\mu_{i,t} = \nu_{i,t} + \lambda_{i,t} * Y_{i,t-1} + \phi_{i,t} * \sum_{j \neq i}(\omega_{ji} * Y_{j,t-1}) \quad (1)$$

The predictors $\lambda_{i,t}$, $\phi_{i,t}$ and $\nu_{i,t}$ can be independently impacted by different covariates, e.g. a covariate may be associated with a reduction in the number of importations, but have little impact on the spread of the virus within the region. Each predictor is estimated using log-linear regressions, containing the following: i) The intercept α (identical across spatial units), and ii) the vector of coefficients β associated with $z_{i,t}$, the vector of covariates at t in i included in each component.

$$log(\lambda_{i,t}) = \alpha^{(\lambda)} + \beta^{(\lambda)} * z_{it}^{(\lambda)} \quad (2)$$

$$log(\phi_{i,t}) = \alpha^{(\phi)} + \beta^{(\phi)} * z_{it}^{(\phi)} \quad (3)$$

$$log(\nu_{i,t}) = \alpha^{(\nu)} + \beta^{(\nu)} * z_{it}^{(\nu)} \quad (4)$$

In recent years, the Epidemic-Endemic model has been applied to a variety of settings, and has been extended to provide additional features. For instance, spatial random effects can be added to the log-linear regressions (equations 2, 3, and 4) to account for spatial heterogeneity in the incidence that would not be explained by the covariates included in the model. The random effects assigned to each region are constant through time [149]. Recent developments allowed for the Epidemic components to account for the impact of cases reported several time steps before t on the current incidence [150]. Indeed, using only the number of cases at the previous time step can exclude transmission events with longer generation times, or unreported generations. The R package *surveillance* also integrates functions that evaluate the calibration of the model by generating one step ahead forecasts, i.e. by fitting the model at each time step of a "calibration period", which usually corresponds to the last n measures, and comparing the prediction of the model at the next time step to what was reported in the data.

Therefore, the Epidemic-Endemic framework provides a flexible type of model that can estimate the association between a variety of covariates and the number of local cases in each region. For instance, Herzog et al. applied this framework to aggregated measles data in Germany, and showed that they were able to capture the incidence in all states using the proportion unvaccinated as a covariate in the autoregressive component [80].

Using the number of cases reported in the past three years as a covariate of the model, one could estimate whether regions eligible for the elimination status (i.e. those that reported low levels of incidence in the past three years) were associated with lower chances of onwards transmission. In other words, whether the values of the three predictors are lower when fewer cases were reported in the past three years. So far, the Epidemic-Endemic has only been applied to temporally aggregated case counts, but using daily case counts may bring more granularity and accuracy to the estimation procedure. The Epidemic section would then have to be modified. Indeed, using only the number of cases on the previous day to compute the current number of cases would lead to bias, since it would imply that direct transmission happens between cases reported on successive dates.

1.4. Summary

The resurgence of measles observed in many countries near elimination since 2017 highlights the importance of developing tools to identify areas with low immunity to the virus, and to evaluate indicators of vulnerability. Programs of routine measles surveillance and vaccination have been implemented for several decades. Therefore, there is a need to assess how these widely available data can be used to gain insights into the dynamics of measles transmission in countries with relatively low national incidence and high vaccine uptake. In this thesis, I aim to introduce methods combining

routinely collected data to describe the sub-national heterogeneity in vulnerability to measles outbreaks, and highlight the association between various indicators and the risks of measles transmission.

1.5. References

[1] Anderson RM, May RM. Infectious Diseases of Humans Dynamics and Control. Oxford: Oxford University Press; 1991.

[2] Perry RT, Halsey NA. The Clinical Significance of Measles: A Review. J Infect Dis 2004;189:S4–16. https://doi.org/10.1086/377712.

[3] World Health Organization. Measles, Fact sheet 2018. https://www.who.int/en/news-room/fact-sheets/detail/measles (accessed May 4, 2021).

[4] World Health Organization (WHO). Vaccine-Preventable Diseases, Surveillance Standard, Measles 2018. https://www.who.int/immunization/monitoring_surveillance/burden/vpd/WHO_SurveillanceVaccinePreventable_11_Measles_R2.pdf (accessed May 5, 2021).

[5] Gastanaduy PA, Redd SB, Clemmons NS, Lee AD, Hickman CJ, Rota PA, et al. Manual for the surveillance of vaccine-preventable diseases. Chapter 7: measles 2019.

[6] The European Commission. Commission Implementing Decision (EU) 2018/945 of 22 June 2018 on the communicable diseases and related special health issues to be covered by epidemiological surveillance as well as relevant case definitions. Off J Eur Union 2018.

[7] Klinkenberg D, Nishiura H. The correlation between infectivity and incubation period of measles, estimated from households with two cases. J Theor Biol 2011;284:52–60. https://doi.org/10.1016/j.jtbi.2011.06.015.

[8] Antia A, Ahmed H, Handel A, Carlson NE, Amanna IJ, Antia R, et al. Heterogeneity and longevity of antibody memory to viruses and vaccines. PLOS Biol 2018;16:e2006601. https://doi.org/10.1371/journal.pbio.2006601.

[9] Guerra FM, Bolotin S, Lim G, Heffernan J, Deeks SL, Li Y, et al. The basic reproduction number (R0) of measles: a systematic review. Lancet Infect Dis 2017;17:e420–8. https://doi.org/10.1016/S1473-3099(17)30307-9.

[10] Moss WJ. Measles. Lancet 2017;390:2490–502. https://doi.org/10.1016/S0140-6736(17)31463-0.

[11] Portnoy A, Jit M, Ferrari M, Hanson M, Brenzel L, Verguet S. Estimates of case-fatality ratios of measles in low-income and middle-income countries: a systematic review and modelling analysis. Lancet Glob Heal 2019;7:e472–81. https://doi.org/10.1016/S2214-109X(18)30537-0.

[12] Langmuir AD. Medical Importance of Measles. Am J Dis Child 1962;103:224–6. https://doi.org/10.1001/archpedi.1962.02080020236005.

[13] Edmunds WJ, Gay NJ, Kretzschmar M, Pebody RG, Wachmann H. The prevaccination epidemiology of measles, mumps and rubella in {E}urope: implications for modeling studies. Epidemiol Infect 2000;125:635–50.

[14] Engelhardt J, Halsey NA, Eddins DL, Hinman AR. Measles mortality in the United States 1971-1975. Am J Public Health 1980;70:1166–9. https://doi.org/10.2105/AJPH.70.11.1166.

[15] Shanks GD, Waller M, Briem H, Gottfredsson M. Age-specific measles mortality during the late 19th-early 20th centuries. Epidemiol Infect 2015;143:3434–41. https://doi.org/10.1017/S0950268815000631.

[16] Snyder MJ, McCrumb FR, Bigbee T, Schluederberg AE, Togo Y. Observations on the Seroepidemiology of Measles. Am J Dis Child 1962;103:250–1. https://doi.org/10.1001/archpedi.1962.02080020262012.

[17] Sencer DJ, Dull HB, Langmuir AD. Epidemiologic basis for eradication of measles in 1967. Public Health Rep 1967;82:253–6. https://doi.org/10.2307/4592985.

[18] Centers for Disease Control and Prevention (CDC). Measles Vaccination 2016. https://www.cdc.gov/measles/vaccination.html (accessed September 11, 2018).

[19] World Health Organization. Measles vaccines: WHO position paper, April 2017 – Recommendations. Vaccine 2019;37:219–22. https://doi.org/10.1016/j.vaccine.2017.07.066.

[20] Lochlainn L, de Gier B, van der Maas N, Rots N, van Binnendijk R, de Melker H, et al. Measles vaccination below 9 months of age: Systematic literature review and meta-analyses of effects and safety. Natl Inst Public Heal Environ 2015:1–109.

[21] Moss W, Scott S. WHO Immunological Basis for Immunization Series. World Heal Organ 2017:1–40.

[22] Carazo S, Billard MN, Boutin A, De Serres G. Effect of age at vaccination on the measles vaccine effectiveness and immunogenicity: Systematic review and meta-Analysis. BMC Infect Dis 2020;20:1–18. https://doi.org/10.1186/s12879-020-4870-x.

[23] Brugha R, Ramsay M, Forsey T, Brown D. A study of maternally derived measles antibody in infants born to naturally infected and vaccinated women. Epidemiol Infect 1996;117:519–24. https://doi.org/10.1017/S0950268800059203.

[24] Waaijenborg S, Hahné SJM, Mollema L, Smits GP, Berbers GAM, Van Der Klis FRM, et al. Waning of maternal antibodies against measles, mumps, rubella, and varicella in communities with contrasting vaccination coverage. J Infect Dis 2013;208:10–6. https://doi.org/10.1093/infdis/jit143.

[25] Linder N, Tallen-Gozani E, German B, Duvdevani P, Ferber A, Sirota L. Placental transfer of measles antibodies: Effect of gestational age and maternal vaccination status. Vaccine 2004;22:1509–14. https://doi.org/10.1016/j.vaccine.2003.10.009.

[26] Guerra FM, Crowcroft NS, Friedman L, Deeks SL, Halperin SA, Severini A, et al. Waning of measles maternal antibody in infants in measles elimination settings – A systematic literature review. Vaccine 2018;36:1248–55. https://doi.org/10.1016/j.vaccine.2018.01.002.

[27] Zhang X, Shirayama Y, Zhang Y, Ba W, Ikeda N, Mori R, et al. Duration of maternally derived antibody against measles: A seroepidemiological study of infants aged under 8 months in Qinghai, China. Vaccine 2012;30:752–7. https://doi.org/10.1016/j.vaccine.2011.11.078.

[28] Metintaş S, Akgün Y, Arslantaş D, Kalyoncu C, Uçar B. Decay of maternally derived measles antibody in central Turkey. Public Health 2002;116:50–4. https://doi.org/10.1038/sj.ph.1900818.

[29] Portnoy A, Jit M, Helleringer S, Verguet S. Impact of measles supplementary immunization activities on reaching children missed by routine programs. Vaccine 2018;36:170–8. https://doi.org/10.1016/j.vaccine.2017.10.080.

[30] Kisangau N, Sergon K, Ibrahim Y, Yonga F, Langat D, Nzunza R, et al. Progress towards elimination of measles in Kenya, 2003-2016. Pan Afr Med J 2018;31:1–11. https://doi.org/10.11604/pamj.2018.31.65.16309.

[31] Simone B, Balasegaram S, Gobin M, Anderson C, Charlett A, Coole L, et al. Evaluation of the measles, mumps and rubella vaccination catch-up campaign in England in 2013. Vaccine 2014;32:4681–8. https://doi.org/10.1016/j.vaccine.2014.05.077.

[32] Patel MK, Goodson JL, Alexander JP, Kretsinger K, Sodha S V, Steulet C. Progress Toward Regional Measles Elimination — Worldwide , 2000 – 2019 2020;69:1700–5.

[33] Perry RT, Gacic-Dobo M, Dabbagh A, Mulders MN, Strebel PM, Okwo-Bele JM, et al. Global

control and regional elimination of measles, 2000-2012. Morb Mortal Wkly Rep 2014;63:103–7.

[34] Perry RT, Gacic-Dobo M, Dabbagh A, Mulders MN, Strebel PM, Okwo-Bele J-M, et al. Progress toward regional measles elimination--worldwide, 2000-2013. MMWR Morb Mortal Wkly Rep 2014;63:1034–8. https://doi.org/10.15585/mmwr.mm6642a6.

[35] World Health Organization. Framework for verifying elimination of measles and rubella. Wkly Epidemiol Rec 2013;88:89–100. https://doi.org/10.1371/jour.

[36] World Health Organization. WHO | Measles. WHO 2018. http://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/active/measles/en/#.W5j1pTrrM64.mendeley (accessed May 18, 2021).

[37] Team E editorial. Expert committee declares WHO Region of the Americas measles-free. Euro Surveill 2016;21. https://doi.org/10.2807/1560-7917.ES.2016.21.39.30360.

[38] Katz SL, Hinman AR. Summary and Conclusions: Measles Elimination Meeting, 16–17 March 2000. J Infect Dis 2004;189:S43–7. https://doi.org/10.1086/377696.

[39] Robert A, Funk S, Kucharski AJ. The measles crisis in Europe—the need for a joined-up approach. Lancet 2019;393:2033. https://doi.org/10.1016/S0140-6736(19)31039-6.

[40] Zimmerman LA, Muscat M, Singh S, Ben Mamou M, Jankovic D, Datta S, et al. Progress Toward Measles Elimination — European Region, 2009–2018. MMWR Morb Mortal Wkly Rep 2019;68:396–401. https://doi.org/10.15585/mmwr.mm6817a4.

[41] European Centre for Disease Prevention and Control. Ongoing outbreak of measles in Romania, risk of spread and epidemiological situation in EU/EEA countries 2017:11.

[42] Georgakopoulou T, Horefti E, Vernardaki A, Pogka V, Gkolfinopoulou K, Triantafyllou E, et al. Ongoing measles outbreak in Greece related to the recent European-wide epidemic. Epidemiol Infect 2018;146:1692–8. https://doi.org/10.1017/S0950268818002170.

[43] Bernadou A, Astrugue C, Méchain M, Le Galliard V, Verdun-Esquer C, Dupuy F, et al. Measles outbreak linked to insufficient vaccination coverage in Nouvelle-Aquitaine region, France, October 2017 to July 2018. Eurosurveillance 2018;23:1–5. https://doi.org/10.2807/1560-7917.ES.2018.23.30.1800373.

[44] Wadman M. Measles epidemic in Ukraine drove troubling European year. Science (80-) 2019;363:677–8. https://doi.org/10.1126/science.363.6428.677.

[45] Organization PAH. Epidemiological Update Measles 2018:20.

[46] World Health Organization (WHO). European Region loses ground in effort to eliminate measles 2019.

[47] Pan American Health Organization. Epidemiological Update: Measles. Paho/ Who 2019;2020:1–12.

[48] Mossong J, O'Callaghan CJ, Ratnam S. Modelling antibody response to measles vaccine and subsequent waning of immunity in a low exposure population. Vaccine 2000;19:523–9. https://doi.org/10.1016/S0264-410X(00)00175-4.

[49] Smith LE, Amlôt R, Weinman J, Yiend J, Rubin GJ. A systematic review of factors affecting vaccine uptake in young children. Vaccine 2017;35:6059–69. https://doi.org/10.1016/j.vaccine.2017.09.046.

[50] Fine P, Eames K, Heymann DL. "Herd immunity": A rough guide. Clin Infect Dis 2011;52:911–6. https://doi.org/10.1093/cid/cir007.

[51] Nokes DJ, Anderson RM. The use of mathematical models in the epidemiological study of infectious diseases and in the design of mass immunization programmes. Epidemiol Infect 1988;101:1–20. https://doi.org/10.1017/S0950268800029186.

[52] Wallinga J, Heijne JCM, Kretzschmar M. A measles epidemic threshold in a highly vaccinated population. PLoS Med 2005;2:1152–7. https://doi.org/10.1371/journal.pmed.0020316.

[53] Fox JP, Elveback L, Scott W, Gatewood L, Ackerman E. Herd immunity: Basic concept and relevance to public health immunization practices. Am J Epidemiol 1995;141:187–97. https://doi.org/10.1093/oxfordjournals.aje.a117420.

[54] McLean AR, Anderson RM. Measles in developing countries. Part II. The predicted impact of mass vaccination. Epidemiol Infect 1988;100:419–42. https://doi.org/10.1017/S0950268800067170.

[55] Mossong J, Muller CP. Modelling measles re-emergence as a result of waning of immunity in vaccinated populations. Vaccine 2003;21:4597–603. https://doi.org/10.1016/S0264-410X(03)00449-3.

[56] Metcalf CJE, Wesolowski A, Winter AK, Lessler J, Cauchemez S, Moss WJ, et al. Using Serology to Anticipate Measles Post-honeymoon Period Outbreaks. Trends Microbiol 2020;28:597–600. https://doi.org/10.1016/j.tim.2020.04.009.

[57] Ramsay M. A strategic framework for the elimination of measles in the European Region. 1997.

[58] Funk S, Knapp JK, Lebo E, Reef SE, Dabbagh AJ, Kretsinger K, et al. Combining serological and

contact data to derive target immunity levels for achieving and maintaining measles elimination. BMC Med 2019. https://doi.org/10.1186/s12916-019-1413-7.

[59] Casiday R, Cresswell T, Wilson D, Panter-Brick C. A survey of UK parental attitudes to the MMR vaccine and trust in medical authority. Vaccine 2006;24:177–84. https://doi.org/10.1016/j.vaccine.2005.07.063.

[60] Brown KF, Kroll JS, Hudson MJ, Ramsay M, Green J, Long SJ, et al. Factors underlying parental decisions about combination childhood vaccinations including MMR: A systematic review. Vaccine 2010;28:4235–48. https://doi.org/10.1016/j.vaccine.2010.04.052.

[61] Brown KF, Long SJ, Ramsay M, Hudson MJ, Green J, Vincent CA, et al. UK parents' decision-making about measles-mumps-rubella (MMR) vaccine 10 years after the MMR-autism controversy: A qualitative analysis. Vaccine 2012;30:1855–64. https://doi.org/10.1016/j.vaccine.2011.12.127.

[62] Chen W, Landau S, Sham P, Fombonne E. No evidence for links between autism, MMR and measles virus. Psychol Med 2004. https://doi.org/10.1017/S0033291703001259.

[63] Dales L, Hammer SJ, Smith NJ. Time trends in autism and in MMR immunization coverage in California. J Am Med Assoc 2001. https://doi.org/10.1001/jama.285.9.1183.

[64] Hornig M, Briese T, Buie T, Bauman ML, Lauwers G, Siemetzki U, et al. Lack of association between measles virus vaccine and autism with enteropathy: A case-control study. PLoS One 2008;3:1–8. https://doi.org/10.1371/journal.pone.0003140.

[65] Health and Social Care Information Centre. NHS Immunisation Statistics. 2014.

[66] Public Health England. Evaluation of vaccine uptake during the 2013 MMR catch-up campaign in England Report for the national measles oversight group About Public Health England 2014.

[67] Le Menach A, Boxall N, Amirthalingam G, Maddock L, Balasegaram S, Mindlin M. Increased measles-mumps-rubella (MMR) vaccine uptake in the context of a targeted immunisation campaign during a measles outbreak in a vaccine-reluctant community in England. Vaccine 2014;32:1147–52. https://doi.org/10.1016/j.vaccine.2014.01.002.

[68] Pearce A, Mindlin M, Cortina-Borja M, Bedford H. Characteristics of 5-year-olds who catch-up with MMR: Findings from the UK Millennium Cohort Study. BMJ Open 2013;3:1–9. https://doi.org/10.1136/bmjopen-2013-003152.

[69] Vivancos R, Keenan A, Farmer S, Atkinson J, Coffey E, Dardamissis E, et al. An ongoing large outbreak of measles in Merseyside, England, January to June 2012. Eurosurveillance 2012;17:1–

5. https://doi.org/20226 [pii].

[70] le Polain de Waroux O, Saliba V, Cottrell S, Young N, Perry M, Bukasa A, et al. Summer music and arts festivals as hot spots for measles transmission: Experience from England and Wales, June to October 2016. Eurosurveillance 2016;21:1–6. https://doi.org/10.2807/1560-7917.ES.2016.21.44.30390.

[71] Datta SS, O'Connor PM, Jankovic D, Muscat M, Ben Mamou MC, Singh S, et al. Progress and challenges in measles and rubella elimination in the WHO European Region. Vaccine 2018;36:5408–15. https://doi.org/10.1016/j.vaccine.2017.06.042.

[72] Muscat M, Bang H, Wohlfahrt J, Glismann S, Mølbak K. Measles in Europe: an epidemiological assessment. Lancet 2009;373:383–9. https://doi.org/10.1016/S0140-6736(08)61849-8.

[73] World Health Organisation Regional Office for Europe. Seventh Meeting of the European Regional Verification Commission for Measles and Rubella Elimination (RVC) 2018.

[74] Pfaff G, Lohr D, Santibanez S, Mankertz A, van Treeck U, Schönberger K, et al. Spotlight on measles 2010: Measles outbreak among travellers returning from a mass gathering, Germany, September to October 2010. Eurosurveillance 2010;15:4–7. https://doi.org/10.2807/ese.15.50.19750-en.

[75] Gautret P, Steffen R. Communicable diseases as health risks at mass gatherings other than Hajj: What is the evidence? Int J Infect Dis 2016;47:46–52. https://doi.org/10.1016/j.ijid.2016.03.007.

[76] Truelove SA, Graham M, Moss WJ, Metcalf CJE, Ferrari MJ, Lessler J. Characterizing the impact of spatial clustering of susceptibility for measles elimination. Vaccine 2019;37:732–41. https://doi.org/10.1016/j.vaccine.2018.12.012.

[77] Gastañaduy PA, Budd J, Fisher N, Redd SB, Fletcher J, Miller J, et al. A Measles Outbreak in an Underimmunized Amish Community in Ohio. N Engl J Med 2016;375:1343–54. https://doi.org/10.1056/NEJMoa1602295.

[78] Under-vaccinated Groups in Europe: Who are they and how to communicate with them in outbreak situations? n.d. http://ecomeu.info/wp-content/uploads/2015/11/ECOM-Under-vaccinated-groups-in-Europe-WP6.pdf (accessed May 6, 2021).

[79] Muscat M, Marinova L, Mankertz A, Gatcheva N, Mihneva Z, Santibanez S, et al. The measles outbreak in Bulgaria, 2009-2011: An epidemiological assessment and lessons learnt. Eurosurveillance 2016;21:no pagination. https://doi.org/10.2807/1560-7917.ES.2016.21.9.30152.

[80] Herzog SA, Paul M, Held L. Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. Epidemiol Infect 2011;139:505–15. https://doi.org/10.1017/S0950268810001664.

[81] McDonald R, Ruppert PS, Souto M, Johns DE, McKay K, Bessette N, et al. Notes from the Field: Measles Outbreaks from Imported Cases in Orthodox Jewish Communities — New York and New Jersey, 2018–2019. MMWR Morb Mortal Wkly Rep 2019;68:444–5. https://doi.org/10.15585/mmwr.mm6819a4.

[82] Toffolutti V, McKee M, Melegaro A, Ricciardi W, Stuckler D. Austerity, measles and mandatory vaccination: Cross-regional analysis of vaccination in Italy 2000-14. Eur J Public Health 2019;29:123–7. https://doi.org/10.1093/eurpub/cky178.

[83] Danis K, Georgakopoulou T, Stavrou T, Laggas D, Panagiotopoulos T. Socioeconomic factors play a more important role in childhood vaccination coverage than parental perceptions: a cross-sectional study in Greece. Vaccine 2010;28:1861–9. https://doi.org/10.1016/j.vaccine.2009.11.078.

[84] Muscat M. Who gets measles in Europe? J Infect Dis 2011;204:353–65. https://doi.org/10.1093/infdis/jir067.

[85] Dubé E, Gagnon D, Nickels E, Jeram S, Schuster M. Mapping vaccine hesitancy-Country-specific characteristics of a global phenomenon. Vaccine 2014;32:6649–54. https://doi.org/10.1016/j.vaccine.2014.09.039.

[86] Olive JK, Hotez PJ, Damania A, Nolan MS. Erratum: The state of the antivaccine movement in the United States: A focused examination of nonmedical exemptions in states and counties (PLoS Med (2018) 15:6 (e1002578) DOI: 10.1371/journal.pmed.1002578). PLoS Med 2018;15:1–10. https://doi.org/10.1371/JOURNAL.PMED.1002616.

[87] Hotez PJ, Nuzhath T, Colwell B. Combating vaccine hesitancy and other 21st century social determinants in the global fight against measles. Curr Opin Virol 2020;41:1–7. https://doi.org/10.1016/j.coviro.2020.01.001.

[88] Callaghan T, Motta M, Sylvester S, Lunz Trujillo K, Blackburn CC. Parent psychology and the decision to delay childhood vaccination. Soc Sci Med 2019;238:112407. https://doi.org/10.1016/j.socscimed.2019.112407.

[89] Health Organization W, Office for Europe R. Guidance on conducting serosurveys in support of measles and rubella elimination in the WHO European Region. World Heal Organ 2013:1–10.

[90] Woudenberg T, Woonink F, Kerkhof J, Cox K, Ruijs WLM. The tip of the iceberg : incompleteness of measles reporting during a large outbreak in The Netherlands in 2013 – 2014. Epidemiol Infect 2018;146:716–22. https://doi.org/https://doi.org/10.1017/ S0950268818002698.

[91] Beaty SM, Lee B. Constraints on the genetic and antigenic variability of measles virus. Viruses 2016;8:1–20. https://doi.org/10.3390/v8040109.

[92] World Health Organisation. Measles virus nomenclature Update: 2012. Wkly Epidemiol Rec 2012;87:73–80. https://doi.org/10.1016/j.actatropica.2012.04.013.

[93] Hiebert J, Severini A. Measles molecular epidemiology: What does it tell us and why is it important? Canada Commun Dis Rep 2014;40:257–60. https://doi.org/10.14745/ccdr.v40i12a06.

[94] Rota JS, Heath JL, Rota PA, King GE, Celma ML, Carabafia J, et al. Molecular Epidemiology of Measles Virus: Identification of Pathways of Transmission and Implications for Measles Elimination. J Infect Dis 1996;173:32–7. https://doi.org/10.1093/infdis/173.1.32.

[95] Brown KE, Rota PA, Goodson JL, Williams D, Abernathy E, Takeda M, et al. Genetic characterization of measles and rubella viruses detected through global measles and rubella elimination surveillance, 2016-2018. Morb Mortal Wkly Rep 2019;68:587–91. https://doi.org/10.15585/mmwr.mm6826a3.

[96] Penedos AR, Myers R, Hadef B, Aladin F, Brown KE. Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks 2015:1–16. https://doi.org/10.1371/journal.pone.0143081.

[97] Bodewes R, Reijnen L, Zwagemaker F, Kohl RHG, Kerkhof J, Veldhuijzen IK, et al. An efficient molecular approach to distinguish chains of measles virus transmission in the elimination phase. Infect Genet Evol 2021;91:104794. https://doi.org/10.1016/j.meegid.2021.104794.

[98] Brown DW, Warrener L, Scobie HM, Donadel M, Waku-Kouomou D, Mulders MN, et al. Rapid diagnostic tests to address challenges for global measles surveillance. Curr Opin Virol 2020;41:77–84. https://doi.org/10.1016/j.coviro.2020.05.007.

[99] Glasser J, Meltzer M, Levin B. Mathematical modeling and public policy: responding to health crises. Emerg Infect Dis 2004;10:2050–1. https://doi.org/10.3201/eid1011.04079708.

[100] Grassly NC, Fraser C. Mathematical models of infectious disease transmission. Nat Rev Microbiol 2008;6:477–87. https://doi.org/10.1038/nrmicro1845.

[101] Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. 2008. https://doi.org/10.1038/453034a.

[102] Armitage P, Berry G, Mathews JN. Statistical Methods in Medical Research. vol. 49. 2008.

[103] Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. Mach Learn 2003;50:5–43. https://doi.org/10.1023/A:1020281327116.

[104] Mclean AR, Anderson RM. Measles in developing countries Part I. Epidemiological parameters and patterns. Epidemiol Infect 1988;100:111–33. https://doi.org/10.1017/S0950268800065614.

[105] Fine PEM, Clarkson JA. Measles in England and Wales--1: An analysis of factors underlying seasonal patterns. Int J Epidemiol 1982;11:5–15. https://doi.org/10.1093/ije/11.1.5.

[106] Grenfell BT, Bjørnstad ON, Kappey J. Travelling waves and spartial hierarchies in measles epidemics. Nature 2001;414:716–23.

[107] Finkenstädt B, Grenfell B. Empirical determinants of measles metapopulation dynamics in England and Wales. Proc Biol Sci 1998;265:211–20. https://doi.org/10.1098/rspb.1998.0284.

[108] Ramsay M, Gay N, Miller E, Rush M, White J, Morgan-Capner P, et al. The epidemiology of measles in England and Wales: rationale for the 1994 national vaccination campaign. Commun Dis Rep CDR Rev 1994:R141-6.

[109] Merler S, Ajelli M. Deciphering the relative weights of demographic transition and vaccination in the decrease of measles incidence in Italy. Proc R Soc B Biol Sci 2014;281:20132676–20132676. https://doi.org/10.1098/rspb.2013.2676.

[110] Xia Y, Bjørnstad ON, Grenfell BT. Measles Metapopulation Dynamics: A Gravity Model for Epidemiological Coupling and Dynamics. Am Nat 2004;164:267–81. https://doi.org/10.1086/422341.

[111] Grenfell BT, Bjørnstad ON, Finkenstädt BF. Dynamics of measles epidemics: Scaling noise, determinism, and predictability with the TSIR model. Ecol Monogr 2002;72:185–202. https://doi.org/10.1890/0012-9615(2002)072[0185:DOMESN]2.0.CO;2.

[112] Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of measles epidemics: Estimating scaling of transmission rates using a Time series SIR model. Ecol Monogr 2002;72:169–84. https://doi.org/10.1890/0012-9615(2002)072[0169:DOMEES]2.0.CO;2.

[113] Wallinga J, Teunis P, Kretzschmar M. Reconstruction of measles dynamics in a vaccinated population. Vaccine 2003;21:2643–50. https://doi.org/10.1016/S0264-410X(03)00051-3.

[114] Blumberg S, Enanoria WTA, Lloyd-Smith JO, Lietman TM, Porco TC. Identifying postelimination trends for the introduction and transmissibility of measles in the United States. Am J Epidemiol 2014;179:1375–82. https://doi.org/10.1093/aje/kwu068.

[115] Blumberg S, Funk S, Pulliam JRC. Detecting Differential Transmissibilities That Affect the Size of Self-Limited Outbreaks. PLoS Pathog 2014;10. https://doi.org/10.1371/journal.ppat.1004452.

[116] Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLoS Comput Biol 2013;9:1–17. https://doi.org/10.1371/journal.pcbi.1002993.

[117] Lau MSY, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, et al. Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. Proc Natl Acad Sci 2017;114:2337–42. https://doi.org/10.1073/pnas.1614595114.

[118] Stein RA. Super-spreaders in infectious diseases. Int J Infect Dis 2011;15:e510–3. https://doi.org/10.1016/j.ijid.2010.06.020.

[119] Endo A, Abbott S, Kucharski AJ, Funk S. Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. Wellcome Open Res 2020;5:67. https://doi.org/10.12688/wellcomeopenres.15842.3.

[120] Lenormand M, Bassolas A, Ramasco JJ. Systematic comparison of trip distribution laws and models. J Transp Geogr 2016;51:158–69. https://doi.org/10.1016/j.jtrangeo.2015.12.008.

[121] Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C. Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 2014;41:260–71. https://doi.org/10.1080/15230406.2014.890072.

[122] Zipf GK. The P 1 P 2/D hypothesis: On the intercity movement of persons. Am Sociol Rev 1946;11:677–86. https://doi.org/10.2307/2657358.

[123] Truscott J, Ferguson NM. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. PLoS Comput Biol 2012;8. https://doi.org/10.1371/journal.pcbi.1002699.

[124] Gog JR, Ballesteros S, Viboud C, Simonsen L, Bjornstad ON, Shaman J, et al. Spatial Transmission of 2009 Pandemic Influenza in the US. PLoS Comput Biol 2014;10. https://doi.org/10.1371/journal.pcbi.1003635.

[125] Stouffer SA. Intervening Opportunities: A Theory Relating Mobility and Distance. Am Sociol Rev

1940. https://doi.org/10.2307/2084520.

[126] Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C. A tale of many cities: Universal patterns in human urban mobility. PLoS One 2012;7. https://doi.org/10.1371/journal.pone.0037027.

[127] Simini F, González MC, Maritan A, Barabási AL. A universal model for mobility and migration patterns. Nature 2012;484:96–100. https://doi.org/10.1038/nature10856.

[128] Fotheringham AS. Spatial flows and spatial patterns. Environ Plan A 1984;16:529–43. https://doi.org/10.1068/a160529.

[129] Bjørnstad ON, Grenfell BT, Viboud C, King AA. Comparison of alternative models of human movement and the spread of disease. BioRxiv 2019:1–15. https://doi.org/10.1101/2019.12.19.882175.

[130] Robert A, Edmunds WJ, Watson CH, Henao-Restrepo AM, Gsell P-S, Williamson E, et al. Determinants of Transmission Risk During the Late Stage of the West African Ebola Epidemic. Am J Epidemiol 2019. https://doi.org/10.1093/aje/kwz090.

[131] Keeling MJ, Hollingsworth TD, Read JM. Efficacy of contact tracing for the containment of the 2019 novel coronavirus (COVID-19). J Epidemiol Community Health 2020;74:861–6. https://doi.org/10.1136/jech-2020-214051.

[132] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B Biol Sci 2007;274:599–604. https://doi.org/10.1098/rspb.2006.3754.

[133] Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal. Am J Epidemiol 2004;160:509–16.

[134] Cauchemez S, Ferguson NM. Methods to infer transmission risk factors in complex outbreak data. J R Soc Interface 2012;9:456–69. https://doi.org/10.1098/rsif.2011.0379.

[135] Cauchemez S, Boëlle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, et al. Real-time estimates in early detection of SARS. Emerg Infect Dis 2006.

[136] Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science (80-) 2014;345:1369--1372. https://doi.org/10.1126/science.1259657.

[137] Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. Nature 2015;524:97–101.

https://doi.org/10.1038/nature14594.

[138] Ruan YJ, Wei CL, Ee LA, Vega VB, Thoreau H, Yun STS, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet 2003;361:1779–85. https://doi.org/10.1016/S0140-6736(03)13414-9.

[139] Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 2009;10:540–50. https://doi.org/10.1038/nrg2583.

[140] Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Comput Biol 2014;10. https://doi.org/10.1371/journal.pcbi.1003457.

[141] Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: A modular platform for outbreak reconstruction. BMC Bioinformatics 2018;19. https://doi.org/10.1186/s12859-018-2330-z.

[142] Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc R Soc B Biol Sci 2012;279:444–50. https://doi.org/10.1098/rspb.2011.0913.

[143] Cottam EM, Thébaud G, Wadsworth J, Gloster J, Mansley L, Paton DJ, et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. Proc R Soc B Biol Sci 2008;275:887–95. https://doi.org/10.1098/rspb.2007.1442.

[144] Gardy JL, Naus M, Amlani A, Chung W, Kim H, Tan M, et al. Whole-genome sequencing of measles virus genotypes H1 and D8 during outbreaks of infection following the 2010 Olympic Winter Games reveals viral transmission routes. J Infect Dis 2015;212:1574–8. https://doi.org/10.1093/infdis/jiv271.

[145] Edmunds WJ, Kafatos G, Wallinga J, Mossong JR. Mixing patterns and the spread of close-contact infectious diseases. Emerg Themes Epidemiol 2006;3:1–8. https://doi.org/10.1186/1742-7622-3-10.

[146] Wallinga J, Teunis P, Kretzschmar M. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. Am J Epidemiol 2006;164:936–44. https://doi.org/10.1093/aje/kwj317.

[147] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 2008;5:0381–91.

https://doi.org/10.1371/journal.pmed.0050074.

[148] Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Stat Modelling 2005;5:187–99. https://doi.org/10.1191/1471082X05st098oa.

[149] Paul M, Held L. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. Stat Med 2011;30:1118–36. https://doi.org/10.1002/sim.4177.

[150] Bracher J, Held L. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. Int J Forecast 2020. https://doi.org/10.1016/j.ijforecast.2020.07.002.

[151] Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355–9. https://doi.org/10.1038/nature04153.

[152] Faye O, Boëlle P-Y, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect Dis 2015;15:320–6. https://doi.org/10.1016/S1473-3099(14)71075-8.

[153] Leclerc QJ, Fuller NM, Knight LE, Funk S, Knight GM. What settings have been linked to SARS-CoV-2 transmission clusters? Wellcome Open Res 2020;5. https://doi.org/10.12688/wellcomeopenres.15889.2.

Chapter 2. *o2geosocial*: Reconstructing who-infected-whom from routinely collected surveillance data

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1704167	**Title**	Mr
First Name(s)	Alexis		
Surname/Family Name	Robert		
Thesis Title	Modelling the risks of measles outbreaks near elimination		
Primary Supervisor	Sebastian Funk		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	F1000 Research		
When was the work published?	January 2021		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	**No**	Was the work subject to academic peer review?	**No**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

SECTION E

Student Signature	[redacted]
Date	31/05/2021

Supervisor Signature	[redacted]
Date	31/5/21

2.1. Abstract

Reconstructing the history of individual transmission events between cases is key to understanding what factors facilitate the spread of an infectious disease. Since conducting extended contact-tracing investigations can be logistically challenging and costly, statistical inference methods have been developed to reconstruct transmission trees from onset dates and genetic sequences. However, these methods are not as effective if the mutation rate of the virus is very slow, or if sequencing data is sparse.

We developed the package o2geosocial to combine variables from routinely collected surveillance data with a simple transmission process model. The model reconstructs transmission trees when full genetic sequences are unavailable, or uninformative. Our model incorporates the reported age-group, onset date, location and genotype of infected cases to infer probabilistic transmission trees. The package also includes functions to summarise and visualise the inferred cluster size distribution.

The results generated by o2geosocial can highlight regions where importations repeatedly caused large outbreaks, which may indicate a higher regional susceptibility to infections. It can also be used to generate the individual number of secondary transmissions, and show the features associated with individuals involved in high transmission events.

The package is available for download from the Comprehensive R Archive Network (CRAN) and GitHub.

2.2. Introduction

The identification of transmission trees and transmission events during infectious disease outbreaks can lead to identifying factors and settings associated with subsequent transmissions [1–4], describing super-spreading events [5,6], or populations and areas more vulnerable to importations and transmission [7–10], and quantifying the impact of control measures [11,12]. The most straightforward approach to reconstructing who-infected-whom is to carry out patient interviews and establish the previous contacts to connect the reported cases. However, contact-tracing investigations are costly and can be challenging to implement. Statistical methods have therefore been developed to infer transmission trees from routinely collected epidemiological data [12–17].

The Wallinga-Teunis method was first developed to infer probabilistic transmission trees from onset dates and generation times in a maximum likelihood framework [12]. Genetic sequencing of pathogens have since become more common, and new tools such as the R package outbreaker2 were created to combine the timing of infection and the genetic sequences in order to improve the accuracy of inferred transmission trees [13,14,18–20]. Nevertheless, the accuracy of these reconstruction methods relies on the proportion of sequenced cases, the quality of the sequences, and the characteristics of the pathogen

[21]. For instance, the measles virus evolves very slowly, and sequences from unrelated cases can be very similar, which makes these methods ineffective for measles outbreaks [22,23].

The package *o2geosocial* was designed to study outbreaks where sequences are uninformative, either because too few cases were sequenced or because the virus evolves too slowly. Building upon the framework presented in *outbreaker2*, *o2geosocial* was developed to infer who-infected-whom from variables routinely collected by surveillance systems, such as the onset date, age, location, and genotype of the cases [7]. Cases from different genotypes cannot be part of a similar transmission chain since differences in genotype illustrate major variations in their genetic sequences [24]. Using age-stratified contact matrices and mobility models, we combined the different variables into a likelihood of connection between cases. In this paper, we first describe the structure of the package. From a use case based on simulated data, we then show how to run the model, evaluate the output, visualise the results of the inference, and customise the input functions to implement different mobility models.

2.3. Methods

2.3.1. Operation

o2geosocial is implemented as an open-source R (version ≥ 3.5.0) package and can be run on all platforms (Windows, Mac, Linux). It incorporates C++ functions into a R framework using *Rcpp* [25]. Package dependencies and system requirements are documented in the *o2geosocial* CRAN repository. A stable version was released on Windows, Mac and Linux operating systems via a CRAN repository. The source code is available through Zenodo [26] and the latest development version is available through a Github repository.

```r
# install from CRAN
install.packages("o2geosocial")

# install from Github
install.packages("devtools")
devtools::install_github("alxsrobert/o2geosocial")
```

The main function of the package, called `outbreaker()`, uses Monte Carlo Markov Chains (MCMC) to sample from the posterior distribution of the underlying model [27] . For each case, it infers the infection date, the infector, and the number of missing generations between the case and their infector. It takes five lists as inputs: i) `moves`, ii) `likelihoods`, iii) `priors`, iv) `data`, and v) `config`. These five lists can be generated and modified using the functions `custom_moves()`, `custom_likelihoods()`, `custom_priors()`, `create_config()` and `outbreaker_data()`.

2.3.2. Implementation

The general implementation of *o2geosocial* follows the structure of *outbreaker2* and builds upon it by adding the effect of the location and the age-stratified contact data to the reconstruction of transmission clusters. However, unlike *outbreaker2*, *o2geosocial* does not take genetic sequences as input. It uses genetic groups (*e.g.* genotype) to exclude connections between cases, *i.e.* two cases cannot be from the same cluster if they belong to different genetic groups [28] . Therefore, *o2geosocial* is adapted to reconstructing transmission clusters from large datasets where genetic sequences are not informative, either because the mutation rate of the virus is slow, or because sequencing is scarce.

In *o2geosocial*, the number of independent clusters in the dataset is inferred using two different processes (Figure 2.1). Firstly, the pre-clustering step aims to group cases before the MCMC runs following three criteria: Only cases reported in a radius of γ km, less than δ days before case i, and from similar or unreported genotype can be classified as potential infectors of i. Cases with overlapping potential infectors, and their potential infectors, are grouped together, and cases from different groups cannot be linked during the MCMC runs. The parameters γ and δ are defined as inputs of the function `create_config()`. Since surveillance datasets can include cases from unrelated outbreaks, the pre-clustering function was developed to remove impossible connections and speed up the MCMC runs.

Secondly, as cases classified in the same group after the pre-clustering step may come from different clusters, we defined a likelihood threshold λ to spot and discard unlikely connections after the MCMC runs: if the likelihood of connection from case j to case i is lower than λ, the connection is discarded and i is an import unrelated to j. In *o2geosocial*, the variable λ can either be an absolute (the log-likelihood threshold will be $\log(\lambda)$) or a relative value (a quantile of the likelihood of all connections in all samples), and is defined by the variables `outlier_threshold` and `outlier_relative` in `create_config()`.

Finally, unlikely connections between cases can alter the inferred infection dates of cases and bias the transmission trees sampled from the MCMC runs. Therefore, we first run a short MCMC to remove these unlikely connections. From this run we compute n, the minimum number of connections with a likelihood lower than λ per sampled tree. We then add n imports to the starting tree and run a longer MCMC. Lastly, we remove the connections with likelihood lower than λ in the final samples and return the infector, infection date and probability of being an import for each case (Figure 2.1).

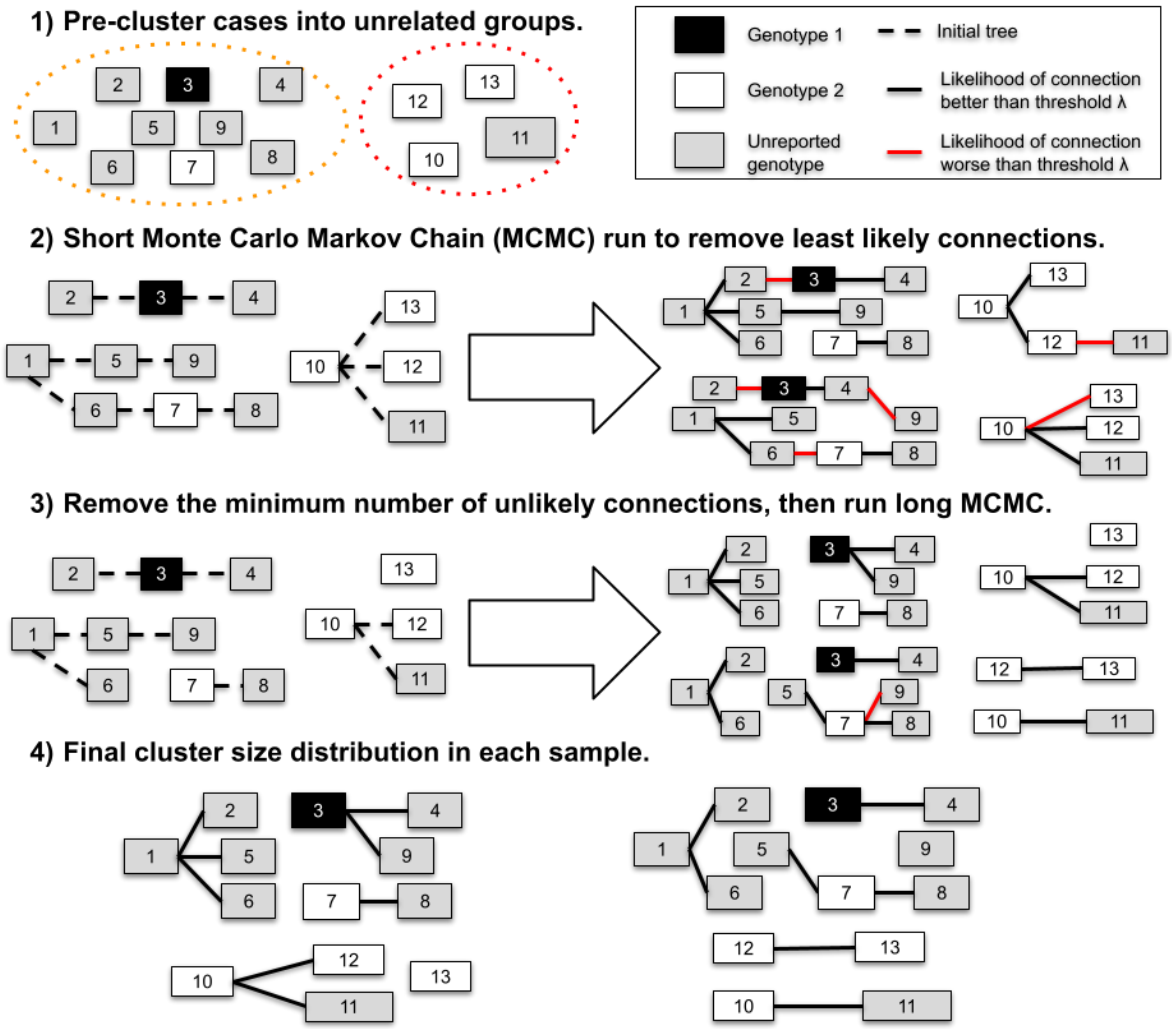

Figure 2.1: Illustration of the process to estimate the cluster size distribution and the import status of 13 *cases. In the first step, cases are split in two groups that do not have overlapping potential infectors (i.e. they were reported in different places, or different times). In step 2, we estimate the minimum number of unlikely transmissions (n) in the samples (right panel). In step 3, we remove n transmissions from the initial tree, and generate samples. Finally, we remove the unlikely connections in each sample of Step 3 and compute the inferred cluster size distribution.*

2.3.3. Likelihoods and priors

The functions `custom_likelihoods()` and `custom_priors()` can be used to edit each component of the likelihood and priors. By default, there are five components in the likelihood:

Genotype component: There can be a maximum of one genotype reported per transmission tree. The genotype of a tree τ is the genotype reported for at least one of the cases belonging to τ. For each genotyped case i_{gen} and at every iteration, only cases from trees with the same genotype as i_{gen}, or without reported genotype can be listed as potential infectors.

Therefore, the genetic component of the likelihood that a case i of genotype g_i was infected by a case j belonging to the tree τ_j is defined as a binary value:

$$G(g_i, g_{\tau_j}) = \begin{cases} 1 \text{ if } g_i \text{ unknown} \\ 1 \text{ if } g_{\tau_j} \text{ unknown} \\ 1 \text{ if } g_i \text{ and } g_{\tau_j} \text{ both known and } g_i = g_{\tau_j} \\ 0 \text{ otherwise} \end{cases}$$

Conditional report ratio: As in the package *outbreaker2*, we allow for missing cases in transmission chains. The number of generations between cases i and j, denoted κ_{ji}, is equal to 1 if j infected i. We define Π as the conditional report ratio of the trees, which differs from the overall report ratio of an outbreak as only unreported cases within transmission chains impact the conditional report ratio. Entirely unreported clusters, or unreported cases infected earlier than the ancestor of a tree do not change the value of Π. By default, the probability of observing κ_{jii} missing generation between i and j from the conditional report ratio $p(\kappa_{ji}|\Pi)$ follows a geometric distribution with mean $(1 - \Pi)/\ \Pi$.

The conditional report ratio is estimated during the MCMC runs using a beta distribution prior. By default, the prior distribution is parametrised as *Beta*(10,1), which is an informative prior of mean 0.9 and standard deviation 0.08.The two parameters of the beta prior can be changed using the variable `prior_pi` in `create_config()`.

Time component: The probability of t_i being the infection date of the case i, given their reported onset date T_i,depends on the distribution of the incubation period f. The incubation period is defined by the variable `f_dens` in the function `outbreaker_data()`.

The probability that *i* was infected by *j* given their respective inferred dates of infection t_i and t_j is defined by the generation time of the disease $w^{\kappa_{ji}}(t_i - t_j)$ (variable `w_dens` in `outbreaker_data()`), and the number of generations κ_{ji} between i and j. The function $w^{\kappa_{ji}}$ was defined as $w^{\kappa_{ji}} = w * w * \ldots * w$, where $*$ is the convolution operator applied κ_{ji} times.

This component of the likelihood follows the framework developed in the Wallinga-Teunis method, and in *outbreaker2*.

Spatial component: The probability of connection between two regions k and l depends on the population sizes m_k and m_l, and the distance between regions d_{kl} . Given spatial parameters a and b, $s(k, l)$ is the probability that a case in the region l was infected by a case reported in k, and is defined using p_{kl}, the connectivity between regions k and l:

$$s(k,l) = \frac{p_{kl}}{\Sigma_h p_{hl}} = \frac{F(d_{kl}, b) * m_k^a * m_l^c}{\Sigma_h (F(d_{hl}, b) * m_h^a * m_l^c)} = \frac{F(d_{kl}, b) * m_k^a}{\Sigma_h (F(d_{hl}, b) * m_h^a)}$$

The package comes with a built-in exponential gravity model: $F(d_{kl}, a) = e^{-b*d_{kl}}$; or a power-law gravity model : $F(d_{kl}, a) = (\frac{1}{d_{kl}})^b$. The exponential gravity model has been shown to be a better representation of short-distance mobility patterns [29]; it is therefore the default option since *o2geosocial* aims at reconstructing transmission in a community or a region. The type of gravity model can be changed by setting the parameter `spatial_method` to "power-law": `create_config(spatial_method = "power_law")`. Other mobility models can be implemented by developing modules. In the use case, we give an example on how to replace the exponential gravity by Stouffer's rank model [30] .

The parameters a and b are estimated during the MCMC run via posterior sampling. This requires re-computing the matrix of spatial connectivity between regions at each iteration and is time-consuming. Therefore, if either a or b is estimated, we allow for a maximum of one missing generation between cases ($\max(\kappa_{ji}) = 2$) and only compute $s^1(k, l)$ and $s^2(k, l)$ for regions that could potentially be connected. By default, the prior distribution of a and b are uniform.

Age component : Given the age group of each case $\alpha_{(1,..,N)}$ and the age-stratified social contact matrix, we introduced $a^{\kappa_{ji}}(\alpha_i, \alpha_j)$, the probability that a case aged α_j infected a case aged α_i. This corresponds to the proportion of contacts to α_i that came from individuals of age α_j . Social contact matrices provided by large scale quantitative investigations such as the POLYMOD study quantify the probability of contact between infectors and infectees of different age groups [31] , and are imported using the R package *socialmixr* [32] . The contact matrix used in the MCMC run is defined by the variable a_dens in outbreaker_data().

Overall likelihood : The overall likelihood that a case i was infected by the case j is equal to $L_i(t_i, j, t_j, \theta) = log(f(t_i - T_i)) + L_{ji}(t_i, t_j, \theta)$ where θ is the parameter set, $f(t_i - T_i)$ is the likelihood that a case with an onset date T_i was infected on t_i , and $L_{ji}(t_i, t_j, \theta)$ is the log-likelihood of connection between i and j defined as:

$$L_{ji}(t_i, t_j, \theta) = log(p(\kappa_{ji}|\Pi) * w^{(\kappa_{ji})}(t_i - t_j) * a^{(\kappa_{ji})}(\alpha_i, \alpha_j) * G(g_i, g_{\tau_j}) * s^{(\kappa_{ji})}(r_i, r_j|a, b))$$

2.3.4. Tree proposals

At every iteration of the MCMC, a set of movements is used to propose an update of the transmission trees. This update is then accepted or rejected depending on the posterior density of the proposed trees. By default, eight movements are tested at each iteration. Three of them were already part of *outbreaker2* and were not modified (`cpp_move_t_inf()` changes the infection date of the cases; `cpp_move_pi()` changes the conditional report ratio; `cpp_move_kappa()` changes the number

of generations between cases). Two movements were edited to scan each transmission tree in order to prevent different genotypes from being in the same tree: (`cpp_move_alpha()` changes the infector; `cpp_move_swap_cases()` swaps infector and infectee). The remaining three are new movements:

- `cpp_move_a()` and `cpp_move_b()` change the spatial parameters *a* and *b* and update the probability of connection between regions.
- `cpp_move_ancestor()` changes the ancestor of the tree. An ancestor is defined as the first case of a transmission tree. For each ancestor i, an index case is drawn from the pool of potential infectors, while another link is randomly picked and deleted.

2.4. Use case

2.4.1. Description of the simulated data

Two simulated datasets are included in *o2geosocial*: `toy_outbreak_short` and `toy_outbreak_long`. Both are lists describing simulated outbreaks and include three elements: i) `cases`: a *data.table* with the ID, location, onset date, genotype, age group, import status, cluster, generation and infector of each case; ii) `dt_regions`: a data table with the ID, population, longitude and latitude of each region; iii) `age_contact`: a numeric matrix of the proportion of contact between age groups. Both simulations were run using distributions of the generation time and the latent period typically associated with measles outbreaks: the incubation period followed a gamma distribution of mean 11.5 days (standard deviation 2.24 days) [33] ; the generation time followed a normal distribution truncated at 0 of mean 11.7 days (standard deviation 2.0 days) [34] .

In order to assess whether the method was able to reconstruct the transmission links between cases, we needed to simulate the transmission trees. Population-level compartmental models cannot be used to generate who-infected-whom. Therefore, we generated the dataset at an individual level, by simulating different transmission trees in the area of interest. The transmission trees were generated using the following process:

1. We created an imported case, with random onset date, region of origin, and age group.
2. We drew the number of secondary cases stemming from this case.
3. If the number of secondary cases was greater than 0, the characteristics of the new cases were drawn using the distributions of the generation time, incubation periods, the spatial kernel, and the proportion of contacts between age groups.
4. We repeated steps 2 and 3, for each new case, until no more secondary cases were drawn (i.e. the random reproduction number drawn in step 2 was 0 for all new case).

5. We repeated steps 1 to 4, until we reached a maximum number of cases, or maximum number of trees, defined before running the simulation.

Numerous factors influencing the transmission dynamics are not included in this simulation framework. However, we do not aim to generate transmission trees which describe the spread of a given pathogen (here measles) in a community with complete accuracy. The main aim of this simulated dataset is to highlight the inference capabilities of the reconstruction method, and to explore causes for discrepancies between the simulations and the model fits, in an ideal setting where all parameters are known and are accounted for in the model.

In this use case, we analyse `toy_outbreak_short`. The dataset contains 75 simulated cases from different census tracts of Ohio in 2014 (variable `cens_tract`). The census tracts represent areas established by the Bureau of Census for analysing populations and generally contain between 2,500 to 8,000 inhabitants. The variable `cluster` describes the transmission tree each case belongs to, and `generation` is equal to the number of generations between the first case of the tree (generation = 1) and the case.

We reconstruct the cluster size distribution of the simulated outbreaks using different models. We then evaluate the agreement between the inferred and the reference transmission clusters in each model, and compare the results obtained with each model. Finally, we assess the geographical heterogeneity of the reconstructed transmission dynamics. We use the package *data.table* for handling data throughout as it is optimised to deal with large datasets [35] . The methods defined in *o2geosocial* would work similarly if we had used the *data.frame* syntax and format.

```
library(o2geosocial)
## We used the data.table syntax throughout this example
library(data.table)
data("toy_outbreak_short")
# Show the first five rows
print(toy_outbreak_short$cases[1:5,])
```

```
##     ID State       Date Genotype  Cens_tract age_group import cluster
## 1: 112  Ohio 2014-01-01       B3 39005970100         6   TRUE      16
## 2:  75  Ohio 2014-01-06       D8 39139002400        11   TRUE      14
## 3: 116  Ohio 2014-01-12       B3 39101000400        11   TRUE      17
## 4: 113  Ohio 2014-01-13       B3 39005970100         6  FALSE      16
## 5: 145  Ohio 2014-01-13       D8 39117965300         8   TRUE      26
##    generation infector_ID
## 1:          1        <NA>
## 2:          1        <NA>
## 3:          1        <NA>
## 4:          2         112
## 5:          1        <NA>
```

```
# Extract dataset
dt_cases <- toy_outbreak_short[["cases"]]
```

In the simulated data, 95% of the clusters contain less than five cases, 47.6% of the clusters are isolated (also called singletons). One larger cluster includes 31 cases (Figure 2.2).

```
# Reference cluster size distribution
hist(table(dt_cases$cluster), breaks = 0:max(table(dt_cases$cluster)),
     ylab = "Number of clusters", xlab = "Cluster size", main = "",
     las = 1)
```

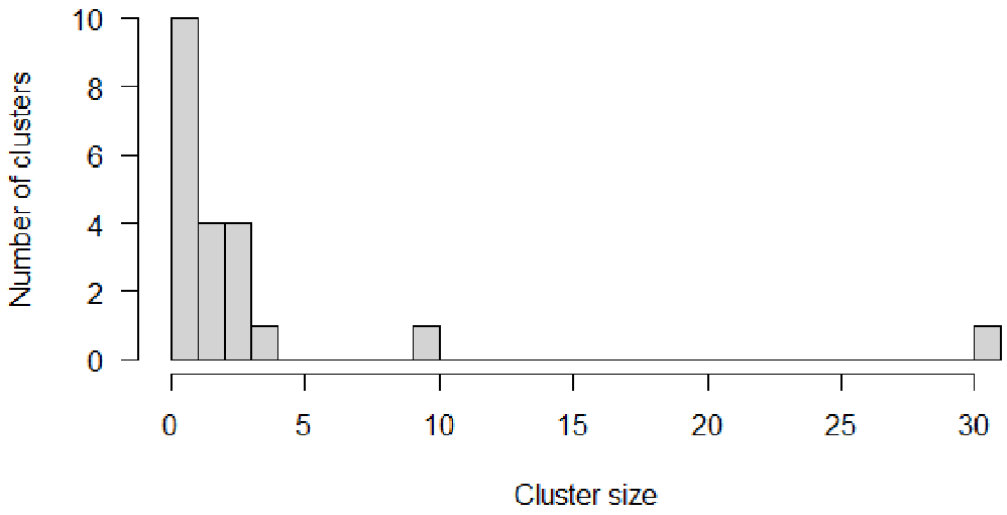

Figure 2.2: Cluster size distribution of the simulated dataset.

2.4.2. Set up and run the models with `outbreaker()`

We set up the distributions the model will use to reconstruct the transmission trees. We define `f_dens` as the duration of the latent period, and `w_dens` as the generation time. These distributions have previously been described for measles outbreaks [33,34,36,37]. In this example, the same distributions were used to generate the simulated data and fit the model. In real-life, there can be discrepancies between the actual distributions and their theoretical estimates. Therefore, we also fitted the model using different distributions of the latent period and generation time, and explored the impact it had on the accuracy of the inferred transmission trees (See Extended Data).

```
# Distribution of the latent period
f_dens <- dgamma(x = 1:100, scale = 0.43, shape = 26.83)
# Distribution of the generation time
w_dens <- dnorm(x = 1:100, mean = 11.7, sd = 2.0)
```

The age specific social contact patterns can be imported from the element `age_contact` of the list `toy_outbreak_short`. Alternatively, one can use the R package *socialmixr* to import a social contact matrix from the POLYMOD survey [32] .

```
# From the list toy_outbreak_short
a_dens <- toy_outbreak_short$age_contact
# Alternatively, from POLYMOD:
polymod_matrix <-
  t(socialmixr::contact_matrix(socialmixr::polymod,
                               countries = "United Kingdom",
                               age.limits = seq(0, 70, by = 5))$matrix)
polymod_matrix <- data.table::as.data.table(polymod_matrix)
# Compute the proportion of connection to each age group
a_dens <- t(t(polymod_matrix)/colSums(polymod_matrix))
```

Finally, the distance matrix between regions is set up from the data table `dt_regions`, element of `toy_outbreak_short`. We use the column population to set up the population vector `pop_vect`. We compute the distance between each region into the distance matrix `dist_mat` using the package *geosphere* [38] .

```
# Extract all regions in the territory
dt_regions <- toy_outbreak_short[["dt_regions"]]
# Extract the population vector
pop_vect <- dt_regions$population
# Create the matrices of coordinates for each region (one "from"; one "to")
mat_dist_from <- matrix(c(rep(dt_regions$long, nrow(dt_regions)),
                          rep(dt_regions$lat, nrow(dt_regions))), ncol = 2)
mat_dist_to <- matrix(c(rep(dt_regions$long, each = nrow(dt_regions)),
                        rep(dt_regions$lat, each = nrow(dt_regions))),
                      ncol = 2)
# Compute all the distances between the two matrices
all_dist <- geosphere::distGeo(mat_dist_from, mat_dist_to)
# Compile into a distance matrix
dist_mat <- matrix(all_dist/1000, nrow = nrow(dt_regions))
# Rename the matrix columns and rows, and the population vector
names(pop_vect) <- rownames(dist_mat) <- colnames(dist_mat) <-
  dt_regions$region
```

We create the lists data, config, moves, likelihoods and priors to run the main function of the package. In this example, we use the default parameters to build moves, likelihoods and priors. The list data contains the distributions `f_dens` and `w_dens`, the population vector and the distance matrix, along with the onset dates, age groups, locations and genotypes of the cases.

Routinely collected surveillance data can include information on the importation status of the cases. In order to investigate the impact of using prior information on the importation status of the cases on cluster reconstruction, we implement two different models: in `out1` the import status is inferred by the model, whereas in `out2` it is set as an input parameter of the model, which only estimates who infected whom.

The first short run in `out1` is run with 10,000 iterations to find the minimum number of importations, and the main run lasts for 20,000 iterations in both models. As the import status of the cases is inferred

in `out1`, we have to set a threshold to quantify what is an unlikely likelihood of transmission between cases. We use a relative outlier threshold at 0.9, which means that the threshold will be the 9^{th} decile of the negative log-likelihoods $L_{ji}(t_i, t_j, \theta)$ in every sample.

```
# Set movement, likelihood and prior lists to default
moves <- custom_moves()
likelihoods <- custom_likelihoods()
priors <- custom_priors()
# Data and config, model 1
data1 <- outbreaker_data(dates = dt_cases$Date, #Onset dates
                         age_group = dt_cases$age_group, #Age group
                         region = dt_cases$Cens_tract, #Location
                         genotype = dt_cases$Genotype, #Genotype
                         w_dens = w_dens, #Generation time
                         f_dens = f_dens, #Latent period
                         a_dens = a_dens, #Age stratified contact matrix
                         population = pop_vect, #Population
                         distance = dist_mat #Distance matrix
)
config1 <- create_config(data = data1,
                         n_iter = 20000, #Iteration number: main run
                         n_iter_import = 10000, #Iteration nb: short run
                         burnin = 5000, #burnin period: first run
                         outlier_relative = T, #Absolute/relative threshold
                         outlier_threshold = 0.9 #Value of the threshold
)
# Run model 1
out1 <- outbreaker(data = data1, config = config1, moves = moves,
                   priors = priors, likelihoods = likelihoods)
# Set data and config for model 2
data2 <- outbreaker_data(dates = dt_cases$Date,
                         age_group = dt_cases$age_group,
                         region = dt_cases$Cens_tract,
                         genotype = dt_cases$Genotype, w_dens = w_dens,
                         f_dens = f_dens, a_dens = a_dens,
                         population = pop_vect, distance = dist_mat,
                         import = dt_cases$import #Importation status
)
config2 <- create_config(data = data2,
                         find_import = FALSE, # Not inferring import status
                         n_iter = 20000,
                         sample_every = 50, # 1 in 50 iterations is kept
                         burnin = 5000)
# Run model 2
out2 <- outbreaker(data = data2, config = config2, moves = moves,
                   priors = priors, likelihoods = likelihoods)
```

The data frames `out1` and `out2` contain the posterior density, likelihood, and prior density of the trees generated at every iteration, along with the values of the spatial parameters `a` and `b`, the conditional report ratio `pi`, and the index, estimated infection date and number of generations for each case.

2.4.3. Compare inferred and reference clusters

The function `summary` prints a summary of the data frame generated by `outbreaker()`. It contains a list with the number of steps, the distributions of the posterior, likelihood and priors, the parameter distributions, the most likely infector and the probability of being an import for each case, and the cluster size distribution.

```
# Summary parameters a and b, removing the burnin-period
#Model 1
print(summary(out1, burnin = 5000)$a)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2144  0.5733  0.8546  0.8497  1.1015  1.4955

print(summary(out1, burnin = 5000)$b)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07172 0.09180 0.09679 0.09835 0.10494 0.12839

# Model 2
print(summary(out2, burnin = 5000)$a)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2248  0.6809  0.9625  0.9359  1.1948  1.4971

print(summary(out2, burnin = 5000)$b)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.08681 0.11978 0.12930 0.13040 0.13973 0.17477
```

In order to compare the reconstructed clusters to the data in each model, we plot the median inferred cluster size distribution in `out1` and `out2` and the credible intervals. First, we group together clusters of similar sizes by defining the breaks of each group in the vector `group_cluster`. In this example, we defined the size categories as 1; 2; 3 – 4; 5 – 9; 10 – 15; 15 – 40 and 40 + cases. The inferred cluster size distributions are shown in the element cluster from the output of `summary(out1)`, and are aggregated using the input variable `group_cluster`.

```
# We create groups of cluster size: initialise the breaks for each group
group_cluster <- c(1, 2, 3, 5, 10, 15, 40, 100) - 1
# Reference data: h$counts
h <- hist(table(dt_cases$cluster), breaks = group_cluster, plot = FALSE)

# Grouped cluster size distribution in each run
clust_infer1 <- summary(out1, group_cluster = group_cluster,
                        burnin = 5000)$cluster
clust_infer2 <- summary(out2, group_cluster = group_cluster,
                        burnin = 5000)$cluster
# Merge inferred and reference cluster size distributions into one matrix
clust_size_matrix <- rbind(clust_infer1["Median",],
                           clust_infer2["Median",],
                           h$counts)
```

The number of isolated cases in the inferred trees in `out1` is lower than in the data (Figure 2.3). We can therefore conclude that when the import status of the cases was inferred, the model underestimated the number of clusters and tended to link together unrelated cases. The cluster size distribution when the import status of the cases is inferred depends on the likelihood threshold set in `outlier_threshold` and `outlier_relative`. Using different values of λ would impact the cluster size distribution in out1. Conversely, the cluster size distribution in `out2` is very similar to the data (Figure 2.3).

```
# Histogram of the inferred and reference cluster size distributions
b <- barplot(clust_size_matrix, names.arg = colnames(clust_infer1), las=1,
             ylab = "Number of clusters", xlab = "Cluster size", main = "",
             beside = T, ylim = c(0, max(c(clust_infer1, clust_infer2))))
# Add the 50% Credible Intervals
arrows(b[1,], clust_infer1["1st Qu.",], b[1,], clust_infer1["3rd Qu.",],
       angle = 90, code = 3, length = 0.1)
arrows(b[2,], clust_infer2["1st Qu.",], b[2,], clust_infer2["3rd Qu.",],
       angle = 90, code = 3, length = 0.1)
# Add legend
legend("topright", fill = grey.colors(3), bty = "n",
       legend = c("Inferred import status",
                  "Known import status", "Simulated dataset"))
```

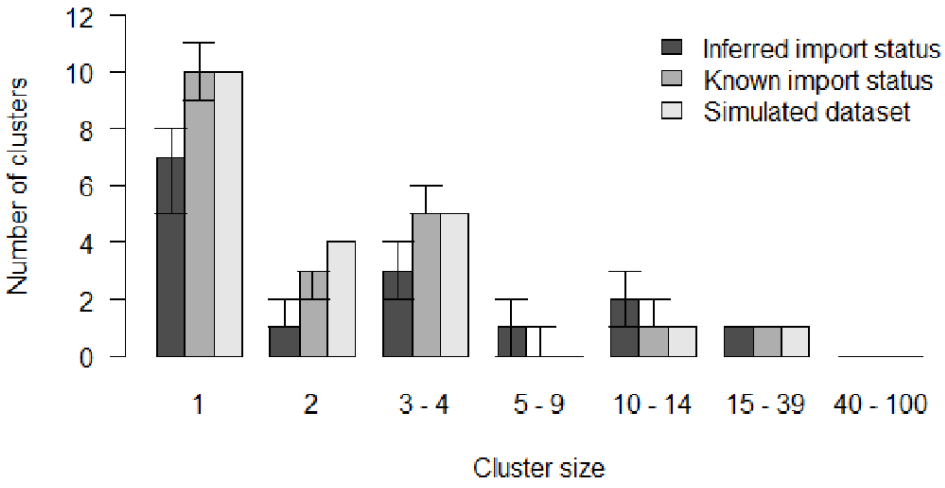

Figure 2.3: Comparison of inferred cluster size distribution in both models with the reference data.

We investigate the reconstructed transmission trees to ensure the index assigned to each case is in agreement with the reference dataset. To do so, we write two functions: in `index_infer` we compute the proportion of iterations where the inferred index of each case matches their actual index (perfect match); in `index_clust` we compute the proportion of iterations where the inferred index is from the same reference cluster as the actual index (close match).

```
#' Title: Compute the proportion of iterations in the outbreaker() output
#` where the inferred index matches the actual index in dt_cases
#'
```

```r
#' @param dt_cases: reference dataset
#' @param out: Matrix output of outbreaker()
#' @param burnin: Numeric, length of the burnin phase
#'
#' @return Numeric vector showing the proportion of iterations pointing to
#' the correct index case
index_infer <- function(dt_cases, out, burnin){
  ## Generate the data frame listing every infector:
  # Select rows above burnin, and columns describing who infected whom
  out_index <- out[out$step > burnin, grep("alpha", colnames(out))]
  # ID of each infector
  ID_index <- matrix(dt_cases[unlist(out_index), ID],
                     ncol = nrow(dt_cases))
  # Match inferred (ID_index) and actual infector (column infector_ID)
  match_infer_data <- t(ID_index) == dt_cases$infector_ID
  # If a case is rightly inferred as an ancestor, set match to TRUE
  match_infer_data[is.na(t(ID_index)) & is.na(dt_cases$infector_ID)] <- T
  prop_correct <- rowSums(match_infer_data,
                          na.rm = T)/ncol(match_infer_data)

  return(prop_correct)
}
# Same as index_infer, except it returns the proportion of inferred indexes
# who are in the same reference cluster as the case
index_clust <- function(dt_cases, out, burnin){
  ## Generate the data frame listing every infector:
  # Select rows above burnin, and columns describing who infected whom
  out_index <- out[out$step > burnin, grep("alpha", colnames(out))]
  # cluster of each infector
  clust_index <- matrix(dt_cases[unlist(out_index), cluster],
                        ncol = nrow(dt_cases))
  # Match inferred (cluster_index) and actual cluster (column cluster)
  match_infer_data <- t(clust_index) == dt_cases$cluster
  # Exclude ancestors
  match_infer_data <- match_infer_data[!is.na(dt_cases$infector_ID),]


  prop_correct <- rowSums(match_infer_data,
                          na.rm = T)/ncol(match_infer_data)

  return(prop_correct)
}
# Run index_infer for each model
index_infer1 <- index_infer(dt_cases = dt_cases, out = out1, burnin = 5000)
index_infer2 <- index_infer(dt_cases = dt_cases, out = out2, burnin = 5000)
# Run index_clust for each model
index_clust1 <- index_clust(dt_cases = dt_cases, out = out1, burnin = 5000)
index_clust2 <- index_clust(dt_cases = dt_cases, out = out2, burnin = 5000)
```

Figure 2.4 shows that the proportion of perfect and close match for most cases is lower in `out1`, which indicates that inferring the import status reduced the accuracy of the inference. Using previous

investigations into the travel history of cases is key to improve the reconstruction of transmission history.

```
# Plot the sorted proportion in each model
par(bty = "n", mfrow = c(1, 2), mar = c(5,4,2,0), oma = c(0, 0, 0, 0))
# Panel A: Perfect match
plot(sort(index_infer1), type = "l", ylab = "Proportion of iterations",
     xlab = "Case", main =  "A", las = 1, col = grey.colors(3)[1], lwd = 3,
     ylim = c(0,1))
lines(sort(index_infer2), col = grey.colors(3)[2], lwd = 3)

# Panel B: Close match
plot(sort(index_clust1), type = "l", xlab = "Case", ylab = "",
     main =  "B", las = 1, col = grey.colors(3)[1], lwd = 3, ylim = c(0,1))
lines(sort(index_clust2), col = grey.colors(3)[2], lwd = 3)
legend("bottomright", col = grey.colors(3)[1:2], lwd = 3, bty = "n",
       legend = c("Inferred import status","Known import status"))
```

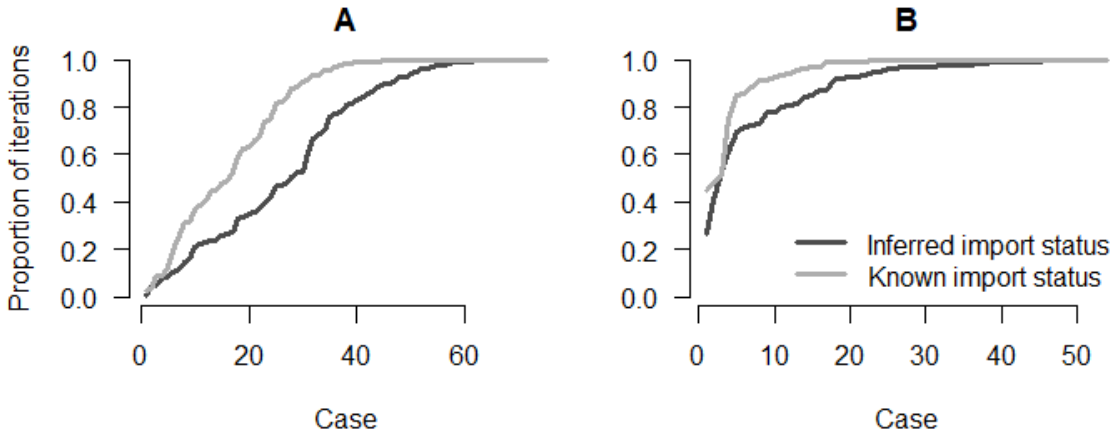

Figure 2.4: Panel A: Proportion of iterations with the correct index for each case; Panel B: Proportion of iterations where the index is from the correct cluster.

We now investigate the geographical distribution of the importations, and the average number of secondary cases per region in `out1` and `out2`. The maps are generated using the package *ggplot2* [39].

First, we retrieve the boundary files of the census tracts in Ohio to generate the background of the maps using the package *tigris* [40] . We import them in a format compatible with the package *sf* and create one background map for each model.

```
library(ggplot2)
# Read the shapefile and create one map for each model
map1 <- tigris::tracts(state = "Ohio", class = "sf", progress_bar = FALSE)
map1$INTPTLON <- as.numeric(map1$INTPTLON)
map1$INTPTLAT <- as.numeric(map1$INTPTLAT)
map2 <- map1
map1$model <- "Model 1"
map2$model <- "Model 2"
```

We are interested in two outputs of the models: i) the number of imports per region, in order to highlight regions where importations of cases are most likely, and ii) the geographical distribution of the number of secondary cases per case, which gives insight into the areas most vulnerable to the spread of the disease.

Number of imports per region: The element `tree` of `summary(out1)` contains the most likely infector, the proportion of iterations where the index is the most likely infector and the median number of generations between the two cases, the most likely infection date and the chances of being an import for each case. We add two columns to `dt_cases` showing the probablity of being an import in `out1` and `out2` for each case. As the import status is not inferred in `out2`, `prop_import2` is a binary value, and is equal to `dt_cases$import`.

```
# Add the proportion of iterations in model 1 where each case is an import
dt_cases[, prop_import1 := summary(out1, burnin = 5000)$tree$import]
# Add the proportion of iterations in model 2 where each case is an import
dt_cases[, prop_import2 := summary(out2, burnin = 5000)$tree$import]
```

We generate the number of imports per region in each model (vectors `prop_reg1` and `prop_reg2`) and add it to the matrices describing the maps.

```
# Number of imports per region in model 1
prop_reg1 <- dt_cases[, .(prop_per_reg = sum(prop_import1)),
                      by = Cens_tract]$prop_per_reg
# Number of imports per region in model 2
prop_reg2 <- dt_cases[, .(prop_per_reg = sum(prop_import2 )),
                      by = Cens_tract]$prop_per_reg
names(prop_reg1) <- names(prop_reg2) <- unique(dt_cases$Cens_tract)

# Add the number of imports in each region to the maps
map1$prop_reg <- prop_reg1[as.character(map1$GEOID)]
map2$prop_reg <- prop_reg2[as.character(map2$GEOID)]
```

We plot the number of imports per region in each model (Figure 2.5). The right panel (`out2`) shows the geographical distribution of importations in the data. We observe discrepancies between the two panels. In `out1`, the inferred number of importations in the central areas is much lower than in the reference data. These maps highlight the uncertainty added when the import status of each case is inferred.

```
# Merge maps
maps <- rbind(map1, map2)
# Crop map to area of interest
lim_lon <- c(-84, -82)
lim_lat <- c(40, 41.5)
maps <- maps[maps$INTPTLON > lim_lon[1] & maps$INTPTLON < lim_lon[2] &
               maps$INTPTLAT > lim_lat[1] & maps$INTPTLAT < lim_lat[2],]
```

```r
# Plot: number of imports per region, two panels
ggplot(maps) +  geom_sf(aes(fill = prop_reg)) + facet_grid(~model)  +
  scale_fill_gradient2(na.value = "lightgrey", midpoint = 0.8,
                       breaks = c(0, 0.5, 1, 1.5), name = "Nb imports",
                       low = "white", mid = "lightblue",
                       high = "darkblue") +
  coord_sf(xlim = c(-83.8, -82.2), ylim = c(40.2, 41.3)) +
  theme_classic(base_size = 9) +
  theme(axis.text = element_blank(), axis.ticks = element_blank(),
        axis.line = element_blank())
```

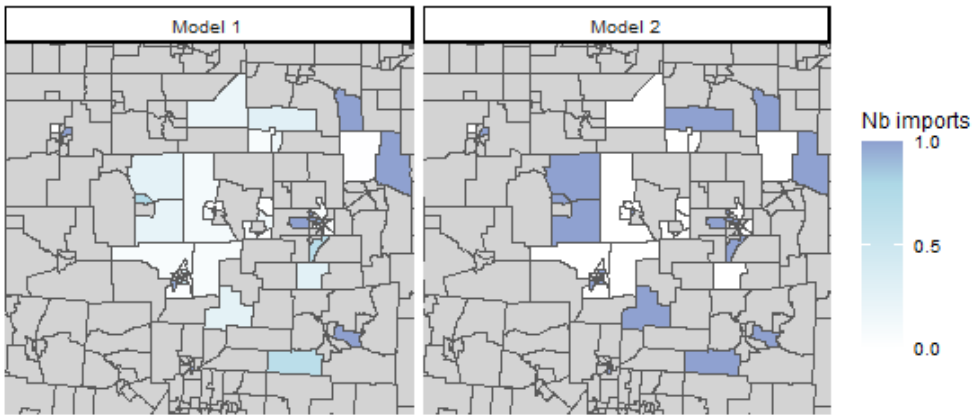

Figure 2.5: Average number of imported cases per census tract, regions where no case was reported are shown in grey.

Average number of secondary cases per region: In this section, we map the number of secondary cases per case in each region to identify which regions were associated with higher levels of transmission. We define the function `n_sec_per_reg` to compute the average number of secondary cases per case and aggregate it per region. We then extract the median number of secondary cases per case in each region.

```r
#' Title: Compute the number of secondary cases per case in each region
#'
#' @param dt_cases: reference dataset
#' @param out: Matrix output of outbreaker()
#' @param burnin: Numeric, length of the burnin phase
#'
#' @return A numeric matrix: the first column is the census tract ID, the
#' other columns show the number of secondary cases per case. Each row
#' corresponds to a different iteration.
n_sec_per_reg <- function(dt_cases, out, burnin){
  ## Number of secondary cases per case
  n_sec <- apply(out[out$step > burnin, grep("alpha", colnames(out))], 1,
                 function(X){
                   X <- factor(X, 1:length(X))
                   return(table(X))})
  ## Aggregate by region
  tot_n_sec_reg <- aggregate(n_sec, list(dt_cases$Cens_tract), sum)
  ## Divide by the number of cases in each region
  tot_n_sec_reg <- cbind(tot_n_sec_reg[, 1],
```

```
                          tot_n_sec_reg[, -1] / table(dt_cases$Cens_tract))
  return(tot_n_sec_reg)
}
## Generate the number of secondary cases per case in each region
n_sec_tot1 <- n_sec_per_reg(dt_cases = dt_cases, out = out1, burnin = 5000)
n_sec_tot2 <- n_sec_per_reg(dt_cases = dt_cases, out = out2, burnin = 5000)
```

We now plot the geographical distribution of the median number of secondary cases in each region according to the models, and compare it with the simulations (Figure 2.6). Despite minor discrepancies, the maps generated by the two models are similar. Both show an important spatial heterogeneity. The eastern and central areas are associated with higher numbers of secondary cases. If we change the vectors `n_sec1` and `n_sec2` to plot different deciles, we show the dispersion of the number of secondary cases in the different iterations of the models. Similarly, we observe minor differences between the maps generated by the models and the simulated data. Most of the regions that repeatedly caused further transmissions in the simulations are identified by the models. In the Extended Data, we compared the regional number of secondary transmissions in the simulated data to the 95% Credible Intervals of both models, and found that the models were able to capture the input values in each region.

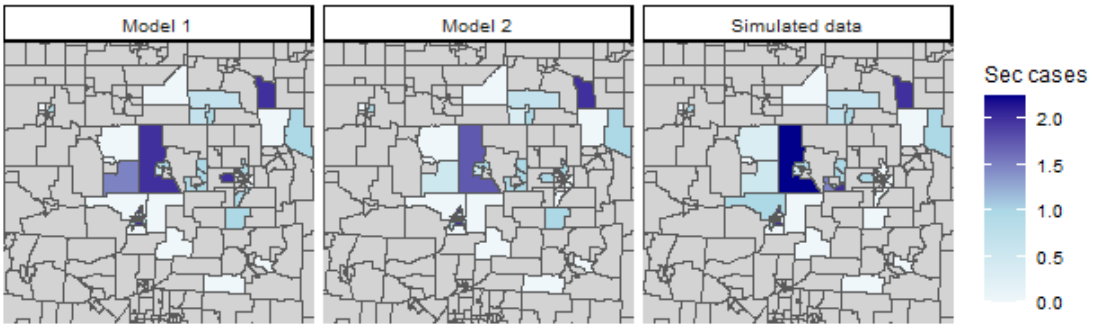

Figure 2.6: Median number of secondary transmission per case in each census tract.

```
   # Merge maps
maps_n_sec <- rbind(map1, map2, map_data)
# Crop map to area of interest
lim_lon <- c(-84, -82)
lim_lat <- c(40, 41.5)
maps_n_sec <- maps_n_sec[maps_n_sec$INTPTLON > lim_lon[1] &
                         maps_n_sec$INTPTLON < lim_lon[2] &
                         maps_n_sec$INTPTLAT > lim_lat[1] &
                         maps_n_sec$INTPTLAT < lim_lat[2],]

# Plot the geographical distribution of the number of secondary cases
ggplot(maps_n_sec) +  geom_sf(aes(fill = n_sec)) + facet_grid(~model)  +
  scale_fill_gradient2(na.value = "lightgrey", mid = "lightblue",
                       low = "white", midpoint = 1, high = "darkblue",
```

```
                      breaks = seq(0, 5, 0.5),name = "Sec cases") +
  coord_sf(xlim = c(-83.8, -82.2), ylim = c(40.2, 41.3)) +
  theme_classic(base_size = 9) +
  theme(axis.text = element_blank(), axis.ticks = element_blank(),
        axis.line = element_blank())
```

2.4.4. Customise the likelihood, prior and movement lists: the Stouffer's rank model

In the previous example, we ran and evaluated two different models to reconstruct transmission clusters from simulated surveillance data, and highlighted the spatial heterogeneity of measles transmission in the region. These models were run using the default likelihood, prior and movement functions. Now we develop a third model, where the spatial connection between regions is based on the Stouffer's rank method [30].

In the Stouffer's rank method, the absolute distance is not used to compute the probability of connection between regions. The connectivity between the regions k and l only depends on the summed population of all the regions closer to l than k. If we define this collection of regions $\Omega_{k,l} = \{i: 0 \leq d(i,l) \leq d(k,l)\}$, Stouffer's distance is then $p_{kl} = m_l^c * \left(\frac{m_k}{\sum_{i\in\Omega_{k,l}} m_i}\right)^a$. From this, we deduce the probability that a case from region l was infected by a case from region k.

$$s(k,l) = \frac{p_{kl}}{\Sigma_h p_{hl}} = \frac{\left(\frac{m_k}{\sum_{i\in\Omega_{k,l}} m_i}\right)^a}{\Sigma_h \left(\frac{m_h}{\sum_{i\in\Omega_{h,l}} m_i}\right)^a}$$

This model is similar to the power-law gravity model with two main differences: i) each cell of the distance matrix should be equal to $\sum_{i\in\Omega_{k,l}} m_i$, and ii) only one spatial parameter a is estimated. First, we create the distance matrix associated with Stouffer's rank:

```
# For every column of the distance matrix, use the cumulative sum of the
# population vector ordered by the distance. Remove the values where
# the distance between the regions is above gamma
dist_mat_stouffer <- apply(dist_mat, 2, function(X){
  pop_X <- cumsum(pop_vect[order(X)])
  omega_X <- pop_X[names(X)]
  # omega_X is set to -1 if the distance between two regions is above gamma
  omega_X[X > config1$gamma] <- -1
  return(omega_X)
})
# The new value of gamma is equal to the maximum of dist_mat_stouffer + 1
gamma <- max(dist_mat_stouffer) + 1
# The values previously set to -1 are now set to the new value of gamma
dist_mat_stouffer[dist_mat_stouffer == -1] <- max(dist_mat_stouffer) * 2
```

Secondly, since the connectivity matrix in the Stouffer's rank model is only computed from one spatial parameter, we write a new movement function `cpp_stouffer` to estimate it. The formula of the Stouffer's rank connectivity matrix is similar to the power law gravity models. Therefore, `cpp_stouffer` is similar to the default movement `cpp_move_a`, and uses the same function to compute the probability matrix (`cpp_log_like()`). This function is written with the package *Rcpp*, and is sourced using the function `Rcpp::sourceCpp` [25].

```
// [[Rcpp::depends(o2geosocial)]]
#include <Rcpp.h>
#include <Rmath.h>
#include <o2geosocial.h>
// This function is used to estimate new values of the spatial parameter.
// It is based on the structure as cpp_move_a in o2geosocial,
// [[Rcpp::export()]]
Rcpp::List cpp_stouffer(Rcpp::List param, Rcpp::List data,
                        Rcpp::List config,
                        Rcpp::RObject custom_ll,
                        Rcpp::RObject custom_prior){
  // Import parameters
  Rcpp::List new_param = clone(param);
  double gamma = config["gamma"];
  int max_kappa = config["max_kappa"];
  Rcpp::List new_log_s_dens = new_param["log_s_dens"];
  Rcpp::NumericMatrix dist = data["distance"], probs = new_log_s_dens[0];
  Rcpp::NumericMatrix ances = data["can_be_ances_reg"];
  Rcpp::NumericVector pop = data["population"], limits = config["prior_a"];
  // Size of the probability matrix
  int nb_cases = pow(probs.size(), 0.5);
  // Draw new value of a
  Rcpp::NumericVector new_a = new_param["a"];
  double sd_a = static_cast<double>(config["sd_a"]);
  double old_logpost = 0.0, new_logpost = 0.0, p_accept = 0.0;
  // proposal (normal distribution with SD: config$sd_a)
  new_a[0] += R::rnorm(0.0, sd_a); // new proposed value
  if (new_a[0] < limits[0] || new_a[0] > limits[1]) return param;
  // Generate new probability matrix
  new_param["log_s_dens"] =
    o2geosocial::cpp_log_like(pop, dist, ances, new_a[0], new_a[0],
                              max_kappa, gamma, "power-law", nb_cases);
  // Compare old and new likelihood values
  old_logpost = o2geosocial::cpp_ll_space(data, config, param,
                                          R_NilValue, custom_ll);
  new_logpost = o2geosocial::cpp_ll_space(data, config, new_param,
                                          R_NilValue, custom_ll);
  // Add prior values
  old_logpost += o2geosocial::cpp_prior_a(param, config, custom_prior);
  new_logpost += o2geosocial::cpp_prior_a(new_param, config, custom_prior);
  // Accept or reject proposal
  p_accept = exp(new_logpost - old_logpost);
  if (p_accept < unif_rand()) return param;
```

```
  return new_param;
}
```

We modify the element a of the list of movements used in the last model. We set up the lists data and config using `dist_mat_stouffer` as the distance matrix. Since there is only one spatial parameter in this model, we set the parameter `move_b` to `FALSE` in `create_config()`, and we set the prior of b to the null function `f_null`.

```
# Edit the lists of movements and priors
moves3 <- custom_moves(a = cpp_stouffer)
# Define null function
f_null <- function(param) {
  return(0.0)
}
priors3 <- custom_priors(b = f_null)


# Set data and config lists
data3 <- outbreaker_data(dates = dt_cases$Date, #Onset dates
                         age_group = dt_cases$age_group, #Age group
                         region = dt_cases$Cens_tract, #Location
                         genotype = dt_cases$Genotype, #Genotype
                         w_dens = w_dens, #Generation time
                         f_dens = f_dens, #Latent period
                         a_dens = a_dens, #Age stratified contact matrix
                         population = pop_vect, #Population
                         distance = dist_mat_stouffer #Distance matrix
)
config3 <- create_config(data = data3,
                         gamma = gamma,
                         init_b = 0, move_b = FALSE, # b is not estimated
                         n_iter = 20000, #Iteration number: main run
                         n_iter_import = 10000, #Iteration nb: short run
                         burnin = 5000, #burnin period: first run
                         outlier_relative = T, #Absolute/relative threshold
                         outlier_threshold = 0.9 #Value of the threshold
)
# Run the model using the Stouffer's rank method
out_stouffer <- outbreaker(data = data3, config = config3, moves = moves3,
                           priors = priors3, likelihoods = likelihoods)
```

We plot the inferred cluster size distribution and compare it to the reference data (Figure 2.7). We observe discrepancies between the inferred distribution and the data: the model over-estimates the number of clusters containing more than 15 cases and underestimates the number of small clusters and isolated individuals.

```
# Grouped cluster size distribution in the Stouffer's rank model
clust_infer_stouf <- summary(out_stouffer, burnin = 5000,
                             group_cluster = group_cluster)$cluster
# Merge inferred and reference cluster size distributions
clust_size_matrix <- rbind(clust_infer_stouf["Median",], h$counts)
```

```
# Plot the two distributions
b <- barplot(clust_size_matrix, names.arg = colnames(clust_infer_stouf),
             beside = T, ylab = "Number of clusters", xlab = "Cluster size",
             main = "", las = 1)
# Add CIs
arrows(b[1,], clust_infer_stouf["1st Qu.",], b[1,],
       clust_infer_stouf["3rd Qu.",], angle = 90, code = 3, length = 0.1)
legend("topright", fill = grey.colors(2), bty = "n",
       legend = c("Inferred import status, Stouffer's rank method",
                  "Simulated dataset"))
```

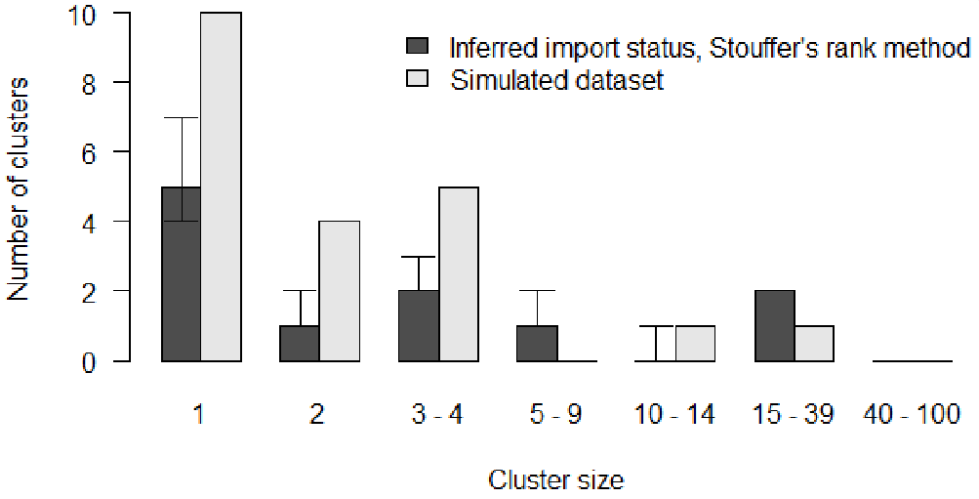

Figure 2.7: Comparison of inferred cluster size distribution with the reference data.

Finally, we plot the proportion of perfect and close matches for each case (Figure 2.8). We observe that the fit obtained with the Stouffer's rank method is consistently worse than the first two models. The Stouffer's rank method did not improve the agreement between the inferred trees and the reference data.

```
# Generate the proportion of perfect and close match for each case in out3
index_infer_stouf <- index_infer(dt_cases = dt_cases, out = out_stouffer,
                                 burnin = 5000)
index_clust_stouf <- index_clust(dt_cases = dt_cases, out = out_stouffer,
                                 burnin = 5000)
# Plot the sorted proportion in each model
par(bty = "n", mfrow = c(1, 2), mar = c(5,4,2,0), oma = c(0, 0, 0, 0))
# Panel A: Perfect match
plot(sort(index_infer_stouf), main = "A", col = grey.colors(2)[1], lwd = 3,
     xlab = "Case", ylab = "Proportion of iterations", type = "l", las = 1,
     ylim = c(0,1))
# Panel B: Close match
plot(sort(index_clust_stouf), type = "l", ylab = "", xlab = "Case",
     main = "B", las = 1, col = grey.colors(2)[1], lwd = 3, ylim = c(0,1))
```

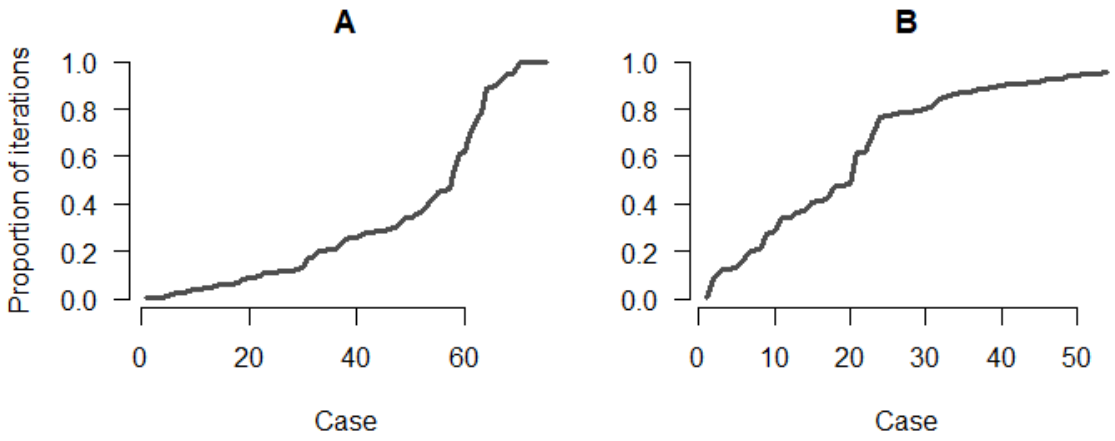

Figure 2.8: Panel A: Proportion of iterations with the correct index for each case; Panel B: Proportion of iterations where the index is from the correct cluster.

The simulated data used in the study were generated using an exponential gravity model, which explains why introducing the Stouffer's rank method did not improve the inference. This is not representative of the performance of each mobility model at reconstructing actual transmission clusters.

In this use case, we only explored customising the spatial component. However, the other components of the likelihoods can also be edited, using the functions `custom_priors()`, `custom_likelihoods()`, or `custom_moves()`. For instance, to account for changes in the distribution of the generation time throughout an outbreak [41], one would have to change the element `timing_infections` of `custom_likelihoods()`. However, the distribution would need to be set prior to running the models.

2.5. Discussion

The R package *o2geosocial* is a new tool for data analysis building upon the framework developed in *outbreaker2*. It uses routinely collected surveillance data to reconstruct transmission networks. It can be used on a broad range of diseases where genetic sequencing is not common, or informative. For instance, it has been applied on national measles surveillance data to reconstruct the cluster size distribution of outbreaks in the United States between 2001 and 2016 [7] . In this study, we presented an application on a simulated dataset using detailed geographic information on the location of cases.

We implemented several models to reconstruct the cluster size distribution of the simulated outbreak. Although each model was able to capture the overall dynamics of transmission, we observed discrepancies between the reference data and the reconstructed cluster size distribution for models where the importation status of the cases was inferred. These discrepancies are linked to the threshold set to define what is considered an unlikely connection. A looser threshold may lead to unrelated cases being connected and a lower number of inferred imports, whereas a stricter threshold increases the

number of short transmission chains. Therefore, the use of epidemiological information describing importation status improves the accuracy of the transmission cluster reconstruction in *o2geosocial*. In case of incomplete epidemiological information, the user can set the importation status for some of the cases, and the others would be inferred. These results highlight that epidemiological investigations are crucial to improve our ability to reconstruct transmission events, particularly when unrelated importations happen concurrently.

The method described in this paper does not account for long-distance transmission, as transmission events are impossible in *o2geosocial* when the distance between regions is above the parameter gamma. In case of long-distance transmission, the infected case would be considered as a new importation. Nevertheless, this limitation is not critical since *o2geosocial* was designed to identify areas most susceptible to local transmission, *i.e.* regions where importations were likely to lead to local outbreaks.

The use of transmission trees and transmission clusters to assess current or future risk of outbreaks comes with various limitations. First, it relies upon the assumption that previous transmission patterns are representative of future outbreaks. Second, it requires past observed transmission events, and does not account for the number of opportunities of transmission per case. Where only sporadic isolated cases have been reported in the country, it is not possible to draw relevant conclusions on communities potentially most vulnerable to transmission. Third, partial detection of cases may bias the cluster size distribution, and under-estimate the number of secondary transmissions per case. Patterns of transmission, and characteristics associated with high-transmission events may still be observable but could introduce a bias if reporting is itself is affected by the same factors as is transmission. Finally, the use of transmission trees for real time modelling can be challenging, given the right-censoring of transmissions caused by recent infectious individuals [42].

The default implementation of the method assumes that the generation times are independent and identically distributed throughout an outbreak, whereas in reality, depletion of susceptibles and competing risk of infection through clustering of contacts would be expected to affect the generation interval. The method can be customised to integrate time varying generation intervals set prior to running the models. However, estimating the distribution of the generation interval during the inference procedure is more challenging to implement in the current framework, which may introduce a bias in our results.

The analyses presented in this paper were run on simulated data, which partly explains the very close match between the inferred and reference cluster size distribution. Indeed, the distributions of the incubation period and generation time used to generate the simulations were the same as the ones

used for cluster inference in the Main Analysis. Using imprecise or inaccurate distributions can lead to biases in the reconstruction of the transmission trees. We re-ran the inference procedure using different distributions (changing the mean or the standard deviation), the results can be seen in the Extended Data. When the distributions were set with lower standard deviations, several links were not observed in the inferred transmission trees anymore. Indeed, these connections had been made impossible since the range of likely values was narrower. In all other examples, the simulated and inferred clusters size distribution remained very close, we only observed a slight drop in the proportion of iterations that contain the right transmission links. Since the likelihood of connection is computed from several components, the discrepancies between the distributions used in the simulations and the model fits did not substantially changed the inferred trees.

We also showed how the model could be edited to implement different mobility models. Describing human mobility during infectious diseases outbreaks is challenging, and the performance of the models depends on the setting [29,43–45]. Future developments in the package will focus on facilitating the integration of new variables in the likelihood of connection, such as workplace or school. Currently, such variables would have to be integrated within one of the components of likelihood. We aim to simplify the addition of new parameters and components in the inference framework. We encourage the development of extensions of *o2geosocial* to study a wide range of pathogens and settings where sequence data are not informative. We hope that wider use of *o2geosocial* can help maximise the information brought by routinely collected data and epidemiological investigations, in order to improve our understanding of outbreak dynamics.

2.6. Reference

[1] Robert A, Edmunds WJ, Watson CH, Henao-Restrepo AM, Gsell P-S, Williamson E, et al. Determinants of Transmission Risk During the Late Stage of the West African Ebola Epidemic. Am J Epidemiol 2019. https://doi.org/10.1093/aje/kwz090.

[2] Faye O, Boëlle P-Y, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect Dis 2015;15:320–6. https://doi.org/10.1016/S1473-3099(14)71075-8.

[3] le Polain de Waroux O, Saliba V, Cottrell S, Young N, Perry M, Bukasa A, et al. Summer music and arts festivals as hot spots for measles transmission: Experience from England and Wales, June to October 2016. Eurosurveillance 2016;21:1–6. https://doi.org/10.2807/1560-7917.ES.2016.21.44.30390.

[4] Leclerc QJ, Fuller NM, Knight LE, Funk S, Knight GM. What settings have been linked to SARS-

CoV-2 transmission clusters? Wellcome Open Res 2020;5. https://doi.org/10.12688/wellcomeopenres.15889.2.

[5] Taube JC, Miller PB, Drake JM. An open-access database of infectious disease transmission trees to explore superspreader epidemiology. MedRxiv 2021:2021.01.11.21249622.

[6] Wang L, Didelot X, Yang J, Wong G, Shi Y, Liu W, et al. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. Nat Commun 2020;11:1–6. https://doi.org/10.1038/s41467-020-18836-4.

[7] Robert A, Kucharski AJ, Gastañaduy PA, Paul P, Funk S. Probabilistic reconstruction of measles transmission clusters from routinely collected surveillance data. J R Soc Interface 2020;17:20200084. https://doi.org/10.1098/rsif.2020.0084.

[8] Gastañaduy PA, Budd J, Fisher N, Redd SB, Fletcher J, Miller J, et al. A Measles Outbreak in an Underimmunized Amish Community in Ohio. N Engl J Med 2016;375:1343–54. https://doi.org/10.1056/NEJMoa1602295.

[9] Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLoS Comput Biol 2013;9:1–17. https://doi.org/10.1371/journal.pcbi.1002993.

[10] Blumberg S, Enanoria WTA, Lloyd-Smith JO, Lietman TM, Porco TC. Identifying postelimination trends for the introduction and transmissibility of measles in the United States. Am J Epidemiol 2014;179:1375–82. https://doi.org/10.1093/aje/kwu068.

[11] Ferguson NM, Donnelly CA, Anderson RM. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. Nature 2001. https://doi.org/10.1038/35097116.

[12] Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal. Am J Epidemiol 2004;160:509–16.

[13] Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Comput Biol 2014;10. https://doi.org/10.1371/journal.pcbi.1003457.

[14] Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: A modular platform for outbreak reconstruction. BMC Bioinformatics 2018;19. https://doi.org/10.1186/s12859-018-2330-z.

[15] Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc R Soc B Biol Sci 2012;279:444–50. https://doi.org/10.1098/rspb.2011.0913.

[16] Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S. A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. PLoS Comput Biol 2012;8. https://doi.org/10.1371/journal.pcbi.1002768.

[17] Kendall M, Ayabina D, Colijn C. Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees 2016:1–22. https://doi.org/10.1214/17-STS637.

[18] Ypma RJF, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 2013;195:1055–62. https://doi.org/10.1534/genetics.113.154856.

[19] Worby CJ, O'Neill PD, Kypraios T, Robotham J V., De Angelis D, Cartwright EJP, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. Ann Appl Stat 2016. https://doi.org/10.1214/15-AOAS898.

[20] Lau MSY, Marion G, Streftaris G, Gibson G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. PLoS Comput Biol 2015. https://doi.org/10.1371/journal.pcbi.1004633.

[21] Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLoS Pathog 2018. https://doi.org/10.1371/journal.ppat.1006885.

[22] World Health Organisation. Measles virus nomenclature Update: 2012. Wkly Epidemiol Rec 2012;87:73–80. https://doi.org/10.1016/j.actatropica.2012.04.013.

[23] Penedos AR, Myers R, Hadef B, Aladin F, Brown KE. Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks 2015:1–16. https://doi.org/10.1371/journal.pone.0143081.

[24] Hiebert J, Severini A. Measles molecular epidemiology: What does it tell us and why is it important? Canada Commun Dis Rep 2014;40:257–60. https://doi.org/10.14745/ccdr.v40i12a06.

[25] Eddelbuettel D, François R. Rcpp: Seamless R and C++ integration. J Stat Softw 2011. https://doi.org/10.18637/jss.v040.i08.

[26] Robert A, Funk S, Kucharski AJ. o2geosocial (Version v1.0.2) 2021.

https://doi.org/https://doi.org/10.5281/zenodo.4818311.

[27] Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. Mach Learn 2003;50:5–43. https://doi.org/10.1023/A:1020281327116.

[28] Worby CJ, Lipsitch M, Hanage WP. Within-Host Bacterial Diversity Hinders Accurate Reconstruction of Transmission Networks from Genomic Distance Data. PLoS Comput Biol 2014;10. https://doi.org/10.1371/journal.pcbi.1003549.

[29] Lenormand M, Bassolas A, Ramasco JJ. Systematic comparison of trip distribution laws and models. J Transp Geogr 2016;51:158–69. https://doi.org/10.1016/j.jtrangeo.2015.12.008.

[30] Stouffer SA. Intervening Opportunities: A Theory Relating Mobility and Distance. Am Sociol Rev 1940. https://doi.org/10.2307/2084520.

[31] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 2008;5:0381–91. https://doi.org/10.1371/journal.pmed.0050074.

[32] Funk S. Socialmixr: social mixing matrices for infectious disease modelling 2018.

[33] Klinkenberg D, Nishiura H. The correlation between infectivity and incubation period of measles, estimated from households with two cases. J Theor Biol 2011;284:52–60. https://doi.org/10.1016/j.jtbi.2011.06.015.

[34] Vink MA, Bootsma MCJ, Wallinga J. Serial intervals of respiratory infectious diseases: A systematic review and analysis. Am J Epidemiol 2014;180:865–75. https://doi.org/10.1093/aje/kwu209.

[35] Dowle M. Package 'data.table.' Cran 2016.

[36] Lessler J, Reich NG, Brookmeyer R, Perl TM, Nelson KE. Incubation periods of acute respiratory viral infections: a systematic review 2015;9:291–300. https://doi.org/10.1016/S1473-3099(09)70069-6.Incubation.

[37] Fine PEM. The Interval between Successive Cases of an Infectious Disease. Am J Epidemiol 2003;158:1039–47. https://doi.org/10.1093/aje/kwg251.

[38] Hijmans RJ. Introduction to the geosphere package (version 1 . 9-92). Cran 2012:1–26.

[39] Wickham H. Ggplot2. Wiley Interdiscip Rev Comput Stat 2011;3:180–5. https://doi.org/10.1002/wics.147.

[40] Walker K. Tigris: An r package to access and work with geographic data from the us census bureau. R J 2016. https://doi.org/10.32614/rj-2016-043.

[41] Svensson Å. A note on generation times in epidemic models. Math Biosci 2007;208:300–11. https://doi.org/10.1016/j.mbs.2006.10.010.

[42] Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One 2007;2. https://doi.org/10.1371/journal.pone.0000758.

[43] Truscott J, Ferguson NM. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. PLoS Comput Biol 2012;8. https://doi.org/10.1371/journal.pcbi.1002699.

[44] Bjørnstad ON, Grenfell BT, Viboud C, King AA. Comparison of alternative models of human movement and the spread of disease. BioRxiv 2019:1–15. https://doi.org/10.1101/2019.12.19.882175.

[45] Merler S, Ajelli M. The role of population heterogeneity and human mobility in the spread of pandemic influenza. Proc R Soc B Biol Sci 2010;277:557–65. https://doi.org/10.1098/rspb.2009.1605.

Chapter 3. Probabilistic reconstruction of measles transmission clusters from routinely collected surveillance data

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1704167	**Title**	Mr
First Name(s)	Alexis		
Surname/Family Name	Robert		
Thesis Title	Modelling the risks of measles outbreaks near elimination		
Primary Supervisor	Sebastian Funk		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	Jounal of the Royal Society Interface		
When was the work published?	July 2020		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	**No**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

SECTION E

Student Signature	[redacted]
Date	31/05/2021

Supervisor Signature	[redacted]
Date	31/5/21

3.1. Abstract

Pockets of susceptibility resulting from spatial or social heterogeneity in vaccine coverage can drive measles outbreaks, as cases imported into such pockets are likely to cause further transmission and lead to large transmission clusters. Characterizing the dynamics of transmission is essential for identifying which individuals and regions might be most at risk. As data from detailed contact-tracing investigations are not available in many settings, we developed an R package called *o2geosocial* to reconstruct the transmission clusters and the importation status of the cases from their age, location, genotype and onset date. We compared our inferred cluster size distributions to 737 transmission clusters identified through detailed contact-tracing in the USA between 2001 and 2016. We were able to reconstruct the importation status of the cases and found good agreement between the inferred and reference clusters. The results were improved when the contact-tracing investigations were used to set the importation status before running the model. Spatial heterogeneity in vaccine coverage is difficult to measure directly. Our approach was able to highlight areas with potential for local transmission using a minimal number of variables and could be applied to assess the intensity of ongoing transmission in a region.

3.2. Introduction

Establishing who infected whom during an outbreak can help inform the design and evaluation of control measures [1–5]. Transmission links can be reconstructed through contact-tracing investigations, whereby cases are asked their movements and contacts during their infectious period. Given that contact-tracing investigations are not always carried out due to the logistical effort and cost involved, inference methods have been developed to use epidemiological data to estimate the probability that a transmission event occurred between any given pair of cases [6–12]. This makes it possible to establish probabilistic transmission trees that link all observed cases. The ensemble of cases belonging to the same transmission tree is called a transmission cluster.

Wallinga & Teunis [2] first developed a likelihood-based estimation procedure to reconstruct probabilistic transmission trees from a given distribution of generation times and observed symptom-onset dates of each case. Since then, genomic, spatial or contact data have been used to supplement the timing of symptoms, which helped identify determinants of transmission, mixing behaviour, individual dispersion, evaluate control measures, anticipate future developments of outbreaks and study viral evolutionary patterns [5,8,9,13–17].

As sequencing of pathogens has become more common, the use of such data to infer transmission trees has increased. Methods developed to add genetic distance to a Wallinga–Teunis algorithm, where cases with lower genetic distance are more likely to be grouped in the same transmission group, showed it substantially increased the accuracy of the reconstructed transmission trees [8,18–21].

The utility of sequence data depends on the characteristics of the pathogen [22,23]. Based on the highly variable 450 nucleotides region of the N gene (N-450) of the measles virus genome, eight measles genotypes have been detected since 2009 [24,25]; these genotype designations are helpful in linking cases, as linked cases must be infected by a virus of the same genotype [25]; however, the diversity of measles genotypes is decreasing [26]. It has been suggested that further sequencing the M-F non-coding region, or full genome sequencing, could help identify measles virus transmission trees, but so far, extended sequencing during measles outbreaks has been scarce [27,28]. In addition, the evolutionary rate of measles virus is very low [29]; therefore, samples from unrelated cases can be very close genetically and genetic sequences from measles cases are not usually indicative of direct transmission links [27,28].

As measles is highly infectious, under-immunized communities (also called pockets of susceptibles) resulting from local heterogeneity in vaccine coverage can lead to large, long-lasting outbreaks [30–34]. Detecting these pockets of susceptibles can be challenging, as historical local values of coverage throughout a given country are rarely available. The number of cases in the transmission trees resulting from each importation during outbreaks, also called the cluster size distribution, will depend both on individual factors (e.g. age of the imported case which might affect contact patterns) and community factors (e.g. the history of coverage in the area) [35,36]. The size of a cluster can, therefore, reflect the level of susceptibility of individuals directly and indirectly connected to the imported case [37,38].

Here, we introduced a model combining age, location, genotype and rash onset date of cases to reconstruct probabilistic transmission trees. We chose these features to make the model applicable to a wide range of settings as they are commonly reported and informative on transmission. We wrote the R package *o2geosocial* to conduct inference on individual-level data using this model. It is based on the package *outbreaker2* and is designed for outbreaks with partial sampling of cases, or uninformative genetic sequences, such as measles outbreaks [9,39]. We used the likelihood of transmission links between different cases to estimate their importation status. We compared the inferred importation status and cluster size distribution to the transmission clusters identified via contact tracing during measles outbreaks in the USA between 2001 and 2016.

3.3. Methods

3.3.1. Presentation of the algorithm

Transmission trees are used to represent who infected whom during an outbreak. They are directed acyclic graphs, where nodes are the reported cases and edges show the connection between them. The root of each transmission tree is an imported case, i.e. a case who was infected in a different transmission setting. The cases placed in the same transmission tree form a transmission cluster. We

estimated the number of cases per cluster (cluster size distribution) and the importation status of the cases from probabilistic transmission trees inferred using routinely collected epidemiological variables.

We used a Metropolis–Hastings algorithm with Markov chain Monte Carlo (MCMC) to classify a set of cases into a set of transmission trees with associated probabilities quantified using a Bayesian model to combine the epidemiological features of the cases. At every iteration of the MCMC algorithm, we proposed a new set of model parameters, infection dates and connections between cases. These three elements formed a tree proposal. We computed the ratio between the posterior probability of this proposal and the current posterior probability. The posterior probability (up to a multiplicative constant which would cancel out when calculating the ratio) was calculated from the likelihood of the trees, and the prior probability of the parameters. The log-likelihood of each tree was equal to the sum of the log-likelihoods of each case. All the notations are defined in Table 3.1.

Table 3.1: Table of notations of all variables and distributions defined in the methods.

Parameter	**Symbol**
Onset date	t_i, t_j
Infection date	T_i
Age	α_i, α_j
Tree	τ_j
Genotype	g_i, g_{τ_j}
Region	r_i, r_j
Number of generations	κ_{ji}
Spatial parameters	a,b,c
Conditional report ratio	ρ
Connectivity	$n_{r_j r_i}$
Population	m_{r_i}, m_{r_j}
Distance	$d_{r_i r_j}$
Parameter set	θ

Importation threshold	λ
Generation time distribution	$w(t_i - t_j)$
Latent period distribution	$f(t_i - T_i)$
Age contact probability	$a(\alpha_i, \alpha_j)$
Genotype probability	$G(g_i, g_{\tau_j})$
Probability of missing generation	$p(\kappa_{ji}\|\rho)$
Spatial probability	$s(r_i, r_j\|\, a, b)$
Log-likelihood of connection between i and j	$\mathrm{L_{ji}}(t_i\,, t_j\,, \theta)$
Individual log-likelihood	$L_i(t_i\,, \mathrm{j}, t_j\,, \theta)$

3.3.1.1. Likelihood function and parameter definition

In a tree proposal, each case i was assigned an infector j and an infection date t_i. We computed the log-likelihood of each case, $L_i(t_i,\ j,\ t_j,\ \theta)$ to calculate the likelihood of the tree. The log-likelihood of i was split in two: (i) the log-probability density of observing the onset date T_i if case *i* had been infected at time t_i $\log(f(t_i\ -\ T_i))$ and (ii) the log-likelihood of connection between i and j $\mathrm{L_{ji}}(t_i\,, t_j\,, \theta)$, with θ the parameter set of the model (1):

$$L_i(t_i\,, \mathrm{j}, t_j\,, \theta) = \log(f(t_i\ -T_i)) + \mathrm{L_{ji}}(t_i\,, t_j\,, \theta) \quad (1)$$

The function *f* represents the distribution of the incubation period. The log-likelihood of connection $\mathrm{L_{ji}}$ was computed from five components reflecting the age group, genotype, location, inferred date of infection of cases i and j, and the report ratio (2). We allowed for an indirect link between cases due to unreported individuals, κ_{ji} corresponds to the number of generations between i and j. If $\kappa_{ji}\ =\ 1$, j infected i, whereas if $\kappa_{ji}\ =\ 2$, an unreported case infected by j infected i, κ_{ji} increases with the number of missing links between i and j

$$\mathrm{L_{ji}}(t_i\,, t_j\,, \theta) = \log\left(p(\kappa_{ji}|\rho)\ \times\ w^{(\kappa_{ji})}(t_i - t_j) \times a^{(\kappa_{ji})}(\alpha_i, \alpha_j) \times G(g_i, g_{\tau_j}) \times s^{(\kappa_{ji})}(r_i, r_j|\, a, b)\right) \quad (2)$$

We calculated the temporal probability of transmission between i and j from the number of days between t_i and t_j and the distribution of the generation time of the disease $w(\mathrm{t})$. This probability was quantified by $w^{(\kappa_{ji})}(t_i - t_j, \kappa_{ji})$, $w^{(\kappa_{ji})} = w * w * \ldots * w$, where $*$ is the convolution operator applied κ_{ji} times. We used a geometric distribution $p(\kappa_{ji}|\rho)$ to quantify the probability of observing κ_{ji} missing

generation between i and j, given the conditional report ratio ρ. The conditional report ratio quantifies the probability of missing generations between two connected reported cases. Entire missing clusters, cases infected after the last cases or cases infected before the ancestor of a cluster would not interfere in the connection between two cases and, therefore, would not affect the value of the conditional report ratio. The conditional report ratio can be higher than the overall report ratio of an outbreak. The 'ancestor' is the earliest identified case in a cluster.

$a(\alpha_i,\ \alpha_j,\ \kappa_{ji})$ was defined as the probability of transmission between age groups α_i and α_j. This probability corresponds to the proportion of contacts to the age group α_i that originated from α_j and can be deduced from studies such as POLYMOD [36]. We defined $G\left(g_i, g_{\tau_j}\right)$ as the probability of observing the pathogen genotype g_i in case i in the tree τ_j containing case *j*. There can only be one measles virus genotype per transmission tree, or cases with unreported genotype. The genotype g_{τ_j} is the genotype contained in the tree τ_j and is known if at least one case in τ_j had a reported genotype.

$$G\left(g_i, g_{\tau_j}\right) = \begin{cases} 1\ if\ g_i \text{ unknown} \\ 1\ if\ g_{\tau_j} \text{ unknown} \\ 1\ if\ g_i \text{ and } g_{\tau_j} \text{both known and } g_i = g_{\tau_j} \\ 0\ otherwise \end{cases} \quad (3)$$

In (3), if $G\left(g_i, g_{\tau_j}\right) = 0$, then the connection between i and j is impossible, and (1) and (2) are equal to $log(0) = -\infty$.

$s(r_i,\ r_j,\ \kappa_{ij})$ was defined as the probability of connection from r_j to r_i, regions of residency of i and j (4). We used an exponential gravity model to quantify the connectivity of the different geographical units [40]. This approach showed good performance at modelling short distance commuting, and was easy to parametrise [40–44]. In the simplest form of the exponential gravity model, the number of connections between r_i and r_j is proportional to the product of the origin population m_{r_j}, the destination population m_{r_i} and an exponential decrease of the distance between r_i and r_j $d_{r_j r_i}$: $\mathrm{n_{r_j r_i}} \propto e^{-a \times d_{r_j r_i}} \times \mathrm{m_{r_j}^b} \times \mathrm{m_{r_i}^c}$, with a, b and c parameters adjusting for the impact of distance and population. From this definition, we deduced $s(r_j, r_i)$, the spatial probability of transmission from i to j:

$$s(r_i, r_j) = \frac{\mathrm{n_{r_j r_i}}}{\sum_h n_{\mathrm{hr_i}}} = \frac{e^{-a \times d_{\mathrm{r_j r_i}}} \times \mathrm{m_{r_j}^b} \times \mathrm{m_{r_i}^c}}{\sum_h e^{-a \times d_{\mathrm{hr_i}}} \times \mathrm{m_h^b} \times \mathrm{m_{r_i}^c}}$$

$$= \frac{e^{-a \times d_{\mathrm{r_j r_i}}} \times \mathrm{m_{r_j}^b}}{\sum_h e^{-a \times d_{\mathrm{hr_i}}} \times \mathrm{m_h^b}} \quad (4)$$

Only the parameters a and b were required to compute the spatial probability of transmission. If $r_i = r_j$, then (4) becomes: $s(r_i, r_j) = \frac{m_{r_i}^b}{\sum_h m_h^b}$, Other distributions than the exponential decrease can be used in this framework if transmission follows a different pattern.

The parameters ρ, a and b were estimated. At each iteration of the MCMC, the log-likelihood of the trees was equal to the sum of all individual log-likelihoods L_i from equation (1). The log-posterior density of the proposed trees was calculated by summing the overall log-likelihood of the trees and the log-priors of the parameters.

3.3.1.2. Tree proposals

We used a Metropolis-Hastings algorithm with MCMC to sample from the posterior distribution of parameters and the transmission trees. To do this, we developed a set of proposal tree updates. These updates were accepted with acceptance probability as defined by the Metropolis-Hastings algorithm [45]. We used eight types of tree proposal to ensure good mixing. Each proposal conserved the overall number of trees, with a maximum of one unique genotype reported per tree.

Five of the proposals had already been implemented in the *outbreaker2* package and were adapted to this setting: (i) change the number of generations between two cases; (ii) change the conditional report ratio ρ; (iii) change the time of infection; (iv) change the infector of a case (if the case is not the ancestor of a tree); (v) swap infector-infectee (if none is the ancestor of a tree).

We added two proposals to change a and b, the spatial kernel parameters. For each proposal, the probability of transmission between every geographical unit was recalculated with the new values. The distance matrix had to be computed for each number of generations between cases, which considerably slowed down the algorithm. As we could not use sequence data, assessing whether a case was isolated or whether it was connected to a reported infector with two missing generations would be very challenging using our model alone. Therefore, we limited the maximal number of missing generations to 1 when a or b were estimated ($\max(\kappa_{ji}) = 2$). Finally, the last proposal was designed to change the ancestor of the tree while conserving the overall number of trees (Figure 3.1).

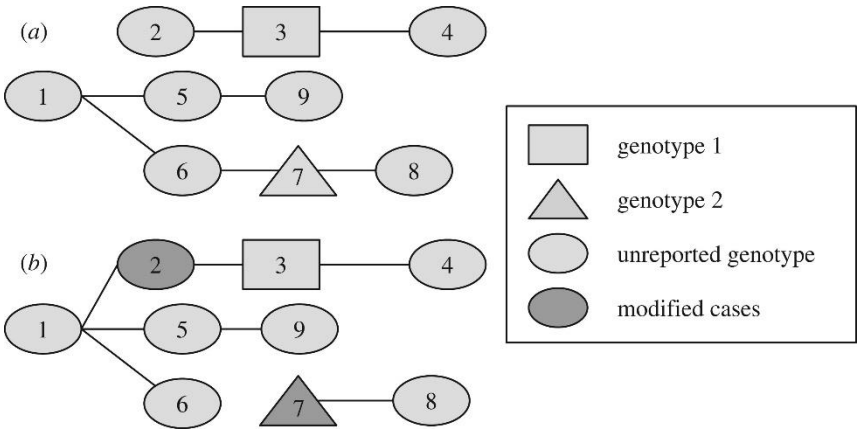

Figure 3.1: Example of the change of ancestors. (a) The initial tree and (b) the new tree proposed after the movement. Initially, there are two ancestors (cases 1 and 2) in a group of nine cases. Cases 3 and 7 have different genotypes and cannot be part of the same tree, the genotypes of the other cases are not reported. The date of infection is in increasing order (1 is the first case, 9 is the last). Therefore, 1 is the only potential infector for 2. One new ancestor was randomly drawn to conserve the number of trees. In this example, 7 is the new ancestor (6 was the only other possibility). The ratio of the posterior densities of (a,b) were then used to determine whether to accept or reject the proposal, according to the Metropolis–Hastings algorithm. This movement ensures good mixing of the potential ancestors of the transmission clusters.

3.3.1.3. Inference of importation status and cluster

Unrelated measles cases stemming from different importations and different regions can be part of the same dataset. Grouping cases and excluding unrealistic transmission links reduces the number of possible trees and speeds up the MCMC runs. To do so, we listed each case's potential infectors using three criteria: (i) the potential infectors must be of the same genotype as the case, or have unreported genotype, (ii) the location of potential infectors must be less than γ km away from the case, and (iii) the potential infectors must have been reported later than δ days before the case. This threshold should be determined from the maximum plausible generation time of the disease. The spatial threshold γ should be defined according to the relevance of long-distance transmissions. Cases with no potential infector were considered as importations. Otherwise, they were grouped together with (i) their potential infectors and (ii) cases with common potential infectors.

After grouping the cases, we estimate their importation status and the cluster size distribution using two runs of MCMC (Figure 3.2). The first run was shorter and aimed at removing the most unlikely connections among each group, as they can reflect unrealistic estimates for incubation periods or generation times and corrupt the estimation of the date of infection. We defined a reference threshold λ, whereby if the individual value of log-likelihood $\mathrm{L_i}$ was worse than λ, then the connection between i and their infector was considered unlikely. In *Outbreaker2,* λ was a relative value, defined from a quantile of the individual log-likelihoods. In *o2geosocial*, λ can be a relative value or an absolute value, chosen from the number of components of the likelihood. For each sample saved from the short

run, we computed the number of unlikely connections *n*. If there was no iteration where all connections were better than λ, *min(n)* new importations were added to the initial tree for the long run (Figure 3.2).

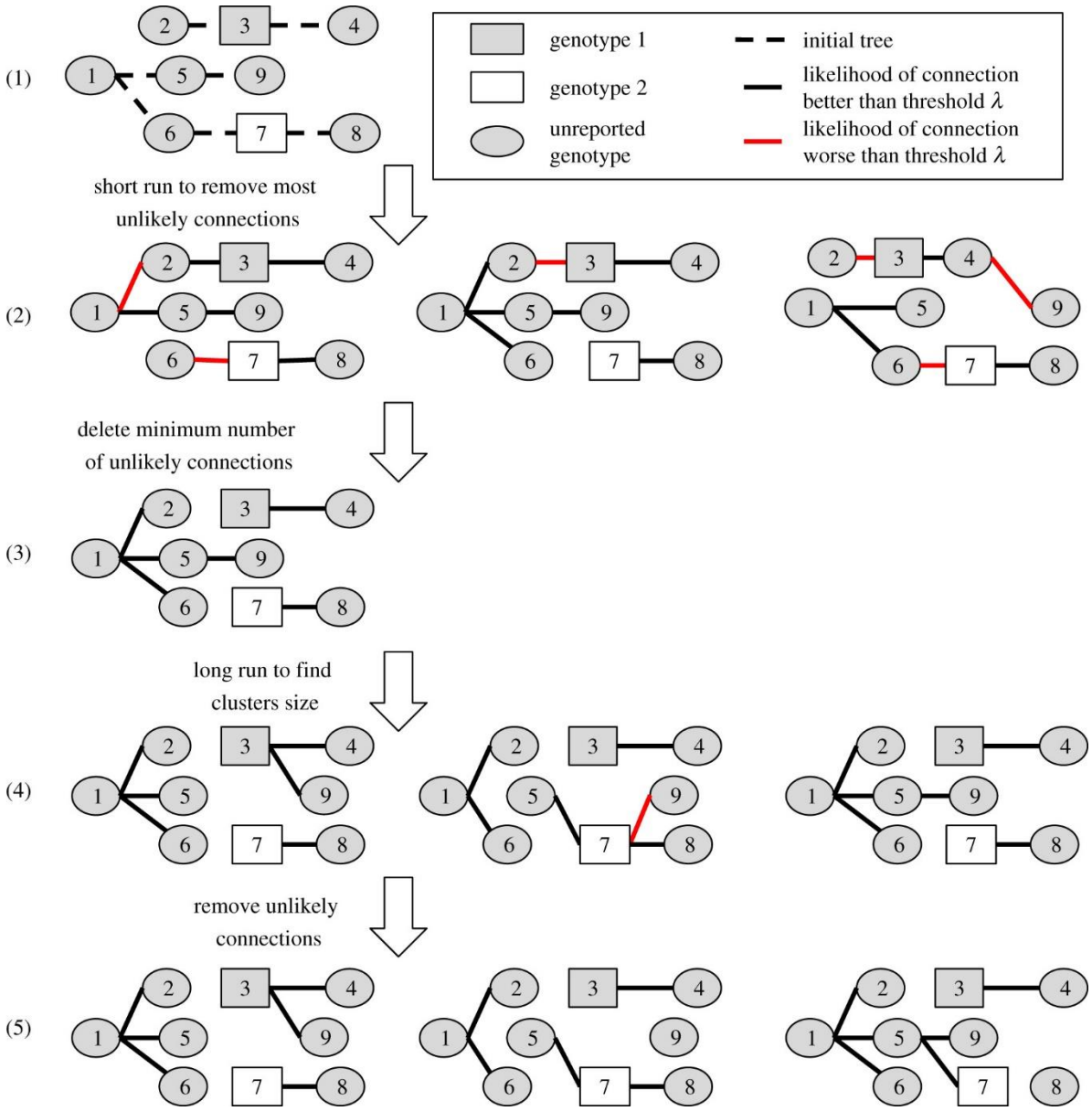

Figure 3.2: Estimating importation status and cluster size distributions in two MCMC runs. Step 1: initial tree obtained after pre-clustering, with the minimum number of importations (here 2, as there are two reported genotypes). Step 2: samples from the first short run, with red lines showing connections worse than the arbitrary threshold λ. Step 3: initial tree for the final run, with one more importation than in step 1, which corresponds to the minimum number of unlikely transmissions at step 2. Step 4: samples from the long run. Step 5: final trees used to compute cluster size distribution and importation status of each case. Case 7 is an importation in one-third of the final samples, whereas case 3 is an importation in all of them.

Finally, we ran a long MCMC chain and obtained samples from the posterior distribution. After removing the burn-in period and thinning the chain, we deleted the unlikely transmission links in each iteration and identified transmission clusters. Therefore, unlike the previous versions of *outbreaker2*, the number of importations in each sample can vary and the individual probability of being an importation can be computed (Figure 3.2).

3.3.2. Validation case study: measles outbreaks in the USA between 2001 and 2016

3.3.2.1. Data

To evaluate the performance of the model, we inferred the transmission clusters from a dataset that also included information on whether measles cases were part of a cluster based on contact-tracing investigations. Measles cases in the USA are reported by healthcare providers and clinical laboratories to their corresponding health department. Each case is investigated by local and state health departments classified according to standard case definitions [46], and linked into clusters epidemiologically (e.g. by establishing a direct contact or a shared location between cases, or when cases are part of a specific community where an outbreak is occurring). Cases are considered internationally imported if at least part of the exposure period (7–21 days before rash onset) occurred outside the USA and rash occurred within 21 days of entry into the USA, with no known exposure to measles in the United States during the exposure period.

Confirmed measles cases are routinely reported by state health departments to the CDC. A total of 2098 measles cases were reported in the USA between January 2001 and December 2016. The number of annual cases did not exceed 700 cases during this time period (Figure 3.3, electronic supplementary material, Figure S1). The importation status, 5-year age group, onset date, county and state of residence were fully reported for 2077 cases. The 21 cases with missing data were discarded. Twenty-five per cent of the cases were classified as importations. Thirty-nine per cent of the cases had their genotype reported.

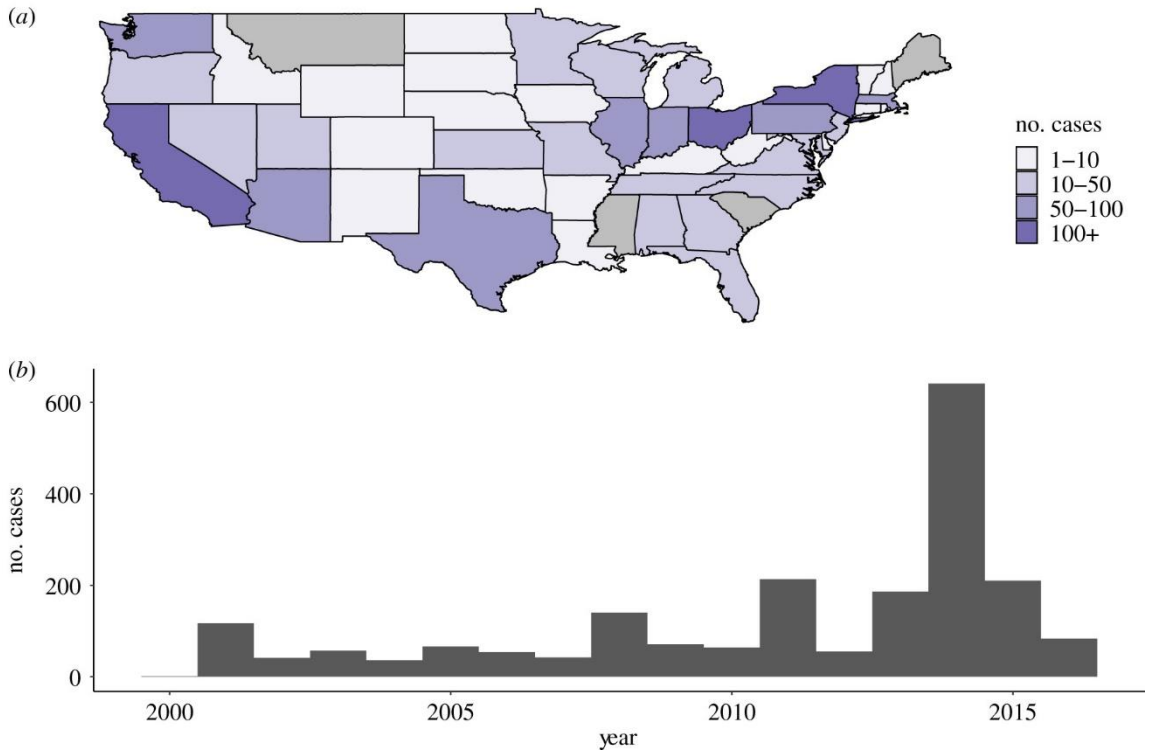

Figure 3.3: (a) Number of cases per state and (b) annual number of cases reported in the USA between 2001 and 2016. Alaska and Hawaii are not shown in (a).

Among cases with complete data, 737 independent clusters, containing 1-380 cases, were reconstructed through contact-tracing investigations. Not every identified case could be linked to an importation, and some transmission clusters contained multiple imported cases (e.g. when related individuals travel together to a foreign country and were infected there). Out of the 737 reference clusters, 38 had several cases classified as importations, 256 had none identified.

3.3.2.2. Model and parameters

The distributions and priors used in the studies are listed in Table 3.2. As no studies quantifying the probability of age-specific contacts have been carried out in the USA, we used the estimates from the POLYMOD study in the UK [36]. The incubation period and the generation time of measles were taken from previous studies [47–49]. We used the population centroid of each county to compute the distance matrix [50]. We used a beta distribution as the prior of the conditional report ratio [8]. The mean of the prior distribution was calculated using the number of clusters whose first case was not classified as an imported case, meaning the investigations were not able to trace back to the first case imported. As there was no prior information on the possible values of the spatial parameters a and b, we used uniform distributions between 0 and 5.

Table 3.2: Values of parameters used to cluster cases declared in the USA.

Parameter	**Symbol**	**Distribution**
Incubation period	$f(t)$	Gamma,

		mean = 11.5, sd = 2.24
Generation time	$w(t)$	Normal, Mean = 11.7, sd = 2.0
Conditional report ratio	ρ	Prior: Beta distribution, Mean = 0.65, sd = 0.15
Spatial parameter 1	a	Prior: Uniform distribution
Spatial parameter 2	b	Prior: Uniform distribution
Spatial pre clustering	γ	Fixed: 100 km
Temporal pre clustering	δ	Fixed: 30 days
Importation threshold	λ	Absolute: • $5 \times \log 0.05 = -15$ • $5 \times \log 0.1 = -11$ Relative: • 5%

For pre-clustering of cases, we set the temporal threshold δ to 30 days, which is above the 97.5% upper quantile of the generation time with a missing generation. We were interested in local transmission to describe the impact of an imported case on a community. But we only had information on the county of residency for each case. Counties are large geographical units: the average county land area is 2911 km^2 and the maximum values reach 50 000 km^2. Therefore, we set the spatial threshold γ to 100 km to exclude long-distance transmission, while still allowing for cross-county transmission.

Finally, we tested several relative and absolute importation thresholds λ. Absolute values were calculated from a factor k, multiplied by the number of components in L_i, excluding the binary genetic component. Tested values were *k = 0.05* ($\lambda = \log(0.05) * 5 = -15$) and *k = 0.1* ($\lambda = -11$). Connections were considered unlikely if the log-likelihood was worse than λ. Relative values were quantiles of all recorded log-likelihoods in the sampled trees (Table 3.2).

3.3.2.3. Inference of importation status

Using the contact-tracing investigations, we considered three different initial distributions of the importation status. In scenario 1, there was no inference of the importation status of cases, and the first case of each epidemiological cluster was classified as importation (ideal importation). In scenario 2: there was no inference of the importation status of cases, and all cases identified as importation in the contact-tracing investigations were classified as importations (epidemiological importation). Finally, in scenario 3, the importation status of cases was inferred, using different thresholds λ, and using no prior information on the importation status of cases or the importation status from the contact-tracing investigations.

3.3.2.4. *Inference of clusters*

In order to compare the inferred and reference clusters, we calculated for each case *i*: (i) the proportion of cases from the same reference cluster as i that were inferred with i (sensitivity) and (ii) the proportion of cases in the same inferred cluster as i that were part of the reference cluster (precision). These values were calculated at every iteration, and the median values were used to evaluate the fit obtained with different values of λ. We also compared the inferred cluster size distribution to the reference data. The credibility intervals for each case are reported in electronic supplementary material, Figure S2.

3.4. Results

We clustered 2077 measles cases reported in the USA between January 2001 and December 2016 using their onset date, age groups, location and genotype. Using the contact-tracing investigations, we considered three different initial importation status distributions: (i) only the ancestors of each epidemiological cluster (first case of each cluster) were importations (ideal importation), (ii) all cases classified as importation in the contact-tracing investigations were importations (epidemiological importation), (iii) no prior information on importation status of cases. The importation status of the cases was, therefore, not probabilistically inferred in scenarios 1 and 2. The length of the short preliminary run was 30 000 iterations and the main run was 70 000 iterations. For each run, the trace of the posterior distribution shows the convergence of the algorithm (electronic supplementary material, Figure S3).

In scenario 1, we did not infer the importation status of cases. The inferred cluster size distribution matched the contact-tracing investigations (Figure 3.4A); 98% of the reference singletons were also isolated in the inferred cluster. For 94% (95% credibility interval: 91–98%) of cases, the inferred cluster had a sensitivity and precision above 75%, meaning more than 75% of the cases in the inferred cluster were in the reference cluster, and more than 75% of the cases in the reference cluster were in the inferred cluster (Figure 3.4B). For 80% (78–93%) of cases, the inferred clusters were a perfect match with the reference clusters. The cluster size distribution stratified by state was similar to the contact-tracing investigations (electronic supplementary material, Figure S4). Therefore, when each ancestor was considered as an importation, the inferred clusters were very close to the reference ones.

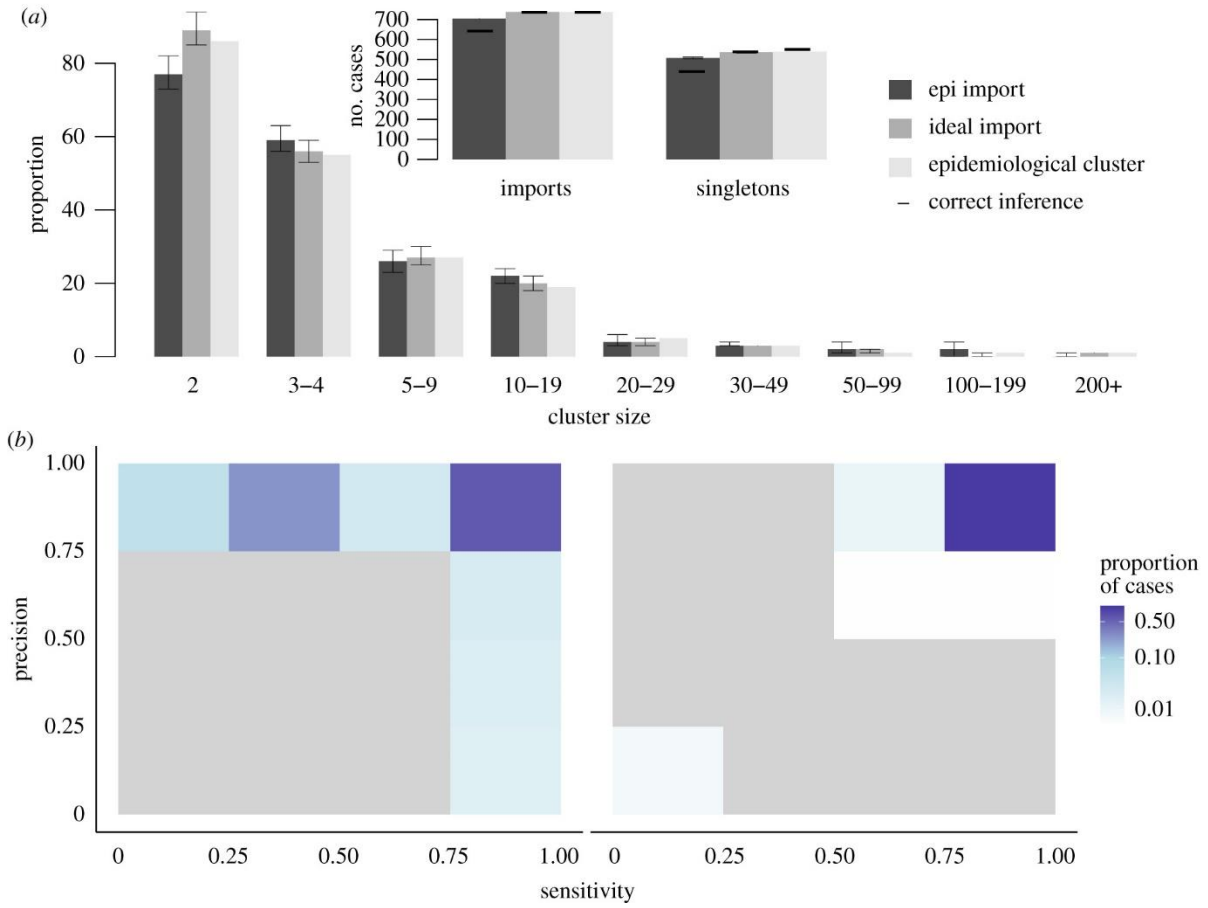

Figure 3.4: Description of transmission clusters inferred using prior knowledge on importation status of cases. (a) Cluster size distribution for scenarios 1 and 2 (grey and dark grey), compared to the reference clusters (light grey). Arrows represent the 95% credibility intervals of each estimate. Only clusters containing at least two cases are represented. Inset: Number of importations and number of isolated cases (singletons) in scenarios 1 and 2, and in the reference clusters. For each scenario, the horizontal dark line represents the number of importations that are also importations in the reference clusters, same for singletons. (b) Heatmap representing the precision and sensitivity of the clusters for each case in scenario 1, cases are classified in a category depending on the proportion of their reference cluster that were inferred in the same cluster (x-axis) and the proportion of mismatches in the inferred cluster. The same for scenario 2.

In scenario 2, we used the importation status distribution of cases reported in the contact-tracing investigations (539 importations). Pre-clustering highlighted 165 cases with no potential infector, which were also classified as importations. We observed discrepancies between the inferred cluster size distribution and the reference one: among the 704 cases inferred as importation, 61 (9%) were not importations in the reference cluster. Furthermore, 94 cases were the ancestor of a reference cluster and were not classified as importations in the inferred clusters (13%). The overall cluster size distribution matched the reference distribution, but 111 reference singletons were inferred as part of transmission clusters (Figure 3.4A; electronic supplementary material, Figure S5). Although the precision of the inferred cluster was above 75% for 93% (88–93%) of the cases, 31% (6–39%) had a sensitivity score below 0.5, meaning they were classified with less than half of the cases from their reference clusters (Figure 3.4C). The discrepancies observed in this scenario are due to inconsistencies between the

importation status distribution and the clustering of cases in the contact-tracing investigations, as reference clusters that gathered several importations were split into different inferred clusters.

In scenario 3, we used different threshold λ to infer the importation status of cases. We tested $\lambda = -15$, $\lambda = -15$ (absolute value), and $\lambda = 95^{th}$ centile of all recorded log-likelihoods (relative value). For each case i, if the log-likelihood L_i was worse than λ, the connection between the case and its infector was removed and the case was considered imported. Firstly, using an absolute factor $\lambda = -15$, 586 (581–593) cases were classified as importations, and 361 (355–369) of them were singletons. These numbers are much lower than the reference dataset that contains 737 clusters, and 539 singletons (Figure 3.5A; electronic supplementary material, Figure S6). However, very few cases inferred as importations or singletons were not classified as such in the reference dataset (15 (10–22) misclassified importations, 4 (0–14) misclassified singletons), and the cluster size distribution for clusters including two cases and more was very similar to the reference one. The precision of the reconstructed cluster was very high (above 75% for 88% (85–93%) of cases) (Figure 3.5B). Overall, the algorithm was not able to accurately identify importations and singletons as the threshold was too low to eliminate some unrealistic connections, but the inferred larger clusters matched their reference counterparts.

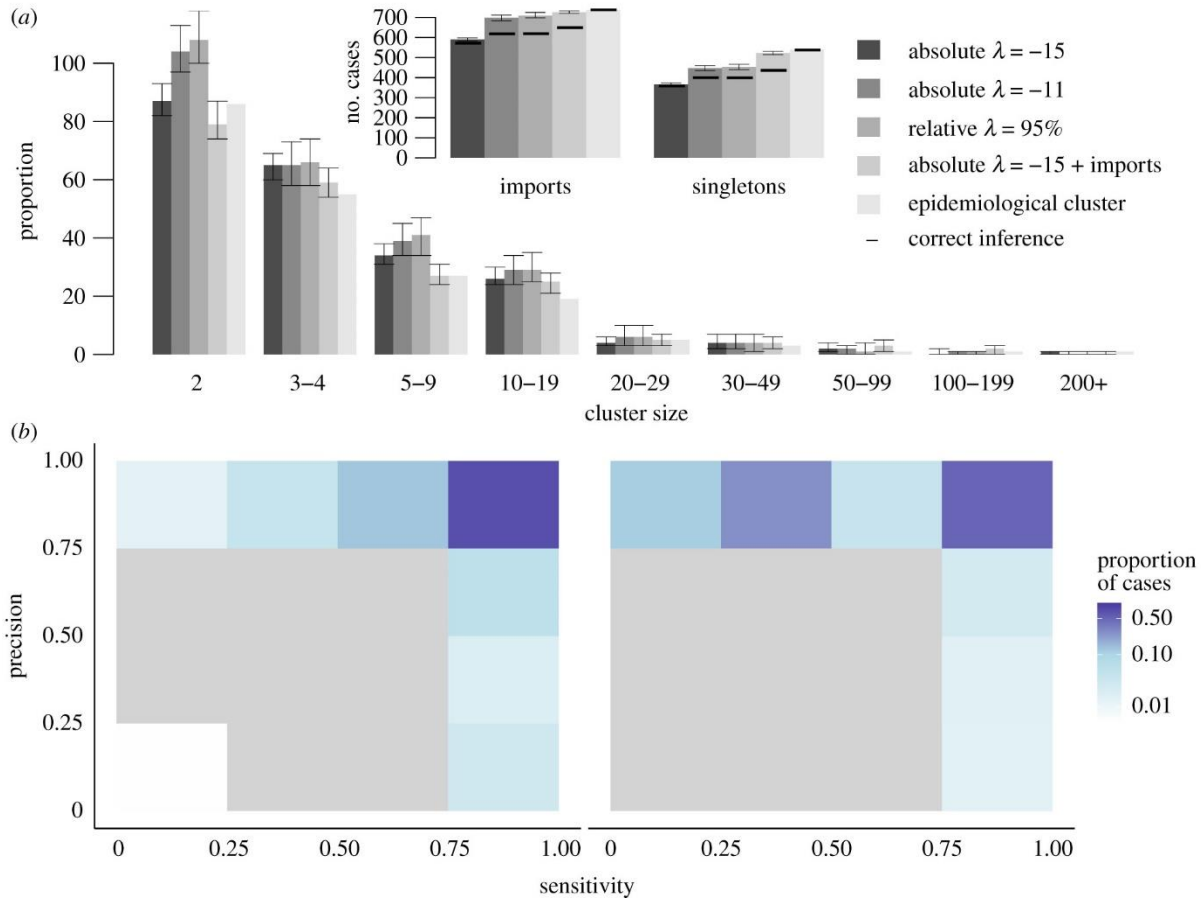

Figure 3.5: Description of transmission clusters generated with inferred importation status of cases. (a) Cluster size distribution for different value of threshold in scenario 3 (sorted by shades of grey), compared to the reference clusters (light grey). Arrows

represent the 95% credibility intervals of each estimate. Only clusters containing at least two cases are represented. Inset: Number of importations and number of isolated cases (singletons). For each scenario, the horizontal dark line represents the number of importations that are also importations in the reference clusters, same for singletons. (b) Heatmap representing the precision and sensitivity of the clusters for each case in scenario 3, with a 5% relative threshold, cases are classified in a category depending on the proportion of their reference cluster that were inferred in the same cluster. (c) Same when importation status is taken from the contact-tracing investigations and inferred using a 5% relative threshold.

We then observed the impact of increasing λ on the inferred cluster size distribution. Runs obtained using an absolute threshold with $\lambda = -11$ and 95% relative threshold yielded very similar results. The number of cases inferred as importations was higher than in previous runs, while all remaining links showed good connection between cases. The number of importations was closer to the reference dataset, and the number of singletons was greater than the reference. Nevertheless, 11% (10–12%) of the inferred importations was not classified as importation in the reference clusters. Furthermore, the number of two-case chains was overestimated, and bigger clusters were likely to be split because of the removal of weaker connections. Therefore, increasing λ did not improve the cluster size distribution, as many importations in the reference clusters were not identified and the number of mismatches increased (electronic supplementary material, Figure S7).

Finally, we combined prior information and inference of importation status to create a scenario where the importation status of only a proportion of the cases is known, because of disparities in the contact-tracing investigations. This scenario is relevant for a dataset combining different outbreaks scattered across a large area or a long period of time. Cases considered as importations in the contact-tracing investigations were set as importations, and we inferred the importation status of the remaining cases. We used a low threshold to remove the least likely transmission links ($\lambda = -15$). Including prior information led to some misclassification of importation status due to the inconsistencies between the epidemiological importation status and the reference clusters. As in scenario 2, some cases were classified with only part of their reference clusters because clusters with several importations were split into different clusters. Indeed, the sensitivity score of 34% (7–51%) of cases was below 0.5. Nevertheless, the cluster size distribution observed in the simulation was the closest to the reference clusters. There were 725 (719–731) clusters, 89% of importations were also ancestors of reference clusters and the number of singletons matched the reference clusters (Figure 3.5A-C). The inferred clusters of 88% (86–94%) of the cases had a precision score of 1, showing they were clustered without any false positives. Despite discrepancies in several states (Massachusetts, Ohio), the cluster size distribution stratified by state showed good agreement with the reference clusters (electronic supplementary material, Figure S8).

The conditional report ratio in the transmission chains ρ and the spatial parameters a and b was estimated in each scenario. The parameter estimates did not depend on the prior importation status distribution or the value of λ. ρ was consistently estimated above 90%, showing a low number of

missing generations between cases (electronic supplementary material, Figure S9). High values of ρ show that most of the reported cases could be connected without missing generations. This is not representative of the overall report ratio, which is usually much lower [51].

There was little variation in the estimates of the spatial parameters between the different scenarios. The population parameter a was estimated between 0.6 and 1 for every scenario, and the distance parameter b was between 0.08 and 0.12. In every scenario, more than 80% of the inferred transmission were between cases distant of less than 10 km, and few long-distance transmissions were recorded (50–100 km); hence, although most of the reconstructed connections were between cases from the same county, the algorithm was able to identify clusters spreading over several counties or states (electronic supplementary material, Figure S10).

We highlighted the added value of including the spatial distance between cases in the likelihood by comparing the cluster size distribution inferred by selecting certain components of L_i (electronic supplementary material, Figure S11). The credibility intervals were much wider when the distance between cases is not part of the likelihood, and the number of chains containing 2–10 cases was overestimated. The important impact of the spatial component of likelihood was also due to the widespread American territory, and could be lower in a different setting.

We used the ratio of the number of importations over the number of subsequent cases per state to evaluate the intensity of transmission in each state between 2001 and 2016 (Figure 3.6). The maps obtained in scenario 1 (ideal scenario) or in scenario 3 (estimation of importation, with epidemiological importations and $\lambda = -15$) were very similar. We only observed minor differences, for example, in South Dakota and in Massachusetts, where the ratios were higher in scenario 3. The highest ratio (31.8 in scenario 1) was observed in Ohio, and is mostly due to a 383 case outbreak in 2014 [32]. We observed major differences between the incidence map (Figure 3.2A) and the ratio per state. Indeed, although 403 cases were reported in California (highest number in the USA), importations caused on average 1.32 subsequent cases in scenario 1 (1.60 in scenario 3), showing a high proportion of reported cases were inferred as importations.

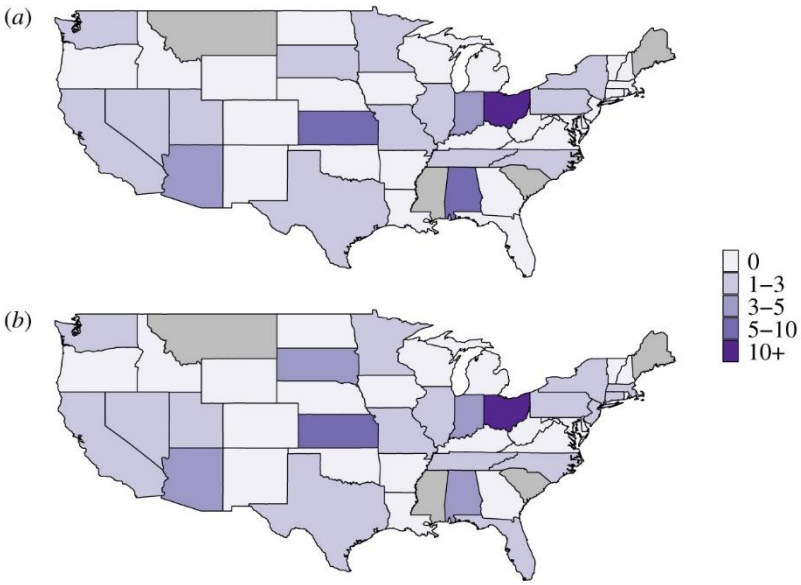

Figure 3.6: Ratio of the number of importations over the number of subsequent cases in each state in (a) scenario 1 (ideal importations) and (b) scenario 3 with epidemiological importations and $\lambda = -15$. Grey states represent states that did not report any case.

Similarly, we used the inferred transmission chain to compute the inferred reproduction number in each state. According to the model, about 60% cases did not cause future transmission, and about 5% caused more than five subsequent cases (electronic supplementary material, Figure S12). These numbers were consistent in each run. The geographical distribution of reproduction number was very similar to the importation–subsequent cases ratio (electronic supplementary material, Figure S13).

3.5. Discussion

We developed the R package *o2geosocial* to classify measles cases into transmission clusters and estimate their importation status using routinely collected surveillance data (genotype, age, onset date and location of the cases). As recently observed during the 2018–2019 measles outbreak in New York, delays in childhood vaccination, local susceptibility and increased contacts can lead to large outbreaks following importations [52,53]. Therefore, we were interested in highlighting the effect of imported cases on communities and we focused on short distance transmission to identify areas where they repeatedly caused subsequent transmission chains. Although this is not predictive of future transmission, it highlights communities with potential for large transmission clusters.

We compared the inferred transmission clusters to the contact-tracing investigations of 2077 confirmed measles cases reported in the USA between 2001 and 2016. We were able to produce reliable estimates of known transmission clusters using epidemiological features with only few misclassifications. Estimating the importation status of cases without prior knowledge was challenging and caused uncertainty on the results. We tested different threshold λ to eliminate unlikely transmissions, and we were able to identify most of the imported cases. Nevertheless, if several cases were imported in the

same region at a similar time, we could not find all of them without discarding valid transmission events, and increasing the number of false positives. When we used the importation status as defined in the contact-tracing investigations without probabilistic inference (scenarios 1 and 2), the reconstructed clusters were similar to the reference ones. Results were also conclusive when we combined prior information and importation inference. The reconstruction of transmission greatly depends on the epidemiological investigations to identify measles importations in a community.

We used the genotype to censor connections between cases when it was reported, as there can be only one reported genotype per transmission cluster. Using a simulated dataset (*toy_outbreak_long* in *o2geosocial*), we explored the impact of increasing the proportion of genotyped cases on clustering and observed it could help identify the number of concurrent transmission trees when multiple genotypes are co-circulating. Moreover, we introduced a spatial component to the likelihood of connection between cases using an exponential gravity model. Previous studies showed this model was able to capture short-distance dynamics better than other gravity models, and was easy to parametrize. Introducing the spatial component greatly improved the precision and the sensitivity of the reconstructed clusters (electronic supplementary material, Figure S11), and the parameter estimates were robust in the different scenarios.

The final results on the clustering of the 2077 cases using *o2geosocial* were obtained in 7 h for each run of 100 000 iterations on a standard desktop computer (Intel Core i7, 3.20 GHz 6 cores), which is much faster than previous implementations of *outbreaker* and *outbreaker2*. With the addition of the pre-clustering step, whereby we reduced the number of potential infectors for each case, the algorithm ran faster. For smaller chains (50 000 iterations), 4 h were needed to estimate the importation status and cluster the cases. The code for the package and the analysis developed in this project is shared on Github (alxsrobert/o2geosocial and *alxsrobert/datapaperMO*), with an illustrative toy dataset, and can be used to analyse recent outbreaks where contact-tracing investigations were not carried out.

Although the results obtained are promising, it should be noted that the dynamics of measles transmission in the USA are likely to be very specific to this location. Indeed, there were less than 700 annual cases between 2001 and 2016. These cases were scattered across a large area, which made the pre-clustering of cases very efficient as we focused on short-distance transmission. In smaller or more endemic settings, the number of potential infectors per cases after the pre-clustering step might be higher, which would increase the running time.

Furthermore, as the location of each case was deduced from the population centroid of counties, we assumed that the distance between cases from the same county was effectively zero. American counties are large and widespread geographical units that can include more than 1 million individuals. For future

use of *o2geosocial,* more accurate information on the location of cases could improve cluster inference by identifying multiple importations in a given county. Because cases are reported by the state of residency, we had to ignore that cases may have been out of the reported county or state during their incubation and infectious period, which has been seen during some outbreaks, such as the 2015 'Disney outbreak' in California [54].

We did not include prior information on the local susceptibility of the different areas affected in *o2geosocial*, and these could be estimated using historical values of local coverage. However, protocols to estimate local vaccination coverage can differ in time and space and be difficult to compare, or unavailable at the local level. Furthermore, these estimates are cross-sectional in nature, and might not take into account catch-up vaccination campaigns, or immunity induced by previous outbreaks. Local seroprevalence surveys could identify pockets of susceptibles, but they have not been carried out on a subnational scale in most countries [55].

There has been no national quantitative analysis of age-specific contact patterns carried out in the USA, so we relied on a contact matrix between age groups available for Great Britain from the POLYMOD study [36]. Nevertheless, little variation in the contact rates between age groups has been observed between European countries, and a previous projection of the social contact matrix in the USA yielded similar results [56]. POLYMOD data were probably the most reliable source of information we could use to deduce an estimate of the contact matrix in the USA.

3.6. Conclusion

Heterogeneity in immunity can cause large outbreaks in countries with high national vaccine coverage, and identifying potential foyers of transmission in post-elimination settings is key for outbreak prevention and control. We have presented a method for estimating the cluster size distribution of past measles outbreaks from routinely collected surveillance data. We found that adding prior knowledge on the importation status of cases improved the inference of the transmission clusters. Although the method was able to identify a proportion of importations, epidemiological investigations on the history of travel and exposure reduced uncertainty on the clustering of cases. We believe these investigations are needed to produce reliable estimates of past transmission clusters. In lieu of the importation status, if multiple genotypes are co-circulating, increasing the proportion of genotyped cases could help discard potential connections and find imported cases. Even with limited information, this method was able to infer probabilistic transmission clusters in a fast and efficient way.

3.7. Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention, US Department of Health and Human Services

3.8. Reference

[1] Ferguson NM, Donnelly CA, Anderson RM. Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. Nature 2001. https://doi.org/10.1038/35097116.

[2] Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal. Am J Epidemiol 2004;160:509–16.

[3] Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355–9. https://doi.org/10.1038/nature04153.

[4] Faye O, Boëlle P-Y, Heleze E, Faye O, Loucoubar C, Magassouba N, et al. Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study. Lancet Infect Dis 2015;15:320–6. https://doi.org/10.1016/S1473-3099(14)71075-8.

[5] Ypma RJF, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. Genetics 2013;195:1055–62. https://doi.org/10.1534/genetics.113.154856.

[6] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc B Biol Sci 2007;274:599–604. https://doi.org/10.1098/rspb.2006.3754.

[7] Cauchemez S, Ferguson NM. Methods to infer transmission risk factors in complex outbreak data. J R Soc Interface 2012;9:456–69. https://doi.org/10.1098/rsif.2011.0379.

[8] Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Comput Biol 2014;10. https://doi.org/10.1371/journal.pcbi.1003457.

[9] Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS Comput Biol 2019. https://doi.org/10.1371/journal.pcbi.1006930.

[10] Haydon DT, Chase-Topping M, Shaw DJ, Matthews L, Friar JK, Wilesmith J, et al. The construction

and analysis of epidemic trees with reference to the 2001 UK foot-and-mouth outbreak. Proc R Soc B Biol Sci 2003. https://doi.org/10.1098/rspb.2002.2191.

[11] Cauchemez S, Boëlle PY, Donnelly CA, Ferguson NM, Thomas G, Leung GM, et al. Real-time estimates in early detection of SARS. Emerg Infect Dis 2006.

[12] Heijne JCM, Rondy M, Verhoef L, Wallinga J, Kretzschmar M, Low N, et al. Quantifying transmission of norovirus during an outbreak. Epidemiology 2012. https://doi.org/10.1097/EDE.0b013e3182456ee6.

[13] Kendall M, Ayabina D, Colijn C. Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees 2016:1–22. https://doi.org/10.1214/17-STS637.

[14] Worby CJ, O'Neill PD, Kypraios T, Robotham J V., De Angelis D, Cartwright EJP, et al. Reconstructing transmission trees for communicable diseases using densely sampled genetic data. Ann Appl Stat 2016. https://doi.org/10.1214/15-AOAS898.

[15] Lau MSY, Marion G, Streftaris G, Gibson G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. PLoS Comput Biol 2015. https://doi.org/10.1371/journal.pcbi.1004633.

[16] Spada E, Sagliocca L, Sourdis J, Garbuglia AR, Poggi V, De Fusco C, et al. Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis C virus infection. J Clin Microbiol 2004. https://doi.org/10.1128/JCM.42.9.4230-4236.2004.

[17] Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, et al. A bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. Proc R Soc B Biol Sci 2014. https://doi.org/10.1098/rspb.2013.3251.

[18] Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science (80-) 2014;345:1369--1372. https://doi.org/10.1126/science.1259657.

[19] Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014-2015 Ebola virus outbreak in West Africa. Nature 2015;524:97–101. https://doi.org/10.1038/nature14594.

[20] Ruan YJ, Wei CL, Ee LA, Vega VB, Thoreau H, Yun STS, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet 2003;361:1779–85. https://doi.org/10.1016/S0140-6736(03)13414-9.

[21] Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet 2009;10:540–50. https://doi.org/10.1038/nrg2583.

[22] Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science (80-) 2004. https://doi.org/10.1126/science.1090727.

[23] Campbell F, Strang C, Ferguson N, Cori A, Jombart T. When are pathogen genome sequences informative of transmission events? PLoS Pathog 2018. https://doi.org/10.1371/journal.ppat.1006885.

[24] Rota PA, Brown K, Mankertz A, Santibanez S, Shulga S, Muller CP, et al. Global distribution of measles genotypes and measles molecular epidemiology. J Infect Dis 2011;204. https://doi.org/10.1093/infdis/jir118.

[25] Hiebert J, Severini A. Measles molecular epidemiology: What does it tell us and why is it important? Canada Commun Dis Rep 2014;40:257–60. https://doi.org/10.14745/ccdr.v40i12a06.

[26] Brown KE, Rota PA, Goodson JL, Williams D, Abernathy E, Takeda M, et al. Genetic characterization of measles and rubella viruses detected through global measles and rubella elimination surveillance, 2016-2018. Morb Mortal Wkly Rep 2019;68:587–91. https://doi.org/10.15585/mmwr.mm6826a3.

[27] Gardy JL, Naus M, Amlani A, Chung W, Kim H, Tan M, et al. Whole-genome sequencing of measles virus genotypes H1 and D8 during outbreaks of infection following the 2010 Olympic Winter Games reveals viral transmission routes. J Infect Dis 2015;212:1574–8. https://doi.org/10.1093/infdis/jiv271.

[28] Penedos AR, Myers R, Hadef B, Aladin F, Brown KE. Assessment of the Utility of Whole Genome Sequencing of Measles Virus in the Characterisation of Outbreaks 2015:1–16. https://doi.org/10.1371/journal.pone.0143081.

[29] World Health Organisation. Measles virus nomenclature Update: 2012. Wkly Epidemiol Rec 2012;87:73–80. https://doi.org/10.1016/j.actatropica.2012.04.013.

[30] Hagemann C, Streng A, Kraemer A, Liese JG. Heterogeneity in coverage for measles and varicella vaccination in toddlers - Analysis of factors influencing parental acceptance. BMC Public Health 2017;17. https://doi.org/10.1186/s12889-017-4725-6.

[31] Glasser JW, Feng Z, Omer SB, Smith PJ, Rodewald LE. The effect of heterogeneity in uptake of the

measles, mumps, and rubella vaccine on the potential for outbreaks of measles: A modelling study. Lancet Infect Dis 2016;16:599–605. https://doi.org/10.1016/S1473-3099(16)00004-9.

[32] Gastañaduy PA, Budd J, Fisher N, Redd SB, Fletcher J, Miller J, et al. A Measles Outbreak in an Underimmunized Amish Community in Ohio. N Engl J Med 2016;375:1343–54. https://doi.org/10.1056/NEJMoa1602295.

[33] Woudenberg T, Van Binnendijk RS, Sanders EAM, Wallinga J, De Melker HE, Ruijs WLM, et al. Large measles epidemic in the Netherlands, May 2013 to March 2014: Changing epidemiology. Eurosurveillance 2017;22:1–9. https://doi.org/10.2807/1560-7917.ES.2017.22.3.30443.

[34] Keenan A, Ghebrehewet S, Vivancos R, Seddon D, MacPherson P, Hungerford D. Measles outbreaks in the UK, is it when and where, rather than if? A database cohort study of childhood population susceptibility in Liverpool, UK. BMJ Open 2017;7. https://doi.org/10.1136/bmjopen-2016-014106.

[35] Kucharski AJ, Edmunds WJ. Characterizing the Transmission Potential of Zoonotic Infections from Minor Outbreaks. PLoS Comput Biol 2015;11:1–17. https://doi.org/10.1371/journal.pcbi.1004154.

[36] Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 2008;5:0381–91. https://doi.org/10.1371/journal.pmed.0050074.

[37] Blumberg S, Lloyd-Smith JO. Inference of R0 and Transmission Heterogeneity from the Size Distribution of Stuttering Chains. PLoS Comput Biol 2013;9:1–17. https://doi.org/10.1371/journal.pcbi.1002993.

[38] Blumberg S, Enanoria WTA, Lloyd-Smith JO, Lietman TM, Porco TC. Identifying postelimination trends for the introduction and transmissibility of measles in the United States. Am J Epidemiol 2014;179:1375–82. https://doi.org/10.1093/aje/kwu068.

[39] Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: A modular platform for outbreak reconstruction. BMC Bioinformatics 2018;19. https://doi.org/10.1186/s12859-018-2330-z.

[40] Lenormand M, Bassolas A, Ramasco JJ. Systematic comparison of trip distribution laws and models. J Transp Geogr 2016;51:158–69. https://doi.org/10.1016/j.jtrangeo.2015.12.008.

[41] Zipf GK. The P 1 P 2/D hypothesis: On the intercity movement of persons. Am Sociol Rev 1946;11:677–86. https://doi.org/10.2307/2657358.

[42] Barthélemy M. Spatial networks. Phys Rep 2011;499:1–79. https://doi.org/10.1016/j.physrep.2010.11.002.

[43] Xia Y, Bjørnstad ON, Grenfell BT. Measles Metapopulation Dynamics: A Gravity Model for Epidemiological Coupling and Dynamics. Am Nat 2004;164:267–81. https://doi.org/10.1086/422341.

[44] Lenormand M, Huet S, Gargiulo F, Deffuant G. A Universal Model of Commuting Networks. PLoS One 2012;7. https://doi.org/10.1371/journal.pone.0045985.

[45] Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. Mach Learn 2003;50:5–43. https://doi.org/10.1023/A:1020281327116.

[46] Centers for Disease Control and Prevention (CDC). National Notifiable Disease Surveillance System: measles/rubeola 2013. https://wwwn.cdc.gov/nndss/conditions/measles/case-definition/2013/ (accessed October 23, 2019).

[47] Lessler J, Reich NG, Brookmeyer R, Perl TM, Nelson KE. Incubation periods of acute respiratory viral infections: a systematic review 2015;9:291–300. https://doi.org/10.1016/S1473-3099(09)70069-6.Incubation.

[48] Klinkenberg D, Nishiura H. The correlation between infectivity and incubation period of measles, estimated from households with two cases. J Theor Biol 2011;284:52–60. https://doi.org/10.1016/j.jtbi.2011.06.015.

[49] Fine PEM. The Interval between Successive Cases of an Infectious Disease. Am J Epidemiol 2003;158:1039–47. https://doi.org/10.1093/aje/kwg251.

[50] US Census Bureau. Centers of Population for the 2010 Census 2010. https://www.census.gov/geographies/reference-files/2010/geo/2010-centers-population.html (accessed August 22, 2019).

[51] Woudenberg T, Woonink F, Kerkhof J, Cox K, Ruijs WLM. The tip of the iceberg : incompleteness of measles reporting during a large outbreak in The Netherlands in 2013 – 2014. Epidemiol Infect 2018;146:716–22. https://doi.org/https://doi.org/10.1017/ S0950268818002698.

[52] Gastañaduy PA, Funk S, Paul P, Tatham L, Fisher N, Budd J, et al. Impact of public health responses during ameasles outbreak in an amish community in Ohio: Modeling the dynamics of transmission. Am J Epidemiol 2018. https://doi.org/10.1093/aje/kwy082.

[53] Patel M, Lee AD, Clemmons NS, Redd SB, Poser S, Blog D, et al. National Update on Measles Cases

and Outbreaks - United States, January 1-October 1, 2019. MMWR Morb Mortal Wkly Rep 2019;68:893–6. https://doi.org/10.15585/mmwr.mm6840e2.

[54] Zipprich J, Winter K, Hacker J, Xia D, Watt J, Harriman K. Measles outbreak--California, December 2014-February 2015. vol. 64. 2015. https://doi.org/10.1016/j.annemergmed.2015.04.002.

[55] Durrheim D. Measles elimination, immunity, serosurveys, and other immunity gap diagnostic tools. J Infect Dis 2018;218:341–3. https://doi.org/10.1093/infdis/jiy138.

[56] Prem K, Cook AR, Jit M. Projecting social contact matrices in 152 countries using contact surveys and demographic data. PLoS Comput Biol 2017. https://doi.org/10.1371/journal.pcbi.1005697.

Chapter 4. The impact of local vaccine coverage and recent incidence on measles transmission in France between 2009 and 2018

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	1704167	**Title**	Mr
First Name(s)	Alexis		
Surname/Family Name	Robert		
Thesis Title	Modelling the risks of measles outbreaks near elimination		
Primary Supervisor	Sebastian Funk		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?			
When was the work published?			
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	Choose an item.	Was the work subject to academic peer review?	Choose an item.

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	PLOS Medicine, MedRXiv (as preprint)
Please list the paper's authors in the intended authorship order:	Alexis Robert, Adam J Kucharski, Sebastian Funk
Stage of publication	**Submitted**

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	Conceptualization, Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

SECTION E

Student Signature	[redacted]
Date	31/05/2021

Supervisor Signature	[redacted]
Date	31/5/21

4.1. Abstract

4.1.1. Background

Despite high levels of vaccine coverage, sub-national heterogeneity in immunity to measles can create pockets of susceptibility, which are hard to detect and may result in long-lasting outbreaks. The elimination status defined by the World Health Organization aims to identify countries where the virus is no longer circulating and can be verified after 36 months of interrupted transmission. However, since 2018, numerous countries have lost their elimination status soon after reaching it, showing that the indicators used to define elimination may not be predictive of lower risks of outbreaks.

4.1.2. Methods and Findings

We quantified the impact of local vaccine coverage and recent levels of incidence on the dynamics of measles in each French department between 2009 and 2018, using mathematical models based on the 'Epidemic-Endemic' regression framework. High values of local vaccine coverage were associated with fewer imported cases and lower risks of local transmissions. Regions that had recently reported high levels of incidence were also at a lower risk of local transmission, potentially due to additional immunity accumulated during these recent outbreaks. Therefore, all else being equal, the risk of local transmission was not lower in areas fulfilling the elimination criteria (i.e., low recent incidence). After fitting the models using daily case counts, we used the parameters' estimates to simulate the effect of variations in the vaccine coverage and recent incidence on future transmission. A decrease of 3% in the three-year average vaccine uptake led to a five-fold increase in the number of cases simulated in a year on average.

4.1.3. Conclusions

Spatiotemporal variation in vaccine coverage because of disruption of routine immunisation programmes, or lower trust in vaccines, can lead to large increases in both local and cross regional transmission. The association found between local vaccine coverage and incidence suggests that, although regional vaccine uptake can be hard to collect and unreliable because of population movements, it can provide insights into the risks of imminent outbreak. Periods of low local measles incidence were not indicative of a decrease in the risks of local transmission. Therefore, the incidence indicator used to define the elimination status was not consistently associated with lower risks of measles outbreak in France. More detailed models of local immunity levels or subnational seroprevalence studies may yield better estimates of local risk of measles outbreaks.

4.2. Introduction

Immunity against infectious diseases accumulates following infection and, if a vaccine is available, routine immunisation programs and vaccination campaigns. Measles is highly infectious and can cause large outbreaks in populations with low immunity [1,2]. Therefore, high levels of vaccine coverage are required to minimise the risks of outbreaks [3]. Furthermore, vaccine uptake must be homogeneously high across the territory to avoid local transmission sustained by regional discrepancies [4,5]. The large-scale implementation of routine immunisation programs led to a drastic reduction in measles cases worldwide, and measles was targeted for elimination in five World Health Organization (WHO) Regions by 2020 under the Global Vaccine Action Plan 2011-2020 [6].

Elimination status, as defined by the WHO, refers to "the absence of endemic measles transmission for ≥12 months in the presence of a well-performing surveillance system" in a given country or region, and is verified "after 36 months of interrupted endemic measles virus transmission"[7]. Although imported cases, or cases directly related to importations could still be expected, there should be no continuous transmission persisting over a long period of time in a region where measles was eliminated. A given WHO region can declare measles eliminated when all countries in the region document interruption of endemic transmission for more than 36 months.

Recently, several countries had their elimination status revoked following large outbreaks less than five years after it was verified. For instance, the United Kingdom achieved elimination in 2017, and lost the status in 2019 along with Albania, Czechia, Greece, Venezuela, and Brazil [8,9]. In these countries, interruption of transmission during a few years was not indicative of reduced risks of major outbreaks. Such occurrences can be explained by several factors, such as a replenishment of susceptible individuals after years without transmission, or importations of cases into subnational areas with lower levels of immunity caused by heterogeneity in vaccine coverage [10–13]. The number and geographical distribution of the susceptible individuals is not routinely monitored in most countries given the perceived cost and logistical challenges of large serological surveys, yet it is a main predictor of outbreak risk [3]. Local values of vaccine coverage can be an alternative measure of heterogeneity, but they are not always available and can be outdated because of the mobility between regions. Furthermore, they only describe vaccine-induced immunity, and therefore ignore the immunity caused by previous outbreaks. In this study, we aim to i) estimate the impact of recent local transmission and local vaccine coverage on the current risk of outbreaks, and the changes in transmission dynamics that would results from variations in these factors, and ii) identify the areas most at-risk for local transmission using France as a case study.

To do so, we implemented an Epidemic-Endemic time-series model using *hhh4*, a framework developed by Held, Höhle and Hofmann to study the separate impact of covariates on importation, cross-regional transmission and local transmissions on aggregated case counts [14,15]. We adapted this framework to daily case counts and applied it to the daily number of measles cases per department (NUTS3 levels) in France reported to the European Center for Disease Prevention and Control (ECDC) between January 2009 and December 2018. We computed the average values of vaccine uptake and the number of cases per department in the past three years to mimic the timeframe used to define the elimination status, and modelled their impact on the local risks of outbreaks.

4.3. Methods

4.3.1. Description of the hhh4 framework

We used the modelling framework implemented in the "hhh4" model, which is part of the R package "*surveillance*"[15], to analyse infectious disease case counts. All the notations are defined in Table 4.. The expected number of cases ($\mu_{i,t}$) reported in the region i at time t depends on three sources of transmission (called "components"):

i. The *autoregressive* component ($\lambda_{i,t}$) represents the impact of $Y_{i,t-1}$, the number of cases in i at the previous time step, on the number of cases in i at t. The number of new cases expected from the autoregressive component is the product of predictors $\lambda_{i,t}$ and $Y_{i,t-1}$. A high value of $\lambda_{i,t}$ indicates that, if there are cases in i, there is potential for high transmission levels. On the other hand, if $\lambda_{i,t}$ is low, cases in i are unlikely to lead to much local transmission.

ii. The *neighbourhood* component ($\phi_{i,t}$) represents the impact of $Y_{j,t-1}$, the number of cases reported in regions around i at the previous time step, on the number of cases in i at t. The exact impact of cases in these regions on cases in i is determined by a distance matrix ω which quantifies the connectivity between the different regions. If $\phi_{i,t}$ is high, cases in regions around i are more likely to cause new cases in i, whereas a low value of $\phi_{i,t}$ indicates that cross regional transmissions towards i are less likely.

iii. The *endemic* component ($\nu_{i,t}$) represents the background number of new cases occurring in region i, regardless of the current number of cases in i, or in the regions around i. If $\nu_{i,t}$ is high, new cases in i are common, regardless of the number of cases in or around i at the previous time step. Since the endemic component does not depend on Y_{t-1}, it represents the background importations that cannot be linked to the mechanistic components. Therefore, these cases either correspond to importations from outside the modelled area (France in our case), or cases that are not otherwise predicted by the other two components.

The full equation for the expected number of cases in region i at time t is:

$$\mu_{i,t} = \nu_{i,t} + \lambda_{i,t} * Y_{i,t-1} + \phi_{i,t} * \sum_{j \neq i}(\omega_{ji} * Y_{j,t-1}) \quad (5)$$

The predictors $\lambda_{i,t}$, $\phi_{i,t}$ and $\nu_{i,t}$ are independently impacted by different covariates, i.e., a covariate may be associated with a reduction of importations, but have little impact on the spread of the virus within the region. We assume that $Y_{i,t}$, the number of observed cases at t in i, follows a negative binomial distribution to allow for overdispersion [16]. The overdispersion parameter ψ is estimated.

The predictors $\lambda_{i,t}$, $\phi_{i,t}$ and $\nu_{i,t}$ are estimated using log-linear regressions. For each predictor, we estimate: i) The intercept α (identical across spatial units), and ii) the vector of coefficients β associated with $z_{i,t}$ the vector of covariates at t in i included in each component.

$$log(\lambda_{i,t}) = \alpha^{(\lambda)} + \beta^{(\lambda)} * z_{it}^{(\lambda)} \quad (6)$$

$$log(\phi_{i,t}) = \alpha^{(\phi)} + \beta^{(\phi)} * z_{it}^{(\phi)} \quad (7)$$

$$log(\nu_{i,t}) = \alpha^{(\nu)} + \beta^{(\nu)} * z_{it}^{(\nu)} \quad (8)$$

Table 4.1: Table of notations of all variables and distributions defined in the methods.

Parameter	**Definition**
i, j	Regions
t	Time
$Y_{i,t}$	Number of cases reported in the region i at time t
$Y'_{i,t}$	Potential for transmission in the region i at time t
$\mu_{i,t}$	Average number of cases predicted in the region i at time t
λ	Autoregressive predictor
ϕ	Neighbourhood predictor
ν	Endemic predictor
ω	Connectivity matrix
α	Intercept
β	Vector of coefficients
z	Matrix of covariates
$f(t)$	Distribution of the serial interval
m_{it}	Number of inhabitants in the region i at time t
d_{ij}	Distance between regions i and j
γ, δ, ϵ	Parameters of the exponential gravity model
u_{it}	Average vaccine coverage in the region i at time t
n_{it}	Recent incidence per million in the region i at time t

N_{it}	Category of recent incidence in the region i at time t
s_{it}	Surface area of the region i at time t

4.3.2. Data

The observed case counts $Y_{i,t}$ was computed from 14,461 cases (10,988 confirmed and 3,473 probable cases) routinely collected in metropolitan France, and reported to the ECDC between January 2009 and December 2018 (Figure 4.1A). This data was retrieved on The European Surveillance System (TESSy) on 22 January 2019. The cases were stratified by the metropolitan department they were reported in. The department correspond to French NUTS3 regions. We excluded three cases where this information was not available. We used the date of symptom onset reported for each case to compute the daily number of cases from 2009 to 2018 per department.

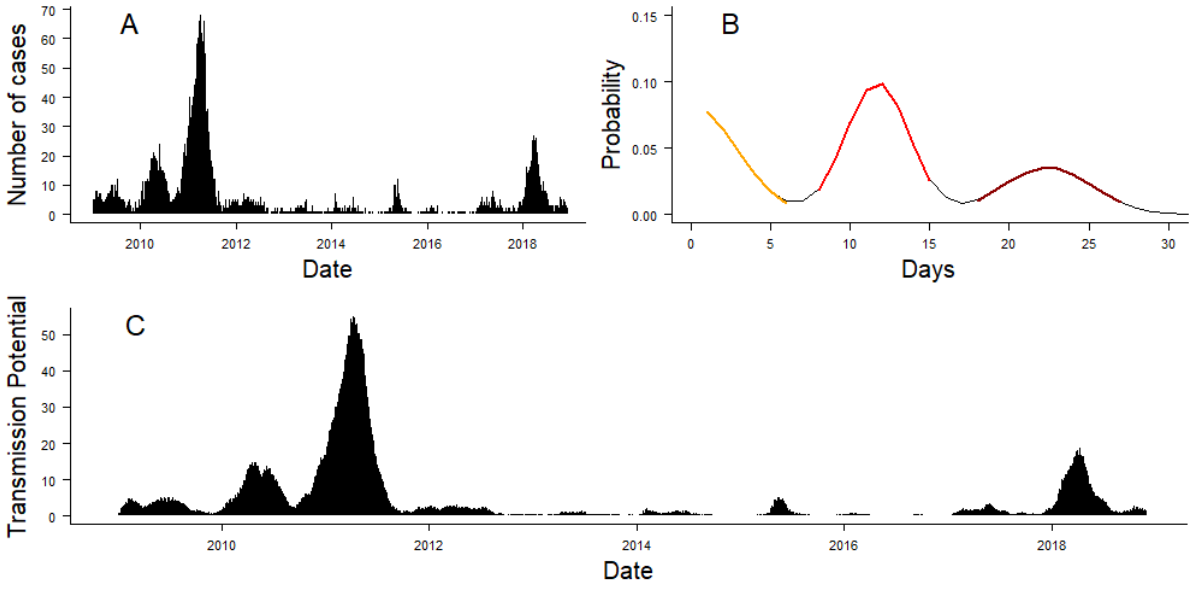

Figure 4.1: Panel A: Daily number of cases reported in France between 1st January 2009 and 30th November 2018. Panel B: Distribution of the composite serial interval used in the model. The different colours of the curve correspond to the three scenarios used to compute the distribution of the serial interval (orange: serial interval when missing ancestor; red: serial interval without unreported case, brown: serial interval when the case between the two reported cases was missing). Panel C: Transmission potential, which was computed by convolving the number of cases in the last 30 days with the composite serial interval.

4.3.3. Adaptation of hhh4 to daily case counts

In *hhh4*, the average number of new cases stemming from the autoregressive and neighbourhood components depends on the number of cases at the previous time step. Therefore, if we use daily case counts, the number of cases at t is only impacted by the number of cases the day before. In reality, however, the serial interval of measles is estimated to be 11 days on average [17]. Previous studies using $hhh4$ relied on temporally aggregated case counts, which partially solved this problem: if the time step is close to the average serial interval, cases of the same generation of transmission can be assumed to be roughly grouped together in the same time point [18]. Nevertheless, studying weekly (or fortnightly)

aggregated cases counts does not reflect the distribution of the serial interval (i.e., it ignores overlapping generations of transmission because of shorter or longer delays between primary and secondary cases). This can lead to directly connected cases being grouped in the same time step, or separated by more than one time step. This aggregation also ignores the potential for unreported cases, which may lead to cases causing transmission two to three weeks after their onset date via an intermediate, unobserved case. Finally, the starting date of aggregation influences how cases are grouped, which can lead to discrepancies in the parameter estimates.

Recent developments in the *surveillance* package included weight estimation to represent the relative impact of previous time steps on the number of cases at t [19]. Since we are using daily case counts, we set the weights of the different time steps from the distribution of the serial interval. We computed Y'_{it}, the transmission potential for each department and time step, by multiplying the number of recent cases by the distribution of the serial interval $f(t)$: $Y'_{it} = \sum_{k=1}^{50} Y_{i,t-k} * f(k)$. Only a subset of measles cases are reported to the surveillance system [20], therefore we accounted for the risks of unreported cases by computing a composite serial interval from three different transmission scenarios (Figure 4.1B):

1. In case of direct transmission between two cases i and j, the number of days between the two cases $f_1(t)$ follows a Normal distribution truncated at 0: $f_1(t) \sim N(11.7, 2)$ [17].
2. In case of unreported cases between i and j, the number of days between the two cases $f_2(t)$ follows a Normal distribution truncated at 0: $f_2(t) \sim N(23.4, \sqrt{8})$. This distribution corresponds to the convolution of $f_1(t)$ with itself.
3. If i and j share the same unreported index case, the number of days between i and j follows a half-Normal distribution (excluding 0) of standard deviation $\sqrt{8}$ days. This distribution corresponds to the distribution of the difference of $f_1(t)$ with itself, excluding values below 1. We added this last scenario to account for multiple concurrent importations stemming from an unreported infector.

We considered that 50% of the composite serial interval reflected direct transmission (scenario 1, without missing generations between cases), and 50% came from the two scenarios with unreported cases (scenarios 2 and 3). The distribution of the composite serial interval is shown in Figure 4.1B. We ran sensitivity analysis to estimate the parameters of the model using composite serial intervals computed with different proportions of direct transmission, and observed it had little influence on the estimation of each parameter (Supplement Section 1).

4.3.4. Connectivity between departments

In the *hhh4* framework, the average number of cases caused in the department i at time t by cases from another department j is quantified by the neighbourhood component. It is equal to $\phi_{i,t} * \omega_{ji} * Y_{j,t-1}$

(Equation 1). Therefore, the number of cases caused by cases from j in i in *hhh4* is influenced by three factors:

- The susceptibility of the department i, quantified by the neighbourhood predictor $\phi_{i,t}$, defined as $log(\phi_{i,t}) = \alpha^{(\phi)} + \beta^{(\phi)} * z_{it}^{(\phi)}$.
- The number of connections from j to i, calculated using an exponential gravity model [21], whereby the number of connections between i and j is proportional to the product of the number of inhabitants in the department of origin m_j, the department of destination m_i and an exponential decrease in the distance between i and j d_{ji}. Therefore, the number of connections from j to i was calculated as $w_{ji} = \mathrm{e}^{-\delta d_{ji}} m_{it}^{\gamma} m_{jt}^{\lambda}$.
- The proportion of the population in j that is infectious.

Therefore, the average number of cases expected from department j to department i at t can be written as the product of these three factors:

$$Y_{ji,t} = \exp\left(\alpha^{(\phi)} + \beta^{(\phi)} * z_{it}^{(\phi)}\right) * \mathrm{e}^{-\delta d_{ji}} m_{it}^{\epsilon} m_{jt}^{\gamma} * \frac{Y_{j,t-1}}{m_{jt}}$$

$$= \exp\left(\alpha^{(\phi)} + \beta^{(\phi)} * z_{it}^{(\phi)} * \epsilon * \log(m_{it})\right) * \frac{\mathrm{e}^{-\delta d_{ji}} m_{jt}^{\gamma}}{m_{jt}} * Y_{j,t-1}$$

Therefore, the log-population $\log(m_{it})$ was added as a covariate of the predictor of the neighbourhood component ϕ. The number of inhabitants in each French department between 2009 and 2018 was taken from the INSEE website [22].

We implemented two models with different methods to compute the distance between departments d_{ji}.

1. In Model 1, every department can be connected to each other, therefore only importations coming from outside the departments included in the study fall into the endemic component. The distance matrix was computed using the distance between the population centroids of each department, which were calculated using the $1km^2$ European Grid dataset [23]. This dataset contains the number of inhabitants in each grid cell covering the country (resolution 1km). We computed the weighted population centre in each department using the R function $zonal$ from the package raster[24] and calculated the distance between population centres.

$$Y_{ji,t} = \phi_{it} * \mathrm{e}^{-\delta d_{ji}} * \frac{m_{jt}^{\gamma}}{m_{jt}} * \mathrm{Y_{j,t-1}}$$

2. In Model 2, the neighbourhood component only takes into account transmission between neighbouring departments, assuming that cross-regional transmissions between non-neighbouring departments would be captured by the baseline number of daily importations (i.e. the endemic component):

$$Y_{ji,t} = \begin{cases} \phi_{it} * \frac{m_{jt}^{\gamma}}{m_{jt}} * \mathrm{Y_{j,t-1}} \text{ if } i \text{ and } j \text{ share a border} \\ 0 \text{ otherwise} \end{cases}$$

Therefore, the neighbourhood component in Model 1 includes both the neighbourhood component and part of the endemic transmission in Model 2.

4.3.5. Covariates

Different covariates can be added in each component of the *hhh4* framework [25]. We implemented the same set of covariates in the two models. The two covariates of interest were the impact of vaccine coverage and the category of incidence in each department in the past three years. We chose this timeframe in order to match the requirements of the elimination status assessment. We also included the number of inhabitants, the surface area of each department, and the seasonality as control variables, as explained below:

4.3.5.1. Vaccine coverage

For each department i and time step t, we computed $u_{i,t}$, the average proportion unvaccinated in the department i over the 3 years prior to t according to local coverage reports. We averaged over the past three years in order to use the same timeframe as the elimination status assessment. We used the yearly first dose uptake among 2-year-old children in each French department between 2006 and 2017. This data is publicly available on the website Santé Publique France [26–28]. The uptake of the second dose was not reported before 2010, and many departments had missing entries after 2010. Therefore, only the local coverage of the first dose was used in the model.

Since 26% of the entries in the coverage dataset were missing, we ran a beta mixed model to infer the missing values. We used the time and squared time (in years) as covariates, and random effects stratified by department. We used the average prediction to infer the missing values from the fitted model and get the complete vaccine coverage dataset. More details on the regression, and the sensitivity analyses that were run are presented in the Appendix (Supplement Section 2). All values of coverage in 2009 were missing, and were not imputed; we computed the average vaccine coverage in 2010, 2011, and 2012 using only two of the three previous years.

Adding the log-proportion of unvaccinated to the model was the most appropriate approach, since it allows the rate of disease spread (i.e. the value of the predictors λ, ν, and ϕ) to be proportional to the

density of susceptibles [25]. Therefore, we calculated the average log-proportion of unvaccinated in the three years before t and added it as a covariate in all three components.

4.3.5.2. Impact of recent incidence

This covariate quantifies the impact of past outbreaks on current transmission. Departments are eligible for WHO certification of elimination status if they have maintained low levels of transmission over the past three years [7]. Therefore, we computed $n_{i,t}$, the number of cases per million reported between a month and three years before t in i. We excluded cases reported in the last month since recent cases may be directly linked to current transmission.

$$n_{i,t} = 1{,}000{,}000 * \sum_{T>(t-30)}^{T<t-365*3} \frac{Y_{it}}{m_{it}}$$

We aggregated $n_{i,t}$ in three categories: i) $N_{i,t}^{(0)} = \begin{cases} 1 \; if \; n_{i,t} < 10 \\ 0 \; otherwise \end{cases}$: very limited transmission in recent years, department potentially eligible for elimination (30% of entries) ; ii)$N_{i,t}^{(1)} = \begin{cases} 1 \; if \; 10 \leq n_{i,t} < 45 \\ 0 \; otherwise \end{cases}$: Moderate transmission in recent years (36% of entries); iii) $N_{i,t}^{(2)} = \begin{cases} 1 \; if \; n_{i,t} \geq 45 \\ 0 \; otherwise \end{cases}$: major outbreak reported in the department in recent years. The threshold of 45 cases per million corresponds to the last tercile of $n_{i,t}$, hence 33% of $n_{i,t}$ fall into this last category.

Computing the level of recent incidence required the number of cases per department in the past three years. Therefore, since this analysis integrates case counts data from 2009, we needed to compute the incidence in each department between 2006 and 2008. Less than 50 cases were reported in France per year in 2006 and 2007 [29], therefore we considered their contribution to the recent level of incidence per department was null. On the other hand, 597 measles cases were reported to the ECDC in France in 2008, but were not stratified by department. Therefore, we used the number of cases reported per department in 2008 on Sante-Publique-France (597 cases overall, mostly reported in the second half of 2008 [30]) and integrated them in the computation of $N_{i,t}$ for $t < 2012$.

The level of recent incidence was a covariate in all three components.

4.3.5.3. Number of inhabitants and surface area

In the subsection “Connectivity between departments”, we discussed the impact of the number of inhabitants on the number of movements between departments. Furthermore, several studies have indicated a potential association between the population density and the number of secondary transmissions [31–33]. Therefore, we controlled for the impact of the number of inhabitants in each department, and the surface area (i.e., the geographical size) on the number of local transmissions.

The log-number of inhabitants $\log(m_{i,t})$ in the department i at time t was added as a covariate in all three components. The log-surface of the department $\log(s_{i,t})$ was added as a covariate in the autoregressive component.

4.3.5.4. Seasonality

We control for the impact of the seasonality of measles outbreaks in France on transmission by adding two covariates (sine-cosine) to all three components.

4.3.5.5. Full model equations for predictors

The covariates are all integrated in the covariate vectors in the equations 2, 3 and 4, yielding:

Autoregressive predictor: $\beta^{(\lambda)} z_{it}^{(\lambda)} = \beta_u^{(\lambda)} log(u_{i,t}) + \beta_{N^{(1)}}^{(\lambda)} N_{i,t}^{(1)} + \beta_{N^{(2)}}^{(\lambda)} N_{i,t}^{(2)} + \beta_m^{(\lambda)} \log(m_{i,t}) + \beta_s^{(\lambda)} \log(s_{i,t}) + \beta_{cos}^{(\lambda)} \cos\left(\frac{2\pi t}{365}\right) + \beta_{sin}^{(\lambda)} \sin\left(\frac{2\pi t}{365}\right)$

Neighbourhood predictor: $\beta^{(\phi)} z_{it}^{(\phi)} = \beta_u^{(\phi)} log(u_{i,t}) + \beta_{N^{(1)}}^{(\phi)} N_{i,t}^{(1)} + \beta_{N^{(2)}}^{(\phi)} N_{i,t}^{(2)} + \beta_m^{(\phi)} \log(m_{i,t}) + \beta_{cos}^{(\phi)} \cos\left(\frac{2\pi t}{365}\right) + \beta_{sin}^{(\phi)} \sin\left(\frac{2\pi t}{365}\right)$

Endemic predictor: $\beta^{(\nu)} z_{it}^{(\nu)} = \beta_u^{(\nu)} log(u_{i,t}) + \beta_{N^{(1)}}^{(\nu)} N_{i,t}^{(1)} + \beta_{N^{(2)}}^{(\nu)} N_{i,t}^{(2)} + \beta_m^{(\nu)} \log(m_{i,t}) + \beta_{cos}^{(\nu)} \cos\left(\frac{2\pi t}{365}\right) + \beta_{sin}^{(\nu)} \sin\left(\frac{2\pi t}{365}\right)$.

4.3.6. Model calibration

A model is deemed well-calibrated if it is able to correctly identify its own uncertainty in making predictions [34]. The most straightforward method to evaluate whether $hhh4$ models are well-calibrated is to generate a one-step-ahead forecast over a chosen test period and compare them with the data [15]. Since we use daily case counts, this method would only assess the ability of the models to capture the number of cases on the next day. We explored the calibration of our models several days ahead. To do so, we selected the last two years of data as the test period, fit the model up to each day, and simulated the number of cases over the next 3, 7, 10 and 14 days for each day of the test period in each department. For each date, we ran at least 100,000 simulations. If the number of cases observed in the data had not been generated in 100,000 simulations, we ran simulations until it was reached.

From these simulations, we generated the predictive probability distribution at each time step in each department. In a model with perfect calibration, the actual number of cases follows the predictive probability distribution ($\mu_{it} \sim P_{it}$ for all predictive distributions P_{it}), i.e., the probability integral transform (PIT) histogram is uniform. We computed the PIT histograms in both models for predictions over 3, 7, 10, and 14 days. The PIT histograms were computed using a non-randomised yet uniform

version of the PIT histogram correcting for the use of discrete values described in Czado et al [35] and implemented in *hhh4*.

The PIT histograms were used to estimate whether the short-term forecasts were in line with the data, and whether the models were consistently missing some scenarios of transmission.

4.3.7. Simulation study

In order to highlight the impact of variations in the local vaccine coverage or the level of recent transmission on the risks of outbreaks, we generated simulations of the number of cases in France across one year under different conditions. To compute these simulations, we used the last values of average vaccine coverage (the average was computed from the values in 2015, 2016, and 2017) and the levels of recent incidence in mid-2018, and simulated the daily number of cases between the 1st of August 2018 and the 31st of December 2019. We started the simulations during the period of the year associated with the lowest number of cases (i.e., on the 1st of August), in order to avoid biases. Indeed, if we had used the last three months of data (until November 2018), some departments may have been repeatedly associated with higher numbers of cases in our simulations, not because they are more at risk of importation or transmission, but because there had been cases reported in these departments at the beginning of the epidemic year. We were only interested in highlighting the impact of variations in coverage and recent transmission, rather than predicting the level of transmission for the entire year of 2019.

We generated 100 samples of the regression coefficients using the variance-covariance matrix and assumed they followed a multivariate normal distribution. For each sample, we computed the values of the three predictors between the 1st of August 2018 and the 31st of December 2019, and simulated the daily number of cases in each department across the year. We ran 100 simulations per sample (i.e. 10,000 simulations were generated per scenario).

We studied four scenarios: i) Using the latest local values of coverage (averaged over the past three years), population and category of recent incidence, ii) Increasing the vaccination coverage in each department by three percent, iii) Decreasing the vaccination coverage in each department by three percent, and iv) setting the recent incidence in each department to minimal levels (i.e. conditions fulfilling the WHO elimination status requirements).

Finally, since tourism and local events can lead to mass gatherings and trigger repeated importations independent of parameters included in the model [36,37], we studied the impact of repeated local importations of cases into specific departments. To do so, we simulated one year of transmission (i.e., until the end of 2019) following the importations of 10 cases in a given department in December 2018.

In these simulations, we did not allow for any other baseline importations throughout the year, in order to assess the potential for geographical spread throughout the country after importation in one department.

4.4. Results

4.4.1. Impact of the covariates on each component

The parameter estimates obtained in both models are shown in Figure 4.2. Values above 0 show aggravating effects associated with an increase in the number of expected cases at the next time step. For both models, departments with a high proportion unvaccinated in the past three years were associated with a higher number of expected cases in the autoregressive (Model 1: 0.14 [0.03 - 0.24] ; Model 2: 0.19 [0.09 - 0.29]) and the endemic component (Model 1: 0.37 [-0.17 - 0.91] ; Model 2: 0.48 [0.17 - 0.80]). This indicates that these departments were at higher risks of background importations, and secondary transmission upon importation. In both components, the effect of vaccination was slightly stronger in Model 2, where cross-regional transmission is restricted to neighbouring departments, than in Model 1, where cross-regional transmission can happen between all departments, although the confidence intervals overlapped. In Model 1, the proportion unvaccinated also had an aggravating effect on the number of cross-departmental transmissions (0.47 [0.23 - 0.71]), whereas in Model 2 there was no clear association between the proportion unvaccinated and an increase in cross-regional transmission (-0.02 [-0.29 - 0.25]). The differences between the models' coefficients were due to the cross-regional transmission in Model 1 corresponding to both the neighbourhood component and some of the endemic transmission in Model 2.

The association between the level of incidence over the past three years (parameters: $immun\,1$ and $immun\,2$ in Figure 4.2) and the components of transmission was similar in both models. In the auto-regressive component, departments that reported high incidence over the past three years ($immun\,2$) were associated with fewer secondary cases per case in the department (Model 1: -0.15 [-0.23 - -0.08]; Model 2: -0.13 [-0.20 - -0.06]). This could be linked to outbreak-induced immunity causing a depletion of susceptibles in departments where incidence was high over the past few years. On the other hand, the parameters associated with $immun\,2$ were above 0 in the neighbourhood and endemic components, which indicates that departments with high incidence in the past three years were more at risk of cross-regional transmission and background importations (Model 1: Endemic 0.89 [0.50 – 1.27]; Neighbourhood: 0.25 [0.09 – 0.41]; Model 2: Endemic 0.67 [0.46 – 0.89]; Neighbourhood: 0.31 [0.11 – 0.51]). The parameter $immun\,1$ was only significantly different from 0 in the endemic component (Model 1: 0.66 [0.22 – 1.10]; Model 2: 0.57 [0.34 – 0.80]), meaning departments that

recently reported moderate levels of transmission were associated with more background importations, but no difference was noticeable in cross-regional or within-region transmission.

The other covariates included in the model showed that the number of inhabitants in a department had an important impact on both the endemic and neighbourhood components: departments with more individuals were more likely to report background importations and cross-regional transmission. On the other hand, the population and the surface area of the departments had no impact on the autoregressive component. We also observed a strong impact of seasonality on the three components (Figure 4.2). Indeed, the peak values of the predictors were 20 to 37% higher than the average value in all components of transmission (Supplement Section 3). The peak of the autoregressive component was in February for both models, the endemic peak was in May for Model 1 (April in Model 2), whereas the neighbourhood component peaked in December in Model 1 (March in Model 2).

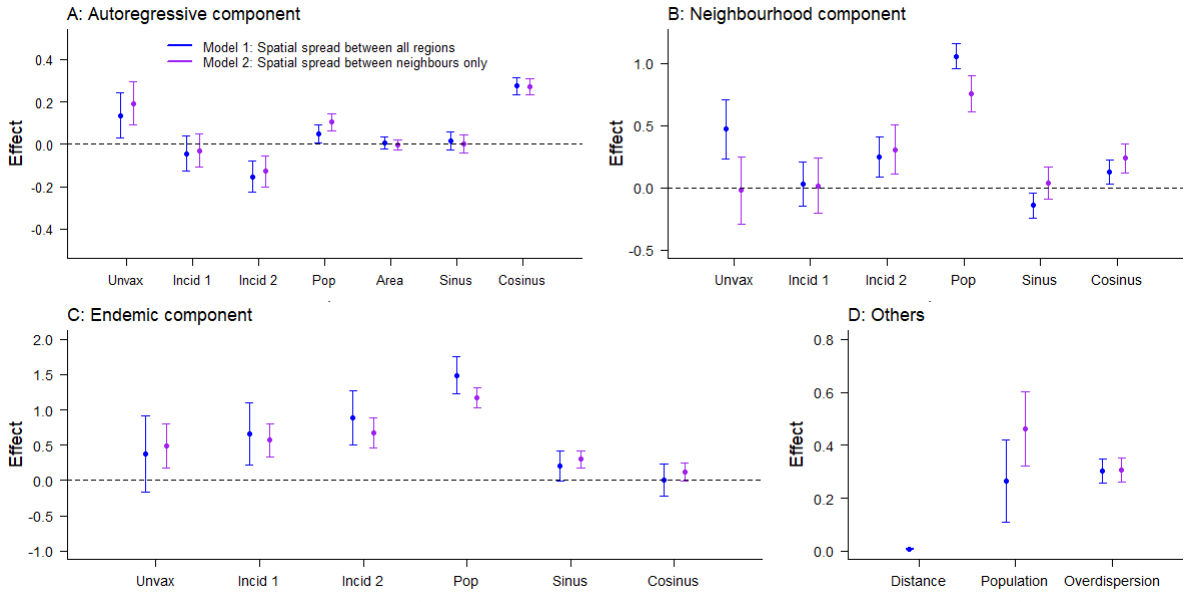

Figure 4.2: Estimates of the parameters in each component of Model 1 (blue) and Model 2 (purple): Panel A: Autoregressive component; Panel B: Neighbourhood component; Panel C: Endemic component; Panel D: Other coefficients. The y-axis. $unvax$ *corresponds to the effect of* $u_{i,t}$*, the mean proportion unvaccinated over the three years before* t *in* i*;* $incid1$ *and* $incid2$ *correspond to the effect of* $N_{i,t}^1$ *and* $N_{i,t}^2$ *the category of incidence in the three years before* t *in* i*;* pop *corresponds to the effect of* $m_{i,t}$ *the number of inhabitants at* t *in* i*;* $area$ *corresponds to the effect of the surface;* sin *and* cos *correspond to the effects of seasonality;* $distance$ *and* $population$ *correspond to the spatial parameters of the connectivity matrix* w *(*δ *and* γ*);* $overdisp$ *is the estimate of the log-overdispersion parameter in the negative binomial distribution of* $Y_{i,t}$*. Dots show the mean values associated with the parameters; arrows show the 95% Confidence interval. Note different y-axes between graphs.*

Using the mean parameter estimates, and the latest values of vaccination coverage, incidence, and number of inhabitants per department, we computed the local predictors ϕ_i, λ_i, and ν_i in both models to highlight the spatial heterogeneity of the transmission risks (Figure 4.3). The predictors were computed ignoring the impact of seasonality, which does not change the geographic distribution of risks since it is not region-dependent in the models. Therefore, the maps correspond to the average local value of the predictors the year following the last data entry (i.e. the 30th of November 2018). The

geographic distributions of the autoregressive predictor are similar in Model 1 and Model 2. This indicates that the same departments were classified as having higher risks of local transmission in both models. Areas with lower values of vaccine uptake such as the South East and South West of France were associated with higher risks of secondary transmission. Indeed, the highest values of within-region transmission were reported in Bouches-du-Rhône and Var (in the South East of France). Populous departments in the North of France were also at risk of secondary transmission despite higher vaccination coverage.

As expected, the overall number of baseline importations in Model 1 was lower than in Model 2, which was compensated by a higher number of cross-regional transmissions (Figure 4.3). This shows that some of the cases that could not be linked to local transmission, or transmission between neighbouring departments in Model 2, were classified as cross-regional transmissions in Model 1, which would indicate long-distance transmission events. In both models, departments with a higher number of inhabitants were most at-risk of cross-regional and baseline importations, which corresponds to the strong association between the number of inhabitants and the endemic and neighbourhood components highlighted in Figure 4.2. Departments like Bouches-du-Rhône that combine a high number of inhabitants with low vaccine coverage were associated with the highest number of baseline and cross-regional importations in both models. The variations in the autoregressive component were smaller than in the importation-related components: For instance, the highest autoregressive predictor value (Var: 0.81 [0.74 - 0.88]) was 35% higher than the lowest value (Lozère: 0.60 [0.53 – 0.66]) in Model 1, whereas the number of baseline importations in Bouches-du-Rhônes was more than 100 times above the number of importations in Lozère (South of France). This can be explained by the coefficients of the autoregressive components being much closer to 0 than the most extreme coefficients in the importation-related components (Figure 4.2).

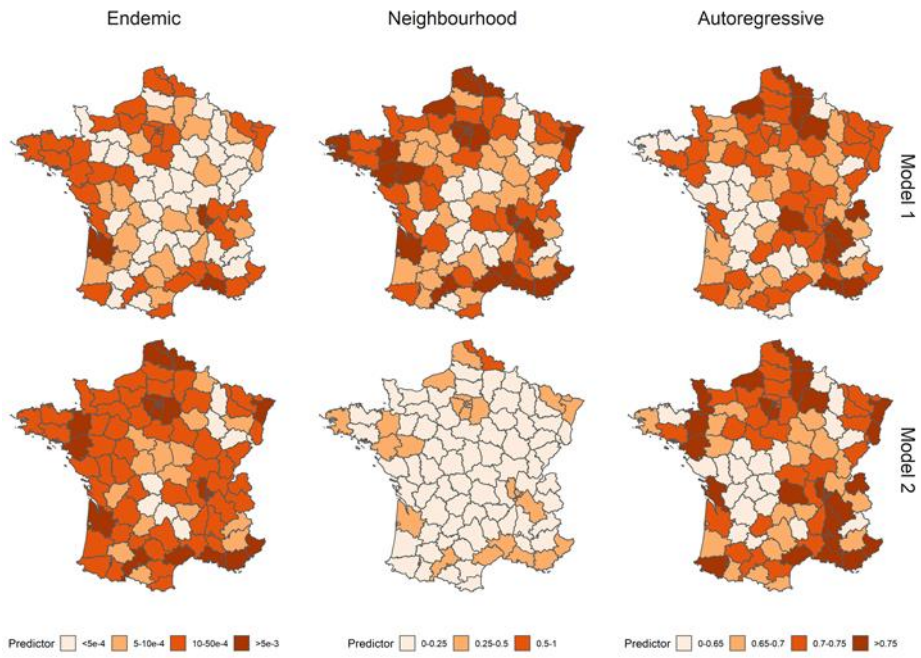

Figure 4.3: Average values of the endemic, neighbourhood, and autoregressive predictors per department in Model 1 (upper row) and Model 2 (lower row) over the year 2019. Since the absolute values are expected to vary over the year because of seasonality, the panels show the relative geographical heterogeneity. The endemic predictor corresponds to the number of importations per day per department, whereas the autoregressive predictor corresponds to the number of secondary cases per case in each department. The absolute value of the neighbourhood predictor is harder to interpret directly since it is multiplied by the connectivity matrix in the equation. Higher values were associated with departments with higher risks of observing cases following population movements.

4.4.2. Model fit and calibration

The daily and weekly fits of Model 1 and Model 2 indicate that they were able to match the transmission dynamics observed in France between 2009 and 2017, despite wide variations in the annual number of cases (Figure 4.4 Panel A and B, Supplement Section 4). In years where active transmission was reported, most of the cases stemmed from the autoregressive component, indicating that the local outbreaks were sustained by transmission within the departments. Indeed, across all years, the autoregressive component accounted for 72.9% of the cases, whereas 23.7% of the cases came from cross-regional transmission, and 3.4% from the endemic component (Supplement Figure S12). This shows that in Model 1, 97.6% of the cases were explained by the transmission stemming from other cases reported in the dataset (93.2% in Model 2). The endemic component described the minority of isolated cases that could not be linked to any concurrent transmission cluster. Therefore, these cases would be more likely to be reported at times of low national levels of transmission when no other case could be linked to them, which explains the shift in seasonality of the endemic component observed in Figure 4.2 and Supplement Section 3.

In order to visually assess the calibration of the model, and its ability to provide reliable short-term predictions for the number of cases per department, we generated PIT histograms showing the

probability integral transform obtained when forecasting the number of cases 3, 7, 10, and 14 days ahead (Figure 4.4, Panels C to F). The PIT histogram is uniform for predictions 3 and 7 days ahead (all groups are above 0.9 and below 1.1), which shows the number of occurrences where the predictions of the model did not capture the number of cases one week ahead was not higher than expected under a uniform distribution. As we increased the number of days of forecast, there were more occurrences of the model mis-predicting the number of cases to come. Indeed, the U-shape observed in Panel F of Figure 4.4 indicates the model was less capable of identifying extreme events two weeks in advance. The calibration study indicated that Model 2 was more prone to under-estimating the number of cases than Model 1, and showed signs of bias for the 7, 10, and 14-day predictions (Supplement Section 4). The national number of cases predicted by Model 1 and Model 2 were similar, and match the data for predictions 7 days ahead (Supplement Figure S11). The AIC scores and the calibration study indicated Model 1 was able to fit the data better than Model 2 and was better calibrated. The rest of the Results section therefore focuses on the conclusions reached using Model 1. The equivalent analysis run on Model 2 is presented in the Supplementary Section 4.

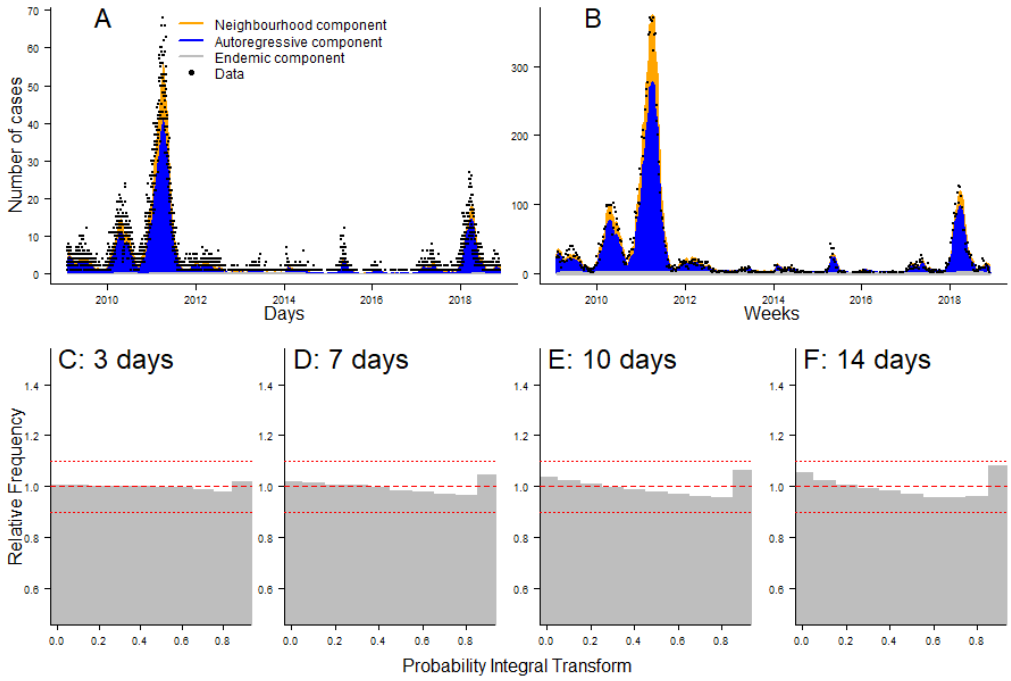

Figure 4.4: Panel A and B: Daily and weekly fit between the data and Model 1. The inferred number of cases is split among the three components of the model. Panel C to F: PIT histograms of Model 1, generated respectively for predictions 3, 7, 10, and 14 days ahead.

4.4.3. Impact of vaccination and recent incidence on onwards transmission

In order to illustrate the impact of recent outbreaks and variations in vaccine coverage on the transmission dynamics in France, we generated 10,000 simulations and computed the number of cases per department in 2019. We ran the simulation from August 2018 (during the historically low

transmission season), until 31st December 2019. We generated four sets of simulations under different initial conditions: using the last measures of average local vaccine coverage, category of recent incidence, and number of inhabitants; increasing or decreasing the vaccine coverage by three percent, and setting the category of recent incidence to 0 in each department.

Under the latest measures of coverage and incidence, the simulated outbreaks display a wide variation in the number of cases in 2019 (minimum 100 cases, median 1,100 cases, maximum 11,100 cases). Active transmission was generated in a wide range of departments. Indeed, across the simulation set, 44 of the 94 French departments reported more than 10 cases in at least 25% of the simulations. There was noteworthy spatial heterogeneity in the levels of incidence. Indeed, in 12 departments, there was no case generated in more than half of the simulations (Figure 4.5, top right panel). The departments most vulnerable to active transmissions were highly populated urban areas, such as Paris, the Bouches-du-Rhône, and the North of France. Because they are highly populated, these departments were susceptible to repeated importations (they reported at least 1 case in more than 95% of the simulations), which could then cause large transmission clusters. This was especially evident in the South-East of France, where we highlighted that the number of secondary cases per case in the department was among the highest in the country (Figure 4.3 and Figure 4.5). Numerous departments were affected by large outbreaks in a subset of the simulated datasets: 27 departments reported more than 50 cases in at least 5% of the simulations (Figure 4.5). Further, at least one major outbreak was generated in the majority of the simulations: in 55% of the simulations, one department reported more than 100 cases (the most commonly affected department were Paris and its surroundings, the Nord, and Bouches-du-Rhône).

Decreasing the average three-year vaccine coverage by three percent led to an important increase in the number of cases per outbreak (median 4,900 cases, more than 95% of the simulations resulted in more than 1,000 cases). This was first due to an increase in importations and cross-regional transmission: all 94 departments had at least one case in more than half of the simulations, 77 in at least 90% of the simulations. Furthermore, the decrease in vaccination coverage resulted in higher chances of uncontrolled transmissions in many departments (Figure 4.5, third row). On the other hand, increasing the vaccine coverage by three percent caused an important drop in the number of cases (median 605 cases, 80% of the simulations generated less than 1,000 cases), caused by both a decrease in the number of importations, and in the potential for secondary transmission following importations. Although outbreaks were still punctually generated, these events are much rarer than in the other two simulation sets: in 25.8% of the simulations, at least one department generated more than 100 cases (54.1% with the baseline scenario, 95.4% when we reduced the local vaccine coverage).

Finally, setting the local recent incidence to the minimum level in each department, which would fulfil the elimination guidelines, had two opposite effects: it led to a decrease in the number of importations and cross-regional transmission, and an increase in the number of infections within each department (Figure 4.2). In this simulation set, the number of departments where no cases were generated in more than half of the simulations was similar to when the vaccine coverage was increased (24 departments in this simulation set, 29 when the vaccine coverage was increased, Figure 4.5), which shows the reduction in the number of cross-regional transmission and background importations. Conversely, the number of large outbreaks was only marginally inferior to the reference simulation set: in 44% of the simulations, there were more than 100 cases generated in at least one department (54% in the reference dataset). The geographical distribution of the risks of large outbreaks was almost identical to the reference simulation set (Figure 4.5). Therefore, although the number of importations was reduced, changing the level of recent incidence did not have a clear impact on the risks of active transmission. More departments became vulnerable to secondary transmission, and despite importations in these departments being rarer, they were more likely to lead to large outbreaks when they happened. The two opposing effects recent incidence had on importation and transmission therefore created a different dynamic of transmission observed in the simulation set, without strongly reducing the risks of outbreaks.

Each of these simulation sets highlighted the wide range of scenarios that could be generated using the parameter distributions inferred by our model. In order to gain more understanding on the spatial spread and consequences of importations, we then explored the impact of localised repeated importations on overall transmission.

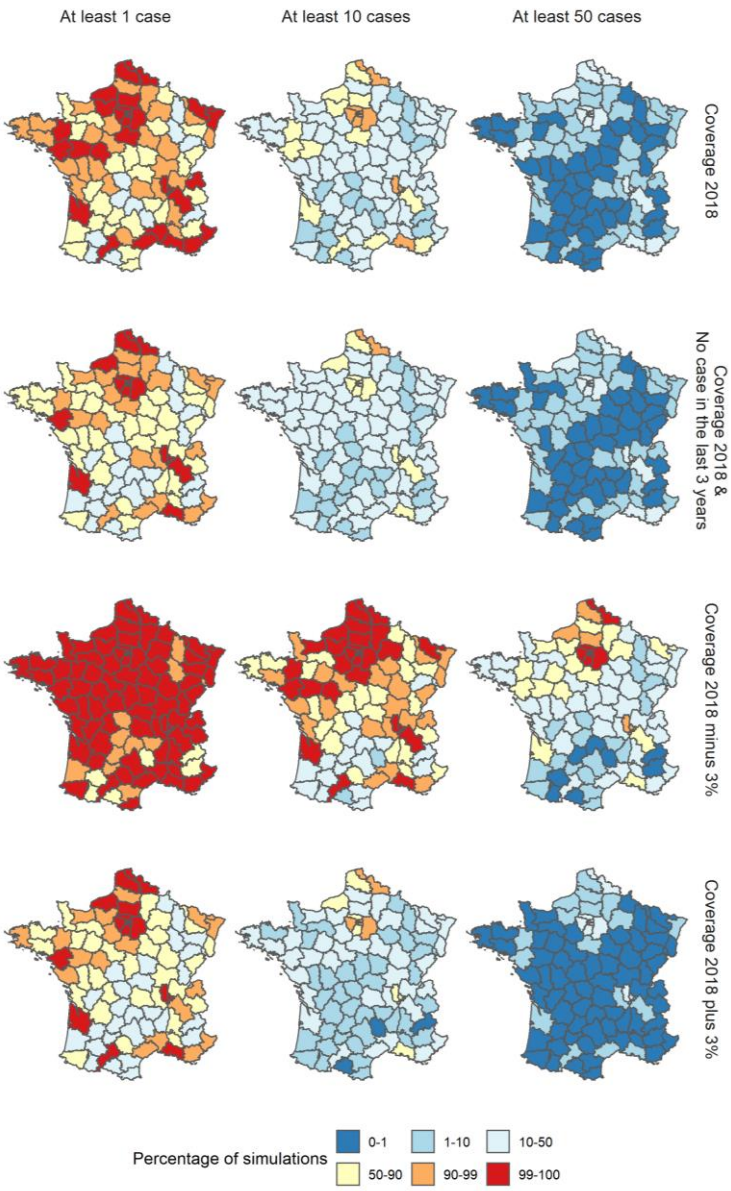

Figure 4.5: Percentage of simulations where the number of cases reported in each department in 2019 was at least 1, 10, and 50 cases for each scenario using parameter estimates from Model 1. Each row corresponds to a different scenario: i) Reference, ii) Minimum level of recent incidence in each department, iii) Local vaccine coverage decreased by three percent in each department, iv) Local vaccine coverage increased by three percent in each department.

4.4.4. Impact of local clusters of transmission

Since the endemic component, which can be interpreted as external importations, represented a minority of the cases in our model (Supplement Figure S12), repeated importations in a given department over a short timespan rarely occurred in the simulations. Furthermore, due to the seasonality of the endemic component, fewer importations are generated early in December to February, which corresponds to the peak period of the other components, and would therefore be more likely to cause secondary transmissions (Supplement Section 3). We simulated one year of transmission following ten importations in December 2018 to illustrate: i) the potential for local outbreaks, and ii)

the spatial spread of transmission following repeated local importations. We selected four departments to compare the impact of repeated importations in a range of settings: Paris (many inhabitants, 91% vaccine coverage, surrounded by urban areas), Bouches-du-Rhône (many inhabitants, 84% vaccine coverage), Haute Garonne (many inhabitants, 91% vaccine coverage but high levels of recent incidence, surrounded by rural areas with lower vaccine coverage), and Gers (Rural area, 79% vaccine coverage) (Figure 4.6).

Firstly, major local outbreaks in the department of importation were generated in all four simulation sets, and especially in Paris and Bouches-du-Rhône, where the proportion of simulations that yielded more than 100 subsequent cases in the department was 40% and 39%, respectively. In the Bouches-du-Rhône, large outbreaks were mostly due to the low vaccination coverage, whereas in Paris, outbreaks were mostly linked to the connectivity to nearby areas and the high number of inhabitants, which meant the department was likely to attract cross-regional transmissions. Major local outbreaks were rarer in the other two scenarios (9% of simulations above 100 in Haute Garonne, 10% in Gers). The lower proportion of large outbreaks resulted from different factors: recent large outbreaks in Haute Garonne reduced the autoregressive predictor, lowering the number of secondary cases per case imported; whereas since Gers is a rural department, with a low number of inhabitants, almost all the local cases were due to local transmission (auto-regressive component), with very few cross-regional transmissions into Gers.

Conversely, the simulations where cases were imported in Gers yielded the largest spatial spread throughout the country: the median number of departments that reported at least 1 case was 53 (16 when the importations were generated in Haute Garonne; 15 in Bouches-du-Rhône; 39 in Paris). As stated in the method, the number of cross-regional transmissions is the product of the predictor and the connectivity matrix, divided by the number of inhabitants in the department of origin, to represent that only a fraction of commuters will be infected. Therefore, populous areas are more likely to attract cross-regional transmissions, whereas more rural departments are more likely to seed outbreaks in other areas. The relatively high spatial spread when cases were imported in Paris is due to the short distance between Paris and its suburbs, which is then more likely to cause cross-regional transmission in the northern departments. Despite the cross-regional spread observed in both of these simulations sets, outbreaks remained local, and occurrences of nation-wide outbreaks were almost null. The departments most at risk of outbreak following cross-regional spread were some of the direct neighbours of the department of importations, or the large urban areas (Figure 4.6). To further explore this, we ran the same simulations decreasing the vaccine coverage by three percent, which greatly increased the number of departments exposed in each simulation set, and increased the risk of local transmission (Supplement Section 6). Therefore, although repeated importations could cause active

transmission in and around the departments of importation, the current values of vaccine coverage and the seasonality of transmission were able to prevent nationwide transmission.

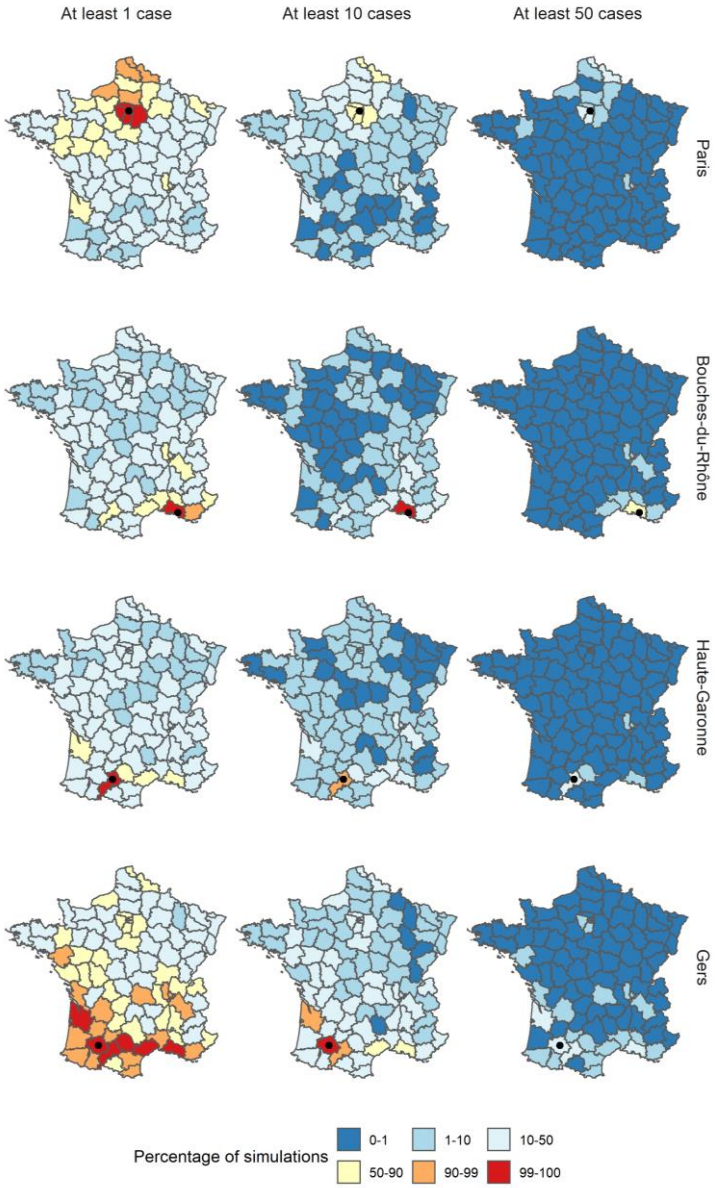

Figure 4.6: Percentage of simulations where the number of cases reported in each department in 2019 was at least 1, 10, and 50 cases following the importations of ten cases in December 2018, and using the parameter estimates from Model 1. For each row, the department of importation is indicated by a black dot.

4.5. Discussion

This analysis explored which local factors were associated with high risks of transmission in France over the last decade. Since 2017, immunity gaps, caused by failures to vaccinate, have been linked to a resurgence of measles in all WHO regions [38]. In countries near elimination, large outbreaks have been linked to heterogeneity in the levels of immunity, with pockets of susceptibles fuelling punctual outbreaks despite high national vaccine uptake [1,2,4,25]. Our study showed that local values of vaccine coverage were linked to lower transmission, whereas lower levels of recent incidence were not

associated with lower risks of local transmission. Furthermore, we highlighted that a drop of 3% in the three-year vaccine coverage triggered a five-fold increase in the number of cases simulated in a year.

The fact that higher vaccine coverage was associated with a lower number of secondary cases is consistent with prior expectations, and would confirm that the local values of first dose vaccine coverage are a good indicator of the actual immunity in the population and risks of future transmission. Reporting accurate values of local vaccine coverage is challenging, for instance because the vaccination status of people moving regions can be hard to track and lead to measurement errors. Furthermore, we did not have access to complete data on the coverage of the second MMR dose, which would be a better indicator of vulnerable areas. Therefore, detecting the association between recent vaccine uptake and incidence is encouraging. The impact of local vaccination coverage on transmission may also be muddled by sub-regional vaccine heterogeneity. For instance, pockets of susceptibles within a region, i.e. areas within the region where the vaccine coverage is substantially lower than the regional average, may be at high risk of transmission and would not be observable in regional coverage [39]. This phenomenon can only be hypothesised here, and could be explored using local data on incidence and vaccine uptake at a sub-regional scale.

Variations in vaccine coverage had a noticeable impact on the number of cases generated in the simulation study. We showed the effects of a three percent increase and decrease of the three-year average vaccine coverage on the number of cases, which highlighted the risks of uncontrolled transmission in the event of a decrease of vaccine-induced protection. Events such as the disruption caused by the SARS-COV-19 pandemic on routine measles vaccination campaigns could therefore highly increase the risks of uncontrolled measles transmission in the years to come [40,41].

The departments that reported few cases per million in the past three years were associated with higher risks of local transmission (autoregressive component). Therefore, according to our model, regions eligible for elimination status were not associated with lower risks of onwards transmission. Conversely, high levels of recent transmission were associated with a lower number of cross-regional transmissions and importations, although we cannot methodologically establish the causality of this association. The impact on the simulation study was clear: when we set the category of recent incidence to the lowest level, departments were less exposed to cases, and spatial spread was rarer, whilst there was little change in the risks of major outbreaks. The simulations showed an ‘all-or-nothing’ situation: departments tended to report very few to no cases, whilst also being more likely to be affected by outbreaks. These results would indicate that looking into the level of incidence to quantify the future risks of outbreaks can be deceptive, and importations in a department with low recent incidence would result in large transmission clusters.

We proposed a new framing of the Epidemic-Endemic model implemented in *hhh4* by adapting it to daily count data using the distribution of the serial interval to compute the local transmission potential. Using daily case counts allowed us to avoid biases associated with aggregated case counts, such as the influence of the arbitrary aggregation date, by accounting for the impact of variation in the serial intervals. We also accounted for the risks of unreported cases by computing a composite multimodal serial interval, thus allowing for transmission with a missing generation, or an unreported ancestor. The model was able to capture the dynamic of transmission better than the 10-day aggregated model, as shown by the calibration study (Supplement Section 7). Nevertheless, our framing of the *hhh4* model introduced new biases: we used a distribution of the serial interval based on previous studies rather than estimating the weights during the fitting procedure and set the proportion of missing generations in the composite serial interval. We explored the impact of the proportion of missing generations by fitting the model with different composite serial intervals and concluded that the impact of each covariate was robust to these changes (Supplement Section 1). We also integrated a potential day-of-the-week effect, and observed that although it had an impact on the auto-regressive component, it did not change the estimates of the other parameters, and therefore did not change the conclusions of the study (Supplement Section 8).

Using the *hhh4* model allowed us to analyse the different impact of various covariates on local and cross-regional transmission, and background importation of cases. According to the models we implemented, an overwhelming majority (>90%) of the transmission came from the cross-regional and local components of the regression. This indicates that in the models, the endemic component only corresponds to rare background cases that could not be linked to concurrent transmission events. This could point towards model misspecifications, for example, connecting unrelated importations to concurrent local transmission. Since endemic transmission tends to refer to cases otherwise unexplained by the mechanistic components, the seasonality of the endemic component is decoupled from the other components, i.e. endemic cases are likely when local and cross-regional transmission are lower.

Since the endemic component accounted for such a small minority of the cases, group importations of cases in a given department were rarely observed in the simulations. However, tourism, and local events lead to large gatherings and can increase the risks of group importations in a limited period of time [36,37]. We simulated the spatial spread following repeated importations in a given department, and highlighted that although large outbreaks in the department of importations were common, nation-wide transmission following these importations was very rare. Only the departments where all cases had been imported, and its neighbours, were at risk of uncontrolled outbreaks. Decreasing the level of vaccination by three percent was associated with a large increase in the level of exposure of all

departments, and in the number of departments where large outbreaks were generated (Supplement Section 6 and 7). The high levels of transmission observed in recent years in France suggest that importations are frequent, and even a small drop in vaccination could dramatically increase measles transmission in the country.

Furthermore, since the number of inhabitants was strongly associated with risks of background importations, most of the endemic importations were reported in urban areas, where the risks of exportations were lower. This could explain the discrepancies between the distribution of the number of cases in the simulations (Figure 4.5, top row), and the actual number of cases reported in France in 2019 [42]. Active transmission was reported in a number of rural areas, notably in the South West of France, and in Savoie (East). This could be due to importations and cross-regional transmission that are under-estimated by our model. Although the model captured the dynamics seen in the data, the calibration study showed it was only able to predict short-term transmission up to one week. The PIT histogram associated to the 14-day calibration displayed signs of bias, which shows that the model was not able to consistently predict variations in the future number of cases in the next two weeks. We identify several factors that could explain the discrepancies observed for longer term predictions: i) the indicator of local immunity we used was flawed: two-dose coverage would be a better indicator of the proportion of the population that is protected; ii) The sub-regional heterogeneity in coverage and past incidence within the department that could be concealed by NUTS3 aggregated data: because of social groups that rarely mix with one another, or large NUTS regions, large outbreaks in a given community would not be a good indicator of the overall level of immunity in a region. Nevertheless, we believe that the results obtained using limited publicly available covariates are encouraging and we intend to apply this method using more complete data.

We identified a number of limitations of this study that have not yet been mentioned: Firstly, potential reactive control measures in case of high transmission were not accounted for. It is likely that if the level of incidence was increasing over a short period of time, control measures would be implemented and the behaviour of the individuals may change (e.g. school closures, catch-up vaccination campaigns). This could impact the number of expected cases after a certain threshold is passed, and impact the dynamics in the simulated outbreaks. Secondly, we did not include information on the age or genotype of the cases. Therefore, unrelated importations in successive time-steps in a given region may be considered as linked by our model, whereas they should be separated. Further development of this method could focus on taking this aspect into account, in order to give information on the number of independent concurrent chains. Thirdly, since this is not a transmission model, some extreme values could trigger unlikely behaviour. For instance, if the vaccination rate would be 100%, we would still expect sporadic transmission. Although this would not be entirely implausible given that only the vaccination coverage

in the past three years was taken into account in the models (i.e. even if it was 100% coverage, there could be susceptible individuals in different age groups). Finally, the impact of the different covariates on the number of cases was constant through time. For instance, the impact of seasonality may depend on factors such as the weather which may vary each year, which would not be accounted for in the model we developed.

We used variables collected in a wide range of settings (regional vaccine coverage, incidence, number of inhabitants, surface), therefore this analysis can be reproduced in other countries to analyse the potential for local transmission as well as the impact of recent incidence and vaccine-induced immunity. Since the case counts data are not publicly available, we share the code used to generate the analysis applied to a simulated dataset on a Github repository: (https://github.com/alxsrobert/measles-regional-transmission).

4.6. Disclaimer

The views and opinions of the authors expressed herein do not necessarily state or reflect those of ECDC. The accuracy of the authors' statistical analysis and the findings they report are not the responsibility of ECDC. ECDC is not responsible for conclusions or opinions drawn from the data provided. ECDC is not responsible for the correctness of the data and for data management, data merging and data collation after provision of the data. ECDC shall not be held liable for improper or incorrect use of the data.

4.7. References

[1] Gastañaduy PA, Budd J, Fisher N, Redd SB, Fletcher J, Miller J, et al. A Measles Outbreak in an Underimmunized Amish Community in Ohio. N Engl J Med 2016;375:1343–54. doi:10.1056/NEJMoa1602295.

[2] Woudenberg T, Van Binnendijk RS, Sanders EAM, Wallinga J, De Melker HE, Ruijs WLM, et al. Large measles epidemic in the Netherlands, May 2013 to March 2014: Changing epidemiology. Eurosurveillance 2017;22:1–9. doi:10.2807/1560-7917.ES.2017.22.3.30443.

[3] Funk S, Knapp JK, Lebo E, Reef SE, Dabbagh AJ, Kretsinger K, et al. Combining serological and contact data to derive target immunity levels for achieving and maintaining measles elimination. BMC Med 2019. doi:10.1186/s12916-019-1413-7.

[4] Glasser JW, Feng Z, Omer SB, Smith PJ, Rodewald LE. The effect of heterogeneity in uptake of the measles, mumps, and rubella vaccine on the potential for outbreaks of measles: A modelling study. Lancet Infect Dis 2016;16:599–605. doi:10.1016/S1473-3099(16)00004-9.

[5] Keenan A, Ghebrehewet S, Vivancos R, Seddon D, MacPherson P, Hungerford D. Measles outbreaks in the UK, is it when and where, rather than if? A database cohort study of childhood population susceptibility in Liverpool, UK. BMJ Open 2017;7. doi:10.1136/bmjopen-2016-014106.

[6] World Health Organization (WHO). Global Vaccine Action Plan Global Vaccine Action Plan. Who 2011:4–7.

[7] World Health Organization. Framework for verifying elimination of measles and rubella. Wkly Epidemiol Rec 2013;88:89–100. doi:10.1371/jour.

[8] World Health Organization (WHO). European Region loses ground in effort to eliminate measles 2019.

[9] Pan American Health Organization. Epidemiological Update: Measles. Paho/ Who 2019;2020:1–12.

[10] Fraser B. Measles outbreak in the Americas. Lancet (London, England) 2018;392:373. doi:10.1016/S0140-6736(18)31727-6.

[11] Litvoc MN, Lopes MIBF. From the measles-free status to the current outbreak in Brasil. Rev Assoc Med Bras 2019;65:1229–30. doi:10.1590/1806-9282.65.10.1129.

[12] Dimala CA, Kadia BM, Nji MAM, Bechem NN. Factors associated with measles resurgence in the United States in the post-elimination era. Sci Rep 2021;11:1–10. doi:10.1038/s41598-020-80214-3.

[13] Bernadou A, Astrugue C, Méchain M, Le Galliard V, Verdun-Esquer C, Dupuy F, et al. Measles outbreak linked to insufficient vaccination coverage in Nouvelle-Aquitaine region, France, October 2017 to July 2018. Eurosurveillance 2018;23:1–5. doi:10.2807/1560-7917.ES.2018.23.30.1800373.

[14] Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Stat Modelling 2005;5:187–99. doi:10.1191/1471082X05st098oa.

[15] Meyer S, Held L, Höhle M. hhh4: Endemic-epidemic modeling of areal count time series. J Stat Softw 2016.

[16] Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355–9. doi:10.1038/nature04153.

[17] Fine PEM. The Interval between Successive Cases of an Infectious Disease. Am J Epidemiol 2003;158:1039–47. doi:10.1093/aje/kwg251.

[18] Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of measles epidemics: Estimating scaling of transmission rates using a Time series SIR model. Ecol Monogr 2002;72:169–84. doi:10.1890/0012-9615(2002)072[0169:DOMEES]2.0.CO;2.

[19] Bracher J, Held L. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. Int J Forecast 2020. doi:10.1016/j.ijforecast.2020.07.002.

[20] Woudenberg T, Woonink F, Kerkhof J, Cox K, Ruijs WLM. The tip of the iceberg : incompleteness of measles reporting during a large outbreak in The Netherlands in 2013 – 2014. Epidemiol Infect 2018;146:716–22. doi:https://doi.org/10.1017/ S0950268818002698.

[21] Lenormand M, Bassolas A, Ramasco JJ. Systematic comparison of trip distribution laws and models. J Transp Geogr 2016;51:158–69. doi:10.1016/j.jtrangeo.2015.12.008.

[22] Institut National de la Statistique et des Etudes Economiques. Estimation de la population au 1er janvier 2020 2020. https://www.insee.fr/fr/statistiques/1893198#consulter (accessed September 7, 2020).

[23] Eurostat. European population grid cells 2011. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/grids (accessed September 12, 2020).

[24] Hijmans RJ, Etten J van, Mattiuzzi M, Sumner M, Greenberg JA, Lamigueiro OP, et al. Package "raster." R 2014.

[25] Herzog SA, Paul M, Held L. Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. Epidemiol Infect 2011;139:505–15. doi:10.1017/S0950268810001664.

[26] Santé Publique France. Données départementales 2007-2012 de couverture vaccinale rougeole, rubéole, oreillons à 24 mois 2019. https://www.santepubliquefrance.fr/determinants-de-sante/vaccination/articles/donnees-departementales-2007-2012-de-couverture-vaccinale-rougeole-rubeole-oreillons-a-24-mois (accessed September 7, 2020).

[27] Santé Publique France. Estimations des couvertures vaccinales à 24 mois à partir des certificats de santé du 24e mois, 2004-2007 2010. https://www.santepubliquefrance.fr/determinants-de-sante/vaccination/articles/donnees-departementales-2013-2017-de-couverture-vaccinale-rougeole-rubeole-oreillons-a-24-mois (accessed September 7, 2020).

[28] Santé Publique France. Données départementales 2013-2017 de couverture vaccinale rougeole, rubéole, oreillons à 24 mois 2019. https://www.santepubliquefrance.fr/determinants-de-

sante/vaccination/articles/donnees-departementales-2013-2017-de-couverture-vaccinale-rougeole-rubeole-oreillons-a-24-mois (accessed September 7, 2020).

[29] Antona D, Lévy-Bruhl D, Baudon C, Freymuth F, Lamy M, Maine C, et al. Measles elimination efforts and 2008-2011 outbreak, France. Emerg Infect Dis 2013;19:357–64. doi:10.3201/eid1903.121360.

[30] Institut de Veille Sanitaire. Données de déclaration obligatoire de la rougeole. 2009.

[31] Fitzpatrick G, Ward M, Ennis O, Johnson H, Cotter S, Carr MJ, et al. Use of a geographic information system to map cases of measles in real-time during an outbreak in Dublin, Ireland, 2011. Eurosurveillance 2012;17:1–11. doi:10.2807/ese.17.49.20330-en.

[32] Yang W, Wen L, Li SL, Chen K, Zhang WY, Shaman J. Geospatial characteristics of measles transmission in China during 2005−2014. PLoS Comput Biol 2017;13:1–21. doi:10.1371/journal.pcbi.1005474.

[33] Andrianou XD, Del Manso M, Bella A, Vescio MF, Baggieri M, Rota MC, et al. Spatiotemporal distribution and determinants of measles incidence during a large outbreak, Italy, september 2016 to july 2018. Eurosurveillance 2019;24:1–12. doi:10.2807/1560-7917.ES.2019.24.17.1800679.

[34] Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area Region of Sierra Leone, 2014–15. BioRxiv 2017:1–17. doi:10.1101/177451.

[35] Czado C, Gneiting T, Held L. Predictive Model Assessment for Count Data 2009:1254–61. doi:10.1111/j.1541-0420.2009.01191.x.

[36] le Polain de Waroux O, Saliba V, Cottrell S, Young N, Perry M, Bukasa A, et al. Summer music and arts festivals as hot spots for measles transmission: Experience from England and Wales, June to October 2016. Eurosurveillance 2016;21:1–6. doi:10.2807/1560-7917.ES.2016.21.44.30390.

[37] Gautret P, Steffen R. Communicable diseases as health risks at mass gatherings other than Hajj: What is the evidence? Int J Infect Dis 2016;47:46–52. doi:10.1016/j.ijid.2016.03.007.

[38] Patel MK, Goodson JL, Alexander JP, Kretsinger K, Sodha S V, Steulet C. Progress Toward Regional Measles Elimination — Worldwide , 2000 – 2019 2020;69:1700–5.

[39] Blumberg S, Enanoria WTA, Lloyd-Smith JO, Lietman TM, Porco TC. Identifying postelimination trends for the introduction and transmissibility of measles in the United States. Am J Epidemiol 2014;179:1375–82. doi:10.1093/aje/kwu068.

[40] Saxena S, Skirrow H, Bedford H. Routine vaccination during covid-19 pandemic response. BMJ 2020;369. doi:10.1136/bmj.m268.

[41] Dinleyici EC, Borrow R, Safadi MAP, van Damme P, Munoz FM. Vaccines and routine immunization strategies during the COVID-19 pandemic. Hum Vaccines Immunother 2021;17:400–7. doi:10.1080/21645515.2020.1804776.

[42] Santé Publique France. Bulletin épidémiologique rougeole. Données de surveillance 2019. 2020. https://www.santepubliquefrance.fr/content/download/231366/2508985 (accessed May 3, 2021).

Chapter 5. Impact of aggregation on the Epidemic-Endemic framework: A simulation study

5.1. Introduction

Reports of outbreaks from national routine surveillance often come in the form of aggregated spatio-temporal data, i.e. they contain the number of cases per region and per unit of time. These publicly available reports are typically weekly, or fortnightly aggregated. Therefore, the information they give on transmission dynamics is coarser than daily case reports, although these are harder to access. Publicly available data are key to understanding how a given virus can spread in a variety of settings. Therefore, mathematical models have been adapted to use aggregated data and to incorporate them in the study of outbreak dynamics [1–3]. However, aggregating incidence data is a simplification that may lead to biases, for instance infectors and infectees may be grouped together at the same time point, and a dependence on the starting date of aggregation can introduce bias. Hence there is a need to assess whether the outputs of models using daily surveillance data are robust if applied to aggregated data.

The Epidemic-Endemic framework, implemented in the R package *surveillance*, was introduced by Held et al. to study infectious disease counts [4], and was further developed in various studies [3,5–8]. It uses aggregated case counts to analyse the number of cases resulting from three different components: the number of importations, local transmission, and cross-regional transmission in each region. It has repeatedly been applied to aggregated measles case counts in countries near elimination, for instance highlighting the impact of heterogeneity in regional coverage on measles transmission in Germany [6]. The Epidemic-Endemic framework commonly computes the average number of cases expected using the number of cases per region at the previous time step. This assumes that the aggregation mirrors the typical number of days between the infector and an infectee's reported onset dates (i.e. the serial interval), and therefore ignores the variability of these intervals and, consequently, the impact of unreported generations between two connected cases or intervals that are shorter than the temporal resolution of reporting. The date of aggregation also influences which cases are clustered together, and which generation each case belongs to. Therefore, using aggregated data may influence how the transmission parameters are estimated, and may introduce biases.

In the fourth chapter of this thesis, I adapted the Epidemic-Endemic framework to daily case counts, describing the risks of transmission over time using the distribution of measles serial interval [9]. In this chapter, I aimed to look into the impact of using aggregated data on the parameter fits and the calibration of the model. To do so, I generated a set of simulated outbreaks, using France as a case

study, and the parameter distribution estimated in Chapter 4. Daily and 10-day aggregated Epidemic-Endemic models were then fitted to the simulated outbreaks. Finally, the outputs were compared to assess whether both sets of models were able to estimate the right parameter values, and whether using daily data significantly improved the calibration of the model.

5.2. Methods

5.2.1. Summary of the Epidemic-Endemic framework

The Epidemic-Endemic framework (also called "hhh4"), introduced in Chapter 4, is used to model the regional case counts Y_{it}. To do so, the model assumes that Y_{it} follows a negative binomial distribution with mean μ_{it}, the expected number of cases at the time step t in the region i, which depends on three sources of transmission, also called components:

- The autoregressive component (predictor λ_{it}) represents the number of secondary cases originating from Y_{it-1}, the number of cases in i at the previous time step.
- The neighbourhood component (predictor ϕ_{it}) represents the number of cases originating from Y_{jt-1} , the number of cases in $j = 1..n$ ($j \neq i$), the regions around i, at the previous time step. The connectivity between the different regions is quantified by the connectivity matrix ω_{ij}.
- The endemic component (predictor ν_{it}) represents the background number of cases expected at t in i that are not linked to recent transmission in or around i.

The full equation for the expected number of cases in region i at time t is:

$$\mu_{i,t} = \nu_{i,t} + \lambda_{i,t} * Y_{i,t-1} + \phi_{\mathrm{i,t}} * \sum_{j \neq i} (\omega_{ji} * Y_{j,t-1}) \quad (9)$$

The predictors $\lambda_{i,t}$, $\phi_{i,t}$ and $\nu_{i,t}$ are estimated using log-linear regressions. Each predictor contains the intercept α (identical across spatial units), and the vector of coefficients β associated with $z_{i,t}$ (the vector of covariates in each component at t in i).

$$log(\lambda_{i,t}) = \alpha^{(\lambda)} + \beta^{(\lambda)} * z_{it}^{(\lambda)} \quad (10)$$

$$log(\phi_{i,t}) = \alpha^{(\phi)} + \beta^{(\phi)} * z_{it}^{(\phi)} \quad (11)$$

$$log(\nu_{i,t}) = \alpha^{(\nu)} + \beta^{(\nu)} * z_{it}^{(\nu)} \quad (12)$$

5.2.2. Generation of simulated outbreaks

I generated 100 simulated datasets, containing the daily number of cases in each of the 94 French metropolitan departments (i.e. NUTS3 regions) between January 2009 and December 2017. All the functions used to generate and analyse the outbreaks, and generate the plots of the analysis, are available on the Github repository *https://github.com/alxsrobert/measles-regional-immunity*. The

outbreaks were simulated using the equations (1) to (4), using the same set of covariates as in Chapter 4, namely:

1. The 3-year average local vaccine coverage: the first dose local vaccination uptake at 2 years-old is available on the website Santé Publique France [10–12]. I retrieved and collated the local data, and inferred the missing entries using a beta mixed model. For more details on the inference methods, and the coverage data, see Section 1 in Chapter 4's Appendix. The functions used to generate the coverage data are in the script *function_coverage.R* on the aforementioned Github repository. The log-proportion of coverage $\log(\mu_{it})$ was added as a covariate in all three components.
2. The category of incidence: This covariate quantifies the impact of local transmission in the past three years on current transmission. Firstly, I computed n_{it}, the number of cases per million reported in each region between a month (to exclude current transmission) and three years before the current time step. The local incidence is then split into three categories: i) $N_{i,t}^{(0)} = \begin{cases} 1 \; if \; n_{i,t} < 10 \\ 0 \; otherwise \end{cases}$: very limited transmission in recent years; ii) $N_{i,t}^{(1)} = \begin{cases} 1 \; if \; 10 \leq n_{i,t} < 45 \\ 0 \; otherwise \end{cases}$: Moderate transmission in recent years; iii) $N_{i,t}^{(2)} = \begin{cases} 1 \; if \; n_{i,t} \geq 45 \\ 0 \; otherwise \end{cases}$: major outbreak reported in the department in recent years. The thresholds were chosen to match the parametrisation used in Chapter 4. At the beginning of the simulation (1st January 2009), the level of incidence was set to the minimum in all regions, i.e. $N_{i0}^{0} = 1, \;\; N_{i0}^{1} = 0$, and $N_{i0}^{2} = 0$, and was then computed at each time step of the simulations. The category of incidence was added as a covariate in all three components.
3. The number of inhabitants: The population per French department between 2009 and 2018 m_{it} is publicly available on the INSEE website [13]. The functions used to generate the population data are in the script *function_distance_population.R* on the Github repository. The log-number of inhabitants was added as a covariate in all three components.
4. The surface area of each region was also computed using the INSEE data [13]. The log-surface of each region was added to the auto-regressive component.
5. The seasonality was generated using two covariates (sine-cosine functions), and added to all three components.

The connectivity between regions was computed using an exponential gravity model, where every region was connected to one another [14]. The connectivity between two regions i and j was computed as follows: $w_{ij} = \mathrm{e}^{-\delta d_{ji}} * \frac{m_{jt}^{\gamma}}{m_{jt}}$, where d_{ji} is the distance in kilometres between i and j, and m_{jt} the number of inhabitants in j and t. The values of the parameters δ and γ were taken from the final

parameter estimates in Model 1. The distance between departments was calculated from the population centroids of each department, which were computed using the $1km^2$ European Grid dataset [15]. This dataset contains the number of inhabitants in each grid cell covering France (resolution 1km). The weighted population centre was calculated using the R function zonal from the package *raster* [16].

For each outbreak, the set of parameters was drawn using Monte Carlo simulations. These simulations were generated from the parameter means and the covariance matrix estimated with Model 1 in Chapter 4, assuming that the estimates follow a multivariate normal distribution (Table 5.2).

In the default aggregated version of the Epidemic-Endemic framework, the number of cases generated at t by the autoregressive and neighbourhood components solely depends on the number of cases at $t-1$. In the daily version of the framework, the transmission potential was computed by convoluting the number of daily cases in the last month with the distribution of the serial interval $f(t)$: $Y'_{it} = \sum_{k=1}^{50} Y_{i,t-k} * f(k)$. The average serial interval for measles was taken to be 11 days, following a 0-truncated normal distribution: $f_1(t) \sim N(11.7, 2)$ [9]. Therefore, in the simulated outbreaks the transmission potential was computed at each iteration in every region.

The full equation of the average number of new cases at t in i was generated from Equation (1), replacing Y_{it} by the transmission potential Y'_{it}, and with the predictors as follows:

Autoregressive predictor: $\beta^{(\lambda)} z_{it}^{(\lambda)} = \beta_u^{(\lambda)} log(u_{i,t}) + \beta_{N^{(1)}}^{(\lambda)} N_{i,t}^{(1)} + \beta_{N^{(2)}}^{(\lambda)} N_{i,t}^{(2)} + \beta_m^{(\lambda)} \log(m_{i,t}) + \beta_s^{(\lambda)} \log(s_{i,t}) + \beta_{cos}^{(\lambda)} \cos\left(\frac{2\pi t}{365}\right) + \beta_{sin}^{(\lambda)} \sin\left(\frac{2\pi t}{365}\right)$

Neighbourhood predictor: $\beta^{(\phi)} z_{it}^{(\phi)} = \beta_u^{(\phi)} log(u_{i,t}) + \beta_{N^{(1)}}^{(\phi)} N_{i,t}^{(1)} + \beta_{N^{(2)}}^{(\phi)} N_{i,t}^{(2)} + \beta_m^{(\phi)} \log(m_{i,t}) + \beta_c^{(\phi)} \cos\left(\frac{2\pi t}{365}\right) + \beta_s^{(\phi)} \sin\left(\frac{2\pi t}{365}\right)$demic predictor: $\beta^{(\nu)} z_{it}^{(\nu)} = \beta_u^{(\nu)} log(u_{i,t}) + \beta_{N^{(1)}}^{(\nu)} N_{i,t}^{(1)} + \beta_{N^{(2)}}^{(\nu)} N_{i,t}^{(2)} + \beta_m^{(\nu)} \log(m_{i,t}) + \beta_c^{(\nu)} \cos\left(\frac{2\pi t}{365}\right) + \beta_s^{(\nu)} \sin\left(\frac{2\pi t}{365}\right)$

The number of new cases at each day and region was generated using a negative binomial distribution, where the mean was computed from Equation (1), and the overdispersion parameter was taken from the parameter set generated in Chapter 4. The mean value of each parameter is listed in Table 5.2. In order to assess the impact of partial reporting of cases on the models' performances, 30% of the cases were removed to account for incomplete reporting of cases, which is common during measles outbreaks. The removed cases were chosen at random. In real-life outbreaks, the proportion of missing cases could vary depending on the number of cases in the area, and the capacities of the surveillance system, this was not accounted for in this example, since I only aimed to quantify how Epidemic-Endemic models can adjust to some level of partial reporting.

Table 5.2: Mean values of the parameters used to generate the simulation set. For each simulation, the parameter set was drawn using the covariance matrix.

Parameter	Symbol	Mean value, autoregressive component	Mean value, neighbourhood component	Mean value, endemic component
Intercept	α	-0.018	7.54	-5.83
Coefficient associated with the log proportion unvaccinated	β_u	0.14	0.47	0.37
Coefficient associated with moderate levels of recent transmission	β_{N^1}	-0.044	0.031	0.66
Coefficient associated with high levels of recent transmission	β_{N^2}	-0.15	0.25	0.89
Coefficient associated with the log number of inhabitants	β_m	0.048	1.06	1.49
Coefficient associated with the surface of the department	β_s	0.0057	Not included	Not included
Coefficient associated with the cosine function of seasonality	β_{cos}	0.27	0.12	0.0082
Coefficient associated with the sine function of seasonality	β_{sin}	0.016	-0.14	0.21
Parameter quantifying the impact of the number of inhabitants on the connectivity between regions	γ	Not applicable	0.27	Not applicable
Parameter quantifying the impact of the distance on the connectivity between regions	δ	Not applicable	0.0079	Not applicable
Overdispersion	ϕ	3.31		

5.2.3. Fitting and evaluating the models

After each simulation, a daily and an aggregated model were fitted to the daily case counts reported per region, i.e. the number of local cases after accounting for the report rate. Both models integrated the same covariates, listed in the previous section (i.e. coverage, incidence, population, surface, seasonality). The thresholds of the incidence categories were not the same as the categories used in the simulations. Indeed, the threshold for high level of incidence in the simulations was defined as 45 cases

per million in the past 3 years, which is similar to the categories used in Chapter 4 and corresponds to the second incidence tercile in the French data. To match the protocol used in Chapter 4, the threshold describing the highest level of incidence in the model was computed using the second tercile of all measures of recent incidence, for each simulation. Therefore, the thresholds differ for each iteration. In the daily model, I accounted for potential missing generations between cases by using a composite serial interval accounting for three scenarios:

1. In case of direct transmission between two cases i and j, the number of days between the two cases $f_1(t)$ follows a Normal distribution truncated at zero: $f_1(t) \sim N(11.7, 2)$.
2. In case of unreported cases between i and j, the number of days between the two cases $f_2(t)$ follows a Normal distribution truncated at zero: $f_2(t) \sim N\left(23.4, \sqrt{8}\right)$, i.e. the convolution of $f_1(t)$ with itself.
3. If i and j share the same unreported index case, the number of days between i and j follows a half-Normal distribution (truncated at zero) of standard deviation $\sqrt{8}$ days, which is equivalent to the distribution of the difference of $f_1(t)$ with itself, excluding values below 1.

As in Chapter 4, 50% of the composite serial interval reflected direct transmission (scenario 1, without missing generations between cases), and 50% came from the two scenarios with unreported cases (scenarios 2 and 3). The aggregated model used a 10-day aggregation, which should correspond to the optimal aggregation for measles outbreaks, given it is close to the average serial interval.

The aggregated and non-aggregated models were compared using different characteristics: Firstly, their ability to accurately estimate the parameters, i.e. whether the mean estimates in a given simulation set were close to the means from the input data, and whether the input parameter fell into the 95% confidence interval of the parameter distribution. The estimated parameters were compared to their "true" values used for the simulations using Probability Integral Transform (PIT) histograms, which were generated by computing the quantile of the input parameter in the cumulative distribution of the fitted distribution [17]. The fitted distribution was calculated using a normal distribution, with the means and standard deviation estimated by the Epidemic-Endemic models.

Secondly, I compared the predictive ability of the two models for each simulation using a strictly proper scoring rule, which is a common way to assess predictive ability when compared to data [18]. To compute the forecast scores of both models, I generated 10-day predictions every 10 days for the last year of data, which corresponded to 35 dates of calibration. For the aggregated model, the built-in function *OneStepAhead* was used to generate the forecasts. In the daily models, analytically computing the distribution of the cumulated number of cases predicted over 10 days was not straightforward. Therefore, a simulation approach was chosen instead, where the daily model was fitted up to each date

of calibration t_c. The parameters estimated during this fit were then used to generate K simulations of the number of cases per region in the next 10 days. The local number of cases across the 10 days were added together to generate the distribution of the number of cases predicted per region. In all calibration sets, K was set to $20{,}000$ simulations, which led to stable predictive distributions. If the number of cases observed in the data had not been observed in the calibration with $K = 20{,}000$, more simulations were generated until the number of observed cases was obtained at least once. The distributions obtained under the aggregated and daily models were then compared to the data to compute the Ranked Probability Score (RPS), a strictly proper scoring rule, as an overall indicator of predictive ability. The RPS encapsulates the paradigm that forecasts should aim to "maximise sharpness subject to calibration" [18], and is defined for count data as [19]:

$$RPS(P_t, x_t) = \sum_{k=0}^{\infty} \big(P_t(k) - \mathbb{1}(k \geq x_t)\big)$$

$P_t(k)$ is the predictive cumulative probability of observing incidence k at time t, and x_t the observed data point.

For a more detailed assessment, I also separately evaluated predictive bias (i.e. the degree to which the models systematically under- or overestimated the number of cases) and sharpness (i.e. the width of the predictive distributions) of the models. The bias is defined as [20]:

$$B_t(P_t, x_t) = 1 - (P_t(x_t) + P_t(x_t - 1))$$

And the sharpness is evaluated using the normalised median absolute deviation about the median (MADN) [20]:

$$S_t(P_t) = \frac{1}{0.675} median(|y - median(y)|)$$

With y a variable with Cumulative distribution function (CDF) P_t.

For each simulated dataset, the *Permutation Test* implemented in the *surveillance* package was used to assess whether the differences in RPS between both models were significant [7].

5.3. Results

5.3.1. Description of the simulated outbreaks

Each dataset was simulated by drawing the number of cases per department from a negative binomial distribution every day between January 2009 and December 2017. The set of simulated outbreaks generated showed a wide variety of transmission dynamics. The overall number of cases generated in nine years ranged from 5,900 to 258,100, while 60% of the simulations had between 10,000 and 20,000

cases (Figure 5.1A). Since only 70% of the cases were assumed to be reported, the median number of cases reported in the simulation was 10,600 cases. Similarly, the number of annual cases greatly varied between the simulation sets (Figure 5.1B). Major outbreaks with more than 2,000 cases were generated in at least one simulation every year, whereas the minimum was around 300 cases per year. Within-simulation variations were also observed: in some simulations the number of yearly cases was relatively constant, with seasonal outbreaks reaching similar daily maximums every year (Figure 5.1C), whereas in others, large outbreaks followed for several years with low levels of transmission (Figure 5.1D).

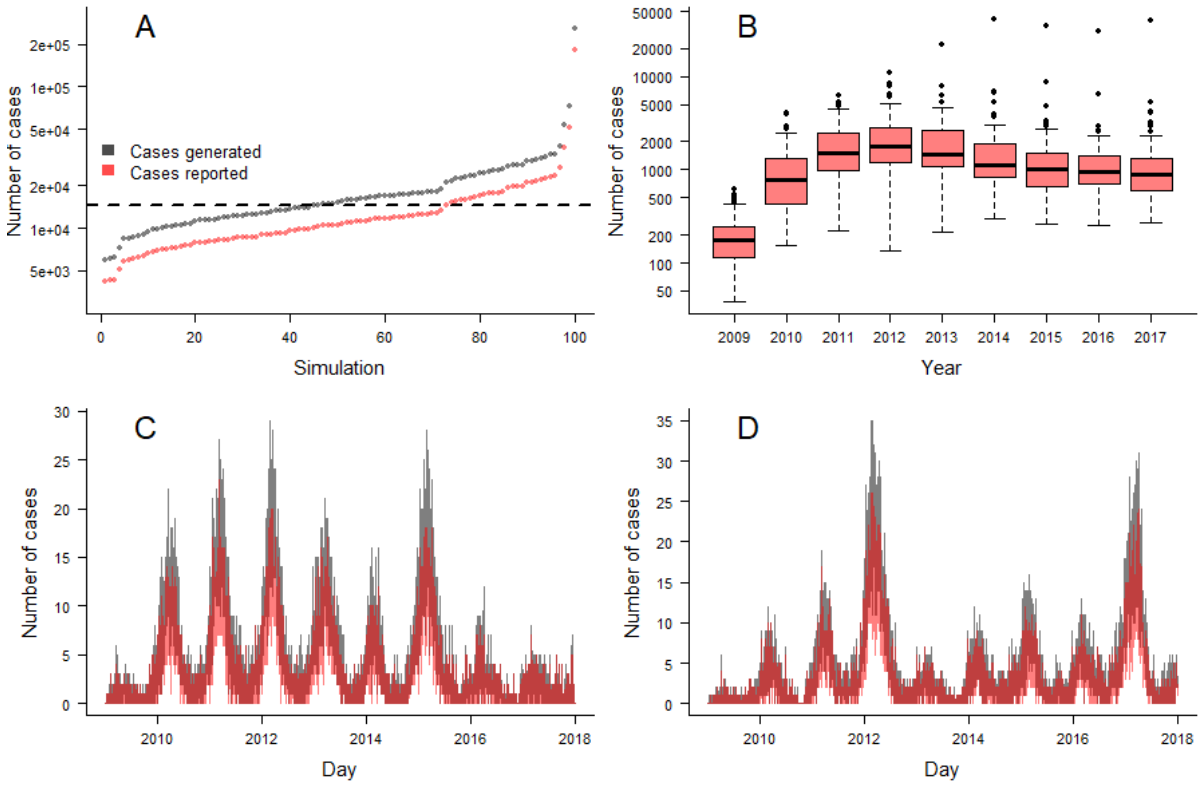

Figure 5.1: Panel A: Overall number of cases generated and reported per simulation, 70% of the generated cases were reported. The black dotted line represents the actual number of cases reported in France in this timespan (approximately 14 000 cases). Panel B: Boxplots of the number of cases reported per year in the simulations, Panels C and D: Daily number of cases in two of the simulations generated. The y-axis in panels A and B are shown in log-scale.

Finally, the spatial distribution of the cases showed that in more than 95% of the simulations, all the departments reported at least one case (left panel Figure 5.2), excepted for Lozère and Arriège (in orange on the map), which were exposed only in more than 75% of the simulations. More spatial heterogeneity was observed in the distribution of departments that reported 10 yearly cases or more in at least one year of the simulations: All departments in the suburbs of Paris and the north of France reported more than 10 yearly cases at least once, whereas various rural regions in the centre and the south reported 10 cases in less than half the simulations (Central panel Figure 5.2). Finally, large outbreaks were sparser: many rural regions never reached 50 cases in a year, where highly populated urban centres (Departments around Paris, Lille, Lyon, or Marseille in orange and red) reported more

than 50 annual cases at least on one occasion in more than half of the simulations. This is most likely because urban centres tend to attract cases from cross-regional transmissions, as observed in Chapter 4. Indeed, since the set of parameters used to generate this simulation set was taken from the model fitted in Chapter 4, this pattern was expected. The R scripts and the functions used to generate the simulations are available in the R folder of the Github repository https://github.com/alxsrobert/measles-regional-immunity (files *generate_simulations.R, function_generate_all_outbreaks.R, and function_generate_outbreak.R*).

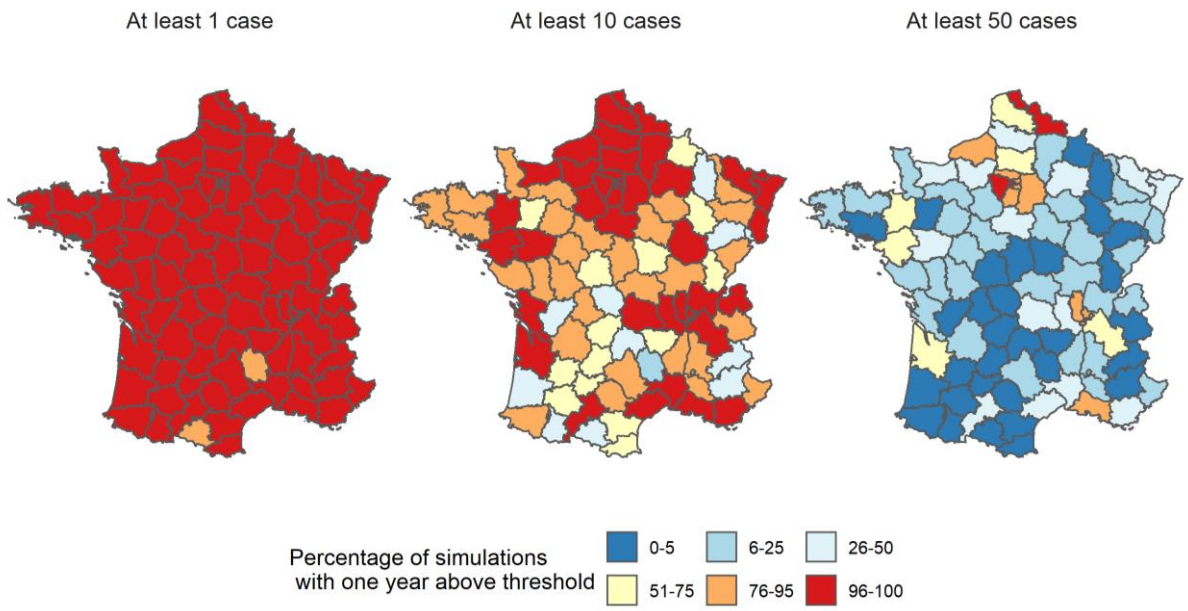

Figure 5.2: Percentage of simulations where the number of cases reported in each department was at least 1, 10, and 50 cases in at least one year.

5.3.2. Parameter fits

For each of the simulations, two models were fitted using the Epidemic-Endemic framework implemented in the R package *surveillance*: a daily model, and an aggregated model using a 10-day aggregation. The code for the implementation of both models is shown in *R/ function_analysis_hhh4* and *R/generate_analysis_simulations.R*, whereas the code to generate all the figures is available in *R/generate_plots_simulations.R*.

Firstly, I compared the ability of both the aggregated and daily models to capture the actual values of the parameters in the 95% confidence intervals. The parameters selected for the comparison were the coefficients associated with each covariate, which quantify whether the model accurately described the impact of the covariates on each component, and the two parameters used to compute the connectivity matrix, which describe whether the models generated the same pattern of cross-regional connectivity as the simulations.

In the set of daily models, the 95% confidence interval captured most of the input parameters used to generate the simulations: 14 of the 21 selected parameters were contained in the 95% confidence interval in at least 90% of the simulations (Figure 5.3A). The impact of the categories of incidence was harder to accurately estimate. Indeed, the incidence parameters in the neighbourhood and autoregressive components were included in the confidence interval in 82 to 90% of the simulations. This could be caused by the combined impact of unreported cases, and discrepancies in the incidence category thresholds between the models and the simulations. Indeed, in the models the threshold defining the last category of incidence was computed using the second tercile of all measures of recent incidence, whereas in the simulation, it was set to 45 cases per million.

In the aggregated models, the proportion of input parameters included in the 95% confidence intervals of the model was lower, and did not exceed 85% for any of the parameters (Figure 5.3B). The seasonality of the neighbourhood and autoregressive components did not correspond to the simulations. Indeed, the "cosine" seasonal parameters in the autoregressive component was never included in the 95% confidence interval, and the "sine" parameter was included in 73% of the models. The seasonal parameters in the neighbourhood models were captured in 24 to 27% of the models. This indicates that the timing of the peaks in transmission differed in the simulations and the aggregated models. Furthermore, discrepancies were observed in the coefficients quantifying the impact of the categories of incidence, which were included in the 95% confidence intervals in 63 to 76% of the simulations.

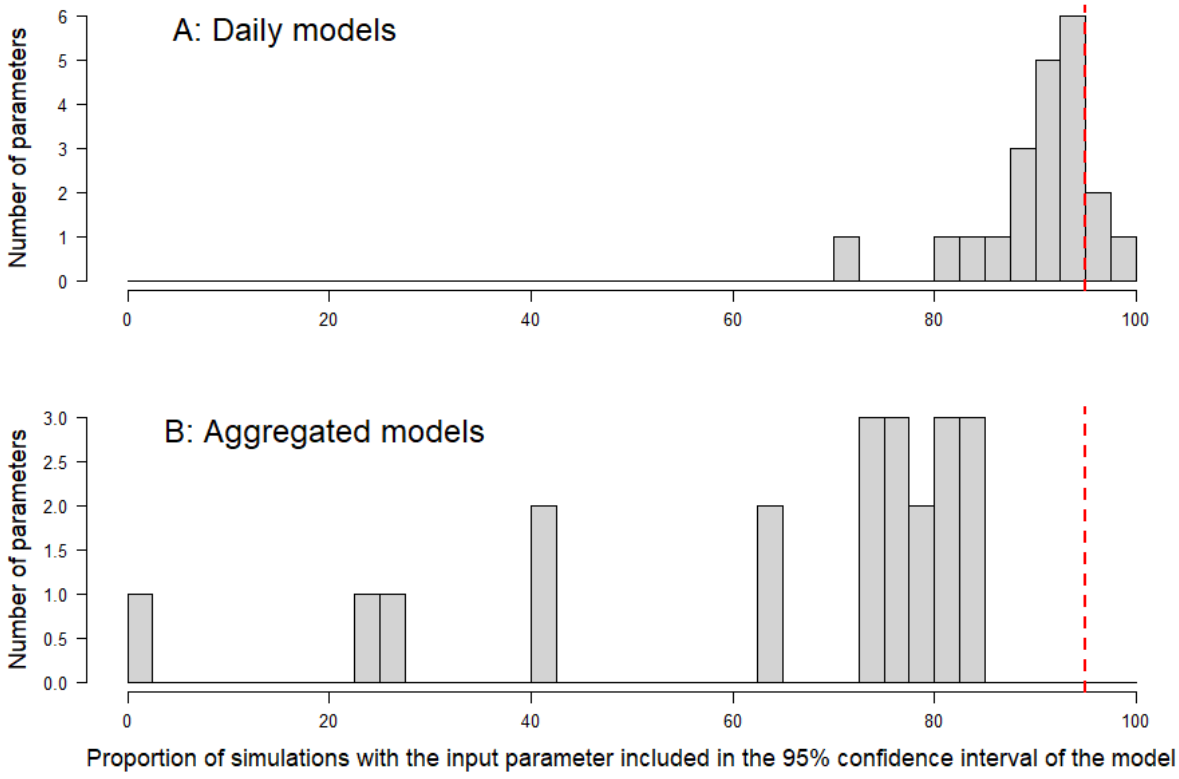

Figure 5.3: Histograms representing the proportion of simulation sets where the input parameters used to generate the simulations are included in the 95% confidence interval of the model. For each parameter, the value should be close to 95%, indicating that 95% of the models included the input parameter in their 95% confidence interval. The top panel shows the results using the daily model, whereas the bottom panel corresponds to the aggregated model fits. The red dotted line corresponds to 95%. The parameters integrated in these figures are the covariates' coefficients in each component, and the parameters of the gravity model.

More specifically, the coefficients associated with the proportion unvaccinated and the recent levels of incidence in each component are of special importance, since they describe the impact of the covariates of interest in Chapter 4. In all three components, the daily model estimated the impact of vaccination with more precision and accuracy than the aggregated models (Figure 5.4). Indeed, the distribution of the difference between the mean estimate and the input parameter had a lower variance around 0 in the daily models than the aggregated set. Because of this bias, the impact of vaccination on the number of local secondary transmissions was estimated to be negligible, or negative, in some of the aggregated models (i.e. the mean estimate in Figure 5.4B was below 0). The PIT histograms highlighted the bias observed in the aggregated models: in all three compartments the PIT histograms were skewed towards extreme values, whereas the PIT histograms generated with the daily models were more uniform. This indicates that the aggregated models tended to underestimate or overestimate the impact of vaccination on the different components. However, the estimation of the impact of vaccination on the endemic component appears imprecise in both models (Figure 5.4, Panels G and H), which could be because cases stemming from the endemic component were hard to identify if they were generated during phases of active transmission. These new importations may then be considered by the models as secondary cases stemming from concurrent transmission chains, and change the estimated impact of the covariates. Alternatively, the uncertainty could be due to the small number of cases stemming from the endemic component.

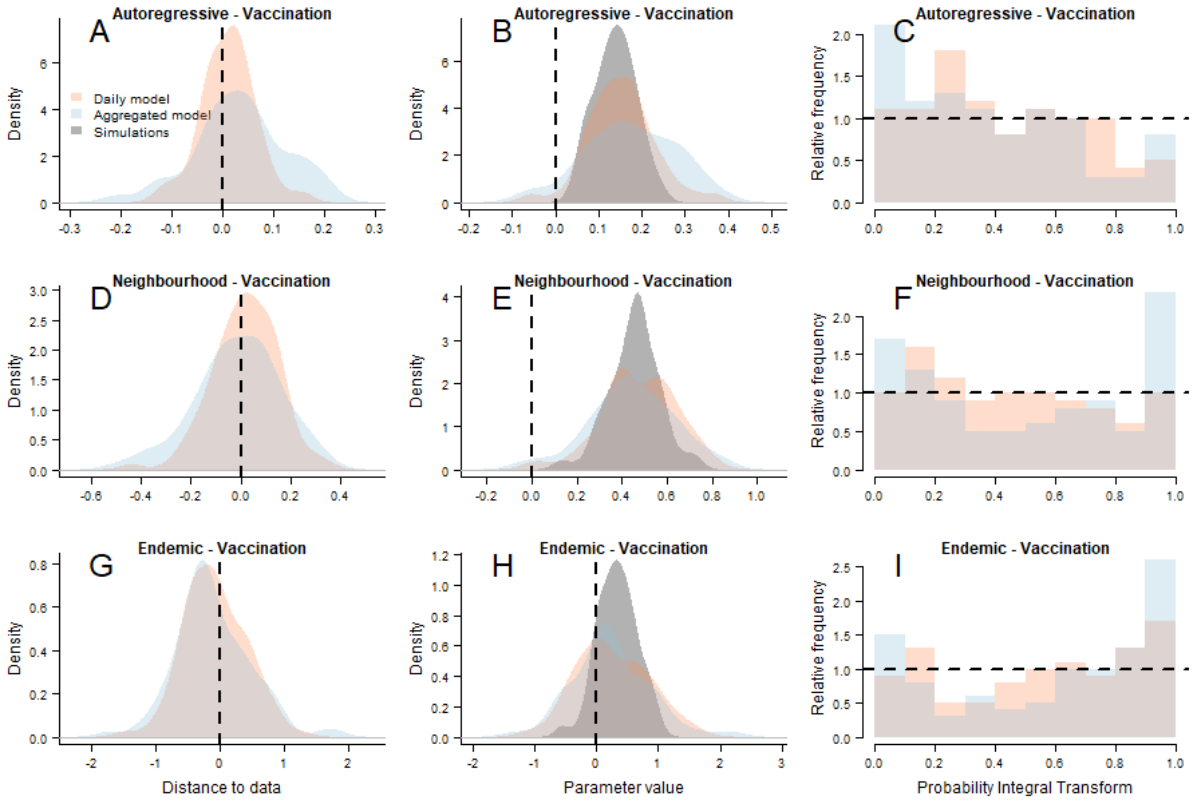

Figure 5.4: Comparison between the estimated impact of the proportion unvaccinated in the daily (red curves) and aggregated (blue curve) models, and the input distribution in the simulations, in each component. Each row corresponds to a different component. The left column (panels A, D, and G) shows the density plots of the difference between the input parameter in the simulations, and the mean parameter estimate in both sets of models, the black dotted vertical line corresponds to 0, where the estimated parameter was equal to the input value. The central column (panels B, E, and H) corresponds to the distribution of the mean parameters in both models, along with the distribution in the simulations. The black dotted vertical line corresponds to 0, where the impact of the parameter on the number of cases is null. Finally, the right column (panels C, F, and I) corresponds to the PIT histograms of the parameter for both models. The dotted horizontal line corresponds to a value of 1, which would correspond to the ideal PIT histogram.

Similarly, the parameters quantifying the effect of recent incidence were more accurately estimated in the daily models of moderate ('incidence 1') and high ('incidence 2') transmission (Figure 5.5 and Figure 5.6). Although the PIT histograms of the daily models were not uniform, the aggregated models were more skewed towards extreme values. The impact of recent incidence was harder to capture for both models because of the report ratio, and the discrepancies in the thresholds defining the categories of incidence in the simulation and the models. However, despite the skewed PIT histograms, the distance between the mean estimate and the input distribution was small, and the overall conclusions were robust: Increases in recent incidence were associated with a reduction in local secondary transmission and an increase in cross-regional transmission. Furthermore, the parameters associated with moderate transmission (incidence 1) had a small impact on transmission, whereas high recent levels of transmission had a larger effect across all the components.

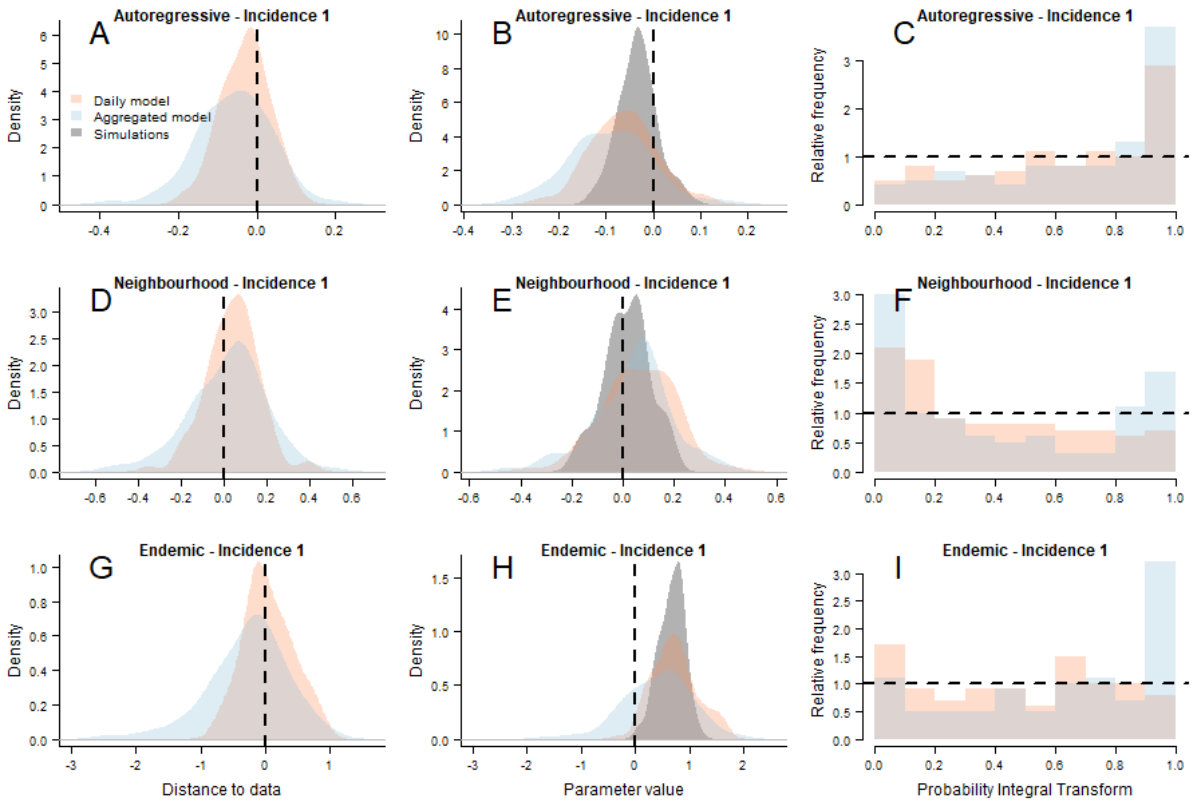

Figure 5.5: Comparison between the estimated impact of medium levels of recent incidence in the daily (red curves) and aggregated (blue curve) models, and the input distribution in the simulations, in each component. Each row corresponds to a different component. The left column (panels A, D, and G) shows the density plots of the difference between the input parameter in the simulations, and the mean parameter estimate in both sets of models, the black dotted vertical line corresponds to 0, where the estimated parameter was equal to the input value. The central column (panels B, E, and H) corresponds to the distribution of the mean parameters in both models, along with the distribution in the simulations. The black dotted vertical line corresponds to 0, where the impact of the parameter on the number of cases is null. Finally, the right column (panels C, F, and I) corresponds to the PIT histograms of the parameter for both models. The dotted horizontal line corresponds to a value of 1, which would correspond to the ideal PIT histogram.

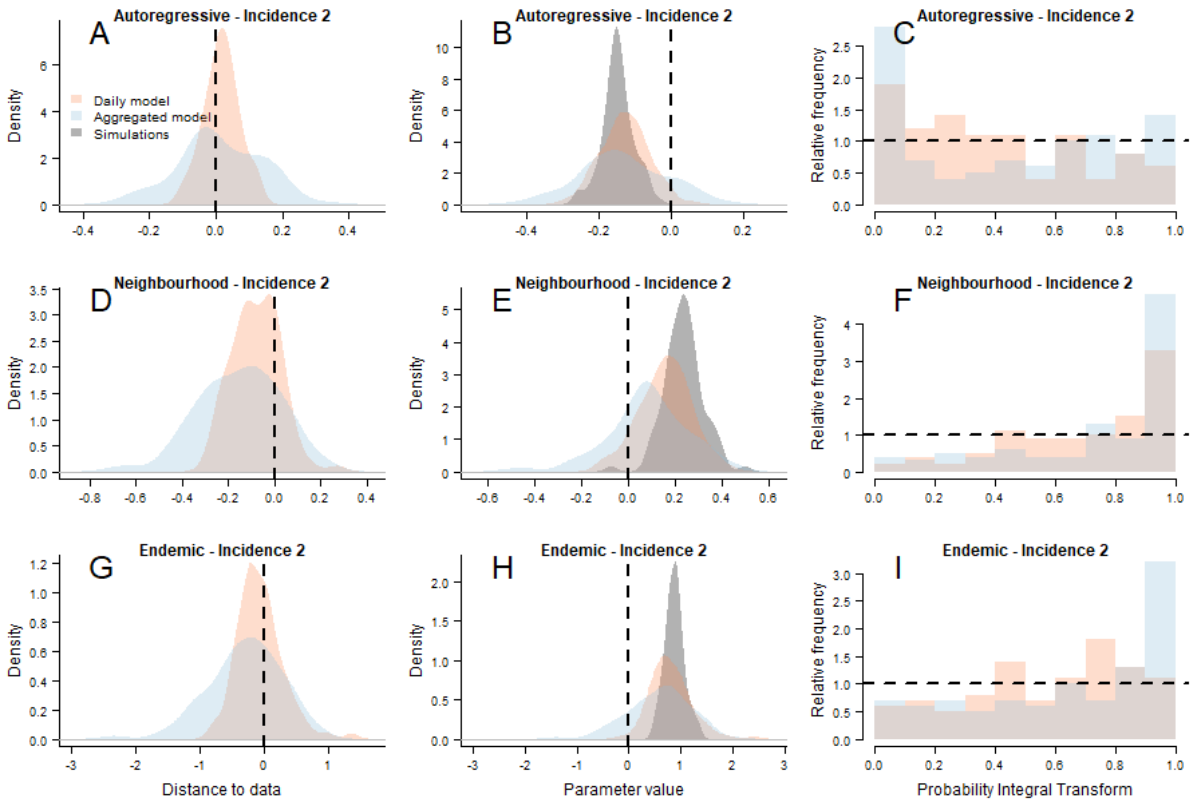

Figure 5.6: Comparison between the estimated impact of high levels of recent incidence in the daily (red curves) and aggregated (blue curve) models, and the input distribution in the simulations, in each component. Each row corresponds to a different component. The left column (panels A, D, and G) shows the density plots of the difference between the input parameter in the simulations, and the mean parameter estimate in both sets of models, the black dotted vertical line corresponds to 0, where the estimated parameter was equal to the input value. The central column (panels B, E, and H) corresponds to the distribution of the mean parameters in both models, along with the distribution in the simulations. The black dotted vertical line corresponds to 0, where the impact of the parameter on the number of cases is null. Finally, the right column (panels C, F, and I) corresponds to the PIT histograms of the parameter for both models. The dotted horizontal line corresponds to a value of 1, which would correspond to the ideal PIT histogram.

I also compared the proportion of cases originating from each component in the simulations, and in both sets of models (Figure 5.7). In the simulations, between 65% to 73% of the cases came from the autoregressive component, whereas cross-regional transmission represented 20% to 34% of the transmissions. The proportions of cross-regional transmissions and endemic importations were estimated to be slightly higher in the daily models. This could be explained by punctual local transmission being classified as cross-regional because of competing infectors. The difference was much more apparent in the set of aggregated models, where only half the cases (50% to 57%) stemmed from the autoregressive component, and the proportion of cross-regional transmissions was higher. Because of the aggregation, some of the local transmissions cannot be classified in the autoregressive component, which could be due to longer serial intervals or unreported generations, and are therefore explained by the model as cross-regional transmission or background importations. This difference is further highlighted by the discrepancies in the distribution of the two spatial parameters (γ and δ) between the simulations and the set of aggregated models: indeed, the population parameter was

overestimated (bottom panel Figure 5.8), whereas the distance parameter tended to be much lower in the aggregated models than in the simulations. Therefore, the decay associated with the increase in distance was slower in the aggregated models (top panel Figure 5.8), and the connectivity matrix estimated in the aggregated models was more uniform, hence cross-regional transmissions between distant regions were more frequent. This shows that in the aggregated models, some of the local transmissions were replaced by cross-regional transmissions.

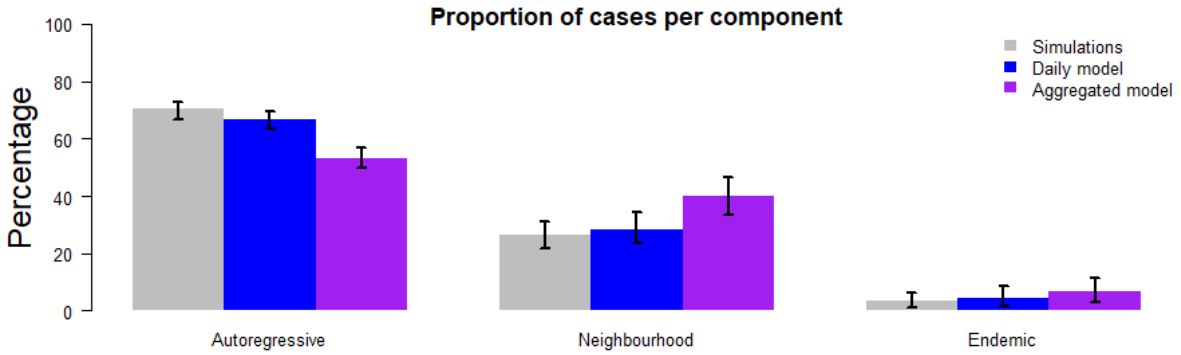

Figure 5.7: Median proportion of cases originating from each component in the simulation sets, and the daily and aggregated models. The arrows correspond to the 95% confidence intervals in each set.

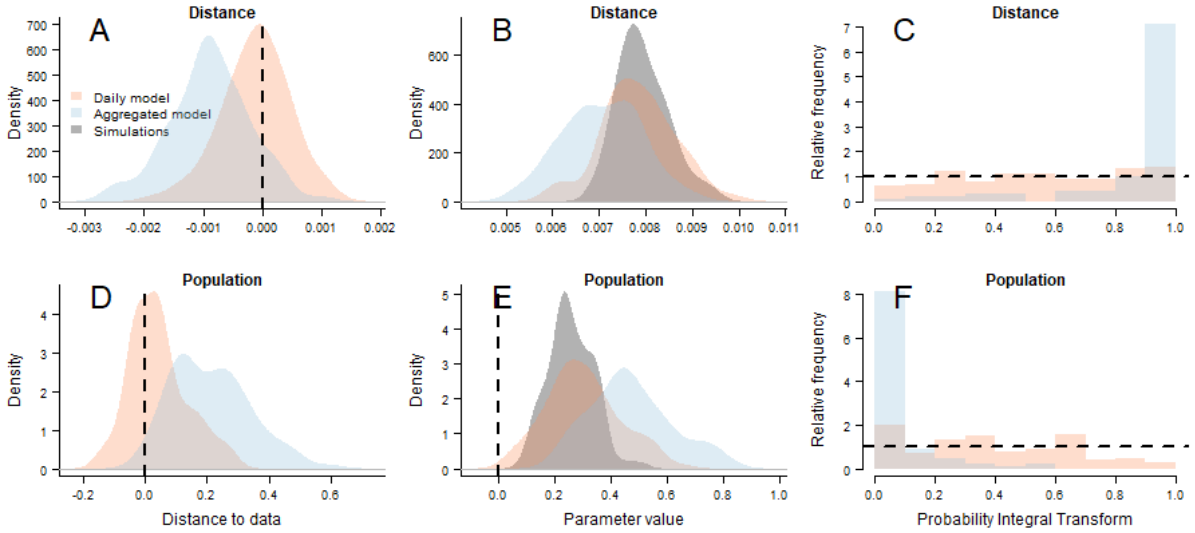

Figure 5.8: Comparison between the estimated parameters of the exponential gravity model in the daily (red curves) and aggregated (blue curve) models, and the input distribution in the simulations, in each component. Each row corresponds to a different component. The left column (panels A, D, and G) shows the density plots of the difference between the input parameter in the simulations, and the mean parameter estimate in both sets of models, the black dotted vertical line corresponds to 0, where the estimated parameter was equal to the input value. The central column (panels B, E, and H) corresponds to the distribution of the mean parameters in both models, along with the distribution in the simulations. The black dotted vertical line corresponds to 0, where the impact of the parameter on the number of cases is null. Finally, the right column (panels C, F, and I) corresponds to the PIT histograms of the parameter for both models. The dotted horizontal line corresponds to a value of 1, which would correspond to the ideal PIT histogram.

5.3.3. Predictive ability

Three indicators were used to compare the 10-day prediction of both sets of models to the data over the calibration period: the bias, which quantifies whether a model systematically over- or underpredicts case numbers; the sharpness, which is independent of the data and quantifies whether the model generates predictions in a narrow range of possible outcomes; and the Ranked Probability Scores (RPS), which is lower if the predictive distribution is close to the one generating the data. The values of each indicator were compared by computing the difference between the predictions of the two models in each simulation. Values of sharpness and RPS are always positive, and lower values are generally preferred, whereas the bias can be negative, and the optimal value is 0. Therefore, the biases were compared by computing the difference in absolute bias between the models.

In most of simulations, the bias value was closer to 0 in the daily models (Figure 5.9A), whereas the values of sharpness were mostly lower in the aggregated models. This indicates that the aggregated models tended to generate narrower predictions, but the daily models were often more balanced. Finally, the RPS was always lower in the daily models. This indicates that using daily incidence improved the predictive ability in every simulation. Nevertheless, to test whether this improvement was substantial, I generated a permutation test on the mean RPS in the aggregated and daily model, for each simulation. In 4% of the simulations, there was limited evidence the daily models improved the predictions ($pvalue > 0.1$ in the permutation test), whereas in 57% of the simulations, there was strong evidence that the predictions generated by the daily models were better than the predictions from the aggregated models ($pvalue < 0.001$ in the permutation test).

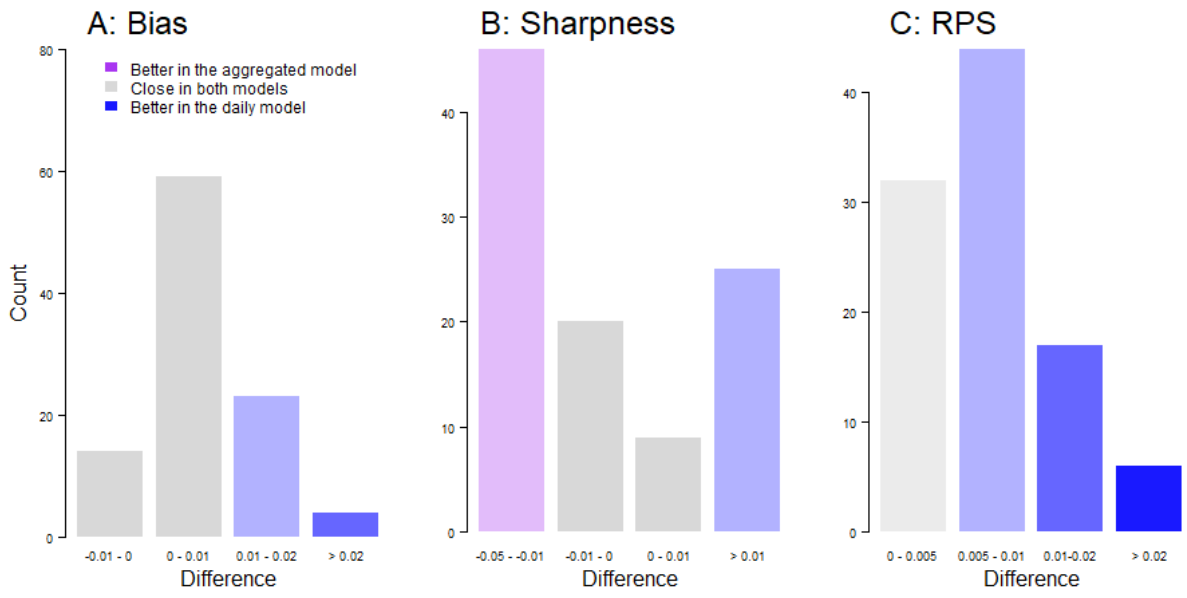

Figure 5.9: Difference in absolute bias, sharpness, and Ranked Probability Scores between the aggregated and daily models. In all three values, lower levels are better. Therefore, positive differences mean that the aggregated values were higher (i.e. the daily model performed better).

5.4. Discussion

The daily models were better able to capture the input parameters of the simulations than the set of aggregated models, and generated predictions significantly better calibrated in most of the simulated outbreaks. In particular, the estimated impact of vaccination coverage corresponded to the input parameters in every component. On the other hand, the impact of the level of recent incidence was harder to capture since the categories of incidence were defined by different thresholds in the simulations and the models. This difference was introduced to match the protocol used in Chapter 4 to define the incidence categories: rather than setting the threshold using an absolute value (45 cases per million in the simulations), I computed the second tercile of incidence in each simulation. Although the mean parameter estimates were biased in some of the simulations, the daily models were able to identify the direction of effect associated with high levels of recent incidence. By using a composite serial interval, the daily models were able to capture the dynamics of transmission despite partial reporting. Therefore, this study highlights the added value of adapting the Epidemic-Endemic framework to daily data.

On the other hand, running the analysis with daily models was more computationally demanding than with the aggregated framework. Similarly, the simulation-based calibration was more time consuming than analytically computing the distribution in the aggregated models. Fitting one daily model took about five minutes (less than one minute per aggregated model), and the calibration runs were about 30 minutes long per simulation (less than two minutes per aggregated model) on a standard desktop computer (Intel Core i7, 3.20 GHz 6 cores). Although the aggregated models were outperformed by daily models, they were also mostly able to capture the value of the input parameters, and their predictions were associated with low bias. Therefore, at least in the scenarios studied, aggregated models could be said to be able to provide insights into the dynamics of transmission if daily data are unavailable. However, this study highlights the importance of developments in the Epidemic-Endemic to integrate cases infected a few time steps before on current transmission, as recently implemented by Bracher and Held [8]. The aggregated models implemented in this study tended to replace local spread by cross-regional transmissions because there was no eligible infector in the region (either because they were not reported, or because of their aggregated onset date). Weighting the potential for transmission over several time steps could allow aggregated models to capture transmission events with longer serial intervals, or to integrate the impact of unreported generations. This could be particularly important with weekly or fortnightly aggregated data because the aggregation period would then be more distinct from the average serial interval for measles, and infectors would be less likely to be grouped exactly one time step before the infectees.

The simulated outbreaks generated in this chapter do not aim to be an accurate description of real-life transmission dynamics. Indeed, factors that were not included in the simulation framework could impact transmission, and the impact of the parameters may change through time. Furthermore, the impact of recent incidence is more complex than the three-category covariate generated here, and the local vaccine coverage may not be an accurate description of the level of immunity in the population. However, the main aim of this study was to explore the loss of information caused by the aggregation of case count across different outbreak scenarios. The variability of outbreaks generated by the simulations allowed an assessment of the added value of daily case counts. The conclusions reached in this study are likely to be robust to the addition of new variables since they would affect both the aggregated and non-aggregated models similarly. Changes in transmission dynamics, such as different serial intervals or reporting rate may have an impact on the conclusions. Indeed, in a fully reported outbreak, with a narrow distribution of the serial interval, on average close to the scale of aggregation, the cases classified in each time step would be more likely to belong to the same generation, with their infectors classified at the previous time step. In this example, the added value of the daily model should be more limited.

Points of improvement in the daily framework would include the use of a time-varying serial interval to compute the transmission potential, or the estimation of the serial interval distribution in the fitting procedure. Indeed, the distribution of the serial interval can vary throughout an outbreak [21,22], meaning it may be better to define the distribution of the serial interval according to the number of local cases recently reported, which would impact the transmission potential. Furthermore, in order to account for time-varying reporting rate, one could account for changes in the proportion of the composite interval that stems from direct transmission during the outbreak. This would be especially relevant with decade-long time series, where the ability of the surveillance system to detect cases may improve through time. In conclusion, the adaptation and application of the Epidemic-Endemic framework to simulated daily case counts resulted in an improvement in the parameter estimates and in the calibration of the models. Integrating daily data into this method is therefore promising and should be further explored to disentangle the local and spatial spread of infectious pathogens.

5.5. References

[1] Finkenstädt BF, Grenfell BT. Time series modelling of childhood diseases: a dynamical systems approach. Appl Stat 2000;49:187–205. doi:10.1111/1467-9876.00187.

[2] Bjørnstad ON, Finkenstädt BF, Grenfell BT. Dynamics of measles epidemics: Estimating scaling of transmission rates using a Time series SIR model. Ecol Monogr 2002. doi:10.1890/0012-9615(2002)072[0169:DOMEES]2.0.CO;2.

[3] Höhle M, Paul M. Count data regression charts for the monitoring of surveillance time series. Comput Stat Data Anal 2008;52:4357–68. doi:10.1016/j.csda.2008.02.015.

[4] Held L, Höhle M, Hofmann M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. Stat Modelling 2005;5:187–99. doi:10.1191/1471082X05st098oa.

[5] Meyer S, Held L, Höhle M. hhh4: Endemic-epidemic modeling of areal count time series. J Stat Softw 2016.

[6] Herzog SA, Paul M, Held L. Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. Epidemiol Infect 2011;139:505–15. doi:10.1017/S0950268810001664.

[7] Paul M, Held L. Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. Stat Med 2011;30:1118–36. doi:10.1002/sim.4177.

[8] Bracher J, Held L. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. Int J Forecast 2020. doi:10.1016/j.ijforecast.2020.07.002.

[9] Fine PEM. The Interval between Successive Cases of an Infectious Disease. Am J Epidemiol 2003;158:1039–47. doi:10.1093/aje/kwg251.

[10] Santé Publique France. Données départementales 2007-2012 de couverture vaccinale rougeole, rubéole, oreillons à 24 mois 2019. https://www.santepubliquefrance.fr/determinants-de-sante/vaccination/articles/donnees-departementales-2007-2012-de-couverture-vaccinale-rougeole-rubeole-oreillons-a-24-mois (accessed September 7, 2020).

[11] Santé Publique France. Estimations des couvertures vaccinales à 24 mois à partir des certificats de santé du 24e mois, 2004-2007 2010. https://www.santepubliquefrance.fr/determinants-de-sante/vaccination/articles/donnees-departementales-2013-2017-de-couverture-vaccinale-rougeole-rubeole-oreillons-a-24-mois (accessed September 7, 2020).

[12] Santé Publique France. Données départementales 2013-2017 de couverture vaccinale rougeole, rubéole, oreillons à 24 mois 2019. https://www.santepubliquefrance.fr/determinants-de-sante/vaccination/articles/donnees-departementales-2013-2017-de-couverture-vaccinale-rougeole-rubeole-oreillons-a-24-mois (accessed September 7, 2020).

[13] Institut National de la Statistique et des Etudes Economiques. Estimation de la population au 1er janvier 2020 2020. https://www.insee.fr/fr/statistiques/1893198#consulter (accessed September 7, 2020).

[14] Lenormand M, Bassolas A, Ramasco JJ. Systematic comparison of trip distribution laws and models. J Transp Geogr 2016;51:158–69. doi:10.1016/j.jtrangeo.2015.12.008.

[15] Eurostat. European population grid cells 2011. https://ec.europa.eu/eurostat/web/gisco/geodata/reference-data/grids (accessed September 12, 2020).

[16] Hijmans RJ, Etten J van, Mattiuzzi M, Sumner M, Greenberg JA, Lamigueiro OP, et al. Package "raster." R 2014.

[17] Dawid a. P. Statistical Theory: The Prequential Approach. J R Stat Soc Ser A 1984;147:278–92.

[18] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 2007;102:359–78. doi:10.1198/016214506000001437.

[19] Czado C, Gneiting T, Held L. Predictive Model Assessment for Count Data 2009:1254–61. doi:10.1111/j.1541-0420.2009.01191.x.

[20] Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area Region of Sierra Leone, 2014–15. BioRxiv 2017:1–17. doi:10.1101/177451.

[21] Svensson Å. A note on generation times in epidemic models. Math Biosci 2007;208:300–11. doi:10.1016/j.mbs.2006.10.010.

[22] Fraser C. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One 2007;2. doi:10.1371/journal.pone.0000758.

Chapter 6. Discussion

In this thesis, I have explored how various routinely collected data sources can be combined and integrated in mathematical models to gain insight into the risks of measles outbreaks in countries near elimination. I have approached this question from two different angles, firstly through the reconstruction and analysis of transmission chains, then by evaluating the association between different indicators of immunity and local case counts to identify reliable predictors of the local risks for outbreaks. I will start this chapter by providing a summary of the results from previous chapters, followed by a discussion of the strengths and limitations of the methods developed in this PhD thesis. I will then summarise the contributions relative to previous research, and finally I will present the future research opportunities that arise from the conclusions of this project.

6.1. Summary of main findings

In Chapter 2, I presented the R package *o2geosocial*, which I developed to infer who infected whom using the onset date, location, age group and genotype of the cases. These variables were chosen because they are routinely collected in most settings, and can all be informative of the history of transmission [1–3]. This novel method was implemented as an R package, building upon previously developed reconstruction methods such as the package *outbreaker2* [4,5], and published on CRAN. Chapter 2 also contains a reproducible example to highlight the flexibility and limitations of the model. This example is based on a simulated dataset of 75 cases reported in a community within 6 months. The method was able to identify the infector for most of the cases, and highlighted the areas more often associated with secondary transmissions. This chapter also showed the added value brought by previous knowledge on the importation status of the cases. Indeed, the inference of the importation status led to inconsistencies between the inferred transmission trees and the simulated data, whereas the transmission trees inferred when the importations were already identified consistently matched the data. This was due to the presence of multiple concurrent independent transmission chains reported in a small geographical area, which are harder to detect since measles genetic data are uninformative in this context (i.e. the method may not find how many independent chains are happening at the same time). Increasing the proportion of importations detected by the model is crucial to correctly assess the cluster size distribution and the number of locally acquired cases per region. For instance, underestimating the number of importations would lead to an underestimation of the number of clusters, and an overestimation of the number of locally acquired cases following each importation. Therefore, the accuracy of the inferred clusters can be improved by case investigations into recent travels, in order to identify at least a proportion of the importations prior to the inference process.

Despite being initially developed to match the specificities of measles virus, *o2geosocial* requires data often collected in infectious disease outbreaks and is therefore applicable to a wide range of settings and pathogens. For instance, reconstructing transmission trees during seasonal flu outbreaks would be useful for identifying routes of transmission repeatedly observed, and showing what locations are typically associated with increases in incidence. Such a study could then be used to design specific vaccination programs and surveillance in the areas most often linked to super-spreading events. On the other hand, this method may have stronger limitations for outbreaks with low report rates, for instance because of asymptomatic transmission. In this situation, *o2geosocial* may over-estimate the number of unrelated importations and under-estimate the missing links between cases. Finally, for pathogens such as the Ebola virus, where genetic sequences are informative, *o2geosocial* presents the opportunity to integrate all the cases (i.e. un-sequenced cases would not be excluded), but may not be able to optimally use the information contained in the reported genetic sequences compared to methods that fully take this into account.

This application of *o2geosocial* on simulated data thus showed its ability to accurately reconstruct who-infected-whom in a simulated measles outbreak and highlighted some of its limitations. However, these simulated data are not perfectly representative of actual outbreak dynamics and can exaggerate how effective an inference method really is. Indeed, the simulation framework perfectly matched the inference framework, which means that all factors responsible for transmissions are integrated in the inference framework, which is not true in applications using real-life data. In the third chapter of the thesis, I applied *o2geosocial* to measles cases reported in the USA between 2001 and 2016. In the United States, contact-tracing investigations are carried out in order to reconstruct the transmission clusters through patient interviews. Therefore, the transmission clusters inferred by various *o2geosocial* models could be directly compared to the clusters from contact-tracing investigations. Overall, the inferred transmission clusters matched the epidemiological clusters: the models were able to correctly infer up to 87% of the imported cases (assuming import status was always correctly ascertained in epidemiological investigation), and captured the overall cluster size distribution. The geographic distance between cases was identified as the most informative variable for clustering, however this may be due to the specificities of measles in the United States (i.e. few cases reported over a very large geographic area), and may not be generalisable to other settings. The match between inferred and epidemiological clusters was further improved when the importation status was known. Since the reconstructed clusters matched the epidemiological contact tracing investigations, this method could be extended to routinely collected data when contact tracing investigations are not carried out. The probabilistic transmission trees could then be studied to bring insights into the areas most at risk of

transmission, estimate transmission parameters [6], and identify the variables associated with increased numbers of secondary cases [7], or super-spreading events [8].

The fourth chapter focused on evaluating the impact of local vaccine coverage and recent incidence on the daily number of measles cases per department (i.e. NUTS3 regions) in France, between 2009 and 2018. The case count data was fitted by adapting the Epidemic-Endemic framework to daily data. In the existing literature, the Epidemic-Endemic model had been applied to aggregated case counts [9–12]. This chapter aimed to evaluate whether areas eligible for WHO's elimination status (i.e. with low recent incidence) were actually associated to lower risks of outbreaks in France, and whether local vaccine coverage was a reliable predictor of future transmission. In the Epidemic-Endemic models, the 3-year average local vaccine uptake was associated with a lower risk of importations (whether cross-regional or baseline importations), and local secondary transmission, indicating that recent values of vaccine coverage were relevant indicators of the risks of outbreak in France. On the other hand, higher levels of local incidence in the past three years were associated with lower risks of local secondary transmission, which indicates that areas that had not reported any cases in the past three years were estimated to be more at risk of outbreaks than the other departments. This could be linked to a replenishment of susceptibles, whereby after years of low incidence, the number of susceptible individuals is sufficient to trigger outbreaks, and thus raises questions regarding the requirements of WHO's elimination status, where countries become eligible for elimination after three years of interrupted transmission [13]. In Chapter 4, I also highlighted the importance of maintaining the national vaccine uptake by exploring the impact of variations in vaccine coverage on the number of simulated cases in a year: a drop in 3% of the three-year average vaccine coverage across the country led to a large increase in the number of annual cases simulated (almost all simulations generated more than 1,000 cases). While predictions of the model lacked accuracy when the calibration period exceeded 10 days ahead and therefore these simulations should not be taken as accurate predictions of future outbreaks the relative impact of variations in coverage is informative of the importance of maintaining the vaccine uptake to limit the risks of outbreaks. The model implemented highlighted the heterogeneous risks of transmission in the different French departments. Indeed urban, populated areas were associated with increased risks of cross-regional transmission, and the risks of local transmission were highest in the South East and the North of France. These areas, identified as vulnerable by the model, could be targeted by serological surveys, in order to verify whether the levels of immunity in the population are sufficient to avoid future outbreaks, and, if not, design catch up vaccination campaigns targeting the local population.

Finally, the fifth chapter of the thesis presented a comparison between aggregated and non-aggregated Epidemic-Endemic models, implemented using the R package *surveillance* and applied to simulated data. Using the parameter sets generated in Chapter 4, I generated 100 simulated outbreaks, and fitted

the simulated case counts with aggregated and non-aggregated models, using the same set of covariates as in the simulations. Although both models were able to capture the dynamics observed in most simulated outbreaks, the daily models outperformed the aggregated models both in their ability to correctly capture the parameters used to generate the simulations, and in their ability to make accurate predictions. The improvements were mostly due to the ability of the daily models to deal with variations in the distribution of the serial intervals and underreporting. The aggregated models tended to compensate for missing local cases at the previous time step by increasing the risks of cross-regional transmission, thereby linking the current case counts to unrelated cases in other regions. This highlights the importance of collecting and using daily data when they are available, and adapting aggregated models to account for missing generations [9].

This set of projects has shown how routinely collected data can help reconstruct local risks of transmission, and highlighted the importance of developing novel methodology to make the best possible use of routinely collected data. Using two case studies in countries near elimination (France), or where measles had already been declared eliminated (USA), it also highlighted that the heterogeneous risks of measles outbreaks can only be captured using multiple data sources, since each measure of risk has strong limitations, and that the indicators of vulnerability need to be evaluated in order to thoroughly assess the risks of measles outbreaks in a country.

6.2. Strengths and limitations

The methods implemented in the different chapters share the strength of being applicable to many settings: I focused on developing inference frameworks aiming to maximise the amount of information that can be inferred from routinely collected measles surveillance and coverage data. As highlighted in Chapter 1, local heterogeneity in immunity against measles is caused by multiple factors, which makes the identification of vulnerable areas challenging. The transmission tree reconstruction method, and the Epidemic-Endemic framework show two complementary approaches to measure and identify areas associated with higher risks of transmission. Both methods are also adaptable to other pathogens where local heterogeneity in risks of transmission is a leading cause of transmission. As mentioned earlier, this would be especially relevant for pathogens with high detectability of the cases, and where several years of data are available, in order to observed repeated transmission patterns. The code implemented to reproduce the analysis in each chapter is documented in publicly available Github repositories in order to be transparent, and make extensions and adaptations as easy as possible. The epidemiological data used in Chapter 3 and Chapter 4 were not publicly available, and thus could not be shared. Therefore, I reproduced the analysis using simulated data in both cases.

The reconstruction method implemented in *o2geosocial* was applied to simulated and real-life data, and both analyses showed the method was able to capture complex dynamics of transmission. Given the limited number of variables required to generate accurate results, these analyses could be repeated in many countries and different settings. Probabilistic transmission trees reconstructed using several years of data could be used to compare the characteristics of the areas associated with importations, or with increases in the number of cases, every year. Increasing the level of immunity in these areas, identified as crucial for limiting the spread of the virus, would be key to reduce the risks of future spread. Therefore, these regions could then be targeted by SIAs, or adaptation in the immunisation program could be implemented, in order to close the immunity gap. Furthermore, integrating different countries into the analysis could bring insight into the number of cross-national transmissions. One could also use the transmission trees reconstructed in different countries to highlight characteristics associated with areas at risk and thus identify indicators of transmission (e.g. low recent incidence, high levels of vaccine hesitancy, high connectivity with other regions), or events commonly associated with increases in the number of cases.

The R package was developed with the purpose of being highly flexible. Indeed, each component in the likelihood of connection between two cases can be edited to implement alternative methods. For instance, in Chapter 2 I showed how the spatial component, by default described using a gravity model, can be changed to a Stouffer's rank model, which was shown to perform well at reconstructing human movements during pre-vaccination era measles outbreaks [14]. The temporal likelihood could also be edited to implement a time-varying serial interval defined before running *o2geosocial*. Indeed, the distribution of the number of days between connected cases can change over the course of the outbreak, so setting a different distribution depending on the number of active cases could improve the ability of *o2geosocial* to accurately capture the history of transmission [15]. Different age-stratified contact matrices can also be used to compute the probability of transmission between age groups. However, although all the analyses implemented in this thesis were able to match the transmission patterns in the input data, it cannot be concluded whether they would work as well in every setting. Indeed, the specificities of the countries used for the analyses may hide certain biases and overstate the abilities of the models implemented. For instance, the small number of cases reported in the United States may maximise the ability of *o2geosocial* to reconstruct transmission clusters, given that cases were spread across such a large spatial area.

Similarly, the analyses ran using the Epidemic-Endemic framework (Chapters 4 and 5) were implemented using a flexible framework, able to accommodate different ways to compute the connectivity between regions, and integrate a wide number of covariates, and were able to capture measles dynamics in both real-life and simulated data. In Chapter 4, the calibration study indicated that

the prediction generated by the model showed signs of bias when the calibration period exceeded 10 days. This can be due to various factors, such as an underestimation of the number of background importations, therefore the number of cases predicted remain low if there is no active transmission in the region at the time of prediction; or a misspecification of how the parameters impact the model. For instance, the model assumes the effect of each parameter is constant throughout the timespan of the study, and many potential factors are excluded (e.g., tourism and mass events). As a result of this bias in calibration, the one-year simulation study should not be seen as a prediction of future outbreaks, and mainly aims to illustrate the relative impact of changes in the level of recent incidence, or in vaccine coverage.

In Chapters 2 and 3, once the probabilistic transmission trees were computed, the identification of transmission events can be used to highlight common scenarios of spread, i.e. whether certain regions tend to be more associated with importations, whereas others are "catalysts" for transmission where the number of cases quickly increases. The reconstructed transmission trees can also be used to identify characteristics or settings commonly associated with high transmission events [7,8,16]. However, one main limitation of the use of transmission trees is that it requires previous recent transmissions, and supposes that the transmission observed in recent years is indicative of future transmission. Indeed, if no case were reported in recent years in a given area, the nationwide transmission trees would not be able to conclude anything about the risks of onwards transmission in the region. One of the conclusions of the Epidemic-Endemic model implemented in Chapter 4 was that regions where high levels of transmission were recently reported were associated with lower risks of secondary cases. Regions with low levels of transmission, on the other hand, had higher risks of local transmission. If this result can be extrapolated to the reconstruction of transmission trees, it would indicate that regions with few isolated cases (i.e. low levels of transmission) would be associated with higher risks of transmission, despite the low number of secondary cases per case in the region according to the transmission chains. This shows the importance of studying patterns of spatial spread over multiple seasons and transmission chains, in order to distinguish between regions repeatedly associated with a high number of transmissions and areas with one-off large transmission clusters.

Another limitation highlighted in both Chapters 2 and 3 was the definition and inference of the importation status in *o2geosocial*. Indeed, in the absence of full information on the importation status of the cases, importations must be inferred by *o2geosocial*, whereby they are defined as "cases who do not have any satisfying potential infector". This means that the threshold after which a connection is deemed implausible must be defined by the user. This definition is already used in other inference methods [4]. I changed the way the threshold was computed, and allowed for different numbers of importations between the inferred trees. However, this value influences the final cluster size

distribution, and sensitivity analysis are therefore required if no epidemiological investigations on the importation status of the cases are carried out. This definition also ignores that, during an outbreak, a given imported case may have plausible potential infectors within the data set. This scenario cannot be captured without the use genotype information (i.e. if the genotype of the imported case is different from the other transmission chains) or sequence data showing the number of independent importations. In order to identify these importations, the threshold would have to be set to a stricter level, causing *o2geosocial* to remove true connections and overestimate the number of imported cases. The identification of independent importations was also challenging in the Epidemic-Endemic framework. Indeed, since the models implemented in Chapters 4 and 5 did not integrate information on the genotype or the importation status of the cases, background importations could not be isolated from active transmissions, and tended to be linked to the ongoing transmission chains. This was highlighted by the fact that the proportion of cases stemming from the endemic component was very low, especially during the peak transmission season.

The accuracy and utility of the inferred transmission trees also depends on the proportion of cases that are detected by the surveillance system. Although missing generations are taken into account in the estimation procedure, *o2geosocial* cannot account for entirely unreported transmission clusters, which could lead to inaccuracies in the cluster size distribution reconstructed by the model if these systematically had characteristics different from observed clusters. As a consequence, regions where transmission clusters were less likely to be detected may also be less likely to be identified as more vulnerable to measles outbreaks. Unreported transmission links may also cause an underestimation of the number of superspreading events. Moreover, if the detection rate of the cases is correlated with factors impacting transmission, the analysis of the reconstructed transmission trees may not be able to accurately estimate the impact of these factors. The reconstructed transmission trees therefore only describe a subset of the transmission happening during an outbreak, especially for pathogens with a high proportion of sub-clinical transmissions. These limitations are better accounted for in the daily adaptation of the Epidemic-Endemic framework. In Chapter 5 I used a simulation study to show that implementing a multimodal composite serial interval accounting for different scenarios of reporting improved the estimations of the factors associated with transmission despite partial detection of cases. However, the Epidemic-Endemic framework does not explicitly measure the proportion of cases detected by the surveillance system, and the composite serial interval was determined prior to running the model. The proportion of the composite serial interval stemming from transmissions with missing generations therefore may have to be changed in settings where measles surveillance is not able to identify a large proportion of the reported cases [17].

6.3. Contributions relative to previous knowledge, and interpretation of results

This program of research provides several meaningful steps forward from previous knowledge:

Firstly, Chapters 2 and 3 showed that measles transmission history could be reconstructed accurately in different settings without using extensive data on the genetic sequences of the cases. Indeed, most recent methods developed to reconstruct transmission trees have relied on combining epidemiological data and genetic sequences [4,6,18–23]. However, these methods are difficult to apply to measles, given that the genetic sequences of measles virus are usually not diverse enough to identify direct transmission [5,24]. The projects presented in Chapters 2 and 3 show that routinely collected epidemiological data can be used to reconstruct probabilistic transmission trees, and identify regions or characteristics which were repeatedly associated with transmission events. This contribution is valuable because it highlights how much information on past transmission dynamics can be extracted from variables routinely collected in most countries. The transmission trees can then be used to assess and describe the heterogeneous risks of measles outbreaks, alongside other routinely collected data such as local vaccine coverage, sero-surveys, and molecular surveillance. Given the different limitations associated with each approach, they should be used complementarily to highlight consistent patterns of immunity in countries with high nationwide vaccine uptake.

This study also highlighted which parts of epidemiological investigations were most important for improving the accuracy of the reconstructed transmission trees. Indeed, the transmission trees were inferred using routinely collected variables such as the onset date, location, and age group of the cases, but the inference of the importation status has proven more challenging. Thus, the accuracy of the inference was improved in both applications by incorporating previous knowledge on the importation status of the cases. This highlights the importance of routinely collecting information on recent travels in surveillance data to detect a portion of the imported cases. Collecting the travel history is especially relevant for pathogens with short incubation periods because cases that travelled immediately before their reported date of symptom onset would be more likely to be imported. The travel history would therefore be more indicative of the importation status and the number of concurrent transmission chains. Therefore, integrating the recent travel history of the cases in the routinely collected surveillance data is crucial to improve the reconstructed transmission trees.

Similarly, if multiple genotypes are co-circulating, the proportion of genotyped cases does impact the ability of the model to find imported cases. However, given the limited number of genotypes reported in recent years, more detailed information on the sequences would be needed to make a substantial difference. Indeed, given that the B3 and D8 genotypes account for a large majority of the genotypes

reported in recent years [25], the risks of independent transmission chains having the same genotype are high.

The application of the Epidemic-Endemic framework model to French measles case count data brought insights into the drivers of the risks of measles transmission in countries near elimination. Indeed, I highlighted that in France, local measures of coverage were associated with lower risks of local and cross-regional transmissions. On the other hand, low regional incidence in recent years was not associated with a reduced risk of secondary transmission. This is an important contribution because it highlights that, if the example of France is representative of other countries near elimination, looking into recent incidence to define the current risks of transmission could lead to misevaluations, whereas local vaccine coverage would be a better indicator of the risks of outbreaks. Given the current guidelines defining the elimination status integrate the number of cases in the past three years [26], this could be one of the factors explaining why the elimination status had to be revoked in certain countries shortly after being declared. The results presented in Chapter 4 show that reliable, detailed values on recent vaccine coverage may be more indicative of the local risks of transmission. Furthermore, the simulation study highlighted how variations in vaccine coverage may influence the number of cases generated per region, with a 3% drop in three-year average uptake resulting in a five-time increase in the average number of cases. Given that many countries near elimination in Europe, and in America have reported very low number of measles cases since March 2020 after year of measles resurgence [27,28], and disruptions in immunisation programs have been observed in various countries during the COVID-19 pandemic [29,30], the results of this study would predict that the risks of imminent measles outbreaks are higher in near elimination settings. This drop could be due to higher levels of immunisation due to past outbreaks, but the fact that it happened simultaneously in many countries that recently reported high levels of transmission, and coincided with the Non-Pharmaceutical Interventions (NPI) implemented to mitigate the spread of SARS-CoV-2, indicate that the risks of measles transmission would remain high if the NPIs were lifted. The low number of cases reported in 2020 and 2021 may have led to a replenishment of susceptibles in areas with relatively lower vaccine coverage. If the potential immunity gap created by these exceptional circumstances is not closed quickly with SIAs, the cohort of children missed by the disrupted immunisation campaign may be associated with higher risks of outbreaks for the years to come. The methods developed in this PhD could be used to identify the areas where the risks of local transmission are highest, and where catch-up campaigns are therefore most needed.

Finally, I extended the application of the Epidemic-Endemic framework to daily incidence data, taking into account the transmissibility of past cases according to the serial interval of the disease. The development of a composite serial interval to integrate different scenarios of transmission improved

the ability of the model to capture the dynamics of transmission and estimate the impact of the covariates. The daily framework I developed outperformed the original aggregate estimates in both the simulated and real-life data. In both chapters 4 and 5, the predictions obtained with the daily models were better than the aggregated models. The simulation study also showed that input parameters were estimated more accurately with the daily models. The loss of performance observed in the aggregated models were mostly due to instances where infectors and infectee were not aggregated in consecutive time steps. Therefore, the use of aggregated models that do not integrate the weighted risks of transmission over several time steps is likely to lead to stronger biases for outbreaks and pathogens with a large proportion of undetected cases, and when the period of aggregation does not match the serial interval of the disease.

The implications of the projects developed in this research go beyond the specific aspects of measles in countries near elimination. Firstly, these conclusions give insight into the colossal effort needed to achieve elimination. Indeed, despite a safe and very effective vaccine having been available for more than forty years, only one WHO region (the Americas) has ever reached the elimination status, and still recently reported large outbreaks following political crises and importations from countries affected by measles resurgence. The regular resurgence of cases and heterogeneous risks of transmission in near elimination settings highlight that immunisation efforts (SIAs, routine or mass vaccination) needs to be linked across regions to be effective. The data on immunisation must be collected and available at a fine scale to be as informative as possible, whereas vaccine policies must be applied at a large scale to avoid the risks of imported cases reaching pockets of susceptibles [31]. As different vaccination strategies are designed and implemented to build immunity against COVID-19, the resurgence of measles outbreaks shows the importance of equal access to vaccination. Each pocket of susceptibles is at risk of outbreak, and outbreaks do not occur in isolation, often crossing borders to affect various connected countries at the same time [32]. Secondly, the last ten years of measles dynamics in countries near elimination can inform the next steps required for countries still affected by endemic transmission of measles, besides increasing their overall vaccine coverage. The regular outbreaks in countries near elimination highlight that the immunisation policies must be developed in coordination with neighbouring countries, and implemented thoroughly at a local level. Otherwise, the progress made during the implementation of the routine coverage can be halted by the heterogeneous risks of outbreaks, and the complex identification of the pockets of susceptibles.

6.4. Future research and data requirements

The methods presented in this thesis were developed to be widely applicable since they require a limited number of variables, all routinely collected in countries near elimination. Therefore, the first approach to further the conclusions presented in the different chapters would be to apply *o2geosocial*

and the Epidemic-Endemic framework to more settings, in order to assess whether the findings and implications I presented are generalisable to other countries. The code and functions developed in each chapter are shared as publicly available Github repositories, and *o2geosocial* was implemented as an R package to facilitate future applications.

As mentioned earlier, the number of measles cases in many near elimination settings has dropped since the beginning of the COVID-19 pandemic, due to a variety of factors including decreased international travel; and a drop in social contacts due to NPI implemented during the pandemic [33]. The low number of cases will lead to a replenishment of susceptibles, and the disruption in immunisation program may create immunity gaps. The main area of future research will be to disentangle the effect of the COVID-19 pandemic on measles transmission in the years to come. Indeed, as presented earlier, many countries near elimination had seen a resurgence of cases since 2017, and another resurgence is likely if the contact patterns between individuals comes back to the pre-pandemic levels. Identifying the most vulnerable regions and designing campaigns to increase the levels of immunity is therefore crucial to avoid observing the levels of measles transmission reported in 2018 and 2019. The methods developed in this PhD project aim to shed light on the heterogenous risks of outbreaks from widely collected data and should therefore be of special interest if such a resurgence is observed. To maximise the ability of these methods to estimate the local risks of transmission, accurate routinely collected data are needed, and various other features could be integrated in both models. Therefore, in this section, I will focus on the variables that could be integrated in both *o2geosocial* and the Epidemic-Endemic framework to improve the accuracy of the models.

Firstly, future applications of the Epidemic-Endemic framework could incorporate the association between vaccine coverage and incidence at different scales. Indeed, increasing the size of the geographical units in the Epidemic-Endemic framework may conceal the heterogeneity in vaccine coverage or immunity within each unit. For instance, in Chapter 4, sub-departmental heterogeneity in vaccine uptake may have biased the association between vaccine coverage and incidence, with vulnerable areas being unnoticeable in a department with high overall coverage. On the other hand, increasing the granularity may lead to inaccurate measures of local vaccine uptake, because of population movements and lack of suitable denominators. Comparing the fits obtained with different geographical scales would give insights into what scale of coverage is most informative of local transmission. If local measures of coverage cannot reliably estimate the level of vaccine-induced immunity in an area, other variables could be collected to act as proxies for vaccine coverage, such as the presence of communities usually associated with lower vaccine coverage (because of vaccine hesitancy or poor access to public health infrastructures and services), socio-economic indicators [34], or the level of local public health expenditure [35]. Tools have also been developed to infer local

vaccination coverage at a detailed geographical scale using household-based surveys [36,37], or aggregated coverage data [38].

The Epidemic-Endemic framework could also be expanded by increasing the number of covariates integrated in each component in order to provide a more complete integration of the factors impacting transmission. For instance, the recent stagnation (or decrease) in vaccine uptake in countries near elimination has been linked to growing levels of vaccine hesitancy [39–41], which were not integrated in the Epidemic-Endemic models. The models could be extended to integrate the impact of changes in vaccine policy, or cuts in funding. Indeed, austerity measures have been linked to decreases in vaccine coverage [35], and adding indicators of local public health expenditure to the covariates would show whether it is also associated with increased risks of incidence. Recently, several European countries, such as France or Italy, have changed their vaccine policy and made measles vaccination mandatory [42]. The effect of this change at a local level could also be evaluated by integrating the change in policy as a covariate of the model. Therefore, the Epidemic-Endemic framework could be used to quantify the relative impact of each of these factors, and analyse which are most correlated with heightened risks of outbreaks.

Furthermore, in the projects presented in Chapters 4 and 5, the vaccine-induced immunity was represented by the coverage of the 1st vaccine dose. Ideally, this immunity would be best described by the coverage at the 2nd dose, however most values of local vaccine uptake at the second dose were missing over the time period considered in the project. A comparison between models integrating each dose might show whether the uptake at one or two doses is more correlated to changes in local incidence, and bring valuable insights into the added value of using the second dose coverage. Finally, the models could account for more sources of immunity if data on the local impact of Supplementary Immunisation Activities (SIAs) were available. Indeed, SIAs can be responsible for a substantial part of immunity in a population [43], and their impact on the risks of transmission could be evaluated by adding this data as a covariate to each component.

Future research could aim to incorporate other key demographic and behavioural variables into *o2geosocial*, and evaluate which epidemiological data are most informative to reconstruct who infected whom. For instance, in the first version of *o2geosocial*, distance between cases only depends on the location reported to the surveillance system. However, connectivity can depend on factors other than distance, such as whether children from under-immunised communities tend to go to the same schools, regardless of the distance between their homes. In this example, their residency may be less informative to reconstruct transmission than their school. Future developments in *o2geosocial* should work towards integrating the impact of the workplace or school in the likelihood procedure. These data are rarely

publicly available but can be routinely collected by national surveillance systems. This could come as an extra factor, estimated by the model, which would quantify the “bonus” in chances of connections between two cases given that they attend the same school or workplace.

In all projects of this thesis, the estimation of the connectivity between regions in both frameworks relied upon using incidence data. However, as highlighted in Chapter 5, this can lead to biases: for instance, the aggregated model over-estimated the number of cross-regional transmissions because it could not identify suitable infectors in the region of origin, which was due to the aggregation and partial detection of the cases. Therefore, the connectivity matrix estimated by this model did not correspond to the input values in the simulations. Mobility patterns are usually described by commuting data, or GPS data taken from mobile phone or social medial. For instance, daily GPS data collected by Facebook through the Data For Good program were recently used to estimate the impact of Non Pharmaceutical Interventions on mobility and the spread of COVID-19 [44,45]. However, using actual mobility data can also lead to selection biases. For instance, children are excluded from commuting and GPS data collected by social media because of data sensitivity. This may be especially relevant for measles outbreaks, during which a large proportion of the reported cases are children. Furthermore, the mobility patterns of cases may not correspond to the daily movements (for instance in case of severe symptoms). Using actual mobility data is therefore especially relevant for pathogens with a proportion of sub-clinical infections (i.e. whose mobility pattern will not affected by the pathogen), and during outbreaks not mainly driven by transmission among children.

Finally, future work could focus on evaluating the impact of integrating different formulations of the Epidemic-Endemic framework in the daily model developed in Chapters and 4 and 5. Firstly, social contact data have been incorporated within the Epidemic-Endemic model to account for heterogeneous contact between age groups in aggregated data [46]. Stratifying by age group could improve the way coverage data is integrated in the model. Indeed, in the current version of the daily model, the three-year average uptake is used to describe the vaccine-induced immunity of the entire population. However, yearly values of vaccine uptake only describe the proportion of children of certain age groups that were vaccinated this year (the reported age groups can vary depending on the country), which may not resemble the vaccine coverage for older age groups, who were vaccinated in previous years. Therefore, the coverage at a given date may be a flawed estimate of the vaccine-induced immunity in a region. In an age-stratified Epidemic-Endemic model, the level of vaccine-induced immunity of a given age group would correspond to the vaccine coverage reported in the year that the age of this cohort was equal to the age at which vaccine coverage is reported, and would remain the same at each year. The vaccine coverage in a given year would not describe the vaccine uptake of the entire population

anymore, but only the coverage among the age group reported in this year, which corresponds to how the yearly vaccine coverage is computed.

Secondly, recent developments of the Epidemic-Endemic model allowed for the inclusion of weights over previous time steps to account for the impact of longer serial intervals on the risks of transmission [9]. A comparison between these models and the daily framework developed in Chapters 4 and 5 would assess whether these weights improve the performance of the aggregated frameworks. Furthermore, in this new version of the framework, the weights can be estimated in the fitting procedure. By adapting this development to the daily framework, one could estimate the proportion of the composite serial interval stemming from direct transmission rather than set a proportion prior to running the model. Indeed, this proportion was arbitrarily set to allow for both direct and indirect transmissions, but sensitivity analyses were needed to show the robustness of the estimates. Estimating the proportion of direct transmission from the case counts would be more sensible and would give insight into the detection rate of the cases.

6.5. Conclusions

This PhD thesis showed that developing models based on routinely collected surveillance data can improve our understanding of local measles transmission dynamics. Using a limited number of widely available data, the methods presented in this project were able to reproduce the spread of measles in different settings, identify local areas most at-risk of outbreaks, and estimate the impact of various factors on transmission. My results suggest that countries near elimination remain at risk of large outbreaks, especially following short-term decreases in vaccination coverage. This is particularly relevant given that the first year of the COVID-19 pandemic was associated with disruption in many routine immunisation programs. Despite the drop in measles transmission reported in various countries during 2020, the risks of measles outbreaks remain high. Methods able to identify pockets of susceptibles and transmission hotspots from limited data are therefore crucial to evaluate the vulnerability of countries near elimination to future outbreaks.

6.6. References

[1] Truscott J, Ferguson NM. Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling. PLoS Comput Biol 2012;8. doi:10.1371/journal.pcbi.1002699.

[2] Edmunds WJ, Kafatos G, Wallinga J, Mossong JR. Mixing patterns and the spread of close-contact infectious diseases. Emerg Themes Epidemiol 2006;3:1–8. doi:10.1186/1742-7622-3-10.

[3] Wallinga J, Teunis P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal. Am J Epidemiol 2004;160:509–16.

[4] Campbell F, Didelot X, Fitzjohn R, Ferguson N, Cori A, Jombart T. outbreaker2: A modular platform for outbreak reconstruction. BMC Bioinformatics 2018;19. doi:10.1186/s12859-018-2330-z.

[5] Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. PLoS Comput Biol 2019. doi:10.1371/journal.pcbi.1006930.

[6] Kenah E, Britton T, Halloran ME, Longini IM. Molecular Infectious Disease Epidemiology: Survival Analysis and Algorithms Linking Phylogenies to Transmission Trees. PLoS Comput Biol 2016;12:1–29. doi:10.1371/journal.pcbi.1004869.

[7] Robert A, Edmunds WJ, Watson CH, Henao-Restrepo AM, Gsell P-S, Williamson E, et al. Determinants of Transmission Risk During the Late Stage of the West African Ebola Epidemic. Am J Epidemiol 2019. doi:10.1093/aje/kwz090.

[8] Taube JC, Miller PB, Drake JM. An open-access database of infectious disease transmission trees to explore superspreader epidemiology. MedRxiv 2021:2021.01.11.21249622.

[9] Bracher J, Held L. Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. Int J Forecast 2020. doi:10.1016/j.ijforecast.2020.07.002.

[10] Meyer S, Held L, Höhle M. hhh4: Endemic-epidemic modeling of areal count time series. J Stat Softw 2016.

[11] Parpia AS, Skrip LA, Nsoesie EO, Ngwa MC, Abah Abah AS, Galvani AP, et al. Spatio-temporal dynamics of measles outbreaks in Cameroon. Ann Epidemiol 2020;42:64-72.e3. doi:10.1016/j.annepidem.2019.10.007.

[12] Herzog SA, Paul M, Held L. Heterogeneity in vaccination coverage explains the size and occurrence of measles epidemics in German surveillance data. Epidemiol Infect 2011;139:505–15. doi:10.1017/S0950268810001664.

[13] World Health Organization. Framework for verifying elimination of measles and rubella. Wkly Epidemiol Rec 2013;88:89–100. doi:10.1371/jour.

[14] Bjørnstad ON, Grenfell BT, Viboud C, King AA. Comparison of alternative models of human movement and the spread of disease. BioRxiv 2019:1–15. doi:10.1101/2019.12.19.882175.

[15] Svensson Å. A note on generation times in epidemic models. Math Biosci 2007;208:300–11. doi:10.1016/j.mbs.2006.10.010.

[16] le Polain de Waroux O, Saliba V, Cottrell S, Young N, Perry M, Bukasa A, et al. Summer music and arts festivals as hot spots for measles transmission: Experience from England and Wales, June to October 2016. Eurosurveillance 2016;21:1–6. doi:10.2807/1560-7917.ES.2016.21.44.30390.

[17] Woudenberg T, Woonink F, Kerkhof J, Cox K, Ruijs WLM. The tip of the iceberg : incompleteness of measles reporting during a large outbreak in The Netherlands in 2013 – 2014. Epidemiol Infect 2018;146:716–22. doi:https://doi.org/10.1017/ S0950268818002698.

[18] Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian Reconstruction of Disease Outbreaks by Combining Epidemiologic and Genomic Data. PLoS Comput Biol 2014;10. doi:10.1371/journal.pcbi.1003457.

[19] Ypma RJF, Bataille AMA, Stegeman A, Koch G, Wallinga J, van Ballegooijen WM. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. Proc R Soc B Biol Sci 2012;279:444–50. doi:10.1098/rspb.2011.0913.

[20] Lau MSY, Marion G, Streftaris G, Gibson G. A Systematic Bayesian Integration of Epidemiological and Genetic Data. PLoS Comput Biol 2015. doi:10.1371/journal.pcbi.1004633.

[21] Vasylyeva TI, Friedman SR, Paraskevis D, Magiorkinis G. Integrating molecular epidemiology and social network analysis to study infectious diseases: Towards a socio-molecular era for public health. Infect Genet Evol 2016;46:248–55. doi:10.1016/j.meegid.2016.05.042.

[22] Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S. A Bayesian Inference Framework to Reconstruct Transmission Trees Using Epidemiological and Genetic Data. PLoS Comput Biol 2012;8. doi:10.1371/journal.pcbi.1002768.

[23] Kendall M, Ayabina D, Colijn C. Estimating transmission from genetic and epidemiological data: a metric to compare transmission trees 2016:1–22. doi:10.1214/17-STS637.

[24] Hiebert J, Severini A. Measles molecular epidemiology: What does it tell us and why is it important? Canada Commun Dis Rep 2014;40:257–60. doi:10.14745/ccdr.v40i12a06.

[25] Brown KE, Rota PA, Goodson JL, Williams D, Abernathy E, Takeda M, et al. Genetic characterization of measles and rubella viruses detected through global measles and rubella elimination surveillance, 2016-2018. Morb Mortal Wkly Rep 2019;68:587–91. doi:10.15585/mmwr.mm6826a3.

[26] World Health Organization. Framework for verifying elimination of measles and rubella. Wkly Epidemiol Rec 2013;88:89–100. doi:10.1371/jour.

[27] Centers for Disease Control and Prevention (CDC). Measles Cases and Outbreaks n.d. https://www.cdc.gov/measles/cases-outbreaks.html (accessed May 22, 2021).

[28] Nicolay N, Mirinaviciute G, Mollet T, Celentano LP, Bacci S. Epidemiology of measles during the COVID-19 pandemic, a description of the surveillance data, 29 EU/ EEA countries and the United Kingdom, January to May 2020. Eurosurveillance 2020;25. doi:10.2807/1560-7917.ES.2020.25.31.2001390.

[29] Statement by the Measles & Rubella Initiative: American Red Cross, U.S. CDC, UNICEF UF and W. More than 117 million children at risk of missing out on measles vaccines, as COVID-19 surges n.d. https://www.who.int/immunization/diseases/measles/statement_missing_measles_vaccines_covid-19/en/ (accessed May 22, 2021).

[30] Santoli JM, Lindley MC, DeSilva MB, Kharbanda EO, Daley MF, Galloway L, et al. Effects of the COVID-19 Pandemic on Routine Pediatric Vaccine Ordering and Administration — United States, 2020. MMWR Morb Mortal Wkly Rep 2020;69:591–3. doi:10.15585/mmwr.mm6919e2.

[31] Robert A, Funk S, Kucharski AJ. The measles crisis in Europe—the need for a joined-up approach. Lancet 2019;393:2033. doi:10.1016/S0140-6736(19)31039-6.

[32] Mankertz A, Mihneva ZR, Gold H, Baumgarte S, Baillot A, Helble R, et al. Spread of measles virus D4-Hamburg, Europe, 2008-2011. Emerg Infect Dis 2011;17:1396–401. doi:10.3201/eid1708.101994.

[33] Durrheim DN, Andrus JK, Tabassum S, Bashour H, Githanga D, Pfaff G. A dangerous measles future looms beyond the COVID-19 pandemic. Nat Med 2021;27:360–1. doi:10.1038/s41591-021-01237-5.

[34] Bocquier A, Ward J, Raude J, Peretti-Watel P, Verger P. Socioeconomic differences in childhood vaccination in developed countries: a systematic review of quantitative studies. Expert Rev Vaccines 2017;16:1107–18. doi:10.1080/14760584.2017.1381020.

[35] Toffolutti V, McKee M, Melegaro A, Ricciardi W, Stuckler D. Austerity, measles and mandatory vaccination: Cross-regional analysis of vaccination in Italy 2000-14. Eur J Public Health 2019;29:123–7. doi:10.1093/eurpub/cky178.

[36] Sbarra AN, Rolfe S, Nguyen JQ, Earl L, Galles NC, Marks A, et al. Mapping routine measles vaccination in low- and middle-income countries. Nature 2021;589:415–9. doi:10.1038/s41586-020-03043-4.

[37] Takahashi S, Metcalf CJE, Ferrari MJ, Tatem AJ, Lessler J. The geography of measles vaccination in the African Great Lakes region. Nat Commun 2017;8:1–9. doi:10.1038/ncomms15585.

[38] Utazi CE, Thorley J, Alegana VA, Ferrari MJ, Nilsen K, Takahashi S, et al. A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping. Stat Methods Med Res 2019;28:3226–41. doi:10.1177/0962280218797362.

[39] Hotez PJ, Nuzhath T, Colwell B. Combating vaccine hesitancy and other 21st century social determinants in the global fight against measles. Curr Opin Virol 2020;41:1–7. doi:10.1016/j.coviro.2020.01.001.

[40] Omer SB, Salmon DA, Orenstein WA, Halsey N. Vaccine Refusal, Mandatory Immunization, and the Risks of Vaccine-Preventable Diseases 2009.

[41] McKee M, Ricciardi W, Siciliani L, Rechel B, Toffolutti V, Stuckler D, et al. Increasing vaccine uptake: confronting misinformation and disinformation. Eurohealth (Lond) 2018;24:35–8.

[42] Vaz OM, Ellingson MK, Weiss P, Jenness SM, Bardají A, Bednarczyk RA, et al. Mandatory vaccination in Europe. Pediatrics 2020;145. doi:10.1542/peds.2019-0620.

[43] Trentini F, Poletti P, Merler S, Melegaro A. Measles immunity gaps and the progress towards elimination: a multi-country modelling analysis. Lancet Infect Dis 2017;17:1089–97. doi:10.1016/S1473-3099(17)30421-8.

[44] Jeffrey B, Walters CE, Ainslie KEC, Eales O, Ciavarella C, Bhatia S, et al. Anonymised and aggregated crowd level mobility data from mobile phones suggests that initial compliance with covid-19 social distancing interventions was high and geographically consistent across the UK. Wellcome Open Res 2021;5:1–10. doi:10.12688/WELLCOMEOPENRES.15997.1.

[45] Chang MC, Kahn R, Li YA, Lee CS, Buckee CO, Chang HH. Variation in human mobility and its impact on the risk of future COVID-19 outbreaks in Taiwan. BMC Public Health 2021;21:1–10. doi:10.1186/s12889-021-10260-7.

[46] Meyer S, Held L. Incorporating social contact data in spatio-temporal models for infectious disease spread. Biostatistics 2017;18:338–51. doi:10.1093/biostatistics/kxw051.

Supplementary Material Chapter 2

S1. Sensitivity analysis

We re-ran the analysis using different distributions of the incubation period and the serial intervals in order to assess the impact of these distributions on the ability of the models to reconstruct the transmission clusters.

Firstly, narrowing both distributions led to increasing the discrepancies between the simulated and inferred transmission trees (Figure S1). Although the cluster size distributions were similar, especially when the importation status was known prior to fitting the model, the proportion of iteration when cases were linked to the "right" infector dropped: about 10 of the cases were never linked to the right infector. However, most cases were linked to an infector from the "right" cluster in most of the simulations. Some transmissions rarely generated in the simulations have become completely impossible in the inferred trees, for instance longer serial intervals, or short incubation periods, which explains why certain simulated links were never generated in the inferred trees.

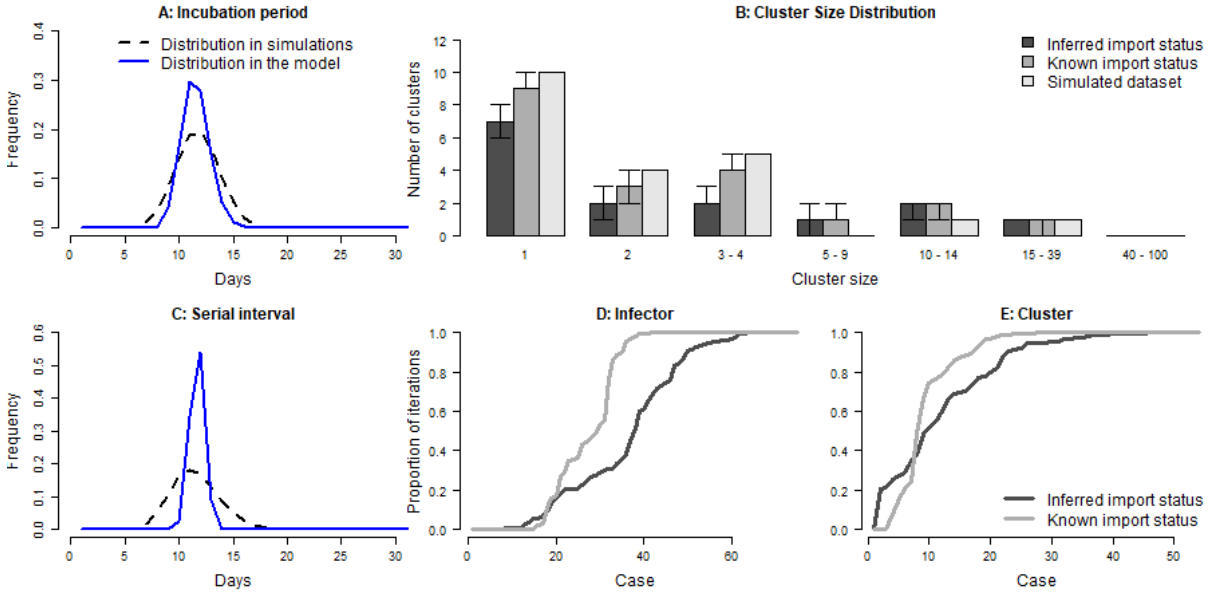

Figure S1: Description of the model fits obtained using lower standard deviations for both the incubation period and the serial interval ($f \sim Gamma(0.14, 80)$ and $w \sim Normal(11.7, 0.7)$). Panel A: Distribution of the incubation period f, used in the simulations and in the model. Panel B: Comparison between the cluster size distribution inferred using narrow distributions and the reference data. Panel C: Distribution of the generation time w. Panel D: Proportion of iterations with the correct index for each case; Panel E: Proportion of iterations where the index is from the correct cluster.

On the other hand, increasing the standard deviation of both distributions made little difference compared to the reference model fits (Figure S2). The cluster size distributions inferred in both models were similar to the fits presented in the Main text. We observed a slight decrease in the proportion of iterations where cases were connected to the right infector. This may be due to the fact that cases have more likely potential infectors than in the reference fits, given that the temporal component is not as

informative. However, these differences remain minor, and the actual infector is part of the pool of potential infectors for almost every case in both models.

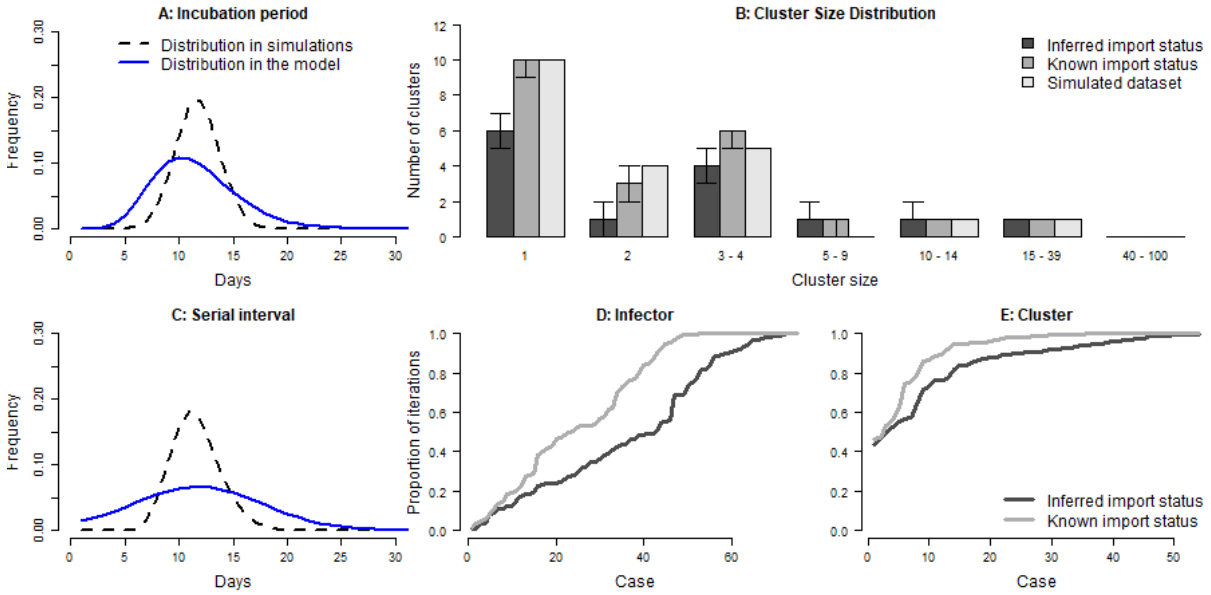

Figure S2: Description of the model fits obtained using higher standard deviations for both the incubation period and the serial interval ($f \sim Gamma(1.29, 8.9)$ and $w \sim Normal(11.7, 6)$). Panel A: Distribution of the incubation period f, used in the simulations and in the model. Panel B: Comparison between the cluster size distribution inferred using wide distributions and the reference data. Panel C: Distribution of the generation time w. Panel D: Proportion of iterations with the correct index for each case; Panel E: Proportion of iterations where the index is from the correct cluster.

Finally, we generated the analysis using distributions with same standard deviation as the reference fits, but changing the mean (Figure S3 and Figure S4). In both cases, we did not observe any major change in the cluster size distribution compared to the Main analysis. The proportion of iterations where cases were connected to an infector from the correct cluster did not change for most cases. Some connections between cases were more rarely represented in the inferred transmission trees, but two third of the cases were connected to the correct infector in more than 60% of the inferred tree (if the importation status was known), which was similar to the reference fits.

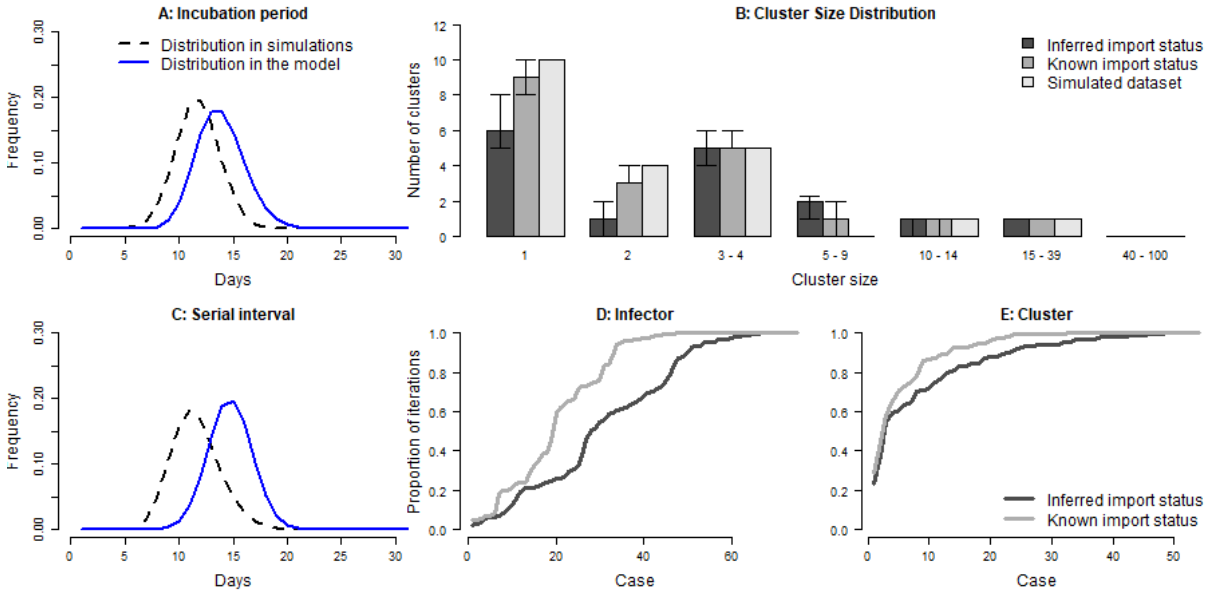

Figure S3: Description of the model fits obtained using later means for both the incubation period and the serial interval ($f \sim Gamma(0.35, 39.5)$ and $w \sim Normal(14.7, 2)$). Panel A: Distribution of the incubation period f, used in the simulations and in the model. Panel B: Comparison between the cluster size distribution inferred using distributions with later means and the reference data. Panel C: Distribution of the generation time w. Panel D: Proportion of iterations with the correct index for each case; Panel E: Proportion of iterations where the index is from the correct cluster.

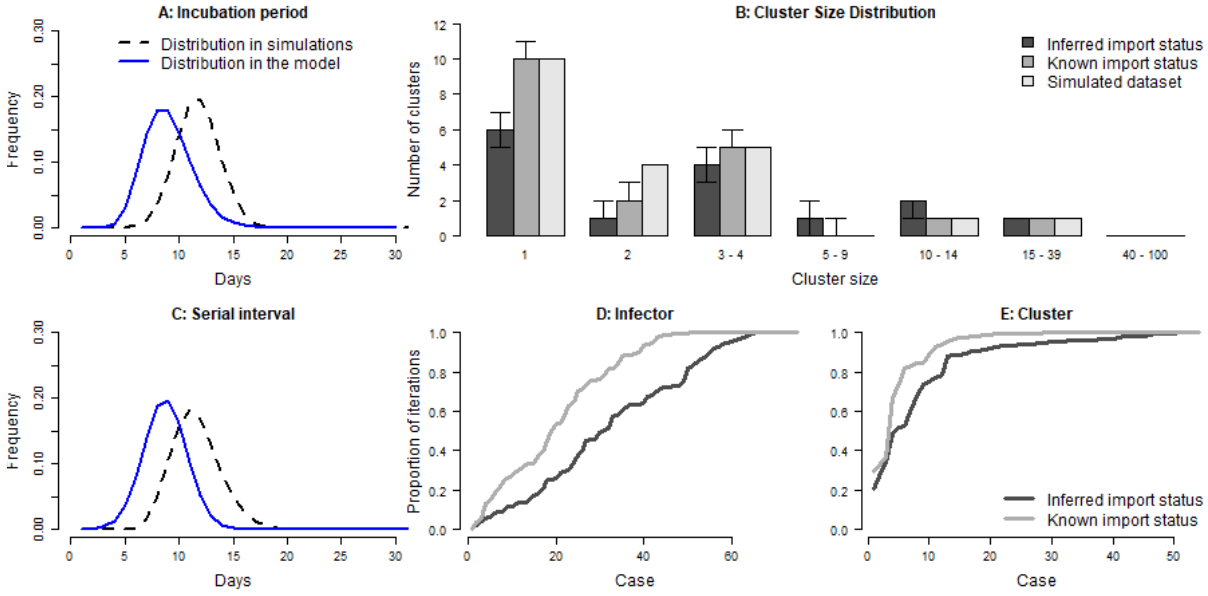

Figure S4: Description of the model fits obtained using earlier means for the incubation period and the serial interval ($f \sim Gamma(0.55, 16.3)$ and $w \sim Normal(8.7, 2)$). Panel A: Distribution of the incubation period f, used in the simulations and in the model. Panel B: Comparison between the cluster size distribution inferred using distributions with earlier means and the reference data. Panel C: Distribution of the generation time w. Panel D: Proportion of iterations with the correct index for each case; Panel E: Proportion of iterations where the index is from the correct cluster.

S2. Comparison local number of secondary cases

In Figure S5, we compared the 95% credible intervals of the average number of secondary transmissions per region in the models to the simulated data. The 95% credible intervals were generated by calculating the average number of secondary transmissions per region in each iteration of the transmission trees

(excluding the burnin phase), and computing the 2.5% and 97.5% quantiles. We observed that both models were able to identify the regions most often associated with secondary transmissions, and in both models, only one simulated regional value did not fall within the 95% Credible Intervals.

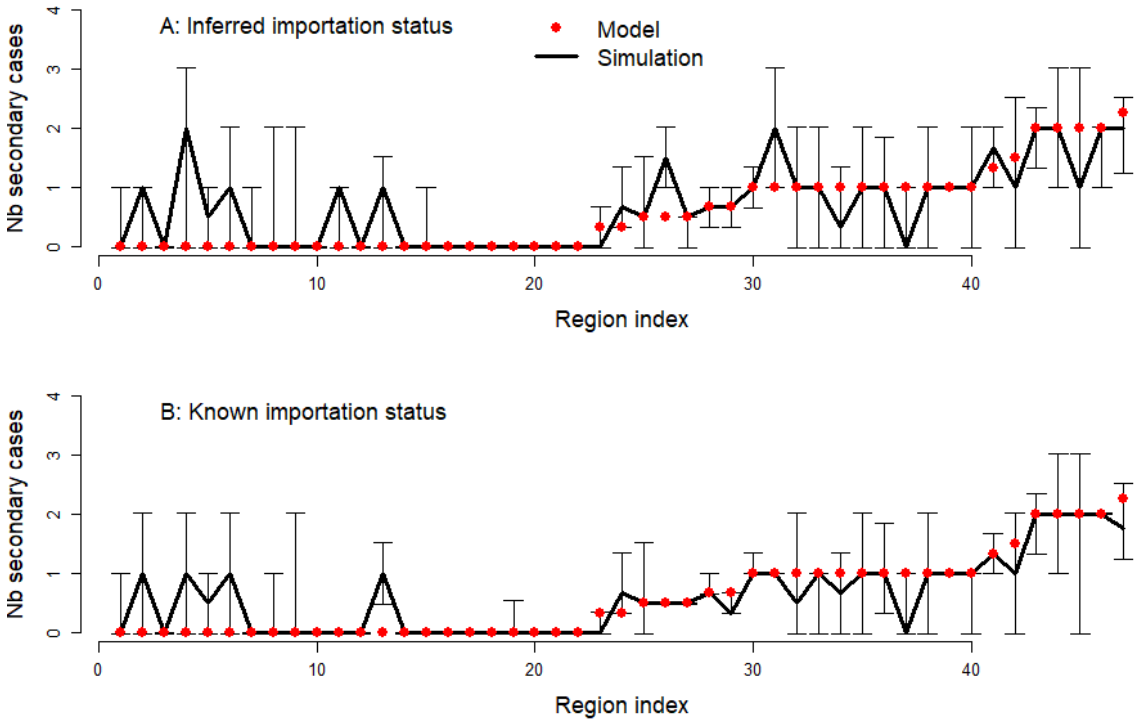

Figure S5: Average number of secondary cases per region in the simulations (red dots) and the models. The black line links the median estimates, the arrows indicate the 95% Credible Intervals. Panel A: Model 1, with inferred importation status; Panel B: Model 2, with known importation status.

Supplementary Material Chapter 3

S1. Description of the US dataset

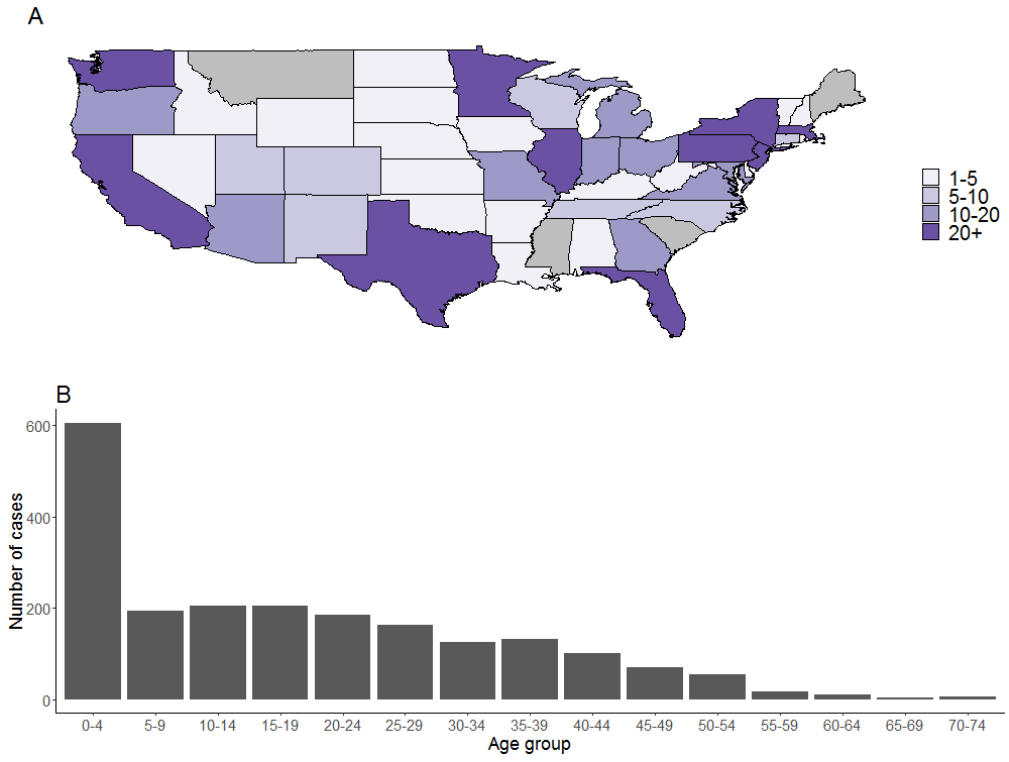

Figure S1: Description of the cases and clusters reported in the United States between 2001 and 2016. Panel A: number of clusters reported per state in the contact tracing investigations (5 clusters cover several states) and Panel B: Age distribution of the cases. Data from CDC's National Notifiable Disease Surveillance System. Hawaii and Alaska and not shown in Panel A.

S2. Evaluation cluster matching

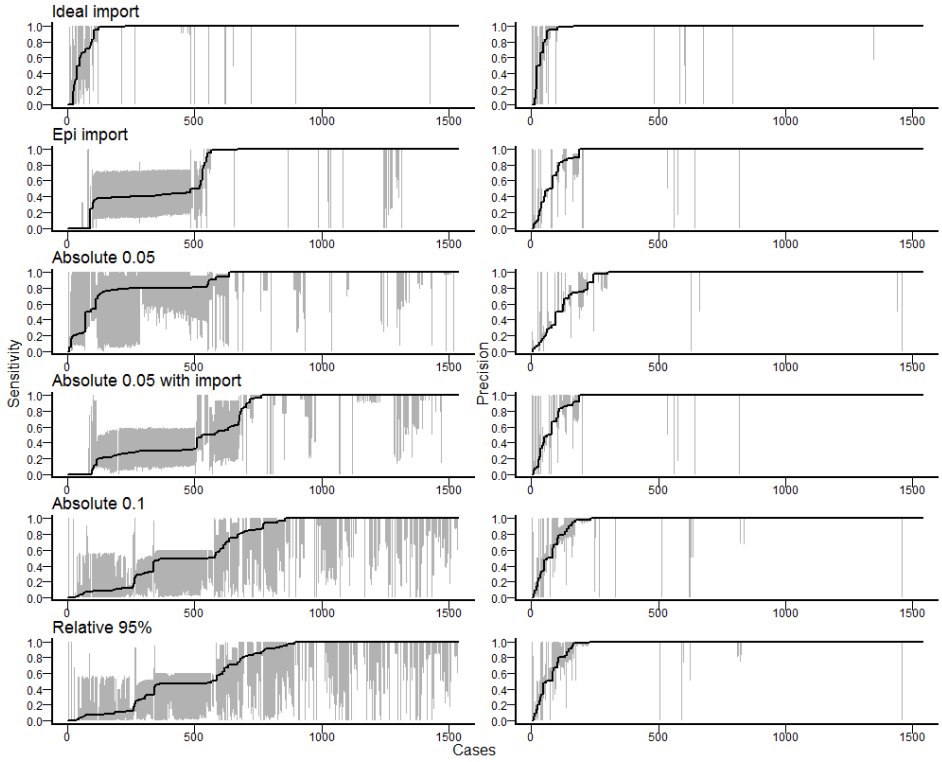

Figure S2: Confidence intervals of the sensitivity (right panels) and precision (left panels) of the clusters of each case in the different runs. For each case, the sensitivity is the proportion of cases from the reference cluster that were correctly inferred, the precision is the proportion of cases from the inferred cluster that were part of reference cluster. Grey areas represent the 95% credibility intervals, and the black line represents the median values of sensitivity or precision across all iterations.

S3. Posterior distribution and convergence

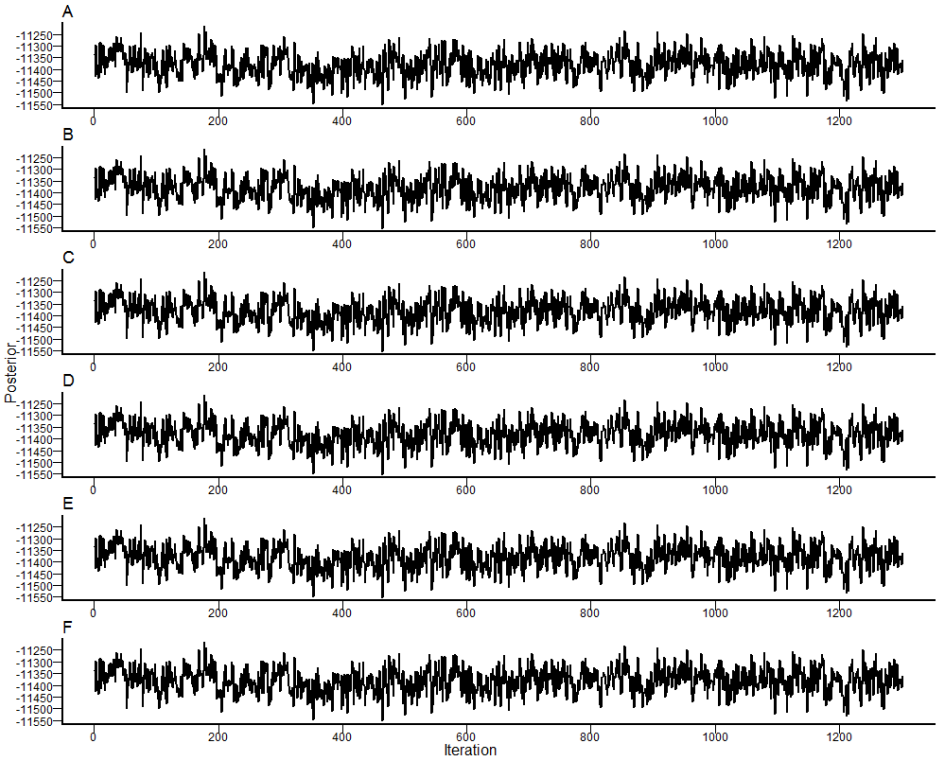

Figure S3: Trace of the posterior value of the MCMC runs after removing the burnin section and thinning. A: Ideal imports; B: Epi imports; C: Absolute threshold, $k = 0.05$ *D: Absolute threshold,* $k = 0.05$ *with prior information on imports; E: Absolute threshold,* $k = 0.1$*; F: Relative threshold upper* $\lambda = 97.5\%$ *quantile.*

S4. Clusters stratified by state

Scenario 1

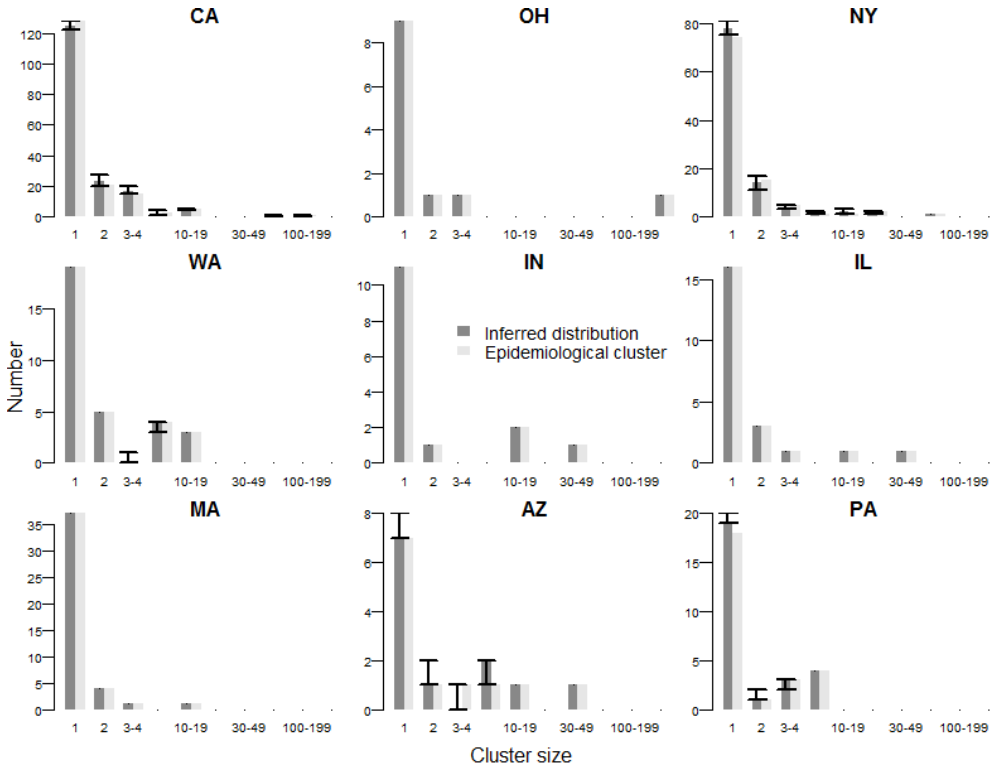

Figure S4: Comparison of the inferred and reference cluster size distributions in the nine states with most cases declared between 2001 and 2016 in scenario 1. Arrows represent the 95% credibility intervals of each estimate.

Scenario 2

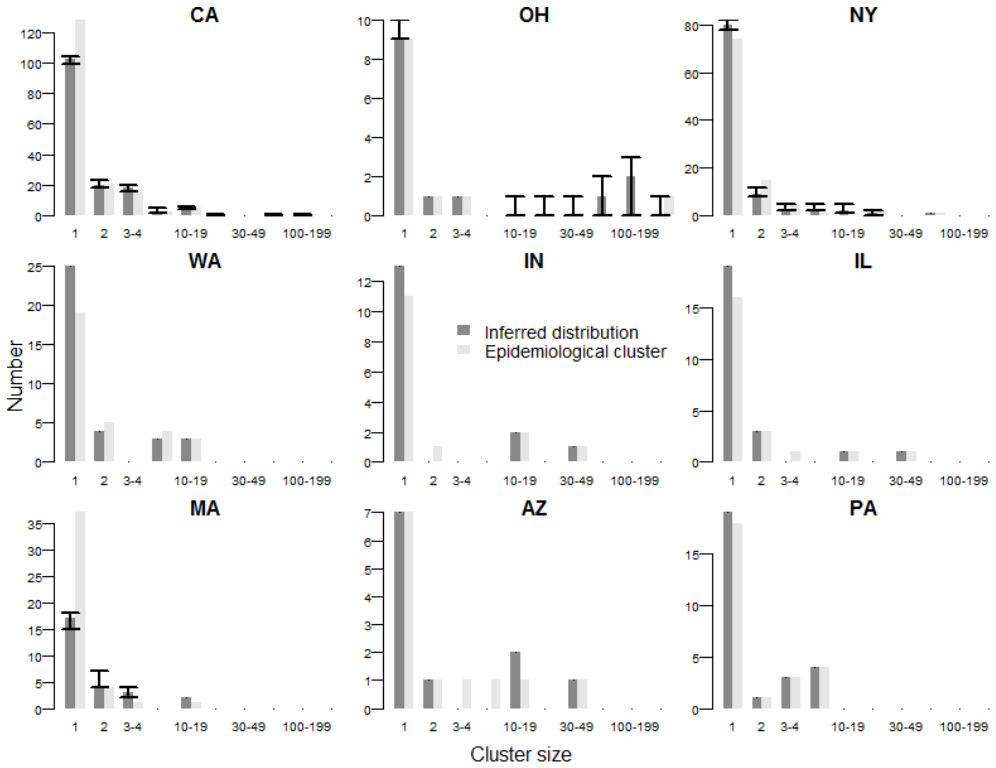

Figure S5: Comparison of the inferred and reference cluster size distributions in the nine states with most cases declared between 2001 and 2016 in scenario 2. Arrows represent the 95% credibility intervals of each estimate.

Scenario 3

Absolute threshold $k = 0.05$

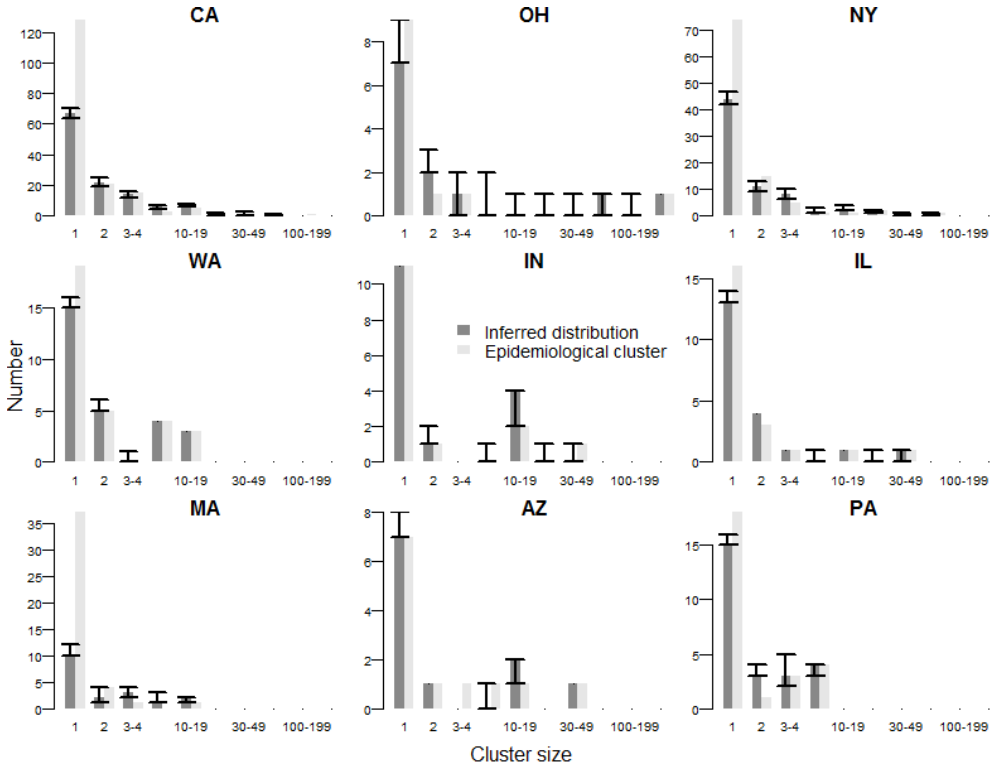

Figure S6: Comparison of the inferred and reference cluster size distributions in the nine states with most cases declared between 2001 and 2016 in scenario 3, with an absolute threshold $k = 0.05$ *and no prior information on the import status of cases. Arrows represent the 95% credibility intervals of each estimate.*

Relative threshold $\boldsymbol{k = 95\%}$

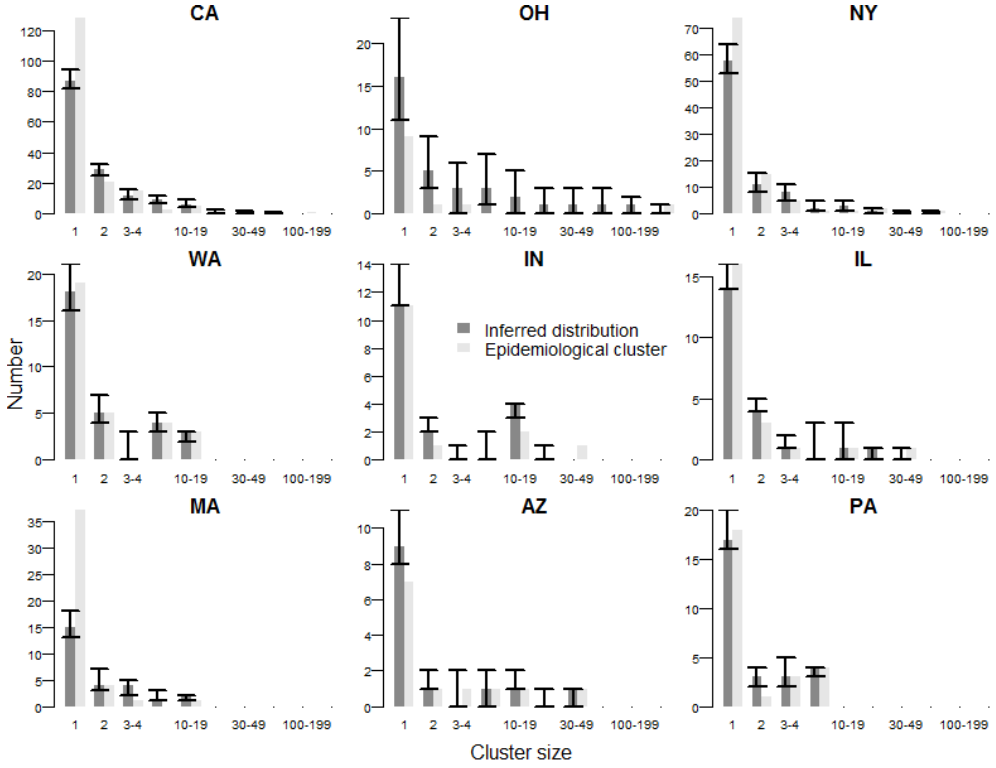

Figure S7: Comparison of the inferred and reference cluster size distributions in the nine states with most cases declared between 2001 and 2016 in scenario 3, with a relative threshold $k = 95\%$ *and no prior information on the import status of cases. Arrows represent the 95% credibility intervals of each estimate.*

Epidemiological imports, and absolute threshold $\boldsymbol{k = 0.05}$

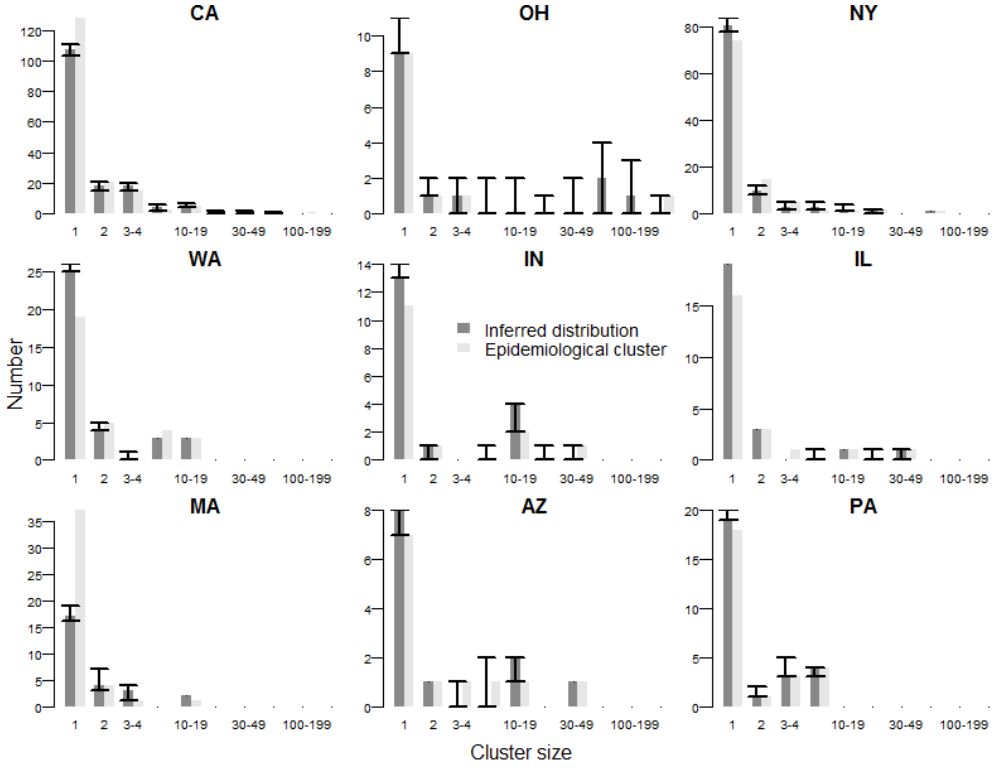

Figure S8: Comparison of the inferred and reference cluster size distributions in the nine states with most cases declared between 2001 and 2016 in scenario 3, with an absolute threshold $k = 0.05$ *and using the import status distribution from the contact tracing investigations as prior information. Arrows represent the 95% credibility intervals of each estimate.*

S5. Parameter estimates

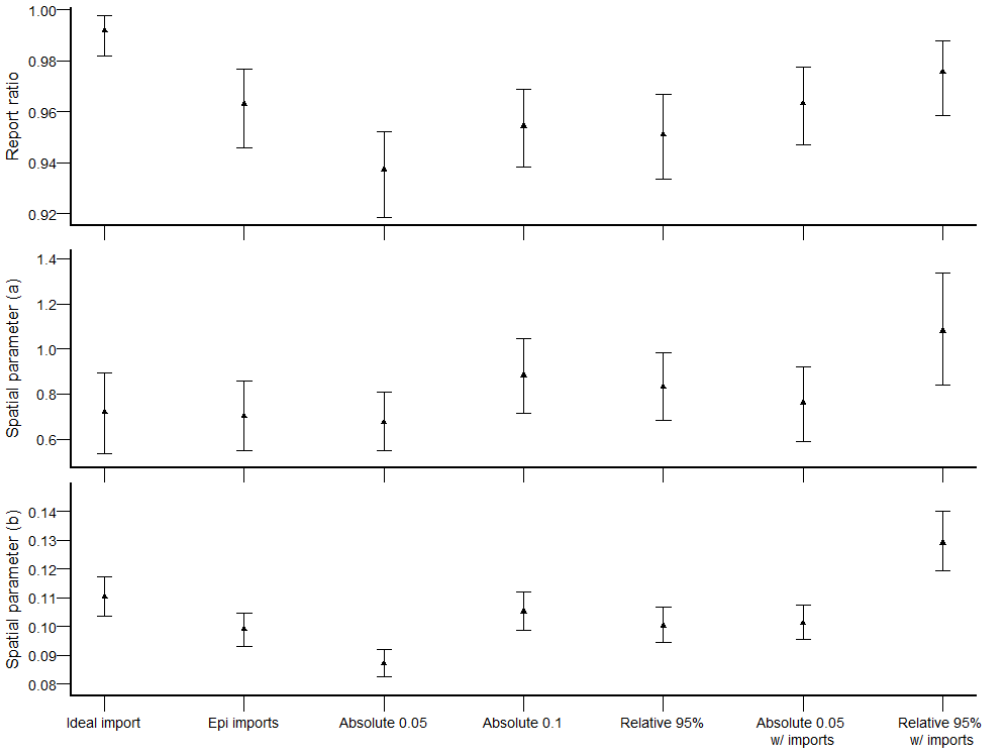

Figure S9: Estimation of A) the report ratio ρ, B) the spatial parameter a and C) the spatial parameter b in each scenario. The dots represent the median estimate, and the arrows correspond to the 95% credibility interval. The estimates were obtained after burnin and thinning.

S6. Distance between transmission

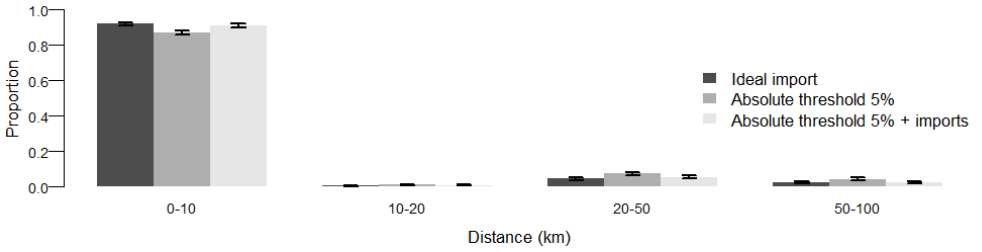

Figure S10: Distribution of the distance between connected cases in scenario 1, scenario 3 without prior and scenario 3 with prior information. Arrows correspond to the 95% credibility interval.

S7. Impact of different components of likelihood

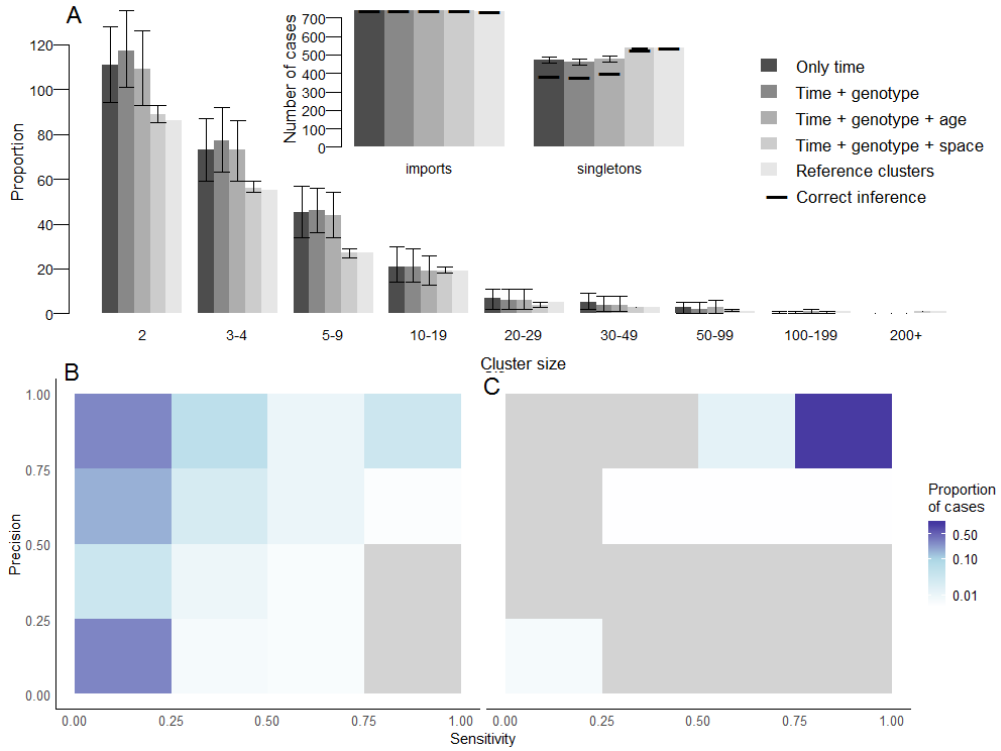

Figure S11: Description of transmission clusters inferred excluding certain components of the likelihood. In this example, the import status of import was not inferred, and taken from the ancestor in each cluster (Scenario 1). Panel A: Cluster size distribution using 1) only the time component of likelihood; 2) time and genotype; 3) time genotype and age; 4) time, genotype and space; compared to the reference clusters (lighgrey). Arrows represent the 95% credibility intervals of each estimate. Only clusters containing at least 2 cases are represented. Inset: Number of imports and number of isolated cases (singletons). For each scenario, the horizontal dark line represents the number of imports that are also imports in the reference clusters, same for singletons. Panel B: Heatmap representing the precision and sensitivity of the clusters for each case when only time and genotype are used to infer the transmission clusters, the cases were classified in a category depending on the proportion of their reference cluster that were inferred in the same cluster (x-axis) and the proportion of mismatches in the inferred cluster. Panel C: Same for time, genotype and space.

S8. Number of secondary transmissions, overall and per state

Overall

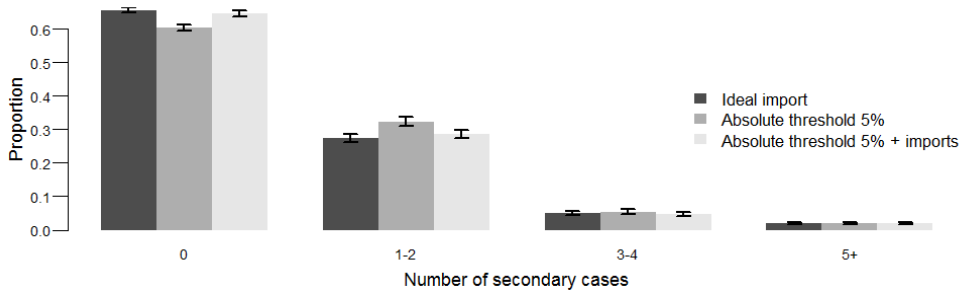

Figure S12: Distribution of the number of secondary cases caused by each case in scenario 1, scenario 3 without prior and scenario 3 with prior information. Arrows correspond to the 95% credibility interval. Cases were classified in groups of transmitter (no further transmission; 1-2 subsequent cases; 3-4 subsequent cases and more than 5 subsequent cases).

Maps per state

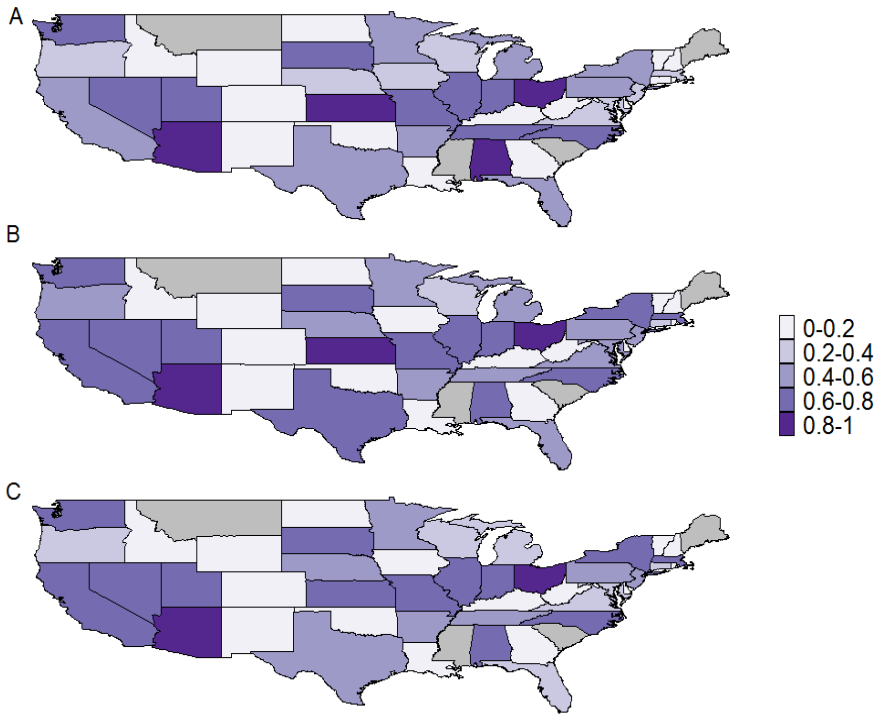

Figure S13: Average number of secondary cases caused by each case stratified by state in A) scenario 1, B) scenario 3 without prior and C) scenario 3 with prior information.

S9. Impact of the proportion of genotype reported. Inference on simulated data

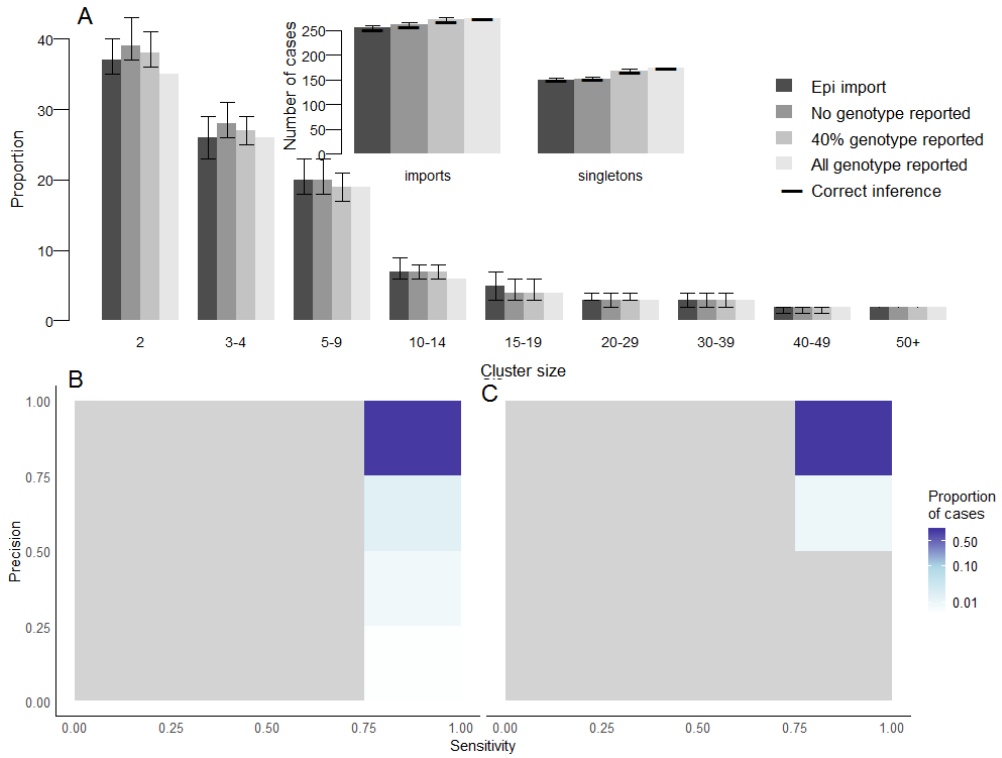

Figure S14: Description of transmission clusters inferred using simulated data (toy_outbreak_long in the o2geosocial package), depending on the proportion of genotyped cases in the data. Panel A: Cluster size distribution when 1) none of the cases was genotyped, 2) 40% of the cases were genotyped (similar to the US dataset) 3) All the cases were genotyped; compared to the reference clusters (lighgrey). Arrows represent the 95% credibility intervals of each estimate. Only clusters containing at least 2 cases are represented. Inset: Number of imports and number of isolated cases (singletons). For each scenario, the horizontal dark line represents the number of imports that are also imports in the reference clusters, same for singletons. Panel B: Heatmap representing the precision and sensitivity of the clusters for each case when no genotype was reported are used to infer the transmission clusters, the cases were classified in a category depending on the proportion of their reference cluster that were inferred in the same cluster (x-axis) and the proportion of mismatches in the inferred cluster. Panel C: Same when all cases were genotyped.

Supplementary Material Chapter 4

S1. Sensitivity analysis: Composite serial interval

In the main analysis, we considered that 50% of the composite serial interval reflected direct transmission (without missing generations between cases), and 50% came from the two scenarios with unreported cases. In order to analyse the impact of the proportion of direct transmission in the composite serial interval, we fitted Model 1 and Model 2 using different composite serial intervals, and reported the fitted distributions of the parameters. We computed ten different composite serial intervals with the proportion of direct transmission increasing from 10% to 100% by increments of 10%.

The impact of the covariates on the risks of transmission was robust to changes in the composite intervals for both Model 1 and Model 2 (Figure S1 and Figure S2). The median estimate of each parameter was included in the 95% confidence interval of the reference results (when the proportion of direct transmission in the composite serial interval is 50%). The only parameter that was impacted was the overdispersion parameter, which increased for most of the fits. This could indicate a wider difference between the mean estimate and the data, which would be taken into account by increasing the dispersion of the negative binomial distribution.

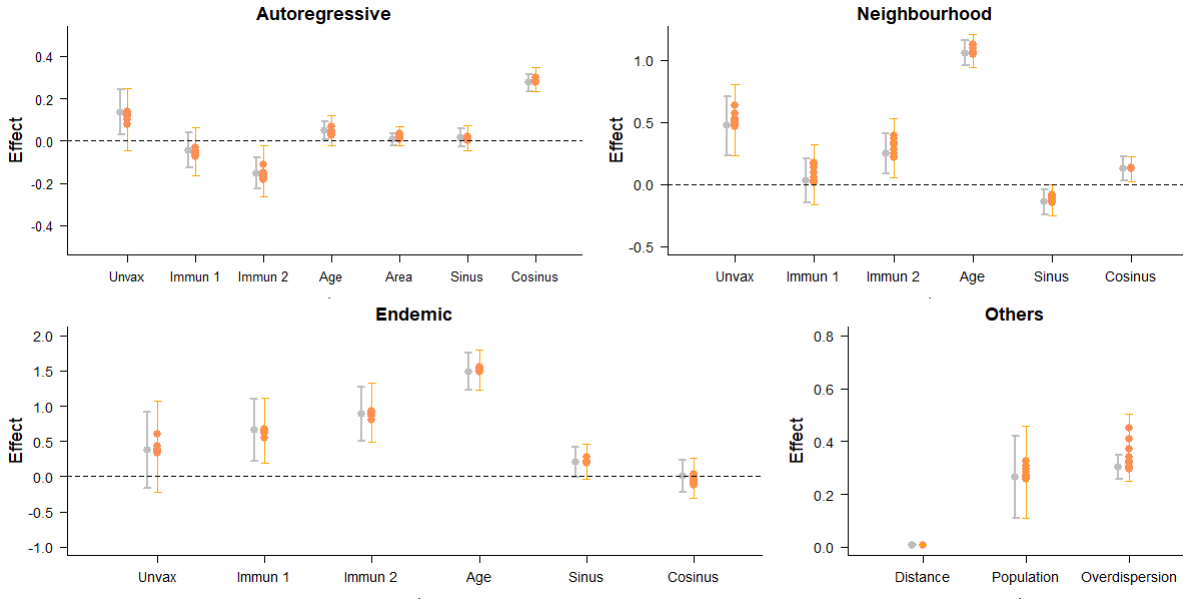

Figure S1: Estimates of the parameters in each component of Model 1, using the reference fit (grey) and ten different values of the composite interval (orange). unvax corresponds to the effect of $u_{i,t}$, the mean proportion unvaccinated over the three years before t in i; incid1 and incid2 correspond to the effect of $N_{i.t}^{1}$ and $N_{i.t}^{2}$ the category of incidence in the three years before t in i; pop corresponds to the effect of $m_{i,t}$ the number of inhabitants at t in i; area corresponds to the effect of the surface; sin and cos correspond to the effects of seasonality; distance and population correspond to the spatial parameters of the connectivity matrix w (δ and γ); overdisp is the estimate of the log-overdispersion parameter in the negative binomial distribution of $Y_{i,t}$. Dots show the mean values associated with the parameters; arrows show the 95% Confidence interval. The orange arrows indicate the extreme values of the 95% confidence interval obtained using different distributions of the composite serial intervals.

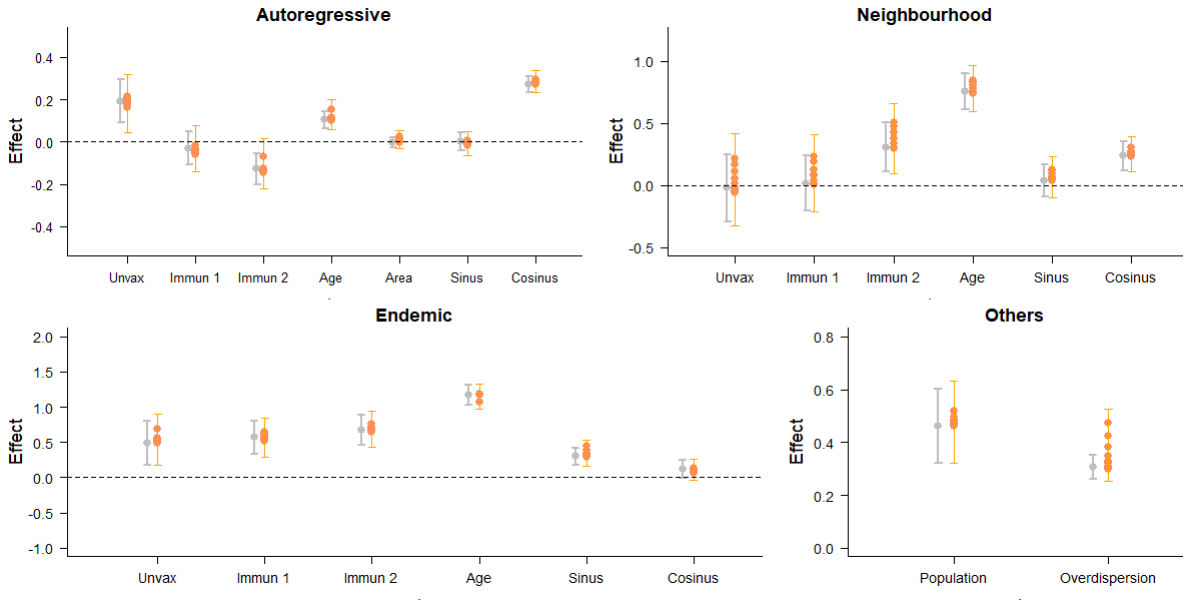

Figure S2: Estimates of the parameters in each component of Model 2, using the reference fit (grey) and ten different values of the composite interval (orange). unvax corresponds to the effect of $u_{i,t}$*, the mean proportion unvaccinated over the three years before t in i; incid1 and incid2 correspond to the effect of* $N_{i,t}^1$ *and* $N_{i,t}^2$ *the category of incidence in the three years before t in i; pop corresponds to the effect of* $m_{i,t}$ *the number of inhabitants at t in i; area corresponds to the effect of the surface; sin and cos correspond to the effects of seasonality; distance and population correspond to the spatial parameters of the connectivity matrix w (δ and γ); overdisp is the estimate of the log-overdispersion parameter in the negative binomial distribution of* $Y_{i,t}$*. Dots show the mean values associated with the parameters; arrows show the 95% Confidence interval. The orange arrows indicate the extreme values of the 95% confidence interval obtained using different distributions of the composite serial intervals.*

S2. Inference of missing data in the regional vaccine coverage

We used publicly available data on 1st dose vaccine uptake between 2004 and 2017 in each department of metropolitan France to calculate the average local vaccine coverage over the past three years. There was no reported value of coverage reported in 2009, therefore the average in 2010, 2011 and 2012 were calculated using only two previous years. Besides 2009, there were 208 missing entries between 2006 and 2017 (17% of all entries), some regions had three years of consecutive unreported coverage, which made it impossible to compute an average without inferring the missing values. Since the incidence dataset starts in 2009, we did not infer the missing coverage data prior to 2006 (Figure S3).

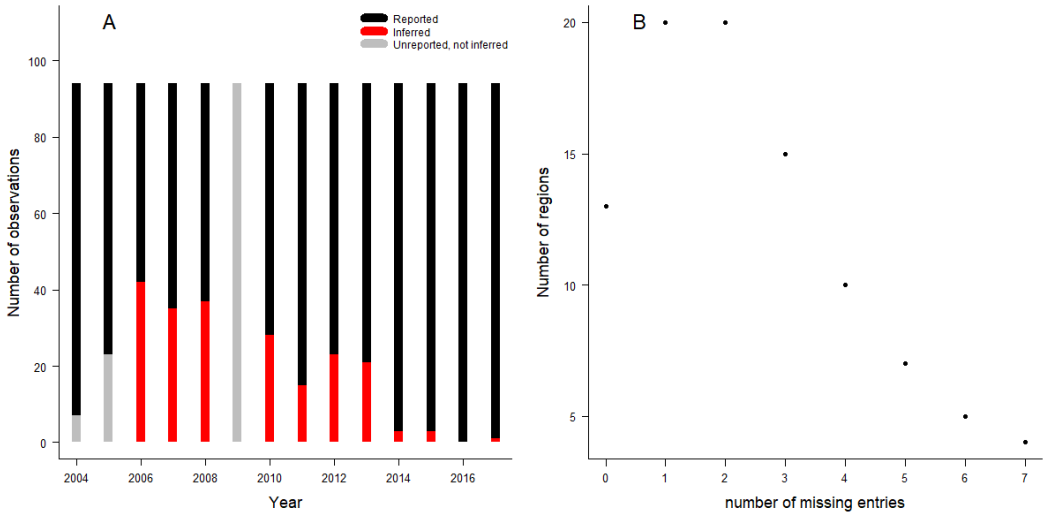

Figure S3: Temporal and spatial distribution of the missing data: Panel A: Number of missing observations per year, missing coverage in 2004 and 2005 did not need to be inferred in the model, and since 2009 was entirely missing, we did not infer any value that year. Panel B: Number of missing entries per region.

We implemented a beta mixed model to fit the annual local values of coverage. We used a beta regression since it is most adapted to modelling proportions or percentages. This model was implemented using the R package *glmmTMB* [1]. Observations are clustered over time within a region. The explanatory variables were orthogonal polynomials of degree 2 over the years covered by the data (t varies between 1 and 14).

$$\text{logit}\,(Y_{ij}) = \beta_{0j} + \beta_{1j}t_i + \beta_{2j}t_i^2$$

Where:

$$\beta_{0j} = \gamma_{00} + U_{0j}$$
$$\beta_{1j} = \gamma_{10} + U_{1j}$$
$$\beta_{2j} = \gamma_{20} + U_{2j}$$

Using two-degree orthogonal polynomials gave more flexibility to the fitted curve than only using linear values of time. The regional values of intercept, and the impact of the orthogonal polynomial varied depending on the area, and were distributed around the fixed effect (Top panel of Figure S4). We show the average fitted trajectory of the vaccine coverage through time, along with the fit in three regions. This highlights that, although the average trajectory slightly increased between 2004 and 2017, there was no abrupt change in the first dose vaccine uptake. Nevertheless, using random effects allows for flexibility in other regions, whereby the fitted trajectories can show greater changes (e.g. region 2 in Figure S4).

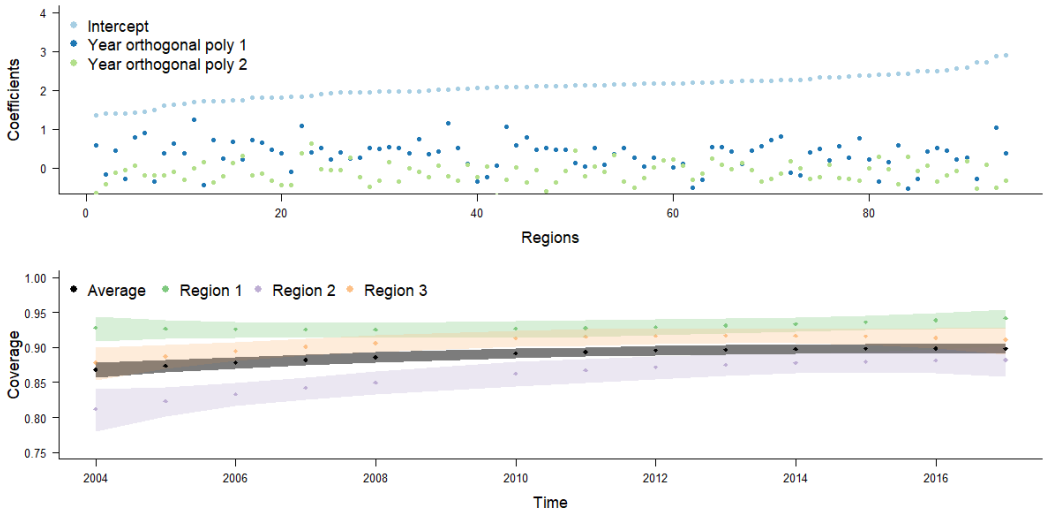

Figure S4: Panel A: Values of the three parameters of the regression for each region. Panel B: Estimated values of coverage between 2004 and 2017, the dots represent the mean estimate, shaded areas correspond to the 95% Confidence Intervals. The purple, orange and green areas representing three departments illustrate how the changes through time can differ depending on the region.

There was no discontinuous jump in the distribution of the random effect distribution for the three parameters of the model (Top panel of Figure S4). The fitted residuals plot did not show any clear trend relative to the dispersion of the data (Figure S5). We used different diagnostic tools provided by the R package *DHARMA* to test whether the model was correctly specified using simulated residuals. The outlier and the dispersion tests did not show any discrepancy, but the Kolgomorov Smirnov test of uniformity showed significant deviation from the expected distribution of residuals. This was expected from the minor discrepancies shown on the QQ plot (right panel of Figure S5).

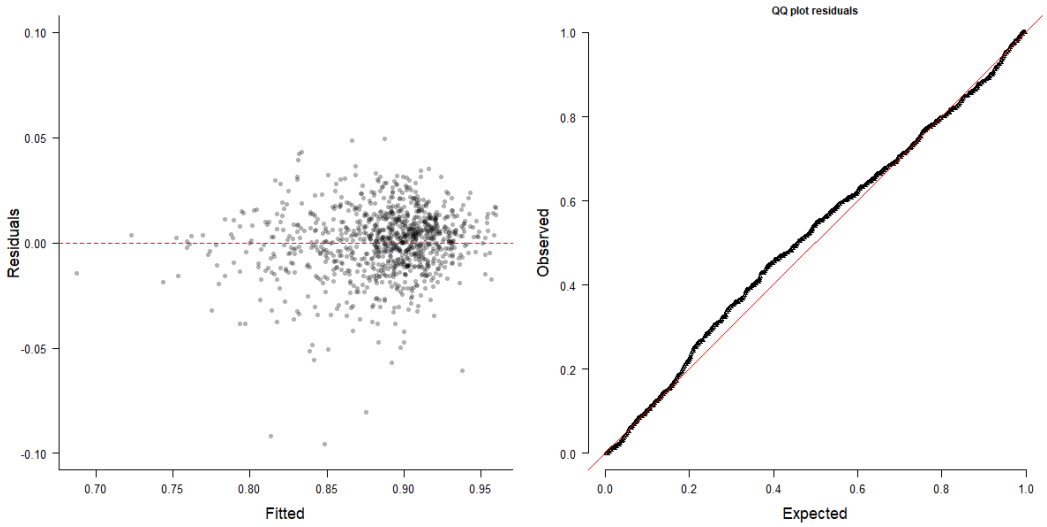

Figure S5: Diagnosis of the regression on the vaccine coverage. Left panel: Fitted vs residuals plot, Left panel: uniform quantile-quantile plot.

Finally, the stratification of residuals by year did not show any trend, which indicates the fit was consistent throughout the different years included in the dataset (Figure S6).

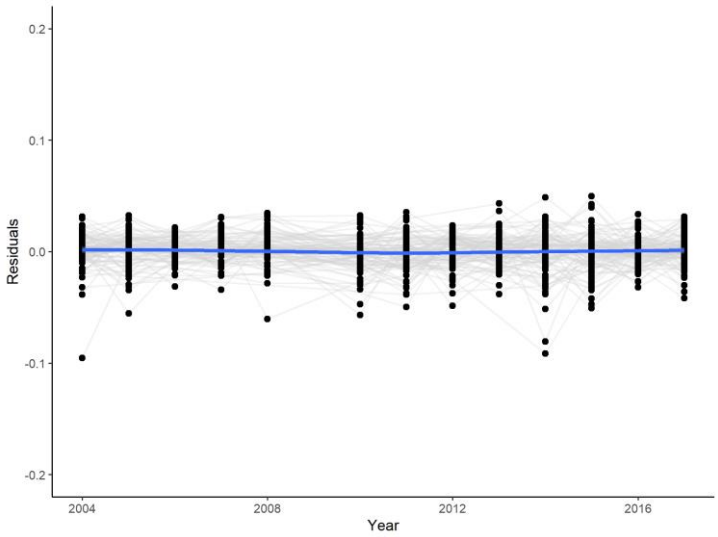

Figure S6: Distribution of the residuals per year of inference. The blue line indicates the mean value every year.

The multiple diagnostics tests and plots we used mostly supported the specification of the beta regression model. We inferred the missing values of coverage by using the mean estimates of the model for the years and regions where data was missing. Since the diagnostics also indicated minor discrepancies between our model and the results, we also generated 100 sets of coverage data by drawing the missing data from the normal distribution of the model (using the mean and standard deviation of the inferred values for the missing entries). We then ran the hhh4 models on each of the full coverage datasets to highlight the influence of the missing entries on parameter estimates. The deviation from the parameter estimates generated with the mean coverage was minimal (Figure S7 and Figure S8). This indicates that our conclusions are robust to changes in the inferred values of the missing data.

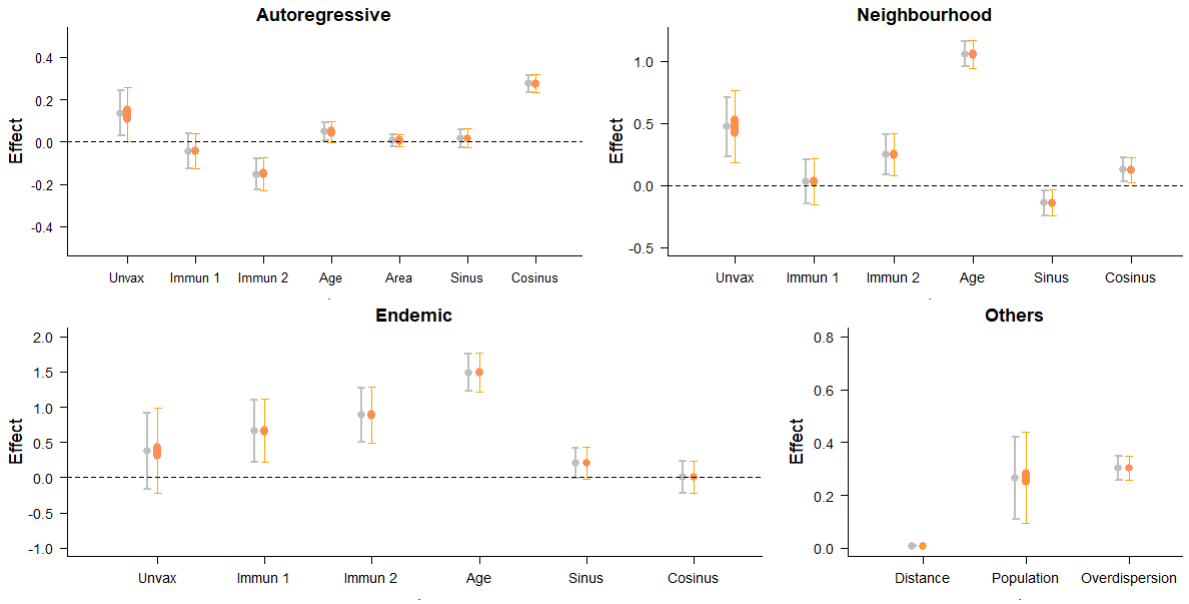

Figure S7: Estimates of the parameters in each component of Model 1, using the reference fit (grey) and 100 different values of coverage for the inferred entries (orange). unvax corresponds to the effect of $u_{i,t}$, the mean proportion unvaccinated over the three years before t in i; incid 1 and incid2 correspond to the effect of $N_{i,t}^1$ and $N_{i,t}^2$ the category of incidence in the three years before t in i; pop corresponds to the effect of $m_{i,t}$ the number of inhabitants at t in i; area corresponds to the effect of the surface; sin and cos correspond to the effects of seasonality; distance and population correspond to the spatial parameters of the connectivity matrix w (δ and γ); overdisp is the estimate of the log-overdispersion parameter in the negative binomial distribution of $Y_{i,t}$. Dots show the mean values associated with the parameters; arrows show the 95% Confidence interval. The orange arrows indicate the extreme values of the 95% confidence interval obtained drawing different values of coverage for the missing entries.

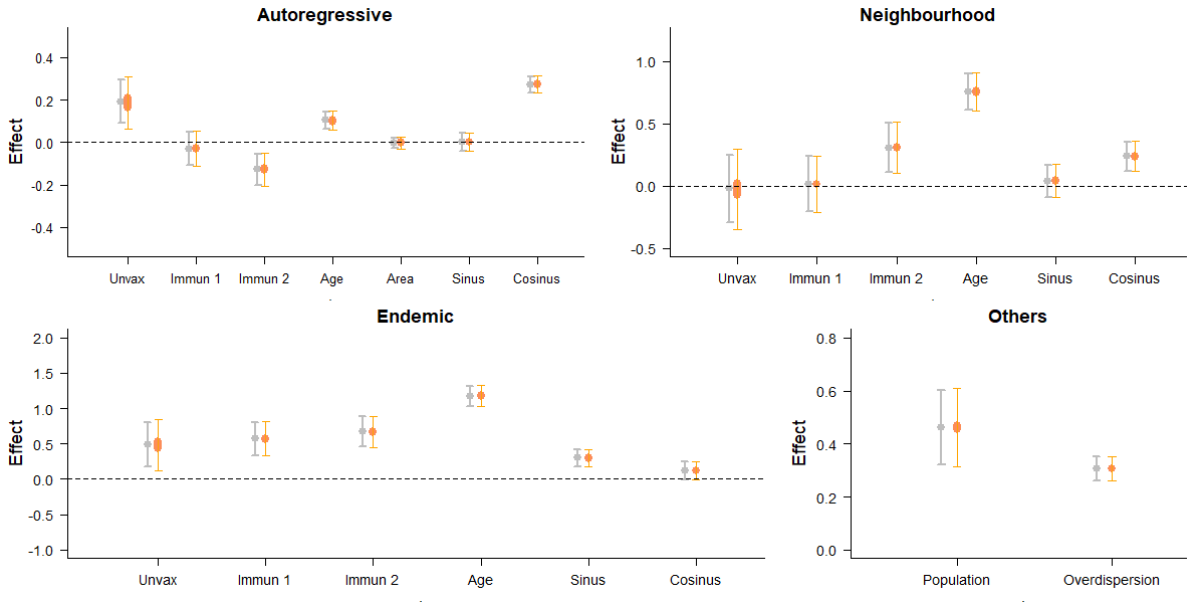

Figure S8: Estimates of the parameters in each component of Model 2, using the reference fit (grey) and 100 different values of coverage for the inferred entries (orange). unvax corresponds to the effect of $u_{i,t}$, the mean proportion unvaccinated over the three years before t in i; incid1 and incid2 correspond to the effect of $N_{i,t}^1$ and $N_{i,t}^2$ the category of incidence in the three years before t in i; pop corresponds to the effect of $m_{i,t}$ the number of inhabitants at t in i; area corresponds to the effect of the surface; sin and cos correspond to the effects of seasonality; distance and population correspond to the spatial parameters of the connectivity matrix w (δ and γ in Equation X); overdisp is the estimate of the log-overdispersion parameter in the negative

binomial distribution of $Y_{i,t}$. Dots show the mean values associated with the parameters; arrows show the 95% Confidence interval. The orange arrows indicate the extreme values of the 95% confidence interval obtained drawing different values of coverage for the missing entries.

S3. Seasonality

Both Model 1 and Model 2 include two parameters per component describing the seasonality of transmission and importations. For each component and each model, we computed $\exp\left(\beta_c^{(\lambda)} cos\left(\frac{2\pi t}{365}\right) + \beta_s^{(\lambda)} sin\left(\frac{2\pi t}{365}\right)\right) - 1$, the multiplicative factor corresponding to the impact of seasonality on the predictor (Figure S9).

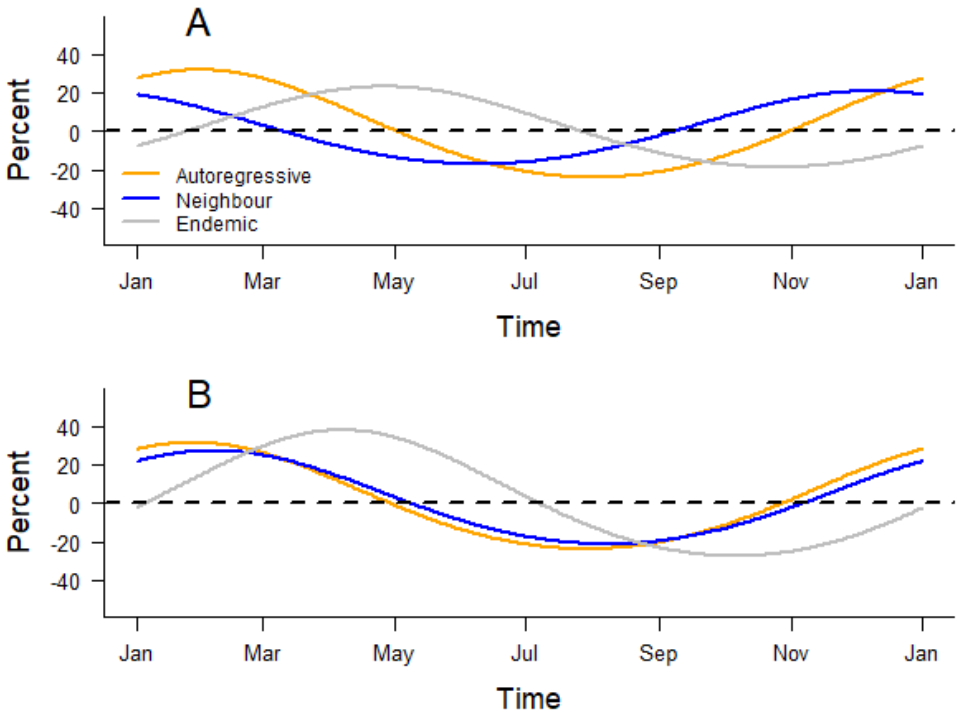

Figure S9:Seasonality of each component in Model 1 (Panel A) and Model 2 (Panel B). We quantified the impact of seasonality using the percent of variation around the mean value every day.

S4. Analysis using the neighbour-based connectivity matrix

The calibration and simulation study presented in the Main text was also run using Model 2. The fits to daily and weekly data were similar to Model 1 (Figure 4.4 and Figure S10). The calibration study indicates that Model 2 was slightly more likely to underestimate the number of cases in short-term predictions than Model 1. We generated the national number of cases predicted 3, 7, 10, and 14 days ahead by Model 1 and Model 2 over the calibration period, and compared the forecasts to the data (Figure S11). The predictions in both models were very similar, with the 95% prediction intervals overlapping on the majority of the calibration period. The data points were included in the 95% prediction intervals for forecasts one week ahead or less, when the period of forecasts was 10 or 14 days, we observed a lag between the predictions and the data when the number of cases started increasing and dropping.

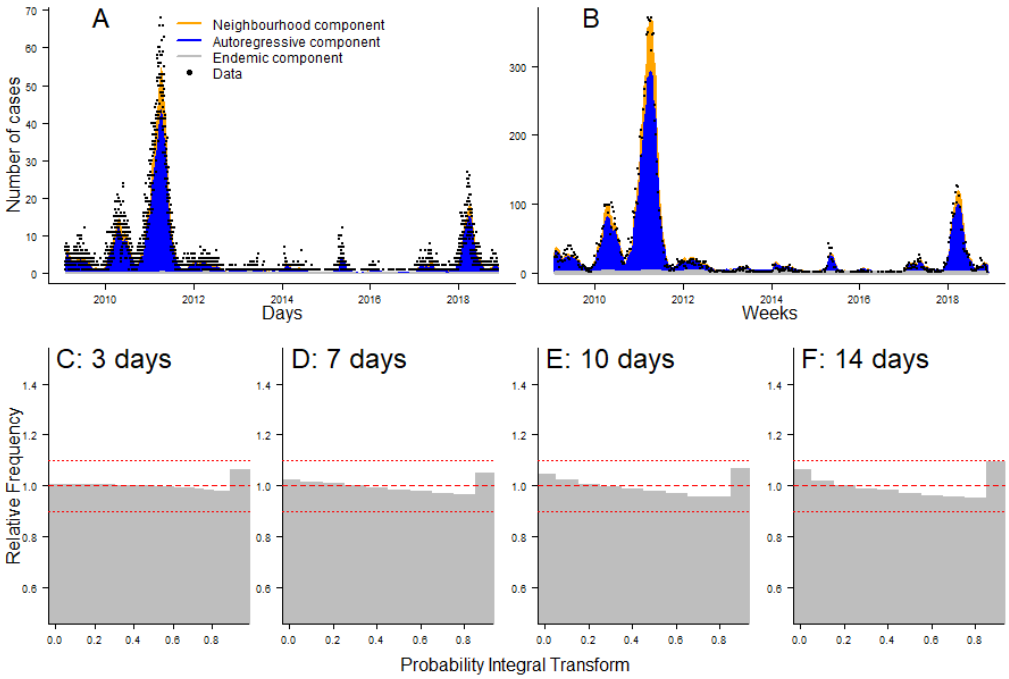

Figure S10: Panel A and B: Daily and weekly fit between the data and Model 2. The inferred number of cases is split among the three components of the model. Panel C to F: PIT histograms of Model 2, generated respectively for predictions 3, 7, 10, and 14 days ahead.

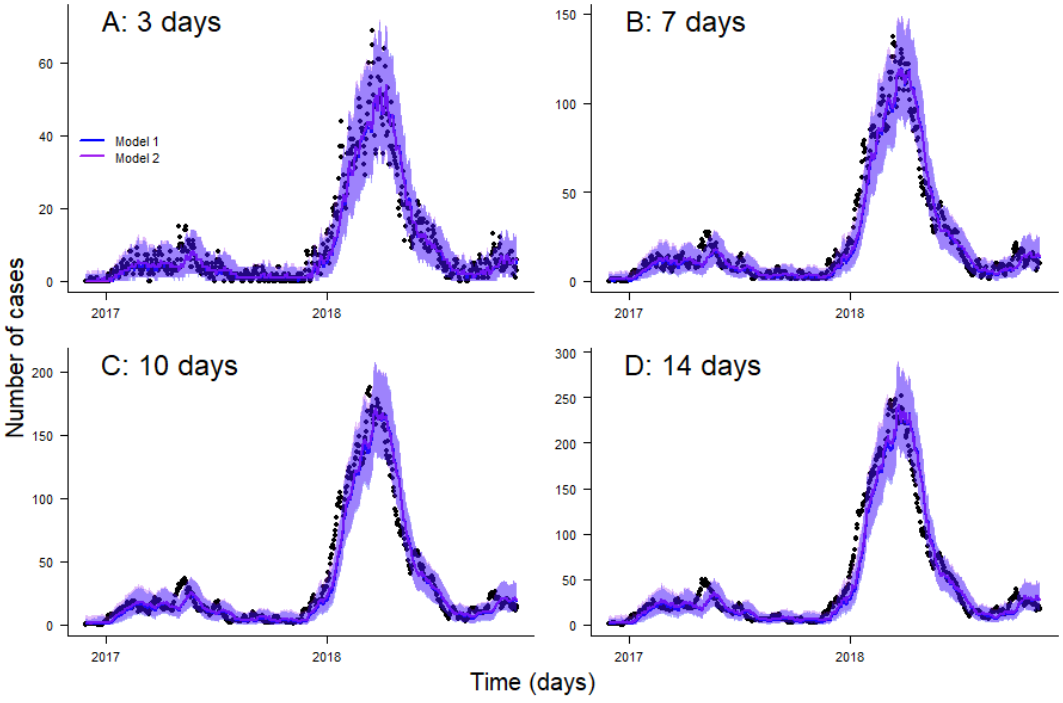

Figure S11: Comparison between predictions 3, 7, 10, and 14 days ahead using model 1 and model 2 and the data. Lines correspond to the median estimates, shaded areas correspond to the 95% predictions intervals. The blue and purple predictions are similar for the entire calibration period, hence the curves overlap in all panels. Black dots represent the number of cases 3, 7, 10, and 14 days ahead in France at each date.

The proportion of cases that stem from the endemic component was slightly higher in Model 2, which is due to the framing of the neighbourhood component (Figure S12). Indeed, in Model 2, the neighbourhood component only contains transmission between neighbours, whereas in Model 1, it describes any cross-regional transmission between the regions included in the study. Therefore, long-distance transmissions fall into the endemic component in Model 2, whereas they would be included in the neighbourhood component of Model 1.

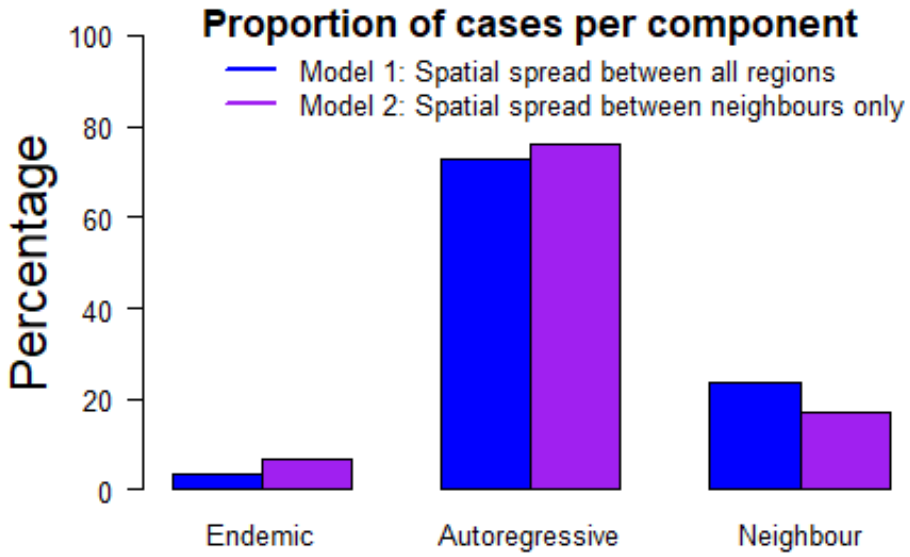

Figure S12: Proportion of cases per component in both models

We generated four simulation sets using the scenarios presented in the Main text with the parameter estimates from Model 2. We observed similar spatial heterogeneity in exposure and risk of large outbreaks. The areas most likely to be affected by large outbreaks were Paris and its suburbs, along with the south of France (Figure S13). Setting the level of recent incidence to the minimum values in each region decrease the number of baseline importations, and therefore reduced the proportion of simulation where most regions were exposed to transmission (i.e. they reported at least one case). Nevertheless, the risks of large outbreaks were similar to the reference simulation set. The effect of variations in vaccine coverage on the risks of importations and local transmission was the same as what was described for Model 1, with a three percent decrease leading to an abrupt increase in the number of cases.

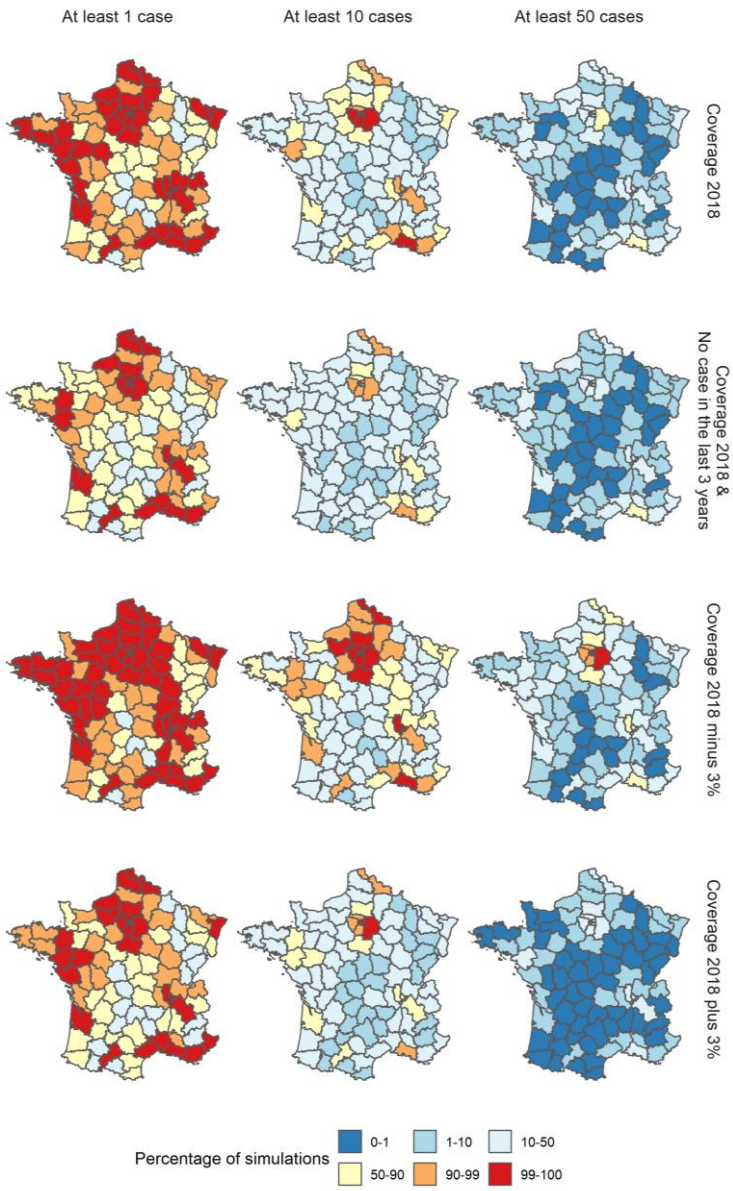

Figure S13: Percentage of simulations where the number of cases reported in each region in 2019 was at least 1, 10, and 50 cases for each scenario using parameter estimates from Model 2. Each row corresponds to a different scenario: i) Reference, ii) Minimum level of recent incidence in each region, iii) Local vaccine coverage increased by 3% in each region, iv) Local vaccine coverage decreased by 3% in each region.

The spatial spread upon repeated importation was more limited than in Model 1 (Figure S14). In all four scenarios, most regions were not exposed to transmission in any of the simulations. This is due to the fact that long-distance transmission would fall into the endemic component of Model 2, which was not used for these simulation sets. On the other hand, large transmission clusters in the region of importations and its neighbours were more common.

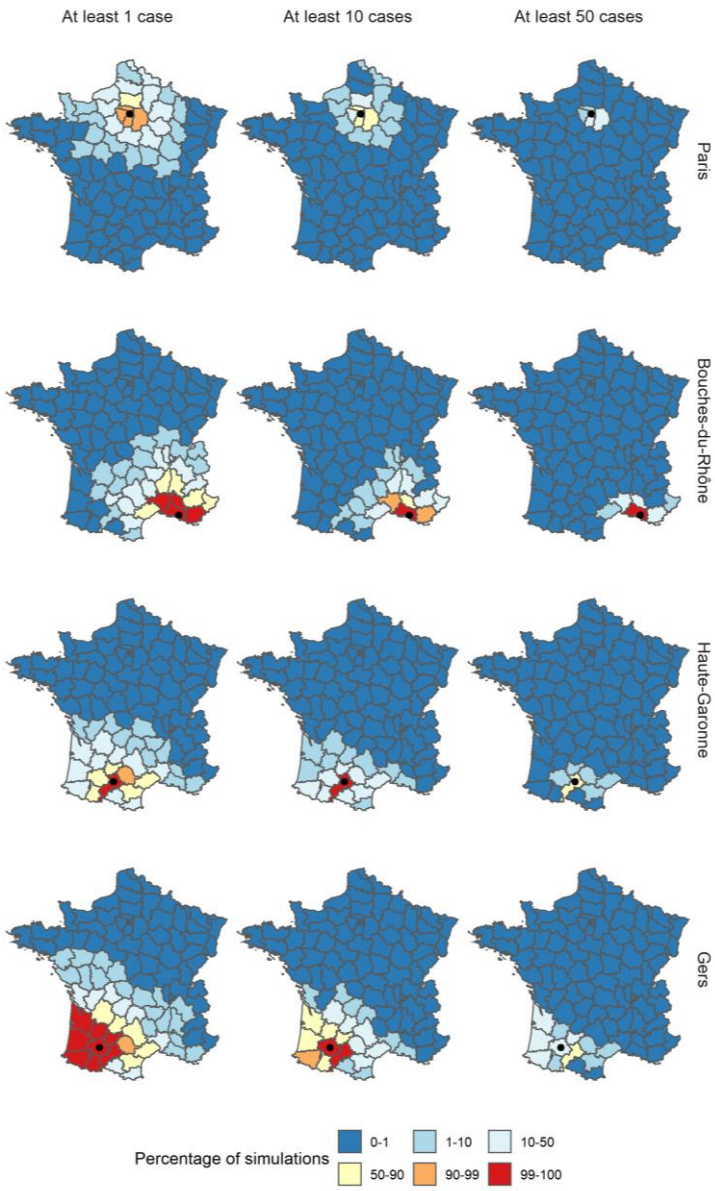

Figure S14: Percentage of simulations where the number of cases reported in each region in 2019 was at least 1, 10, and 50 cases following the importations of ten cases in December 2018, and using the parameter estimates from Model 2. For each row, the region of importation is indicated by a black dot.

S5. Last values of the covariates

The simulation sets were generated using the last measures of the variables included in our models. We show the geographic distributions of each variable (Figure S15):

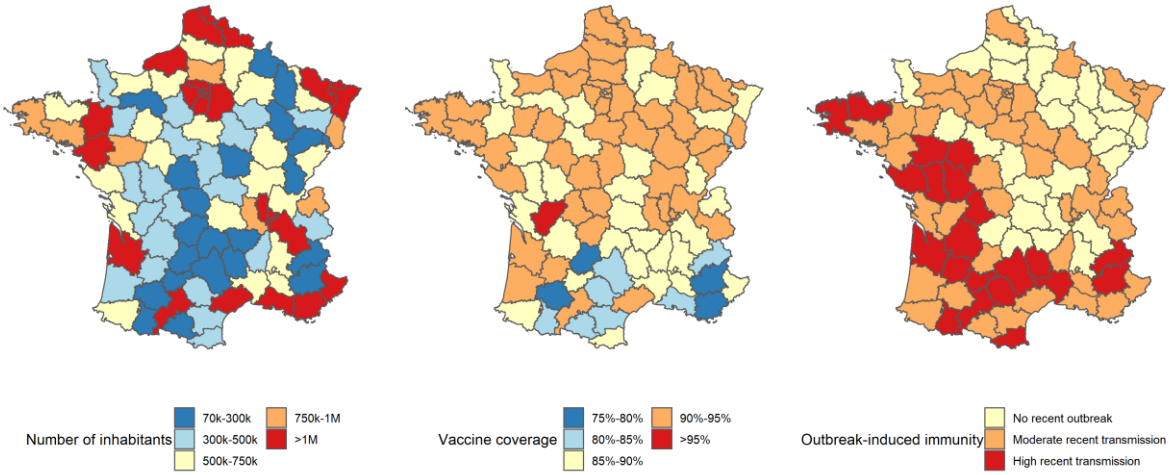

Figure S15: Geographic distribution of the number of inhabitants (right panel), average vaccine coverage (central panel), and recent incidence (left panel) at the end of 2018.

S6. Local importations with different vaccine coverage

Decreasing the vaccine uptake in each region by three percent led to an increase in the number of regions exposed to transmission following a local group importation (Figure S16). The risks of generating large outbreaks were higher both in the region of imports and in other areas, especially in highly populated urban regions. This was due to an increase in the number of cases generated in the neighbouring component, which led to more cross-transmission into regions with higher number of inhabitants.

On the other hand, a reduction of vaccine uptake led to a decrease in the overall number of cases generated per outbreak (Figure S17). Spill overs from the region of origin were much rarer: in the case of group importations in Haute-Garonne or Bouches-du-Rhône, no department outside the department of importation was exposed to one case in more than half of the simulations. Risks of large outbreaks were also reduced: in all four simulations, no department reported more than 50 cases in at least half the simulations, and few apart from the department of importation reported more than 10 cases.

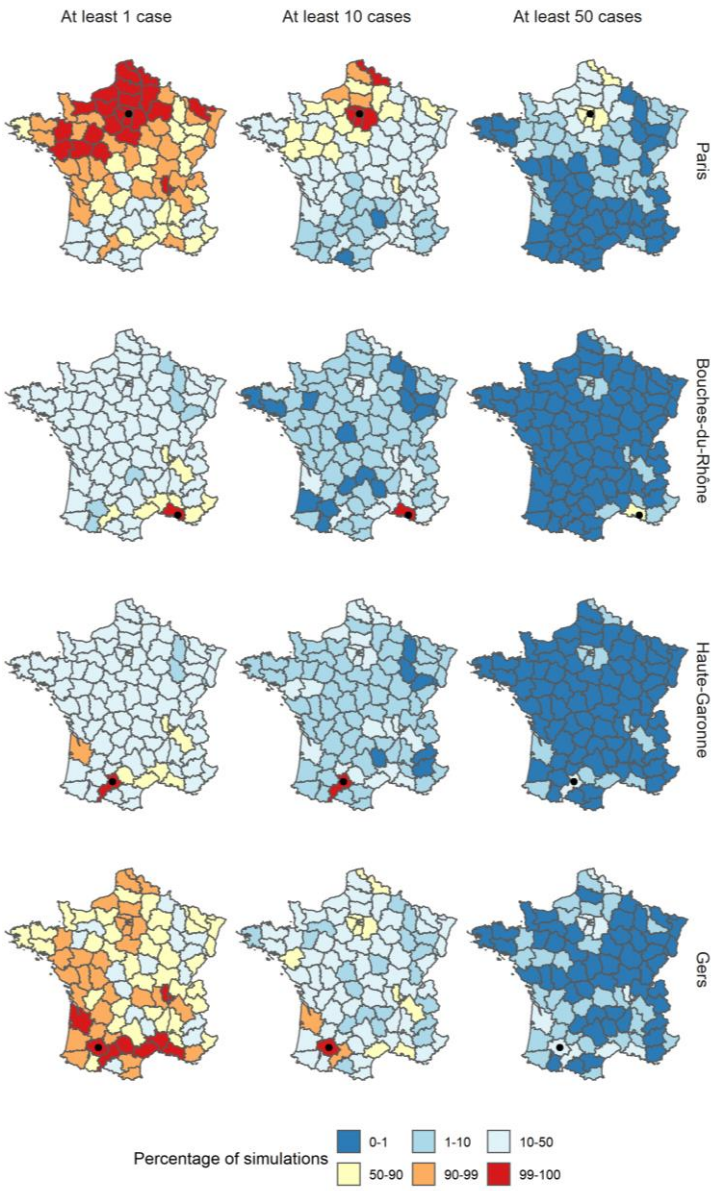

Figure S16: Percentage of simulations where the number of cases reported in each region in 2019 was at least 1, 10, and 50 cases following the importations of ten cases in December 2018, and using the parameter estimates from Model 1 and a three percent decrease in vaccine coverage in each region. For each row, the region of importation is indicated by a black dot.

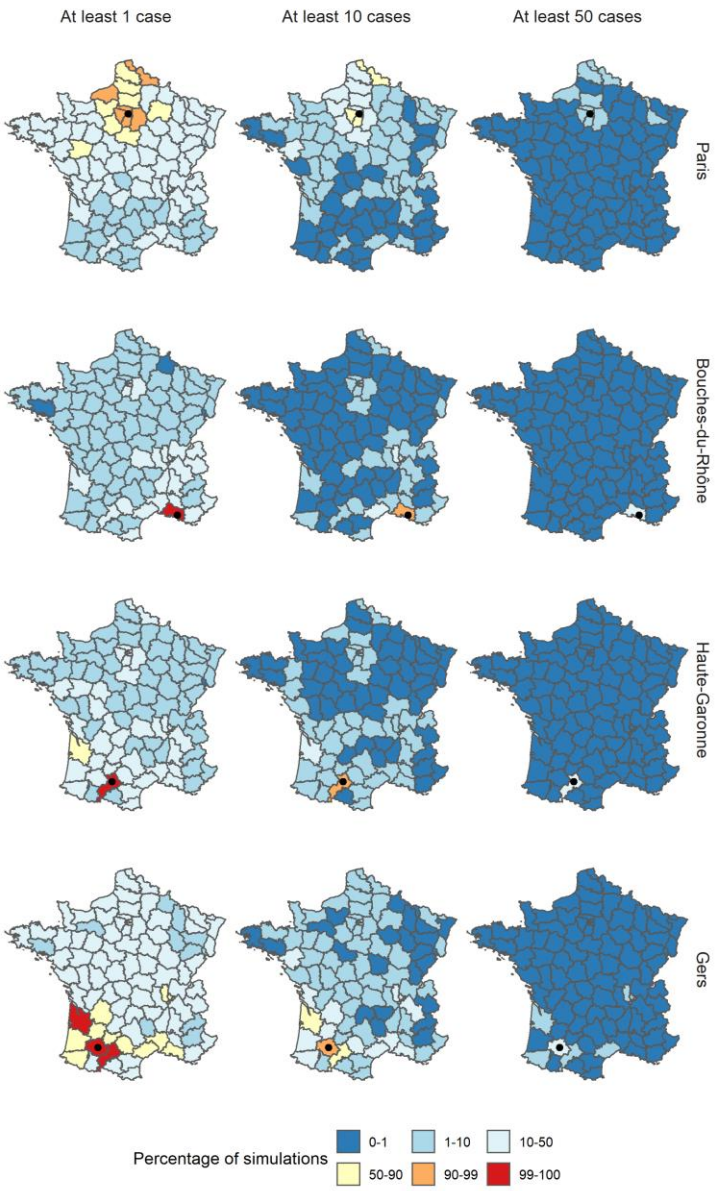

Figure S17: Percentage of simulations where the number of cases reported in each region in 2019 was at least 1, 10, and 50 cases following the importations of ten cases in December 2018, and using the parameter estimates from Model 1 and a three percent increase in vaccine coverage in each region. For each row, the region of importation is indicated by a black dot.

S7. Comparison with aggregated models and impact random effects

We assessed the impact of using daily case count and random effects on the calibration of the models by the Ranked Probability score. We were interested in three models: The daily model without random effects (presented in the Main Text), the daily models with random effects, and the aggregated model, fitted using a 10-day aggregation. Although random effects allow for more flexibility in the model, they significantly slow down the fitting procedure (40 times slower [2]). All models were run with the same specifications as Model 1, where cross-regional transmission can happen between every region, and with using vaccine coverage, recent levels of transmission, number of inhabitants, surface, and seasonality as covariates.

For each of the three models, we generated 10-day predictions every 10 days for the last two years of data. This corresponded to 72 calibration dates for 94 regions, hence 6,768 data points. We computed three different indicators using the R package scoringutils [3,4]:

1. The sharpness shows the ability of the model to generate predictions in a narrow range of possible outcomes, which means that the sharpness score is independent of the data. We used the normalised median absolute deviation about the median.
2. The bias: indicates whether a model systematically under or over predicts. Least biased models will get a mean value around 0, whereas completely biased models will get a value of -1 or 1.
3. The average Ranked Probability Score (RPS) for Count Data, proper scoring rule minimised if the predictive distribution is the same as the one generating the data.

The daily model without random effect had the lowest value of sharpness, indicating that the forecasts were generated in a narrower range of outcomes than the other models (Figure S18). The mean value of bias was closest to 0 in the aggregated model, which show that these forecasts were well balanced and did not tend to under or over-estimate the number of cases. We performed a permutation test on the RPS score, which indicated that the calibration of the two daily models was significantly better than the aggregated model (p-value < 0.02) [2].

Adding the random effects to the daily model did not lead to major improvements: The sharpness and bias were better in the model without random effects, whereas the difference between the RPS scores was not significant (p-values > 0.1). Since using random effects slows down the fitting procedure, without contributing a major improvement to the calibration, random effects were not integrated into the main analysis.

The fact that random effects did not substantially improve the calibration may be due to the time span of this analysis: random effects can be used to account for heterogeneity between the regions that would not be explained by the covariates included in the analysis. Since a random effect is applied to a given region throughout the entire time span of the data, they quantify constant effects on the region that was otherwise unobserved. As this analysis covers almost ten years of data, the impact of factors not included in this analysis may have varied over time, which would explain the little added value of the random effects.

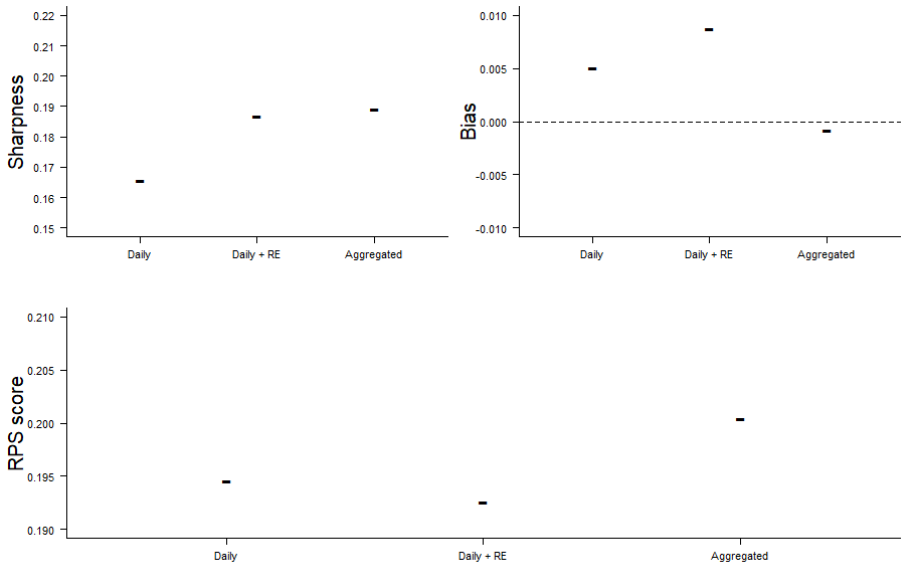

Figure S18: Sharpness, bias and RPS scores of the Daily model with and without Random Effects (RE), and of the aggregated model

S8. Control for day-of-the-week effect

Since we use daily onset dates, we explored the impact of potential reporting bias based on the day of the week. Indeed, delays in reporting can cause bias in the date of declaration of cases, and can explain why more cases would be reported on weekdays than on weekends. The number of cases with an onset date on Saturdays or Sundays was slightly lower than for the other days (1,970 cases on Saturdays and Sundays, whereas the average value on weekday is 2,100 cases, Figure S19).

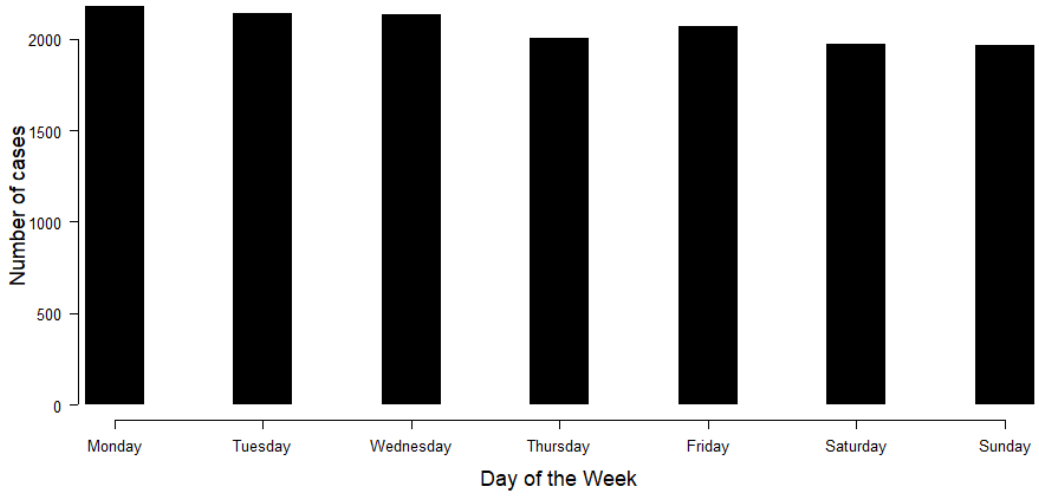

Figure S19: Number of cases by day of the week. We used the onset date for each case reported to the ECDC between 2009 and 2018.

We implemented a model controlling for the weekend effect in each component. This model contained the same covariate and distance matrix as Model 1. The covariate "Weekday" was a binary variable,

equal to 1 on Saturdays and Sundays, and 0 otherwise. In the neighbourhood and endemic components, the covariate "weekday" had no significant impact on the value of the predictor (Figure S20). On the other hand, the number of cases stemming from the autoregressive component was smaller on weekends (coefficient estimate: -0.09 [-0.15- -0.03]). Introducing this covariate brought no change to the values of the other coefficients in the model.

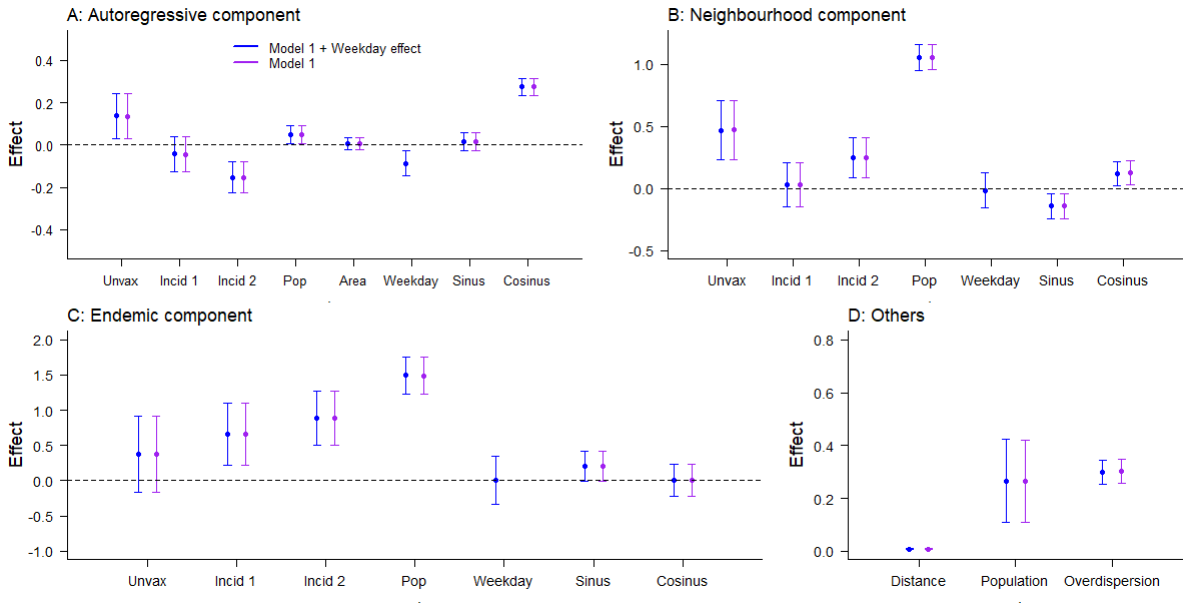

Figure S20: Comparison of the parameter estimates obtained in Model 1 (similar to Figure 4.2) with or without a weekday covariate added in each compartment. Weekday was computed as a binary covariate, whose value was 1 on Saturdays and Sundays, and 0 otherwise.

References

[1] Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, et al. glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J 2017;9:378–400. doi:10.32614/rj-2017-066.

[2] Meyer S, Held L, Höhle M. hhh4: Endemic-epidemic modeling of areal count time series. J Stat Softw 2016.

[3] Funk S, Camacho A, Kucharski AJ, Lowe R, Eggo RM, Edmunds WJ. Assessing the performance of real-time epidemic forecasts: A case study of Ebola in the Western Area Region of Sierra Leone, 2014–15. BioRxiv 2017:1–17. doi:10.1101/177451.

[4] Bosse NI, Abbott S, EpiForecasts, Funk S. scoringutils: Utilities for Scoring and Assessing Predictions 2020. doi:10.5281/zenodo.4618017.