

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Zarebski, Alexander Eugene

Title:

Quantitative Epidemiology: A Bayesian Perspective

Date:

2019

Persistent Link:

<https://hdl.handle.net/11343/227413>

Terms and Conditions:

Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.

Quantitative Epidemiology: A Bayesian Perspective

Alexander Eugene Zarebski
ORCID 0000-0003-1824-7653

Doctor of Philosophy

August 2019

Department of School of Mathematics and Statistics

The thesis is being submitted in total fulfilment of the degree.
The degree is not being completed under a jointly awarded degree.

Except where acknowledged in the allowed manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for another degree in any university.

Alexander Eugene Zarebski

Abstract

Influenza inflicts a substantial burden on society but accurate and timely forecasts of seasonal epidemics can help mitigate this burden by informing interventions to reduce transmission. Recently, both statistical (correlative) and mechanistic (causal) models have been used to forecast epidemics. However, since mechanistic models are based on the causal process underlying the epidemic they are poised to be more useful in the design of intervention strategies. This study investigate approaches to improve epidemic forecasting using mechanistic models. In particular, it reports on efforts to improve a forecasting system targeting seasonal influenza epidemics in major cities across Australia.

To improve the forecasting system we first needed a way to benchmark its performance. We investigate model selection in the context of forecasting, deriving a novel method which extends the notion of Bayes factors to a predictive setting. Applying this methodology we found that accounting for seasonal variation in absolute humidity improves forecasts of seasonal influenza in Melbourne, Australia. This result holds even when accounting for the uncertainty in predicting seasonal variation in absolute humidity.

Our initial attempts to forecast influenza transmission with mechanistic models were hampered by high levels of uncertainty in forecasts produced early in the season. While substantial uncertainty seems inextricable from long-term prediction, it seemed plausible that historical data could assist in reducing this uncertainty. We define a class of prior distributions which simplify the process of incorporating existing knowledge into an analysis, and in doing so offer a refined interpretation of the prior distribution. As an example we used historical time series of influenza epidemics to reduce initial uncertainty in forecasts for Sydney, Australia. We explore potential pitfalls that may be encountered when using this class of prior distribution.

Deviating from the theme of forecasting, we consider the use of branching processes to model early transmission in an epidemic. An inhomogeneous branching process is derived which allows the study of transmission dynamics early in an epidemic. A generation dependent offspring distribution allows for the branching process to have sub-exponential growth on average. The multi-scale nature of a branching process allows us to utilise both time series of incidence and infection networks. This methodology is applied to data collected during the 2014–2016 Ebola epidemic in West-Africa leading to the inference that transmission grew sub-exponentially in Guinea, Liberia and Sierra Leone.

Throughout this thesis, we demonstrate the utility of mechanistic models in epidemiology and how a Bayesian approach to statistical inference is complementary to this.

Declaration

This is to certify that:

1. the thesis comprises only of my original work towards the PhD except where indicated in the Preface;
2. due acknowledgement has been made in the text to all other material used; and
3. the thesis is less than 100, 000 words in length, exclusive of tables, maps, bibliographies and appendices.

Preface

This thesis emerged from work I did while assisting in the development of an influenza forecasting system for Melbourne, Australia, a system which is now being used by several major cities around Australia. The system is intended to both generate forecasts of influenza epidemics and improve our understanding of the process underlying these epidemics. Consequently, it uses a mechanistic model for how influenza is transmitted through a population. Prof. James McCaw, and Dr Robert Moss of the University of Melbourne and Dr Peter Dawson from the Defence Science and Technology Group orchestrated this project, with Dr Moss bringing it to fruition. This thesis reports my efforts to solve methodological problems that arose in the development of this forecasting system.

Chapters 1 and 2 provide some context for the rest of this thesis. The first provides an introduction to influenza and epidemiology before covering the mathematical and statistical techniques used in the forecasting system referred to above. The second reviews literature describing recent applications of these techniques and further details for the interested reader.

Chapter 3 consists of a publication written by myself (primary author), Peter Dawson, James McCaw and Robert Moss. In this chapter, we propose an approach to model selection targeted towards selecting a model for predictive skill. In doing so we establish that influenza forecasts are improved by accounting for variation in absolute humidity.

Chapter 4 presents a derivation of a family of prior distributions which simplify the process of incorporating existing knowledge into epidemic forecasts. As an example, we develop a simple predictive model for summary statistics of influenza epidemics in Sydney, Australia, based on historical time series. Predictions from this simple model are then used to construct a prior distribution for retrospective forecasts using a mechanistic model. This example demonstrates our approach for incorporating existing knowledge (or opinions) into forecasts, even if this knowledge does not pertain to the parameterisation used by the forecasting model. Consequently, it may also be of interest to the broader statistical community working with Bayesian methods.

Chapter 5 presents a study in which we used an inhomogeneous branching processes to model the early stages of transmission of Ebola in the 2014–2016 epidemic in West Africa. We demonstrate how to estimate epidemic dynamics from data in the form of either time series or infection networks. From this, we infer that the initial growth of Ebola virus transmission in West Africa in 2014–2016 was sub-exponential. This work was part of a broader attempt to understand how additional data sources could inform epidemic characterisation.

Finally, we summarise and discuss the work presented in this thesis and offer some opinions as to what may be fruitful lines of further enquiry.

Acknowledgements

Writing this thesis was difficult. The last four years have been difficult. I would not have been able to get through either without help from some amazing people.

Some of them have taught me how to live in the world of academia and helped me survive the process of learning this. My supervisors and advisory panel, James, Rob, Peter and Jodie, have and continue to teach and inspire me. The support staff at the University of Melbourne who have kept an eye out for me: Alex and the amazing Kirsten. My amazing study buddies: Ada, Claire, Gerry, and Jackson who are probably sick to death of my rambling and should be spared the contents of this thesis.

Some have helped me live in the world outside of academia. My friends and family who keep me going: Rachel, Kath, Jesse and Emily, and Sally. The housemates who put up with me: Pat, Hugh, Kate, Phil and Manda. The Cadmus Team, who taught me so much: Robbie, Herk, and the Ricks.

These people made my world a better place. I hope they understand how much they mean to me and how wonderful they are.

Contents

Abstract	v
Declaration	vii
Preface	ix
Acknowledgements	xi
Contents	xiii
List of Figures	xv
List of Tables	xix
1 Introduction	1
1.1 Influenza	1
1.2 The need for epidemic forecasts	2
1.3 Mathematical models of epidemics	3
1.4 Transmission models	4
1.5 Observation models	12
1.6 Bayesian statistics	14
1.7 Subjective vs Objective: it's all Bayesian to me	16
1.8 How do these methods allow me to do better quantitative epidemiology?	19
2 Literature Review	21
2.1 Introduction	21
2.2 Recent applied work	21
2.3 Theoretical work	23
2.4 Discussion	26
3 Model selection for seasonal influenza forecasting	27
3.1 Introduction	27
3.2 Publication	27
3.3 Contribution to the goals of this thesis	43
4 Prior distributions	45
4.1 Abstract	45
4.2 Introduction	45
4.3 Somewhat informative prior	47
4.4 Retrospective forecasting example	49

4.5	Results	52
4.6	Conclusion	58
4.7	Discussion	60
4.8	Acknowledgements	61
4.9	Supplementary Materials	61
5	Branching processes	65
5.1	Abstract	65
5.2	Introduction	65
5.3	Model	66
5.4	Method	69
5.5	Results	73
5.6	Conclusion	78
5.7	Discussion	79
5.8	Acknowledgements	80
5.9	Supplementary materials	80
6	Summary	83
7	Discussion	85
	References	89

List of Figures

1.1	In the SIR model susceptible individuals are infected, become infective themselves, and eventually recover. This can be seen in the monotonic decrease in the proportion of the population that is susceptible: the red line labelled S , the peak in the infectious proportion: the green line labelled I , and the monotonic increase in the proportion recovered: the blue line labelled R . . .	5
1.2	The SIR model partitions a population by disease status: susceptible to (s), infectious with (i), or immune (r), to the pathogen. The arrows between the compartments represent how members of the population transition between these states along with the rates of these transitions.	6
1.3	The SEIR model partitions a population by disease status: susceptible to (s), exposed to (e), infectious with (i), or immune (r), to the pathogen. The arrows between the compartments represent how members of the population transition between these states along with the rates of these transitions. . . .	7
1.4	The SIRS model with forcing which now has a time-dependent β and allows for the loss of immunity via the arrow from r back to s	8
1.5	The stochastic SIR model allows the proportions to fluctuate around their deterministic analogues. This allows for the infectious proportion to reach zero in a finite amount of time, i.e., it allows this model to describe the extinction of the pathogen. There are 50 realisations of the stochastic process shown on top of the deterministic solutions from Figure 1.5; although the realisations differ, they all have a similar shape to the deterministic solution.	9
1.6	Initially, the branching process approximation provides a useful approximation to the stochastic SIR model since it is far easier to work with analytically. Once the susceptible population is significantly diminished the approximation breaks down leading to the divergent behaviour seen in this figure. The deterministic solution for Figure 1.1 is shown (as a solid black line) with 50 realisations (thin grey lines) of the branching process approximation and the expected value to the branching process, (the dashed line).	10
4.1	Trajectories from the SIR model, shown as a spaghetti plot. The parameters have a somewhat informative prior distribution and agree with the prior beliefs. The cross-hairs show the mean and two standard deviations of the distribution corresponding to those beliefs which are based upon a predictive model using data from the Sydney influenza epidemics 2010–2016.	52
4.2	Distribution of solutions to the SIR model under parameters sampled from the literature and reference priors. Both distributions attribute non-negligible prior probability to outcomes which are clearly unreasonable. From 2010 to 2016 the number of cases had never exceeded 3000 in a single week. The cross hairs have been included to ease comparison to Figure 4.1.	53

- 4.3 Samples of the peak statistics, (the timing and magnitude of the peak,) appear uncorrelated under the SIP (the Pearson correlation coefficient is -0.03). Histograms demonstrate that the marginal distributions of the statistics are uni modal and vaguely Gaussian. Correlation ellipses shown (in red) are a one-sigma region of a bivariate normal distribution fit to the samples with the mode indicated by the dot in the centre. 54
- 4.4 Histograms and scatter plots of samples from the somewhat informative prior (SIP). The numbers above the diagonal indicate the Pearson correlation under the prior distribution due to the constraint on the distribution of the peak time and magnitude. Correlation ellipses show (in red) are a one-sigma region of a bivariate normal distribution fit to the samples with the mode indicated by the dot in the centre. 55
- 4.5 Histograms and scatter plots of samples from the literature informed and reference priors. Under these prior distributions the parameters are independent. The additional features of these figures are the same as described in Figure 4.4. Note: the axes vary between the three scatter plot matrices. . . . 56
- 4.6 Forecasts of the 2017 influenza epidemic in Sydney, Australia, using the somewhat informative prior. The first 20 observations of case counts from the epidemic were used to inform this forecast. The cross-hairs indicate the prior distribution of the epidemic peak. The posterior distribution of the epidemic curve underestimates the four observations around the peak, but otherwise matches the observed and unobserved data well. 56
- 4.7 Forecasts of the 2017 influenza epidemic in Sydney, Australia, using a.) the literature informed prior and b.) the reference prior. The first 20 observations of case counts from the epidemic were used to inform this forecast. The cross-hairs indicate the prior distribution of the epidemic peak. In each case, the posterior distribution of the epidemic curve matches the observed and unobserved data well. 57
- 4.8 Time series of the number of confirmed cases on influenza from 2010–2017 in Sydney, Australia, coloured by year. The magnitude of the number of cases observed each year is clearly increasing across this period. The peak in 2017 is approximately three times greater than the largest peak previously observed. 62
- 4.9 The week of the year in which confirmed cases peaked is plotted for the 2010–2016 influenza epidemics in Sydney (see Figure 4.8 for the full time series). There is no significant trend in the peak week from 2010–2016. The solid lines show both a linear (blue) and constant (red) models fit to the data along with 95% confidence intervals. The similarity between the linear and constant models suggests we should model this as a constant. 63
- 4.10 The maximum number of confirmed cases in a single week is plotted for the 2010–2016 influenza epidemics in Sydney (see Figure 4.8 for the full time series). There is a clear trend in the peak magnitude from 2010–2016. We modelled this using quasi-Poisson regression. The maximum number of confirmed cases in a single week for the 2017 epidemic is shown as a diamond at the upper edge of the predictive confidence interval. 63

5.1	The infection tree from Faye et al [1]. The colour of the nodes indicates whether the data were included in the analysis and the labels indicate where the infection occurred.	70
5.2	The branching process fit to time series of confirmed cases of EVD from Guinea, Liberia and Sierra Leone. The expected generation sizes (the model fit) are shown as a solid line with the 50% and 95% credible interval on this estimate shown as a grey ribbon. The observed case counts are shown as red points.	73
5.3	Histograms of posterior samples under the hierarchical model for a.) Guinea, b.) Liberia and c.) Sierra Leone. The marginal prior distribution is included as a solid line to assess convergence.	74
5.4	Histograms representing the posterior distribution of the model parameters conditional upon the secondary infections data and the time series data from Conakry. The solid lines show the prior distribution for each of the parameters (obtained via numerical integration). The growth rate and dispersion parameters are shown on a log-scale.	75
5.5	Simulated time series from the Reed-Frost epidemic model and the mean of these time series. The blue portion of the time series was used in the simulation re-estimation study.	76
5.6	A scatter plot of the maximum posterior probability estimate of the growth rate obtained from the time series data and the secondary infections data. There is a single point for each simulation, and the solid line shows a linear fit with a 95% confidence interval, the dashed line shows the parity line. Each facet shows the estimate conditional upon a different number of observations in each generation: 2, 5, or 10.	76
5.7	A scatter plot of the maximum posterior probability estimate of the deceleration parameter obtained from the time series data and the secondary infections data. There is a single point for each simulation, and the solid line shows a linear fit with a 95% confidence interval, the dashed line shows the parity line. Each facet shows the estimate conditional upon a different number of observations in each generation: 2, 5, or 10.	77
5.8	A scatter plot of the maximum posterior probability estimate of the dispersion parameter obtained from the time series data and the secondary infections data. There is a single point for each simulation, and the solid line shows a linear fit with a 95% confidence interval, the dashed line shows the parity line. Each facet shows the estimate conditional upon a different number of observations in each generation: 2, 5, or 10.	77

List of Tables

4.1	The details of the three prior distributions used for the retrospective forecasting. The notation $SIP(a, b)$ is used to indicate that the support of the parameter is the interval (a, b) . Under the SIP the elements of the parameter vector have an <i>a priori</i> correlation. Unlike the SIP, the literature and reference priors have independent components (U and Exp denote the uniform and exponential distribution respectively). The “Literature” based prior is taken from our previous work, [2] and the “Reference” prior is a uniform distribution over a large range of plausible values.	51
4.2	The approximation to the SIP — obtained via a generalised method of moments — provides a good fit for the properties of the desired distribution on the peak summary statistics: the week with the most cases, and the maximum number of cases per week.	54
5.1	Notation used for the branching process.	67
5.2	Prior distributions used for the model parameters in the hierarchical model described in Section 5.4.2.	71
5.3	Prior distributions used for the model parameters in the comparison of the two data types from Conakry.	72

1

Introduction

Disease has always played a major role in human history [3]. This thesis emerged from the development of an epidemic forecasting system for Melbourne, Australia. In particular, epidemics of influenza at the population scale, i.e., transmission through a city during an epidemic. Methodological gaps were encountered and my attempts to bridge these gaps form the scientific contribution of this thesis. This chapter describes the background material needed to understand these bridges.

1.1 Influenza

Pandemics of influenza are devastating; the 1918 pandemic resulted in tens of millions of deaths [4]. Perhaps less recognised is the major impact that *seasonal* influenza epidemics have. Each year 5% – 10% of adults and 20% – 30% of children are infected, leading to hundreds of thousands of deaths [5]. In Australia alone, there are several thousand deaths due to influenza and its complications each year [6].

The influenza virus spreads between humans in several ways; direct contact, and large virus-laden droplets expelled during coughing and sneezing are thought to be the predominant modes of transmission [7]. There is experimental evidence which suggests that absolute humidity affects the survival of the virus in droplets and consequently, its ability to transmit between hosts [8–10]. However, the effects of humidity on transmission at the population scale are complex, appearing to differ between temperate and tropical climates [11, 12]. The challenge of studying the impact of climatic factors at the population level is compounded by the difficulty of separating the effect on virus and the effect on human behaviour. In other words, when transmission increases in a temperate climate during winter, it is difficult to distinguish if this is due to a physiological effect (for instance due to humidity), or a change in behaviour of the human hosts since both may lead to similar outcomes [13].

Two measures of humidity are frequently used: absolute and relative humidity. One must be mindful of the difference, as both are used in the literature. Absolute humidity relates to the actual water vapour content in the air; relative humidity relates to water vapour content relative to the saturation point at that temperature. In temperate climates, absolute

humidity is at its lowest, both indoors and outdoors, during winter. In these climates, there is also an (almost inevitable) influenza epidemic during winter. Despite the regularity of these epidemics, their timing and size can vary dramatically.

Influenza viruses can be classified into types: A, B, C and D¹. Types A and B are the most prevalent and are responsible for the most associated morbidity and mortality among humans. The type A virus is further classified into subtypes based on surface molecules. Haemagglutinin and neuraminidase assist the virus to enter and exit host cells, the precise structure of these molecules on the influenza virus leads to the descriptors, such as H1N1 and H3N2. H1N1 was responsible for the 2009 pandemic, and H3N2 is often associated with severe infection [4].

Further refinement of this classification, into strains, is based upon the genome of the virus. Typically there are many strains of influenza in circulation. There have been attempts to understand the interactions between strains at multiple scales: the global evolutionary scale [15], the population scale [16, 17], and the scale of the individual host [18]. Understanding the interactions between strains of influenza is difficult, both in terms of conceptual understanding in virology and immunology, and the way to describe these interactions formally as either equations or computer code. Subsequently, as is common, in this thesis we aggregate co-circulating strains for mathematical convenience.

1.2 The need for epidemic forecasts

Influenza vaccination underwent clinical trials in the United States in the 1930s, becoming available to the public there in the 1940s [19]. However, vaccination is not the only intervention available to mitigate the impact of influenza, antivirals are used to manage infection and a variety of social measures (e.g., school closures) are used to reduce transmission [20, 21]. However, these interventions are costly and require the use of finite resources, and, as with vaccination, may depend on the specific strain.

When faced with a virus with significant mortality, such as Avian influenza (H5N1) [22] or Ebola virus disease [23], preventing even a small fraction of infections can save many lives. Determining an appropriate intervention strategy to deal with an outbreak of infectious disease when there is substantial uncertainty about the situation being faced is challenging. There are several aspects contributing to this difficulty. *There is likely to be a substantial degree of uncertainty.* Even establishing the current situation, i.e. how many people have been infected and how many are susceptible to infection can be incredibly difficult, or infeasible on ethical grounds. In addition to this, there can be uncertainty in the dominant modes of transmission, and what the likely effects of interventions would be. *Stochastic effects are likely to play an important role.* As addressed below, stochastic effects appear to play an important role in the evolution of an epidemic, so even having a detailed understanding of the initial condition of the epidemic may be insufficient to predict its outcome² [25]. *The range of potential responses is huge.* Even for a single intervention measure, actually deciding on the details of how to carry this out is a complicated task with many sub-decisions, as demonstrated in *in silico* for anthrax [26, 27].

Timely and accurate characterisation of an influenza epidemic and the impacts of available

¹Influenza D was only recently discovered in cattle [14].

²Of course, even if there were no stochastic effects, even simple models of epidemics exhibit chaotic behaviour [24].

interventions would facilitate improved decision-making on the part of health care practitioners [28]. Such a characterisation would require two components: improved awareness of the current and future states of the epidemic, and the ability to utilise this awareness to make better decisions. Accurate and timely forecasts of how the epidemic is likely to unfold would contribute to improving awareness. However, it is important to keep in mind that this is ultimately being done with the intention of these forecasts feeding into the decision-making process, i.e., assisting in the development of support systems to optimise the use of finite resources. We will limit ourselves to the forecasting of seasonal influenza epidemics for the current work. In particular, forecasting using mechanistic models, i.e., models deduced from assumptions regarding the causal factors underlying the epidemic.

The use of the term “mechanistic” is to distinguish these models from “phenomenological” (a.k.a. statistical, or correlative) models. Both approaches are used to forecast epidemics, with each having its merits [29, 30]. While the mechanistic approach strives to derive a mathematical form to describe the system being considered, phenomenological models assume a convenient mathematical form which is assumed to contain an appropriate description among its possible parameterisations. The predictive power of phenomenological models, under the moniker of “machine learning” has had a significant impact on the world. However, understanding why these models make the predictions they do is an active area of research [31]. This can be problematic, it is difficult to trust instructions about *what* to do if you cannot understand *why* they are correct. Since the ultimate goal is to improve understanding, in this thesis we primarily consider mechanistic models, however, there is some overlap. In the remainder of this chapter, we will survey the development and use of mechanistic models of epidemics.

1.3 Mathematical models of epidemics

Before we consider the ways in which epidemics have been modelled, it may be useful to consider what it means to model something and what can be achieved by doing so? To model a system mathematically requires making explicit what we think are the significant processes in that system. The choice of what these processes are and how to express them can be thought of as the axioms of our theory of how the system behaves. This process can be enlightening since in choosing these axioms the modeller is forced to recognise the limits of their understanding of the system under study. Moreover, given the model has been derived correctly from these axioms, formal analysis of this model will provide a way to ensure consistent reasoning about this system. You configure the model to represent the system you are interested in, and then it becomes just a mechanical process to derive the consequences of this configuration under the assumptions of the model.

For example, if you have a model of an epidemic, and you configure the model to represent a population of interest, then by evaluating the model you can answer all sorts of questions about how the population will be affected by the epidemic. Moreover, if you have a hypothesis about *potential* mechanisms influencing that epidemic, representing them in a model is a natural way to see if this hypothesis contradicts previous observations about epidemics: in which case this gives evidence to reject that hypothesis.

Hopefully, this has provided some reason *why* you might want a mathematical model, but we have not answered what a mathematical model *is*. A mathematical model is a set of objects which represent certain components of a system, and rules for how these objects interact and change. In a *good* model, the possible states of these objects which can be

reached by application of the rules will all be reasonable and the application of these rules will be easy. The result is a model that is useful for the reasons outlined above. A model can be *bad* if it admits impossible configurations or the applications of the rules is particularly challenging. The use of “good” and “bad” is a matter of taste to a degree: “appropriate” and “inappropriate” may be more accurate, but this conveys a useful way to think about models.

The target of our modelling efforts in this thesis is the spread of a pathogen through a population, in particular, the spread of seasonal influenza virus through major cities in Australia, we will also consider the spread of Ebola virus in West Africa during the 2014–2016 epidemic. A good model of an epidemic will require representation of both the pathogen and the population in which it spreads, whether implicit or explicit, but there is another component which is important: the process which gives rise to our observations.

In the physical sciences, experiments are typically designed and the technology used to make measurements is well understood. In epidemiology, ethical and practical constraints mean that much of the data used must be obtained opportunistically. The data collected to monitor epidemics often comes from health care practitioners, although increasingly other data sources are being used, such as social media [32] and internet search queries [33]. Since the measurement process may not be understood precisely this can lead to unrecognised biases in the measurements. It is important to take steps to mitigate this bias by accounting for the manner in which the data was collected.

The part of the model representing the spread of the pathogen through the population we refer to as a *transmission model* and the part representing how the data is collected we refer to as the *observation model*. Transmission models seek to describe a physical process which generalises over contagion and population; when people speak of an epidemic model, they usually mean a model of transmission. Observations models are more specific to the type of, and way in which data is collected. In the following sections, we look in more detail at both transmission and observation models.

1.4 Transmission models

Attempts to model transmission of disease mathematically can be traced back to Daniel Bernoulli [34], however, what is widely considered the foundational work would not come until nearly 200 years later. During the late 1920s, Kermack and McKendrick laid what is essentially the foundations of all mathematical theories of transmission [35, 36]. A special case of their work is a system of ordinary differential equations (ODE) which has come to be known as the SIR model [37]. The SIR model and variants of it have been extensively studied by different fields: applied mathematicians have studied it as a dynamical system [24], probabilists as a stochastic process [38], and epidemiologists as a tool to inform control measures [39].

We will now review SIR-type models of two varieties: deterministic and stochastic. There are other very important varieties, such as spatially explicit models and structured population models but these are beyond the scope the current work; the interested reader will find an excellent survey of these varieties in [37].

1.4.1 Deterministic models of transmission

The SIR model partitions a population by disease status; the elements of the partition are referred to as *compartments*. The *susceptible* compartment consists of all the individuals

who can be infected with the disease. Susceptible individuals can transition to the *infectious* compartment after being in contact with an infectious individual, i.e., an individual already in the infectious compartment. This transition from being susceptible to being infectious is referred to as the “incidence” of infection. When infectious individuals recover they cease to be infectious and are no longer susceptible to re-infection (i.e., have developed immunity to re-infection); at this point, they transition to the *recovered* compartment.

Figure 1.1 shows an example solution of the SIR model, i.e., how the size of each of the compartments changes over time. The population starts with all of its members susceptible, except for one who is infectious. Initially, the number of infectious people increases as the infection spreads among those who are susceptible. Since infectious people recover and there are only a finite number of people susceptible to infection, eventually, the number of infectious individuals decreases.

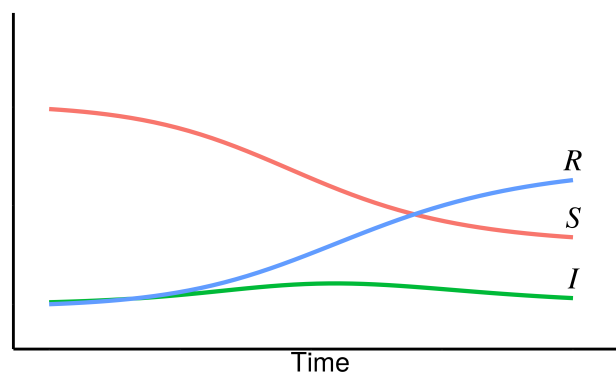


Figure 1.1: In the SIR model susceptible individuals are infected, become infective themselves, and eventually recover. This can be seen in the monotonic decrease in the proportion of the population that is susceptible: the red line labelled S , the peak in the infectious proportion: the green line labelled I , and the monotonic increase in the proportion recovered: the blue line labelled R .

It is often more convenient to study this system in terms of the proportion of the population assigned to each compartment rather than counting the number of individuals in each compartment. Typically — and as we have done here — one will denote the absolute number of individuals in a compartment by an upper case letter, and the proportion of the population in that compartment by the lower case letter. The state of the mechanism can be represented by three variables: s , i , and r , which denote the proportion of the population in each compartment at time t . At all times $s + i + r = 1$, since these compartments partition the population. The movement of the population through these compartments can be described using the schematic in Figure 1.2, where the arrows between the compartments represent the changes in disease state.

The rate individuals interact (in such a way that infection will occur) in this population is β , and the rate at which infectious individuals recover is γ . The following assumptions are usually made to simplify the model for how these quantities evolve:

1. There is no migration into or out of the population.
2. The characteristic duration of infection is much shorter than generations of the population, (i.e., fertility and mortality rates) and is constant.
3. The population is large enough that it is reasonable to represent proportions as real numbers.

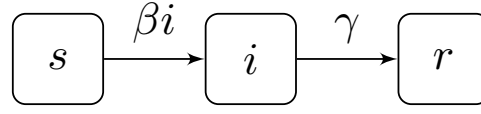


Figure 1.2: The SIR model partitions a population by disease status: susceptible to (s), infectious with (i), or immune (r), to the pathogen. The arrows between the compartments represent how members of the population transition between these states along with the rates of these transitions.

4. The members of the population are identical with respect to the infectious agent, and they mix homogeneously, i.e., each individual is equally likely to come into contact with each other individual.

This leads to the following non-linear system of ODEs:

$$\frac{ds}{dt} = -\beta si, \quad \text{and} \quad \frac{di}{dt} = \beta si - \gamma i. \quad (1.1)$$

Assumptions 1 and 2 are used to justify fixing the size of the population and using proportions in each compartment to represent the state. Assumptions 2 and 3 are used to justify the use of autonomous ODEs to model the system. Assumption 4 is used to justify the use of mass-action kinetics (the functional form of the terms (1.1)). It is a simulation³ of the solution to these equations that is shown in Figure 1.1.

A fundamental quantity of epidemiology is the *basic reproduction number*, denoted $\mathcal{R}_0$. This is defined as the average number of individuals an infectious individual will infect in a population which is otherwise comprised of susceptible individuals. For this model the basic reproduction number is β/γ . The derivation of $\mathcal{R}_0$ is most natural in stochastic models, however, it also has a sensible definition in terms of eigenvalues for many deterministic models [37]. Its importance comes from the role it plays in “threshold theorems”: an epidemic will *only* occur in this mathematical model if $\mathcal{R}_0 > 1$.

While the basic reproduction number is a static property of the model, there is a dynamic equivalent, the *effective* reproduction number, $\mathcal{R}_{\text{eff}}$, which accounts for the state of the system. The number of infectious individuals will only increase while $\mathcal{R}_{\text{eff}} = \mathcal{R}_0 s(t) > 1$. To see where this comes from, observe that the rate of change of i is proportional to $s\beta/\gamma - 1$. Therefore, the infectious proportion either increases or decreases depending on whether $\mathcal{R}_{\text{eff}}$ is greater than or less than one.

Two additional quantities of interest to epidemiology are the *incidence* and the *prevalence*. The incidence is the rate of new infections, in the equations above this is βs , the incidence term is sometimes used to refer to the whole expression βsi in Equation (1.1). The prevalence is the proportion of the population that is affected by the disease at a given time, in the model above this is i . It should be noted that in this model we have assumed that being affected by the disease, i.e., experiencing symptoms, is equivalent to being infectious. More elaborate models consider asymptomatic cases where there can be infectious individuals who do not experience symptoms [40].

³Due to the nonlinear term in Equation (1.1) the system cannot be solved analytically and hence numerical methods must be used.

In the context of seasonal influenza, there are some important variants of the SIR model. Once exposed to influenza virus, there is a delay before people become infectious, and they do not instantaneously experience symptoms. The delay between exposure (being infected) and actually becoming infectious is modelled by the addition of an *exposed* compartment, E as shown in Figure 1.3. Exposed individuals transition from the *susceptible* compartment to the *exposed* compartment, and then to the *infectious* compartment. The transition from exposed to infectious occurs at a rate σ . The system of ODEs to describe this is referred to as the SEIR model:

$$\frac{ds}{dt} = -\beta si, \quad \frac{de}{dt} = \beta si - \sigma e, \quad \text{and} \quad \frac{di}{dt} = \sigma e - \gamma i$$

now with the conservation law $s + e + i + r = 1$. The basic reproduction number is still given by $\mathcal{R}_0 = \beta/\gamma$, however, there is no longer a guarantee that $di/dt > 0$ since individuals must transition through the exposed state before they become infectious.

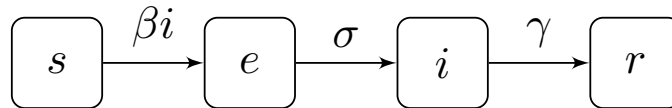


Figure 1.3: The SEIR model partitions a population by disease status: susceptible to (s), exposed to (e), infectious with (i), or immune (r), to the pathogen. The arrows between the compartments represent how members of the population transition between these states along with the rates of these transitions.

As mentioned above, influenza activity demonstrates significant seasonal variation in temperate climates: suggesting transmission may be influenced by climate⁴. To incorporate this seasonal variation, the parameter β is modulated. In the models above, this is typically realised by relaxing Assumption 2, making β a function of time:

$$\beta(t) = \beta_1 \left(1 + \beta_2 \sin \left(\frac{t}{2\pi T} \right) \right).$$

The period of this oscillation, T , is the period of the seasonal effect. The resulting system of ODEs is referred to as an SIR model with *forcing*. Since the forcing is time dependent the system is no longer autonomous; one needs to be aware of the fact that the solution is dependent on *when* it was solved (in model time, t).

Forced SIR models can be extended to allow periodic solutions by including transition from the R compartment back to the S compartment at a rate δ as shown in Figure 1.4. The periodic solutions to forced models can exhibit complex and chaotic behaviour [24]. The biological reasoning for this additional transition, i.e., the loss of immunity among the recovered individuals, is that the virus can evolve, leaving people vulnerable to reinfection. However, this can also be brought about by including demography into the model, i.e., the birth of susceptible individuals and death of individuals in each compartment.

The models presented above are based on strong assumptions and their applicability is dependent upon these assumptions capturing the dominant processes of the real world system. In particular, assumptions 3 and 4 are frequently questioned [41, 42]. Network models are

⁴While one must not confuse correlation with causation; in this case, it is probably justified.

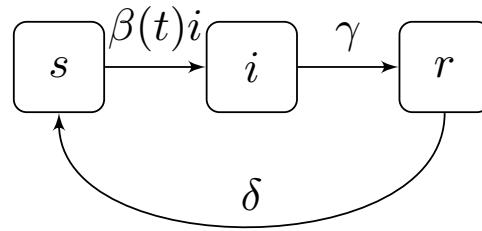


Figure 1.4: The SIRS model with forcing which now has a time-dependent β and allows for the loss of immunity via the arrow from r back to s .

seen as a promising way to relax the assumption of homogeneity (assumption 4) however there are still substantial theoretical and computational challenges involved in their use [43, 44]. A more established approach, primarily used to relax the assumption of a large population (assumption 3) is to use stochastic models. Stochastic models can represent an outcome which is difficult to obtain with ODE models: the possibility of entering a state where there is no disease in the population. Such states are usually precluded from ODE models due to the so-called “atto-fox” [45] which refers to non-physical states in which prevalence is positive but so small it corresponds to far less than a single individual in the population.

1.4.2 Stochastic models of transmission

Before we delve into a more formal treatment of stochastic epidemic models, let us first informally consider the SIR model (represented in Figure 1.1) when it is recast as a stochastic process. For simplicity, we will consider a fixed initial condition, again with a single infectious individual in an otherwise susceptible population. In the deterministic model, the state of the population is represented by a pair of real numbers which vary continuously as described by the ODEs in Equation (1.1). In the stochastic formulation of the SIR model we treat individuals as indivisible units, i.e., the state of the system is represented by the integer number of individuals in each of the compartments. These individuals now jump between the disease states; the jumps are random variables and their rate of occurrence is analogous to the rates in the deterministic model. Since the jumps are random variables, every realisation of this model is a random variable.

Figure 1.5 shows the deterministic model solution from Figure 1.1 but with 50 realisations of a stochastic model layered on top. The deterministic solution approximates the average behaviour of the stochastic model [46]. An important difference though is that the stochastic model can *stop*, in the sense that once there are no more infectious individuals, it is not possible for anything to change. This can be seen in some of the realisations that stop early.

Continuous-time Markov chains (CTMC) are a natural way to formulate a stochastic SIR model, and where we will start our introduction to stochastic epidemic models [37, 47]. Relaxing assumption 3 from the ODE formulation above, i.e., the assumption of a large population, and assuming transitions from susceptible to infectious, and then to recovered occur after exponentially distributed intervals of time leads one to the CTMC version of the SIR model. Since the population in this model is finite and discrete, it can reach a disease-free state in a finite amount of time. However, while the CTMC is more expressive (i.e. it admits a wider range of behaviour) this comes at the cost of additional theoretical

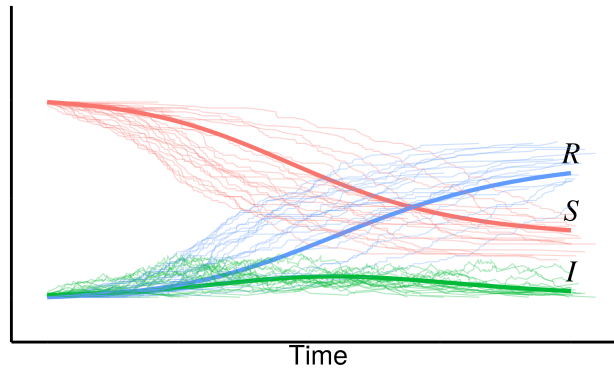


Figure 1.5: The stochastic SIR model allows the proportions to fluctuate around their deterministic analogues. This allows for the infectious proportion to reach zero in a finite amount of time, i.e., it allows this model to describe the extinction of the pathogen. There are 50 realisations of the stochastic process shown on top of the deterministic solutions from Figure 1.5; although the realisations differ, they all have a similar shape to the deterministic solution.

and computational challenges in its use. In particular simulation of the process becomes computationally expensive as the model is scaled up to larger populations. This computational expense can lead to difficulties when trying to estimate the parameters of such models [48]. Because of this, simpler stochastic models which are cheaper to simulate are often used as an approximation of the full model.

Drawing samples from a CTMC is conceptually simple: the stochastic simulation algorithm (SSA) (a.k.a. the Doob-Gillespie algorithm [49]) has been used to sample from CTMCs since the 1940s [50] and has been applied to the simulation of epidemics since as early as 1953 [51]. The difficulty arises when scaling up to large populations because the number of events to be simulated becomes large. Yet individually, these events often do not have a substantial impact on the state of the system: the exception being when there are only a small number of infectious individuals, and when *stochastic extinction* is a plausible event. A computational approach to handling this is the tau-leap approximation which can substantially reduce the simulation cost by aggregating events [52]. Provided the individual events do not have a substantial impact on the state of the system, the approximation this introduces is minor. However, as already mentioned, this is not the case when there are only a small number of infectious individuals. A promising solution to this is the use of hybrid methods which switch between solution strategies depending on the state of the system [53]. The idea being, that you invest computational resources in regions of state space where the details matter, and use an approximate solution the rest of the time. While these methods have been used in the context of epidemiology they have not yet seen widespread adoption [54, 55].

An alternative to approximating the *computation* is to approximate the *model*. One such approach to this approximation is to use *stochastic differential equations*. In particular, approximating the number of events with a Gaussian random variable — with a unit coefficient of variation — leads to the following chemical Langevin equation [56]:

$$d\mathbf{X}(t) = \mathbf{f}(\mathbf{X}(t))dt + \mathbf{G}(\mathbf{X}(t))d\mathbf{W}(t),$$

where $\mathbf{X}(t) = [S(t), I(t), R(t)]^T$. The drift vector is given by

$$\mathbf{f}(\mathbf{X}) = \begin{bmatrix} -\alpha SI \\ \alpha SI - \beta I \\ \beta I \end{bmatrix},$$

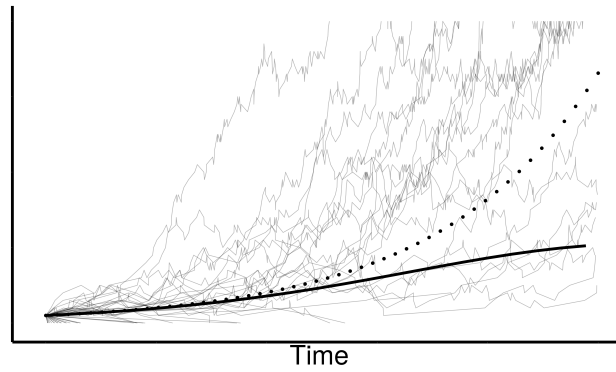


Figure 1.6: Initially, the branching process approximation provides a useful approximation to the stochastic SIR model since it is far easier to work with analytically. Once the susceptible population is significantly diminished the approximation breaks down leading to the divergent behaviour seen in this figure. The deterministic solution for Figure 1.1 is shown (as a solid black line) with 50 realisations (thin grey lines) of the branching process approximation and the expected value to the branching process, (the dashed line).

the diffusion matrix by

$$\mathbf{G}(\mathbf{X}) = \begin{bmatrix} -\sqrt{\alpha SI} & 0 \\ \sqrt{\alpha SI} & -\sqrt{\beta I} \\ 0 & \sqrt{\beta I} \end{bmatrix}$$

and $d\mathbf{W}(t)$ is a two-dimensional Wiener process. Sampling from this process can be achieved via the Euler-Maruyama method [57] (a stochastic analogue of the Euler method for numerically solving ODEs). To control the level of stochasticity in this model, one might introduce a multiplicative factor into the diffusion matrix [58].

A special case described in the work of Kermack and McKendrick — already present in McKendrick's earlier work — considers the initial stages of an epidemic in a large population. The simplifying assumption is that each infection does not substantially influence the number of susceptible individuals in the population. In a large population, it seems reasonable to assume that the initial infections do not significantly affect the number of susceptible individuals. Following this line of reasoning one quickly arrives at the idea that a branching process [59] is an appropriate way to model the early stages of an outbreak. Branching processes are particularly well suited to studying the basic reproduction number, and the probability of a population reaching a disease-free state due to stochastic effects, i.e., the pathogen going extinct.

Before we delve into the details of branching processes, first let us consider an example of how this can be used to approximate the initial behaviour of the stochastic model used in Figure 1.5. Figure 1.6 shows 50 realisations of a branching process for the number of infectious individuals in the population. It can be seen in these realisations that initially, the branching process seems to be a reasonable approximation to the stochastic model used in Figure 1.5. However, the assumption that the number of susceptible individuals is constant eventually becomes unreasonable and hence the approximation breaks down.

Chapter 5 of this thesis demonstrates the use of branching processes in epidemiological modelling and statistics. The theory of branching processes is extensive and it has been used to model many biological systems, however, only a few ideas are necessary to understand the material in this thesis. To demonstrate these ideas, consider the situation that motivated the Galton-Watson process: inheritance of family names among male members of a population

[60, 61]. In the following, we loosely follow the generating function treatment given by Feller in [59].

Consider a man with a unique family name which he will pass onto his sons⁵, we will use a branching process to investigate the number of men with this family name in subsequent generations of the population. The number of men in the n th generation with the family name is denoted Z_n . Initially, he is the only man with this family name, so the zeroth generation only has one man with this family name, $Z_0 = 1$. If he has k sons, there will be k men in the first generation with the family name and so on. The probability of a man passing on the family name to k sons is p_k and the number of sons each man has is independent of the number his brothers may have had. If X_g^i is the number of sons of the i th man in the g th generation, then

$$Z_g = \sum_{i=1}^{Z_{g-1}} X_{g-1}^i.$$

The process is said to have gone extinct if $Z_n = 0$. Galton and Watson were interested in the potential extinction of aristocratic family names, however, branching processes have many applications. In the current epidemiological context, the X_g^i represent the number of people infected by the i th individual in the g th generation of the outbreak. This leads to a particularly simple definition of the basic reproduction number: $\mathcal{R}_0 = \mathbb{E}X$.

For a random variable X where $\mathbb{P}(X = k) = p_k$ the probability generating function (PGF) of X is the function

$$P(s) = \sum_k p_k s^k.$$

PGFs, and generating functions more generally, allow probabilistic and combinatorial constructions to be studied algebraically: simplifying many calculations [62]. For example

$$\mathbb{P}(X = 0) = P(0) \quad \text{and} \quad \mathbb{E}X = \left. \frac{dP}{ds} \right|_{s=1}.$$

The PGF of Z_n , which we denote G_n , is given by

$$G_n(s) = P(G_{n-1}(s)) \quad \text{with the initial condition} \quad Z_0 = 1 \quad \text{giving} \quad G_0(s) = s.$$

Now in epidemiological terms, the probability of the pathogen being extinct by the n th generation is $g_n = G_n(0)$ and hence $g_n = P(g_{n-1})$. With generating functions it is elementary — though tedious [59] — to show the following results:

- If $\mathbb{E}X \leq 1$ then the pathogen almost surely goes extinct, if $\mathbb{E}X > 1$ it goes extinct with probability g where $g = \lim_{n \rightarrow \infty} g_n$.
- If the pathogen does *not* go extinct, Z_n grows without bound.
- $\mathbb{E}Z_n = (\mathbb{E}X)^n$, which also follows trivially from the law of total expectation.

⁵For potential readers in the distant future, at the time of writing in Western society, an individual's family name was typically patrilineal. Despite important (and ironic) exceptions, this rule is used in the example for simplicity's sake.

As the last result shows, the branching process described above has geometric/exponential growth in the expected number of infections in each generation (the incidence). Observational studies from real epidemics suggest that early on in an epidemic the changes in incidence are polynomial rather than exponential [63–65], i.e., even in the early stages the incidence does not grow exponentially in time, it is better described by a low order polynomial. That this has been observed during the *early* transmission is important, since it suggests that it is not merely a result of depleting the pool of susceptible individuals [63]. It appears the assumption of homogeneity in the population needs revision; this motivated the work described in Chapter 5.

1.4.3 Alternative approaches to modelling transmission

In this thesis we only consider deterministic and stochastic models of transmission. However, there are alternative mechanistic approaches to modelling transmission which may be of interest to the reader, in particular, agent based modelling (ABM) and quasi-mechanistic modelling. The deterministic and stochastic approaches described above are motivated by the notion that a model should be built up from assumptions regarding the most important aspects of the transmission process. ABM and quasi-mechanistic take harder and softer approaches (respectively) to mechanism.

The agent based approach to modelling involves representing a system as a collection of interacting autonomous agents. The focus here is on describing the actions and interactions of the agents correctly and then investigating what manifests at the population level. For example, the branching process described above may be viewed as a simple ABM. For the most part though the use of these models relies on being able to simulate them in a computer. The GLocal Epidemic and Mobility computational model (GLEAM) is a notable example of the application of ABMs to the modelling and forecasting of epidemics [66, 67]. While GLEAM is bespoke software for modelling epidemics, there is a range of general purpose simulators which have also been used for this task [68].

While the ABM approach seeks to describe the transmission process explicitly the so-called, quasi-mechanistic approach acknowledges that there are significant exogenous factors and introduce additional stochasticity to account for this. For example, a common approach used to account for unknown factors which may influence a parameter is to represent the value of that parameter with random walk (through time) or use a non-parametric estimator for it [69, 70]. While this introduces additional variance into estimates, it can be a useful way to assess potential changes in system dynamics [71].

There is a vast literature on epidemic models with many different modelling approaches in use. Given the dizzying number of possibilities, one must be selective in their choice of model, since often the best one will be context specific. Choose the simplest one capable of describing the important aspects of transmission and hence answering questions about these aspects. By “simplest” we mean simple both in terms of parsimony and ease of use by the modeller.

1.5 Observation models

For both ethical and practical reasons, it is not feasible to carry out the level of surveillance required to accurately estimate the state of a population with regards to the number of people infected with or susceptible to infection by a pathogen like influenza. Instead, there are

surveillance systems which generate data which is then used to estimate what that state is, although this inference can be challenging. Depending on how the data is collected, the same epidemic can produce very different data sets. Subsequently, when estimating the state of the transmission model which gave rise to a specific data set, it is important to appropriately model the specific way in which the data was collected. Since the surveillance system provides our observations of the epidemic, we refer to this part of the model as the *observation model*.

There are many types of surveillance systems. Loosely speaking, surveillance systems can be classified as either *active* or *passive*. In passive surveillance systems, data is recorded when individuals meeting certain criteria come to the attention of the system. In active surveillance systems, individuals are recruited from the population and tested to see if they meet those criteria. For example, if I recruit employees from the University of Melbourne and test whether they are infected with influenza, this would be active surveillance. If I went to the nearby Royal Melbourne Hospital and performed the same test on people who came to the hospital seeking care for influenza-like illness, it would be passive surveillance.

Which approach to surveillance will be most appropriate depends on what we hope to learn from the surveillance. Continuing with the previous example, if I were interested in the proportion of the population that have influenza, the prevalence, the active approach would be more appropriate. If I wanted to understand which strains of influenza were circulating, the passive approach would be more appropriate. In the following sections, we consider examples of models for both active and passive surveillance systems.

1.5.1 Active surveillance

As one might expect, active surveillance systems typically require more work and subsequently, often produce smaller data sets. During the SARS [72] and Ebola epidemics [73] there was active surveillance in the form of contact tracing. Contact tracing involves monitoring contacts of individuals known to be infectious in the hopes of quickly identifying individuals who have been infected. This has the obvious benefit of meaning that these individuals get treatment quickly and through careful management of their infection, can prevent them from infecting other individuals. There is the added benefit, that in recording these activities we obtain data about who-infected-whom. Or, in the case when people have *not* been infected, we get information about how likely infection is to occur.

Contact tracing was used to contain Ebola during the epidemic in West Africa from 2014–2016. Another important example of active surveillance is what is known as a *first few hundred* study (FF100) [74]. In FF100 studies households of confirmed infections are monitored with regular testing of all individuals in the residence. In Chapter 5 we will make use of data collected during active surveillance.

1.5.2 Passive surveillance

In the SIR model, prevalence is the proportion of the population that is infected at a given time, ($i(t)$ the proportion in the infectious compartment at time t .) However, observations of prevalence for acute infections (such as influenza) are rare; it is more common to observe incidence. Recall from above that incidence is the rate at which infections occur. In the SIR model, this refers to the flow from the s to the i compartment, i.e., the term βsi in Equation (1.1). If one were considering the incidence, $\mathcal{I}$, over a period of time, $[t_a, t_b)$, e.g., the weekly incidence, this would be the integral of this quantity over that period of time:

$$\mathcal{I} = \int_{t_a}^{t_b} \beta s(u) i(u) du. \quad (1.2)$$

For influenza, we assume people begin to experience symptoms at the same time they become infectious. And, that if they are going to seek medical services, they will do so then. In the SEIR model, people become infectious when they transition from the exposed, e , compartment to the infectious, i , compartment: which occurs at a rate is σe .

The data we are primarily concerned with is the incidence of influenza, as represented by the number of confirmed cases of influenza each week, i.e., the weekly incidence. Since there are a plethora of aspects of influenza epidemics not represented in our model, there will undoubtedly be discrepancies between the incidence in our model and the observed case counts. We model these discrepancies as statistical fluctuations about the incidence in the model of transmission. For example, in our SIR model if the *actual* number of observed cases over $[t_a, t_b)$ is Y this could be modelled as coming from a Poisson distribution,

$$Y \sim \text{Poisson}(N p_{\text{obs}} \mathcal{I}),$$

where N is the size of the population, p_{obs} is the probability of an individual being observed when they are infected, and $\mathcal{I}$ is the incidence from Equation (1.2). When considering a sequence of such observations, each is treated as independent random variables conditional upon the model incidence in their respective week.

In reality, not all cases are observed, there is substantial heterogeneity in observation, and there are exogenous infections which produce a background of observations throughout the year, e.g., cases where people have been infected in one location and sought care in another. The observation model used in the forecasts in Chapters 3 and 4 makes use of such an observation model. However, it is an elaboration of the one presented above: a negative binomial distribution is used, and a linear model is used to account for the observation probability and the background observations [75]. In the next section, we present the statistical framework used to fit these models.

1.6 Bayesian statistics

Above we have discussed mathematical models used to represent epidemics, and how we might represent the process of observing epidemics. Models of epidemics allow us to formalise our reasoning about their behaviour and study the impact of potential interventions. For instance, by studying how various interventions affect the effective reproduction number we can determine if they would avert an epidemic (under the assumptions of the model). In any modelling exercise, we make deductions about the behaviour of the model and, assuming the model is a faithful representation of reality, trust the same deductions will be approximately true of the real world. Usually, these deductions are dependent upon the parameterisation of the model, so it is essential to choose parameters which mean the model representation is as close to the real world as possible. One of the main goals of statistics is to provide the theoretical tools needed to discover the values of the parameters this is likely to be⁶.

While mathematical modelling is an exercise in deduction, drawing inferences based upon these models is, to some extent, inductive: we are attempting to expand our knowledge,

⁶A precise definition of “closest”, and the precise interpretation of what is meant by a scenario being likely to be the closest to the real world differs among approaches to statistics [76].

based upon our model and observations, some fact about the real world. Obtaining threshold theorems about the behaviour of epidemic models is deduction; estimating the parameters of an epidemic model from observations is induction. Of course, statistical theories are still as logically sound as those of mathematics. This is only to say that there are differences in the meaning of their results.

In this thesis, we use the Bayesian approach, which is characterised by treating unknown quantities as random variables [77]. For example, if I were to select a card at random from a pack without looking at it, the suit of the card would be unknown to me, and I would consider each suit equally likely. From a Bayesian point of view, the suit of the card is a random variable which has a uniform distribution over the outcomes: spades, clubs, diamonds and hearts. The distribution which represents our initial beliefs about the unknown is called the *prior* distribution.

Additional information may subsequently alter our beliefs about unknown quantities. For example, if I drew another card from the pack and observed it to be the ace of spades, this would suggest that the first card I had drawn is more likely to be any suit other than spades. In which case I would no longer believe the suit of my card to be uniformly distributed over the four options: I would consider clubs, diamonds and hearts to be more likely than spades. The result of updating our prior distribution based on observed data is called the *posterior* distribution.

We will now consider something so obvious that it is easy to overlook and is rarely addressed explicitly. The description above did not specify *how* to update the distributional representation of beliefs. Bayes' theorem provides an equation for computing conditional probabilities, a task which is notoriously difficult for humans [78]. The conditional distribution provides a rational, and convenient way to update these distributions and Bayes' theorem has become the de facto approach to doing so. It should be noted that the conceptual leap of Bayesian statistics was not the discovery of an equation for manipulating conditional probabilities. The conceptual leap came in representing beliefs about unknown quantities as distributions in the first place.

Let θ denote the unknown quantity we care about. Initially, we represent the distribution of possible values of θ with the prior density, $\pi(\theta)$. If we then perform an experiment and collect data, $\mathcal{D}$, the question we must ask ourselves is "what has $\mathcal{D}$ told us about θ ?" Developing a model for the experiment leads to the distribution of $\mathcal{D}$ given θ , with density $f(\mathcal{D}|\theta)$ but the Bayesian is interested in the distribution of θ given $\mathcal{D}$, denoted $p(\theta|\mathcal{D})$. Bayes' theorem tells us

$$p(\theta|\mathcal{D}) = \frac{f(\mathcal{D}|\theta)\pi(\theta)}{f(\mathcal{D})}$$

where $f(\mathcal{D})$ is the marginal density of $\mathcal{D}$. The majority of the time it will not be possible to evaluate $f(\mathcal{D})$, so it is much more common to see this written as

$$p(\theta|\mathcal{D}) \propto f(\mathcal{D}|\theta)\pi(\theta).$$

This proportionality emphasises that the dependence on θ only enters through the likelihood, f , and the prior distribution, π .

In an epidemiological context, data is often scarce, which makes the utilisation of existing knowledge desirable since it allows us to obtain more precise estimates. The Bayesian approach to statistics provides a natural way to account for this existing knowledge via the prior distribution, though this is easier in theory than in practice. Chapter 4 presents

an approach to simplify the task of incorporating this existing knowledge into the prior distribution.

While it is somewhat tangential to the application of epidemic forecasting, the discussion of prior distributions will be easier if a bit of context is provided regarding the two main interpretations of Bayesian probability: the subjective and objective. The distinction can seem irrelevant to applied statistics, but as we will see, the different interpretations suggest different choices of prior and which can lead to different outcomes. In the following section, we will see a worked example of the impact it can have.

1.7 Subjective vs Objective: it's all Bayesian to me

As George Santayana says, “Those who cannot remember the past are condemned to repeat it.” In a sense, the prior distribution, and subsequently the posterior, plays the role of memory, recording what knowledge we have *before* we carried out some experiment. We start with one distribution for an unknown quantity, then after observing some relevant data, based on this observation, we update the distribution of this unknown quantity. Whether you consider this distribution to be a representation of subjective belief about the value of the unknown or the result of a deduction from first principles may seem irrelevant, but it can lead to substantial differences in the inferences drawn. Since being condemned to repeat the same statistical analysis ad infinitum does not appeal to me we will now look a little closer at the prior distribution.

1.7.1 Interpretations of Bayesian probability

We begin by exploring approaches taken to choosing prior distributions. The choice of prior is a controversial issue which has fuelled much of the development of Bayesian theory [79, 80]. Upon casual inspection the prior distribution may seem to be an algebraic requirement, merely there to balance an equation, however, it has its roots in the interpretation of Bayesian probability. The interpretation of what probability *is* may seem irrelevant to the working scientist — only of interest to philosophers of statistics — *however*, it has repercussions which manifest in the way we choose prior distributions. Here we are interested in two particular interpretations of Bayesian probability: the subjective interpretation of Bruno De Finetti [81] and the objective interpretation of Edwin Thompson Jaynes [82]. There are subtleties to these interpretations that we will not delve into here; instead, we will consider just the components required to understand the practical implications.

The objective interpretation, á la Jaynes, considers probability theory as an extension of logic, i.e., an extension which can handle uncertainty in the truth of a statement. This school of thought seeks to establish a universal way to handle uncertainty, and as a result, it pursues algorithmic ways to select prior distributions, to avoid needing to resort to personal preference. Popular approaches include: appeals to the principle of maximum entropy, championed by Jaynes [82], seeking priors which are invariant under re-parameterisation, typically, through the use of Jeffreys priors [83], the reference priors of Bernardo [84, 85], and the use of improper priors [86]. These approaches usually lead to non-informative prior distributions, i.e., a function which plays the role of a prior distribution but has as little influence on the results as possible. One might reasonably suggest the ultimate goal of the objective school is to reduce the choice of prior to a matter of convention.

In contrast, the subjective interpretation treats probability as a degree of belief, which can

reasonably differ between individuals. However, for such beliefs to be taken seriously, it is usually assumed that they are *coherent*. A set of beliefs about certain events is *not* coherent if and only if it is possible to make a Dutch book against them, i.e., it is possible to construct a set of wagers which individually appear rational (to the belief holder), but which guarantee an overall loss [81, 87]. The stereotypical subjective prior is one which concentrates probability mass in accordance with the individual's beliefs. The details of how the probability mass is distributed presumably informed by a review of relevant literature, or existing understanding of the system under consideration [88, 89]. The subjective school does not have a goal in the way the objective school does, i.e., under their interpretation, there is not a unique universal prior distribution one should seek. However, it seems reasonable to assume they would advocate for the construction of a prior distribution making use of all available knowledge.

1.7.2 Requisite gambling example

As is customary in the statistical literature, we will use a gambling example to illustrate how the philosophical distinction described above manifests in applications. Although the example is somewhat tongue-in-cheek, it is useful for the purposes of illustration. And along the way we can contrast the Bayesian and frequentist approaches⁷. In this example, proponents of each approach are offered a wager and must decide whether to accept this wager. By following their reasoning we see an example of how their approach influences their decision-making.

An individual — scheming to make some quick money — has a dollar coin which *always* comes up heads. They approach a statistician and flips the coin twice, of course, both time it comes up heads. The individual offers the statistician the following wager (obviously without revealing the true nature of the coin):

If the coin comes up tails on the next flip, the statistician keeps it (the statistician gains 1 dollar). If the coin comes up heads the statistician must pay x dollars (the statistician loses x dollars).

First, we consider how a frequentist might handle this situation. Their null hypothesis is that the coin is fair; the probability that the coin comes up heads, p , is $1/2$. The two heads observed does not provide sufficient evidence to reject their null hypothesis, so they make their decision assuming $p = 1/2$. Treating the coin as a Bernoulli trial, they expect a gain of $(1 - x)/2$ dollars if they accept the wager, so they will accept it if $x < 1$ because in that case, they expect to make a profit.

A Bayesian statistician (again treating the coin as a Bernoulli trial), expects to gain $1 - (1 + x)\mathbb{E}[p]$ dollars from accepting the wager. Note, that for expectation, $\mathbb{E}[p]$ is taken with respect to the distribution on the probability of heads *after* witnessing the first two heads. If the statistician follows the objective approach of Jaynes, they may initially suppose the probability of the coin coming up heads, p , has a uniform distribution (this being the maximum entropy prior). Conditioning upon the two heads observed this would become Beta(3, 1) [77]. Therefore, they expect to win $1 - (3/4)(1 + x)$ dollars and will accept the wager for $x < 1/3$.

Supposing instead that our statistician follows the subjective approach of De Finetti; they know that coins are usually fair, but they have a hunch that the individual may be trying to

⁷This is not an attempt to argue for the superiority of either method, it is just to remind the reader there are alternatives.

deceive them; what fair and rational individual would offer such a wager for $x < 1$? This hunch is difficult for the objective Bayesian to justify incorporating into their analysis. Since the subjectivist treats their prior as a personal belief, they can justify using a prior in which the coin is fair with probability $1/2$ and fixed to always come up on one side with probability $1/2$. Therefore, their prior belief regarding p has point masses at 0 , $1/2$ and 1 with probability $1/4$, $1/2$ and $1/4$ respectively. Upon seeing the initial two heads they update the probabilities at the point masses to 0 , $1/3$ and $2/3$. They expect to win $1 - (5/6)(1 + x)$ dollars and hence will only accept the wager if $x < 1/5$.

The subjectivist stands to lose less than the objectivist in this scenario because of their distrust of strangers offering wagers about coin flips. The objectivist could make use of the additional information, however, it would be harder to do so while maintaining their style of Bayesianism. I constructed this scenario to demonstrate how the subjectivist approach can be beneficial, but this is not always the case. If challenged to *estimate* the bias on the coin after observing more coin flips, the subjectivist would be in trouble for any $p \notin \{0, 1/2, 1\}$. To be fair to the frequentist, let's consider the situation where the coin is flipped a further 8 times. Now when the frequentist is offered the wager, they reject the null hypothesis. They infer that the most likely situation is that the coin *always* comes up heads and will not accept any wager where $x \geq 0$.

We now consider the decision of yet another Bayesian. This one has extensive experience with coins. They started with the same beliefs as the objectivist but had already seen 200 flips of a fair coin in the past (which had resulted in 100 heads and 100 tails). Consequently, they have a $\text{Beta}(101, 101)$ prior distribution for the probability that a coin comes up heads, i.e., they have assumed coins are all identical. Upon hearing of the 10 consecutive heads they update their distribution to $\text{Beta}(111, 101)$; they will accept the wager if $x < 101/111$.

Provided the individual with the loaded coin can do some algebra they stand to walk away with the statisticians' money since the coin *always* comes up heads. But what does this example tell us about the different styles of Bayesianism, and why was the frequentist involved? Well, we have seen that in the low data setting, only observing the first 2 heads, the subjectivist would have lost less than the objectivist, because they incorporated more information into their decision making. However, drawing on prior knowledge which was not appropriate led the final Bayesian to ruin. They transferred too much of their prior knowledge into a new setting and it drowned out the evidence of 10 heads in a row.

What should we learn from the frequentist? Initially, they had insufficient data; it would take 5 consecutive heads to reach a p -value below 0.05. However, once sufficient data became available, they made what would be considered a good choice. Moreover, at no point did they need to perform any integration. In this example, the computation is trivial, but this is rarely the case for modern applications where the required integration is usually intractable.

This example is not an attempt to argue for any particular style of statistics; it is a demonstration of the potential issues with each style. As a final note, perhaps we should note the former derivatives trader [90], who, upon observing these events recalls an important study on the dynamics of flipped coins [91] and wonders why none of the statisticians questioned the assumption of independence in the coin flips.

1.7.3 A note on computation

No discussion of Bayesian statistics is complete without a mention of the algorithms used to carry it out. The rise of digital computers and the use of Markov chain Monte Carlo (MCMC) methods brought Bayesian statistics to the masses [92]. It is worth noting that computational

difficulties still exist, i.e., while MCMC will converge to the correct distribution, this can take a *long* time. Weakly-informative priors are used to provide enough information about the plausible values of parameters so as to guide MCMC toward sensible regions of parameter space — perhaps even resolve identifiability issues — but the intention is to do so without influencing the results significantly. If this can be justified by knowledge of the data generating process, then it seems to be admissible under both the subjective and objective approaches. A word of caution: MCMC is the de facto way to perform Bayesian analysis, however, the choice of algorithm should not be used as a justification for a choice of prior.

To round off this note on computation, let us take a step back and consider the role of algorithms in applied statistics. In an applied context, the problem should dictate the method, not the other way around. That said, it would be naive to ignore the influence familiar methodology has on the way we approach problems. To increase the variety of our approaches, we note the following algorithms:

- Filter based algorithms are designed to handle problems with a strong sequential structure, which they exploit to decompose the problem into a sequence of sub-problems. These algorithms often derive from the Kalman filter [93], which performs an exact computation of an approximate problem, or the particle filter [94], which performs an approximate computation of the exact problem.
- Approximate Bayesian computation (ABC), which uses simulation to approximate the posterior distribution [95]. Given observed data D , ABC draws samples from a distribution, $p(\theta|B)$, used to approximate $p(\theta|D)$. The data B is a *set* of possible observations which are close — typically, though not necessarily in a Euclidean sense — to the actual data D . Since this approach does not require the likelihood function to be evaluated it can be useful when working with complex models.
- Variational methods start by assuming the posterior distribution is within a parametric family and then find the optimal family member [96]. An appealing aspect of this approach is that re-casting the problem in terms of optimisation allows one to make use of techniques from mathematical optimisation.

1.8 How do these methods allow me to do better quantitative epidemiology?

The material surveyed above spans applied mathematical modelling, stochastic processes and the foundations of Bayesian inference. It is important to remember, the research presented in this thesis was performed in the context of a larger project aimed at improving our ability to characterise and forecast influenza epidemics. In the spirit of the applied nature of this project, the problem dictated the method. The resulting ensemble of techniques developed and the process of developing them should generalise to other applied projects.

2

Literature Review

2.1 Introduction

This literature review is divided into two sections. The first section surveys recent applications of mathematical epidemiology. The second section, on theoretical work, surveys the technology applied in the previous section. Both sections are of course incomplete, but they provide a taste of the literature that influenced this thesis.

2.2 Recent applied work

These references have been selected to demonstrate some of the areas in which mathematical methods have been applied to problems in epidemiology. In particular, references which demonstrate how modelling can be used to understand transmission at the population level.

Before we consider models of pathogen transmission, it is useful to have an understanding of the influenza virus: the pathogen we are most interested in. Influenza is important to global health; the 1918 pandemic caused the deaths of tens of millions of people [97]. The reader may find the review by Nicholson *et al* [4] useful. It provides much of the background on influenza required to understand this thesis. Shaman and Melvin [8] re-analyse experimental data and found that absolute humidity has a stronger effect on the survival and transmission of the influenza virus than temperature or relative humidity.

The modelling and forecasting of influenza transmission were encouraged in 2008 when Google made public their Google Flu Trends (GFT) project. Google used trends in search engine queries to predict time series of influenza-like illness (ILI) curated by the CDC [33]. We used the output of GFT as the data source for our first mechanistic forecasts in [98]. Despite initial optimism, Google Flu Trends ended up being used [99] as an example of the dangers of assuming “big data” is the solution to all problems.

Not long after the launch of Google Flu Trends the 2009 H1N1 pandemic began. Shaman *et al* [9] find a strong correlation between the onset of influenza epidemics and decreased absolute humidity in populations around the United States of America. Yaari *et al* [10] analyse time series of influenza in Israel. Model selection finds weather has a significant

effect on transmission. Chowell *et al* [100] analyse time series of influenza cases in Mexico to understand the impact of the school cycle on transmission. Ong *et al* [101] report on their work setting up an ILI surveillance system in response to the 2009 pandemic. They used a particle filter to generate real-time forecasts.

In a scoping review [28] Chretien *et al* used the query `influenza AND (forecast* OR predict*)` to search for studies which tested forecasting methods against independent data. They advocate for the development of best practices for influenza forecasting with standardisation of performance metrics, and additional effort put into model selection and sensitivity analysis. They encourage a closer relationship between modellers and public health officials so theoretical development matches real-world needs, a sentiment reiterated by Viboud and Vespignani in [102]. They describe the requirement of “methods to incorporate subjective input into quantitative forecasting models”. Recent work by Moss *et al* [103] reports on engagement with public health practitioners and an approach to incorporating such input.

Despite the underwhelming performance of GFT, there is still a very strong desire to make use of big data. Recently Lu *et al* [104] combined Google Trends technology with a network model to improve nowcasting capability.

Another popular approach is to fuse multiple data sources to reduce the biases that may be present in any single data source. In [2] we fused data from multiple epidemiological surveillance systems, and in [105] we incorporated climatological data to improve forecasts. A more nuanced analysis of the data generating process is given in [106], where we use crowdsourced data to understand the output of traditional surveillance systems. See [103] for an account of the real-world implementation of these methodologies.

In early work by Koelle *et al* [17] an SIR-type differential equation model was described which has an explicit representation of influenza strains. This work is important because it provides a link between genetics and epidemiology. Du *et al* [16] further develops these ideas in a simplified setting, to forecast influenza epidemics. Hadfield *et al* [15] describe their platform for tracking viral strains (based on genetic data) with an emphasis on sharing data and methodology.

Reed *et al* [107] use data collected by a sentinel surveillance system to refine estimates of the burden of influenza in the United States. They conclude existing approaches underestimate the burden of influenza, demonstrating how one data sets can correct for biases in another.

For several years, the CDC in the United States has run an influenza forecasting competition. Biggerstaff *et al* [108] describe results from the 2013–2014 CDC influenza forecasting competition. Ben-Nun *et al* [109] describe their participation in the 2016/2017 CDC Flu Challenge. They used a combination of models, used an informative prior constructed in a method similar to the one we reported in [106].

A successful competitor in the CDC competitions has been ensembles of other contributed models [30, 110]. In the spirit of such ensemble approaches Farrow *et al* [111] set up a website to collect predictions of influenza activity and then combined these to generate their forecasts.

There are many pathogens that burden humanity. The year 2013 saw the beginning of the West-African Ebola epidemic. Fisman *et al* [112] propose an SIR-type model with a modified force of infection which decreases in time, governed by a discount factor. Chowell *et al* [63] analyse time series from the Ebola epidemic concluding transmission at the local level was sub-exponential, but aggregating to larger areas can make it appear exponential. In related work, Viboud *et al* [65] introduce a phenomenological model to study sub-exponential transmission for a variety of historic epidemics.

In 2002, severe acute respiratory syndrome (SARS) became the first severe epidemic of

the 21st century. Choi *et al* [113] fit an exponential growth model to SARS case counts. The approach is presented as a simple way to arrive at an estimate of the burden if no efforts are taken to reduce transmission. In a more refined analysis, Lipsitch *et al* [114] analyse time series of SARS cases in order to estimate the reproduction numbers and investigate the potential impacts of interventions. In [115] we see a modelling study of SARS in Hong Kong and Toronto notable for its early use of a time-dependent (decreasing) contact rate. Klinkenberg *et al* [72] perform a study demonstrating the potential utility of contact tracing as a means to control epidemics with implications for a SARS-like pathogen.

The Zika epidemic began in 2015. Chowell *et al* [64] use phenomenological models to perform a retrospective forecast of Zika. A clever choice of statistical model, the generalised-Richards model, provides a natural estimate of the final size. This work also builds the case for the use of sub-exponential models.

2.3 Theoretical work

The previous section listed some recent applications of mathematical models to study contemporary epidemics. In this section, we consider the development of these mathematical models in their own right. The material is loosely divided into three sections: the models used to describe epidemics, the underlying theory of probability and stochastic processes, and the Bayesian statistics which combines the models with data.

2.3.1 Modelling of epidemics

As described in Chapter 1, quantitative approaches to epidemiology go back to Bernoulli and D'Alembert [34, 116, 117]. However, it is the work of Kermack and McKendrick that is usually recognised as laying the foundation of modern mathematical epidemiology [35, 36]. Starting with mild assumptions about the transmission and infection processes they developed the SIR model in a general form and then considered some special cases, in the process they demonstrated many of the fundamental ideas of mathematical epidemiology.

The interested reader will find comprehensive treatments of many epidemic models in the texts of Anderson and May [118] and Keeling and Rohani [119]. The text by Andersson and Britton on stochastic models [38], while slim, gives a very technical treatment from the perspective of probabilists and statisticians. The collection edited by Brauer *et al* [37], covers a broad range of important modelling techniques.

Differential equations are one of the fundamental tools of mathematical modelling. The spread of pathogens through space and time can be modelled with partial differential equations as described in the second volume of Murray's famous text [120]. However, typically ordinary differential equations (ODEs) are more convenient. Stroud *et al* [41] analysed output from a spatially explicit agent-based model to investigate how to adjust models without a spatial component to reduce their bias. They found they can adjust ODEs to account for spatial effects by throttling the incidence term. Kuznetsov and Piccardi show that even simple ODE models of epidemics can exhibit chaotic solutions [24]. Dushoff *et al* [25] observe that even very small fluctuations in parameter values of SIR-type models can lead to resonance which has strong effects. For example, minor changes to the force of infection can have a large effect on the size of an epidemic which is an important observation for the work in this dissertation. The review of Chowell *et al* [121] (and the references therein) survey evidence that early growth of epidemics is frequently sub-exponential, they review a range of models which have

sub-exponential growth of incidence.

2.3.2 Probability theory and stochastic processes

The previous section covers references for different models of epidemics. The following section covers references for the mathematical techniques for manipulating such models, with an emphasis on Markovian processes.

The very readable text by Ross [47] is an excellent introduction to probability theory. A field steeped in tradition, probability has many classic texts such as Feller's [59]. Volume 1 provides the majority of the mathematical techniques used in this thesis, however, it is substantially more technical than other introductory texts.

Continuous-time Markov chains (CTMCs) are a popular approach to modelling epidemics. They are conceptually simple and very flexible. However, they can be difficult to handle numerically when used to model large populations.

In [49] Gillespie popularised the Gillespie algorithm (a.k.a., the Doob-Gillespie algorithm) to simulate exactly from a CTMC. Initially developed for simulating chemical reactions, this has become a staple of simulation-based studies. However, the Doob-Gillespie algorithm can be computationally expensive when simulating a large number of particles (e.g., large populations). Tau-leaping is an approximate method which reduces the computational expense [52].

An alternative method is to re-formulate the model as a system of stochastic differential equations (SDE). The theory of SDE is vast and largely beyond the scope of this thesis, however, a couple of references are particularly useful. Allen *et al* [56] demonstrates a correspondence between different, but equivalent formulations of stochastic differential equations and how to construct such equations to model chemical reactions: an approach which easily can be applied to modelling transmission of pathogens. Higham [57] provides an introduction to the numerical simulation of stochastic differential equations. Together these provide a solid approach to stochastic models with large populations.

2.3.3 Bayesian statistics

We now consider the literature on Bayesian theory and its associated computational techniques. First, we will consider Bayesian theory, with an emphasis on prior selection, since this is the topic of a later chapter. Then we will consider the algorithms and techniques used to understand posterior distributions.

One of the most cited Bayesian texts is Jeffreys's [122, 123], which gives an objective perspective of Bayesian statistics and the invariant prior distributions named after him [83]. Another classic is Jaynes' magnum opus [82], which also takes an objective stance, arguing for the maximum entropy prior [124]. Bernardo [84] provides an initial foray into reference priors, using simple examples to check whether they are a sensible approach. A more refined treatment is given in [85] which, being part of a handbook contains an extensive bibliography.

A more recent text concerned with applied Bayesian statistics is the text by Gelman *et al* [77]; it focuses on the practical application of Bayesian statistics. In [125] Gelman *et al* advocate for a holistic approach to prior selection. They challenge the notion that the prior distribution can feasibly represent beliefs about parameters without reference to the likelihood. In [126] Gabry *et al* describes iterative model construction revolving around visualising the predictive distributions. In a sense, this could be considered similar to empirical Bayes. These works blur the line between the objective and subjective approaches.

At the other end of the spectrum, Goldstein provides a purely subjective perspective in [89]. He argues that a subjective approach to data analysis becomes increasingly beneficial as the complexity of problems increases, in agreement with some of the recommendations of [28].

The simulation of random variables is an essential tool for modern Bayesian statistics, so we will now consider computational simulation as it relates to Bayesian statistics. Once again, Ross provides a highly readable introductory text [127]. Starting with elementary probability theory and random number generation (inverse method, accept-reject, importance sampling, et cetera) in a slim text, he also covers variance reduction and MCMC techniques.

Much of the computation required by Bayesian statistics is simulating random variables from the posterior distribution. The Metropolis-Hastings algorithm [128, 129] is a generalisation of the Gibbs sampler which has revolutionised Bayesian methods. The expository article by Chib and Greenberg provides a highly readable derivation of the algorithm and how the Gibbs sampler arises as a special case [130].

Many pieces of software have been developed to simplify the use of algorithms to sample from the posterior distribution. An important example is the output of the Bayesian inference using Gibbs sampling (BUGS) project: WinBUGS [131] and OpenBUGS [132]. The BUGS language uses directed acyclic graphs (DAGS) to describe statistical models and then uses Gibbs sampling to sample from the posterior distribution. An important dialect of BUGS is JAGS [133] which is a re-implementation of BUGS designed to run on more platforms. A modern take on the use of DAGS to describe statistical models is Stan [134] which makes use of a Hamiltonian Monte Carlo [135, 136].

While BUGS, JAGS and Stan are generic, there are languages specialised for applying Bayesian statistics to evolutionary biology: BEAST/2 [137, 138], RevBayes [139]. More generally these languages are part of their own paradigm of probabilistic programming.

A problem related to posterior sampling is the filtering problem. The filtering problem refers to the estimation of the current state of a stochastic process based on noisy observations up until the present [140]. While the filtering problem could be solved using the posterior sampling techniques mentioned above, there has been an emphasis on finding recursive solutions which are far more efficient when new information is constantly becoming available [94].

Kalman [141] found a solution to the linear Gaussian filtering problem which now bears his name. A nice pedagogical reference for the Kalman filter is [93]. The assumptions of the Kalman filter are restrictive. The extended Kalman filter and unscented Kalman filter are two generalisations used to tackle a wider family of problems ([142] and references therein).

The particle filter is a stochastic algorithm to approximately solve the filtering problem [143]. Doucet *et al* [144] is *the* reference for sequential Monte Carlo methods (such as the particle filter). Parts 1 and 2 provide context and describe the theoretical foundations of the particle filter, Part 3 describes modifications to improve the performance of the algorithm, and Part 4 considers applications of these methods. Of particular relevance to our use of particle filtering is the chapter by Liu and West concerned with simultaneously estimating the state and parameters of a process which contains useful recommendations for smoothing parameter inference ([145] and references therein).

For a gentler introduction to the world of particle filtering see Arulampalam *et al* [94]. This provides simple derivations and implementations of several common variants of the particle filter while providing some context for the filtering problem. Kitagawa [146] demonstrates the role of particles filters in both filtering and smoothing. Several worked examples are included and the appendix contains a detailed description of various re-sampling techniques.

For clear examples of the use of a particle filter in epidemic prediction see Skvortsov *et al* [58] and King *et al* [147].

While using particle filters to forecast influenza epidemics, Yang and Shaman [148] developed a technique, *space re-probing*, which helps to avoid degeneracy in the particles (and subsequently allows the filter to use a smaller number of particles, improving efficiency).

2.4 Discussion

Mathematical epidemiology draws on much of applied mathematics and statistics, computer science, and epidemiology. As such this review contains only a small portion of the literature in these fields, however, it should provide all of the necessary background material required to understand this thesis. Of course, there are other fields that are highly relevant to mathematical epidemiology which are not represented here: geography and network science, genetics and phylodynamics, and health economics to name a few. Each can provide insights into different aspects of transmission and the impact of various interventions. For this reason, I suspect we will see increasing collaboration between them leading to a more comprehensive understanding of the complexities of epidemics in the future.

3

Model selection for seasonal influenza forecasting

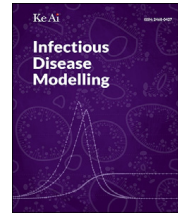
3.1 Introduction

Given multiple ways to generate forecasts of an epidemic, one might naturally want to use the one which gives the best forecasts¹. However, when the forecast specifies a distribution of possible outcomes it becomes harder to evaluate which distribution is best and even what “best” means. In the publication included below, we generalised the idea of Bayes factors (a Bayesian model selection technique) to predictive models as a way to select a model for the purposes of forecasting.

We considered the SEIR model and two variations: one with a modified incidence term to account for heterogeneous mixing in the population, and one where the incidence is influenced by climate (absolute humidity). As an application of our approach, we tested which of the three models performed best at forecasting seasonal influenza epidemics in Melbourne, Australia. We found that accounting for the effects of absolute humidity improved forecasts (relative to the standard SEIR model) while accounting for heterogeneous mixing degraded forecasts. Importantly, the improvement from accounting for absolute humidity could be obtained even with a simple prediction of this signal, i.e., one does not need to forecast absolute humidity perfectly to benefit from accounting for it.

3.2 Publication

¹Of course there will be other considerations too such as the time and effort taken to generate the forecasts, but assuming these are roughly equivalent, it seems natural to opt for the one that produces the best forecasts.



Model selection for seasonal influenza forecasting



Alexander E. Zarebski ^a, Peter Dawson ^b, James M. McCaw ^{a, c, d}, Robert Moss ^{c, *}

^a School of Mathematics and Statistics, The University of Melbourne, Melbourne, Australia

^b Land Personnel Protection Branch, Land Division, Defence Science and Technology Organisation, Melbourne, Australia

^c Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia

^d Modelling & Simulation, Murdoch Childrens Research Institute, Royal Childrens Hospital, Melbourne, Australia

ARTICLE INFO

Article history:

Received 31 October 2016

Received in revised form 16 December 2016

Accepted 16 December 2016

Available online 10 January 2017

ABSTRACT

Epidemics of seasonal influenza inflict a huge burden in temperate climes such as Melbourne (Australia) where there is also significant variability in their timing and magnitude. Particle filters combined with mechanistic transmission models for the spread of influenza have emerged as a popular method for forecasting the progression of these epidemics. Despite extensive research it is still unclear what the optimal models are for forecasting influenza, and how one even measures forecast performance.

In this paper, we present a likelihood-based method, akin to Bayes factors, for model selection when the aim is to select for predictive skill. Here, “predictive skill” is measured by the probability of the data *after* the forecasting date, conditional on the data from *before* the forecasting date. Using this method we choose an optimal model of influenza transmission to forecast the number of laboratory-confirmed cases of influenza in Melbourne in each of the 2010–15 epidemics. The basic transmission model considered has the susceptible-exposed-infectious-recovered structure with extensions allowing for the effects of absolute humidity and inhomogeneous mixing in the population. While neither of the extensions provides a significant improvement in fit to the data they do differ in terms of their predictive skill. Both measurements of absolute humidity and a sinusoidal approximation of those measurements are observed to increase the predictive skill of the forecasts, while allowing for inhomogeneous mixing reduces the skill.

We discuss how our work could be integrated into a forecasting system and how the model selection method could be used to evaluate forecasts when comparing to multiple surveillance systems providing disparate views of influenza activity.

© 2017 KeAi Communications Co., Ltd. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Influenza causes regular but unpredictable seasonal epidemics in temperate climes. Due to the difficulties of large scale data collection there is increasing interest in “now-casting” the state of influenza to improve situational awareness (Ginsberg et al., 2009; Lazer, Kennedy, King, & Vespignani, 2014). Experimental evidence suggests a decrease in absolute humidity increases the influenza virus’ ability to transmit between hosts (Shaman & Kohn, 2009), potentially driving the distinctive seasonality of influenza epidemics in temperate climes. There is also much interest in understanding the impact of contact

* Corresponding author.

networks on disease transmission, and how these effect the dynamics at a population level. Many infectious disease models assume the population mixes homogeneously (Allen, Brauer, van den Driessche & Wu, 2008; Stroud et al., 2006). An alternative approach is to restrain the rate of transmission in the model to account for inhomogeneity in the mixing of real populations (Chowell, Sattenspiel, Bansal, & Viboud, 2016; Ristic, Skvortsov, & Morelande, 2009; Roy & Pascual, 2006).

This paper investigates how allowing for the effects of absolute humidity and inhomogeneous mixing in the transmission process can improve our ability to: explain the observed influenza activity (now-casting) and predict future incidence (forecasting). Building on previous work with mechanistic models (Moss et al., 2015, 2016a, 2016b), a particle filter is used to predict the number of lab-confirmed cases of influenza observed by the Victorian Department of Health and Human Services (Australia), and to determine which model is most suitable for now-casting and forecasting. A Bayesian approach is used which can be applied more generally to the problem of selecting a model with the most predictive skill, where by “predictive skill”, we refer to the likelihood of “future” data (to be forecast), conditional on the data already observed.

Section 2 contains a description of the materials used in this analysis: the data for influenza activity and absolute humidity, the basic “transmission” model describing how influenza spreads in the community and its alternatives, and the “observation” model which connects the time series of notifications to the transmission model. Section 3 contains a description of the statistical techniques used to fit and interrogate the model along with the statistical basis for the model selection.

Section 4 contains the results of these analyses and in Section 5 we discuss the implications of this work for forecasting seasonal epidemics and how this methodology can be used for model selection when working with multiple surveillance systems providing disparate views of influenza activity.

2. Materials

2.1. Data

Influenza is a nationally notifiable disease in Australia, subsequently the Victorian Department of Health and Human Services (VDHHS) receives a notification for each specimen which tests positive for influenza in Victoria, Australia (Lambert et al., 2010). These notifications form the time series investigated in this paper. While the VDHHS is notified of positive tests, there are no data for negative tests. As a result it is difficult to distinguish between high levels of influenza activity and high ascertainment. The VDHHS captures only a small fraction of the total incidence of influenza in Victoria, the peak of the “burden of illness pyramid” (O'Brien et al., 2010; Wheeler et al., 1999). Despite these limitations, previous work ((Thomas, McCaw, Kelly, Grant, & McVernon, 2015) and (Moss et al., 2016b)) suggests that — of the data pertaining to influenza-like illness (ILI) and influenza activity generated by systems surveying this population — the VDHHS data are the least variable and most amenable to prediction. Subsequently, these data are thought to provide the best possible source of information surrounding the underlying dynamics.

While available at a daily resolution, the influenza notifications were aggregated by week to smooth the signal. Time series of relative humidity and temperature in Melbourne were obtained from the (Australian) Bureau of Meteorology (measurements taken every 3 hours). The absolute humidity (AH) was calculated from these and the results averaged over each day. These averages were then smoothed using a cubic spline (default `smooth.spline` in R) and scaled so the minimum and maximum values (over the whole 6 years) were -1 and 1 respectively. Fig. 1 displays the AH and notification time series for each of the years considered in this study.

2.2. Model

The VDHHS notification time series has previously been modelled (Moss et al., 2015, 2016b) as the observations from a hidden Markov model (HMM) as represented in Fig. 2. The hidden Markov chain, and the model for the observations are described below. The hidden “transmission model” describes how influenza spreads in the community. This is represented by the middle layer of Fig. 2 and is described in Section 2.2.1. Section 2.2.2 describes the priors for the transmission models. The “observation model” links transmission to the data collected by the VDHHS; it is represented by the bottom layer of Fig. 2 and is described in Section 2.2.3. Absolute humidity is included in the top layer of the figure with each of the AH_t indicating the time series of AH between the observations Y_{t-1} and Y_t .

2.2.1. Transmission model

The transmission model describes the spread of influenza in the population. The model is a susceptible-exposed-infectious-recovered (SEIR) type compartmental model (Anderson & May, 1992; Keeling & Rohani, 2008), where the state at time t (measured in days) is the number of people in each of the compartments, $\mathbf{X}(t) = [S(t), E(t), I(t)]^T$. The evolution of the state vector is governed by a system of stochastic differential equations described below. A closed population of $N = 4,108,541$ is used; this figure was derived from population statistics for metropolitan Melbourne (Department of Health & Human Services, 2013). Since the population is closed, the number of people in the “recovered” compartment can be obtained from the conservation law, $R = N - (S + E + I)$.

Initially everyone in the population is assumed to be susceptible to the virus, hence the initial condition for the state vector is $\mathbf{X}(0) = [N, 0, 0]^T$. In real populations there will be people who are immune to the circulating strains (John et al., 2009; McCaw et al., 2009), this is not problematic; the particle filter will converge in regions of parameter space where the basic

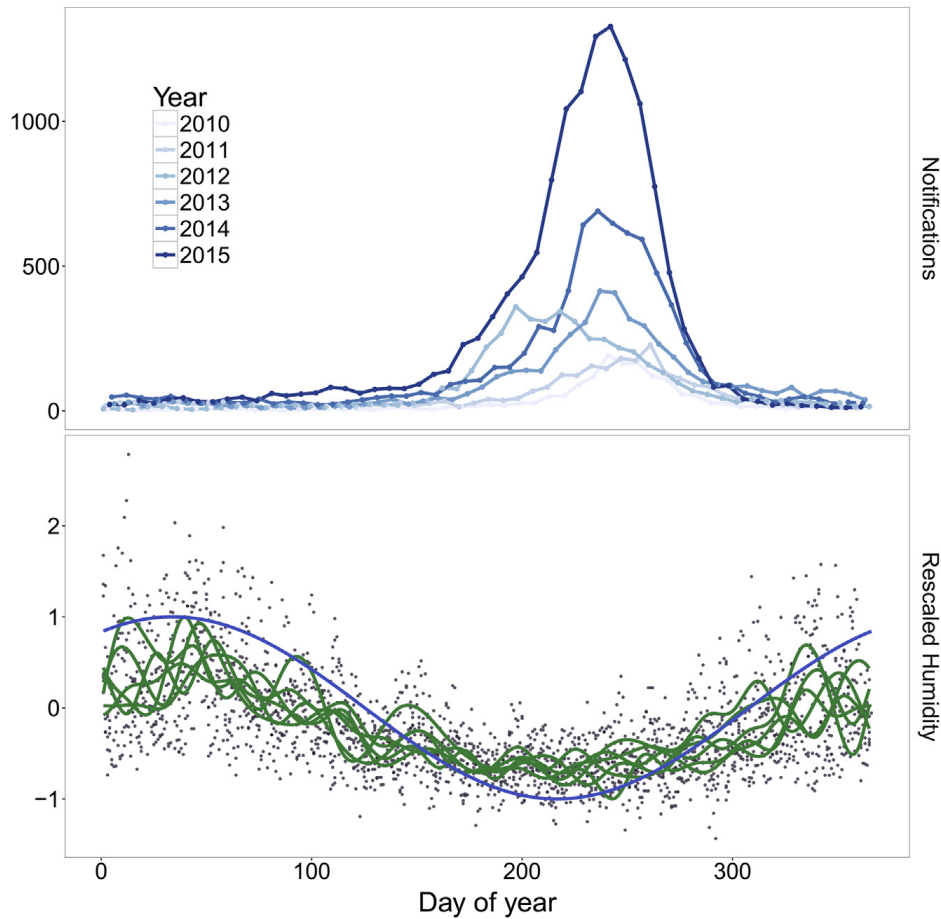


Fig. 1. (Top) Time series of the number of laboratory confirmed cases of influenza in Melbourne for 2010–15 aggregated by week. (Bottom) Scaled time series of the measurements of absolute humidity in Melbourne for 2010–15 in grey, with cubic spline smoothing in green and a sinusoidal approximation in blue. The minimum and maximum values over the whole six years were set to -1 and 1 respectively.

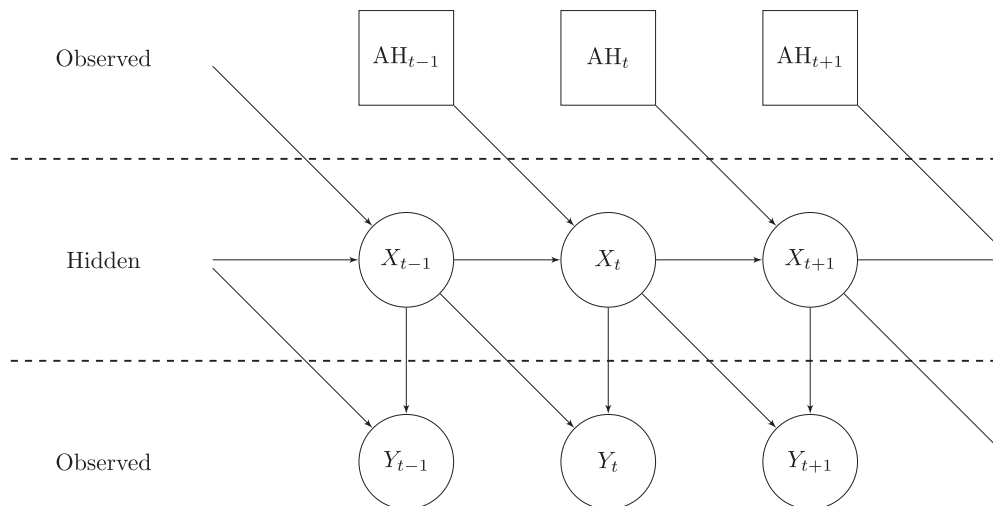


Fig. 2. Graphical representation of the hidden Markov model in which the hidden state, X_t , represents the state of the SEIR transmission model at time t and the observations, Y_t , the number of notifications over the week prior. The absolute humidity signal, AH_t , is assumed to be a deterministic function of time. The arrows indicate that: the evolution of the hidden state is dependent on its current state and the AH signal, and the observations are dependent on the current state of the hidden state and its state at the previous measurement.

reproductive number is lower to account for the increased transmission. This model supports a single outbreak assumed to start after a geometrically distributed number of days. During each day there is a fixed probability p_{exp} (set to $1/36$ (Moss et al., 2016a)) that a single individual will be exposed to the virus. This results in the state jumping to $[N - 1, 1, 0]^T$. After the initial exposure the states evolve by a system of stochastic differential equations (Allen et al., 2008a) of the form

$$d\mathbf{X}(t) = \underbrace{\mathbf{D}_1(\mathbf{X}(t))}_{\text{Drift}} dt + \varepsilon \underbrace{\mathbf{D}_2(\mathbf{X}(t))}_{\text{Diffusion}} d\mathbf{W}(t). \quad (1)$$

The drift vector is given by

$$\mathbf{D}_1(\mathbf{X}) = \begin{bmatrix} -\beta N^{-1}SI \\ \beta N^{-1}SI - \sigma E \\ \sigma E - \gamma I \end{bmatrix}. \quad (2)$$

The average behaviour of the system over a small interval of time is the same as for the deterministic (ODE) SEIR model: $\beta I/N$ is the rate at which susceptible people are exposed to the virus, σ is the rate at which people become infectious after exposure, and γ is the rate at which people recover from being infectious. The diffusion matrix is given by

$$\mathbf{D}_2(\mathbf{X}) = \begin{bmatrix} -\sqrt{\beta N^{-1}SI} & 0 & 0 \\ \sqrt{\beta N^{-1}SI} & -\sqrt{\sigma E} & 0 \\ 0 & \sqrt{\sigma E} & -\sqrt{\gamma I} \end{bmatrix}. \quad (3)$$

Stochasticity is introduced into the transmission model by $d\mathbf{W}(t)$, a three-dimensional Wiener process. The diffusion matrix transforms this stochasticity so the variance is proportional to the mean behaviour (as specified by the drift vector). As in previous work (Moss et al., 2016a, 2016b; Skvortsov & Ristic, 2012) the parameter ε specifies this proportionality and is set to 0.025. This value was selected heuristically. A degree of stochasticity is required to avoid numerical issues (such as impoverishment) when using the particle filter, however too much reduces forecast skill. Experimentation with a larger value of ε did not qualitatively change the results but created a substantial bias towards underestimation as stochastic extinction is more likely with larger values of ε (data not shown).

The model described above constitutes the null hypothesis, $\mathcal{H}_0$, for the transmission process, i.e., that transmissibility is constant and the rate of exposure varies linearly with the number of people who are infectious. Two ways of extending this model are considered: the first allows for the effect of absolute humidity (AH) on the transmissibility parameter, β , and is denoted by $\mathcal{H}_{\text{AH}}$ when using the spline smoothed AH measurements and $\mathcal{H}_{\text{sine}}$ when using a sinusoidal approximation of the AH. The second allows for the effect of inhomogeneous mixing in the population, and is denoted by $\mathcal{H}_{\text{mix}}$.

The effects of AH are introduced by allowing β to vary with a humidity signal (Shaman, Pitzer, Viboud, Grenfell, & Lipsitch, 2010; Yaari, Katriel, Huppert, Axelsen, & Stone, 2013). We consider two such signals; the smoothed measurements of AH and a sinusoidal approximation to these measurements. This produces a time dependent rate of infectious contact, β_t , that varies linearly with the AH signal so that

$$\beta_t = \beta_1(1 + \beta_2 \text{AH}(t)) \quad (4)$$

To allow for inhomogeneous mixing, the factor $N^{-1}S$ in Equations (2) and (3) becomes $(N^{-1}S)^\eta$ for $1 \leq \eta \leq 2$ (Roy & Pascual, 2006; Stroud et al., 2006). Since $0 < N^{-1}S < 1$ this has the effect of reducing the transmission rate. This modification encapsulates the idea that the number of contacts an individual has with distinct people in the population saturates as the size of the population grows, and so the probability of encountering a new susceptible individual is diminished.

2.2.2. Prior distributions for transmission model parameters

The average incubation and infectious periods, σ^{-1} and γ^{-1} respectively, have uniform priors, $\mathcal{U}(1/2, 3)$ (Beauchemin & Handel, 2011; Nicholson, Wood, & Zambon, 2003). The ratio β_t/γ is the time-dependent basic reproduction number, $\mathcal{R}_0$, which governs much of the behaviour of this model. Initial samples of β_1 (equivalent to β in the null model) are obtained by sampling the ratio β_1/γ (given γ) from $\mathcal{U}(1, 3/2)$ and solving for β_1 . As a result β_1 initially takes values from $1/3$ to $3/2$, while this may seem restrictive, it only holds for the initial samples. Due to the use of regularisation in the particle filter (described in Section 3.2.1) the full range of values $\mathcal{R}_0$ can take is $1/3$ to $9/2$ (Keeling & Rohani, 2008).

When allowing for the effects of absolute humidity in $\mathcal{H}_{\text{AH}}$ and $\mathcal{H}_{\text{sine}}$, a time dependent β_t is used which can vary by as much as 20% of β_1 . This is achieved by putting a uniform distribution, $\mathcal{U}(-1/5, 0)$, on β_2 , (Equation (4)). In $\mathcal{H}_{\text{mix}}$ the parameter η is given a uniform prior distribution, $\mathcal{U}(1, 2)$, (Stroud et al., 2006). In the null model, $\mathcal{H}_0$, the mixing is homogeneous and absolute humidity plays no role so the parameter β_2 is set to 0 and η is set to 1.

2.2.3. Observation model

Incidence, as captured by epidemiological surveillance systems, is the rate at which people enter a diseased (symptomatic) state. For the data considered here, this is the number of notifications received by the VDHHS per week, which for week n we denote Y_n . Let W_n denote the day during the n^{th} week when Y_n is observed, i.e., Y_n is the number of laboratory-confirmations

received by the VDHHS in the period $(W_{n-1}, W_n]$. The observation model specifies the distribution of Y_n given the cumulative model incidence over $(W_{n-1}, W_n]$.

Model incidence is the rate at which people move from the exposed to the infectious state, i.e., at time t this is $\sigma E(t)$. Integrating the model incidence over $(W_{n-1}, W_n]$, gives the probability, $p_{\text{inf}}(n)$, a random individual becoming infectious during that period of time, hence

$$p_{\text{inf}}(n) = \frac{1}{N} \int_{(W_{n-1}, W_n)} \sigma E(t) dt.$$

Since people can only move through the compartments sequentially, this integral simplifies to the difference in the proportion of people who *could* become infectious at the start of the week and the proportion of people who could *still* become infectious at the end of the week. Therefore,

$$p_{\text{inf}}(n) = \frac{S(W_{n-1}) + E(W_{n-1}) - (S(W_n) + E(W_n))}{N}.$$

Model incidence only accounts for those who became infectious during the week, ignoring exposures that did not progress to the infectious stage. Those who were *only* exposed are counted in the week in which their incubation is completed and they become infectious. This reflects our assumption that people begin to experience symptoms, and subsequently appear in surveillance systems, at the same time they become infectious. The observation model assumes that of the people who become infectious, on average, a constant proportion, p_{obs} , will be observed. Therefore, the expected number of notifications over $(W_{n-1}, W_n]$ is

$$\mu_{\text{epi}}(n) = N p_{\text{obs}} p_{\text{inf}}(n).$$

While this explains notifications during the flu season, influenza notifications are received throughout the year. The additional notifications are attributed to a background signal, i.e., over the period $(W_{n-1}, W_n]$ there are, on average, $\mu_{\text{bg}}(n)$ background notifications. It is assumed that everyone who is not part of the model incidence has some fixed probability, p_{bg} , of generating such a notification. Therefore, the expected number of background notifications in week n is given by

$$\mu_{\text{bg}}(n) = N p_{\text{bg}} (1 - p_{\text{inf}}(n)).$$

The number of notifications, Y_n , given the expected number of notifications, $\mu_n = \mu_{\text{epi}}(n) + \mu_{\text{bg}}(n)$, is modelled as a negative binomial random variable (Lindén & Mäntyniemi, 2011; Thomas et al., 2015), i.e.,

$$Y_n \mid \mu_n, k \sim \text{NB} \left(\mu_n, \mu_n + \frac{\mu_n^2}{k} \right), \quad (5)$$

where the variance depends upon the dispersion parameter, k . Therefore the observation model requires three parameters to be completely specified, p_{bg} , p_{obs} and k .

3. Methods

In Section 2.2.1 a null model, $\mathcal{H}_0$, for the transmission of influenza was defined. Alternatives were also given: humidity modulated transmission, $\mathcal{H}_{\text{AH}}$ and $\mathcal{H}_{\text{Sine}}$, and transmission restrained by inhomogeneous mixing, $\mathcal{H}_{\text{mix}}$. By integrating the system over the interval between observations, the transmission model is treated as a discrete time Markov chain, enabling application of standard particle filtering techniques for hidden Markov models (Doucet, Freitas, & Gordon, 2001; Sanjeev Arulampalam, Maskell, Gordon, & Clapp, 2002). The observation process is defined by the observation model of Section 2.2.3. Together these constitute the hidden Markov model shown in Fig. 2. Parameter estimation is carried out in a sequential Bayesian framework in which a sequence of approximate posterior densities are computed. Each posterior gives the distribution of the transmission model state and its parameters, conditional on the observations $Y_{1:j}$ for $j = 1, \dots, n$. A mathematical description of the problem is given in Section 3.1. The computational scheme used to generate these approximations is the particle filter (PF) (Doucet & Johansen, 2009; Doucet et al., 2001; Sanjeev Arulampalam et al., 2002). Section 3.2 describes the PF used, how forecasts are generated with it and the specifics of its implementation. Section 3.3 provides implementation details and Section 3.4 describes the estimation of the data likelihood and how this is used in model selection. Code used to generate all of the results presented in this manuscript are available online (<http://dx.doi.org/10.4225/49/5851d9ea54c65>).

3.1. Problem statement

Except where specified otherwise, we assume the parameters of the observation model, p_{bg} , p_{obs} and k , are known. Let $Y_{1:T}$ denote the first T observations $Y_1, \dots, Y_T$. The filtering problem involves computing the distribution of the hidden state, $\mathbf{X}(W_T) = [S(W_T), E(W_T), I(W_T)]^\top$, and its parameters, $\Theta = [\beta, \sigma, \gamma]^\top$, conditional upon the available data at day W_T : $p(\mathbf{X}(W_T), \Theta | Y_{1:T}, \mathcal{H}_X)$, which is referred to as the “filtering density”. The inclusion of $\mathcal{H}_X$ in the notation makes explicit the assumption of a particular hypothesis for the transmission model. The *forecasting problem* requires computing the *predictive* distribution of the hidden state for the next H weeks, which is

$$p(\mathbf{X}(W_{T+i}) : i = 1, \dots, H | Y_{1:T}, \mathcal{H}_X). \quad (6)$$

The predictive distribution of the hidden state can then be used to compute the predictive distribution of future observations, which is

$$p(Y_{(T+1):(T+H)} | Y_{1:T}, \mathcal{H}_X). \quad (7)$$

Two model selection problems are considered: selecting the model which provides the best explanation of all the data, and the model which provides the best predictions of unobserved data given limited information from the start of the epidemic. The measure of “best” being the *a priori* probability of the data given the model, $\mathbb{P}(Y_{1:T} | \mathcal{H}_X)$ in the former and $\mathbb{P}(Y_{(T+1):(T+H)} | Y_{1:T}, \mathcal{H}_X)$ in the latter.

3.2. Particle filter

3.2.1. Filtering

A bootstrap particle filter (PF) (Doucet & Johansen, 2009; Doucet et al., 2001; Sanjeev Arulampalam et al., 2002) was used to generate approximate samples (“particles”) from the filtering density. During the filtering process re-sampling was used to avoid degeneracy, a numerical issue which occurs when the majority of the probability mass accumulates in a small subset of the particles. The particles were re-sampled if their effective number dropped below 25% of their total number using deterministic re-sampling (Douc & Cappé, 2005; Kitagawa, 1996).

Re-sampling causes particles with large weights to be duplicated, and those with low weights to be removed from the sample. Therefore, each re-sampling event reduces the number of distinct particles. Since the PF simultaneously estimates the hidden state, $\mathbf{X}(t)$, and the parameters, Θ , it is important that there is sufficient diversity among the particles to properly explore parameter space. Post-regularisation is used to maintain particle diversity, as described in (Musso, Oudjane, & Gland, 2001, pp. 247–271) using a Gaussian kernel. This involves randomly perturbing the particles in a systematic way during re-sampling to ensure they are distinct, while reducing the loss of information this causes.

After conditioning on the first T observations the PF consists of a weighted set of particles, $\mathcal{C}(W_T) = \{(\mathbf{P}^{(i)}, w^i) : i = 1, \dots, M\}$. Together the particles and their weights define a discrete distribution which approximates the filtering density at day W_T . The particles, $\mathbf{P}^{(i)} = (\mathbf{X}^{(i)}(t), \Theta^{(i)})$, consists of approximate samples from the filtering density over state and parameter space. These approximate samples make up the support points of the distribution, and the associated weights, w^i , their probabilities. The number of particles used for both filtering and forecasting was 7500.

3.2.2. Forecasting

Forecasting involves computing the distribution of future observations, i.e., approximating the distribution in Expression (7). This is done in two steps. First the filtering density is used to estimate the future hidden states using the set of particles and weights in $\mathcal{C}(W_T)$ to approximate the density in Expression (6). For each particle, $\mathbf{P}^{(i)}$, a trajectory $\hat{\mathbf{X}}^{(i)}(\hat{t})$ for $\hat{t} \geq t$ is sampled (conditioning on $\hat{\mathbf{X}}^{(i)}(t) = \mathbf{X}^{(i)}(t)$ and the parameters $\Theta^{(i)}$). The ensemble of trajectories generated, the $\hat{\mathbf{X}}^{(i)}$ and their corresponding weights w^i , are then treated as a weighted sample from the density in Expression (6).

The second step involves integrating over the hidden state to obtain the predictive distribution for the observations (Expression (7)). Using the discrete approximation from the first step the integral becomes a sum, and the distribution of the future observations can be expressed as a mixture of negative binomials (Equation (5)). The joint distribution of future observations is summarised by central credible intervals (CIs), i.e., a set the observations will fall with some specified probability. For example, the 100 $\alpha\%$ CI for a random variable Y (which takes values in $\mathbb{N}$) is the set of integers $[y_{\min}, y_{\max}] \subset \mathbb{N}$ such that $\mathbb{P}(y_{\min} \leq Y \leq y_{\max}) \geq \alpha$. Therefore the values of $y_{\min}$ and $y_{\max}$ are given by

$$y_{\min} = \max_{y \in \mathbb{N}} \left\{ y : \mathbb{P}(Y \geq y) \geq \alpha + \frac{1 - \alpha}{2} \right\} \quad (8)$$

and

$$y_{\max} = \min_{y \in \mathbb{N}} \left\{ y : \mathbb{P}(Y \leq y) \geq \alpha + \frac{1 - \alpha}{2} \right\}. \quad (9)$$

Since $y_{\min}$ and $y_{\max}$ are computed independently, this is not necessarily the smallest interval into which Y will fall with probability α , however it is simpler to calculate and the difference between the two methods is negligible. The computational cost of computing $y_{\min}$ and $y_{\max}$ can be significantly reduced by approximating each component in the mixture with a Gaussian distribution with the same mean and variance. An iterative method (Newton-Raphson) was used to compute the relevant percentiles in Equations (8) and (9).

3.3. Simulation parameters

The parameterisation of the observation model is treated separately to the estimation of the parameters in the transmission model. To assess the optimal performance of the transmission models a range of parameterisations of the observation model were tried and best parameters selected. This process of optimising the observation model, the method used to select the simulations dates, and the integration method for the transmission model are described in the following sections.

3.3.1. Simulation dates

Given the presence of the background signal in the notifications, there is no definitive method for identifying the start and end dates of a flu season. This raises the question of when one should start to generate forecasts. For each of the epidemic time series, 2010–15, we started the filtering process on the 1st of May. The end date of the filtering is defined to be the date of the first observation when the cumulative case count is at least 95% of the total cases for that year. Forecasts are generated for each of the 8 weeks prior to and including the week in which notifications peaked. The different forecasting dates are used to investigate how the performance changes over the course of an epidemic. An example of the key dates in the filter/forecast process is given in Fig. 3. A full listing of the simulation dates is provided in Table 1.

3.3.2. Integration procedure

The system of SDEs describing the transmission model is integrated using the Euler-Maruyama method (Higham, 2001) with a time step of $\Delta t = 0.2$, i.e., the stochastic version of forward Euler. To ensure that during each day an epidemic is seeded with probability p_{exp} there is a Bernoulli trial at each time step with probability $1 - (1 - p_{\text{exp}})^{\Delta t}$ of seeding an epidemic. Once an epidemic has been seeded its evolution is governed by Equation (1). If a step results in any of the state variables leaving the

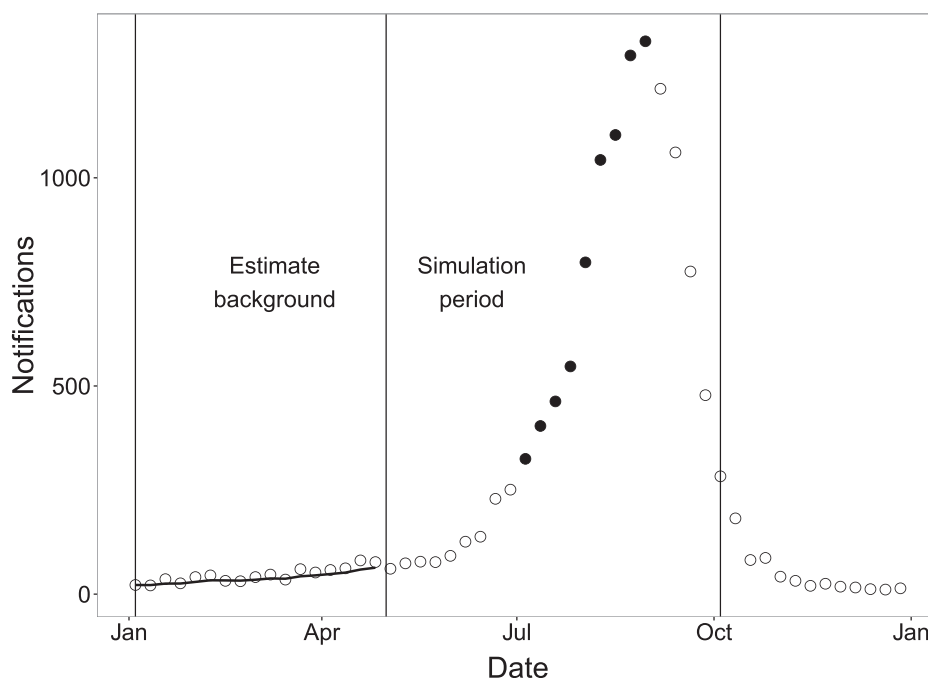


Fig. 3. Simulation periods for 2015. The first portion of the data (circles) is used to estimate the background notification rate via the exponentially weighted moving average (solid line). The second portion of the data is the target of the filtering and forecasting. The solid circles indicate the dates at which a forecast was generated.

Table 1

The dates of the first observation of the year, the date the filtering process was started, the date of the week with the most notifications and the date at which the filtering process ended. The filtering process was terminated once 95% of the year's cases had been observed. See Fig. 3 for an example of the division of the time series into a period for background estimation, and the filtering/forecasting period.

First observation	Start date	Peak date	End date	Background estimate
2010-01-03	2010-05-01	2010-08-29	2010-12-05	6
2011-01-02	2011-05-01	2011-09-18	2011-11-06	17
2012-01-01	2012-05-01	2012-07-15	2012-11-04	17
2013-01-06	2013-05-01	2013-08-25	2013-12-01	14
2014-01-05	2014-05-01	2014-08-24	2014-10-26	28
2015-01-04	2015-05-01	2015-08-30	2015-10-04	64

range $[0, N]$ then they are clipped; any subsequent change in population size is corrected by adjusting the size of the recovered compartment.

3.3.3. Observation parameters

The observation model described in Section 2.2.3 has three parameters: p_{bg} , p_{obs} and k . The performance of the PF appears to be relatively insensitive to the dispersion parameter, k , and previous work suggests a value of 100 is appropriate (Moss et al., 2015, 2016b). Subsequently k is fixed at 100.

An estimate of the background notification rate can be obtained from the elements of the time series prior to the start date. This rate specifies the background notification probability, p_{bg} . Let $B_{1:M}$ be the number of notifications in the M weeks prior to the start of the simulation period. An exponentially weighted moving average of the signal, at the i^{th} week, A_i , is given by

$$A_1 = B_1 \quad \text{and} \quad A_n = \lambda B_n + (1 - \lambda)A_{n-1}$$

with $\lambda = 0.25$ (Hunter, 1986). It is assumed that in the pre-simulation period the epidemic has not begun and so every notification is part of the background signal. This leads to the following running estimate of the background probability: $p_{bg} = [A_M]/N$, where A_M has been rounded to the nearest integer. To assess the sensitivity of the models to this estimate, several perturbations of the probability were used: $p_{bg} + j\Delta/N$ for $j = 0, 1, 2$ with $\Delta = 5$. The weighting constant, λ , was selected following the recommendations of Hunter (Hunter, 1986). However, its influence on p_{bg} is small relative to the perturbations, hence its selection is not expected to affect the fit.

Previous work suggests the performance of the PF is most sensitive to the observation probability, p_{obs} , (Moss et al., 2015). A range of plausible values can be obtained from an order of magnitude estimation. The population of Melbourne is in the millions, and each year there are thousands of notifications. Estimates of the annual attack rate of seasonal influenza range from 5 – 10% in adults and 20 – 30% in children (World Health Organization, 2014). This leads to an estimated observation probability of $\approx 10^{-2}$. Subsequently, values in $[10^{-3}, 5 \times 10^{-2}]$ were considered. The PF appears to be most sensitive to p_{obs} at the lower end of the range so the set of values considered was uniformly spaced in the logarithm.

3.4. Model selection

In Section 2.2.1 alternative transmission models were defined: the null model, $\mathcal{H}_0$, the model accounting for absolute humidity, $\mathcal{H}_{AH}$ and the model accounting for inhomogeneous mixing, $\mathcal{H}_{mix}$. This paper aims to solve two model selection problems: the first is to determine the best fitting model, the second to determine the model which performs best at prediction. The former is about understanding the importance of climate and contact structure on transmission, the latter purely pragmatic. Bayes factors (BF) are used for the model selection in both cases; the estimation procedure for the BF and its application to each problem are described below.

3.5. Likelihood estimation

The key quantity for the model selection is the likelihood of the data, $Y_{1:T}$. The PF can estimate this quantity recursively. First observe that the likelihood can be factorised as

$$\mathbb{P}(Y_{1:T}) = \mathbb{P}(Y_1) \prod_{i=2}^T \mathbb{P}(Y_i | Y_{1:(i-1)}). \quad (10)$$

While not explicit, in the equation above and for the rest of the derivation it is assumed that we are conditioning on a given hypothesis, $\mathcal{H}$. Each factor in the product can be expressed as an integral by conditioning on the relevant hidden state, therefore

$$\mathbb{P}(Y_i | Y_{1:(i-1)}) = \int \mathbb{P}(Y_i | x_{(i-1):i}) p(x_{(i-1):i} | Y_{1:(i-1)}) dx_{(i-1):i}.$$

The PF provides an estimate of the integral with

$$\hat{\mathbb{P}}\left(Y_i \mid Y_{1:(i-1)}\right) = M^{-1} \sum_{\ell=1}^M \mathbb{P}\left(Y_i \mid \hat{X}_{(i-1):i}^{(\ell)}\right),$$

where $\hat{X}_{(i-1):i}^{(\ell)}$ is the ℓ^{th} sample from the density $p(x_{(i-1):i} \mid Y_{1:(i-1)})$. Since the distribution of the observation given the hidden state is known, each term of the sum can be computed. The PF does this by using stratified re-sampling to obtain a set of $\hat{X}_{(i-1):i}^{(\ell)}$. This ensures that we have a uniformly weighted set of samples and reduces the variance in the estimator (Kitagawa, 1996; Ross, 1990). An estimator, $\hat{\mathbb{P}}(Y_{(T+1):(T+H)} \mid Y_{1:T})$, is constructed in the same way, however the trajectory of each sample is extended for H weeks.

3.5.1. Fitting

Bayes factors were estimated to determine which of $\mathcal{H}_0$, $\mathcal{H}_{\text{AH}}$, or $\mathcal{H}_{\text{mix}}$ provides the best fit. The Bayes factor is given by

$$B_{m0} = \frac{\mathbb{P}(Y_{0:T} \mid \mathcal{H}_m)}{\mathbb{P}(Y_{0:T} \mid \mathcal{H}_0)}, \quad (11)$$

for $\mathcal{H}_m \in \{\mathcal{H}_{\text{AH}}, \mathcal{H}_{\text{mix}}\}$. The probabilities in the numerator and denominator of Equation (11) are estimated using the PF as described above.

3.5.2. Forecasting

As with the model selection for fit, the forecasting ability of the PF using the various transmission models was assessed via Bayes factors, which in this situation is given by

$$\tilde{B}_{m0}(T) = \frac{\mathbb{P}(Y_{(T+1):(T+H)} \mid Y_{0:T}, \mathcal{H}_m)}{\mathbb{P}(Y_{(T+1):(T+H)} \mid Y_{0:T}, \mathcal{H}_0)}. \quad (12)$$

To estimate $\mathbb{P}(Y_{(T+1):(T+H)} \mid Y_{0:T}, \mathcal{H}_m)$ the particle approximation for the joint density of $\mathbf{X}(W_T), \Theta \mid Y_{0:T}, \mathcal{H}_m$ was integrated until W_{T+H} . Then the probability of the observations $Y_{(T+1):(T+H)}$ was estimated by summing the conditional probabilities of the observations for each of the trajectories in the ensemble.

The aggregate Bayes factors for the alternative models considering all the epidemics is calculated by assuming that each epidemic is independent and has its own parameters. As in the case of individual epidemics we assume that the optimal parameters of the observation model are known. Let $Y_{A:B}^{(y)}$ denote the notifications time series from year y , then the probability in the numerator of Equation (11) becomes

$$\prod_{y=2010}^{2015} \mathbb{P}\left(Y_{1:H_y}^{(y)} \mid \mathcal{H}_m\right)$$

and the probability in the numerator of Equation (12) becomes

$$\prod_{y=2010}^{2015} \mathbb{P}\left(Y_{(T_y+1):(T_y+H_y)}^{(y)} \mid Y_{0:T_y}, \mathcal{H}_m\right)$$

where the T_y and H_y are now year dependent. The former due to the selection of the forecasting dates which relies on aligning the epidemics by peak week, and the latter because of the method used to determine the end of the forecasting period.

4. Results

This paper considers two model selection problems, the first is to determine which of the transmission model hypotheses provides the best fit to the notification data (Section 4.1) and the second is to determine which produces the best forecast based on limited data from the start of an epidemic. The results for 2015 are given in Section 4.2 and the aggregate results over 2010–15 in Section 4.3.

4.1. Filtering results

Fig. 4 shows an example of the evolution of the filtering density for the null and sinusoidally forced models over the 2015 epidemic (See Supplementary Material Text 1 Fig. S1-1–S1-6 for the results from 2010 to 15). The whole notification time series for 2015 is shown in the figure, although the filtering density was only estimated for the observations falling in the

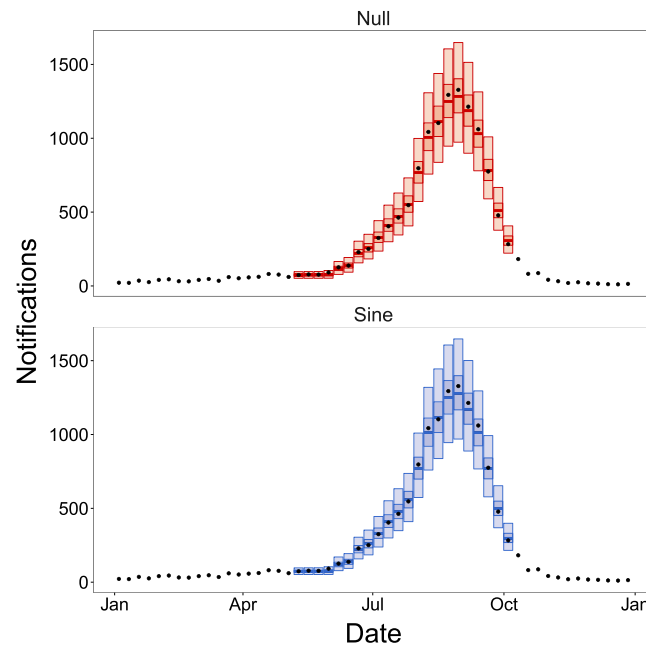


Fig. 4. The 50% and 95% credible interval for the observations under the filtering distribution for the null and sinusoidally forced models for the 2015 notification data. These running summaries of the observation distribution demonstrate both the null and sinusoidally forced models have near identical ability to assimilate new data, i.e., they have equal now-casting capabilities.

simulation period as described in Section 3.3.1. The number of notifications received for each of the weeks are represented by solid points. The box plots summarise the approximate distribution for each of the observations under the filtering distribution by showing the 50% and 95% credible intervals. As such, they act as a running fit of the model to the data and represent the most up-to-date posterior at any point in the simulation period. Supplementary Material Text 2 Fig. S2-1–S2-24 present the posterior samples for the transmission model parameters.

Using the method described in Section 3.5 the probability of the time series from the years 2010–15 arising from each of the transmission models can be estimated. The ratio of these probabilities are the Bayes factors, and the logarithm of these ratios are shown for each of the alternative models in Fig. 5. This figure shows that the transmission model that uses smoothed measurements of AH gives the best explanation of the data, however the improvement over the null model is not statistically significant. Allowing the transmissibility to vary sinusoidally, or allowing for inhomogeneous mixing, reduces the explanatory ability of the model. Again for the sinusoidal transmission hypothesis the difference is not statistically significant, however there is strong evidence that inhomogeneous mixing provides a weaker fit.

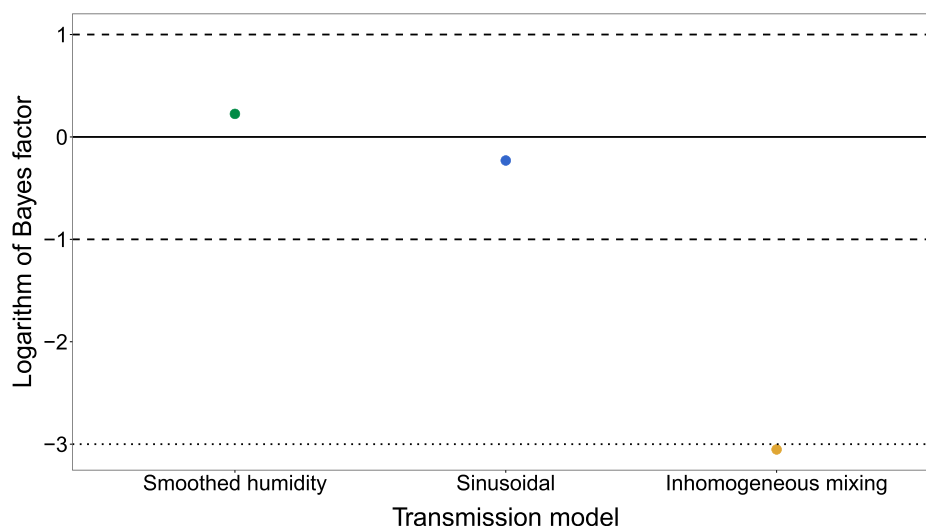


Fig. 5. The logarithm of the aggregate Bayes factor (across all the epidemics 2010–15) for each of the alternative transmission models. The solid horizontal line indicates parity with the null model, anything above this line is an improvement in model fit over the null, and below the fit is weaker. The dashed horizontal lines indicate the significance threshold.

4.2. Retrospective forecast for 2015

Much of the motivation for this work is in generating forecasts based on limited data from the start of an epidemic. Fig. 6 demonstrates some of these forecasts for the 2015 epidemic. It shows a sequence of forecasts produced by the null and sinusoidal models using increasing amounts of data from the start of the epidemic. The sinusoidal model is presented as it demonstrates a realistic forecasting tool, i.e., without the requirement of a long term, detailed forecast of AH. The full set of forecasts for all the models and years are available in Supplementary Material Text 1 Fig. S1-7–S1-30. As in Fig. 4 the number of notifications are shown as points. The solid points at the start of the epidemic were used to fit the model, which was then used to predict the values of the subsequent hollow points. This time the box plots represent a summary of the posterior distribution of the “future” observations. The rectangles collectively form the 50% and 95% credible interval (CI) of the forecast, i.e., the regions in which all future notifications are expected to fall with probabilities 0.5 and 0.95 respectively.

The first column of Fig. 6, (“Sinusoidal”) contains the forecasts generated when the transmissability varies sinusoidally over the year, the second column, (“Null”) those from the null model. For each row, the models were fit using all the data in the simulation period up until the date shown on the right. For example the first row contains the forecast generated using all the observations available on the 19th of July. For each row the model is fit to an additional element of the time series. As more data is used to train the PF the CIs should converge as the particles concentrate in regions of high posterior likelihood, resulting in narrower rectangles.

The estimated Bayes factors for the sinusoidally forced model's forecasts are also shown. The Bayes factor is largest in the second row, showing the improvement of the sinusoidal model over the null was the greatest for the forecasts generated on the 26th of July (for the weeks shown). The next section describes how the models performed when judged in this way considering all of the epidemics 2010–15.

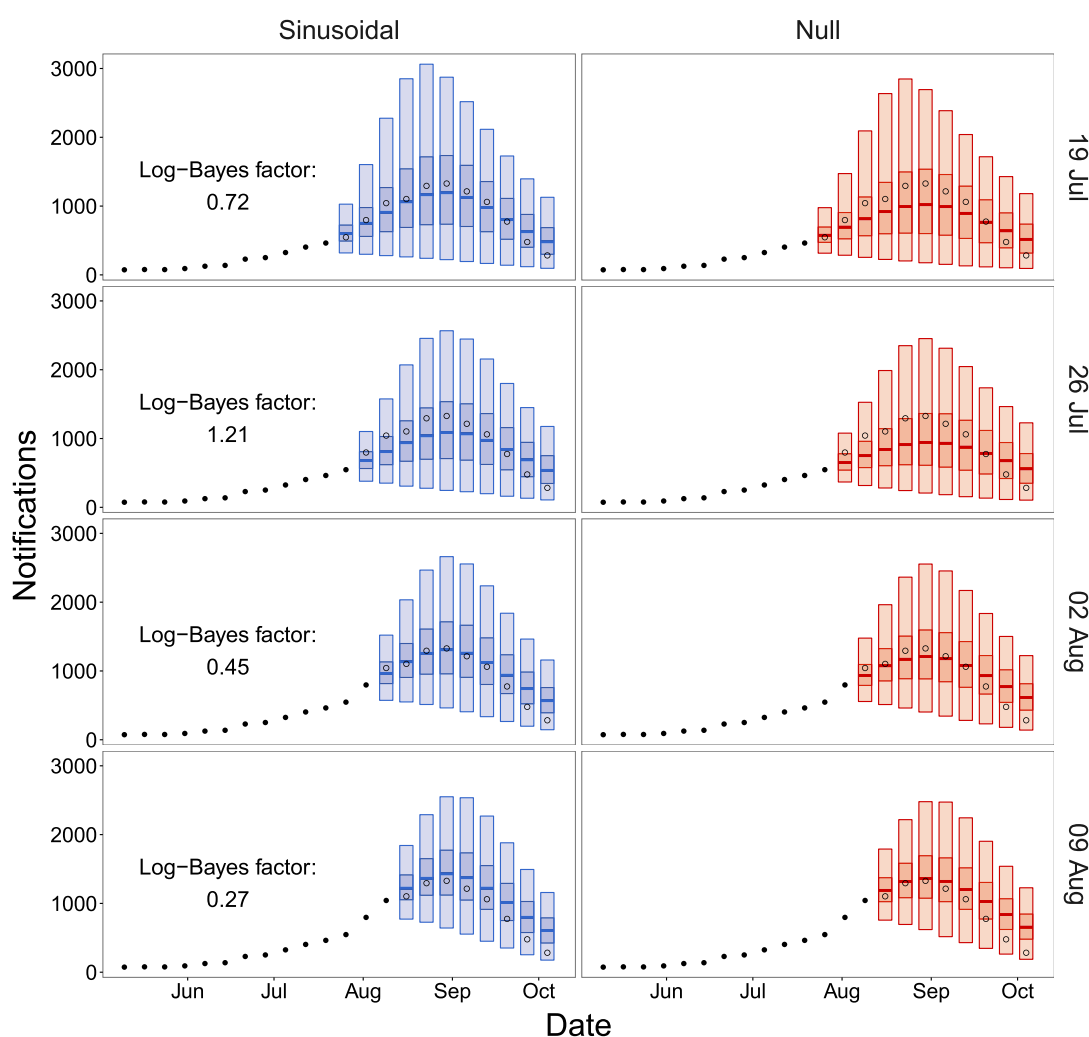


Fig. 6. Comparison of the forecasts from the null and sinusoidally forced transmission models using increasing amounts of data from the 2015 epidemic. The solid points represent “observed” data used to fit the model and the hollow points represent the “future” data, the target of the forecast. The logarithms of the Bayes factors reported describe the improvement in forecast performance by the sinusoidally forced model over the null for each of the forecasts generated.

4.3. Forecasting performance 2010–15

Fig. 7 shows the Bayes factors of the forecasts (across all the years) when aligning the epidemics by their peak week. For example, the values on the blue line were obtained by fixing the forecasting date (for all the years) relative to the peak week and computing the Bayes factor for the sinusoidally forced transmission model. Doing so gives one of the values on the curve, the full set of values coming from varying the forecasting date from 8 to 0 weeks prior to the peak week. The coloured curves represent the Bayes factors for the different alternative transmission models. This figure demonstrates that, given the correct observation model, there is strong evidence that the spline-smoothed absolute humidity model provides better forecasts than the null model over the 8 weeks leading up to the peak week. The improvement over the null model by the spline-smoothed model is largest when forecasting approximately one month prior to the peak. The sinusoidal model also outperforms the null model when generating forecasts over an interval of approximately a month, a month prior to the peak week. However, when generating forecasts more than 6 weeks prior to the peak or within a week of it, the performance of the sinusoidal model is weaker than that of the null model. As with the model fit, the forecasting performance of the inhomogeneous mixing model is poor for the majority of the season, although it does improve around the time of the peak.

An alternative method for investigating the differences between the forecasts is to consider the errors. Fig. 8 shows the average error in the forecast median as a function of the number of observed positive tests for all of the observations across the years 2010–2015. A point at (x, y) indicates that when forecasting an observation of x cases the average error in the prediction was y , so negative and positive values of y indicate underestimation and overestimation respectively. Since forecasts are generated at multiple weeks we average this error over the different forecasting dates. The solid lines show a LOESS smoothing of the scatter plot for each of the models to highlight the general trend. These show that for small numbers of positive tests the forecasts are reasonably unbiased, however all the models tend to underestimate larger observations.

5. Discussion

5.1. Principal findings

Accounting for the effects of absolute humidity (AH) does not significantly alter model fit and allowing for inhomogeneous mixing leads to worse fits. However, with a well parameterised observation model, allowing for the effects of AH improves forecasts of seasonal influenza. Moreover, even with a simple approximation of the AH (i.e., a sine wave) forecast performance is still improved. While the model accounting for inhomogeneous mixing leads to poor forecasts in the ascent phase of the epidemic once the peak has been reached this model appears to produce better forecasts of the descent phase. These results for model fit, and predictive skill relate to the aggregate performance of the models over all the epidemics from 2010 to 2015.

5.2. Study strengths and weaknesses

5.2.1. Strengths

While it is interesting to see that accounting for AH can improve forecast performance, a substitute signal (sine wave) has also been shown to improve forecast performance. This is an important observation as even the best predictor may be useless if we cannot obtain/predict it reliably.

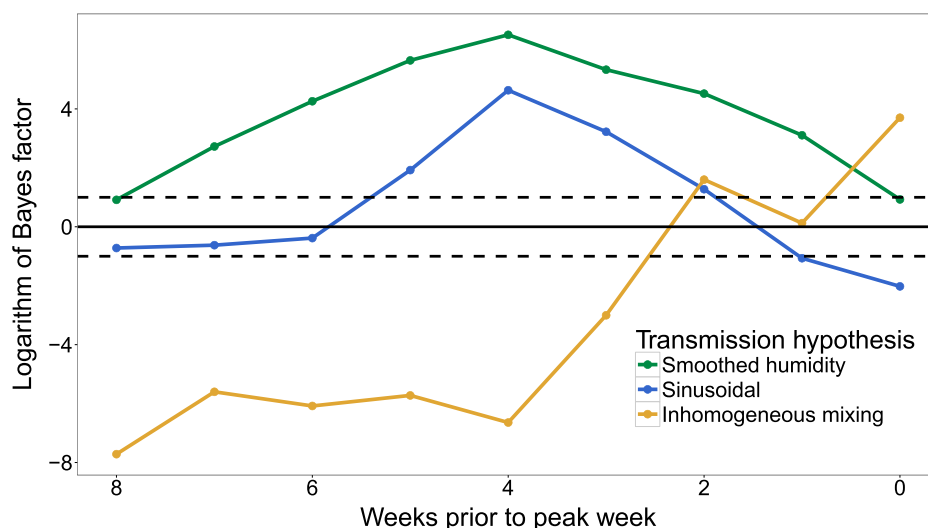


Fig. 7. The logarithm of the aggregate forecast Bayes factor (across all the epidemics 2010–15) for each of the alternative transmission models. The solid horizontal line indicates parity with the null model, anything above this line is an improvement in predictive skill. The dashed horizontal lines indicate the significance threshold.

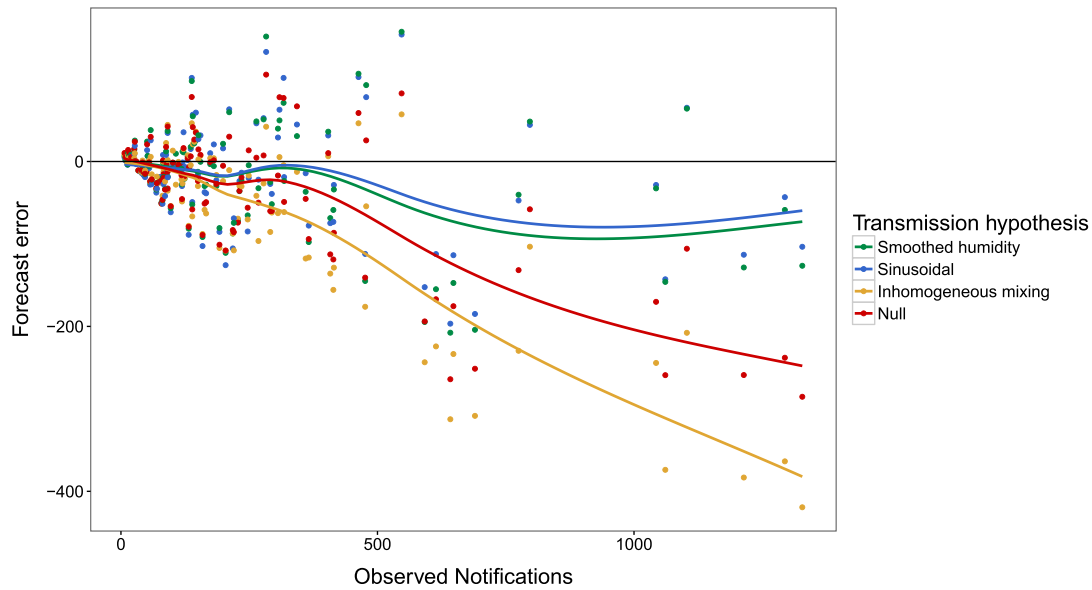


Fig. 8. Forecast error plotted against the size of the observation being forecast. Each point represents the error in attempts to forecast a single observation (averaged over the forecasts made at different points in the season). A point at (x, y) indicates that when forecasting an observation of x cases the average error in the prediction was y , so negative and positive values of y indicate underestimation and overestimation respectively. The colour of each point indicates which model was used to generate the forecast. The solid lines represents a LOESS smoothing of the data.

Attempts to forecast epidemics are often judged by their ability to predict one-dimensional summary statistics such as the timing of the peak of the epidemic (Chretien et al., 2014; Nsoesie, Brownstein, Ramakrishnan, & Marathe, 2014), in this paper a likelihood based measure is used. This measure is built upon the probabilistic model of the data and therefore avoids the need for further decision making in what to optimise for i.e., ability to predict peak time, magnitude or final size. Such an approach has several desirable properties, such as providing a natural way to assess forecast performance when there are multiple data streams. For instance, when using data from two surveillance systems showing the epidemic peaking at different times, it is unclear how to judge the accuracy of the forecast's peak time prediction. If forecast performance is assessed by the same metric as model fit, this method could be used to estimating forecast performance based on current fit to the data (i.e., does the model which best fits the most recent observation also best predict the next one?). Furthermore, the use of Bayes factors implicitly accounts for model complexity through the integration over the whole parameter space.

The PF lends itself naturally to estimation of credible intervals (CIs) of the parameters. In previous analyses of these data, forecasts consisted of the median estimate of the expected number of observations and only the CIs of this estimate. This analysis improves upon this by providing a summary of the predictive distribution of the observations. An iterative method has been presented to efficiently estimate the relevant quartiles. While the difference is negligible when case counts are high, when they are low, such as at the beginning and end of an epidemic, there is a substantial difference. Showing the uncertainty in the observation model gives a more realistic representation of the forecasts and should improve communication of these results.

5.2.2. Weaknesses

The greatest weakness of this work is the assumption of a known observation model. While estimation of these parameters was not a goal of this paper it should be kept in mind that the results presented here do require some knowledge of these parameters. Previous attempts at live forecasting in Melbourne, Australia (Moss et al., 2015) have highlighted that the observation probability, p_{obs} , in particular plays an important role in forecast performance. This presents a challenge since it appears that for Melbourne's influenza surveillance systems, this parameter changes from year to year (Moss et al., 2016b). Moreover, an important assumption of this work is that the observation probability is constant *within* a season. Without such an assumption, or a plausible alternative, it would be difficult to distinguish changes in transmission from changes in observation.

Another weakness of this study is the quantity of data. Each of the six years presents only a single time series, making it difficult to draw strong conclusions about either the transmission or observation processes. Due to this dearth of data the Bayesian framework is particularly attractive as it provides a more satisfactory quantification of the uncertainty.

When fitting nested models an effect size analysis can be informative. While the model selection revealed allowing for AH effects (i.e., allowing $\beta_2 > 0$) improved forecast performance, no analysis of the size of β_2 was performed. In part this is because of the sensitivity of the model to this parameter, which makes such an investigation difficult.

The transmission model assumes there is only a single circulating influenza strain and that initially everyone in the population is susceptible. While this is clearly false it is a useful assumption and with the existing data it would be challenging to parameterise a multi-strain model.

5.3. Comparison with other studies

5.3.1. Transmission model

Modelling the rate of transmission as dependent on AH has been done in several studies. The precise way in which the dependence is expressed varies though, with some models (such as the one presented here) using linear dependence (Bock Axelsen et al., 2014), while others using exponential dependence (Shaman & Kohn, 2009). To analyse this data set it has been assumed that the rate of infection should increase as AH decreases. This is because Melbourne has a temperate climate. In tropical climates the interaction between humidity and influenza transmission appears to be more complex and such an assumption may not be valid.

The method used to account for inhomogeneous mixing is also not the only one which appears in the literature (Chowell et al., 2016). However the majority of these approaches are similar at the core: the transmission rate should be retarded by saturation of local contacts, and this idea is realised in the functional form used here.

In the models considered in this paper it is only possible for individuals to progress through the compartments once. As a result an individual cannot be infected twice and since the population is closed the null model only supports a single epidemic. Extensions to this model allow for a loss of immunity resulting in transition from the recovered compartment back into the susceptible compartment and the incorporation of births and deaths which achieves similar changes in the dynamics. However, over the time scales considered here accounting for births and deaths will have a negligible effect.

5.3.2. Computational method

In this paper a particle filter (PF) has been used to fit and forecast the notification time series. The PF is an attractive method for this sort of analysis, and for practical forecasting work for several reasons. It is flexible in allowing for an arbitrary observation model, and maintains the non-linearity of the transmission model, whereas the Kalman filter and variants thereof usually require some form of linearisation and normality assumptions. However, this flexibility comes at the cost of potential numerical issues that arise from the use of a finite sample of particles, such as particle impoverishment and degeneracy. We have used re-sampling and regularisation to address these issues, however alternatives exist (Yang & Shaman, 2014). Alternative methods which do not require modification of the model include: particle MCMC (Doucet et al., 2001) and iterated filtering (Ionides et al., 2015). The former is significantly more computationally expensive and for the later it is less clear how one would incorporate the uncertainty in the parameter estimates into forecast generation.

5.4. Further work

The metric for forecast performance presented here provides a natural way to assess the performance of forecasts when comparing them to the data from multiple surveillance systems. This will assist in our future efforts to assimilate data from multiple sources, each with their own biases. Another approach to improve predictive skill is to incorporate more prior knowledge into the forecasts, a task to which the Bayesian framework is well suited. Specifically, by constructing an informative prior based on previous epidemics it is reasonable to expect that initial uncertainty in the forecast can be reduced.

In addition to the main challenges listed above there are a number of attractive changes which could be made to the model. For instance, the background signal is a phenomenological modification and ignores the infectious potential of these individuals, this could be corrected for by forcing a proportion of the non-incident population into the infectious class. The analysis in (Mercer, Glass, & Becker, 2011) suggests this may improve the initial estimation of the reproduction number, potentially improving forecasts early in the season. Moreover, by allowing the dispersion parameter k to vary in Equation (5) it is possible to have quite a flexible mean-variance relationship in the observation model.

5.5. Meaning and implications

We have demonstrated that our existing forecast technology (Moss et al., 2015, 2016a, 2016b) can be improved by allowing for the effects of absolute humidity (AH) in the transmission model. While the true values of AH provide the largest improvement, even a simple approximation to the AH data (e.g. sinusoidal) is sufficient to improve predictive skill. The greatest improvement using a sine wave is seen when forecasting 5 to 2 weeks prior to the peak week. After the peak it is still possible to improve upon the null model by accounting for inhomogeneous mixing. However, there is little to gain in terms of now-casting by modifying the transmission model.

Methodologically this paper presents an objective function for forecast optimisation and an iterative scheme for approximating the credible intervals of the forecast. The former allows for a model selection based on a more comprehensive comparison of forecast and data, and provides a sensible way to optimise the forecasting tool to multiple data streams. The latter, by more naturally describing forecast uncertainty, improves the communication of the results.

Acknowledgements

This work was funded by the DSTG project “Bioterrorism Preparedness Strategic Research Initiative 07/301”. James M. McCaw is supported by an ARC Future Fellowship (FT110100250). We thank Nicola Stephens, Lucinda Franklin and Trevor Lauer (VDHHS) for providing access to, and interpretation of, Victorian influenza surveillance data.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.idm.2016.12.004>.

References

- Allen, Edward J., Allen, Linda J. S., Arciniega, Armando, & Greenwood, Priscilla E. (2008). Construction of equivalent stochastic differential equation models. *Stochastic Analysis and Applications*, 26(2), 274–297.
- Allen, Linda J. S., Brauer, Fred, van den Driessche, Pauline, & Wu, Jianhong (2008). *Mathematical epidemiology*. Springer.
- Anderson, Roy M., & Robert, May M. (1992). *Infectious diseases of Humans: Dynamics and control*, 28. Wiley Online Library.
- Beauchemin, Catherine A. A., & Handel, Andreas (2011). A review of mathematical models of influenza A infections within a host or cell culture: Lessons learned and challenges ahead. *BMC Public Health*, 11(1), 1.
- Bock Axelsen, Jacob, Yaari, Rami, Grenfell, Bryan T., & Stone, Lewi (2014). Multiannual forecasting of seasonal influenza dynamics reveals climatic and evolutionary drivers. *Proceedings of the National Academy of Sciences*, 111(26), 9538–9542.
- Chowell, Gerardo, Sattenspiel, Lisa, Bansal, Shweta, & Viboud, Cécile (2016). Mathematical models to characterize early epidemic growth: A review. *Physics of Life Reviews*.
- Chretien, Jean-Paul, George, Dylan, Shaman, Jeffrey, Chitale, Rohit A., & Ellis McKenzie, F. (2014). Influenza forecasting in human populations: A scoping review. *PloS ONE*, 9(4), e94130.
- Department of Health & Human Services. (2013). *2013 local government area profiles*. Technical report, Victorian Government (Accessed 3 September 2014).
- Douc, Randal, & Cappé, Olivier (2005). Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on image and signal processing and analysis, 2005* (pp. 64–69). IEEE.
- Doucet, Arnaud, Freitas, Nando de, & Gordon, Neil (2001). *Sequential monte carlo methods in practice*. Springer Science & Business Media.
- Doucet, Arnaud, & Johansen, Adam M. (2009). A tutorial on particle filtering and Smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12(656–704), 3.
- Ginsberg, Jeremy, Mohebbi, Matthew H., Patel, Rajan S., Brammer, Lynnette, Smolinski, Mark S., & Brilliant, Larry (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–1014.
- Higham, Desmond J. (2001). An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3), 525–546.
- Hunter, Stuart J. (1986). The exponentially weighted moving average. *Journal of Quality Technology*, 18(4), 203–210.
- Ionides, Edward L., Nguyen, Dao, Atchadé, Yves, Stoev, Stilian, & King, Aaron A. (2015). Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences*, 112(3), 719–724.
- Mathews, John D., Chesson, Joanne M., McCaw, James M., & McVernon, Jodie (2009). Understanding influenza transmission, immunity and pandemic threats. *Influenza and Other Respiratory Viruses*, 3(4), 143–149.
- Keeling, Matt J., & Rohani, Pejman (2008). *Modeling infectious diseases in humans and animals*. Princeton University Press.
- Kitagawa, Genshiro (1996). Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1), 1–25.
- Lambert, Stephen B., Faux, Cassandra E., Grant, Kristina A., Williams, Simon H., Bletchly, Cheryl, Catton, Michael G., et al. (2010). Influenza surveillance in Australia: We need to do more than count. *Medical Journal of Australia*, 193(1), 43–45.
- Lazer, David, Kennedy, Ryan, King, Gary, & Vespignani, Alessandro (2014). The parable of google Flu: Traps in big data analysis. *Science*, 343(14 March).
- Lindén, Andreas, & Mäntyniemi, Samu (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92(7), 1414–1421.
- McCaw, James M., McVernon, Jodie, McBryde, Emma S., & Mathews, John D. (2009). Influenza: Accounting for prior immunity. *Science*, 325(5944), 1071.
- Mercer, Geoff N., Glass, Kathryn, & Becker, Niels G. (2011). Effective reproduction numbers are commonly overestimated early in a disease outbreak. *Statistics in Medicine*, 30(9), 984–994.
- Moss, Robert, Fielding, James E., Franklin, Lucinda J., Kelly, Heath A., Stephens, Nicola, McVernon, Jodie, et al. Live forecasting of the 2015 Melbourne influenza season using lab-confirmed influenza cases. Under review.
- Moss, Robert, Zarebski, Alexander, Dawson, Peter, & McCaw, James M. (2016). Forecasting influenza outbreak dynamics in Melbourne from Internet search query surveillance data. *Influenza and Other Respiratory Viruses*.
- Moss, Robert, Zarebski, Alexander, Dawson, Peter, & McCaw, James M. (2016). Retrospective forecasting of the 2010–2014 Melbourne influenza seasons using multiple surveillance systems. *Epidemiology & Infection*, 1–14.
- Musso, Christian, Oudjane, Nadia, & Gland, Francois Le (2001). Improving regularised particle filters. In *Sequential monte carlo methods in practice*. Springer.
- Nicholson, Karl G., Wood, John M., & Zambon, Maria (2003). *Influenza*. *Lancet*, 362, 1733–1745.
- Nsoesie, Elaine O., Brownstein, John S., Ramakrishnan, Naren, & Marathe, Madhav V. (2014). A systematic review of studies on forecasting the dynamics of influenza outbreaks. *Influenza and Other Respiratory Viruses*, 8(3), 309–316.
- Ristic, Branko, Skvortsov, Alex, & Moreland, Mark (2009). Predicting the progress and the peak of an epidemic. In *Acoustics, Speech and signal processing, 2009. ICASSP 2009. IEEE International conference on* (pp. 513–516). IEEE.
- O'Brien, Sarah J., Rait, Greta, Hunter, Paul R., Gray, James J., Bolton, Frederick J., et al. (2010). Methods for determining disease burden and calibrating national surveillance data in the United Kingdom: the second study of infectious intestinal disease in the community (IID2 study). *BMC Medical Research Methodology*, 10(39).
- Ross, Sheldon M. (1990). *A course in simulation*. Prentice Hall PTR.
- Roy, Manojit, & Pascual, Mercedes (2006). On representing network heterogeneities in the incidence rate of simple epidemic models. *Ecological Complexity*, 3, 80–90.
- Sanjeev Arulampalam, M., Maskell, Simon, Gordon, Neil, & Clapp, Tim (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2), 174–188.
- Shaman, Jeffrey, & Kohn, Melvin (2009). Absolute humidity modulates influenza survival, transmission, and seasonality. *Proceedings of the National Academy of Sciences*, 106(9), 3243–3248.
- Shaman, Jeffrey, Pitzer, Virginia E., Viboud, C. écile, Grenfell, Bryan T., & Lipsitch, Marc (2010). Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biol*, 8(2), e1000316.
- Skvortsov, Alex, & Ristic, Branko (2012). Monitoring and prediction of an epidemic outbreak using syndromic observations. *Mathematical Biosciences*, 240, 12–19.
- Stroud, Phillip D., Sydoriak, Stephen J., Riese, Jane M., Smith, James P., Mniszewski, Susan M., & Romero, Phillip R. (2006). Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing. *Mathematical Biosciences*, 203, 301–318.
- Thomas, Emma G., McCaw, James M., Kelly, Heath A., Grant, Kristina A., & McVernon, Jodie (2015). Quantifying differences in the epidemic curves from three influenza surveillance systems: A nonlinear regression analysis. *Epidemiology and Infection*, 143, 427–439.
- Wheeler, Jeremy G., Sethi, Dinesh, Cowden, John M., Wall, Patrick G., Rodrigues, Laura C., Tompkins, David S., et al. (1999). Study of infectious intestinal disease in England: Rates in the community, presenting to general practice, and reported to national surveillance. *The BMJ*, 318, 1046–1050.
- World Health Organization. (2014). *Influenza (seasonal) fact sheet N° 211*. <http://www.who.int/mediacentre/factsheets/fs211/en/>.
- Yaari, Rami, Katriel, Guy, Huppert, Amit, Axelsen, Jacob, & Stone, Lewi (2013). Modelling seasonal influenza: The role of weather and punctuated antigenic drift. *Journal of The Royal Society Interface*, 10(84), 20130298.
- Yang, Wan, & Shaman, Jeffrey (2014). A simple modification for improving inference of non-linear dynamical systems. arXiv preprint arXiv:1403.6804.

3.3 Contribution to the goals of this thesis

In the publication above, we developed a measure of the quality of probabilistic forecasts by extending the idea of a Bayes factor. This measure was used to select a model for forecasting of seasonal influenza epidemics in Melbourne, Australia. Forecasts of seasonal influenza were improved by accounting for variation in absolute humidity, which is not surprising as there is experimental evidence to suggest absolute humidity is an important factor in the ability of influenza to transmit between hosts. Initially, this might seem to be a daunting result, as it suggests we should simultaneously forecast both the absolute humidity and the influenza transmission, but we demonstrated that even a basic forecast of absolute humidity is sufficient to significantly improve the forecasts. We also investigated the impact of accounting for deviations from the assumption of homogeneous mixing by modifying the incidence term; this did not lead to an improvement in the forecasts.

Above, we have only considered model selection from within a set of SIR-type models. This is because we are primarily interested in generating interpretable forecasts, and hence using mechanistic models. We have not considered the relative performance of more traditional null models (e.g., a one-step ahead prediction, or the average of previous seasons) in the current work. While of course this can be done within the framework presented above, to do so would be somewhat tangential to the current investigation and is hence beyond its scope.

Although accounting for absolute humidity (via seasonal forcing) leads to a significant improvement in the forecasts, this improvement is minor, i.e., the size of the effect is small. Since existing models already do a good job of fitting time series of cases, I suspect that substantial improvements to epidemic forecasts will not come from further refinement of the transmission model but from refinement to the observation model and the statistical analysis. Consequently, in the next chapter, I opt for a simpler (mechanistic) transmission model and focus on some of the details of the statistical analysis.

4

Prior distributions

4.1 Abstract

In Bayesian statistics the prior distribution allows us to incorporate existing knowledge or beliefs when analysing data. Prior distributions representing substantial amounts of existing knowledge/beliefs are called informative and can have a strong influence on resulting inferences. Informative priors have great potential when faced with complex, potentially unidentifiable models where they can constrain the posterior. However, the use of informative priors can be challenging and there are many pitfalls.

We present the *somewhat informative prior* (SIP) distribution to simplify the use of informative priors; making it easier to incorporate existing knowledge into an analysis. The SIP encodes prior knowledge for a generic statistical model and can be constructed systematically using an algorithm we describe. We demonstrate this with an example in epidemic forecasting. We incorporate prior knowledge mined from historical epidemics to retrospectively forecast the 2017 influenza epidemic in Sydney, Australia.

Our algorithm constructed an SIP which accurately represents prior knowledge about the parameters of our model for observed disease incidence with minimal user input. Forecasts conditioned upon the SIP are more closely aligned with prior knowledge than those generated using more typical prior distributions, i.e., the SIP improves forecasts in a low-data setting.

Removing the need to manually select the parameters for an informative prior simplifies the process of subjective Bayesian analysis. This enables practitioners to spend more time thinking about *what* they know about their data and spend less time about *how* to represent it in a prior distribution.

4.2 Introduction

In this chapter, we describe an approach to the construction of prior distributions, or more specifically, how to select the hyperparameters of a prior. This approach is motivated by the notion that all knowledge about a system is relevant to inference, i.e., the prior distribution should be constructed with the whole statistical model in mind, rather than just considering

each parameter separately. A result of this approach is that the resulting prior distributions account for correlations between parameters, correlations that may only be known implicitly. Accounting for these correlations is non-trivial, so we describe a systematic way to construct such distributions and an algorithm to assist in the computation. Essentially, this process is concerned with the representation of knowledge in a probability distribution.

The *somewhat informative prior* (SIP) is a distribution over the parameters chosen such that summary statistics of the model (i.e., observable behaviour) under this distribution follow a specific distribution (i.e., behave similarly to previous observations). Importantly, this is done in such a way that an approximate solution can be constructed algorithmically. This allows the analyst to work with their model at a level of abstraction closer to the level at which we usually think of (and observe) the process being modelled. There are also additional benefits: it provides a method to share knowledge between different models of the same system, and avoid logical inconsistencies which can occur due to poorly specified priors.

We use the term “statistical model” in a broad sense, referring to anything with a sample space and a set of distributions over that sample space which are indexed by a set of parameters [149]. This encompasses both traditional statistical models (e.g., linear regression) and more complex mechanistic models (e.g., systems of differential equations and agent-based models). In each case, there is a set of parameters and a model, and each instance of the parameters describes a distribution over the possible observations of the model.

For the majority of analyses, the de facto method for constructing the prior distribution is to go through the components of the parameter vector and to select an independent distribution for each component. The selection of these distributions usually involves selecting a parametric family appropriate for that component and then selecting a member of the family based on what is known about the parameter. As a simple example, when performing linear regression, one might assume that each coefficient is drawn from an independent normal distribution. In the case of hierarchical models, this process may be repeated, but for the most part, it is similar. There are exceptions to this approach, such as the birth-death tree priors used when inferring phylogenies [150].

Constructing prior distributions in the manner described above can be difficult: when using an unfamiliar or highly complex model it may be unclear how existing knowledge pertains to the parameters, and there can be relationships between parameters in complex models which will not be reflected in a prior distribution with independent components. To demonstrate, consider a moderately complex model for the flight of a shuttlecock: you may guess that air resistance will bring it to a halt over several meters when launched at a moderate speed. However, unless you have some familiarity with fluid dynamics, it may be harder to guess a plausible range for the parameters of the drag equation, or how different combinations of angle of attack and launch speed might result in equivalent flight times. If your understanding of a system is based on observations of its *behaviour*, it can be difficult to construct a prior distribution for *parameters* of a model which are consistent with that behaviour if there is not a clear link between the two: the parameters and the behaviour.

One approach to reconcile knowledge of parameters and behaviour is history matching [151]. History matching is a method to partition parameter space into two regions based on whether the parameters produce plausible or implausible behaviour. This method was developed to study the output of functions which are computationally demanding to evaluate; for example, it was used to calibrate an agent-based model of HIV transmission [152]. History matching has also been used to choose prior distributions [153]. This approach takes plausible and implausible values of summary statistics and a degree of confidence in these values, and uses them to construct a set of constraints. It then finds potential prior distributions by

searching for members of a parametric family satisfying these constraints.

Given history matching is a way to study computationally demanding problems, the study of history matching as an approach to prior choice has focused on finding computationally efficient ways to carry out the necessary computation [153]. There are important open questions remaining: how to select the parametric family from which to choose the prior, and how to handle the case where we have a distribution for the summary statistics, rather than just constraints on its values. We address these questions with the SIP. Loosely speaking, the SIP is informative on *summary statistics*, but less informative on the actual parameters of the model. The SIP can be constructed algorithmically, and we describe an algorithm to do so. The computational requirement to construct this mapping is simple: we only need to be able to compute the summary statistic for a sample from the statistical model.

In the epidemic forecasting context, typically, there will only be a small amount of data available from the current epidemic, however, there may be a substantial amount of expert knowledge and historical data. This makes epidemic forecasting a good candidate to benefit from the inclusion of expert (domain) knowledge via informative priors [28, 103, 109, 111].

To demonstrate the utility of the SIP, we consider an example of epidemic forecasting, specifically retrospective forecasting the 2017 influenza epidemic in Sydney, Australia. Our algorithm accurately represents historical seasonal influenza epidemics in a prior distribution. With only small amounts of data from the start of the epidemic, forecasts generated using the SIP appear substantially better than those generated using a less informative prior distribution. Whereby “better” we are referring to closer alignment with what seems plausible and avoiding attributing significant probability to clearly unreasonable outcomes.

4.3 Somewhat informative prior

We refer to the SIP (as rigorously defined in terms of measures), and its approximation interchangeably. In practice, the difference will be inconsequential. Uncertainty in the distribution of summary statistics elicited from experts will, in most cases, outweigh the error by the approximation. A formal definition of the SIP is given below, however, this rigorous definition is only used to establish notation and motivate our variational approximation. Readers more interested in the application of this method may skim this section and the remainder of the work should still be comprehensible.

4.3.1 Construction

Let X be a *statistical model*, i.e., a set of potential observations and an indexed set of distributions over this set; the set of indices of the distributions is Θ [149]. The prior distribution is a distribution over the elements of Θ . Let S be another statistical model, with Λ as its index set. The statistical model, X , represents all potential observations, and the statistical model, S , represents the summary statistics of the observations in X .

The models X and S are connected by a function, $V : X \rightarrow S$, representing the process of summarising an observation from X into an observation in S . We assume the summary loses information, i.e., the function V is not injective (it is many-to-one). The function V induces a (possibly empty) binary relation between Θ and Λ which we denote by $\equiv$. Parameters, $\theta \in \Theta$ and $\lambda \in \Lambda$ are equivalent, $\theta \equiv \lambda$, when $P_\theta(V^{-1}[e]) = P_\lambda(e)$ for all events e in the σ -algebra over outcomes in S where P_x is the measure indexed by x and $V^{-1}[e]$ denotes the preimage of e under the summary function. This equivalence means that the distributions indexed by

θ and λ agree on the probability of events up to the resolution of the summary statistic.

We want a prior distribution which accurately represents what is known about the summary statistic in S . If two distributions over X (points in Θ) lead to the same distribution of summary statistics, then if we only know about the summary statistics we have no reason to favour one over the other as more likely. This is an appeal to both the principle of insufficient reason (if multiple things have the same summary we have no reason to favour one over the rest) and invariance (under a deterministic map the distribution should be preserved where possible).

If we knew the distribution of the summary statistic exactly, this would correspond to a specific $\lambda \in \Lambda$ and the SIP would be a uniform distribution over any θ equivalent to λ in the sense defined above. When the distribution of the summary statistic is not known exactly, but instead has its own distribution (a distribution on Λ), the SIP is a compound distribution. There is no guarantee that the SIP will exist, but when it does, it can be computed by integrating the uniform distributions on Θ over the distribution on Λ .

As a final remark regarding this, it should be noted that there is nothing in this definition precluding V from mapping components of X directly to S and including an element of Θ in X . Consider for example $\Theta = \mathbb{R}^n$ and $\Lambda = \mathbb{R}^2$. If for $\theta = (\theta_1, \dots, \theta_n)$ one possessed strong beliefs about θ_1 , it would be reasonable to consider $\theta \mapsto (\theta_1, f(\theta_2, \dots, \theta_n))$ for some observation function f . This means that, if you have prior knowledge about both parameters and summary statistics both of these can be incorporated into a prior distribution using this method.

4.3.2 Approximation

For most applications, it will not be possible to determine whether such a distribution even exists. However, it is straight forward to find an approximation to the SIP which always exists and often will be useful. The retrospective forecasting example below makes use of such an approximation. However, before we describe this approximation we begin by outlining the general approach to its construction.

Since prior distributions are distributions over Θ we begin with a parametric family, $\mathcal{F}$, on Θ , of the statistical model X . The approximation is the member of $\mathcal{F}$ which best approximates the SIP. Each prior distribution, F in $\mathcal{F}$ induces a distribution over X . This, in turn, induces a distribution on S through the summary function V . We denote this resulting distribution on S by F^* .

Taking a more constructivist approach, we can sample from F^* with the following steps: (1) draw a sample $\theta \sim F_\Theta$, (2) since θ determines a distribution over X (by definition) we can then sample $x \sim F_{X|\Theta=\theta}$, finally (3) we can then use V to get $V(x) \in S$. This defines a distribution over the image of X in S under V which we can extend to F^* over all of S by setting the density to zero outside of the image of X .

If prior knowledge about the summary statistic suggests it follows a distribution G (over S), we approximate the SIP by

$$\hat{F} = \arg \min_{F \in \mathcal{F}} d(F^*, G) \quad (4.1)$$

for some statistical distance d . The choice of the family $\mathcal{F}$ and the statistical distance d are addressed in the example below.

For the more computationally oriented reader, pseudocode for the evaluation of the statistical distance is given by Algorithm 1.

Algorithm 1: Pseudocode to estimate the statistical distance between a candidate distribution and the target distribution based on a generalised method-of-moments.

```
def statistical_distance(candidate_distribution):
    # Construct a list of summary samples
    summary_samples = []
    for _ in range(num_samples):
        theta = sample_theta(candidate_distribution)
        x = sample_x(theta)
        summary = compute_summary(x)
        summary_samples.append(summary)
    # Calculate the absolute error
    sample_mean = mean(summary_samples)
    sample_std = std(summary_samples)
    return absolute_error(sample_mean, sample_std,
                          target_mean, target_std)
```

4.4 Retrospective forecasting example

Populations in temperate locations regularly have epidemics of influenza during winter months which place a substantial burden healthcare systems. Subsequently, forecasting these epidemics has received global attention [28, 108]. We performed a retrospective forecast of a seasonal influenza epidemic to demonstrate the utility of the SIP and to compare it to a couple of alternatives. Our target for the retrospective forecasts was case counts of influenza during the 2017 epidemic in metropolitan Sydney, an Australian city with a population of approximately five million people.

The count data was provided by the New South Wales Public Health Network and Communicable Diseases Branch, Health Protection NSW; the counts are presence-only for confirmed cases of influenza. Based on a study using data from the public health pathology services of three other Australian states it is likely the data represent only a small sample of the true *incidence* (i.e., the total number infected each week) [154]. The use of this data was approved by The Health Sciences Human Ethics Sub-Committee, Office for Research Ethics and Integrity, The University of Melbourne (Application ID 1646516.3).

We modelled the transmission of influenza using a system of ordinary differential equations: the SIR model [37]. The SIR model describes the proportion of the population that are: susceptible to infection ($s(t)$), infectious ($i(t)$), or recovered ($r(t)$) and subsequently no longer infectious or capable of being infected at time t . We parameterise this system in terms of the basic reproduction number, $\mathcal{R}_0$, the recovery rate, γ , and the seed time, t_{seed} , as shown in Equation (4.2). An *observation model* is used to describe the relationship between the observed number of influenza cases and the incidence as calculated through the SIR model. The number of observed cases is modelled with a negative binomial distribution where the expected number of observed cases in a window of time is proportional to the incidence in the SIR model over that window as shown in Equation (4.4). Together, the SIR transmission model and the observation model constitute a hidden Markov model, with the observed cases conditional upon the *unobserved* transmission of influenza (see the Supplementary Materials for details of the hidden Markov model [105]).

We now outline the steps involved in constructing the SIP over the parameters of the SIR

model: $\mathcal{R}_0, \gamma, t_{\text{seed}}$. Steps 1 and 2 are already part of the modelling process, we describe them here to emphasise their role in the construction of the SIP. The selection of the parametric family, $\mathcal{F}$, and the statistical distance, d , are addressed in Steps 3 and 4. The latter two steps are prime candidates for automation.

Step 1: Choose the summary statistics. Key summary statistics of an influenza epidemic are the magnitude and timing of the peak caseload, i.e., the maximum number of cases seen in a single week, and the week in which this occurs [155]. Subsequently, we chose the expected timing and magnitude of peak incidence as the summary statistics, i.e., the values in S are pairs consisting of the week in which cases are expected to peak and the size of the peak.

Step 2: Elicit a distribution over the summary statistics. In lieu of eliciting a distribution for the peak statistics from an expert, we constructed a statistical model for their distribution based on historical Sydney influenza data from 2010–2016, (see the Supplementary Material for the details). This model (erroneously) predicted that in 2017, the epidemic in Sydney would peak during week number 33 ± 7.8 with 6100 ± 2800 cases where the bounds represent two standard deviations. The data upon which these estimates are made and the model are shown in Figures 4.9 and 4.10. We will discuss the error in these initial predictions later.

Step 3: Choose a family of distributions. The family from which we select the prior distribution, $\mathcal{F}$, is multivariate normal distributions. The hyper-parameters of this family are the mean vector of the multivariate normal distribution and a symmetric tri-diagonal matrix, A , which specifies the covariance matrix through $\Sigma = AA^T$. Plausible ranges for the parameters of the SIR model are given in Table 4.1. Since the multivariate normal distribution has $\mathbb{R}^3$ as its support, we transformed the parameters using the logit function. Since this change of variable is applied separately to each dimension, the Jacobian is tractable.

Step 4: Selecting the optimal family member. The optimal family member is selected using Equation (4.1). The multivariate normal induces a distribution on the parameters of the SIR model ($\mathcal{R}_0, \gamma, t_{\text{seed}}$) and subsequently a distribution on the summary statistics chosen in **Step 1**. A generalised method-of-moments approach is used to estimate the hyper-parameters of $\mathcal{F}$. In the notation above, the statistical distance $d(f, g)$, is the sum of the absolute difference in the mean and standard deviation of the distributions f and g . In this example, these distributions are the result of the predictive model in **Step 2** and the F^* induced on the peak statistics. The calculation of the mean and standard deviation involves an intractable integral, so we used Markov Chain Monte-Carlo (MCMC) to estimate them.

Steps 1–4 lead to an approximation which is easy to sample from and evaluate the density of, making it convenient to use. The prior distribution allows for correlations between parameters, an important property rarely seen in the literature where it is more common to specify an independent prior on each component of the parameter vector [109]. The computation of the mean and standard deviation in Step 4 is intractable (since it involves the solution to the SIR model), so we used an MCMC method to estimate it. Subsequently, the optimisation in this step is a stochastic optimisation problem. We solved this optimisation with an evolutionary strategy after experimentation with deterministic optimisation algorithms failed (data not shown). Since the objective function is bounded below, we can gauge the quality of the minimum found which provides a helpful diagnostic of performance.

To demonstrate the impact of the SIP on forecasting we generated retrospective forecasts of the 2017 influenza epidemic in Sydney. As a point of comparison, we also generated forecasts using more typical prior distributions: a weakly informative prior from our previous work [2], and an uninformative reference prior [77]. The specifications of the alternative prior distributions are also provided in Table 4.1. To see how the forecasts performed with

Table 4.1: The details of the three prior distributions used for the retrospective forecasting. The notation $\text{SIP}(a, b)$ is used to indicate that the support of the parameter is the interval (a, b) . Under the SIP the elements of the parameter vector have an *a priori* correlation. Unlike the SIP, the literature and reference priors have independent components (U and Exp denote the uniform and exponential distribution respectively). The “Literature” based prior is taken from our previous work, [2] and the “Reference” prior is a uniform distribution over a large range of plausible values.

Prior	SIR parameter	Distribution
SIP	$\mathcal{R}_0$	$\text{SIP}(1, 4)$
	γ	$\text{SIP}(1/10, 7)$
	t_{seed}	$\text{SIP}(14, 40)$
Literature	$\mathcal{R}_0$	$\sim \text{U}[1, 2]$
	γ	$\sim \text{U}[1/2, 3]^{-1}$
	t_{seed}	$\sim \text{Exp}(7/36) + 14$
Reference	$\mathcal{R}_0$	$\sim \text{U}[1, 2]$
	γ	$\sim \text{U}[0, 7]$
	t_{seed}	$\sim \text{U}[0, 52]$

different amounts of data we generated them using 0 and 20 weeks worth of data from the start of the time series. Consistent with published approaches to forecasting influenza epidemics in Australian cities [2], we assumed a known observation process in which 1 in 20 influenza infections are observed, with a background rate of 100 cases per week. The forecasts were generated by sampling from the posterior distribution (which when using 0 weeks worth of data is still the prior distribution).

The posterior samples were obtained using the Metropolis-Hastings algorithm with a Gaussian kernel. A total of 10^5 samples were drawn in each of 4 chains with the first half of these being discarded as burn-in. Visual inspection of the chains suggested they had converged and the $\hat{R}$ -statistic was ≤ 1.01 for each of the three forecasts.

4.5 Results

Evidence that the SIP improves the predictive prior distribution

The predictive model used to construct the prior does a poor job of predicting the peak in 2017, when there was an anomalously large number of cases. Subsequently, the peak is at the edge of the predictive credible interval (Figure 4.10). Figure 4.1 shows trajectories sampled from the SIP and the range of values proposed by the predictive model based on the previous epidemics.

There are a couple of points to emphasise here. The first is that the trajectories shown in the figures show a sample of the *expected* value of the incidence. This expected value then has an additional level of uncertainty represented by the observation model¹. The second is that the SIP does not preclude peak values outside of the two sigma range, it just attributes little prior probability to their occurrence. So, while the majority of the sampled trajectories peak within the proposed range, if the data suggests the epidemic will peak outside of this range, the posterior is still capable of this, (provided the *model* was capable of such a peak, clearly, it cannot exceed the population size). For comparison, we show analogous samples from the weakly informative and reference prior in Figure 4.2.

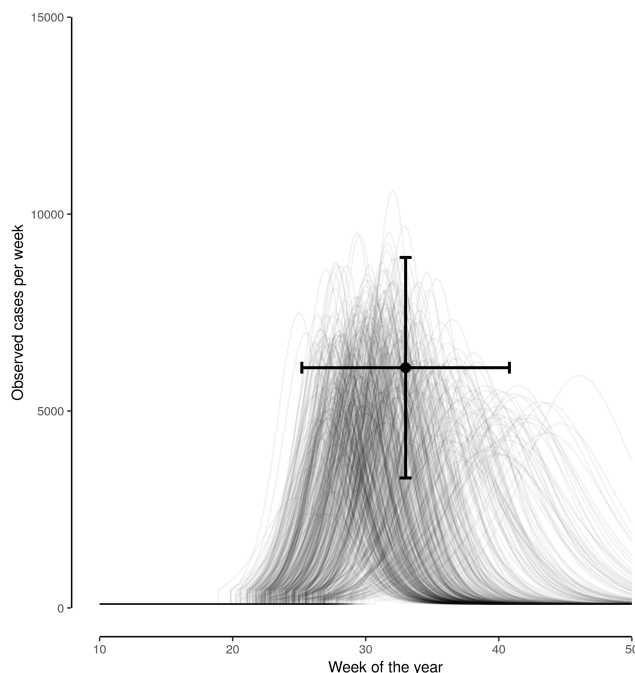
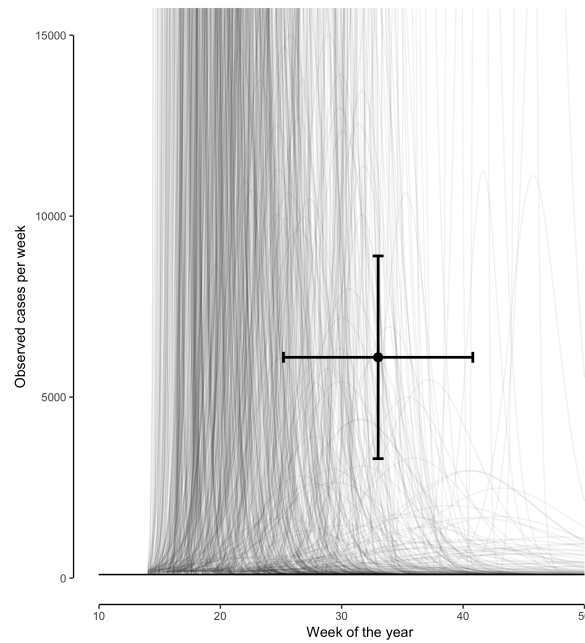
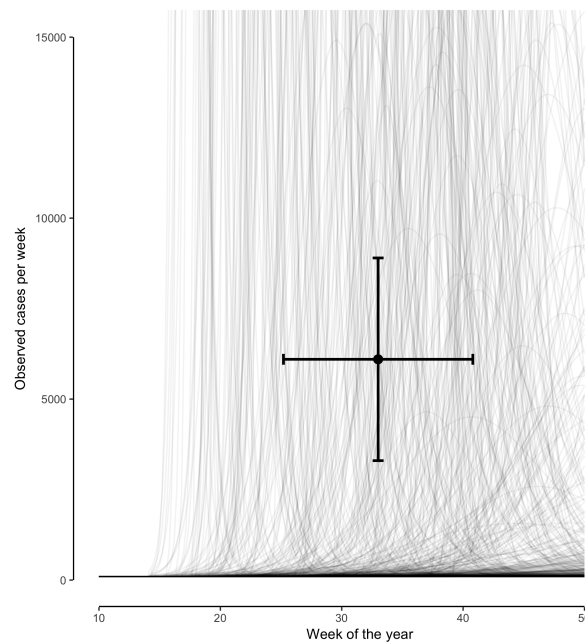


Figure 4.1: Trajectories from the SIR model, shown as a spaghetti plot. The parameters have a somewhat informative prior distribution and agree with the prior beliefs. The cross-hairs show the mean and two standard deviations of the distribution corresponding to those beliefs which are based upon a predictive model using data from the Sydney influenza epidemics 2010–2016.

¹See the Introduction for discussion regarding the distinction between transmission models and observation models.



(a) Weakly informative (literature based) prior



(b) Reference prior

Figure 4.2: Distribution of solutions to the SIR model under parameters sampled from the literature and reference priors. Both distributions attribute non-negligible prior probability to outcomes which are clearly unreasonable. From 2010 to 2016 the number of cases had never exceeded 3000 in a single week. The cross hairs have been included to ease comparison to Figure 4.1.

Evidence that the SIP satisfy the desired properties

Table 4.2 shows the mean and two standard deviations of the peak summary statistics under two distributions: the distribution predicted by the statistical model from Step 2, and the SIP used to approximate this distribution. Samples of peak statistics drawn from the SIP are shown in Figure 4.3, demonstrating that, as desired, there is little correlation between the summary statistics under the SIP (the Pearson correlation coefficient is -0.03).

Peak	Prior	SIP
Magnitude	6100 ± 2800	6079 ± 2398
Week	33 ± 7.8	31 ± 7.6

Table 4.2: The approximation to the SIP — obtained via a generalised method of moments — provides a good fit for the properties of the desired distribution on the peak summary statistics: the week with the most cases, and the maximum number of cases per week.

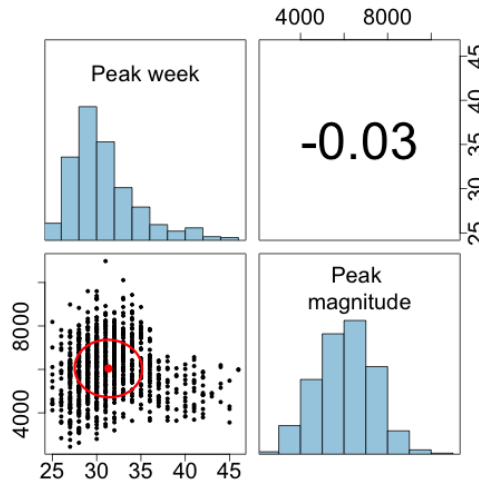


Figure 4.3: Samples of the peak statistics, (the timing and magnitude of the peak,) appear uncorrelated under the SIP (the Pearson correlation coefficient is -0.03). Histograms demonstrate that the marginal distributions of the statistics are uni modal and vaguely Gaussian. Correlation ellipses shown (in red) are a one-sigma region of a bivariate normal distribution fit to the samples with the mode indicated by the dot in the centre.

To demonstrate how the SIP and the other two distributions actually describe the parameters a panel plot of samples from each of the priors is shown in Figures 4.4 and 4.5. In Figure 4.4 the SIP can be seen to concentrate the probability mass for $\mathcal{R}_0$ in a narrow range and leads to a substantial dependence between $\mathcal{R}_0$ and the recovery rate γ . Under the SIP there are strong correlations between all three of the parameters. Figure 4.5 shows how in stark contrast to the SIP, under the literature based and reference priors the parameters are independent (by construction) and the prior mass is distributed over a much wider range of values.

The effect of the SIP on the posterior distribution

Figures 4.6 and 4.7 show trajectories from the posterior distribution after conditioning on the first 20 elements of the time series starting from the SIP, literature informed prior and

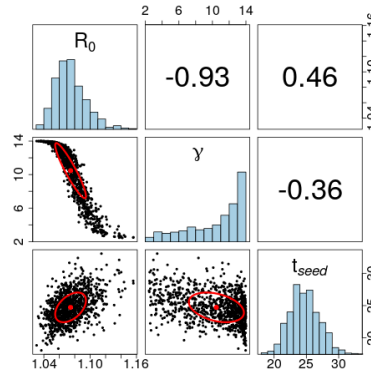


Figure 4.4: Histograms and scatter plots of samples from the somewhat informative prior (SIP). The numbers above the diagonal indicate the Pearson correlation under the prior distribution due to the constraint on the distribution of the peak time and magnitude. Correlation ellipses show (in red) are a one-sigma region of a bivariate normal distribution fit to the samples with the mode indicated by the dot in the centre.

reference prior respectively. In each case, the posterior distribution has converged, leading to a tight cluster of trajectories. The forecasts accurately fit the week in which cases peaked, but tend to underestimate peak magnitude. The level of underestimation of the peak is largest for the forecasts generated using the SIP where the prior still appears to have a non-trivial effect on the posterior despite the substantial amount of data. This is the cost of an informative prior when predicting an anomalous epidemic; our predictive model of the summary statistics ascribed very little likelihood to such a large peak and subsequently, it was given little prior probability.

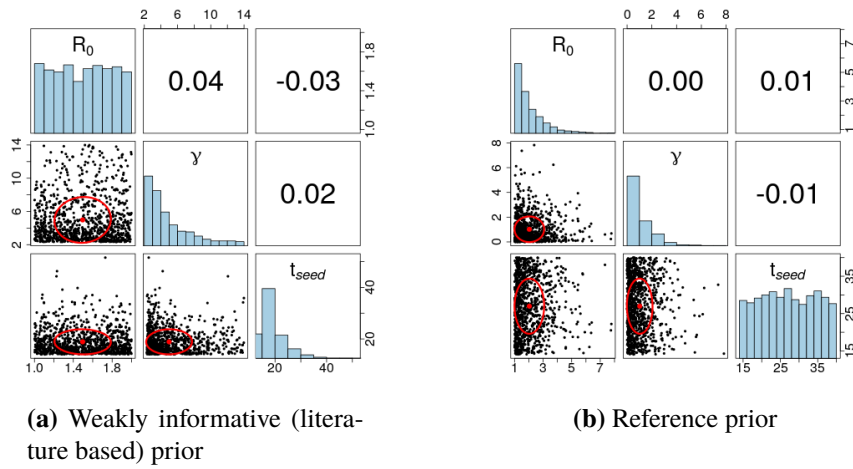


Figure 4.5: Histograms and scatter plots of samples from the literature informed and reference priors. Under these prior distributions the parameters are independent. The additional features of these figures are the same as described in Figure 4.4. Note: the axes vary between the three scatter plot matrices.

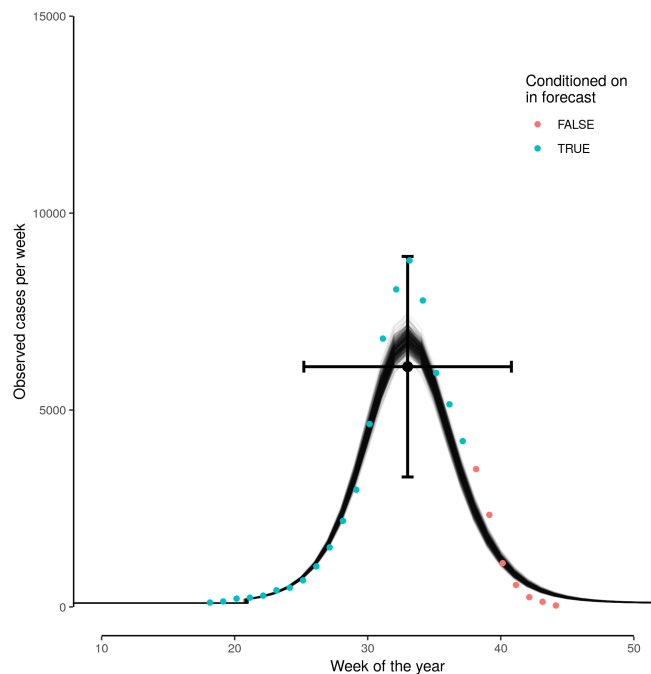
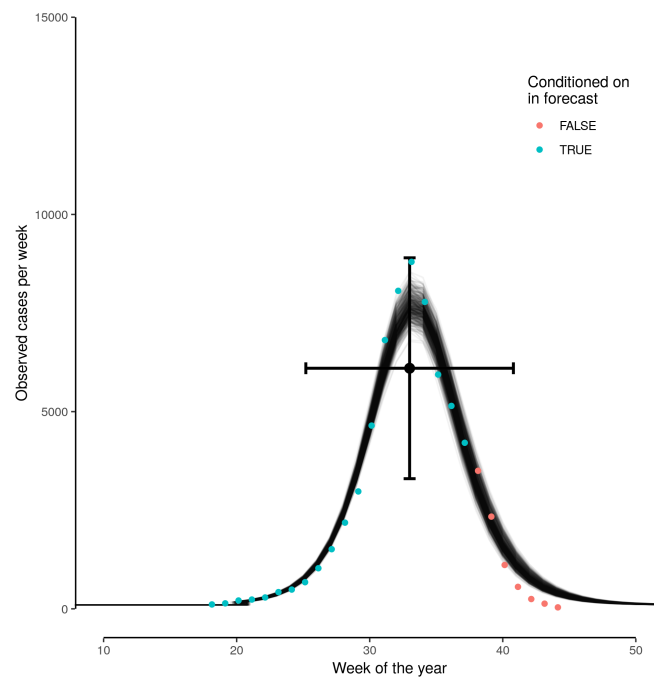
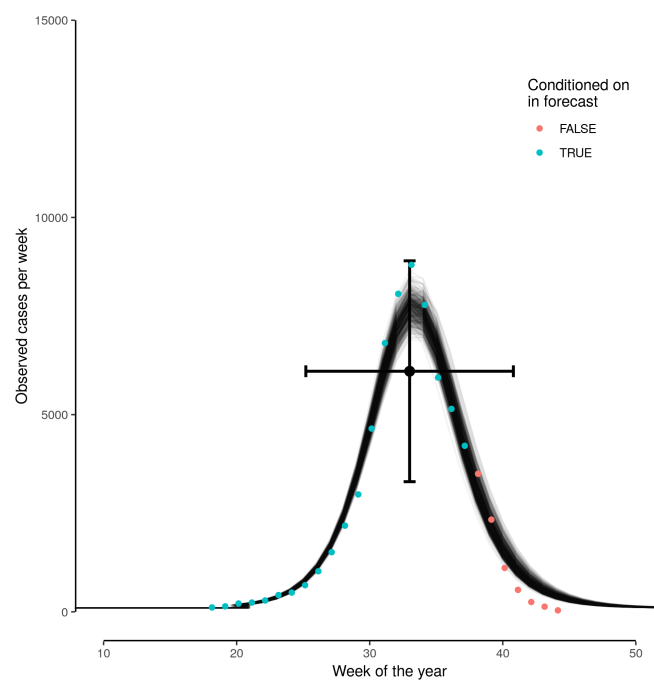


Figure 4.6: Forecasts of the 2017 influenza epidemic in Sydney, Australia, using the somewhat informative prior. The first 20 observations of case counts from the epidemic were used to inform this forecast. The cross-hairs indicate the prior distribution of the epidemic peak. The posterior distribution of the epidemic curve underestimates the four observations around the peak, but otherwise matches the observed and unobserved data well.



(a) Weakly informative (literature based) prior



(b) Reference prior

Figure 4.7: Forecasts of the 2017 influenza epidemic in Sydney, Australia, using a.) the literature informed prior and b.) the reference prior. The first 20 observations of case counts from the epidemic were used to inform this forecast. The cross-hairs indicate the prior distribution of the epidemic peak. In each case, the posterior distribution of the epidemic curve matches the observed and unobserved data well.

4.6 Conclusion

4.6.1 Findings

In a retrospective epidemic forecast, we used a somewhat informative prior (SIP) to improve early season predictions of seasonal influenza. An algorithm is described for constructing a SIP which makes it easier to utilise domain knowledge and can improve the performance of epidemic forecasts.

Leveraging existing knowledge from multiple sources can be very useful for forecasting [156]; Bayesian statistics and informative priors provide natural way to accomplish this. Often historical data can suggest a distribution of plausible behaviour for the system being modelling, in our seasonal influenza example this is a distribution for the time and magnitude of the epidemic peak. However, incorporating this into a prior distribution can be difficult; the SIP simplifies the process of incorporating this knowledge into a Bayesian analysis. In particular, it has a *prior predictive distribution* which matches any existing knowledge about summary statistics of the model.

This work has two primary contributions: methodologically, it extends the history matching method, enabling it to handle domain knowledge in the form of a full distribution for summary statistic, and in a more immediate way, it improves on an existing epidemic forecasting system, a challenge of international interest [108, 157, 158].

In terms of the contribution to Bayesian methodology, the algorithm we described is another step towards systematising the art of crafting a prior distribution [153]. While most prior distributions are specified as joint distributions with independent elements, our construction incorporates correlations between the parameters. These correlations make it the SIP suitable for complex models where interactions between the parameters should not be ignored. The construction of a SIP, the translation of prior knowledge to prior distribution, can be expressed as a stochastic optimisation problem, which can be solved with a simple stochastic optimisation method (in our case a genetic algorithm).

When forecasting epidemics, there is often very little contemporary data available to inform the forecasts, moreover, noise in the available data is often substantial [106]. This makes it difficult for mechanistic models to forecast epidemics. However, despite this, forecasts based on human judgement can do surprisingly well at forecasting epidemics, suggesting there is important information in expert knowledge which could be utilised [111]. This situation suggests there is potential to improve the performance of mathematical models (in a Bayesian context) through the use of subjective approaches. Information from previous epidemics (encoded in a SIP) allowed us to substantially reduce uncertainty in the forecasts generated using data from only the initial stages of the epidemic. The SIP is more closely aligned with our understanding of the data than the weakly informative prior taken from the literature, or the reference prior considered. There was a cost in making use of this prior information though. Since 2017 had an unusually high case count at its peak the prior distribution attributed low probability to such high peaks and this biased the forecasts. While one would not expect this to be the typical outcome — not all years can be anomalous — there is a fine balance between utilising prior knowledge and ignoring new evidence.

With respect to our forecasts, shown in Figures 4.6 and 4.7, there is a substantial portion of the time series from time before substantial epidemic activity. This may have a disproportionately strong influence on the posterior. To elaborate, since there are more elements of the time series at relatively low levels (i.e., early in the epidemic) than there are around the peak, and since each element contributes equally to the likelihood, we may not be giving enough

weight to the portion of the epidemic that we actually care about: the peak. It is unclear from the present investigation whether pre-processing the data in some way, e.g., removing elements from the start of the time series, or down-weighting them, could be beneficial, however this could be an interesting line of further investigation.

4.6.2 Recommendations

We now briefly consider some of the recommendations we believe are sensible in light of our findings.

Methodology

The slew of libraries for MCMC has increased the accessibility of Bayesian analysis. To assist in the use of MCMC there are also many tools to diagnose poor performance and assess the quality of the algorithm's performance. Bundling these tools together is transitioning the art of MCMC into a black box allowing applied statisticians to spend more time thinking about their analysis and less about the methods they use to carry it out. In the same vein, we view our contribution as a way to simplify the construction of bespoke prior distributions. Automating the choice of prior distribution allows the practitioner to trade one decision about *method* for a decision about the *model*.

It is still important to perform a sensitivity analysis to establish the impact of the prior distribution on the inferences made. As we have seen above, if the prior knowledge is substantially biased, it can have a detrimental effect. If one was concerned about such biases, the elicited distribution for the summary statistics could be made more diffuse to reduce its impact. However, establishing a protocol for how one might do this systematically is beyond the scope of the current work.

We recommend that sensitivity analyses should try to consider using a prior distribution incorporating *all* available expert knowledge; this allows us to determine whether the data is actually providing significant novel information about the process under study. There are cases where this will not be feasible, for instance, when using the models in an exploratory way to explore potential behaviour of a system. Regardless, the process of determining whether the use of an informative prior is appropriate and how one might be constructed can still be useful.

Epidemic forecasting

Bayesian forecasting can benefit substantially from the judicious application of prior knowledge, as we saw in the forecasts generated at the start of the influenza epidemic in Sydney in 2017. The SIP produces forecasts which more accurately reflect existing knowledge when only data from the start of the epidemic is available. However, one should constantly be aware that informative priors can lead to biased inferences if there is a substantial mismatch between prior beliefs and observed data, as we saw in the forecasts from later in the season. This does not disqualify the SIP as a tool for epidemic forecasting, it just means that there is no free lunch. If you want to incorporate prior knowledge, the SIP is a tool to do so, however in doing so you need to be aware of the risks.

In the example above we used the distribution of summary statistics exactly as predicted by a statistical model. If one had less confidence in their prior predictions they could inflate the variance of this distribution so as to reduce its impact on the posterior. Presumably, this

would enable the forecasts to adapt better to anomalous epidemics such as the 2017 influenza season.

4.7 Discussion

Much ink has been used – and misused – writing about prior distributions, and sadly we are contributing to this. However, it would be remiss if we were to define a prior selection method without providing some interpretation of it. We claimed that what we are trying to achieve with the somewhat informative prior is allow the prior to influence the results, but to do so without significantly violating our prior beliefs. The interpretation of the SIP is simple, the predictive prior distribution on summary statistics should not violate existing beliefs about the plausible values of these summary statistics.

Our approach is similar to history matching where constraints on summary statistics have been used to inform parameter choice and prior distributions [151–153]. However, the SIP extends history matching by handling a full distribution of the summary statistic. Moreover, we have demonstrated how to obtain such a distribution from historical data in the context of epidemic forecasting, and how using a transformed multivariate normal distribution can be used to construct a prior distribution which accounts for correlation between model parameters.

The applied statistician can benefit from this in two ways. Obviously, there is the specification of a prior distribution which accurately captures existing domain knowledge. For large data sets, the impact of the prior distribution will likely be small, however many important analyses are still performed on small data sets, and in these cases, a poorly chosen prior distribution can invalidate inferences drawn [125]. Making it easier to choose an appropriate prior distribution frees up the analyst to spend time on additional diagnostics, such as sensitivity to that choice of prior. Second, and more subtly, there is benefit from the process of constructing the prior. Following the steps we described encourages a clear specification of the goals of the analysis. Choosing the summary statistics forces the analyst to consider what is the main output of the model. It also makes it easier to incorporate knowledge that may not directly pertain to the parameters of the model, since this can also be introduced through the summary.

A common issue in the application of Monte Carlo methods is the computational cost of evaluating the target density (typically a posterior distribution up to a normalisation constant). There is the initial optimisation to obtain the parameters of the SIP which may be expensive depending on the problem at hand, but after that evaluating the resulting prior distribution is fast, so it is suitable for use in an MCMC algorithm. Since the SIP is represented by a transformed multivariate normal distribution it has the same computational cost as more standard prior distributions, which is typically negligible relative to the cost of evaluating the likelihood function anyway.

We encourage the use of prior distributions which are consistent with existing knowledge. The use of such priors guards against drawing conclusions which are *a priori* indefensible but have been obtained due to a poor (and somewhat arbitrary) choice of prior distribution. Such priors allow us to make use of existing domain knowledge, potentially increasing the power of the resulting analysis [159, 160]. Finally, it encourages communication between domain experts and those carrying out the statistical analysis as the process of eliciting this knowledge can clarify subtleties of the data generating process [161]. By removing the step of manually selecting parameters for the prior, practitioners can spend more time thinking

about *what* they know about their data and its generative process and less time thinking about *how* to represent this knowledge.

4.8 Acknowledgements

I would like to acknowledge the insightful comments from Christopher Drovandi which had a significant impact on this work.

4.9 Supplementary Materials

Hidden Markov model

The SIR model describes a population in terms of which of its members are susceptible to, infectious with, or immune to an infectious disease. We consider a formulation of the SIR model as a system of ordinary differential equations. These equations describing the dynamics of the proportion of the population that is: susceptible, s , infectious with, i , and immune, r . Since these proportions are exhaustive, there is a conservation quantity in this system, $s + i + r = 1$. The dynamics are described in terms of the following non-linear differential equations:

$$\frac{ds}{dt} = -\frac{\mathcal{R}_0}{\gamma}si \quad \text{and} \quad \frac{di}{dt} = \frac{\mathcal{R}_0}{\gamma}si - \gamma i, \quad (4.2)$$

where $\mathcal{R}_0i/\gamma$ is the rate at which susceptible individuals are infected and γ is the rate at which infectious individuals recover from infection, (and cease to be infectious and gain immunity to re-infection).

Initially the whole population is susceptible to infection, which corresponds to the initial condition $(s(0), i(0)) = (1, 0)$. Since the incidence, $si\mathcal{R}_0/\gamma$, in this state is zero, this is an equilibrium of the system. To leave this initial state, a seeding event occurs at time t_{seed} when a proportion, δ , of the susceptible individuals become infectious. This discontinuity takes the system to the state $(s(t_{\text{seed}}), i(t_{\text{seed}})) = (1 - \delta, \delta)$. Obviously, a continuous solution to this system cannot be found, but for the purposes of the current work, the weak solution is sufficient. In fact, one could re-write this system with a different initial condition and obtain a continuous solution, but this would require a parameterisation of the system in terms of the initial condition. The current parameterisation in terms of the event time and rate parameters makes the system easier to reason about.

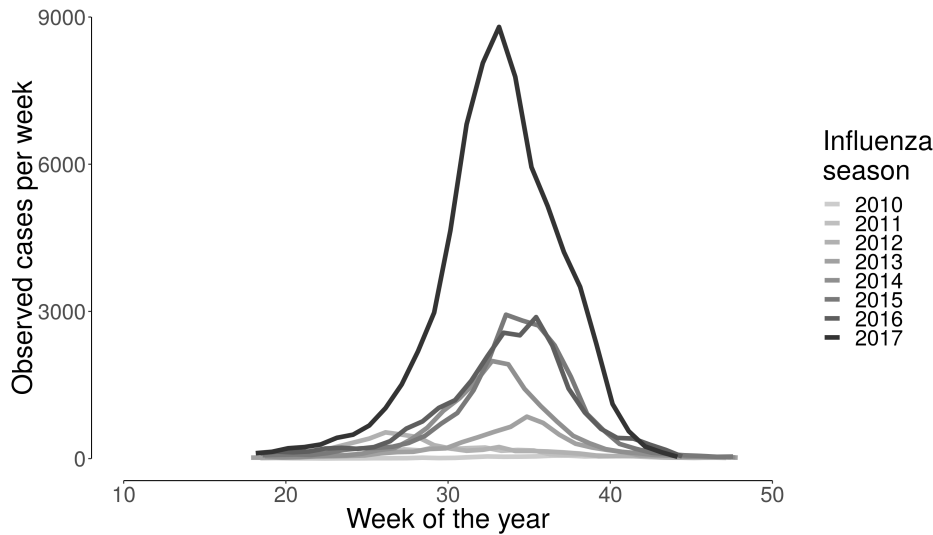
An important functional of this system is the proportional incidence, which, in the n th week is,

$$\mathcal{I}_n = \int_{\text{week } n} \frac{\mathcal{R}_0}{\gamma} s(t)i(t)dt. \quad (4.3)$$

In a population of N_{pop} individuals who will have an infection observed with probability p_{obs} , and with importations occurring at a rate of N_{import} per week, the expected number of cases to be observed in the n th week, μ_n is $N_{\text{pop}}p_{\text{obs}}\mathcal{I}_n + N_{\text{import}}$. The number of cases observed in the n th week, Y_n , is modelled as a random variable:

$$Y_n | \mu_n \sim \text{NegBinomial}(\mu_n, \phi), \quad (4.4)$$

Figure 4.8: Time series of the number of confirmed cases on influenza from 2010–2017 in Sydney, Australia, coloured by year. The magnitude of the number of cases observed each year is clearly increasing across this period. The peak in 2017 is approximately three times greater than the largest peak previously observed.



where ϕ is the dispersion parameter, set to 100. We assume conditional independence of the observations given the μ_n . The SIR model describes the transmission of the infection through the population and acts as the hidden process of our hidden Markov model and the observed number of infections is the observed data.

Sydney Exploratory Data Analysis

The time series from Sydney consists of the number of confirmed cases of influenza reported during each week from 2010–2017 as shown in Figure 4.8. The distribution of peak statistics we used, i.e., one we thought may have seemed plausible *before* the 2017 epidemic, was obtained from a simple analysis of the time series from the epidemics from 2010 to 2016. In practice, the specification of this distribution could draw on domain knowledge from a range of sources (including expert opinion). Since the peak timing and magnitude are frequently used as a way to judge the accuracy of an epidemic forecast, we choose to use them as the summary statistics for the SIP.

Figure 4.9 shows that there is little evidence of a trend in the timing of the peak week across the years. Subsequently, we summarise it by its mean and standard deviation: 33 and 3.9 respectively. As Figure 4.10 shows there evidence of a trend in the peak height across the years. Since biological count data is frequently over-dispersed with respect to the Poisson distribution we model the peak height with a quasi-Poisson generalised linear model, i.e., a generalised linear model with a log link function and a variance which is proportional to the mean. Simulation was used to establish that the predictive distribution for the peak height in 2017, under this model, has a mean of 6100 and a standard deviation of 1400. Of course, we now know that in 2017 the maximum number of cases in a single week would drastically surpass this, peaking at 8798, however, since we are doing a retrospective forecast we should not utilise this fact. Therefore, conditional upon the data that was available at the start of 2017, it is reasonable to believe the peak would occur in 33 ± 7.8 and that 6100 ± 2800 cases might be observed in that week.

Figure 4.9: The week of the year in which confirmed cases peaked is plotted for the 2010–2016 influenza epidemics in Sydney (see Figure 4.8 for the full time series). There is no significant trend in the peak week from 2010–2016. The solid lines show both a linear (blue) and constant (red) models fit to the data along with 95% confidence intervals. The similarity between the linear and constant models suggests we should model this as a constant.

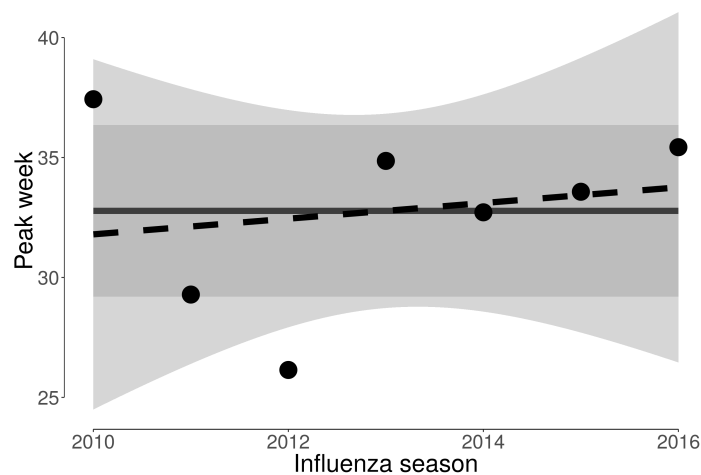
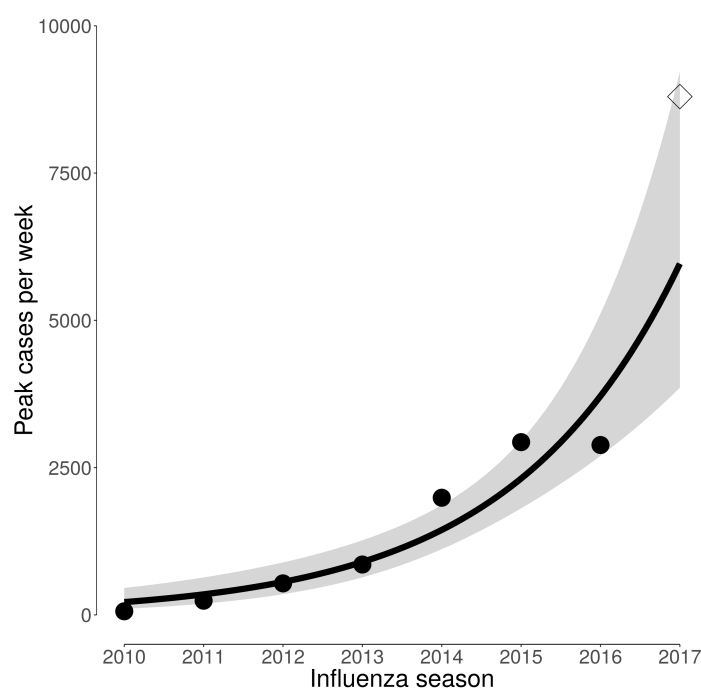


Figure 4.10: The maximum number of confirmed cases in a single week is plotted for the 2010–2016 influenza epidemics in Sydney (see Figure 4.8 for the full time series). There is a clear trend in the peak magnitude from 2010–2016. We modelled this using quasi-Poisson regression. The maximum number of confirmed cases in a single week for the 2017 epidemic is shown as a diamond at the upper edge of the predictive confidence interval.



Branching processes

5.1 Abstract

Exponential growth is a mathematically convenient model for the early stages of an outbreak of an infectious disease. However, for many pathogens (such as Ebola virus) the initial rate of transmission is sub-exponential, even before transmission is affected by depletion of susceptible individuals.

We present a stochastic multi-scale model capable of representing sub-exponential transmission: an in-homogeneous branching process extending the generalised growth model. To validate the model, we fit it to data from the Ebola epidemic in West Africa (2014–2016). We demonstrate how a branching process can be fit to both time series of confirmed cases and chains of infection derived from contact tracing. Our estimates of the parameters suggest transmission of Ebola virus was sub-exponential during this epidemic. Both the time series data and the chains of infections lead to consistent parameter estimates. Differences in the data sets meant consistent estimates were not a foregone conclusion.

We used a simulation study to investigate the properties of our methodology. In particular, to determine the extent to which the estimates obtained analysing time series data and those obtained analysing chains of infection data agree. We were able to use this simple branching process to handle data collected during contact tracing and to answer questions about the epidemiology of the disease, making it a useful tool for preliminary outbreak investigations.

5.2 Introduction

Physical systems can rarely support exponential growth for extended periods; during an epidemic, depletion of susceptible individuals leads to reduced transmission and, if intervention measures have not already done so, cause incidence to decline. Despite recent work showing the initial transmission of many diseases is sub-exponential, it is still common to see epidemics represented by models in which transmission grows exponentially [65]. This is concerning because exponential growth is extremely sensitive to its growth rate parameter, which can inflate the variance of forecasts. During an outbreak of a novel pathogen, uncertainty in the

growth rate is almost guaranteed.

The quantitative models used in epidemiology vary, from simple phenomenological models [162, 163] to complex agent-based simulations [164, 165].

Typically, the simpler phenomenological models lack the mechanistic underpinning to answer relevant questions, e.g., what will be the effect of vaccinating 20% of the population? It is difficult to quantify the impacts of such an intervention if there is no explicit representation of a susceptible population in the model. At the other end of the complexity spectrum, agent-based models are essentially black-boxes. It is conceptually simple to explore the impact of interventions in agent-based models, however, there is a cost to this simplicity. The complexity of agent-based models make them difficult to reason about mathematically, and since they are computationally intensive, even statistical analysis can be challenging.

In this chapter, it is demonstrated how a branching process can overcome some of the challenges described above: the mismatch between exponential growth of transmission and observations, and difficulty of finding a model with a mechanistic basis which is still mathematically tractable. A temporal in-homogeneity in the branching process ensures the generation sizes grow algebraically (in expectation), instead of the typical geometric/exponential growth. Branching processes can be viewed as either a tree, where it can describe who-infected-whom or as time series to describe the total number of cases through time, as such, they are a good example of a multi-scale model. Unlike many of the complex mechanistic models, the simplicity of the branching process means it possible to reason about them quantitatively and work with them computationally.

We explore the use and properties of this model from three perspectives. First, we use the branching process in a hierarchical model of transmission of Ebola virus in West Africa. Using data collected by the WHO we demonstrate the branching process can faithfully describe observed epidemics. Second, we fit the branching process to two different types of data. Fitting the process to chains of infection and time series of cases of Ebola virus disease (EVD) demonstrates the model provides broadly consistent parameter estimates using either data type, despite differences between the data sets. While the sub-exponential transmission of Ebola virus has been previously noted, [63], the branching process allows us to go further, supporting this claim through the interrogation of a new data set: a fully resolved infection tree inferred by Faye *et al* [1]. Third, to investigate the extent to which one might expect the previous result (i.e., obtaining similar estimates from each data type) to generalise, we performed a simulation study. The goal of this case study was not to investigate the utility of each data type for estimating the parameters per se, but whether, when both data types are derived from the same epidemic they produce concordant estimates. Applying our method to simulations from a Reed-Frost model we can compare the estimates from each data type for many different realisations and examine the degree to which consistency between the estimates is a general property.

5.3 Model

Below we derive the branching process in terms of a generic cumulative incidence function, i.e., a function describing the total number of cases that have occurred by a given time. We then consider the special case of a cumulative incidence function previously used to analyse time series of Ebola in West Africa [65]. Finally, we construct a likelihood function for this model, both in terms of a time series of cases and for observations of the number of secondary cases generated by individuals.

5.3.1 Construction

Let X_g^i denote the number of secondary infections due to individual i in generation g and Z_g the total number of infectious individuals in that generation, i.e., the sum of the X_{g-1}^i . We derive an in-homogeneous branching process where the *expected* generation sizes are $f_g = \mathbb{E}Z_g$.

Usually, the expected number of infectious individuals in a branching process grows exponentially/geometrically in the number of generations of transmission. For example, if $\mathbb{E}X = \mu$ then $\mathbb{E}Z_g = \mu^g$. The branching process derived below has expected generation sizes (i.e., the $\mathbb{E}Z_g$) which can follow any given monotonically increasing function. The notation used in this construction is summarised in Table 5.1.

Variable	Symbol	Variable type
Generation index	g	Constant
Generation times	Δ_g	Constant
Expected cumulative size	C	Constant
Expected generation size	f_g	Constant
Expected secondary infections	μ_g	Constant
Secondary infections	X_g	Random
Generation size	Z_g	Random
Growth rate	r	Parameter
Deceleration parameter	p	Parameter
Dispersion parameter	k	Parameter
Extinction	$\mathcal{E}_g$	Random event

Table 5.1: Notation used for the branching process.

Let $C(t)$ be the *expected* cumulative incidence by time t , i.e., the number of infections we would expect to occur by time t . Evaluated at multiples of the serial interval, C yields the generation sizes, f_g for $g = 1, 2, \dots$:

$$f_g = C(\Delta_g) - C(\Delta_{g-1})$$

where Δ_g is the time of the g th generation. The first value of this sequence is $f_0 = Z_0$, the number of infectious individuals in the first generation. Then, assuming the X_g^i are independent with mean $\mu_g = f_{g+1}/f_g$ we observe,

$$\begin{aligned}
 \mathbb{E}Z_g &= \mathbb{E} \left[\mathbb{E} \left[Z_g | Z_{g-1} \right] \right] \\
 &= \mathbb{E} \left[\mathbb{E} \left[\sum_i^{Z_{g-1}} X_{g-1}^i | Z_{g-1} \right] \right] \\
 &= \mathbb{E} \left[Z_{g-1} \mathbb{E} \left[X_{g-1}^i | Z_{g-1} \right] \right] \\
 &= \mathbb{E} \left[Z_{g-1} \right] \mu_{g-1}.
 \end{aligned}$$

The solution to this recurrence is

$$\mathbb{E}Z_g = Z_0 \prod_{i=0}^{g-1} \mu_i.$$

So $\mathbb{E}Z_g = f_g$ from the definition of μ_g .

In summary, by fixing the expected value of the offspring distribution (in terms of the generation) we obtain a branching process which, on average, has an expected cumulative incidence C . This means we can get the behaviour of a phenomenological model which is known to fit observations better than exponential/geometric growth, but still has a mechanistic foundation. i.e., since it explicitly represents the individuals in the population.

5.3.2 Cumulative incidence function

The construction above assumes a cumulative incidence function, C . We use the generalized growth model [65] defined by

$$\frac{dC}{dt} = rC^p \quad \text{which has the solution} \quad C(t) = \left(\frac{r}{m}t + A \right)^m$$

where $m = 1/(1 - p)$ and $A = Z_0^{1/m}$, with initial condition $C(0) = Z_0$. The growth rate, r , is as for standard exponential growth. The *generalisation* enters through the inclusion of the exponent p .

The parameter p is referred to as the *deceleration parameter*; it influences the dynamics of transmission. For $0 < p < 1$ the incidence interpolates through polynomials limiting to exponential growth as $p \rightarrow 1$. For $p < 1$ there is a diminishing increase in the force of infection with each additional infection. When $p = 0$ the force of infection is constant, for $p = 1/2$ (when $m = 2$) the incidence grows linearly (since the incidence is the derivative of the cumulative incidence by definition), $p = 2/3$ provides quadratic growth and with $p = 1$ we recover exponential growth in incidence.

Previous analysis suggests the spread of diseases, such as Measles, HIV/AIDS, and FMD, can be explained by values of $p < 1$; 0.51 (0.47, 0.55), 0.5 (0.47, 0.54), and 0.42 (0.27, 0.58) respectively [65].

5.3.3 Offspring distribution

Since epidemiological count data is frequently over-dispersed (with respect to the Poisson distribution) we use the negative binomial distribution for the offspring distribution. Over-dispersion in count data can occur for many reasons [75], for case counts in an epidemic, *superspreaders* can play an important role [166]. We parameterise the negative binomial in terms of its mean, μ , and a shape parameter, k , (a.k.a. the “dispersion parameter”). Under this parameterisation the variance, σ^2 , grows quadratically in μ :

$$\sigma^2 = \mu + \mu^2/k, \tag{5.1}$$

so as $k \rightarrow \infty$ we recover the Poisson distribution. Since the mean value is determined by the cumulative incidence function this choice of offspring distribution only introduces a single additional parameter, k .

5.3.4 Population view

Realisations from the branching process are naturally viewed as a tree, with the edges indicating who infected whom. However, as the notation suggests, this process can also be viewed as a sequence of generation sizes, $Z_{0:g}$. We refer to this representation of the process

as the *population view*. As we will see, the ability to represent a process as both a tree and a time series is very useful when making use of multiple data types.

Since the offspring distribution is negative binomial, the generation sizes are just sums of negative binomial random variables where the number of terms in the sum is the size of the previous generation. The sum of independent negative binomial random variables is still negative binomial; this can be seen from the moment generating function. Hence, the process is a Markov chain where each state is drawn from a negative binomial distribution conditional upon the size of the previous generation. We refer to this Markov chain as the population view of the process.

5.3.5 Conditional likelihood

Early work by Wald in the 1940's demonstrated the importance of survivorship bias. The importance of subtleties in the provenance of data, and how to account for this via conditioning is well understood in phylogenetics [167] yet does not appear to have permeated to the same degree into the epidemiology literature (notable exceptions being the work of Mercer [168] and Rida [169]).

Popular estimators of the basic reproduction number, R_0 , are biased towards *overestimation* in the early stages of an epidemic, [168]. We condition the process against extinction in the likelihood during fitting to mitigate this bias. This is because — by virtue of being observed — the outbreak must have avoided stochastic extinction [169].

Let $\mathcal{E}_g$ denote extinction of the pathogen by generation g , ($\mathcal{E}_g = \{Z_g = 0\}$) and $\mathcal{E}_g^c$ its complement, ($\mathcal{E}_g^c = \{Z_g > 0\}$). For data, $\mathcal{D}$, the likelihood function of the branching process described above (where extinction *can* occur) is $\mathcal{L}(\theta|\mathcal{D}) = \mathbb{P}(\mathcal{D}|\theta)$ with θ the parameter vector: $\theta = (r, p, k)$. When we condition the process against extinction we get the likelihood function for the conditioned branching process (CBP), $\mathcal{L}_{\text{CBP}}$:

$$\mathcal{L}_{\text{CBP}}(\theta|\mathcal{D}) = \mathbb{P}(\mathcal{D}|\theta, \mathcal{E}_g^c) = \frac{\mathcal{L}(\theta|\mathcal{D}, \mathcal{E}_g^c)}{\mathbb{P}(\mathcal{E}_g^c|\theta)}$$

The denominator is the probability the process is *not* extinct at generation g . When $\mathcal{D}$ is a time series we compute the extinction probability recursively using generating functions. All estimates in this paper which are based on times series make use of the CBP.

For secondary infections data, the probability of extinction conditional upon partial observations is prohibitively expensive to evaluate since it requires integrating over all the possible hidden infection trees. Subsequently, when working with secondary infections data we do not condition the process against extinction. Instead we treat each count of secondary infections as an independent sample from the offspring distribution.

5.4 Method

5.4.1 Data

Data of cases of EVD in Guinea, Liberia and Sierra Leone from 2014–2016 were obtained from the WHO [170]. We extracted confirmed cases from the patient data and then selected the longest stretch of consecutive weeks (the temporal resolution of the data) in which there was at least one confirmed case for each country. This process was repeated to generate a time series for each of the countries considered. The longest stretches occurred at the beginning

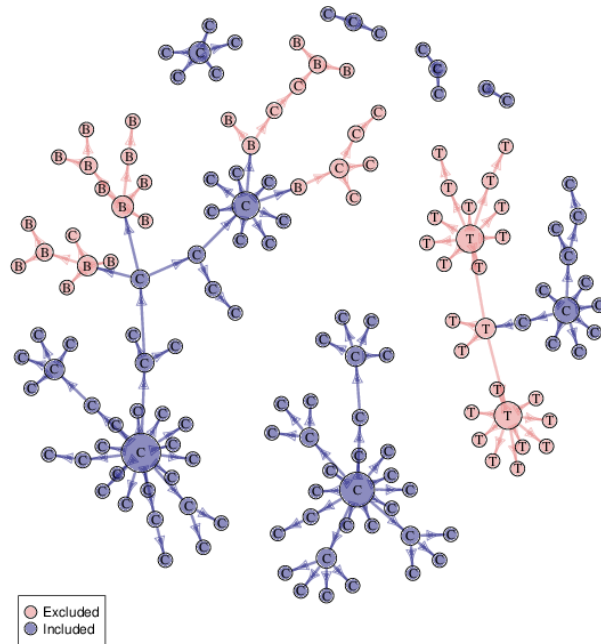


Figure 5.1: The infection tree from Faye et al [1]. The colour of the nodes indicates whether the data were included in the analysis and the labels indicate where the infection occurred.

of the epidemic for both Guinea and Sierra Leone, while several isolated cases were removed from the start of the Liberian time series. These time series were aggregated by fortnight as a proxy for generations of transmission since the Ebola virus has an approximate 14 day generation time [171]. The first 20% of cases were used in the analysis (as the $Z_{0:G}$) to represent transmission during the initial stage of the epidemic.

The WHO data also includes approximate locations for each case. Using this information, we extracted another time series specific to Conakry, the capital city of Guinea.

Faye *et al* [1] resolved an infection tree for cases from Conakry and the towns of Boffa and Telimele resulting in the data shown in Figure 5.1. Of the 193 confirmed and probable cases reported from these locations, 152 were placed in the tree with 106 of these from Conakry. To avoid the effects of re-importation we only used cases from Conakry that were not re-introductions from Boffa or Telimele, leaving 98 cases in the tree.

In the case of Conakry, it is important to note that there are important differences between the data sets. The time series is specific to confirmed cases where the tree contains both confirmed and probable cases. And the number of cases in the time series is far greater than the number in the infection tree.

5.4.2 Time series model

The confirmed cases of EVD in the three West African countries were modelled as time series of generation sizes using the population-level formulation of the branching process. We considered a hierarchical model in which the model parameters for each country come from a common prior distribution which is also estimated. The prior distributions used

for the parameters in this model are shown in Table 5.2. We computed the marginal prior distributions of the model parameters numerically to visually inspect the difference between the prior and posterior distributions.

The model was implemented in Stan and Hamiltonian Monte Carlo (HMC) was used to sample from the posterior distribution. Four HMC chains were run; the first 1000 samples of each chain were discarded as burn-in before a further 5000 samples were taken. Of the 5000, this was thinned by a factor of 5 to obtain the final 1000 samples for each chain. The chains appeared to have converged and mixed well: this was established via visual inspection and the $\hat{R}$ -statistic (< 1.01 for all variables). The effective sample size was appropriate given the dimensionality of the problem: for all variables in excess of 80% of the full number of iterations. Subsequently, the posterior samples likely provide a good representation of the posterior distribution.

Type	Parameter	Prior
Hyperparameter	α_p	Uniform(1, 5)
	β_p	Uniform(1, 5)
	μ_r	Normal(0, 1)
	μ_k	Normal(0, 1)
Parameter	p	Beta(α_p, β_p)
	r	Lognormal(μ_r, σ^2)
	k	Lognormal(μ_k, σ^2)
Constant	σ^2	1/6

Table 5.2: Prior distributions used for the model parameters in the hierarchical model described in Section 5.4.2.

5.4.3 Comparison of time series and chain of infection data from Conakry (Guinea)

As shown in Section 5.3.4, the branching process can be viewed at the individual or population scale. This prompts the question of whether data collected at each of these scales is equally informative about the parameters of the process, i.e., whether there is any advantage one over the other. We consider two data sets collected in Conakry (the capital city of Guinea) from the Ebola epidemic of 2014–2016: a time series of the number of confirmed cases each week (population scale data), and an infection tree describing who infected whom in a subset of cases (individual scale data). We fit the branching process to both data sets in order to determine whether they would lead to concordant parameter estimates. Note, similarity of the estimates was not guaranteed a priori, since while they are both observations of the same epidemic, the data sets consist of different cases. The time series has all the confirmed cases from Conakry, the infection tree contains only a subset of the confirmed cases but it also contains suspected cases which were excluded from the time series [1].

We used the population view of the branching process to model the time series of confirmed cases from Conakry. For the secondary infections tree from Conakry (described in Section 5.4.1), we modelled the number of secondary infections from each individual as an independent sample from the offspring distribution. This takes the form of pairs, (g, X_g^i) , one for each individual, where g is their infection generation (the node's depth in the tree) and X_g^i is their number of secondary infections (the out-degree of the node).

The prior distribution used is shown in Table 5.3. Fitting the model to each data set allows us to investigate whether these views of the same epidemic are consistent.

The models were implemented in Stan and Hamiltonian Monte Carlo (HMC) was used to sample from the posterior distribution. Four HMC chains were run; the first 10000 samples of each chain were discarded as burn-in before a further 10000 samples were taken. Of the 10000, this was thinned by a factor of 10 to obtain the final 1000 samples for each chain. The chains appeared to have converged and mixed well: this was established via visual inspection and the $\hat{R}$ -statistic (< 1.01 for all variables, with most < 1.001). The effective sample size was sufficiently large, in excess of 90% of the true sample size for all variables. Subsequently, the posterior samples likely provide a good representation of the posterior distribution.

Type	Parameter	Prior
Parameter	p	Beta(1.5, 1.5)
	r	Lognormal(1, σ^2)
	k	Lognormal(1, σ^2)
Constant	σ^2	1/6

Table 5.3: Prior distributions used for the model parameters in the comparison of the two data types from Conakry.

5.4.4 Simulation re-estimation

We carried out a simulation study to investigate whether estimates derived from time series and secondary infections data are concordant and how this depends on the number of secondary infections observed. The goal of this study is to determine the regularity with which the estimates agree, rather than the accuracy with which they capture the dynamics of the epidemic. We simulated a Reed-Frost (RF) epidemic model 1000 times, recording who-infected-whom in each generation [37], as described in Supplementary Materials. Note that the RF-model assumes a finite population while the branching process implicitly assumes an infinite population. Subsequently, in the RF-model the susceptible pool can be depleted during the epidemic – retarding transmission – eventually causing incidence to decline to zero. In addition to allowing us to investigate agreement between the estimates, fitting the branching process to realisations of the RF-model demonstrates how the model handles deviations from the assumptions used in its construction.

The models were implemented in Stan and L-BFGS was used to approximate the maximum a posteriori probability (MAP) for each of the simulations. Due to the large number of replications considered it was not feasible to check the output of each optimisation manually, instead it was left to the implementation of the optimisation algorithm to determine whether the computation had converged or whether a numerical issue had been encountered (in which case the simulation and optimisation were repeated).

5.5 Results

5.5.1 Hierarchical model fit

Figure 5.2 shows the fit of the hierarchical model to time series of confirmed cases of EVD from Guinea, Liberia and Sierra Leone. The credible intervals on the figures show the uncertainty in the *expected incidence*, i.e., the 50% and 95% credible intervals for $\mathbb{E}Z_g$.

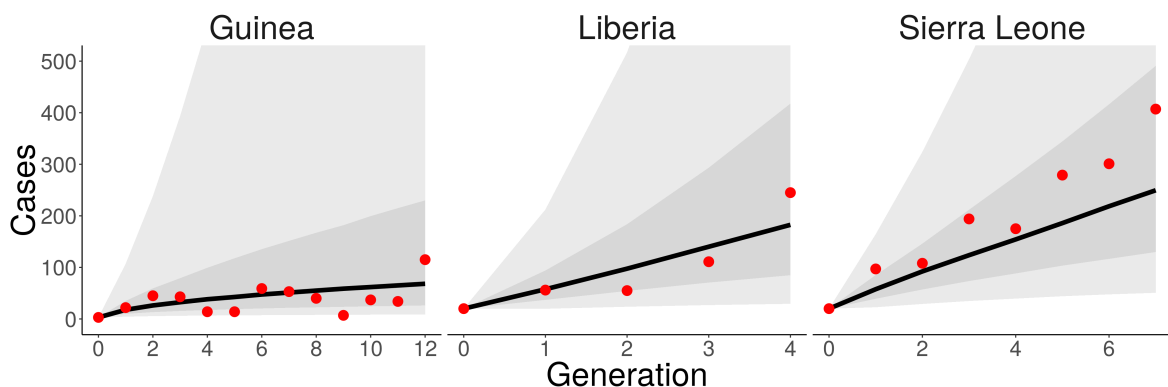


Figure 5.2: The branching process fit to time series of confirmed cases of EVD from Guinea, Liberia and Sierra Leone. The expected generation sizes (the model fit) are shown as a solid line with the 50% and 95% credible interval on this estimate shown as a grey ribbon. The observed case counts are shown as red points.

Figure 5.3 shows the marginal posterior distributions of the *logarithm* of the growth rate, the deceleration parameter and the *logarithm* of the dispersion parameter respectively. Figure 5.3b shows the posterior mass for p has accumulated around 0.5 for all three countries; in the model, this corresponds to approximately linear growth in the incidence. Another way to view this would be that the cumulative incidence had quadratic growth. Recall from Equation (5.1) that the variance scales with the inverse of the dispersion parameter. For each of the countries the dispersion parameter, k , has converged to small values indicating that the variance scales quickly with the mean incidence. This suggests stochasticity played an important role in the initial transmission in these countries.

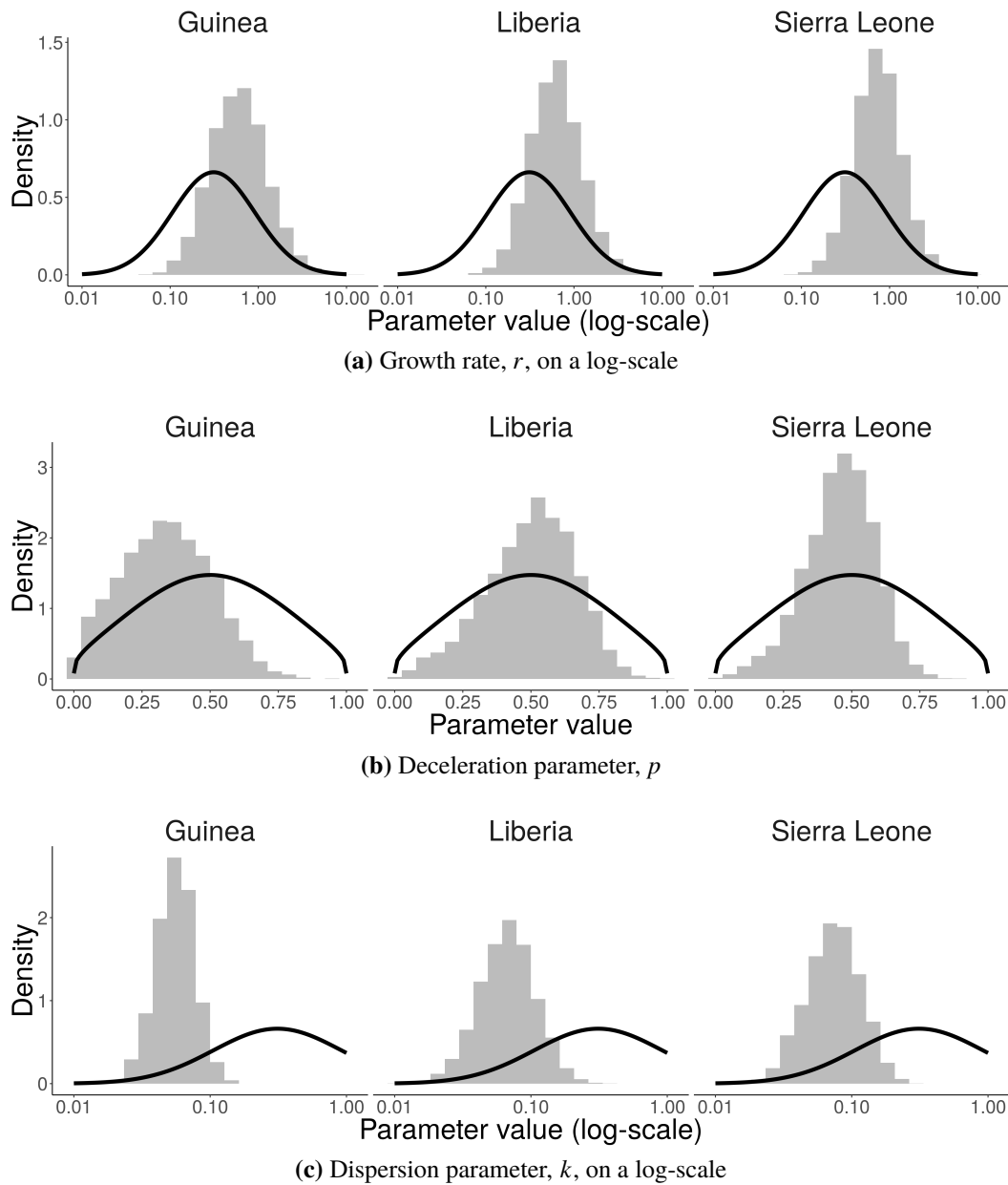


Figure 5.3: Histograms of posterior samples under the hierarchical model for a.) Guinea, b.) Liberia and c.) Sierra Leone. The marginal prior distribution is included as a solid line to assess convergence.

5.5.2 Comparison of time series and infection chain data from Conakry (Guinea)

Figure 5.4 shows the marginal posterior distributions for the growth rate, deceleration and dispersion respectively, conditioning upon the time series and secondary infections data from Conakry. The posterior distributions differ from their prior, indicating information was extracted from the data. The parameter estimates inferred from each data set are broadly consistent, suggesting, in this instance, that both data types provide a consistent representation of the dynamics. The time series data suggested a smaller growth rate (mean= 0.38, CI= 0.06 – 0.47) than the tree data (mean= 0.75, CI= 0.17 – 2.52). This trend is reversed for

the deceleration parameter, which are 0.30, CI= 0.04 – 0.67 for the time series and 0.13, CI= 0.01 – 0.36 for the tree data. Overall, the time series data suggests slower, but more rapidly accelerating growth than the secondary infections data. We consider potential causes for these differences below.

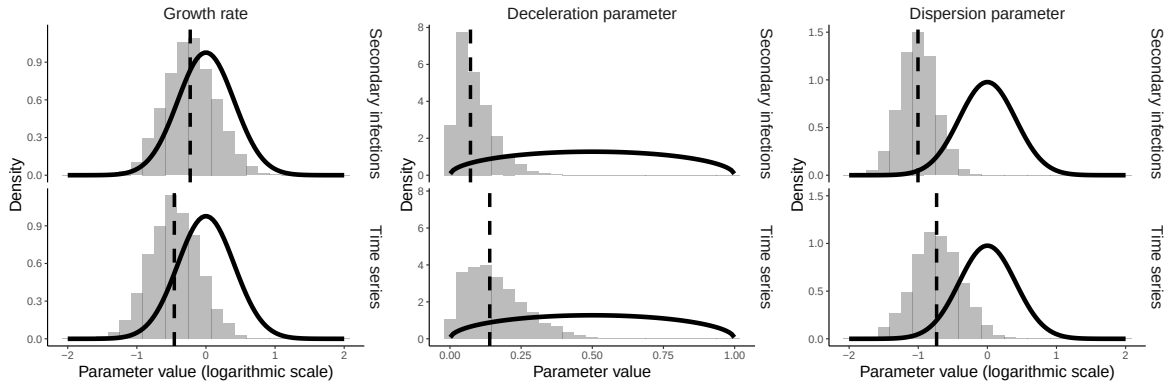


Figure 5.4: Histograms representing the posterior distribution of the model parameters conditional upon the secondary infections data and the time series data from Conakry. The solid lines show the prior distribution for each of the parameters (obtained via numerical integration). The growth rate and dispersion parameters are shown on a log-scale.

5.5.3 Simulation re-estimation

Figure 5.5 shows the simulations of the number of infectious individuals in each generation of the RF-model (Section 5.4.4). For most of these simulations, the incidence is still increasing during the first 7 generations suggesting the epidemic peak has not yet been reached for the majority of these simulated epidemics. The purpose of the simulation study is to demonstrate the degree to which both time series and secondary infections observations lead to consistent estimates of model parameters, and how this depends upon the number of secondary cases observed.

Figures 5.6, 5.7 and 5.8 shows the relationships between the maximum a posteriori probability (MAP) estimates of the growth rate and the deceleration parameter (respectively) obtained using either data type. In the case of secondary infections data the number of infectious people to “contact trace” is a tuning parameter: a property of the actual observation process. For this study we inspected the number of secondary infections at three levels of observation, i.e., we recorded the number of secondary infections from 2, 5 and 10 individuals in each generation.

Considering the MAP conditional upon each data type, there is a strong correlation between the estimates obtained with each data type for both the growth rate and the deceleration parameter, and this correlation grows stronger as more secondary infections are observed. In the case of the deceleration parameter, once ten individuals have had their secondary infections observed both data types lead to essentially the same estimates. There is a clear bias and increased variability in the estimates derived from the secondary infections data for both the growth rate and the dispersion parameter. As with any Bayesian analysis, it is important to understand the impact of the prior distribution; in the absence of any data, the MAP would be the mode of the prior distribution. In each case, there is a consistent shift in the MAP estimate away from the mode of the prior (as shown in the figures.)

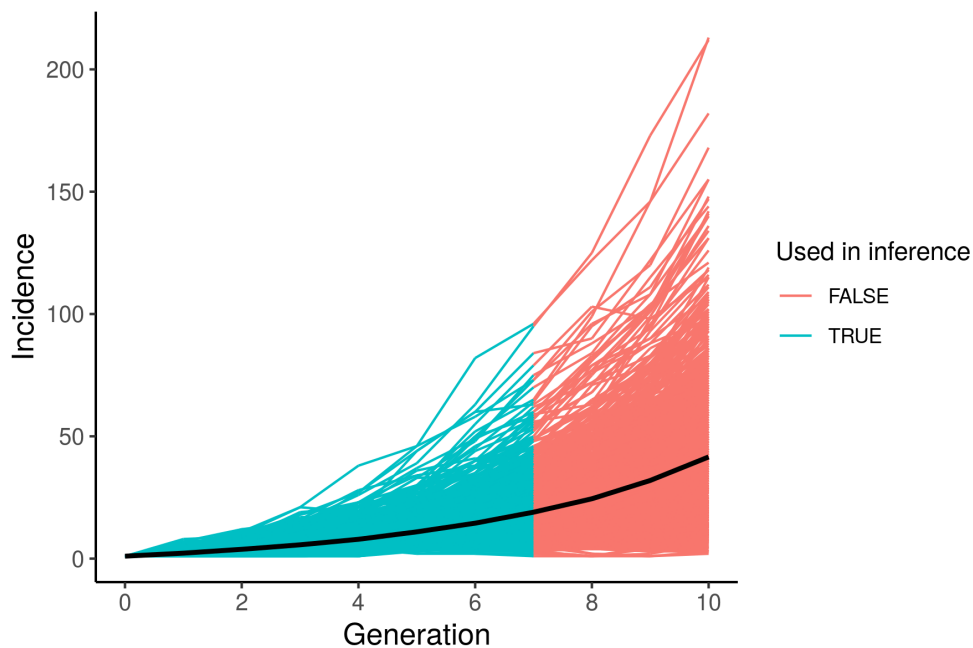


Figure 5.5: Simulated time series from the Reed-Frost epidemic model and the mean of these time series. The blue portion of the time series was used in the simulation re-estimation study.

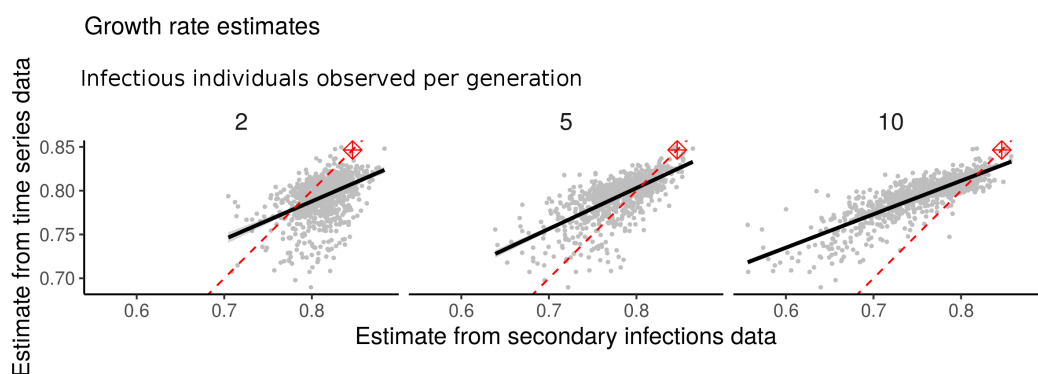


Figure 5.6: A scatter plot of the maximum posterior probability estimate of the growth rate obtained from the time series data and the secondary infections data. There is a single point for each simulation, and the solid line shows a linear fit with a 95% confidence interval, the dashed line shows the parity line. Each facet shows the estimate conditional upon a different number of observations in each generation: 2, 5, or 10.

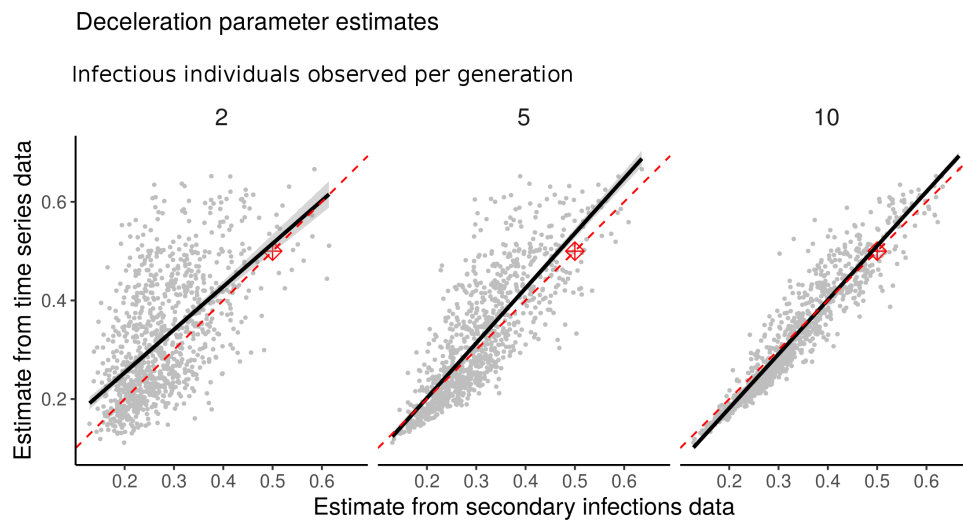


Figure 5.7: A scatter plot of the maximum posterior probability estimate of the deceleration parameter obtained from the time series data and the secondary infections data. There is a single point for each simulation, and the solid line shows a linear fit with a 95% confidence interval, the dashed line shows the parity line. Each facet shows the estimate conditional upon a different number of observations in each generation: 2, 5, or 10.

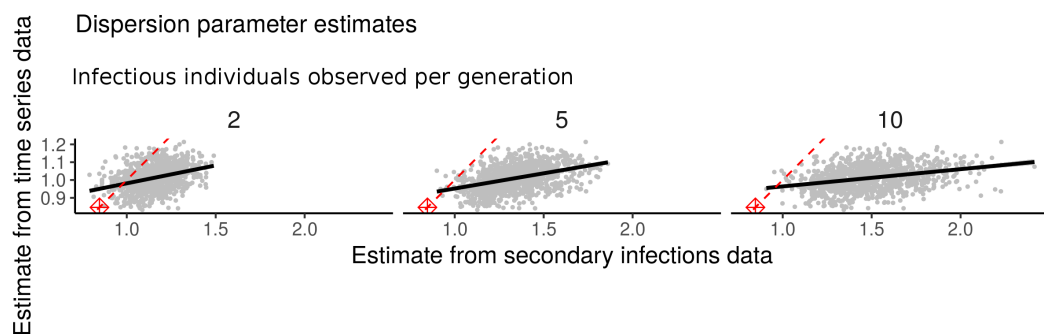


Figure 5.8: A scatter plot of the maximum posterior probability estimate of the dispersion parameter obtained from the time series data and the secondary infections data. There is a single point for each simulation, and the solid line shows a linear fit with a 95% confidence interval, the dashed line shows the parity line. Each facet shows the estimate conditional upon a different number of observations in each generation: 2, 5, or 10.

5.6 Conclusion

5.6.1 Hierarchical model fit

The analysis of the EVD time series from Guinea, Liberia and Sierra Leone demonstrates that the in-homogeneous branching process is capable of faithfully describing disease transmission at the population level. The posterior distribution of the deceleration parameter, which controls the scale of the growth, suggests that initially, the incidence grew approximately linearly (and the cumulative incidence quadratically). This differs from the results presented by Chowell *et al* [63], who observed that transmission at a sub-national level grew sub-exponential, but that at the national level it grew approximately exponentially. While it is tempting to attribute these differences to the differences in the modelling approach, the most likely explanation is the different pre-processing of the time series data. The previous analysis considered a portion of the time series from later in the epidemic, to mitigate the influence of stochastic effects. Since we are using a stochastic model, it is appropriate to make use of such data, despite the influence stochasticity plays.

5.6.2 Conakry time series and chains of infection

Using either time series data or secondary infections data from Conakry, Guinea led to similar parameter estimates demonstrating that either data set could be used to characterise transmission. The time series estimates have a smaller growth rate and a larger deceleration parameter than those from the secondary infections data. The difference in the estimates could be partially attributed to the estimates trading off faster growth (i.e., higher growth rate) for less acceleration of growth (i.e., smaller deceleration parameter.) Since this trade-off should yield similar dynamics over short time spans it is unclear whether this difference would pose substantial issues to interpretation of the parameters.

In the case of this Ebola epidemic, the time series data was available long before the infection tree. However, obtaining comprehensive time series of disease is challenging, and it is interesting to know that there are alternative data sets which are useful and already part of the data collected during intervention measures such as contact tracing. Moreover, this observation does not guarantee that we can rely on the agreement between the inference methods in general which is why we also carried out the simulation study.

5.6.3 Simulation re-estimation

The simulations from the Reed-Frost (RF) model (shown in Figure 5.5) emphasise the variability between realisations of stochastic epidemic models, and consequently, the substantial role stochasticity plays during outbreaks. The parameter estimates derived from the time series data and secondary infections data generated by these epidemics have a strong correlation which increases with the number of secondary infections observed. However, for the growth rate there is a clear trend that the secondary infections data tends to yield lower point estimates of the growth rate. A difference of this kind should not be ignored, however, given there will also be a level of uncertainty on these estimates, they will still give broadly consistent characterisations of the epidemic. The simulations used were generated with an RF model so there is not an obvious ground truth to compare these values to in order to further investigate which of the estimators is biased.

Together, this demonstrates that characterisations derived from each data type will be similar given a sufficient number secondary observations, however (particularly in the case of the growth rate) there are systematic differences in the estimates that we were unable to explain. Conditioned the process against extinction in the case of the time series estimator, but not in the case of the secondary infections estimator may be contributing to this systematic difference.

5.7 Discussion

We presented an in-homogeneous branching process to model outbreaks. The simplicity of the process means we can construct both a population and an individual scale view and subsequently assimilate data from either scale. Assimilating both data types simultaneously was beyond the scope of the current work. Since the conditional distribution of Poisson variables given their sum is multinomial, it would be feasible to perform simultaneous assimilation with a Poisson offspring distribution, however, the matter becomes more complicated when using a negative binomial distribution, as we use here.

Our model admits a closed form for the likelihood for the time series and we have supplied an approximation for the likelihood of the secondary infections data in Section 4.3.5. These closed forms make it feasible to do a Bayesian analysis and handle subtleties of the fitting process (in the case of the time series data), such as conditioning the process against extinction to account for the implicit observation bias. We do not address unobserved cases in the secondary infections data, nor do we have a sophisticated method for aggregating cases into generations. However, in our analysis of the Ebola data this does not appear to cause problems with the inference. This difference may be the cause of the systematic difference in the estimators obtained in the simulation study, however a more in-depth study of this was beyond the scope of the current work.

This work extends work carried out during the 2014 Ebola epidemic by Chowell *et al* [63] and a comprehensive study of the dynamics of several pathogens' transmission [65]. We used the same phenomenological model as the phenomenological backbone of the branching process. The resulting process has the same dynamics (on average) but with a mechanistic underpinning. This enables us to handle a wider range of data types, for example, the tree data from Faye *et al* [1].

Lags in time series data cause substantial problems when forecasting incidence [103]. "First Few Hundred" (FF100) studies collect the same type of data as contact tracing and are heralded as a way to rapidly provide a characterisation of transmission dynamics [48]. While for the 2014 Ebola epidemic the time series was available before the secondary infections tree, there does not seem to be anything intrinsic to the data collection process that precludes this being reversed. In fact, it seems plausible that in active surveillance programs and with increased use of sequencing, secondary infections data may become available before time series. Of course, there are ethical, procedural, and technical challenges that are introduced by collecting, analysing and storing this sort of data.

The work of Black *et al* [48] and Walker *et al* [172] considers transmission in the household, and to some extent, population level transmission using FF100 type data. In addition to estimating the transmission dynamics, these authors also account for uncertainty in the observation process. However, they do not validate their approach on real data; instead, they validated their inference on data generated using an agent-based model [173, 174]. Focusing on pandemic influenza, their model requires data at a finer temporal resolution than

the generation-based one we present and requires substantial computational resources.

Most pertinent to improving the value of our approach is establishing how to handle incomplete secondary infections data. We investigated the consequences of partial observation of the infectious population, but with perfect ascertainment of the number of infections due to each individual. A natural extension then is to consider partial observation of the population with imperfect resolution, i.e., observe a random subset of the infectious population and observe only a subset of their infections. This additional way in which data can be missing is particularly important in airborne disease, such as influenza, where the source of an infection may be harder to ascertain. If the goal is to characterise the transmission dynamics of a pathogen for which sub-clinical cases are rare, such as Ebola virus disease, then the assumption of complete observation among those observed does not seem unreasonable. As sequencing data becomes more readily available we will have improved capability to determine who-infected-whom and models such as the one presented in this work are poised to take advantage of this additional information.

5.8 Acknowledgements

I would like to acknowledge the helpful discussions with Gerardo Chowell and Peter Taylor during the conceptualisation of this work.

5.9 Supplementary materials

5.9.1 Reed-Frost simulation

The standard Reed-Frost process considers a closed population of N people. The process starts with a single member of the population infectious on day 0, the remaining $N - 1$ people are susceptible to the infection. Time is discrete, and individuals infected on day $n - 1$ become infectious on day n and are no longer susceptible or infectious (i.e., they are recovered) on day $n + 1$. If there are I_{n-1} infectious members and S_{n-1} susceptible members on day $n - 1$ then the number of individuals infected is a random variable:

$$I_n \sim \text{Bin}(S_{n-1}, \lambda I_{n-1}/N) \quad \text{and} \quad S_n = S_{n-1} - I_n. \quad (5.2)$$

The number of recovered individuals can be recovered from the conservation equation $N = S + I + R$. The parameter λ controls the number of secondary infections an individual generates in an otherwise susceptible population. The process terminates when either there are no more infectious individuals or there are no more susceptible individuals left.

Keeping track of the number of infectious individuals in each of the generations provides the time series data sets used in the simulation re-estimation study. Since each individual in the process is equally likely to infect each susceptible individual we assume that the infections can be distributed among the infectious individuals uniformly at random. The secondary infections data sets used in the simulation re-estimation study are obtained by taking a random sample of the infectious individuals and recording their generation number and the number of infections they generated. The relationship of who-infected-whom allows us to map this process to a distribution over a space of trees, where node a being the parent of node b indicates that a infected b . From this view, the secondary infections data consists of taking a random sample of the nodes and measuring its depth in the tree and the number

of child nodes it has. The secondary infections data consisted of the number of cases due to a sample of individuals in the first 6 generations, (the information available at the start of the 7th generation). This sampling mimics the (limited) data collected during contact tracing.

When sampling from this process we fixed the population at $N = 100000$ with $\lambda = 1.2$. We recorded the process through the first 7 generations of transmission. Since we are primarily interested in outbreaks that will go on to cause a substantial number of infections we conditioned the process against extinction during the first 10 generations, and that there should be at least 30 cases in these initial 10 generations.

6

Summary

The investigations required to accurately forecast epidemics with mechanistic models will improve our understanding of the underlying transmission process and enable us to anticipate its burden. As is currently done, [103], this can produce forecasts which are communicated to public health practitioners who may then use these as another piece of data in their decision making. For instance, knowing the number of influenza cases requiring hospitalisation at the peak of the epidemic would assist in managing hospital surge capacity. Mechanistic approaches also enable us to investigate the impacts of hypothetical intervention measures, e.g., vaccinating 30% of the population. In this thesis, we considered some of the challenges involved in forecasting epidemics using mechanistic models, primarily epidemics of seasonal influenza in Australian cities.

In Chapter 3 we reported on a Bayesian approach to model selection when the goal is to select the model with the best predictive skill. We proposed that a Bayesian measure is preferable for forecasting since it accounts for both the variance and the bias (rather than simpler methods which only consider the accuracy of point predictions.) Applying this methodology we found that incorporating even a rudimentary forecast of absolute humidity improved forecasts of seasonal influenza epidemics in Melbourne, Australia. Using the actual values of absolute humidity improved the forecasts further, but obviously, this data would not be available when generating the forecasts.

In Chapter 4 we reported an approach for constructing prior distributions. This approach makes it easier to incorporate information which is difficult to connect to parameters of a model. In doing so it was demonstrated how historical data can be used to construct a prior distribution which concentrates the probability mass in regions of parameter space which lead to plausible model outputs. We have provided a way that this could be done generically (finding a multivariate normal which leads to approximately correct system-level behaviour.) Applying this methodology to a retrospective forecast of influenza epidemics in Sydney, Australia, we found that this method improves forecasts generated early in the season. However, if there are substantial differences between the historical epidemics and the one being forecast this approach can introduce biases which reduce predictive skill.

In Chapter 5 we deviated from the theme of forecasting, reporting instead on the use of branching processes to model early transmission in an epidemic and how to estimate the

parameters of this model from either time series of incidence or infection trees. This was facilitated by considering a branching process as both a tree, in which there are explicit connections between who-infects-whom and as a time series of the total number of cases in each generation of the process. Consistent inferences about the dynamics of Ebola transmission were drawn from two distinct types of data from the epidemic. In particular, we found that initially, the incidence of cases grew sub-exponentially (rather than the exponential growth implicit in many models of epidemics). In a simulation study, we found separate analyses of time series data and infection trees derived from the same epidemic simulations led to broadly consistent parameters. This suggests that the two data sources could lead to consistent estimates more generally.

7

Discussion

What should the reader walk away with after reading this text? Well, I would like to hope it has encouraged them to think about several things: what it means for a prediction to be good or bad (after reading the chapter on model selection for forecasting), what we mean when we talk about prior knowledge (after reading the chapter on prior distributions), and finally what it means for a model to be multi-scale and how this relates to the use of multiple data types (after reading the chapter on branching processes). While carrying out the work reported in each of these chapters I found myself asking these questions (and many others) and not being completely satisfied with the answers. I still haven't answered these questions to my satisfaction, but I think I have pointed out some interesting things to think about for the next person tackling them. I will now consider each of these questions in a bit more detail.

Meaningfully judging the quality of a prediction is surprisingly difficult. There are obvious answers, involving quantitative measures of variance and bias which come to mind, but they tend to ignore the context in which these predictions are being used. As the field of epidemic forecasting matures and forecasts become more reliable (as has occurred with weather prediction) we will hopefully see them utilised to improve decisions about how to protect the public. For this to become a reality we need to consider the speed with which these predictions can be made, in the sense of how much data must we gather before we can start making useful predictions. Then there is the matter of who is consuming these predictions; if they are just being fed into the next component in a computational pipeline then perhaps having a full distribution of potential outcomes might be essential. But if forecasts are to be interpreted by people, a full distribution of outcomes could be overwhelming. This suggests that the larger question, *how good is a prediction?*, can only be answered on a case-by-case basis. In this thesis, we have provided an answer for one such case, model selection.

I have demonstrated how to incorporate existing knowledge into a prior distribution when that knowledge is a distribution of plausible outcomes. Typically, prior knowledge is thought of as beliefs about the values of parameters of a model; however, I have argued that this prior knowledge can pertain to any behaviour of a system. In the context of epidemic forecasting, this additional knowledge (about the timing and magnitude of the epidemic peak) can be useful as a way of constraining predictions to a plausible range of values. Importantly, you can encode system level knowledge into a prior distribution without needing to understand

how the parameters affect the system.

Multi-scale modelling and the use of multiple data sources is difficult but has great potential. I have demonstrated how a branching process can be considered an example of a multi-scale model and shown how parameters of this model can be estimated from different data types. As we continue to measure more aspects of our environment, the ability to make use of that data will improve the utility of our models. I hope the next time the reader encounters multiple sources of data they see this as an opportunity to develop a richer model which provides a natural connection to multiple data sources.

There are two prominent directions for future work which seem promising: closer collaboration between those producing forecasts and the intended end users, and improved understanding of the connection between epidemics and the data that is collected to characterise them. There is already valuable work being done in these areas (in particular that of my collaborators), as such this may be more of a promotion for their work than a discussion of novel future directions. I am confident that, if successful, the first would be of substantial value, however, it also seems the most challenging and the furthest from the type of material presented above (and hence the one I am least qualified to comment on!) The second is a more natural extension of the work presented above and as such, I have a better idea of how to pursue it. Unless progress is made in the first direction though, the value of progress on the latter will be diminished.

While improving our understanding of the processes driving epidemics is useful for developing a more thorough understanding of how to handle future outbreaks of infectious disease, there is the potential for this field to contribute more, giving a clearer understanding of the situation during an epidemic and providing evidence to inform decisions about how to react. For this potential to be realised trust and lines of communication need to be established between forecasters and public health practitioners: to allow information to flow, and be understood by both parties upon receipt (see for example the ongoing efforts of Moss *et al* [103].) This will require establishing a level of reliability in the forecasts and standard methods of communicating them which are easy to understand and use (see for example the ongoing efforts from the CDC [108, 175]). For many statistical techniques there are diagnostics which warn the user that the output (or input) needs close scrutiny. The widespread use of these diagnostics expands the pool of potential users and means that we can put more trust in the output of these techniques because we assume they will have prevented their user from making incorrect inferences. I think that, in the same manner, developing these sort of diagnostics for epidemic forecasts would make them easier to use and more trustworthy.

Retrospective forecasting of epidemics provides insight into the possible level of performance for current methodology. In doing so it is common, as we have done in this thesis, to simplify some of the complications of real-time forecasting. In particular observations are treated as static; however, during the epidemic there are frequently modifications to previous observations as additional data becomes available [176]. The process of collecting epidemiological data is complex and there are many points where bias can creep in. Improving the way in which we account for the observation process would reduce this bias and could inform the design of more effective observation processes. For this reason, I think that effort should be directed towards the development of observation models which thoroughly account for the complexity of the observation process. While an appropriate observation model can reduce the bias in a data set this would likely involve a bias-variance trade-off. Even moderately complex mechanistic models are usually multi-dimensional and data sources are usually only informative for one of its state variables. Consequently, I think that effort should be put into integrating data from multiple sources both to counter bias in any one data source and to

provide information about more aspects of the epidemic. Both refining existing observation models and incorporating information from additional sources will be made easier if there is closer collaboration with those who collect the data as they will be able to offer insights into the peculiarities of particular data sets.

The information age offers us the potential to drastically reduce the burden of infectious disease. While it may seem presumptuous, I hope the efforts reported here, to assist in the development of an influenza forecasting system for Australia, have in some way contributed to this.

References

- [1] O. Faye, P.-Y. Boëlle, E. Heleze, O. Faye, C. Loucoubar, N. Magassouba, B. Soropogui, S. Keita, T. Gakou, E. H. I. Bah, L. Koivogui, A. A. Sall, and S. Cauchemez. *Chains of transmission and control of Ebola virus disease in Conakry, Guinea, in 2014: an observational study*. The Lancet Infectious Diseases **15**(3), 320 (2015).
- [2] R. Moss, A. Zarebski, P. Dawson, and J. M. McCaw. *Retrospective forecasting of the 2010–2014 Melbourne influenza seasons using multiple surveillance systems*. Epidemiology & Infection **145**(1), 156 (2017).
- [3] J. M. Diamond. *Guns, Germs and Steel: A Short History of Everybody for the last 13,000 Years* (Random House, 1998).
- [4] K. G. Nicholson, J. M. Wood, and M. Zambon. *Influenza*. The Lancet **362**(9397), 1733 (2003).
- [5] World Health Organization. *Influenza (Seasonal)*. <http://www.who.int/mediacentre/factsheets/fs211/en/>. Accessed: 2016-08-17.
- [6] C. Dalton, D. Durrheim, J. Fejsa, L. Francis, S. Carlson, E. T. d’Espaignet, and F. Tuyl. *Flutracking: A weekly Australian community online survey of influenza-like illness in 2006, 2007 and 2008*. Communicable Diseases Intelligence Quarterly Report **33**(3), 316 (2009).
- [7] G. Brankston, L. Gitterman, Z. Hirji, C. Lemieux, and M. Gardam. *Transmission of influenza A in human beings*. The Lancet Infectious Diseases **7**(4), 257 (2007).
- [8] J. Shaman and M. Kohn. *Absolute humidity modulates influenza survival, transmission, and seasonality*. Proceedings of the National Academy of Sciences **106**(9), 3243 (2009).
- [9] J. Shaman, V. E. Pitzer, C. Viboud, B. T. Grenfell, and M. Lipsitch. *Absolute Humidity and the Seasonal Onset of Influenza in the Continental United States*. PLOS Biology **8**(2), 1 (2010).
- [10] R. Yaari, G. Katriel, A. Huppert, J. B. Axelsen, and L. Stone. *Modelling seasonal influenza: the role of weather and punctuated antigenic drift*. Journal of The Royal Society Interface **10**(84) (2013).
- [11] Tamerius, James D. AND Shaman, Jeffrey AND Alonso, Wladimir J. AND Bloom-Feshbach, Kimberly AND Uejio, Christopher K. AND Comrie, Andrew AND Viboud, Cécile. *Environmental Predictors of Seasonal Influenza Epidemics across Temperate and Tropical Climates*. PLOS Pathogens **9**(3), 1 (2013).

- [12] W. Yang, M. J. Cummings, B. Bakamutumaho, J. Kayiwa, N. Owor, B. Namagambo, T. Byaruhanga, J. J. Lutwama, M. R. O'Donnell, and J. Shaman. *Dynamics of influenza in tropical Africa: Temperature, humidity, and co-circulating (sub)types*. INFLUENZA AND OTHER RESPIRATORY VIRUSES **12**(4), 446 (2018).
- [13] B. D. Dalziel, S. Kissler, J. R. Gog, C. Viboud, O. N. Bjornstad, C. J. E. Metcalf, and B. T. Grenfell. *Urbanization and humidity shape the intensity of influenza epidemics in US cities*. SCIENCE **362**(6410, SI), 75 (2018).
- [14] L. Ferguson, A. K. Olivier, S. Genova, W. B. Epperson, D. R. Smith, L. Schneider, K. Barton, K. McCuan, R. J. Webby, and X.-F. Wan. *Pathogenesis of Influenza D Virus in Cattle*. Journal of Virology **90**(12), 5636 (2016).
- [15] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher. *Nextstrain: real-time tracking of pathogen evolution*. Bioinformatics **34**(23), 4121 (2018).
- [16] X. Du, A. A. King, R. J. Woods, and M. Pascual. *Evolution-informed forecasting of seasonal influenza A (H3N2)*. Science Translational Medicine **9**(413), eaan5325 (2017).
- [17] K. Koelle, S. Cobey, B. Grenfell, and M. Pascual. *Epochal Evolution Shapes the Phylodynamics of Interpandemic Influenza A (H3N2) in Humans*. Science **314**(5807), 1898 (2006).
- [18] P. Cao, A. W. C. Yan, J. M. Heffernan, S. Petrie, R. G. Moss, L. A. Carolan, T. A. Guarnaccia, A. Kelso, I. G. Barr, J. McVernon, K. L. Laurie, and J. M. McCaw. *Innate Immunity and the Inter-exposure Interval Determine the Dynamics of Secondary Influenza Virus Infection and Explain Observed Viral Hierarchies*. PLOS Computational Biology **11**(8), 1 (2015).
- [19] I. Barberis, P. Myles, S. Ault, N. L. Bragazzi, and M. Martini. *History and evolution of influenza control through vaccination: from the first monovalent vaccine to universal vaccines*. Journal of Preventive Medicine and Hygiene **57**(3), E115 (2016).
- [20] S. A. Harper, K. Fukuda, T. M. Uyeki, N. J. Cox, and C. B. Bridges. *Prevention and Control of Influenza: Recommendations of the Advisory Committee on Immunization Practices (ACIP)*. Morbidity and Mortality Weekly Report: Recommendations and Reports **54**(8), 1 (2005).
- [21] C. Jackson, E. Vynnycky, J. Hawker, B. Olowokure, and P. Mangtani. *School closures and influenza: systematic review of epidemiological studies*. BMJ Open **3**(2) (2013).
- [22] World Health Organization. *Influenza (Avian and other zoonotic)*. [https://www.who.int/en/news-room/fact-sheets/detail/influenza-\(avian-and-other-zoonotic\)](https://www.who.int/en/news-room/fact-sheets/detail/influenza-(avian-and-other-zoonotic)). Accessed: 2019-02-19.
- [23] World Health Organization. *Ebola virus disease*. <https://www.who.int/en/news-room/fact-sheets/detail/ebola-virus-disease>. Accessed: 2019-02-19.
- [24] Y. A. Kuznetsov and C. Piccardi. *Bifurcation analysis of periodic SEIR and SIR epidemic models*. Journal of Mathematical Biology **32**(2), 109 (1994).

- [25] J. Dushoff, J. B. Plotkin, S. A. Levin, and D. J. D. Earn. *Dynamical resonance can account for seasonality of influenza epidemics*. Proceedings of the National Academy of Sciences **101**(48), 16915 (2004).
- [26] D. N. Kyriacou, D. Dobrez, J. P. Parada, J. M. Steinberg, A. Kahn, C. L. Bennett, and B. P. Schmitt. *Cost-Effectiveness Comparison of Response Strategies to a Large-Scale Anthrax Attack on the Chicago Metropolitan Area: Impact of Timing and Surge Capacity*. Biosecurity and Bioterrorism **10**(3), 264 (2012).
- [27] G. S. Zaric, D. M. Bravata, J.-E. Cleophas Holty, K. M. McDonald, D. K. Owens, and M. L. Brandeau. *Modeling the logistics of response to anthrax bioterrorism*. Medical Decision Making **28**(3), 332 (2008).
- [28] J.-P. Chretien, D. George, J. Shaman, R. A. Chitale, and F. E. McKenzie. *Influenza Forecasting in Human Populations: A Scoping Review*. PLOS ONE **9**(4), 1 (2014).
- [29] S. Kandula, T. Yamana, S. Pei, W. Yang, H. Morita, and J. Shaman. *Evaluation of mechanistic and statistical methods in forecasting influenza-like illness*. Journal of The Royal Society Interface **15**(144), 20180174 (2018).
- [30] N. G. Reich, L. C. Brooks, S. J. Fox, S. Kandula, C. J. McGowan, E. Moore, D. Osthus, E. L. Ray, A. Tushar, T. K. Yamana, M. Biggerstaff, M. A. Johansson, R. Rosenfeld, and J. Shaman. *A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States*. Proceedings of the National Academy of Sciences **116**(8), 3146 (2019).
- [31] M. T. Ribeiro, S. Singh, and C. Guestrin. *Why Should I Trust You?: Explaining the Predictions of Any Classifier*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144 (ACM, 2016).
- [32] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. *Predicting Flu Trends using Twitter data*. In *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 702–707 (2011).
- [33] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. *Detecting influenza epidemics using search engine query data*. Nature **457**(7232), 1012 (2009).
- [34] K. Dietz and J. Heesterbeek. *Daniel Bernoulli’s epidemiological model revisited*. Mathematical Biosciences **180**(1-2), 1 (2002).
- [35] W. O. Kermack and A. G. McKendrick. *Contributions to the Mathematical Theory of Epidemics I*. Proceedings of the Royal Society A **115A**, 700 (1927).
- [36] W. O. Kermack and A. G. McKendrick. *Contributions to the Mathematical Theory of Epidemics II: The problem of endemicity*. Proceedings of the Royal Society **138**(834), 55 (1932).
- [37] F. Brauer, P. van den Driessche, and J. Wu. *Mathematical Epidemiology* (Springer, Berlin, Heidelberg, 2008).
- [38] H. Andersson and T. Britton. *Stochastic Epidemic Models and Their Statistical Analysis*, vol. 151 (Springer, Berlin, Heidelberg, 2012).

- [39] J. A. Lewnard, M. L. N. Mbah, J. A. Alfaro-Murillo, F. L. Altice, L. Bawo, T. G. Nyenswah, and A. P. Galvani. *Dynamics and control of Ebola virus transmission in Montserrado, Liberia: a mathematical modelling analysis*. The Lancet Infectious Diseases **14**(12), 1189 (2014).
- [40] K. Y. Leung, P. Trapman, and T. Britton. *Who is the infector? Epidemic models with symptomatic and asymptomatic cases*. Mathematical Biosciences **301**, 190 (2018).
- [41] P. D. Stroud, S. J. Sydorik, J. M. Riese, J. P. Smith, S. M. Mniszewski, and P. R. Romero. *Semi-empirical power-law scaling of new infection rate to model epidemic dynamics with inhomogeneous mixing*. Mathematical Biosciences **203**(2), 301 (2006).
- [42] M. Roy and M. Pascual. *On representing network heterogeneities in the incidence rate of simple epidemic models*. Ecological Complexity **3**(1), 80 (2006).
- [43] L. Pellis, F. Ball, S. Bansal, K. Eames, T. House, V. Isham, and P. Trapman. *Eight challenges for network epidemic models*. Epidemics **10**, 58 (2015).
- [44] I. Z. Kiss, J. C. Miller, and P. L. Simon. *Mathematics of Epidemics on Networks* (Springer, 2017).
- [45] D. Mollison. *Dependence of epidemic and population velocities on basic parameters*. Mathematical Biosciences **107**(2), 255 (1991).
- [46] T. G. Kurtz. *Limit Theorems for Sequences of Jump Markov Processes*. Journal of Applied Probability **8**(2), 344 (1971).
- [47] S. M. Ross. *Introduction to Probability Models* (Academic press, 2014).
- [48] A. J. Black, N. Geard, J. M. McCaw, J. McVernon, and J. V. Ross. *Characterising pandemic severity and transmissibility from data collected during first few hundred studies*. Epidemics **19**, 61 (2017).
- [49] D. T. Gillespie. *A general method for numerically simulating the stochastic time evolution of coupled chemical reactions*. Journal of Computational Physics **22**(4), 403 (1976).
- [50] J. L. Doob. *Markoff Chains-Denumerable Case*. Transactions of the American Mathematical Society **58**(3), 455 (1945).
- [51] M. S. Bartlett. *Stochastic Processes or the Statistics of Change*. Journal of the Royal Statistical Society. Series C **2**, 44 (1953).
- [52] D. T. Gillespie. *Approximate accelerated stochastic simulation of chemically reacting systems*. The Journal of Chemical Physics **115**(4), 1716 (2001).
- [53] A. Moraes, R. Tempone, and P. Vilanova. *Hybrid Chernoff Tau-Leap*. Multiscale Modeling & Simulation **12**(2), 581 (2014).
- [54] M. J. Keeling and J. V. Ross. *Efficient methods for studying stochastic disease and population dynamics*. Theoretical Population Biology **75**(2-3), 133 (2009).

- [55] P. G. Ballard, N. G. Bean, and J. V. Ross. *The probability of epidemic fade-out is non-monotonic in transmission rate for the Markovian SIR model with demography*. *Journal of Theoretical Biology* **393**, 170 (2016).
- [56] E. J. Allen, L. J. S. Allen, A. Arciniega, and P. E. Greenwood. *Construction of Equivalent Stochastic Differential Equation Models*. *Stochastic Analysis and Applications* **26**(2), 274 (2008).
- [57] D. J. Higham. *An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations*. *SIAM Review* **43**(3), 525 (2001).
- [58] A. Skvortsov, B. Ristic, and C. Woodruff. *Predicting an epidemic based on syndromic surveillance*. In *2010 13th International Conference on Information Fusion* (2010).
- [59] W. Feller. *An Introduction to Probability Theory and Its Applications*, vol. 2 (John Wiley & Sons, 2008).
- [60] H. W. Watson and F. Galton. *On the Probability of the Extinction of Families*. *The Journal of the Anthropological Institute of Great Britain and Ireland* **4**, 138 (1875).
- [61] *Strengths and weaknesses*. <https://xkcd.com/1545/>. Accessed: 2018-10-14.
- [62] P. Flajolet and R. Sedgewick. *Analytic Combinatorics* (Cambridge University Press, 2009).
- [63] G. Chowell, C. Viboud, J. M. Hyman, and L. Simonsen. *The Western Africa Ebola Virus Disease Epidemic Exhibits Both Global Exponential and Local Polynomial Growth Rates*. *PLOS Currents* **7** (2015).
- [64] G. Chowell, D. Hincapie-Palacio, J. Ospina, B. Pell, A. Tariq, S. Dahal, S. Moghadas, A. Smirnova, L. Simonsen, and C. Viboud. *Using Phenomenological Models to Characterize Transmissibility and Forecast Patterns and Final Burden of Zika Epidemics*. *PLOS Currents* **8** (2016).
- [65] C. Viboud, L. Simonsen, and G. Chowell. *A generalized-growth model to characterize the early ascending phase of infectious disease outbreaks*. *Epidemics* **15**, 27 (2016).
- [66] Duygu Balcan and Bruno Gonçalves and Hao Hu and José J. Ramasco and Vittoria Colizza and Alessandro Vespignani. *Modeling the spatial spread of infectious diseases: The GLEaMviz computational model*. *Journal of Computational Science* **1**(3), 132 (2010).
- [67] Broeck, Wouter Van den and Gioannini, Corrado and Gonçalves, Bruno and Quagiotto, Marco and Colizza, Vittoria and Vespignani, Alessandro. *The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale*. *BMC Infectious Diseases* **11**(1), 37 (2011).
- [68] J. B. Dunham. *An Agent-Based Spatially Explicit Epidemiological Model in MASON*. *Journal of Artificial Societies and Social Simulation* **9**(1), 3 (2005).
- [69] Held, Leonhard and Hofmann, Mathias and Höhle, Michael and Schmid, Volker. *A two-component model for counts of infectious diseases*. *Biostatistics* **7**(3), 422 (2006).

- [70] D. Osthus, J. Gattiker, R. Priedhorsky, and S. Y. Del Valle. *Dynamic Bayesian Influenza Forecasting in the United States with Hierarchical Discrepancy (with Discussion)*. Bayesian Analysis **14**(1), 261 (2019).
- [71] S. Funk, A. Camacho, A. J. Kucharski, R. M. Eggo, and W. J. Edmunds. *Real-time forecasting of infectious disease dynamics with a stochastic semi-mechanistic model*. Epidemics **22**, 56 (2018).
- [72] D. Klinkenberg, C. Fraser, and H. Heesterbeek. *The Effectiveness of Contact Tracing in Emerging Epidemics*. PLOS ONE **1**(1), 1 (2006).
- [73] World Health Organization. *Implementation and management of contact tracing for Ebola virus disease* (2015).
- [74] Public Health England. *"The First Few Hundred (FF100)" Enhanced Case and Contact Protocol* (2013).
- [75] A. Lindén and S. Mäntyniemi. *Using the negative binomial distribution to model overdispersion in ecological count data*. Ecology **92**(7), 1414 (2011).
- [76] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference* (Springer Science & Business Media, 2013).
- [77] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis* (Chapman and Hall, 1995).
- [78] S. Selvin, M. Bloxham, A. I. Khuri, M. Moore, R. Coleman, G. R. Bryce, J. A. Hagans, T. C. Chalmers, E. A. Maxwell, and G. N. Smith. *A Problem in Probability [Letters to the Editor]*. The American Statistician **29**(1), 67 (1975).
- [79] D. V. Lindley. *Theory and Practice of Bayesian Statistics*. Journal of the Royal Statistical Society. Series D **32**(1/2), 1 (1983).
- [80] C. Walters and D. Ludwig. *Calculation of Bayes Posterior Probability Distributions for Key Population Parameters*. Canadian Journal of Fisheries and Aquatic Sciences **51**(3), 713 (1994).
- [81] B. De Finetti. *Theory of Probability: A Critical Introductory Treatment*, vol. 6 (John Wiley & Sons, 2017).
- [82] E. T. Jaynes. *Probability Theory: The Logic of Science* (Cambridge University Press, 2003).
- [83] H. Jeffreys. *An invariant form for the prior probability in estimation problems*. Proceedings of the Royal Society A **186**(1007), 453 (1946).
- [84] J. M. Bernardo. *Reference Posterior Distributions for Bayesian Inference*. Journal of the Royal Statistical Society. Series B **41**(2), 113 (1979).
- [85] J. M. Bernardo. *Reference Analysis*. Handbook of Statistics **25**, 17 (2005).
- [86] J. B. Haldane. *The Precision of Observed Values of Small Frequencies*. Biometrika **35**(3/4), 297 (1948).

- [87] W. Talbott. *Bayesian Epistemology*. In E. N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Metaphysics Research Lab, Stanford University, 2016), Winter 2016 ed.
- [88] A. Gelman, F. Bois, and J. Jiang. *Physiological Pharmacokinetic Analysis Using Population Modeling and Informative Prior Distributions*. *Journal of the American Statistical Association* **91**(436), 1400 (1996).
- [89] M. Goldstein. *Subjective Bayesian Analysis: Principles and Practice*. *Bayesian Analysis* **1**(3), 403 (2006).
- [90] N. N. Taleb. *Black Swans and the Domains of Statistics*. *The American Statistician* **61**(3), 198 (2007).
- [91] P. Diaconis, S. Holmes, and R. Montgomery. *Dynamical Bias in the Coin Toss*. *SIAM Review* **49**(2), 211 (2007).
- [92] C. Robert and G. Casella. *A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data*. *Statistical Science* **26**(1), 102 (2011).
- [93] R. Faragher. *Understanding the Basis of the Kalman Filter Via a Simple and Intuitive Derivation*. *IEEE Signal Processing Magazine* **29**(5), 128 (2012).
- [94] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. *A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking*. *IEEE Transactions on Signal Processing* **50**(2), 174 (2002).
- [95] M. A. Beaumont, W. Zhang, and D. J. Balding. *Approximate Bayesian Computation in Population Genetics*. *Genetics* **162**(4), 2025 (2002).
- [96] C. M. Bishop. *Pattern Recognition and Machine Learning* (Springer, Berlin, Heidelberg, 2006).
- [97] J. K. Taubenberger and D. M. Morens. *1918 Influenza: the Mother of All Pandemics*. *Emerging Infectious Diseases* **12**(1), 15 (2006).
- [98] R. Moss, A. Zarebski, P. Dawson, and J. M. McCaw. *Forecasting influenza outbreak dynamics in Melbourne from Internet search query surveillance data*. *Influenza and Other Respiratory Viruses* **10**(4), 314 (2016).
- [99] D. Lazer, R. Kennedy, G. King, and A. Vespignani. *The Parable of Google Flu: Traps in Big Data Analysis*. *Science* **343**(6176), 1203 (2014).
- [100] G. Chowell, S. Echevarría-Zuno, C. Viboud, L. Simonsen, J. Tamerius, M. A. Miller, and V. H. Borja-Aburto. *Characterizing the Epidemiology of the 2009 Influenza A/H1N1 Pandemic in Mexico*. *PLOS Medicine* **8**(5), 1 (2011).
- [101] J. B. S. Ong, M. I.-C. Chen, A. R. Cook, H. C. Lee, V. J. Lee, R. T. P. Lin, P. A. Tambyah, and L. G. Goh. *Real-Time Epidemic Monitoring and Forecasting of H1N1-2009 Using Influenza-Like Illness from General Practice and Family Doctor Clinics in Singapore*. *PLOS ONE* **5**(4), 1 (2010).
- [102] C. Viboud and A. Vespignani. *The future of influenza forecasts*. *Proceedings of the National Academy of Sciences* **116**(8), 2802 (2019).

- [103] R. Moss, J. E. Fielding, L. J. Franklin, N. Stephens, J. McVernon, P. Dawson, and J. M. McCaw. *Epidemic forecasts as a tool for public health: interpretation and (re) calibration*. Australian and New Zealand Journal of Public Health **42**(1), 69 (2018).
- [104] F. S. Lu, M. W. Hattab, C. L. Clemente, M. Biggerstaff, and M. Santillana. *Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches*. Nature Communications **10**(1), 147 (2019).
- [105] A. E. Zarebski, P. Dawson, J. M. McCaw, and R. Moss. *Model selection for seasonal influenza forecasting*. Infectious Disease Modelling **2**(1), 56 (2017).
- [106] R. Moss, A. E. Zarebski, S. J. Carlson, and J. M. McCaw. *Accounting for Healthcare-Seeking Behaviours and Testing Practices in Real-Time Influenza Forecasts*. Tropical Medicine and Infectious Disease **4**(1) (2019).
- [107] C. Reed, S. S. Chaves, P. Daily Kirley, R. Emerson, D. Aragon, E. B. Hancock, L. Butler, J. Baumbach, G. Hollick, N. M. Bennett, M. R. Laidler, A. Thomas, M. I. Meltzer, and L. Finelli. *Estimating Influenza Disease Burden from Population-Based Surveillance Data in the United States*. PLOS ONE **10**(3), 1 (2015).
- [108] M. Biggerstaff, D. Alper, M. Dredze, S. Fox, I. C.-H. Fung, K. S. Hickmann, B. Lewis, R. Rosenfeld, J. Shaman, M.-H. Tsou, P. Velardi, A. Vespignani, L. Finelli, and the Influenza Forecasting Contest Working Group. *Results from the centers for disease control and prevention’s predict the 2013–2014 Influenza Season Challenge*. BMC Infectious Diseases **16**(1), 357 (2016).
- [109] M. Ben-Nun, P. Riley, J. Turtle, D. Bacon, and S. Riley. *National and Regional Influenza-Like-Illness Forecasts for the USA*. bioRxiv (2018).
- [110] E. L. Ray and N. G. Reich. *Prediction of infectious disease epidemics via weighted density ensembles*. PLOS Computational Biology **14**(2), 1 (2018).
- [111] D. C. Farrow, L. C. Brooks, S. Hyun, R. J. Tibshirani, D. S. Burke, and R. Rosenfeld. *A human judgment approach to epidemiological forecasting*. PLOS Computational Biology **13**(3), 1 (2017).
- [112] D. N. Fisman, T. S. Hauck, A. R. Tuite, and A. L. Greer. *An idea for short term outbreak projection: Nearcasting using the basic reproduction number*. PLOS ONE **8**(12), 1 (2013).
- [113] B. C. K. Choi and A. W. P. Pak. *A simple approximate mathematical model to predict the number of severe acute respiratory syndrome cases and deaths*. Journal of Epidemiology & Community Health **57**(10), 831 (2003).
- [114] M. Lipsitch, T. Cohen, B. Cooper, J. M. Robins, S. Ma, L. James, G. Gopalakrishna, S. K. Chew, C. C. Tan, M. H. Samore, D. Fisman, and M. Murray. *Transmission Dynamics and Control of Severe Acute Respiratory Syndrome*. Science **300**(5627), 1966 (2003).
- [115] E. Massad, M. N. Burattini, L. F. Lopez, and F. A. Coutinho. *Forecasting versus projection models in epidemiology: The case of the SARS epidemics*. Medical Hypotheses **65**(1), 17 (2005).

- [116] D. Bernoulli. *Réflexions sur les avantages de l'inoculation*. Mercure de France pp. 439–482 (1760).
- [117] J. L. R. d'Alembert. *Sur l'application du calcul des probabilités à l'inoculation de la petite vérole*. Opuscles mathématiques **2**, 26 (1761).
- [118] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control* (Oxford University Press, 1992).
- [119] M. J. Keeling and P. Rohani. *Modeling Infectious Diseases in Humans and Animals* (Princeton University Press, 2011).
- [120] J. D. Murray. *Mathematical Biology II: Spatial Models and Biomedical Applications* (Springer-Verlag New York Incorporated New York, 2001).
- [121] G. Chowell, L. Sattenspiel, S. Bansal, and C. Viboud. *Mathematical models to characterize early epidemic growth: A review*. Physics of Life Reviews **18**, 66 (2016).
- [122] H. Jeffreys. *The Theory of Probability* (Oxford University Press, 1998).
- [123] C. P. Robert, N. Chopin, and J. Rousseau. *Harold Jeffreys's Theory of Probability Revisited*. Statistical Science **24**(2), 141 (2009).
- [124] E. T. Jaynes. *Prior Probabilities*. IEEE Transactions on Systems Science and Cybernetics **4**(3), 227 (1968).
- [125] A. Gelman, D. Simpson, and M. Betancourt. *The Prior Can Often Only Be Understood in the Context of the Likelihood*. Entropy **19**(10) (2017).
- [126] J. Gabry, D. Simpson, A. Vehtari, M. Betancourt, and A. Gelman. *Visualization in Bayesian workflow*. Journal of the Royal Statistical Society: Series A (Statistics in Society) **182**(2), 389 (2019).
- [127] S. M. Ross. *A Course in Simulation* (Prentice Hall PTR, 1990).
- [128] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. *Equation of State Calculations by Fast Computing Machines*. The Journal of Chemical Physics **21**(6), 1087 (1953).
- [129] W. K. Hastings. *Monte Carlo sampling methods using Markov chains and their applications*. Biometrika **57**(1), 97 (1970).
- [130] S. Chib and E. Greenberg. *Understanding the Metropolis-Hastings Algorithm*. The American Statistician **49**(4), 327 (1995).
- [131] D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. *WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility*. Statistics and Computing **10**(4), 325 (2000).
- [132] D. Lunn, D. Spiegelhalter, A. Thomas, and N. Best. *The BUGS project: Evolution, critique and future directions*. Statistics in Medicine **28**(25), 3049 (2009).

- [133] M. Plummer. *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, vol. 124 (Vienna, Austria, 2003).
- [134] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. *Stan: A Probabilistic Programming Language*. *Journal of Statistical Software* **76**(1) (2017).
- [135] R. M. Neal. *MCMC using Hamiltonian dynamics*. *Handbook of Markov Chain Monte Carlo* **2**(11), 2 (2011).
- [136] M. D. Hoffman and A. Gelman. *The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo*. *Journal of Machine Learning Research* **15**(1), 1593 (2014).
- [137] M. A. Suchard, P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond, and A. Rambaut. *Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10*. *Virus Evolution* **4**(1), vey016 (2018).
- [138] R. Bouckaert, J. Heled, D. Kühnert, T. Vaughan, C.-H. Wu, D. Xie, M. A. Suchard, A. Rambaut, and A. J. Drummond. *BEAST 2: A Software Platform for Bayesian Evolutionary Analysis*. *PLOS Computational Biology* **10**(4), 1 (2014).
- [139] S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. *RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model-Specification Language*. *Systematic Biology* **65**(4), 726 (2016).
- [140] A. Doucet and A. M. Johansen. *A tutorial on particle filtering and smoothing: Fifteen years later*. *Handbook of Nonlinear Filtering* **12**(656-704), 3 (2009).
- [141] R. E. Kalman. *A New Approach to Linear Filtering and Prediction Problems*. *Journal of Basic Engineering* **82**(1), 35 (1960).
- [142] S. J. Julier and J. K. Uhlmann. *Unscented Filtering and Nonlinear Estimation*. *Proceedings of the IEEE* **92**(3), 401 (2004).
- [143] B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman filter: Particle filters for tracking applications* (Artech House, 2003).
- [144] A. Doucet, N. De Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice* (Springer, 2001).
- [145] J. Liu and M. West. *Combined Parameter and State Estimation in Simulation-Based Filtering*. In *Sequential Monte Carlo Methods in Practice*, pp. 197–223 (Springer, 2001).
- [146] G. Kitagawa. *Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models*. *Journal of Computational and Graphical Statistics* **5**(1), 1 (1996).
- [147] A. King, D. Nguyen, and E. Ionides. *Statistical Inference for Partially Observed Markov Processes via the R Package pomp*. *Journal of Statistical Software, Articles* **69**(12), 1 (2016).

- [148] W. Yang and J. Shaman. *A simple modification for improving inference of non-linear dynamical systems*. arXiv preprint arXiv:1403.6804 (2014).
- [149] P. McCullagh. *What is a Statistical Model?* The Annals of Statistics pp. 1225–1267 (2002).
- [150] B. Rannala and Z. Yang. *Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference*. Journal of Molecular Evolution **43**(3), 304 (1996).
- [151] P. S. Craig, M. Goldstein, A. H. Seheult, and J. A. Smith. *Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments*. In *Case Studies in Bayesian Statistics*, pp. 37–93 (Springer, 1997).
- [152] I. Andrianakis, I. R. Vernon, N. McCreesh, T. J. McKinley, J. E. Oakley, R. N. Nsubuga, M. Goldstein, and R. G. White. *Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda*. PLOS Computational Biology **11**(1), 1 (2015).
- [153] X. Wang, D. J. Nott, C. C. Drovandi, K. Mengersen, and M. Evans. *Using History Matching for Prior Choice*. Technometrics **60**(4), 445 (2018).
- [154] S. B. Lambert, C. E. Faux, K. A. Grant, S. H. Williams, C. Bletchly, M. G. Catton, D. W. Smith, H. A. Kelly, *et al.* *Influenza surveillance in Australia: we need to do more than count*. The Medical Journal of Australia **193**(1), 43 (2010).
- [155] The Centre for Disease Control and Prevention. *FluSight 2017–2018* (2018). [Online; accessed 27-August-2018], URL <https://predict.phiresearchlab.org/>.
- [156] P. E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction* (Random House, 2016).
- [157] S. Y. Del Valle, B. H. McMahon, J. Asher, R. Hatchett, J. C. Lega, H. E. Brown, M. E. Leany, Y. Pantazis, D. J. Roberts, S. Moore, A. T. Peterson, L. E. Escobar, H. Qiao, N. W. Hengartner, and H. Mukundan. *Summary results of the 2014-2015 DARPA Chikungunya challenge*. BMC Infectious Diseases **18**(1), 245 (2018).
- [158] C. Viboud, K. Sun, R. Gaffey, M. Ajelli, L. Fumanelli, S. Merler, Q. Zhang, G. Chowell, L. Simonsen, and A. Vespignani. *The RAPIDD ebola forecasting challenge: Synthesis and lessons learnt*. Epidemics **22**, 13 (2018).
- [159] P. M. Kuhnert, T. G. Martin, and S. P. Griffiths. *A guide to eliciting and using expert knowledge in Bayesian ecological models*. Ecology Letters **13**(7), 900 (2010).
- [160] T. G. Martin, M. A. Burgman, F. Fidler, P. M. Kuhnert, S. Low-Choy, M. McBride, and K. Mengersen. *Eliciting Expert Knowledge in Conservation Science*. Conservation Biology **26**(1), 29 (2012).
- [161] J. Kadane and L. Wolfson. *Experiences in elicitation*. Journal of the Royal Statistical Society Series D - The Statistician **47**(1), 3 (1998). Royal Statistical Society Meeting on Elicitation, London, England, April 16, 1997.

- [162] J. Lega and H. E. Brown. *Data-driven outbreak forecasting with a simple nonlinear growth model*. *Epidemics* **17**, 19 (2016).
- [163] P. Nouvellet, A. Cori, T. Garske, I. M. Blake, I. Dorigatti, W. Hinsley, T. Jombart, H. L. Mills, G. Nedjati-Gilani, M. D. V. Kerkhove, C. Fraser, C. A. Donnelly, N. M. Ferguson, and S. Riley. *A simple approach to measure transmissibility and forecast incidence*. *Epidemics* **22**, 29 (2018). The RAPIDD Ebola Forecasting Challenge.
- [164] M. Ajelli, S. Merler, L. Fumanelli, A. Pastore y Piontti, N. E. Dean, I. M. Longini, M. E. Halloran, and A. Vespignani. *Spatiotemporal dynamics of the Ebola epidemic in Guinea and implications for vaccination and disease elimination: a computational modeling analysis*. *BMC Medicine* **14**(1), 130 (2016).
- [165] S. Merler, M. Ajelli, L. Fumanelli, M. F. Gomes, A. P. y Piontti, L. Rossi, D. L. Chao, I. M. Longini Jr, M. E. Halloran, and A. Vespignani. *Spatiotemporal spread of the 2014 outbreak of Ebola virus disease in Liberia and the effectiveness of non-pharmaceutical interventions: a computational modelling analysis*. *The Lancet Infectious Diseases* **15**(2), 204 (2015).
- [166] J. O. Lloyd-Smith, S. J. Schreiber, P. E. Kopp, and W. M. Getz. *Superspreading and the effect of individual variation on disease emergence*. *Nature* **438**(7066), 355 (2005).
- [167] T. Stadler. *How Can We Improve Accuracy of Macroevolutionary Rate Estimates?* *Systematic Biology* **62**(2), 321 (2013).
- [168] G. N. Mercer, K. Glass, and N. G. Becker. *Effective reproduction numbers are commonly overestimated early in a disease outbreak*. *Statistics in Medicine* **30**(9), 984 (2011).
- [169] W. N. Rida. *Asymptotic Properties of Some Estimators for the Infection Rate in the General Stochastic Epidemic Model*. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(1), 269 (1991).
- [170] World Health Organization. *Ebola data and statistics*. <http://apps.who.int/gho/data/node ebola-sitrep.quick-downloads?lang=en>. Accessed: 2018-01-15.
- [171] G. Chowell and H. Nishiura. *Transmission dynamics and control of Ebola virus disease (EVD): a review*. *BMC Medicine* **12**(1), 196 (2014).
- [172] J. N. Walker, J. V. Ross, and A. J. Black. *Inference of epidemiological parameters from household stratified data*. *PLOS ONE* **12**(10), 1 (2017).
- [173] N. Geard, J. M. McCaw, A. Dorin, K. B. Korb, and J. McVernon. *Synthetic Population Dynamics: A Model of Household Demography*. *Journal of Artificial Societies and Social Simulation* **16**(1), 8 (2013).
- [174] N. Geard, K. Glass, J. M. McCaw, E. S. McBryde, K. B. Korb, M. J. Keeling, and J. McVernon. *The effects of demographic change on disease transmission and vaccine impact in a household structured population*. *Epidemics* **13**, 56 (2015).

-
- [175] M. Biggerstaff, M. Johansson, D. Alper, L. C. Brooks, P. Chakraborty, D. C. Farrow, S. Hyun, S. Kandula, C. McGowan, N. Ramakrishnan, R. Rosenfeld, J. Shaman, R. Tibshirani, R. J. Tibshirani, A. Vespignani, W. Yang, Q. Zhang, and C. Reed. *Results from the second year of a collaborative effort to forecast influenza seasons in the United States*. *Epidemics* **24**, 26 (2018).
- [176] R. Moss, A. E. Zarebski, P. Dawson, L. J. Franklin, F. A. Birrell, and J. M. McCaw. *Anatomy of a seasonal influenza epidemic forecast*. *Communicable Diseases Intelligence* **43** (2019).