Modelling the 2010-2019 Haiti Cholera Epidemic

A Study in Simulation-Based Inference for Infectious Disease Transmission

Anna Rosengart

Thesis Advisor
Professor Edward Ionides

Department of Statistics, University of Michigan

Abstract

At the onset of an epidemic, it is common for disease intervention methods to be evaluated for efficacy via mathematical modelling prior to deployment. Model specification and construction must be informed by both the case data of the disease under study and the scientific principles underlying its transmission. For any given epidemic, there are countless possible models that can be formulated and used for motivating public health action, thus underscoring the importance of model criticism and comparison. Although there may be many models that vary in implementation, complexity, and mechanistic and stochastic elements, it is imperative that models with the best available forecasting accuracy be used for informing policy in real-life health crises. To exemplify this point we use the 2010-2019 cholera epidemic in Haiti as a case study. Through the analysis of three different stochastic models, we show the wide variability in model quality and forecasting that can result from minor changes in model specification and calibration.

All materials and code can be downloaded at `https://github.com/aerosengart/haiti_thesis`.

1 Introduction

In late October of 2010, the first cases of cholera in over one hundred years were observed in Haiti. Within the next several weeks, case reports skyrocketed into the thousands and reached over 230,000 by mid-February of the next year [17, 2]. Exacerbated by the poor water infrastructure and the fact that the highly virulent strain introduced to the nation, O1 biotype El Tor serotype Ogawa, had associated antibiotic resistance, the threat the disease posed to public health was large [17]. Presently, there have been well over 800,000 cumulative cases and almost 10,000 deaths in Haiti since the onset of the epidemic [1]. Unfortunately, the opportunity for using vaccination to prevent cholera from becoming a chronic problem has passed, but cholera's decade-long presence in Haiti has provided much more data for the study of the progression of cholera in a population. Therefore, continued efforts in modelling this epidemiological event may prove instructive in responses to future outbreaks.

There was a multitude of opinions on the best and most effective course of action to take to mitigate cholera's spread at the start of the epidemic and throughout the disease's continued presence in Haiti. Vaccination was considered a promising response as two killed oral cholera vaccines (OCVs) were available at the time. However, the duration and level of protection acquired by these vaccines is dependent upon dosage, national coverage, and the age of the vaccine recipient as young children shed immunity faster than adults [2]. Questions continue to arise when considering the logistics of vaccine distribution and administration: When is the best time to begin vaccination? Should vaccine delivery be prioritized according to location-based risk? How many vaccines must be administered for the optimal amount of protection, and are there enough vaccines currently available for this? With all of these questions in mind, it was, and still is, challenging to determine whether investment in widespread vaccination would be an effective measure against cholera.

One means by which to overcome these uncertainties is epidemiological modelling, a powerful method of analysis of disease dynamics that can be used to inform vaccination policy and public health decision-making. In December of 2010, the CDC's modelling study predicted that the impact of vaccination would be relatively minor with the number of vaccines available at the time [4]. However, a later study by Chao et al. found that even relatively low vaccination coverage (around 30% of the population) was effective at controlling the spread of the disease if the administration was informed by the relative risk and exposure of a given community [2]. Fung et al. also showed that OCVs used in conjunction with improved sanitation practices and water infrastructure was most effective at reducing the

number of cases [6]. This paper provides an analysis of a state-space model proposed by Elizabeth C. Lee, Andrew S. Azman, and Justin Lessler of the Johns Hopkins Bloomberg School of Public Health. We begin with an analysis of the methods and results of Lee et al. and go on to propose improvements to its implementation. Through this case study, we aim to exhibit the dependence of the utility and quality of simulation-based inference in epidemiological contexts upon model examination and refinement.

2 Background

In many epidemiological settings, statistical modelling can be of immense use in motivating public health decision-making. For example, modelling reported cases of a given infectious disease can help inform the subsequent actions taken to mitigate the disease's spread. Due to the inherent randomness and complexity of population dynamics, there are numerous different methods for modelling disease, all of which have advantages and disadvantages. Yet the contributions of epidemiological models to our understanding of how infectious diseases evolve in a population are of high value, and it is worth tackling the challenge of developing a good and useful model.

It is well established that state-space models are appropriate and effective models when studying environmental and biological processes. At its core, a state-space model has two components: an unobserved state process and a dependent observation process [9]. The ability to use a state-space model to inform policy and public action depends upon the model's quality, which itself is dependent upon the ease of statistical inference with respect to the model's parameters. Fortunately several methods have been developed to facilitate estimation of unknown parameter values, one of which is maximum likelihood via iterated filtering (MIF) proposed by Ionides et al., a variant of which is addressed in section 2.3 [8].

Compartment models are another standard tool for modelling infectious diseases. By dividing a population into compartments, for example as (S)usceptible, (I)nfectious, (R)ecovered in the standard SIR compartment model, the spread of a disease can be described with much more specificity because attention is given to all stages of host infection. However, it is nearly impossible to know how many individuals populate a compartment or are transitioning between compartments at a given time. To overcome this uncertainty, compartment models can be coupled with state-space models to form a comprehensive representation of a disease's progression in which parameters can be more easily estimated via inference methods for state-space models. In the next few sections, we provide a brief overview of the foundational

concepts needed to understand these models.

2.1 Time Series and Markov Processes

Consider a sequence of N time points, $t_{1:N} = \{t_1, t_2, \ldots, t_N\}$, and a sequence of N observations made at each time point, $y_{1:N} = \{y_{t_1}, y_{t_2}, \ldots, y_{t_N}\}$. We call $Y_{1:N}$ a time series model with jointly defined random variables $Y_n, \quad \forall n \in 1:N$, and we can conceive of the data, $y_{1:N}$, as one realization of $Y_{1:N}$ [14].

We then describe a time series model, $X_{1:N}$, where $X_n = X(t_n)$ is a random process at time $n, \quad \forall n \in 1:N$. Should this time series model satisfy the condition that its state at time $n+1$ is conditional only on its state at time n, $X_{1:N}$ is called a Markov process model. Mathematically, this can be represented as the following equation stating that the conditional density of the process X_n given the processes $X_{1:n-1}$ is equivalent to the conditional density given only the process X_{n-1} [5]:

$$f_{X_n|X_{1:n-1}}(x_n|x_{1:n-1}) = f_{X|X_{n-1}}(x_n|x_{n-1}) \tag{1}$$

2.2 Partially Observed Markov Processes

Often the details of the mechanisms underlying the evolution of a natural system are unknown. In epidemiology, the exact number of individuals exposed to disease at a given time is usually unknown. We can work around the issue of missing information using partially observed Markov (POMP) models. We create a POMP model by joining two processes, one that is unobservable (latent) but of interest and one that is observable and dependent upon the first.

Let the random variables $X_{1:N}$ represent the latent state process where X_1 serves to initialize the process model, $f_{X|X_{n-1}}(x_n|x_{n-1})$. With the random variables $Y_{1:N}$ representing the observable measurement process, the measurement model is $f_{Y_n|X_n}(y_n|x_n)$, and the collected data $y_{1:N}$ are observations of this process. We assume that each Y_n depends only upon the latent process at time n, X_n, and is conditionally independent of the other variables representing the measurement and latent processes, Y_m and $X_m, \quad \forall m \in 1:N, \quad m \neq n$ [15]. Together, $X_{1:N}$ and $Y_{1:N}$ form our POMP model.

2.3 Likelihood and Iterated Filtering

In problems of statistical inference, it is common to use likelihood to inform parameter estimation and model selection. Given a model parameterized by vector θ in the m-dimensional parameter space Θ_m, the likelihood function is the joint probability density of the data, $y_{1:N}$, at θ:

$$\mathcal{L}(\theta) = f_{Y_{1:N}}(y_{1:N}; \theta) \tag{2}$$

We then aim to find an estimate of θ, $\hat{\theta}$, which maximizes this function, $\mathcal{L}(\hat{\theta})$, or its natural logarithm, $\ell(\hat{\theta})$ [13].

The utility of an epidemiological model of disease spread is dependent upon its ability to be used for forecasting cases or incidence. This ability is itself dependent upon our confidence in the model's prediction accuracy and our understanding of the ways in which the latent states change with time. Thus, we have two linked problems: identifying the distribution of X_n at time n given $y_{1:n}$ and finding parameter values, $\hat{\theta}$, which maximize the likelihood of our data. These problems are known as the filtering problem and the inference problem, respectively [3, 13].

Especially in the case of highly complex environments, both the likelihood function and the transition density of a POMP model can be difficult to write analytically, making these two problems quite hard. Many methods have been developed to surmount the inference and filtering problems, one of which is the particle filter. For the particle filter, we need only supply data, simulators for the initial density and the one time-step transition density of the latent process, and an evaluator for the density of the observation process conditional on the latent process to get maximum likelihood estimates for the model parameters.

We first initialize a swarm of M particles at time 1, $\{X_1^m; m \in 1:M\}$, each containing the necessary state information along with a vector of parameter values, θ. Then for each time $n \in 1:N$, we push the particles forward one time-step by drawing from the one time-step transition density, giving us an ensemble of particles representing the prediction distribution at time n, $f_{X_n|X_{n-1}}(\cdot|X_{n-1}^m; \theta)$. We weight the particles according to our data by evaluating the measurement density, so $w_{n,m} = f_{Y_n|X_n}(y_n|x_n^m)$. Finally we resample the particles according to these weights, which leads to an ensemble of particles representing the filtering distribution at time n, $f_{X_n|Y_{1:n}}(x_n|y_{1:n}; \theta)$.

Because of the assumed independence of the measurement process variables and their

dependence upon the latent process variables in a POMP model, we have that:

$$\begin{aligned}
\mathcal{L}(\theta) &= f_{Y_{1:N}}(y_{1:N};\theta) \\
&= \prod_{n=1}^{N} f_{Y_n|Y_{1:n-1}}(y_n|y_{n-1};\theta) \\
&= \prod_{n=1}^{N} \int f_{Y_n|Y_{1:n-1},X_n}(y_n|y_{1:n-1},x_n;\theta) f_{X_n|Y_{1:n-1}}(x_n|y_{1:n-1};\theta) dx_n \\
&= \prod_{n=1}^{N} f_{Y_n|X_n}(y_n|x_n)
\end{aligned} \tag{3}$$

Notice that the weights used in the particle resampling are $w_{n,m} = f_{Y_n|X_n}(y_n|x_n^m)$ for each particle m at time n. If we take the average of $f_{Y_n|X_n}(y_n|x_n^m)$ over all M particles, we can approximate $f_{Y_n|X_n}(y_n|x_n;\theta)$. Therefore:

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} f_{Y_n|X_n}(y_n|x_n;\theta) \approx \prod_{n=1}^{N} \frac{1}{M} \sum_{m=1}^{M} f_{Y_n|X_n}(y_n|x_n^m) \tag{4}$$

In other words, by the Monte Carlo principle we can approximate the conditional likelihood at time n with $w_{n,m}$. Thus the particle filter provides a much easier way to estimate the likelihood of the data given our model and to approximate the distribution of X_n at time n given $y_{1:n}$ [11, 10, 8].

An extension of the particle filter is the improved iterated filtering algorithm (IF2) developed by Ionides et al. [9]. As a plug-and-play method, IF2 is a computationally efficient, simulation-based means for maximum likelihood estimation and inference. IF2 takes an initialized swarm of particles and, using a combination of particle filtering, small changes to the parameter values, and particle resampling, estimates the parameter values which achieve the maximum likelihood [10]. With the particle filter and IF2, we are able to approximate solutions to the inference and filtering problems.

3 Methodology

3.1 Model Structure

Lee et al. used an SEIAR compartmental model (S: Susceptible, E: Exposed, I: Infectious, A: Asymptomatic Infectious, R: Recovered) for the Haiti cholera epidemic. In their formulation, at a given time point t each compartment contains some unobserved number

of individuals from the total population of Haiti. Between two time points t and $t+1$, individuals can transition into the system by birth, out of the system by death, or between compartments at rates that are either specified or estimated. We define these transition rates with the following series of equations:

$$q_{S_k E_k} = \lambda(t) \tag{5}$$

$$q_{E_k I_k} = \sigma(1 - \theta_0(t)) \tag{6}$$

$$q_{E_k A_k} = \sigma\theta_0(t) \tag{7}$$

$$q_{I_k R_k} = q_{A_k R_k} = \lambda \tag{8}$$

$$q_{R_k S_k} = \alpha q_{S_0 S_k} = q_{E_0 E_k} = q_{I_0 I_k} = q_{A_0 A_k} = q_{R_0 R_k} = \eta_k(t) \tag{9}$$

$$q_{\cdot S_0} = \mu q_{S_k \cdot} = q_{E_k \cdot} = q_{I_k \cdot} = q_{A_k \cdot} = q_{R_k \cdot} = \delta \tag{10}$$

where $q_{X_k Y_k}$ indicates the one time-step transition rate from compartment X to compartment Y, and $k \in [0, 10]$ denotes vaccination cohort with $k = 0$ indicating the cohort that did not receive vaccinations. At time t, $\eta_k(t)$ is the vaccination rate of cohort k, and $\lambda(t)$ is the force of infection, calculated as $\lambda(t) = \frac{\beta(I(t)+(1-\kappa)A(t))^\nu}{N(t)}$. The seasonal transmission term is $\beta = \sum_{i=1}^{6} \beta_i s_i$, which consists of six degree six periodic B-spline terms, $s_{1:6}$, multiplied by estimated seasonality parameters, $\beta_{1:6}$. $I(t)$ is the proportion of the population that is infectious at time t, $A(t)$ is the proportion of the population that is asymptomatic at time t, $N(t)$ is the population of Haiti at time t with $N_0 = 10911819$, $\kappa = 0.95$ is the assumed reduction in infectiousness of asymptomatic individuals, ν is an estimated population mixing coefficient, and $\theta_0(t) = 0$ is the proportion of non-vaccinated, exposed individuals who become infected but are asymptomatic. Not dependent on time are $\frac{1}{\alpha} = 8$, the mean duration of natural immunity in years; $\frac{1}{\sigma} = 1.4$, the latent period of cholera in days; $\frac{1}{\gamma} = 2$, the infectious period of cholera in days; $\mu = 0.43$, the birth rate per 1000 individuals per week; and $\delta = 0.14$, the natural death rate per 1000 individuals per week. $q_{\cdot S_0}$ and $q_{X_k \cdot}$ denote the transition rates into and out of the system's compartments via birth and death, respectively. Below is a figure based upon the model diagram from Lee et al. [12]. It illustrates the compartmental model with one vaccination cohort. Transitioning out of the system due to death is omitted for legibility.

3.2 Reproduction

The preference for complex over simple models has been growing for several decades despite the fact that it has been shown that complexity is associated with decreases in forecast-

ing accuracy [7]. Because epidemiological modelling is motivated by the need to accurately forecast disease prevalence to inform policy, we first establish a point of comparison for the evaluation of our model fit and quality. We elect to use a linear, Gaussian autoregressive moving average (ARMA) model of order (2,1) as it is a fairly simple model in which the current state depends only on previous states and white noise [16]. We can then compare the likelihood of the data under this model to the likelihoods achieved under our proposed models to evaluate whether the additional complexity is truly beneficial.

Lee et al. divided the case data into two periods: epidemic (October 23rd, 2010 through March 31st, 2015) and endemic (April 1st, 2015 through January 12th, 2019) [12]. We adopted this breakpoint in our analyses. The ARMA(2,1) benchmark model achieved log-likelihoods of -1616.678, -1139.238, and -2800.808 for the epidemic, endemic, and the combined time period, respectively.

After establishing benchmark log-likelihoods, we attempted to reproduce the results of Lee et al. as closely as possible in order to facilitate the evaluation of their model and parameter estimates. Lee et al. implemented their model in the R package `pomp` v1.19 and started the model calibration by generating 300 different sets of starting parameter values. They then used trajectory matching followed by iterated filtering to find a maximum likelihood estimate for the parameter values using each of the 300 sets. From the epidemic calibration, they pruned away sets resulting in filtering failures or extreme outlying values. The remaining sets were used as starting values for the endemic calibration in which all parameters were reestimated, excluding the initial state values (E_0 and I_0) [12].

We repeated most of this process with some minor changes. We did not perform trajectory matching as it assumes a deterministic latent process, which is not assumed in the forecasting model. Additionally, Lee et al. did not publish their initial starting sets, so we created our own using the schema provided in their supplemental code. We left weeks with missing data as `NA` rather than `0` as the `pomp` package is capable of working with missing data. We also filtered out epidemic parameter sets with $\nu \leq 0.9$ and $\beta_1 \geq 100$ and endemic parameter sets with log-likelihoods of -3000 units or less to avoid outlying parameter values similar to Lee et al.'s pruning process. Our reproduction (fig. A1) does seem to visually match the results of Lee et al. in figure S7 of their supplement [12].

References

[1] Achieving coordinated national immunity and cholera elimination in Haiti through vaccination: a modelling study. *The Lancet Global Health*, 8(8):e1081–e1089, 2020.

[2] Dennis L. Chao, M. Elizabeth Halloran, and Ira M. Longini. Vaccination strategies for epidemic cholera in Haiti with implications for the developing world. *Proceedings of the National Academy of Sciences of the United States of America*, 108(17):7081–7085, 2011.

[3] Dan Crisan. The stochastic filtering problem: a brief historical account. *Journal of Applied Probability*, 51(A):13–22, 2014.

[4] Kashmira A. Date, Andrea Vicari, Terri B. Hyde, Eric Mintz, M. Carolina Danovaro-Holliday, Ariel Henry, Jordan W. Tappero, Thierry H. Roels, Joseph Abrams, Brenton T. Burkholder, Cuauhtémoc Ruiz-Matus, Jon Andrus, and Vance Dietz. Considerations for oral cholera vaccine use during outbreak after earthquake in Haiti, 2010-2011. *Emerging Infectious Diseases*, 17(11):2105–2112, 2011.

[5] Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov Chains*. Cham, 2018.

[6] Isaac Chun Hai Fung, David L. Fitter, Rebekah H. Borse, Martin I. Meltzer, and Jordan W. Tappero. Modeling the effect of water, sanitation, and hygiene and oral cholera vaccine implementation in Haiti. *American Journal of Tropical Medicine and Hygiene*, 89(4):633–640, 2013.

[7] Kesten C. Green and J. Scott Armstrong. Simple versus complex forecasting: The evidence. *Journal of Business Research*, 68(8):1678–1685, 2015.

[8] Edward L. Ionides, Anindya Bhadra, Yves Atchadé, and Aaron King. Iterated filtering. *Annals of Statistics*, 39(3):1776–1802, 2011.

[9] Edward L. Ionides, C. Bretó, and Aaron A. King. Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 103(49):18438–18443, 2006.

[10] Edward L. Ionides, Dao Nguyen, Yves Atchadé, Stilian Stoev, and Aaron A. King. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences*, 112(3):719 LP – 724, jan 2015.

[11] Aaron A. King, Dao Nguyen, and Edward L. Ionides. Statistical inference for partially observed markov processes via the R package pomp. *Journal of Statistical Software*, 69:1–43, 2016.

[12] Elizabeth C. Lee, Dennis L. Chao, Joseph C. Lemaitre, Laura Matrajt, Damiano Pasetto, Javier Perez-Saez, Flavio Finger, Andrea Rinaldo, Jonathan D. Sugimoto, M. Elizabeth Halloran, Ira M. Longini, Ralph Ternier, Kenia Vissieres, Andrew S. Azman, Justin Lessler, and Louise C. Ivers. Supplementary appendix 3: Achieving coordinated national immunity and cholera elimination in Haiti through vaccination: a modelling study. *The Lancet Global Health*, 8(8):e1081–e1089, 2020.

[13] Russell B. Millar. *Maximum Likelihood Estimation and Inference: with exmaples in R, SAS, and ADMB*. John Wiley & Sons, Ltd., 2011.

[14] Robert H. Shumway and David S. Stoffer. Time Series Analysis and Its Applications. In George Casella, Stephen Fienberg, and Ingram Olkin, editors, *Time Series Analysis and Its Applications*, chapter Characteristics of Time Series. Springer Science+Business Media Inc., New York, second edition, 2006.

[15] Robert H. Shumway and David S. Stoffer. Time Series Analysis and Its Applications. In George Casella, Stephen Fienberg, and Ingram Olkin, editors, *Time Series Analysis and Its Applications*, chapter State-Space Models. Springer Science+Business Media Inc., New York, second edition, 2006.

[16] Robert H. Shumway and David S. Stoffer. Time Series Analysis and Its Applications. In George Casella, Stephen Fienberg, and Ingram Olkin, editors, *Time Series Analysis and Its Applications*, chapter ARIMA Models. Springer Science+Business Media Inc., New York, second edition, 2006.

[17] John Zarocostas. Cholera outbreak in Haiti-from 2010 to today. *Lancet (London, England)*, 389(10086):2274–2275, 2017.