Estimating the Attack Ratio of Dengue Epidemics under Time-varying Force of Infection using Aggregated Notification Data

Flavio Codeço Coelho[1*], Luiz Max de Carvalho[2],
1 Escola de Matemática Aplicada , Fundação Getulio Vargas (FGV), Rio de Janeiro – RJ, Brazil.
2 Programa de Computação Científica (PROCC), Fundação Oswaldo Cruz, Rio de Janeiro – RJ, Brazil.
∗ **E-mail: fccoelho@fgv.br**

Abstract

Quantifying the attack ratio of disease is key to epidemiological inference and public health planning. For multi-serotype pathogens, however, different levels of serotype-specific immunity make it difficult to assess the population at risk. In this paper we propose a Bayesian method for estimation of the attack ratio of an epidemic and the initial fraction of susceptibles using aggregated incidence data. We derive the probability distribution of the effective reproductive number, R_t , and use MCMC to obtain posterior distributions of the parameters of a single-strain SIR transmission model with time-varying force of infection. Our method is showcased in a data set consisting of 18 years of dengue incidence in the city of Rio de Janeiro, Brazil. We demonstrate that it is possible to learn about the initial fraction of susceptibles and the attack ratio even in the absence of serotype specific data. On the other hand, the information provided by this approach is limited, stressing the need for detailed serological surveys to characterise the distribution of serotype-specific immunity in the population.

Introduction

Dengue is an arthropod-borne febrile disease caused by a flavivirus with four serotypes (DEN-1, 2, 3 and 4) which causes an estimated 50 million infections each year [1]. In humans, immunity against a particular serotype is considered permanent after the exposure and cross immunity to other serotypes is considered short lived [2] although some studies argue for a longer duration of cross-immunity [3, 4]. As a consequence, the proportions of viral serotypes co-circulating at any point in time are strongly dependent on previous incidence patterns of the disease, which determine the number of individuals susceptible to each serotype at any point in time.

Dengue transmission is also modulated by environmental conditions, among which, temperature, due to its effects on the vector reproduction, stands out as a strong predictor of incidence [5, 6]. In places with sufficient seasonal temperature variation, dengue is predominantly a summer disease. So it is fair to say that these environmental fluctuations play a key role in determining beginning and end of epidemic periods. This climatic influence is exerted mainly through its effects on the force of infection, which cannot be taken as constant [7] but rather as a seasonal (oscillating) function of time. The long term dynamics of dengue is also modulated by the alternation of virus types in circulation. Demographics also plays a role in replenishing the population of susceptibles.

The attack ratio (AR) of a disease is a measure of morbidity defined as the number of new cases divided by the population at risk. For dengue epidemics, it can be difficult to calculate the AR due to the lack of knowledge of the population at risk. The population at risk in this case is the number of susceptibles to the circulating virus type(s) before a given epidemic. Thus, in order to calculate the attack ratio, we need to determine the number of susceptibles to the circulating virus types right before the epidemic, which is virtually impossible without regular virological surveys.

The attack ratio is also influenced by the reproductive number of the disease [8, 9], which is closely associated with the force of infection. Thus the incorporation of the effective reproduction number, R_t , as a function of time, is crucial to an accurate estimation of the AR of seasonal diseases like dengue and Influenza.

Other methods for estimating the number of susceptibles while accommodating time-varying force of infection have been proposed before, for measles [10, 11], a disease that shows remarkable seasonality. These methods try to reconstruct the entire series of infectious and susceptibles from case data using deterministic models and generally work well for measles because there is a one-to-one relationship between exposure and immunity, since measles is caused by a single-strain pathogen. Recently, methods in the same fashion were developed for dengue when serotype-specific data is available [12]. When such data is not available, the series of susceptibles to all possible serotypes, cannot be reconstructed based solely on a deterministic transmission model, since the arrival/re-emergence of new serotypes, an intrinsically stochastic event, can drastically change the pool of susceptibles, throwing off any sequential estimation based on the incidence dynamics.

Dengue has been reintroduced in Rio de Janeiro in 1986 after being absent for 68 years [13]. During the last decades of the 20th century only DEN-1 and DEN-2 serotypes were in circulation. The remaining serotypes DEN-3 and DEN-4, arrived respectively in 2000 and 2011 [14, 15]. Due to this patterns of recent re-introduction of the disease, its incidence dynamics is still dominated by the introduction events and environmental determinants of transmission.

In this paper, we propose a new approach to estimate the number (fraction) of susceptibles using a simplified model of dengue transmission based on a single-strain Susceptible-Infectious-Removed (SIR) model with time-varying infection rate. In order to bypass the limitations of not knowing the serotype-specific seroprevalence and the exact behaviour of the force of infection through time, we propose to inform the time-varying transmissibility using the R_t series derived from the notification data [16]. We extend a Bayesian framework previously used to estimate the number of susceptibles in Influenza epidemics in Europe [17] to include time-varying force of infection and derive a probability distribution for R_t to accommodate uncertainty in the estimates. Then, from the incidence series and the population at risk, we calculate the attack ratio for each epidemic. We apply our method to estimate S_0 before every major dengue epidemic in the city of Rio de Janeiro, Brazil in the last 18 years.

Methods

In this section we will start by describing the data and then the method used to estimate the effective reproductive number, R_t, from the data and obtain its posterior distribution. We then proceed to describe the Susceptible-Infectious-Recovered (SIR) model used to represent the aggregated disease incidence and how R_t can be integrated into the model to allow for time varying force of infection. Next, an approach to approximate the posterior distributions of the numbers of susceptible to the main circulating dengue viruses for each epidemic is detailed. Finally, we discuss how to estimate the attack ratio of each epidemic using the estimated susceptible fraction and the observed incidence.

Data

The data used in this paper consists of time series of weekly notified cases of dengue for the city of Rio de Janeiro from 1996 to 2014. The cases are notified based only on clinical criteria. Laboratory confirmation and serotype information are available only for a very small sample and only on recent years (2010-2013). For the parameter estimation procedures incidence was normalized by dividing the number of cases reported by the total city population at each year as given by the census (Census Bureau, Brazilian Institute of Geography and Statistics, `http://www.ibge.gov.br/english/`).

Estimating the effective reproductive number (R_t)

In monitoring of infectious diseases, it is important to assess whether the incidence of a particular disease is increasing significantly, in order to decide to take preventive measures. The effective reproductive number at time t, R_t , can be understood as a real-time estimate of the basic reproductive number (R_0) and is defined as the average number of secondary cases per primary case at time t.

Let Y_t be the number of reported disease cases for a particular time $t \in (0, T)$. Nishiura el al. (2010) [16] extend the theory developed by Stallybrass et al. (1931) [18] and propose to estimate R_t as

$$R_t = \left(\frac{Y_{t+1}}{Y_t}\right)^{1/n} \tag{1}$$

where n is taken to be the ratio between the length of reporting interval and the mean generation time of the disease. Here we are interested in the simpler case $n = 1$. If R_t is to be used as a decision tool, however, one needs to be able to quantify the uncertainty about estimate in equation 1. Here we detail how to obtain credibility intervals for R_t under the assumption that the counts Y_t are Poisson distributed for all t.

We explore the approach of Ederer and Mantel [19], whose objective was to obtain confidence intervals for the ratio of two Poisson counts. Let $Y_t \sim Poisson(\lambda_t)$ and $Y_{t+1} \sim Poisson(\lambda_{t+1})$ and define $W = Y_t + Y_{t+1}$. The authors note that by conditioning on the sum W

$$Y_{t+1}|S \sim Binomial(W, \theta_t) \tag{2}$$

$$\theta_t = \frac{\lambda_{t+1}}{\lambda_t + \lambda_{t+1}} \tag{3}$$

Let $c_\alpha(\theta_t) = \{\theta_t^{(L)}, \theta_t^{(U)}\}$ be such that $Pr(\theta_t^{(L)} < \theta_t < \theta_t^{(U)}) = \alpha$. Analogously, define $c_\alpha(R_t) = \{R_t^{(L)}, R_t^{(U)}\}$ such that $Pr(R_t^{(L)} < R_t < R_t^{(U)}) = \alpha$. Ederer and Mantel (1974) [19] show that one can construct a $100\alpha\%$ confidence interval for R_t by noting that

$$R_t^{(L)} = \frac{\theta_t^{(L)}}{(1 - \theta_t^{(L)})} \quad \text{and} \quad R_t^{(U)} = \frac{\theta_t^{(U)}}{(1 - \theta_t^{(U)})} \tag{4}$$

Because the transform from θ to R_t is monotonically increasing, the result holds for confidence and credibility intervals alike.

Many authors have chosen to quantify the uncertainty about θ following orthodox approaches (see for example [20] and [21]) mainly for simplicity. We choose instead to take a Bayesian approach and use the $100\alpha\%$ posterior credibility interval for θ_t as $c_\alpha(\theta_t)$. If we choose the conjugate beta prior with parameters a_0 and b_0 for the binomial likelihood in (2), the posterior distribution for θ_t is

$$p(\theta_t|Y_{t+1}, W) \sim Beta(Y_{t+1} + a_0, Y_t + b_0) \tag{5}$$

Combining equations (4) and (5) tells us that the induced posterior distribution of R_t is a beta prime (or inverted beta) with parameters $a_1 = Y_{t+1} + a_0$ and $b_1 = Y_t + b_0$ [22]. The density of the induced distribution is then

$$f_P(R_t|a_1, b_1) = \frac{\Gamma(a_1 + b_1)}{\Gamma(a_1)\Gamma(b_1)} R_t^{a_1 - 1}(1 + R_t)^{-(a_1 + b_1)} \tag{6}$$

Thus, the expectation of R_t is $a_1/(b_1 - 1)$ and its variance is $a_1(a_1 + b_1 - 1)/\left((b_1 - 2)(b_1 - 1)^2\right)$. Note that this result holds only for $n = 1$. Sampling from the posterior in (6) can be made straightforward by first sampling from (5) and then applying the transform in (4). Also, one can choose a_0 and b_0 so as to elicit meaningful prior distributions for R_t. We show how to elicit the prior for R_t from specified prior mean and variance or coefficient of variation in the **Appendix**.

Also, since $R_t > 1$ indicates sustained transmission, one may be interested in computing the probability of this event. This can be easily achieved by integrating (6) over the appropriate interval. By noting that

$$Pr(R_t > 1) = 1 - \int_0^1 f_P(r)dr \tag{7}$$

$$= 1 - Pr(\theta_t < \frac{1}{2}) \tag{8}$$

one can compute the desired probability while avoiding dealing with the density in (6) directly.

Mathematical modelling

A Susceptible-Infectious-Removed (SIR) model is proposed to model dengue dynamics. In the traditional formulation of the model, transmission is governed by a constant transmission rate β and recovery happens at a rate τ.

For our analysis we chose to let the force of infection vary with time, just as it does in the actual epidemics, as seen in the data. So as the epidemic progresses, the effective transmission rate changes and is given by

$$\beta(t) = \frac{R_t \cdot \tau}{S_0} \tag{9}$$

where R_t is the effective reproductive number, estimated as in 1 and S_0 is the initial fraction of susceptible individuals. The complete model with the time-varying force of infection is given by the system of ordinary differential equations:

$$\begin{aligned} \frac{dS}{dt} &= -\beta(t)SI \\ \frac{dI}{dt} &= \beta(t)SI - \tau I \\ \frac{dR}{dt} &= \tau I \end{aligned} \tag{10}$$

where $S + I + R = 1 \,\forall\, t$. Of course, this is a rather simplified model, in which, for instance, the vector is omitted. The rationale for this simplification is based on the ability of the empirically derived R_t to incorporate the effects of the fluctuating vector populations. Although demography is not included in the model – population is assumed to remain constant throughout an epidemic – the population variation from year to year is taken into consideration as the prevalences are calculated for each year by dividing the number of cases by the official population estimated by the census. In Brazil, dengue affects all age groups which still haven't been exposed to all 4 serotypes. This is in contrast with countries in which the four serotypes have been endemic for a very long period of time, where dengue mostly affect the youth.

Also, although there are multiple circulating serotypes, our approach can not discriminate between them due to the lack of serotype-specific data. Nevertheless, this modelling strategy can still provide some insight into the disease dynamics and allows us to estimate the initial fraction of susceptibles S_0, a key epidemiological parameter.

Bayesian parameter estimation

We take a Bayesian approach to the estimation of S_0. First the incidence time series was divided into $J = 13$ epidemic windows that corresponded to significant raises in incidence and normalized to lie on the $[0, 1]$ interval. For a given interval $j = \{t_j^{\text{start}}, t_j^{\text{end}}\}$ we observe an incidence time series $\mathbf{Y_j}$. We are thus interested in the posterior distribution

$$p(S_{0j}|\mathbf{Y_j}) \propto l(\mathbf{Y_j}|S_{0j}, R_t, \tau)\pi(S_{0j}) \tag{11}$$

The likelihood $l(\mathbf{Y_j}|\cdot)$ is a approximated by a Normal distribution with fixed variance σ^2. This approximation is a numerical convenience since it confers better stability to the MCMC sampling. In this estimation procedure we kept R_t fixed at fixed at the posterior mean obtained as described above and fixed $\tau = 1/7\,\text{days}^{-1}$. To complete the inference, we need to specify prior distributions for the parameters of interest. We place a flat Beta$(1, 1)$ prior on $S_{0j}\ \forall j$

To approximate the posterior in (11) we use Markov chain Monte Carlo techniques implemented in the Bayesian inference with Python (BIP) [17] available at `http://code.google.com/p/bayesian-inference/`. BIP uses a Differential Evolution Adaptive Metropolis (DREAM) [23] scheme that efficiently samples from high-dimensional joint distributions using multiple adaptive chains running in parallel with delayed rejection (see **Appendix** for details). Also, as the numerical integration routine implemented within BIP needs $\beta(t)$ to be available at arbitrary values of t, i.e., as continuous function of time because of the variable step size, we used linear interpolation to obtain values of R_t for any time point. In this study we used one chain per parameter, i.e, 3 chains for each run. The chains were run until 5000 samples were obtained after discarding 500 burn-in samples. Convergence of the parallel chains was verified at every 100 iterations by the calculation of the Gelman-Rubin's R (potential scale reduction factor), which approaches 1 at convergence [24].

Calculating the attack ratio

The attack ratio of an epidemic is defined by the number of infections divided by the size of the population at risk.

$$A = \frac{\#\text{ cases}}{\text{Population at risk}} \tag{12}$$

Based on what has been discussed so far, we can rewrite (12) for each epidemic j as

$$A_j = \frac{\sum Y_j}{S_{0j}} \tag{13}$$

where S_{0j} is the number of susceptibles before each epidemic j, which we estimated before.

Python and R code to perform all the analyses described above is publicly available at `https://github.com/fccoelho/paperLM1`.

Results and Discussion

In this paper we propose a method to bypass the lack of serotype-specific case data by informing the time-varying force of infection with the instantaneous reproductive number, R_t which we calculate from aggregated data. The main contribution of this paper can be summarized in the following items: (i) we develop a method to quantify uncertainty about R_t that is readily applicable to other diseases and; (ii) R_t is used to inform a dynamic epidemic model with time-varying force of infection in order to gain insight into the attack ratio of each epidemic; (iii) we propose an estimation procedure for circulating serotype's S_0 from aggregate case data, which is robust to epidemic sizes; (iv) AR estimates are provided for 18 years of Dengue epidemics in Rio de Janeiro, Brazil.

Figure 1 shows the R_t series, according to Equation 1 [16] along with the confidence bands derived in this paper. It can be seen that the inter-epidemic periods are characterized by R_t being indistinguishable from 1. Due to the intrinsic variability of the R_t series, the examination of its credible intervals is essential to identify periods of sustained transmission. The wider intervals between epidemics are due to the scarcity of cases during these periods. credibility intervals, and therefore offers protection against false alarms (see the section on tail behaviour in the **Appendix** for a detailed explanation).

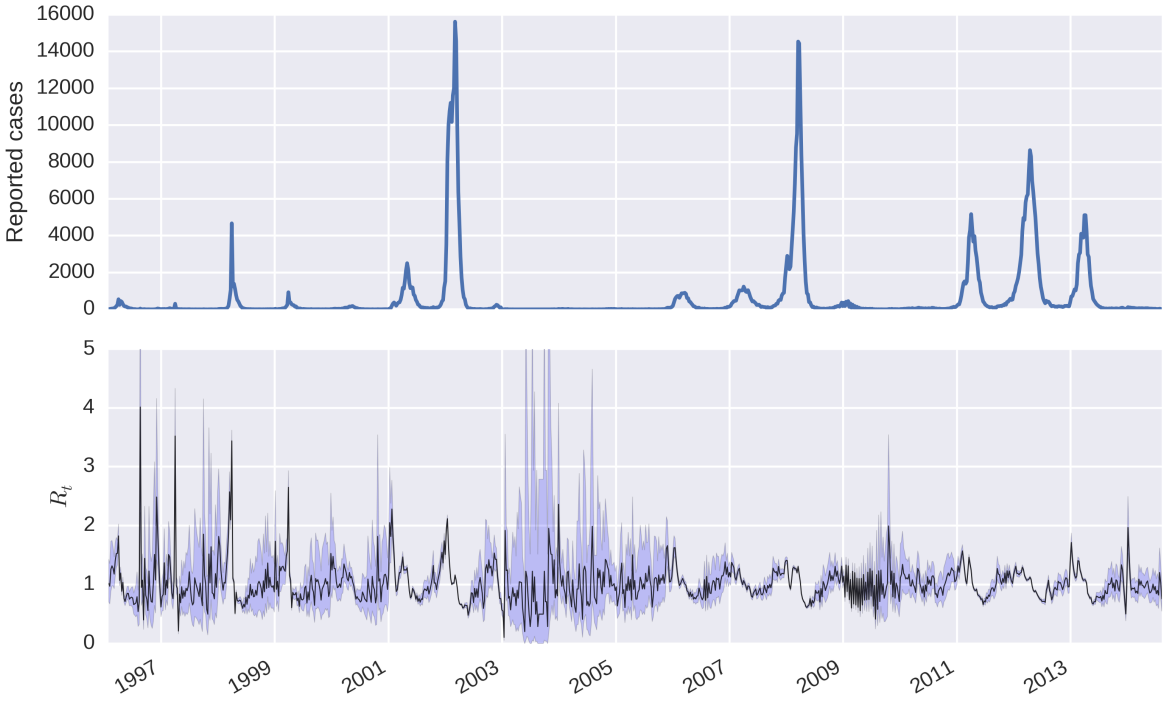


Figure 1. Estimated time-series for R_t, along with 95% credible intervals. Top panel show reported cases from which R_t is estimated. As expected, uncertainty about R_t is greater when the case counts are low, for instance in the period 2003–2006, which represented a big hiatus between major epidemics. The intrinsic variability of R_t can be used to inform the time-varying force of infection, since it reflects variation in the vector population and other environmental factors such as temperature and seasonal variation.

A key epidemiological quantity is the attack ratio (AR) of an epidemic, a measure of morbidity and speed of spread which can be used to predict epidemic size and help efficient Public Health planning. The AR depends fundamentally on the population at risk, which in the case of dengue is every naive (to a particular serotype) individual in the population. Estimating the initial susceptible fraction S_0 for each epidemic is thus central to the estimation of the AR. Methods for estimating the number of susceptibles have been proposed before, for other diseases [10, 11]. These methods attempt to reconstruct the entire series of infectious and susceptibles for measles outbreaks from case data. In the case of dengue, the full (multi-year/multi-epidemic) series of susceptibles to all possible serotypes, cannot be reconstructed based solely on a deterministic transmission model, since the arrival/re-emergence of new serotypes (which are a stochastic events) can change drastically the pool of susceptibles throwing off any sequential estimation based on the incidence dynamics.

Since there is very limited information regarding the actual proportions of each virus in circulation and most information available is about the predominant serotypes for some epidemics in the period of study only [25], we propose the use of a simplified a single strain model. The main argument we put

forward is that by conditioning on the R_t series, we implicitly take into account the variability introduced by the co-circulation of multiple serotypes and heterogeneous levels of immunity in the general population. We sought to deal with all important sources of uncertainty impinging on the estimation of the AR of a dengue epidemic, but not all could be satisfactorily addressed in this analysis. For instance, in any given epidemic there is a large number of mild and asymptomatic cases, which nevertheless acquire immunity. It is estimated that for every case reported, up to 10-20 are not seen by health authorities [26]. Another source of uncertainty is under-reporting of diagnosed cases, which is a serious issue in the health care systems of many developing countries such as Brazil. Duarte and França (2006) [27], estimated the sensitivity of Dengue reporting for hospitalized patients in Belo-Horizonte, Brazil to be of 63%, meaning that approximately 37% of the suspected Dengue cases go unreported. Lastly, demography and migrations affect the number of susceptible in ways which are not easy to fully determine.

Figure 2, shows the model from (10) fitted to the data. Despite its limitations, our simplified model fits the data well. In it we can see that the susceptibles series in each epidemic starts at the estimated level of S_0. The proportion of susceptibles may seem low, but we must remember that these estimates are being affected by an unknown under-reporting factor, which experts suggest is somewhere between 5 and 10, i.e. for every case observed there are 5 or 10 unobserved. Since this under-reporting affects both the numerator and denominator of (13), its effects should cancel out, giving us an unbiased attack ratio estimate. One other possible source of bias which would lead to the underestimation of S_0 could come from a significant part of the population not being exposed to the disease. However, as we can see in Figure 3, despite the differences in intensity (incidence), the entire city seems to be at risk, with no particularly "protected" areas, at least in the last four epidemics.

Table 1 contains the attack ratios and medians of the S_0 estimated for each epidemic/outbreak. Underreporting of cases, which is known to exist but of which exact figures cannot be determined, will lead to underestimation S_0. However, the attack ratio shall remain unbiased as the underreporting affects both the numerator and the denominator of equation 13. It is interesting to notice that the larger epidemics, in terms of peak size are not the one with the greater attack ratios. This stresses the importance of knowing the immunological structure of the population. Knowing the S_0 for the circulating viruses we can order to more accurately assess the potential impact of a coming epidemic, since particularly virulent types, can be rendered less of a threat by a low S_0. Honório et al.[28] conducted a serological survey in three separate localities within the city, right before the 2008 epidemic, the authors report seroprevalences varying from 56-77.4% which is compatible with our prediction of 87.5% $(1 - S_{0,2007})$ for the entire city, considering that we underestimate S_0 due to the underreporting of cases.

We hope that the results presented in paper will motivate public health authorities to invest in annual serological surveys, to determine the susceptibility profile to each dengue virus as well as to estimate the under-reporting factor of the notification system.

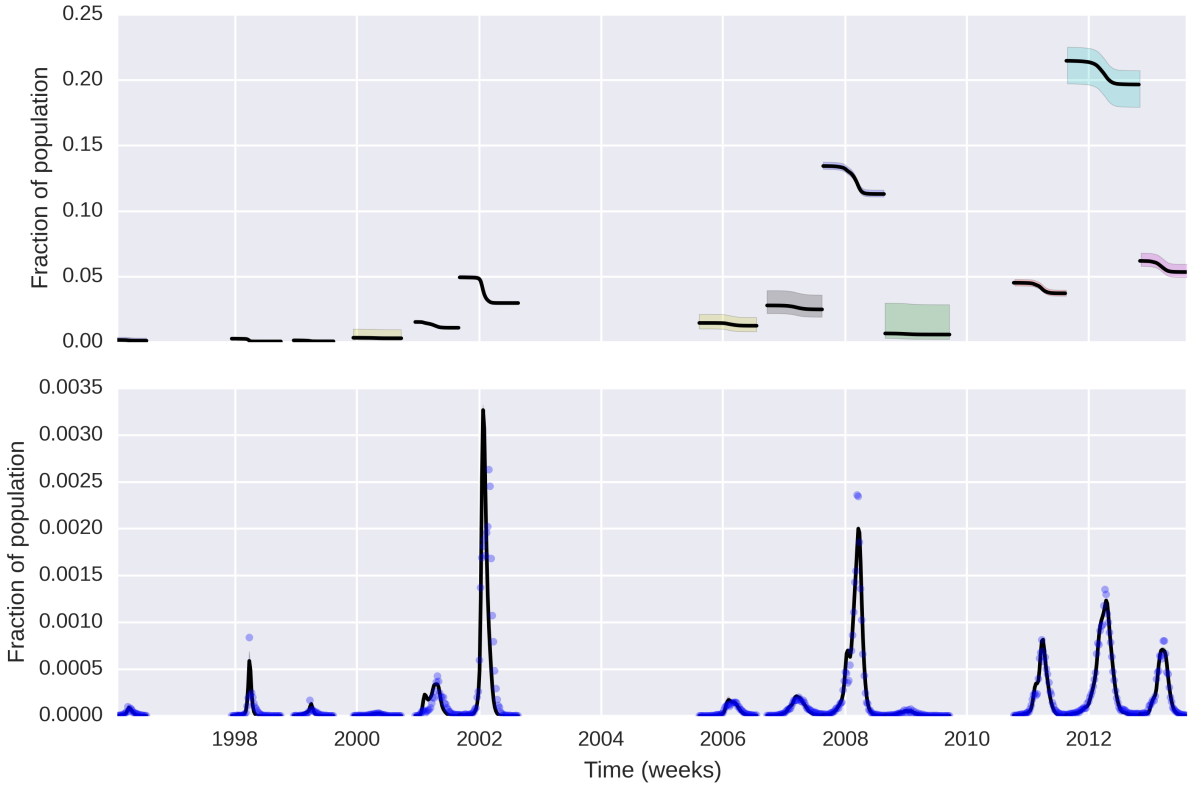
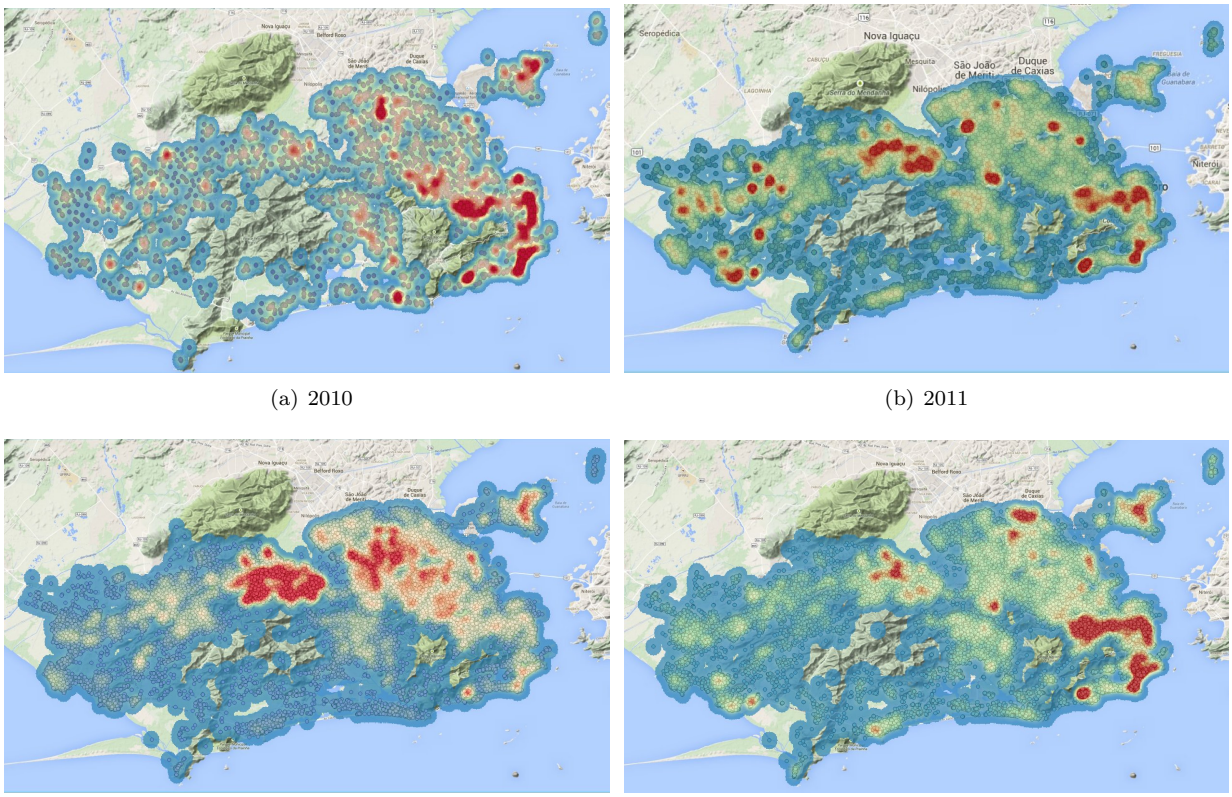


Figure 2. Susceptibles and Infectious posterior curves. The curves were estimated only for the periods where $R_t > 1$. The susceptible curves in the top panel reflect the prevalence of fraction of susceptibles to circulating strain(s) for each epidemic/outbreak. In the lower panel, we see the posterior distribution of infectious curves, represented by its median and 95% credible interval. Credible intervals are very narrow, and can be hard to distinguish from the median line. Dots show the observed cases, scaled as fractions of the entire population.

Acknowledgements

LMC is grateful to Dr. Leonardo Bastos for useful discussions on the posterior inference for R_t . The authors are also grateful to Claudia T. Codeço for helpful discussions about the manuscript.



(a) 2010 (b) 2011

(c) 2012 (d) 2013

Figure 3. Maps showing the incidence of dengue in the city of Rio de Janeiro from 2010 to 2013. Circles indicate individual notified cases. A heatmap is overlayed on the maps showing absolute density of cases. It can be seen that several areas of the city were affected and no region seems to be free of transmission risk. This suggests that although transmission risk varies spatially, there is significant exposure over the entire city.

Table 1. Median attack ratio and 95% credibility intervals calculated according to (13). Values are presented as percentage of total population. †: Year corresponds to the start of the epidemic, but the peak of cases may occur in the following year. ‡: percentage of population. These results show considerable variation in AR between epidemics, consistent with the accquiring and loss of serotype-specific immunity.

†Year	‡Cases	median Attack Ratio	‡S_0	Circulating Serotypes
1996	0.066	0.39 (0.17-0.54)	0.171(0.12-0.38)	DEN-1 and 2[13]
1997	0.238	0.87 (0.74-0.87)	0.273(0.27-0.32)	DEN-1 and 2[13]
1998	0.0708	0.5 (0.49-0.5)	0.142(0.14-0.14)	DEN-1 and 2[13]
1999	0.0371	0.11 (0.037-0.2)	0.345(0.18-1.0)	DEN-1, 2 and 3[14]
2000	0.394	0.25 (0.24-0.27)	1.55(1.5-1.6)	DEN-1, 2 and 3[14]
2001	2.38	0.48 (0.47-0.49)	4.95(4.8-5.1)	DEN-1, 2 and 3[14]
2005	0.217	0.15 (0.1-0.21)	1.47(1.0-2.1)	
2006	0.315	0.11 (0.08-0.14)	2.81(2.2-3.7)	DEN-2[15]
2007	2.03	0.15 (0.15-0.15)	13.5(13.0-14.01)	DEN-2[15]
2008	0.0923	0.14 (0.031-0.31)	0.672(0.3-2.4)	DEN-2[15]
2010	0.831	0.18 (0.17-0.19)	4.54(4.3-4.8)	DEN-2[15]
2011	1.85	0.086 (0.082-0.094)	21.5(20.0-23.0)	DEN-1, 2 and 4[15]
2012	0.864	0.14 (0.13-0.15)	6.21(5.8-6.8)	

References

1. Guzman, M. G. *et al.* Dengue: a continuing global threat. *Nat. Rev. Microbiol.* **8**, 7–16 (2010).

2. Halstead, S. B. Dengue. *The Lancet* **370**, 1644–1652 (2007). URL `http://www.sciencedirect.com/science/article/pii/S0140673607616870`.

3. Reich, N. G. *et al.* Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. *J R Soc Interface* **10**, 20130414 (2013).

4. Salje, H. *et al.* Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. *Proceedings of the National Academy of Sciences* **109**, 9535–9538 (2012).

5. Honório, N. A., Codeço, C. T., Alves, F. C., Magalhães, M. & Lourenço-De-Oliveira, R. Temporal distribution of aedes aegypti in different districts of rio de janeiro, Brazil, measured by two types of traps. *Journal of Medical Entomology* **46**, 1001–1014 (2009). URL `http://www.bioone.org/doi/abs/10.1603/033.046.0505`.

6. Wu, P.-C. *et al.* Higher temperature and urbanization affect the spatial patterns of dengue fever transmission in subtropical taiwan. *Science of The Total Environment* **407**, 2224–2233 (2009). URL `http://www.sciencedirect.com/science/article/pii/S0048969708011509`.

7. Reiner, R. C. *et al.* Time-varying, serotype-specific force of infection of dengue virus. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E2694–E2702 (2014). URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4084484/`.

8. Bacaër, N. & Gomes, M. G. M. On the final size of epidemics with seasonality. *Bulletin of Mathematical Biology* **71**, 1954–1966 (2009). URL `http://link.springer.com/article/10.1007/s11538-009-9433-7`.

9. Katriel, G. & Stone, L. Attack rates of seasonal epidemics. *Mathematical Biosciences* **235**, 56–65 (2012). URL `http://www.sciencedirect.com/science/article/pii/S0025556411001532`.

10. Bjrnstad, O., Finkenstdt, B. & Grenfell, B. Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs* **72**, 169–184 (2002).

11. Wallinga, J., Teunis, P. & Kretzschmar, M. Reconstruction of measles dynamics in a vaccinated population. *Vaccine* **21**, 2643–2650 (2003).

12. Reiner, R. C. *et al.* Time-varying, serotype-specific force of infection of dengue virus. *Proc. Natl. Acad. Sci. U.S.A.* **111**, E2694–2702 (2014).

13. Nogueira, R. M. R. *et al.* Dengue in the state of rio de janeiro, brazil, 1986-1998. *Memórias do Instituto Oswaldo Cruz* **94**, 297–304 (1999).

14. De Simone, T. *et al.* Dengue virus surveillance: the co-circulation of DENV-1, DENV-2 and DENV-3 in the state of Rio de Janeiro, Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **98**, 553–562 (2004).

15. Nogueira, R. M. & Eppinghaus, A. L. Dengue virus type 4 arrives in the state of Rio de Janeiro: a challenge for epidemiological surveillance and control. *Memórias do Instituto Oswaldo Cruz* **106**, 255–256 (2011).

16. Nishiura, H., Chowell, G., Heesterbeek, H. & Wallinga, J. The ideal reporting interval for an epidemic to objectively interpret the epidemiological time course. *J R Soc Interface* **7**, 297–307 (2010).

17. Coelho, F. C., Codeço, C. T. & Gomes, M. G. A Bayesian framework for parameter estimation in dynamical models. *PLoS ONE* **6**, e19616 (2011).

18. Stallybrass, C. O. *et al.* The principles of epidemiology and the process of infection. *The Principles of Epidemiology and the Process of Infection.* (1931).

19. Ederer, F. & Mantel, N. Confidence limits on the ratio of two poisson variables. *American Journal of Epidemiology* **100**, 165–167 (1974).

20. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212 (1927).

21. Clopper, C. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 404–413 (1934).

22. Dubey, S. D. Compound gamma, beta and f distributions. *Metrika* **16**, 27–31 (1970).

23. Vrugt, J. A. *et al.* Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation* **10**, 271–288 (2008).

24. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics* **7**, 434–455 (1998).

25. Macedo, G. A. *et al.* Virological surveillance for early warning of dengue epidemics in the state of Rio de Janeiro, Brazil. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **107**, 141–146 (2013).

26. Luz, P. M., Grinsztejn, B. & Galvani, A. P. Disability adjusted life years lost to dengue in Brazil. *Tropical Medicine & International Health* **14**, 237–246 (2009). URL `http://onlinelibrary.wiley.com/doi/10.1111/j.1365-3156.2008.02203.x/abstract`.

27. Duarte, H. H. P. & França, E. B. Data quality of dengue epidemiological surveillance in belo horizonte, southeastern Brazil. *Revista de Saúde Pública* **40**, 134–142 (2006). URL `http://www.scielosp.org/scielo.php?script=sci_abstract&pid=S0034-89102006000100021&lng=en&nrm=iso&tlng=en`.

28. Honório, N. A. *et al.* Spatial evaluation and modeling of dengue seroprevalence and vector density in Rio de Janeiro, Brazil. *PLoS Negl Trop Dis* **3**, e545 (2009).

29. Mira, A. On metropolis-hastings algorithms with delayed rejection. *Metron* **59**, 231–241 (2001).

Appendix

A remark on prior distributions and tail behaviour of the distribution of R_t

There are a number of approaches to deriving the distribution of R_t . Alternatively to the approach described in the main text [19], one could use the conditional distribution of R_t on Y_{t+1} and Y_t as defined in equation A7 of Nishiura et al. [16]:

$$f_R(R_t) = (Y_t R_t)^{Y_{t+1}} e^{-Y_t R_t} \tag{14}$$

Noticing the kernel of (14) is that of a gamma distribution with $a_2 = Y_{t+1} + 1$ and $b_2 = Y_t$, we obtain a proper density from which to construct $c_\alpha(R_t)$, simply by computing the appropriate quantiles of said distribution. This density is

$$f_N(R_t|a_2, b_2) = \frac{b_2^{a_2}}{\Gamma(a_2)} R_t^{a_2-1} e^{-b_2 R_t} \tag{15}$$

In order to decide which approach to take, it may be of use analysing the tail behaviour of the derived distributions for R_t . Consider the case of using a flat Uniform$(0, 1)$ prior for θ_t. With $a_0 = b_0 = 1$, $a_1 = a_2$ and $b_1 = b_2 + 1$. The beta prime (inverse beta distribution) will have heavier tails compared to the conditional distribution proposed by [16], thus providing more conservative confidence/credibility intervals. To see that one needs simply take the ratio of the Beta prime and Gamma (unnormalized) densities and evaluate the limit as R_t goes to infinity:

$$\lim_{R_t \to \infty} \frac{f_P(R_t|a_1, b_1)}{f_N(R_t|a_2, b_2)} = \lim_{R_t \to \infty} \frac{e^{Y_t R_t}}{(1 + R_t)^{Y_t + Y_{t+1} + 2}} = \infty \tag{16}$$

Note also that we deliberately construct $c_\alpha(R_t)$ as a equal-tailed $100\alpha\%$ credible set, rather than a less conservative highest posterior density (HPD) interval. Finally, we performed a simple simulation study to assess the result in equation 16. The simulations were carried out as follows:

1. Create a two-dimensional grid of values for λ_1 and λ_2, with rate values from 2 to 1000;

2. For each point $(\lambda_1^{(j)}, \lambda_2^{(j)})$ in the grid,

   - Generate 1000 realisations of the random variables $Y_1^{(ij)} \sim \text{Poisson}(\lambda_1^{(j)})$ and $Y_2^{(ij)} \sim \text{Poisson}(\lambda_2^{(j)})$, with $i = 1, 2, \ldots, 1000$ and compute $R_t^{(ij)} = Y_2^{(ij)}/Y_1^{(ij)}$;

- Compute the α% credibility/confidence intervals using the method proposed here (β) and using the distribution proposed by Nishiura (2010) [16] (Γ), denoted by $c_\beta(R_t^{(ij)};\alpha)$ and $c_\Gamma(R_t^{(ij)};\alpha)$ respectively;
- Determine whether
  (i) $c_\Gamma(R_t^{(ij)};\alpha) \subseteq c_\beta(R_t^{(ij)};\alpha)$;
  (ii) $c_\beta(R_t^{(ij)};\alpha) \subseteq c_\Gamma(R_t^{(ij)};\alpha)$ or;
  (iii) there is overlap
- Calculate $Pr(R_t^{(ij)} > 1)$ using both distributions

The results show that the credibility intervals obtained using the distribution proposed here were larger than (i.e. contained) those computed using the distribution in 15 in 90.69% of the simulations (integrating over the grid), while the converse was observed in 1.14% of the cases. Moreover, our method yielded lower $Pr(R_t > 1)$ in 79.48% of the simulations, indicating it is indeed more robust to false alarm. An R script to perform the simulation study described above can be found at `https://github.com/fccoelho/paperLM1/blob/master/R/credibility_comparison.R`.

As a side note, the Bayesian approach presented in this paper will give similar results to orthodox confidence intervals [20] and [21] for Y_{t+1} and $Y_t >> 1$. Under the flat uniform prior for θ_t, the Bayesian posterior credibility interval is nearly indistinguishable from the confidence interval proposed by Clopper & Pearson (1931) [21] for $Y_{t+1}, Y_t > 20$. Note also that the uniform prior ($Beta(1,1)$) for θ_t constitutes a poor prior choice mainly because the induced distribution for R_t is only well-defined for $b_0 > 2$.

An advantage of the Bayesian approach is that one can devise prior distributions for θ_t taking advantage of the intuitive parametrization and flexibility of the beta family of distributions. Prior elicitation can also be done for R_t and the hyper-parameters directly plugged into the prior for θ_t. One can, for example, choose prior mean and variance for R_t and find a_0 and b_0 that satisfy those conditions. Let m_0 and v_0 be the prior expectation and variance for R_t. After some tedious algebra one finds

$$a_0 = \frac{m_0 v_0 + m_0^3 + m_0^2}{v_0} \tag{17}$$

$$b_0 = \frac{2v_0 + m_0^2 + m_0}{v_0} \tag{18}$$

If one wants only to specify m_0 and the coefficient of variation $c = \sqrt{v_0}/m_0$ for R_t *a priori*, some less boring algebra gives:

$$a_0 = \frac{m_0^3 c^2 + m_0^3 + m_0^2}{m_0^2 c^2} \tag{19}$$

$$b_0 = \frac{2m_0^2 c^2 + m^2 + m}{m_0^2 c^2} \tag{20}$$

This approach thus makes it possible to incorporate epidemiological knowledge about disease Biology (e.g. the magnitude of R_0) into the computation of R_t. This may prove particularly important when disease counts are low and/or close to the detection threshold. We provide an R script to perform the elicitation described above at `https://github.com/fccoelho/paperLM1/blob/master/R/elicit_Rt_prior.R`.

Estimating S_0 from simulated data

In order to determine whether the inference methodology proposed can recover the true parameter values of the underlying simulation model, we have devised a simple simulation experiment. Using the SIR

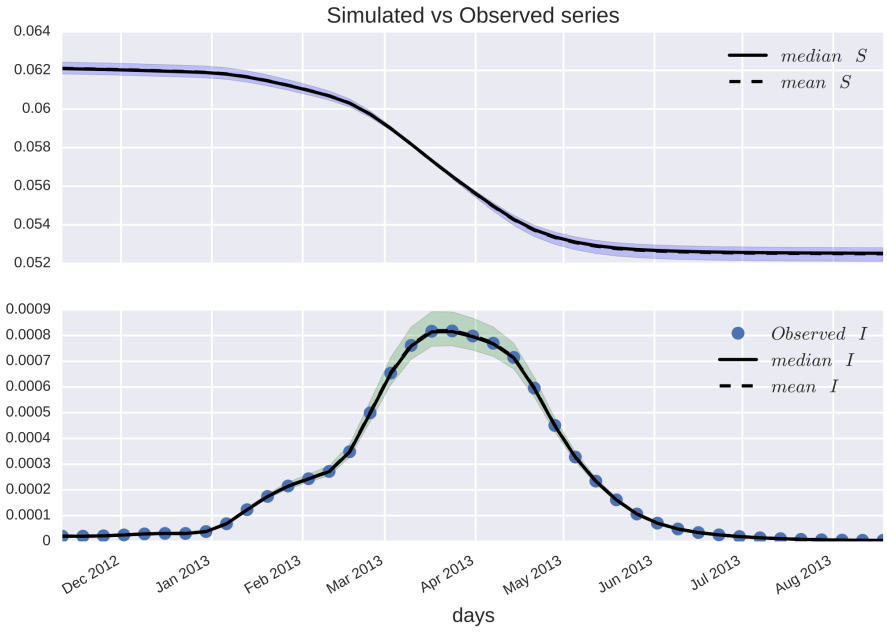


Figure 4. S and I series fitted to simulated incidence data (blue dots) with $S_0 = 0.0621$, and R_t estimated from real incidence data.

model presented in the main paper we have simulated the incidence curve of the 2013 epidemic (2012 in table 13), using $S_0 = 0.0621$ and R_t , as estimated from actual incidence data, and $\tau = 1$. Figure 4 shows the posterior S and I alongside with simulated incidence data. It is clear that the method can recover the correct value for S_0 used to simulate incidence. The script to generate the simulated data and run the inference is available at the paper's Github repository (`https://github.com/fccoelho/paperLM1/blob/master/python/fit_simulated_data.py`).

DiffeRential Evolution Adaptive Metropolis (DREAM)

Since the posterior distributions desired in this paper are not available in closed-form, we need to resort to numerical methods to obtain approximations. We employ an adaptive Markov chain Monte Carlo (MCMC) algorithm, proposed by [23], called DiffeRential Evolution Adaptive Metropolis (DREAM).

DREAM draws on the basic structure of Differential Evolution Markov Chain (DE-MC) which runs N independent chains in parallel and accepts moves proportional to the difference of two randomly sampled members (chains), thus differentially evolving conditional on the proposal variances. An additional aspect of DREAM is the so-called delayed rejection. This procedure adapts the original Metropolis-Hastings algorithm by attempting new moves whenever a proposal is rejected (instead of setting the current state to the previously accepted one), thus delaying the rejection of proposals. This potentially decreases autocorrelation and improves mixing (see [29] for details).

A desirable property of DREAM is that one can scale the algorithm with model complexity, meaning one can initialise as many parallel chains as there are parameters in the model, thus exploiting the multi-core architecture of modern computers to speed up convergence of the MCMC.

A Python implementation of DREAM for dynamic models is available in the package Bayesian inference with Python (BIP) [17] available at `http://code.google.com/p/bayesian-inference/`. The code to produce the results presented in this paper is available at `https://github.com/fccoelho/paperLM1/tree/master/python`.