

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

Flavio Codeço Coelho<sup>1\*</sup>, Luiz Max de Carvalho<sup>2</sup>,

**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 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 there serotypes is considered short lived [2]. 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 [3, 4]. 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 [5] 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 [6, 7], 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 [8, 9], 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 [10]. 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 events, can drastically change the pool of susceptibles, throwing off any sequential estimation based on the incidence dynamics.

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 [11]. We extend a Bayesian framework previously used to estimate the number of susceptibles in Influenza epidemics in Europe [12] 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 et al. (2010) [11] extend the theory developed by Stallybrass et al. (1931) [13] and propose to estimate  $R_t$  as

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

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 [14], whose objective is to obtain confidence intervals for the ratio of two Poisson counts. Let  $Y_t \sim \text{Poisson}(\lambda_t)$  and  $Y_{t+1} \sim \text{Poisson}(\lambda_{t+1})$  and define  $S = Y_t + Y_{t+1}$ . The authors note that by conditioning on the sum  $S$

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

$$\theta_t = \frac{\lambda_{t+1}}{\lambda_t + \lambda_{t+1}} \quad (3)$$

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) [14] show that one can construct a 100 $\alpha\%$  confidence interval for  $R_t$  by noting that

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

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

Many authors have chosen to quantify the uncertainty about  $\theta$  following orthodox approaches (see for example [15] and [16]) mainly for simplicity. We choose instead to take a Bayesian approach and use the 100 $\alpha\%$  posterior credibility interval for  $\theta_t$  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  $\theta_t$  is

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

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  $b_1 = Y_t + b_0$  [17]. The density of the induced distribution is then

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

Thus, the expectation of  $R_t$  is  $a_1/(b_1 - 1)$  and its variance is  $a_1(a_1 + b_1 - 1)/((b_1 - 2)(b_1 - 1)^2)$ . 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

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

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

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  $\beta$  and recovery happens at a rate  $\tau$ .

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

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

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

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

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. 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

$$p(S_{0j} | \mathbf{Y}_j) \propto l(\mathbf{Y}_j | S_{0j}, R_t, m, \tau) \pi(S_{0j}) \quad (11)$$

The likelihood  $l(\mathbf{Y}_j | \cdot)$  is assumed to be a Normal distribution with fixed variance  $\sigma^2$ . In this estimation procedure we kept  $R_t$  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) [12] available at <http://code.google.com/p/bayesian-inference/>. BIP uses a Differential Evolution Adaptive Metropolis (DREAM) [18] scheme that efficiently samples from high-dimensional joint distributions using multiple adaptive chains running in parallel with delayed rejection. 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 [19].

## 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.

$$A = \frac{\# \text{ cases}}{\text{Population at risk}} \quad (12)$$

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

$$A_j = \frac{\sum Y_j}{S_{0j}} \quad (13)$$

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 show a method to quantify uncertainty about  $R_t$  that is readily applicable to other diseases and; (ii) we propose to use  $R_t$  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) We estimate the AR for 18 years of Dengue epidemics.

Figure 1 shows the  $R_t$  series, according to (1) [11] 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. The method to quantify uncertainty proposed here provides more conservative credibility intervals, and therefore offers protection against false alarms.

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 [8, 9]. These methods try 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.

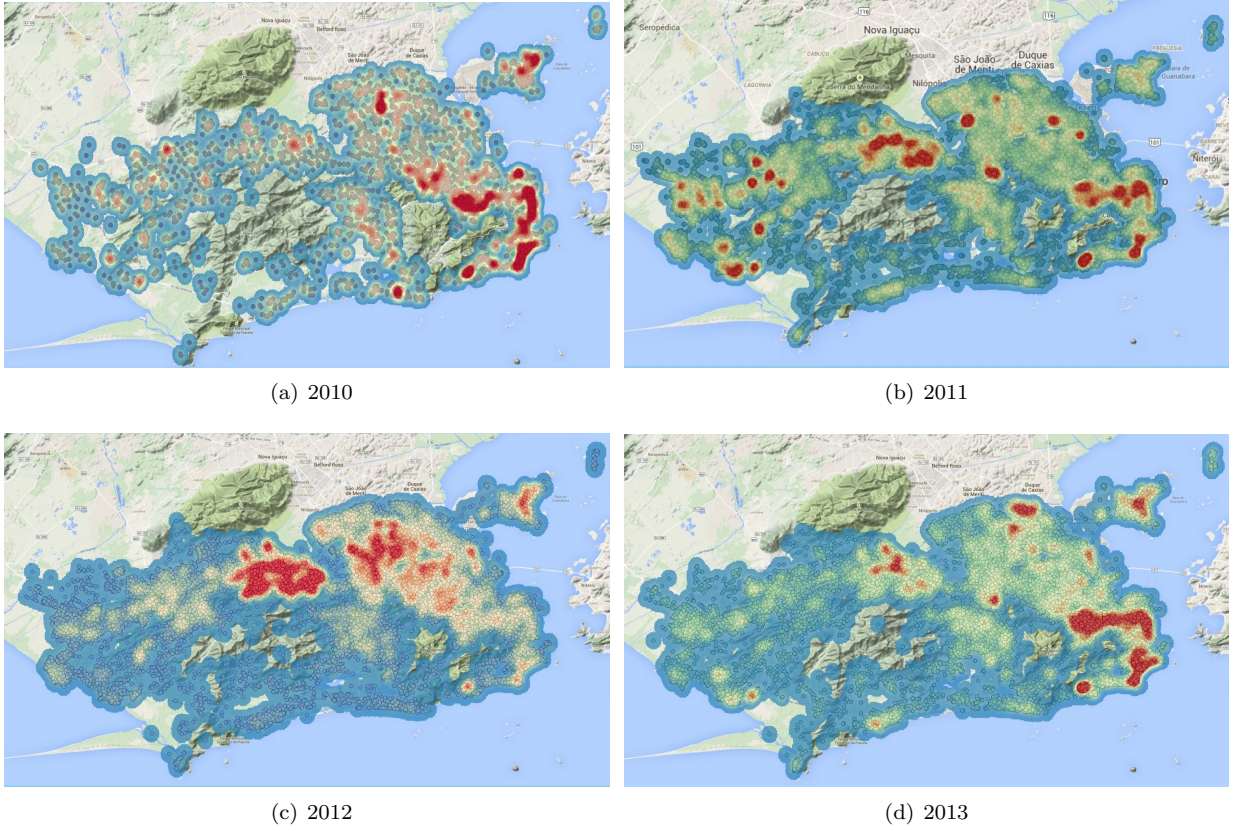
Since there is very limited information regarding the actual proportions of each virus in circulation and most information available is about the predominant serotypes for some epidemics in the period of study only [20], we propose the use of a simplified a single strain model. The main argument we put 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.



**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.

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 [21]. 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) [22], 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 3, 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

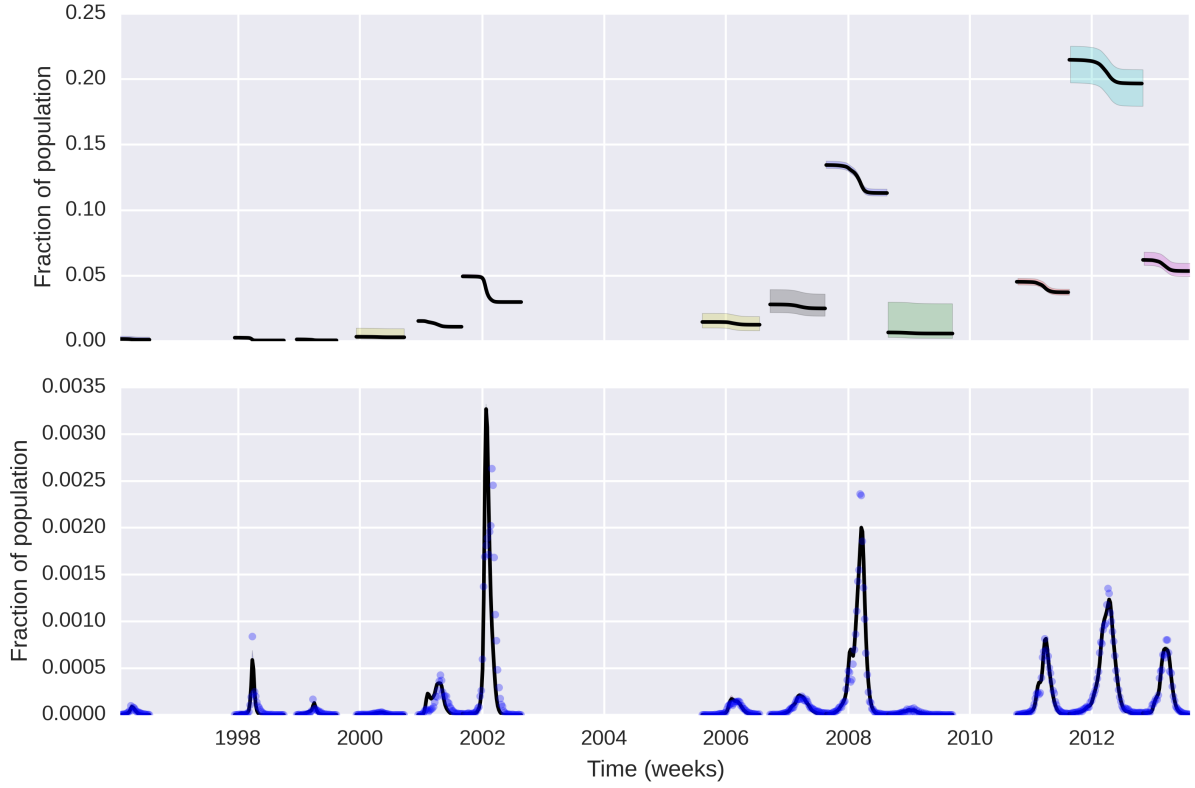


**Figure 2. Maps showing the incidence of dengue in the city of Rio de Janeiro from 2010 to 2013.** Circles indicate individual notified cases. A heatmap is overlaid on the maps showing absolute density of cases. It can be seen that several areas of the city were affected and no region seems to be free of transmission risk. This suggests that although transmission risk varies spatially, there is significant exposure over the entire city.

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 2, 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. 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  $S_0$ .

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 3. 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.



**Table 1. Median attack ratio and 95% credibility intervals calculated according to (13).**

Values are presented as percentage of total population.  $\dagger$ : Year corresponds to the start of the epidemic, however the peak of cases may occur in the following year.  $\ddagger$ : Susceptible fraction. These results show considerable variation in AR between epidemics, consistent with the acquiring and loss of serotype-specific immunity.

Year $^\dagger$	median Attack Ratio	$S_0^\ddagger$
1996	0.39 (0.17-0.54)	0.00171(0.0012-0.0038)
1997	0.87 (0.74-0.87)	0.00273(0.0027-0.0032)
1998	0.5 (0.49-0.5)	0.00142(0.0014-0.0014)
1999	0.11 (0.037-0.2)	0.00345(0.0018-0.01)
2000	0.25 (0.24-0.27)	0.0155(0.015-0.016)
2001	0.48 (0.47-0.49)	0.0495(0.048-0.051)
2005	0.15 (0.1-0.21)	0.0147(0.01-0.021)
2006	0.11 (0.08-0.14)	0.0281(0.022-0.037)
2007	0.15 (0.15-0.15)	0.135(0.13-0.14)
2008	0.14 (0.031-0.31)	0.00672(0.003-0.024)
2010	0.18 (0.17-0.19)	0.0454(0.043-0.048)
2011	0.086 (0.082-0.094)	0.215(0.2-0.23)
2012	0.14 (0.13-0.15)	0.0621(0.058-0.068)

## 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. 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>.
4. 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>.
5. 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/>.
6. 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>.
7. Katriel, G. & Stone, L. Attack rates of seasonal epidemics. *Mathematical Biosciences* **235**, 56–65 (2012). URL <http://www.sciencedirect.com/science/article/pii/S0025556411001532>.
8. Bjørnstad, O., Finkenstädt, B. & Grenfell, B. Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs* **72**, 169–184 (2002).

9. Wallinga, J., Teunis, P. & Kretzschmar, M. Reconstruction of measles dynamics in a vaccinated population. *Vaccine* **21**, 2643–2650 (2003).
10. 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).
11. 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).
12. Coelho, F. C., Codeço, C. T. & Gomes, M. G. A Bayesian framework for parameter estimation in dynamical models. *PLoS ONE* **6**, e19616 (2011).
13. Stallybrass, C. O. *et al.* The principles of epidemiology and the process of infection. *The Principles of Epidemiology and the Process of Infection*. (1931).
14. Ederer, F. & Mantel, N. Confidence limits on the ratio of two poisson variables. *American Journal of Epidemiology* **100**, 165–167 (1974).
15. Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22**, 209–212 (1927).
16. Clopper, C. & Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 404–413 (1934).
17. Dubey, S. D. Compound gamma, beta and f distributions. *Metrika* **16**, 27–31 (1970).
18. Vrugt, J. A. *et al.* Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Tech. Rep., Citeseer (2008).
19. Brooks, S. P. & Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics* **7**, 434–455 (1998).
20. 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).
21. 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>.
22. 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.scielo.org/scielo.php?script=sci\\_abstract&pid=S0034-89102006000100021&lng=en&nrm=iso&tlng=en](http://www.scielo.org/scielo.php?script=sci_abstract&pid=S0034-89102006000100021&lng=en&nrm=iso&tlng=en).

## 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 [14], 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.* [11]:

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

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

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

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  $\theta_t$ . 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 [11], 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:

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

Finally, note 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.

As a side note, the Bayesian approach presented in this paper will give similar results to orthodox confidence intervals [15] and [16] for  $Y_{t+1}$  and  $Y_t \gg 1$ . Under the flat uniform prior for  $\theta_t$ , the Bayesian posterior credibility interval is nearly indistinguishable from the confidence interval proposed by Clopper & Pearson (1931) [16] for  $Y_{t+1}, Y_t > 20$ . Note also that the uniform prior (*Beta*(1, 1)) for  $\theta_t$  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  $\theta_t$  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  $\theta_t$ . 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

$$a_0 = \frac{m_0 v_0 + m_0^3 + m_0^2}{v_0} \quad (17)$$

$$b_0 = \frac{2v_0 + m_0^2 + m_0}{v_0} \quad (18)$$

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:

$$a_0 = \frac{m_0^3 c^2 + m_0^3 + m_0^2}{m_0^2 c^2} \quad (19)$$

$$b_0 = \frac{2m_0^2 c^2 + m_0^2 + m_0}{m_0^2 c^2} \quad (20)$$

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 above elicitation at [https://github.com/fccoelho/paperLM1/blob/master/R/elicit\\_Rt\\_prior.R](https://github.com/fccoelho/paperLM1/blob/master/R/elicit_Rt_prior.R).