Multiple-serotype models of dengue virus transmission: simulation study and perspectives for the application of inference in epidemiological surveillance

Caetano Souto-Maior[‡*],

Instituto Gulbenkian de Ciência, Oeiras 2780-156, Portugal

‡ current address: National Heart Lung and Blood Institute, NIH, Bethesda, MD, United States of America
* caetanosoutomaior@protonmail.com

Abstract

With around 3 billion people at risk, dengue virus is endemic to many parts of the world. In the Brazilian city of Rio de Janeiro, surveillance measures require notification of new dengue virus cases, and are supplemented by serum collection from patients and sequencing of viral RNA. Phylogenetic analyses have been performed for all serotypes circulating in the country to identify viral genotypes, potentially identify new introductions, and compare viruses presently circulating in the country with those in the past, and of other countries. As a separate type of analysis, a number of mathematical models have been developed to describe dengue virus transmission – particularly qualitative incidence or prevalence patterns – although few have been tested. In this chapter, I show how different mathematical formulations could represent transmission of dengue virus by mosquitoes to humans, how the different model structures entail assumptions about the process, and how these affect outputs qualitatively. Inference from simulated data is used as proof of principle that the kind of data available could be used to accurately estimate all model parameters; however, it is shown that stochasticity may severely hamper efforts to test the models quantitatively. I further implement inference from sequence data for the different models, and compare the performance to that of time series. The methods are applied to the data available for the city of Rio de Janeiro.

1   Background

The persistence of dengue fever (as well as more severe syndromes caused by dengue virus) constitutes the most extensive viral epidemic transmitted by arthropods, with around 3 billion people at risk worldwide, and 300 million annual cases estimated (Bhatt et al. 2013). The recently recorded expansion in the range of the main transmission vectors, *Aedes aegypti* and *Aedes albopictus* (Kraemer et al. 2015) – presumably due to higher temperatures at temperate regions resulting from climate change – in combination with the emergence of other *Aedes*-transmitted diseases further increased attention to vector control.

Although among the vector transmitted viruses dengue itself has arguably lost some of the attention to Chikungunya and especially to Zika virus (due especially to the previously unknown relationship between the latter and birth defects) at the population level the study of either of these diseases is to a very large extent the study its human and mosquito hosts. The fact that licensed vaccines for these disease was essentially absent – dengue virus had a vaccine in phase 3 clinical trials (Eisen & Moore 2013; Hadinegoro et al. 2015; Villar et al. 2014) that was just recently licensed (WHO 2016) – further highlights the importance of vector control, and of the knowledge about dengue virus transmission in the control of all of the diseases transmitted by the *Aedes* mosquitoes.

One half of the of cycle of dengue virus ($DENV$) – that is, the mosquito-to-human transmission – happens through the bite of an infected *Aedes aegypti* or *Aedes albopictus* mosquito (i.e. the vector in *vector-transmitted disease*); the transmission cycle is completed when an infected human is bitten by a mosquito that in turn becomes infected (Halstead 2007). Although importation from other geographical areas (Salje et al. 2012; Stoddard et al. 2013; Vazquez-Prokopec et al. 2010) as well as sylvatic cycles between non-human primates and other *Aedes* species may play a role in sustaining transmission (Vasilakis et al. 2011), these basic steps of the human-*Aedes* cycle should be enough to create chains of transmission that allow endemicity, and this is considered the primary cycle in explaining dengue virus persistence.

A few details are noteworthy in a general model of dengue virus transmission, which would otherwise conform nicely to that of a generic vector-transmitted mode of propagation. Dengue virus has four antigenically distinct variants – types, or strains – commonly referred to as serotypes (DENV-1 through 4). Anything from a single one to all four of them can be circulating in any one place. If only one serotype is present, a simple description of transmission where a susceptible host gets infected, recovers, and becomes immune to further infection is generally adequate, since infection with a serotype is accepted to confer human hosts lifelong immunity against that same type. If more than one serotype is circulating, infection can happen at least twice (but not more than four times, because unlike influenza, for instance, evolution of the virus does not allow it to escape immunity built against it), one for each previously unseen serotype. In this case multiple infections may need to be accounted for. Also, it could be important to differentiate between strains, as a secondary infection can only be caused by a serotype different from the previous.

Dengue virus transmission has been extensively explored through mathematical models (Johansson et al. 2011). As usual, the disease states of the human hosts have been described by simple extensions of the susceptible-infected-recovered framework, often (but not always) coupled with susceptible-infected description of the mosquito hosts. A mix-and-match of other known or suspected features specific (although possibly not exclusive) to dengue virus have been further added: secondary infections, temporary strain-transcending immunity (cross-protection), enhanced (or reduced) susceptibility to secondary infections, increased lethality in case of severe presentations (often associated to secondary infections, as well as other risk factors such as age or

blood-related dysfunctions) (Johansson et al. 2011).   72

Therefore, dengue virus transmission can be described mathematically by multiple   73
explicit serotypes, which we denote by multiple-serotype models (e.g. a two-serotype   74
model means two different strains of dengue are explicitly described), or by a single   75
explicit virus type, hereafter denoted by SIRX models (which include the classic SIR   76
and SIRS models, as well as intermediate formulations, as described in the methods   77
section). Both formulations purport to describe settings where one or more serotypes   78
may be circulating; in the former description each explicit serotype causes infection once,   79
while in the latter the ensemble of unspecified serotypes causes infection twice or more.   80

Although seemingly subtle, the conceptual difference between the two modeling   81
approaches is profound: while in the SIRX vector models secondary infections depended   82
exclusively on waiting for recovered individuals to become susceptible again, in the   83
multi-type models strains compete for multiply susceptible individuals. This feature can   84
causes serotype alternation and induce oscillation even in the absence of seasonal   85
forcing. More importantly, this model is less of a caricature of the process of disease   86
transmission, since it is widely accepted that an individual infected with a serotype   87
cannot be infected again by the same strain, rendering the explicit description of the   88
SIRX models technically impossible. Whether one approach or the other is more   89
suitable to describe real dengue epidemics, however, cannot in principle be decided   90
without confronting both models to epidemiological data.   91

On the epidemiological records side of dengue, some unique patterns are often   92
highlighted in dengue virus epidemics, particularly the oscillations with multianual   93
periods and serotype replacement in successive epidemics (Adams et al. 2006). These   94
can be verified, respectively, from incidence records that show greater number of cases   95
usually around the rainy seasons, and through serological surveys or, more recently,   96
sequencing of circulating viruses. Mathematical models of dengue transmission therefore   97
are built to reproduce these broad patterns; nevertheless, different combinations of   98
anyone's favorite model components may reproduce them, in a way that is   99
indistinguishable from someone else's choice of building blocks. A non-exhaustive list of   100
processes that could produce realistic outputs in a computer simulation include: 101
stochasticity(Otero & Solari 2010), spatial structure (Favier et al. 2005), enhanced   102
secondary infectivity (Nagao & Koelle 2008), "unnatural" transmission routes (Chikaki   103
& Ishikawa 2009).   104

One of the most hyped effects among the many incorporated one way or another into   105
the mathematical models is that of antibody dependent enhancement, by which a   106
secondary infection would be more severe than the first in the presence of titers of   107
heterologous antibodies against *DENV* (Kliks et al. 1989). The inclusion of the effect   108
has been shown to drive chaotic dynamics even in deterministic mathematical models   109
(Bianco et al. 2009), so as a result it has been suggested that it could be the most   110
important effect modulating the observed somewhat erratic epidemic patterns. The   111
plausibility of the effect is asserted through the observation that in the presence of   112
subneutralizing antibodies invasion of the cell by viruses is facilitated (Guzman &   113
Vazquez 2010); however, in terms of a mathematical model it is not clear if that would   114
translate into increased susceptibility, increased infectivity, or simply a   115
transmission-unrelated increase in virus lethality. If the magnitude of the enhancement   116
could ever be as large as claimed in modeling studies is not established either.   117
Furthermore, it is not clear whether, if present, the effect would be the dominant factor,   118
or if it would be important to the transmission dynamics at all.   119

Many other effects and combinations would still not exhaust the list of tens of   120
models that purport to explain dengue transmission (Johansson et al. 2011);   121
nevertheless, a quantitative evaluation of the conformity of these models to real data   122
was not done systematically [but see Rasmussen et al. (2014b) for a rare exception].   123

Here, I put together a model of the human and vector population with secondary infections, either one or two serotypes, and temporary immunity after any infection, but otherwise minimal in what regards any other asymmetries, enhancements, or alternative routes of transmission. I show that minimalistic multi-serotype models can sustain oscillations in the absence of any of the latter effects or stochasticity; I also implement an individual-based model that can simulate both epidemiological and viral evolution. I further develop inference methods to fit these models to time series and multi-serotype sequence data, compare the inference results for the different kinds of simulated data, and apply the estimation method to real data.
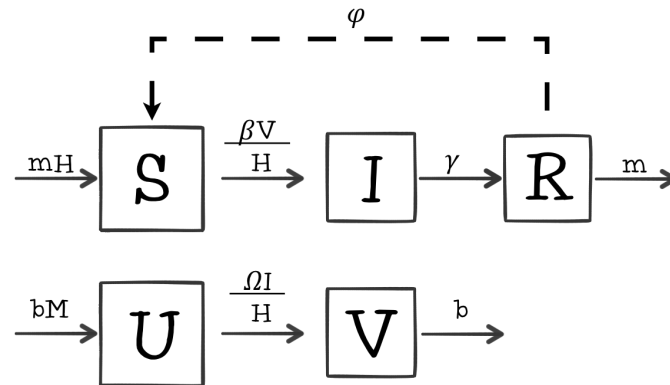

Figure 1. Structure of an SIRS plus vector model, with possible loss of immunity indicated by the dashed arrow. All compartments are subject to natural mortality m, but the arrows corresponding to those processes are ommitted in all but the last compartments to avoid repetition and clutter.

2 Methods

2.1 SIR model extensions for dengue virus transmission

2.1.1 SIRX plus vector models

The simplest model to describe dengue transmission is arguably the vector SIR model, not unlike the first basic models of malaria transmission with a human and mosquito population, although malaria may have an indefinite number of reinfections, making it a SIS model (Ross 1916; Smith et al. 2012). The SIR model assumes human hosts are only susceptible once (S), and after infection (I) enter the recovered compartment (R) permanently, being immune to any further infection by dengue afterwards.

The vector compartment is modeled as a susceptible mosquito compartment (U), and an infected one (V), from which a mosquito host never exits once it enters – noting that the female mosquitoes are the only ones transmitting disease, so the male population is absent, or implicit.

Alternatively, human hosts may be allowed to lose immunity acquired from past infections and become susceptible again; under that assumption a single host can potentially be reinfected an unlimited number of times before it dies.

The schematic drawing of both model formulations is shown in figure 1, the only difference between the two being the arrow representing hosts that exit the recovered compartment and reenter the initial susceptible compartment. In this latter case, the structure of the human compartments is that of what is dubbed the SIRS model – the first and last susceptible compartments being the same.

The compartments and parameters are very much standard: m being the human host mortality rate; β the mosquito to human transmission coefficient, or rate; γ the human recovery rate; ϕ is the immunity loss rate (which is equal to zero in the SIR version of the model); b is the mosquito host mortality rate; Ω is the human to mosquito transmission rate. Additionally, it is assumed that the birth rates are the same as the mortality rates for each host species; therefore the population sizes stay constant: H is the value set for the human population, while M is the size of the female mosquito population. The mathematical formulation of the SIRS dynamics of transmission is given by the system of equations (1).

$$
\begin{aligned}
\frac{dS}{dt} &= mH + \phi R - \frac{\beta}{H}VS - mS \\
\frac{dI}{dt} &= \frac{\beta}{H}VS - \gamma I - mI \\
\frac{dR}{dt} &= \gamma I - \phi R - mR \\
\frac{dU}{dt} &= bM - \frac{\Omega}{H}IU - bU \\
\frac{dV}{dt} &= \frac{\Omega}{H}IU - bV
\end{aligned}
\tag{1}
$$

In addition to vector and human basic demographic and epidemiological parameters, it is also common to assume seasonal forcing of the vector population, emulating changing conditions from more to less favorable throughout the year, usually attributable to either hot/cold, or humid/dry seasons. The result of that function is then added either to the birth or death rate of the mosquito population, resulting in a deterministic sinusoidal oscillation.

For clarity the seasonality function is not introduced in this first display of the mathematical system (eqs. 1), but is detailed in the following system (2) instead.
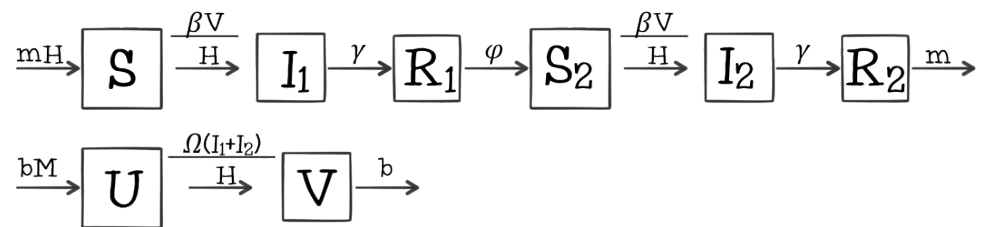

Figure 2. Structure of an SIR plus vector model, with explicit number of possible reinfections (in the particular case illustrated the vector SIRx2).

2.1.2 SIR-vector models and multiple infections

Given that dengue virus has a finite number of serotypes, it is expected that any one host can only be infected with dengue a few times in a lifetime; therefore, secondary infections can be modeled by explicit compartments for the secondarily or further infected.

In this case, the total number of times a single individual can be reinfected has a hard limit given by the number of infected compartments in the model, which are never

revisited. This is straightforwardly modeled by a series of SIR compartments chained together (i.e. an SIR structure followed by another SIR).

The identities of individual serotypes are not explicit in this model, but sequential infections implicitly describe this feature of dengue virus transmission. The case with two consecutive infections is shown in figure 2.

The compartmental structure of the vector population is unaltered, while the human host population follows a susceptible-infected-recovered path, being immune to reinfection for the time they stay in the first recovered compartment (R_1). After that, loss of immunity takes individuals to a second susceptible state (S_2, unlike the SIRS model where the first state is revisited), where individuals are again susceptible to infection by infected mosquitoes.

In case of infection they move to the secondarily infected compartment (I_2), and after recovery they move to the last compartment (R), where they are recovered and can only exit by the ultimate process of death.

The model parameters are the same as the previous model and describe exactly the same processes as before, with the single and only slight exception being that ϕ describes a path of waning immunity through different compartments due to the general model structure. The mathematical formulation of the model with two sequential infected compartments is given by the system of equations (2).

As mentioned above, system (2) also has a seasonality term acting on mosquito birth rates. This consists of a time dependent cosine function with argument 2π multiplied by time itself plus a phase variable δ; this assumes these variables are given in years, resulting in an oscillation period of one year, but can trivially be transformed into months, weeks, or days, for instance, by dividing by the appropriate factor. The cosine multiplies the amplitude scaling parameter α; this factor is added to the constant birth rate, resulting in a population whose size oscillates between $(1 \pm \alpha)M$.

$$
\begin{aligned}
\frac{dS}{dt} &= mH - \frac{\beta}{H}VS - mS \\
\frac{dI_1}{dt} &= \frac{\beta}{H}VS - (\gamma + m)I_1 \\
\frac{dR_1}{dt} &= \gamma I_1 - (\phi + m)R_1 \\
\frac{dS_2}{dt} &= \phi R_1 - \frac{\beta}{H}VS_2 - mS_2 \\
\frac{dI_2}{dt} &= \frac{\beta}{H}VS_2 - (\gamma + m)I_2 \\
\frac{dR_2}{dt} &= \gamma I_2 - mR_2 \\
\frac{dU}{dt} &= bM[1 + \alpha \cdot cos(2\pi(t + \delta))] - \frac{\Omega}{H}(I_1 + I_2)U - bU \\
\frac{dV}{dt} &= \frac{\Omega}{H}(I_1 + I_2)U - bV
\end{aligned}
\tag{2}
$$

In contrast to the three-compartment SIRS model I refer to this model as the SIRx2 model (given that the human population states are described by a SIR plus SIR, or two times the SIR structure), where once-recovered individuals would again become susceptible to infection. Because $DENV$ has 4 human-infecting serotypes infections may further be tertiary or quaternary; unless explicitly stated, I hereafter refer to all infections after the first simply as secondary, as opposed to primary. SIRx3 and SIRx4 models where individuals infected twice or three times, respectively, become once again

susceptible can be built by straightforwardly extending the SIRx2; therefore, I do not   211
show specific schemes or systems of equations for those.   212

In the infinity limit these models become the SIRS model, except human hosts enter   213
an infinite number of new compartments instead of entering the same compartments an   214
infinite number of times, provided of course hosts stay alive long enough. Depending on   215
rates of infection and death, a smaller number of compartments may be enough for a   216
human host to have enough new compartments for many lifetimes of repeated infection,   217
in which case the model will also approach the SIRS model even with a finite number of   218
reinfections.   219

I refer to this entire class of models as SIRX, which include the shown SIR, SIRS, as   220
well as the SIRx2 (or any other number between two and infinity) models. In any of   221
those cases, however, the identity of multiple serotypes are only implicit in the fact that   222
human hosts can have secondary infections, because there is only one class of infected   223
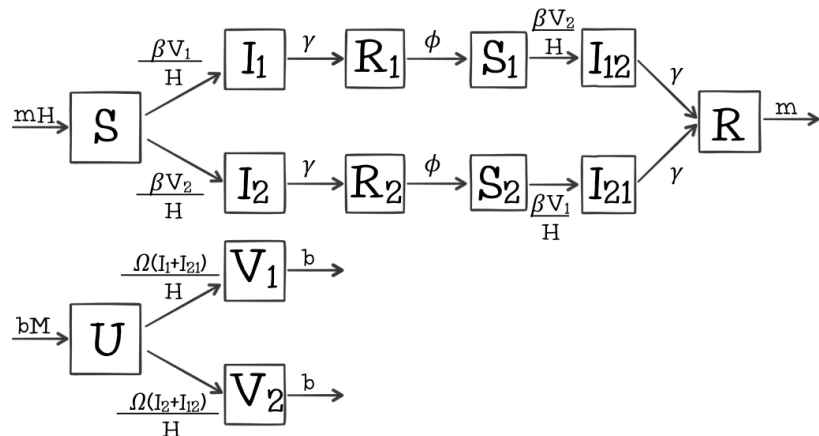mosquitoes that transmit to all susceptible humans regardless.

Figure 3. Structure of an explicit two-serotype model and its rate parameters.

224

2.1.3 Multiple-serotype infections in the descriptions of dengue virus incidence   225 226

A more complete description of multiple-serotype transmission is one that differentiates   227
not only between primary and secondary infections but also disinguishing which   228
serotype causes infection each time. This requires not only a series of compartments,   229
but also parallel paths that describe the order in which the multiple serotypes cause   230
infection. For two serotypes, for instance DENV-1 and DENV-2, it is accepted that a   231
human host could be infected twice, once for each serotype; that is accounted for by the   232
two sequential infected compartments of the SIRx2 model described previously.   233

Here I wish to account for the order of infection: a human host can either be   234
infected by DENV-1 and then DENV-2, or by DENV-2 and then DENV-1; this creates   235
two alternative paths which are shown in figure 3 – e.g. I_{12} denotes individuals first   236
infected with serotype 1 and now infected with serotype 2.   237

Unlike the SIRx2 models the mosquito hosts can either harbour one or the other   238
serotype; therefore, also in contrast to the previous formulations, a secondary human   239
infection depends on transmission from a mosquito with different serotype from that of   240
the primary. The mathematical description of this explicit two-serotype model is given   241

by the system of equations (3).

$$\frac{dS}{dt} = mH - \frac{\beta}{H}(V_1 + V_2)S - mS$$

$$\frac{dI_1}{dt} = \frac{\beta}{H}V_1 S - (\gamma + m)I_1$$

$$\frac{dR_1}{dt} = \gamma I_1 - (\varphi + m)R_1$$

$$\frac{dS_1}{dt} = \varphi R_1 - \frac{\beta}{H}V_2 S_1 - mS_1$$

$$\frac{dI_{12}}{dt} = \frac{\beta}{H}V_2 S_1 - (\gamma + m)I_{12}$$

$$\frac{dI_2}{dt} = \frac{\beta}{H}V_2 S - (\gamma + m)I_2$$

$$\frac{dR_2}{dt} = \gamma I_2 - (\varphi + m)R_2 \qquad (3)$$

$$\frac{dS_2}{dt} = \varphi R_2 - \frac{\beta}{H}V_1 S_2 - mS_2$$

$$\frac{dI_{21}}{dt} = \frac{\beta}{H}V_1 S_2 - (\gamma + m)I_{21}$$

$$\frac{dR}{dt} = \gamma(I_{12} + I_{21}) - mR$$

$$\frac{dU}{dt} = bM[1 + \alpha \cdot cos(2\pi(\delta + t))] - \frac{\Omega}{H}(I_1 + I_{21})U - \frac{\Omega}{H}(I_2 + I_{12})U - bU$$

$$\frac{dV_1}{dt} = \frac{\Omega}{H}(I_1 + I_{21})U - bV_1$$

$$\frac{dV_2}{dt} = \frac{\Omega}{H}(I_2 + I_{12})U - bV_2$$

Although the model described by the system of equations (3) does contain (more than a couple) SIR-like components, and could possibly be seen as a parallelization of the SIRx2 model, the level of complexity arising from further introducing $DENV$-specific features is considerably higher. It may be more useful to look at it as two interacting epidemics (Rohani et al. 2003), since it is unlikely that either the compound output or the individual serotype dynamics can be predicted to trivially conform to that of its more well known building blocks.

Regarding the basic reproductive number, R_0, because it is calculated with regard to a fully susceptible population, the secondary compartments do not cause the models to differ in that matter, and the quantity is given by $R_0 = \frac{\beta\Omega}{b(m+\gamma)}\frac{M}{H}$. If such differences are actually verified in the output of the model, it would be due to the structure of the immunological states believed to represent a population at risk of a multi-serotype dengue epidemic, since all other processes are present in the previous models as well.

On top of the structuring of basic epidemiological and demographic processes (Keeling & Rohani 2011), additional complexity may appear by introducing asymmetries between serotypes (e.g. serotype 1 more infective than serotype 2, or causing infection for a longer period of infection) or order-dependent rates of infection (e.g. secondary infections more or less likely than primary), which are plausible for various biological or medical reasons, and are likely in comparison to the narrow null hypothesis of perfect symmetry.

Other common extensions that are known to be present to some degree, unlike more complex hypothesized immunological or epidemiological effects, include exposed compartments (describing individuals that harbor the pathogen but cannot transmit it yet), spatial structure of transmission, heterogeneity in contact rates, susceptibility to

infection, or in infectivity, gamma-shaped (as opposed to exponential) host survival, and many others. Nevertheless, none of these are included in the models used here, I discuss some of these later in the text, although it may become clear by the section on inference which are the difficulties of including too many parameters however simple and concrete the processes may be.

2.1.4 Implementation

The models were initially implemented as continuous ordinary differential equations (ODEs), solved by numerical methods to approximating the deterministic solutions of the system; commonly available as ODE solver functions in multiple programming languages such as Matlab, Python, and R languages, which were used at different times with no particular preference for either.

2.2 Individual-based models

Discrete, stochastic, individual-based versions of the models described above were also implemented. Besides the importance of the randomness of the events in the epidemiological model, it was important to be able to simulate not only the epidemiological outputs, but also the evolution of viral sequences. Because there is no direct continuous approximation neither to the appearance of a random mutation, nor to a genetic sequence of nucleotides, the most straightforward way of simulating evolution is to explicitly attribute viral sequences to infected individuals, and allowing them to randomly acquire new mutations as they get transmitted.

Implementations were done in both C++ programming language by using a previous implementation (Gordo et al. 2009; Gordo & Campos 2007), and later by adapting the algorithm to the Python programming language to take advantage of the random number implementations in the latter.

In brief, an "Individual" class was created to have a "sequences" attribute (which was empty if the individual was not infected), and each human or mosquito host was an instance of that class. At each time step (Δt, which multiplies all probabilities hereafter mentioned), the number of new infections was drawn from a random binomial distribution, since the maximum number of infections is bounded by the total number of susceptible individuals; the probability parameters were equal to the force of infection (e.g. $\lambda = \beta V_1 / H$ for infections caused by mosquitos infected with serotype 1) and number of trials equal to the susceptible population (e.g. the susceptible to all S, or to type 1 S_2, equivalently.) Because the infectivity of all individuals is assumed to be the same, the sequence infecting each new host is randomly drawn from the pool of all existing sequences from the previous time step. After a successful infection of a new individual, new mutations have the opportunity to arise with probability determined by a per-genome mutation rate – mutations are assumed to follow an infinite alleles and sites model, so every new mutation is one that was not previously present in the population. Mutations do not affect any of the model parameters, so evolution is completely neutral. The number of sequences inside a single infected host can be greater than one, in which case this within-host population undergoes a Wright-Fisher sampling step (Wakeley 2009).

The number of new births of the mosquito population, apart from the sinusoidal additional factor, is drawn from a poisson distribution with mean $\mu_{vec} = b$; deaths are drawn from a binomial distribution with probability parameter also $\mu_{vec} = b$, so the total population is expected to fluctuate around the initial value. The human population is assumed to be strictly constant; that is enforced by the number of births being exactly equal to the number of stochastic deaths – this condition can be easily relaxed, however.

Otherwise, as a general rule, the number of events at each time step was drawn from ₃₁₆
a random binomial (and when applicable, multinomial) distribution where the number ₃₁₇
of trials was given by the number of individuals in the compartment, and the parameter ₃₁₈
for probability of success in each trial was given by the rates in the model (e.g. the ₃₁₉
number of individuals recovered from an infected compartment I_1 is given by a random ₃₂₀
binomial distribution draw with parameters I_1 and γ). If competing processes were ₃₂₁
present, a multinomial was used instead. If the number of events was not bounded, for ₃₂₂
instance by the size of the compartment, a poisson distribution was used instead (e.g. ₃₂₃
the number of mosquito births, or number of new mutations). ₃₂₄

The output of this implementation is both a stochastic time series of susceptibles, ₃₂₅
infected, recovered, and incidences (i.e. the randomly drawn number of new infections ₃₂₆
recorded at each step in the human and vector populations), and the pool of all extant ₃₂₇
pathogen sequences for each serotype at all or selected time points (for convenience, ₃₂₈
split into mosquito and human harbored sequences). ₃₂₉

2.3 Epidemiological surveillance data (and pseudodata) ₃₃₀

In countries like Brazil, where notification of dengue is compulsory to doctors, the ₃₃₁
records commonly consist of periodically reported new cases into the Information ₃₃₂
System for Incident Notification [Sistema de Informação de Agravos de Notificação] ₃₃₃
(SINAN). As with many other common endemic diseases, laboratory confirmation is not ₃₃₄
routine, so diagnostic relies mainly on clinical criteria; neither the serotype causing the ₃₃₅
infection is normally recorded, nor if the infection is primary or not. Therefore, typical ₃₃₆
time series do not distinguish between serotypes or sequential infections; what would be ₃₃₇
available would be a series of equally spaced, discrete values representing the number of ₃₃₈
cases of "dengue fever" generically defined reported every month or week (SINAN). ₃₃₉

The time series data set used here is from the Brazilian SINAN, with the absolute ₃₄₀
number of new weekly cases of dengue from the year 2009 until 2013 in the city of Rio ₃₄₁
de Janeiro, when three large incidence peaks are observed. More specific diagnostics ₃₄₂
data have also been occasionally produced in the form of serological surveys, although ₃₄₃
these were obtained for specific studies and small cohorts. Because this kind of data is ₃₄₄
sparse and difficult to access, I do not use any such data, but merely note that it exists ₃₄₅
for the disease and location I am (mainly) concerned with, even if in a fragmented way. ₃₄₆

Relatively recently, routine surveillance of dengue started to include genetic data of ₃₄₇
the virus. It is now routine activity to isolate samples from patients and obtain ₃₄₈
nucleotide sequence from the virus in the isolate; an immediate result is the ₃₄₉
identification of serotypes, and possibly of more specific variants such as genotype (a ₃₅₀
finer grained distinction within each serotype) as well as the relationship to strains ₃₅₁
previously found elsewhere (dos Santos et al. 2002). This data set therefore consists of ₃₅₂
somewhat sparsely sampled viral isolates sequenced along several years – the total ₃₅₃
number is in the order of tens for the city of Rio de Janeiro – and is to a great extent ₃₅₄
available as part of published studies (Araújo et al. 2009; de Bruycker-Nogueira et al. ₃₅₅
2015; Castro et al. 2013, 2012; Miagostovich et al. 2003, 2006; DeSimone et al. 2004), as ₃₅₆
well as in public databases for genetic sequences such as GenBank (Benson et al. 2015). ₃₅₇

The individual-based models described in the previous subsection were designed in a ₃₅₈
way that could directly reproduce the form observed in the real data available. Unless ₃₅₉
there is specific interest in greater details, whenever a model is simulated I try to store ₃₆₀
the output in a format that mimics the amount, type and level of aggregation, period ₃₆₁
and interval of collection, and any other feature pertinent to a specific data set. When ₃₆₂
used in the same way as the real data (e.g. for parameter inference), I call a data set of ₃₆₃
this sort synthetic, or pseudo-data. ₃₆₄

Generally summarizing the two types of pseudo data sets used here, the time series ₃₆₅
are weekly records of all new cases (the weekly sum of daily-generated incidences), and ₃₆₆

the second type are genetic sequences. The genetic equivalent of the all-inclusive time series data would be a sequence (or group of sequences) for every newly infected individual time-stamped with the week (or any other time step) when it appeared. That is impractical even for simulated data set, and for a series of reasons it is essentially impossible in real epidemics.

Instead, the pseudo-sequence data is a sample from different times, where the number of sequences from each time point is proportional to the number of cases then, i.e. it is a sample of sequences over one long period, weighted by epidemic size at each of multiple small intervals. A total number of 100 sequences per serotype over the course of a few years was assumed to be a sufficient number, comparable to that of previous data sets used for similar purposes (Rasmussen et al. 2014b), and considered feasible in a city with a population on the order of 10 million, and outbreaks on the order of a few tens of thousands (SINAN). It is also comparable (though larger) in size to data sets collected for other purposes in the city of Rio de Janeiro (DeSimone et al. 2004) When needed, the specific pseudo and real data sets are detailed at the pertinent results sections.

The objective is to obtain pseudo data sets suitable for inference purposes. It is difficult to establish beforehand what the most informative sampling scheme would be (Frost et al. 2015); that is a question on its own right that is not explored here.

2.4   Bayesian inference from time series

Analytical solutions for the systems adopted in this chapter are not available; therefore, a numerical approximation to the continuous solution was obtained through an ODE solver whenever needed.

Because the number of new cases in any given week is an integer number I chose to use a poisson distribution: the likelihood of the observed value is computed using the sum of all possible human infected states as the poisson parameter (as modeled by compartments, i.e. the total number of dengue cases of any kind in the model output), and the total likelihood is therefore the product of that over all time points in the series – or, more conveniently, the sum of their logarithms.

A binomial distribution could as well be used, in which case its probability parameter could be given by the forces of infection and the number of trials would be given by the susceptible populations at risk. While that would be possible, and possibly mimic more accurately our simulation model and the bounds in the maximum number of infections, it is more cumbersome to add the parameters coming from the different compartments and, more importantly, the poisson distribution expects greater or equal variance when compared to the binomial, and therefore can accommodate any overdispersion in the data.

The bayesian Markov Chain Monte Carlo (*MCMC*) inference algorithm was implemented in Python language using the PyMC module (Patil et al. 2010). Unless detailed otherwise, as a norm gamma-shaped priors were used for parameters that have independently estimated or commonly accepted values, and priors uniform over a wide range were used otherwise. As technical criteria for quality of the inference, the following criteria were used (Gelman et al. 2013, chap. 11): Markov chains were run until the likelihood and posterior traces converged to a maximum and attained stationarity with that regard; for all results shown, replicates of the chains were run to assure mixing (unless otherwise specified, model fit and posterios are computed and shown for individual chains only); initial sampling corresponding to one tenth of the total number of iterations was discarded as warm-up or burn-in period. Correlation between parameters were computed at the end of the chain for at least one replicate when parameters seemed systematically biased.

Most critical was the time limit of around one month that was imposed for practical reasons; that allowed chains as long as a few million iterations on a dedicated

high-capacity computer. Most implementations did not seem to have issues with the   418
likelihood converging with a few hundred thousand iterations; nevertheless, more subtle   419
issues were observed in some cases and are discussed in the results and discussion   420
sections.   421

I applied this estimation method to both simulated data sets, as well as to a time   422
series of dengue incidence from the city of Rio de Janeiro running from 2009 to 2013   423
(SINAN).   424

The estimation method used does not therefore account for noise in the system state.   425
Stochastic solutions can be used instead; however, doing so is not as simple as replacing   426
a deterministic solution by a single stochastic one (which may be only slightly slower to   427
obtain), but requires a considerably more sophisticated and a computationally much   428
more intensive method method to estimate the likelihood Andrieu & Doucet (2010);   429
Ionides et al. (2006), with possibly a couple of thousand simulations at each iteration of   430
the Markov Chain. I discuss these so called Sequential Monte Carlo or particle filters as   431
perspectives in the end of this chapter.   432

2.5   Population genetics and phylodynamic inference   433

Although the individual-based model can produce a simulated data set that mimics a   434
set of viral sequences sampled at arbitrary time points along time, unlike with the time   435
series data there is no simple way of calculating the likelihood of the model parameters   436
given that kind of data.   437

The conceptually most straightforward method is probably the following: make up   438
metrics that are assumed to be representative of the data; simulate the model; compute   439
the same metrics for a large number of model outputs; and try to find the model   440
parameters that best approximate the real data. In spite of the gross oversimplification   441
of this description, this is the basis of Approximate Bayesian Computation (or ABC)   442
methods.   443

An alternative framework to compute the likelihood of a model given sequence data –   444
and arguably a more elegant one, at least in the sense that it is based on a full   445
likelihood expression – relies on the bifurcating properties of the trees that connect   446
related sequences, or conversely (with time flowing backwards) the coalescence of a set   447
of related samples into a common ancestor. The latter gives the name to the coalescent   448
theory, or simply, *the coalescent*, as described by Kingman (1982). Since then the   449
problem of estimating a Wright-Fisher (or Moran) population size from sequence data   450
using the coalescent has been extended to varying environments (Griffiths & Tavaré   451
1994), to implicitly defined population functions (Frost & Volz 2010), and more   452
generally to structured populations (Volz 2012).   453

The Beast 2 software (Bayesian Evolutionary Analysis by Sampling Trees) contains   454
many basic as well as advanced implementations of coalescent-based $MCMC$ inference   455
(Bouckaert et al. 2014); it is written in the Java programming language, and uses an   456
XML file to input the actual sequence data and the specifications to the multiple classes   457
involved in the calculations. A phylodynamics package is also available, which among   458
many things includes implementations of the SIR model (Kühnert). Other software for   459
that purpose is also available, notably as packages for the R programming language   460
(Paradis; Volz et al.). Beast 2 was chosen due to the existence of a general community   461
of users, its openness and extensibility, apparent propensity to phylodynamics   462
implementations, and especially the helpfulness of some of its earliest developers.   463

Nevertheless, there were no tools built in to the core software, nor any extension   464
packages (including the phylodynamics package developed by Kühnert) that were   465
suitable for direct application to the inference using the models pertinent to dengue   466
virus transmission as describe above. For one, a structured implementation is needed –   467
in the simplest case, viral sequences may be either in the human or in the mosquito   468

host, and intuitively coalescence can only happen if they are inside the same hosts (unfortunately, further details about structure in the coalescent for this kind of model are way out of scope, but see Volz (2012) for a complete description of how structure is incorporated in the context of disease transmission models). The implementation was greatly facilitated by code shared from a previous implementation from Rasmussen et al. (2014b), which consisted of an epidemiological model Java class and a structured coalescent likelihood computation class; these extensions could almost directly be applied to the SIR-vector model, and could straightforwardly be used for the SIRS model as well, and more or less easily adapted to all SIRX models.

The implementation of the two-serotype models, however, required additional tinkering. Because $DENV$ serotypes are only 60% similar by sequences, they are usually treated separately (de Bruycker-Nogueira et al. 2015; Castro et al. 2013, 2012). Therefore, they are not expected to find a common ancestor in the recent period of a few years of dengue transmission, in which case two separate trees are needed. The two-serotype epidemiological model produces the poulation dynamics function for both serotypes, so a single Java class was implemented with this model; however, the likelihoods have to be calculated for each tree, so two separate classes were implemented that used separate trees and substitution model parameters, but fetched (different, serotype-specific parts) of the same population model output. The combination of the two tree likelihoods was then the global likelihood of the one-epidemiology, two-tree model, and from then on could be used by the $MCMC$ algorithm in Beast 2 with no additional tinkering needed to perform inference. The same quality checking criteria as in the time series inference were used; one month of real time amounted to a few tens of millions of iterations in Beast 2 for the most complex models. Additionally, Beast 2 computed the effective number of samples (Gelman et al. 2013, chap. 11) as an additional metric to assess convergence.
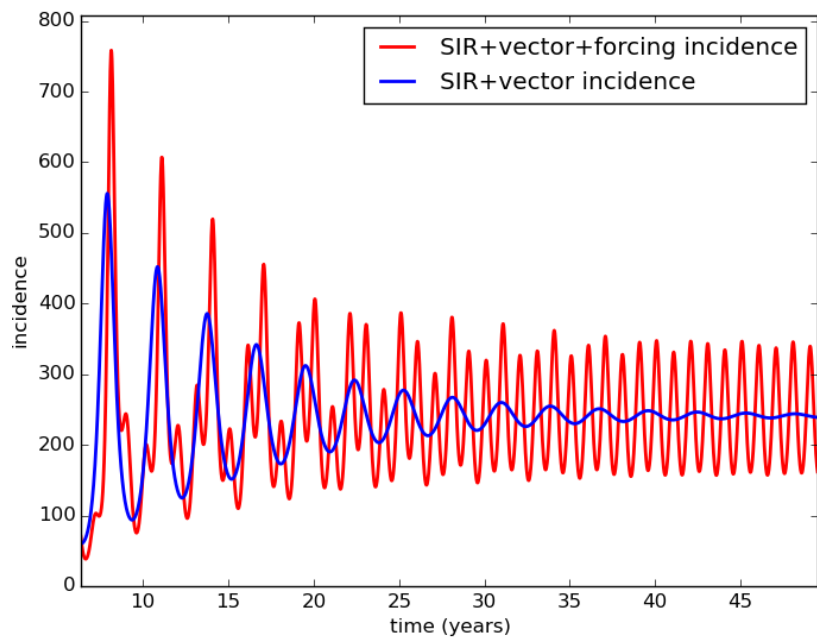

Figure 4. SIR+vector models with (red) and without seasonal forcing (blue) Parameter values in are: $\Omega = \beta = 1.0, b = 0.1, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, \alpha = 0.1, H = 1000000, M = 513789$.

3 Results

3.1 Forward model simulation

3.1.1 Predictions for disease incidence and prevalence

Figure 4 shows a deterministic simulation of the SIR-vector model without, and with seasonal forcing in the birth rates. It is worth noting, that although the amplitude of seasonal forcing is of the magnitude of 10%, oscillations can quite easily be greater than that through resonance of the natural (damped) oscillations and the forcing (Dushoff et al. 2004). The average incidence, however, is roughly the same in the presence or absence of forcing.

In the case of waning immunity, secondary infections are also possible, which potentially increases the total incidence at any point in time. Figure 5 shows the total incidence for the vector transmitted SIR, SIRx2, SIRx4, and SIRS models; for a set of parameter values, changing only the transmission rate within a certain range can result in a very similar incidence profile even in a deterministic setting.
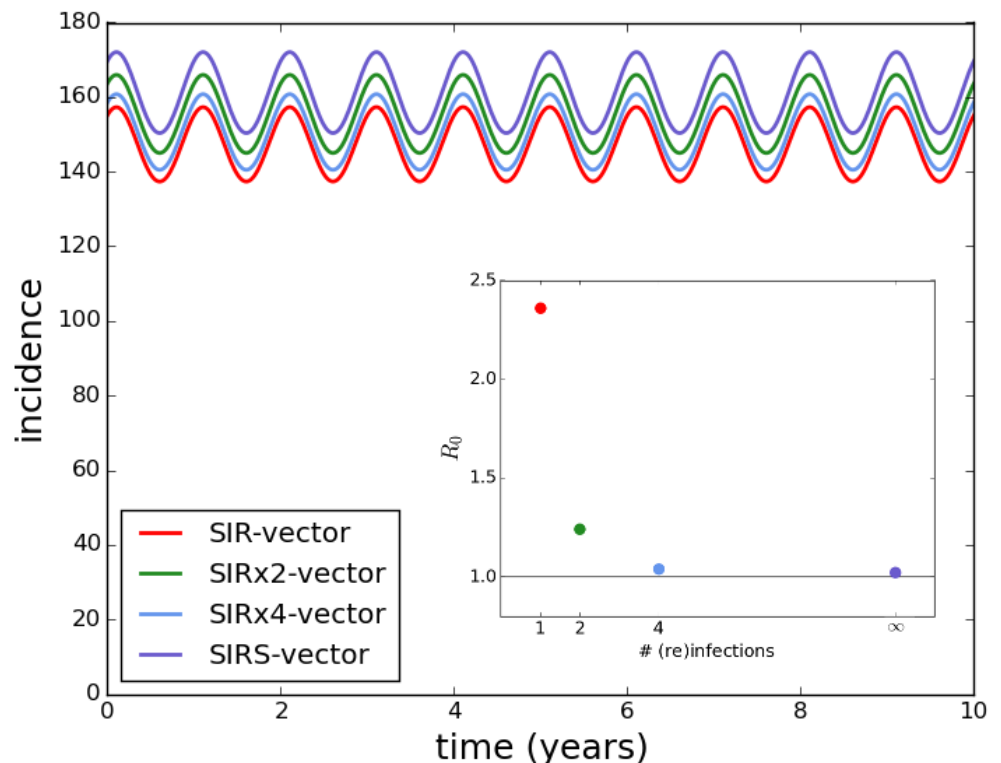

Figure 5. SIR+vector models in some of its variants: standard vector SIR (red), SIRx2 (green), SIRx4 (light blue), and SIRS (purple), showing approximately the same incidence levels at or near near the oscillatory equilibrium. The single changing parameter is transmission intensity for the SIR, SIRx2, SIRx4 and SIRS models, respectively: $\Omega = \beta = 0.2536; 0.1835; 0.1680; 0.1665$. Additional parameter values in are: $b = 0.1, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, \phi = 0.00165, \alpha = 0.02, \delta = 0, H = 1000000, M = 513789$.

Although for larger transmission rates the number of susceptibles may dominate and cause the models to output distinguishable time series, it must be acknowledged that for some parameter sets, model structure is not easily identifiable from the incidence pattern. Therefore, not only parameter values, but also model structure, are likely to be

critical to describe the transmission of dengue virus, and can have a great impact in the parameters estimated.
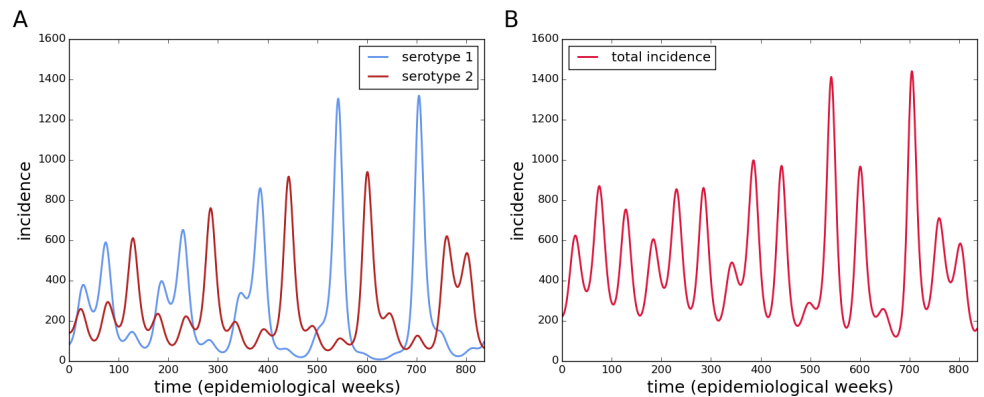
<div style="text-align: right">513</div>

Figure 6. Incidence dynamics for an explicit two-serotype dengue virus model. Shown incidence is that for each serotype separately (a), and the sum of new infections for both serotypes (b). Parameter values are: $b = 0.1, \Omega = 0.7, \beta = 0.7, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, \varphi = 0.00165, \alpha = 0.02, \delta = 0, H = 1000000, M = 513789$.

<div style="text-align: right">514</div>

The structure of the two-serotype model is not directly comparable to the previous models; nevertheless, the output of the model with the same common set of parameters (except for β) is shown in figure 6, except for transmission.

<div style="text-align: right">515
516
517</div>

Besides the obvious difference that this model has explicit series for both serotypes, the incidence pattern is less regular. A more striking qualitative difference is the fact that the numbers of infected individuals get much closer to zero (at least for individual serotypes); in a deterministic model it means just that, but in a stochastic model that may mean increased probability of extinction.

<div style="text-align: right">518
519
520
521
522</div>

3.1.2 Discrete time and individual-based implementations of vector transmission of DENV

<div style="text-align: right">523
524</div>

Discrete time and stochasticity in the system state transitions can significantly affect the oscillations in resonating systems (Dushoff et al. 2004). This is illustrated by the incidence outputs of both the vector SIR, and more extremely the vector SIRS models, in figure 7.

<div style="text-align: right">525
526
527
528</div>

The probability of extinction of the pathogen is generally low in those cases, although it is clear that numbers are much lower for the case without reinfection (vector SIR), such that if extinction doesn't commonly occur, the trajectories are nevertheless much more prone to the general process noise. Similarly, the intermediate models with two and four infections are prone to effects of discretization and stochasticity. Figure 8 shows the results for the remaining SIRX models.

<div style="text-align: right">529
530
531
532
533
534</div>

For the chosen parameters, extinction is more likely in the two-serotype model Figure 9 shows the results for this model.

<div style="text-align: right">535
536</div>

3.1.3 Summary of forward modeling predictions

<div style="text-align: right">537</div>

Models that are structured differently but with the same parameters, and therefore same basic reproductive number, will have different outputs and observed incidence and prevalence levels. Conversely, qualitative aspects of the observed disease data can be reproduced by different combinations of model structure and parameters. It is not at all trivial to determine what aspects of the model are known, or best approximate reality,

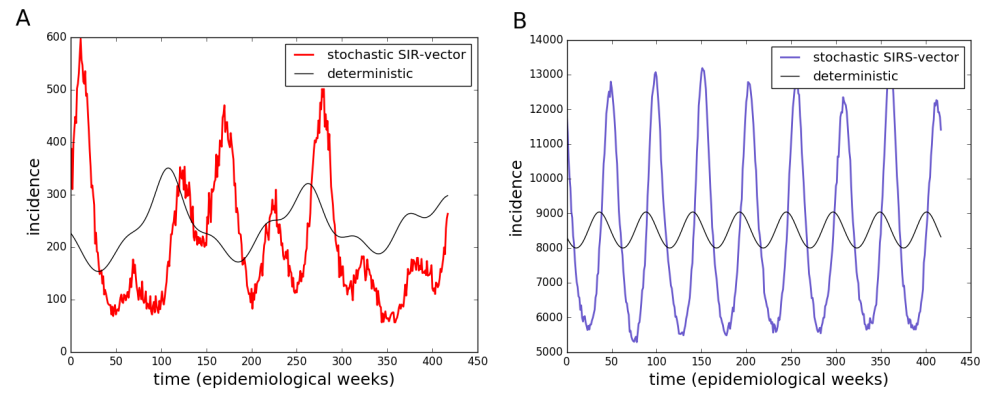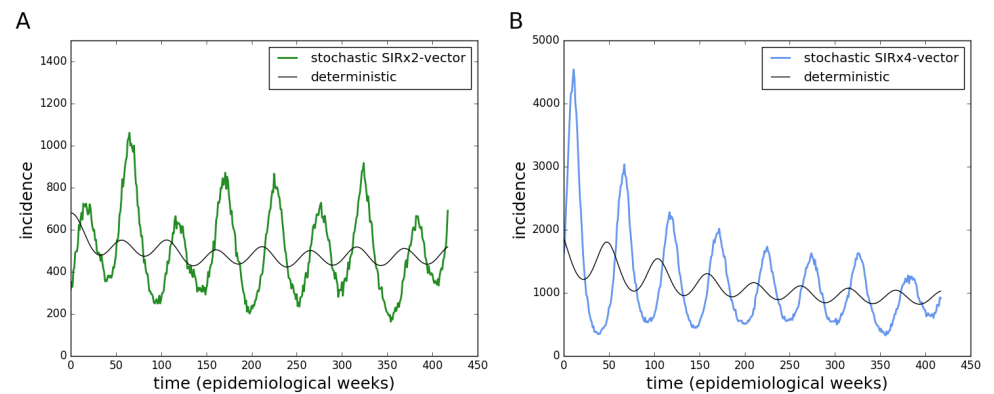<div style="text-align: right">538
539
540
541
542</div>

Figure 7. Stochastic output of vector SIR model (a), and of SIRS model (b). Gray lines show solutions to the deterministic system. Parameter values are: $b = 0.1, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, \phi = 0.00165, \alpha = 0.02, \delta = 0, H = 1000000, M = 513789.$, and $\Omega = \beta = 0.7$ for the SIR, and $\Omega = \beta = 0.47$, for the SIRS.

Figure 8. Stochastic output of vector SIRx2 model (a), and of SIRx4 model (b). Gray lines show solutions to the deterministic system. Parameter values are: $b = 0.1, \Omega = \beta = 0.7, m = 3.65 \cdot 10^{-5}, \gamma = 0.14, \phi = 0.00165, \alpha = 0.02, \delta = 0, H = 1000000, M = 513789.$

which can be treated as nuisances, and which of them are robust to structure or parametrization changes.

In the case of genetic diversity, it is even harder to distinguish clear relationships between the real data and the model predictions. Summaries are useful to make broad assessments about the data, but are usually not suitable for finer grained comparisons, nor model testing. A quantitative comparison is therefore needed to assess which model best represents reality, and what parameter values explain the processes of disease transmission.

The forward simulation approach is quite convenient when good estimates are available for all or most parameters, and there is good confidence in the model structure, or the outputs are robust or easy to evaluate for unknown parameters. Even for a reasonably large amount of synthetic (or pseudo-) data it is feasible to perform computation of genetic summaries in a reasonably short time without much algorithm optimization effort.

For complex, non-linear models it may, however, be difficult to thoroughly perform sensitivity analyses for more than a couple parameters and identify disease dynamics compatible with real data. For genetic diversity computations that rely on individual-based simulations it is even more costly and time-consuming to adopt the
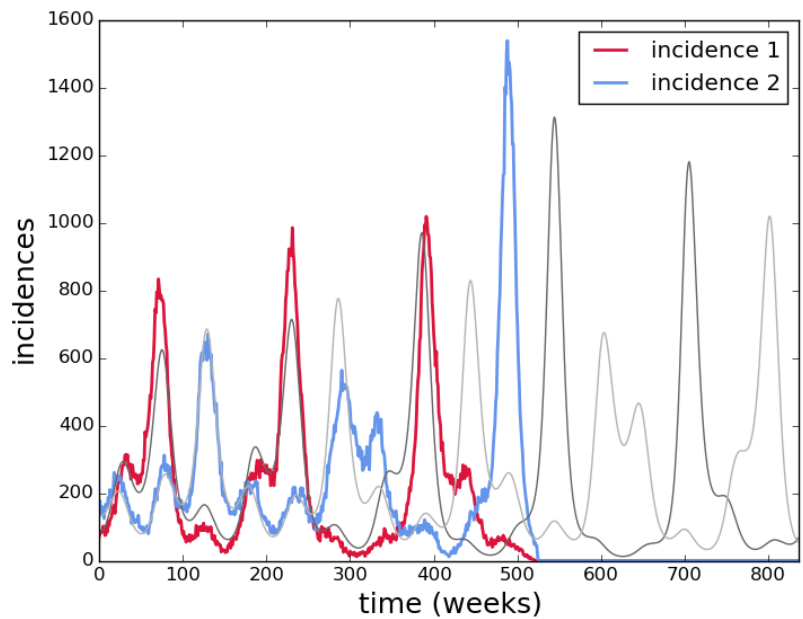
Figure 9. Stochastic output of two-sreotype model. Gray lines show solutions to the deterministic system. Parameter values are the same as before.

forward approach.   561

3.2   Inference: quantitative hypothesis testing   562

3.2.1   Time series-based inference   563

The *MCMC* algorithm produces samples of the joint posterior, which consist of a series   564
of parameter sets accepted by the method, the fit can be empirically computed by   565
simulating the model for a representative subset of these samples, and computing the   566
mean or median values of the model output. Credibility intervals can be similarly   567
computed by taking the score at some percentile (e.g. a 95% CI results if 2.5% and   568
97.5% are chosen). We present the mean computed as described above and call it the   569
"fit" hereafter, together with the 95% credibility intervals, unless otherwise stated.   570
    For inference purposes, the ratio of mosquito population M to human hosts H is the   571
estimated parameter, denoted hereafter as M_{ratio}. The mean time of immunity before   572
becoming susceptible, that is $1/\varphi$ or $1/\phi$ are the parameters actually estimated, and are   573
denoted just as such, the inverse of the original parameter.   574
    The easiest data set the inference method can fit is a deterministic simulation with   575
some noise added afterwards to the continuous solution. This best case but unlikely   576
scenario serves as proof of concept that there would be enough information in a data set   577
with this format. Figure 10, panel A shows the fit of the two-serotype model to   578
pseudodata produced by a deterministic simulation of 209 weeks (approximately four   579
years) with poisson noise added to each of the 209 time points, i.e. the model used for   580
estimation is the same as that used to produce the simulated data.   581
    The fit shown in is extremely good, considering the pseudodata is the output of a   582
highly nonlinear model, and that only an aggregated and partial observation of the   583
system is used (i.e. the sum of the changes in the infected compartment). Therefore,   584
while this is the best case scenario, it is not a given that it would be possible to   585
adequately fit the model, and furthermore estimate the parameters accurately – the   586
parameters could be structurally unidentifiable, or the amount or type of data could not   587

allow accurate estimation of the parameters. The estimates of the model parameters are 588
shown in figure 10, panel B, which displays the prior and posterior distributions. Most 589
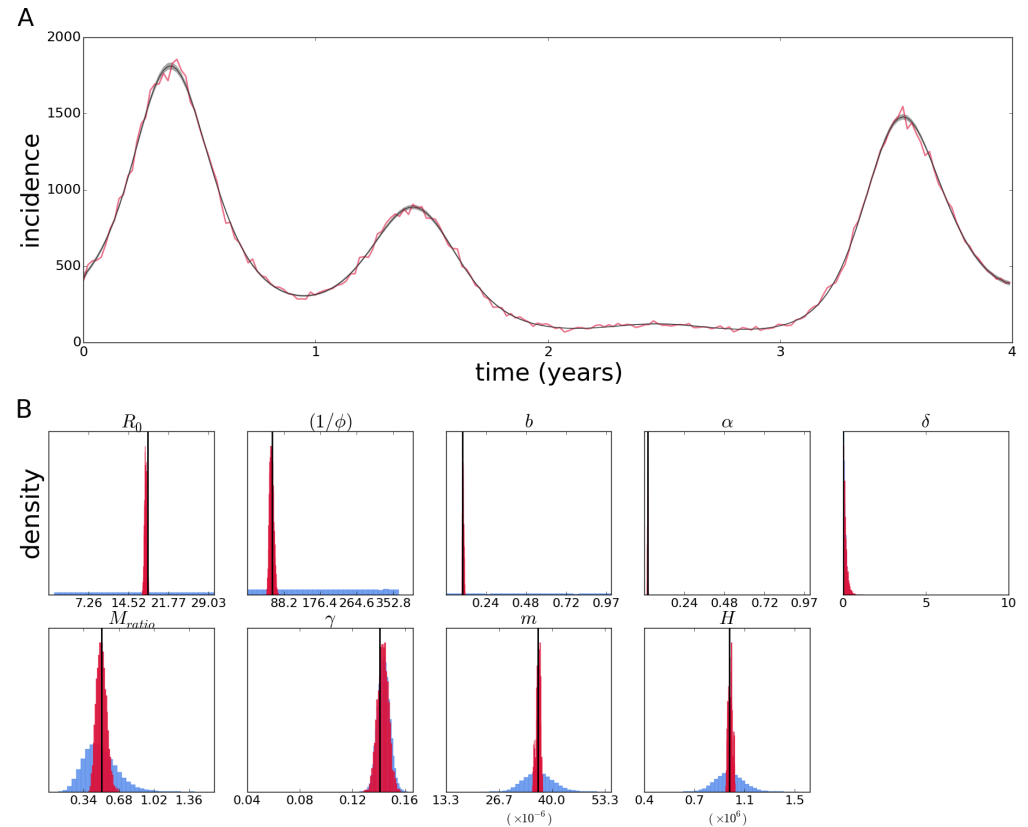estimates are quite precise even in the absence of informative priors.

Figure 10. Fit of the continuous two-serotype model with poisson likelihood to data simulated from the continuous model with poisson noise added afterwards (A). Posterior distributions (crimson) of epidemiological parameters (B) – prior distributions are shown in light blue (possibly not visible if the densities are very low compared to the posterior density). True values, i.e. parameter values used to simulate the data set, are shown as vertical black lines.

590

The case of the exact same estimation method applied to a data set produced by the 591
stochastic individual-based model simulation instead is shown in figure 11. While the fit 592
to the data is still quite good, the effect of stochasticity on the posterior estimates can 593
be observed as pronounced biases in some posterior distributions. 594
Unlike the previous case, gamma priors were used for the initial population states, 595
and they are nevertheless not as well estimated as before, but contain the true values 596
(not shown). Particularly important is the estimate for R_0, which is significantly lower 597
than the true value and the estimate of the temporary cross protection time (i.e. 598
strain-transcending immunity – the inverse of the rate of immunity waning φ – only 599
present in models with secondary infections), which is precisely estimated. 600
The observed biases are similar to those observed with a simpler vector SIR model 601
(see appendix,), and the fact that these are not observed with the continuous 602
simulation suggests that model complexity or data structure are not the main factors. 603
Indeed, fixing some parameters and/or initial conditions can improve estimation of 604
the remaining parameters. Appendix A shows that fixing all other epidemiological 605
parameters except for R_0 improves it estimate and fixing initial conditions improves it 606

further (figure A.2) However, it is not necessarily true that the more parameters fixed, the better the estimates, as leaving a few of the epidemiological parameters results in the best estimate of R_0 when compared to the above (figure A.3).

The *MCMC* algorithm allows empirical computation of the correlation between the parameters, since it relies on repeatedly sampling the posterior distribution of parameters. The biases could therefore be at least in part attributed to correlation between the estimates (not shown, but see discussion section); however, the problem seems to go beyond lack of identifiability of specific parameter combinations, since allowing some parameters to accommodate uncertainty can improve inference results. I further discuss parameter correlation, fixing parameters, uncertainty in estimates, and possible methods to get around these issues in the next section.
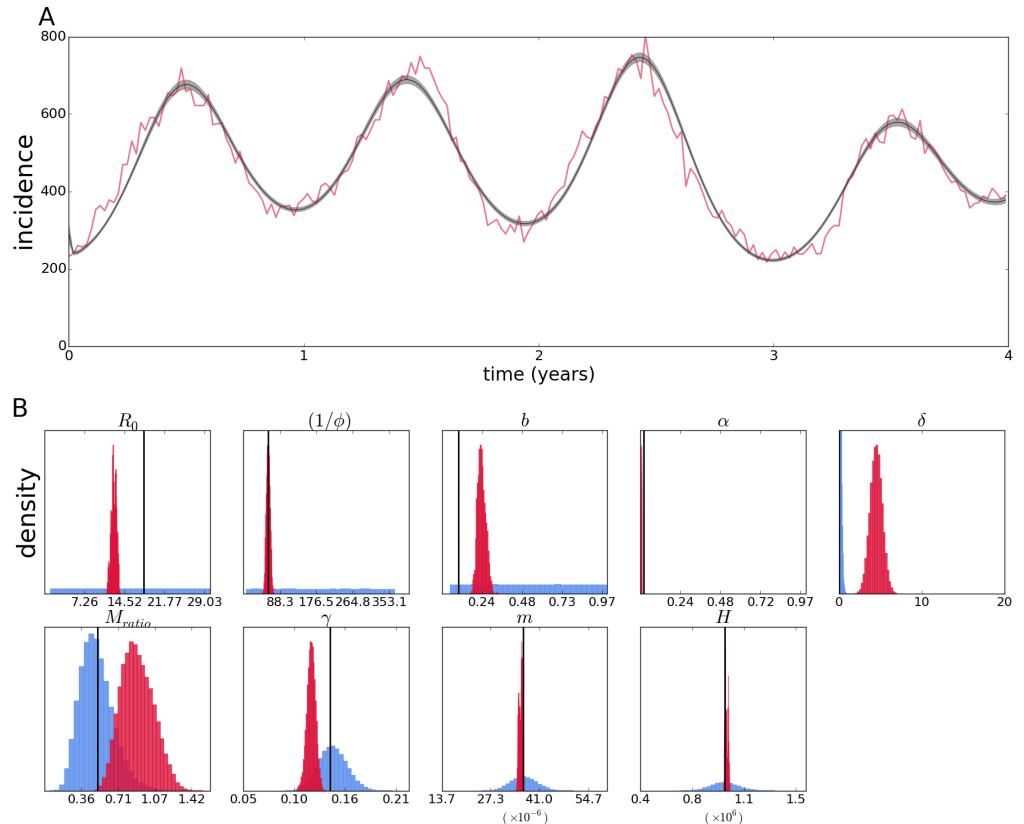

Figure 11. Fit of the two-seroytpe continuous model with poisson likelihood to data simulated from its stochastic version (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

Because incidence data most likely does not differentiate between infecting serotypes, a time series gives no information about the number of circulating serotypes, and therefore it is not possible to decide on single or multi-serotype models based on that alone. For inference purposes a number of circulating serotypes must be assumed when a transmission model is used.

The previous results assume the correct model structure (although not the correct error model) is known; with real data other factors such as model misspecification can introduce further sources of errors. Appendix A shows that using the vector SIRx2 model to estimate parameters from a time series simulated from the two-serotype model can throw estimates off (particularly that of R_0); using the same model on data simulated from the vector SIR model can have equally problematic results. That

highlights the importance of testing different alternative models for the same data set, when it is not possible to favor any specific model based on the data alone.

It is in principle possible to distinguish the infecting serotype, and even the order of infection, for time series data; however, that would require elaborated tests and/or record keeping for current and previous infections for every single recorded case. Sequence based inference relies on a sample of cases, not all possible records, and differentiates infecting serotype by default, which can get around some of these issues.

3.2.2 Genealogy-based inference

The aim of this section is to show results of methods comparable to that of time series-based inference. In Rio de Janeiro, the available dengue virus sequence data seems to be a result of concentrated efforts to obtain data representative of each single epidemic period, not an overall data set with particular features. To my knowledge, there is no specific goal for the entirety of the available data as to the total size, sampling interval, or representativity of both epidemic and inter-epidemic periods, for instance. Therefore, the simulated data set used for inference here is created to mimic more of an ideal yet feasible data set to collect. In practice that was done by deciding on the total number of sequences desired and sampling with a constant probability that would yield that expected value – given that binomial sampling was done every week, periods with greater number of cases would be more represented in the data.

For a constant probability to be specified, the total number of cases in the entire interval should be known before sampling, which cannot be the case in the real world; even if that was known a probability of sampling a sequence cannot be directly decided on by researchers. Nevertheless, the computational sampling scheme mimics the real world in that the greater the number of infected individuals, the more report to hospitals and health centers, and the probability of obtaining consent to get biological samples and sequencing them can be set by an arbitrary rate (e.g. a goal to obtain a sample from one out of 100 patients) that would yield a total number of samples over a usual epidemic.

For the vector SIR model, a pseudodata set was sampled from the last four years of a sixteen year run resulting in 106 simulated sequences with random mutations. The results from estimation based on that data are shown in figure 12. The posterior contains the true values of the parameters, although some are slightly biased, not centered around the true value.

Again an expected value and confidence intervals are computed from the sets of parameters in each step of the Markov Chain; however, figure 12A does not show a fit, since the time series data is not used for estimation, but rather a reconstruction of the incidence patterns. Nevertheless, this reconstruction may accurately depict the mean incidence if the sequence data is informative enough, and here it indeed reproduces oscillations on the same order as those in the data. The phase of the oscillations are not perfectly synchronized, and the amplitude is more regular than the actual incidence (although this is expected due to the stochastic data as opposed to the deterministic nature of the inference method; this is further discussed in the next section).

For the two-serotype model, the pseudo-sequences used here for inference were produced by the same run of the individual-based model as the time series pseudodata (in the previous subsection) – as would be the case with a real epidemic, where all possible new cases are recorded as incidence, and some of them are sampled for sequencing. The pseudo-sequence data is sampled from the same 209 weeks for which pseudo-incidence time series was recorded; the total simulation length was 16 years (therefore, the pseudo-time series in the previous subsection actually consists of the last 4 years of this longer run. Only the sequence data is used to produced the results that follow.
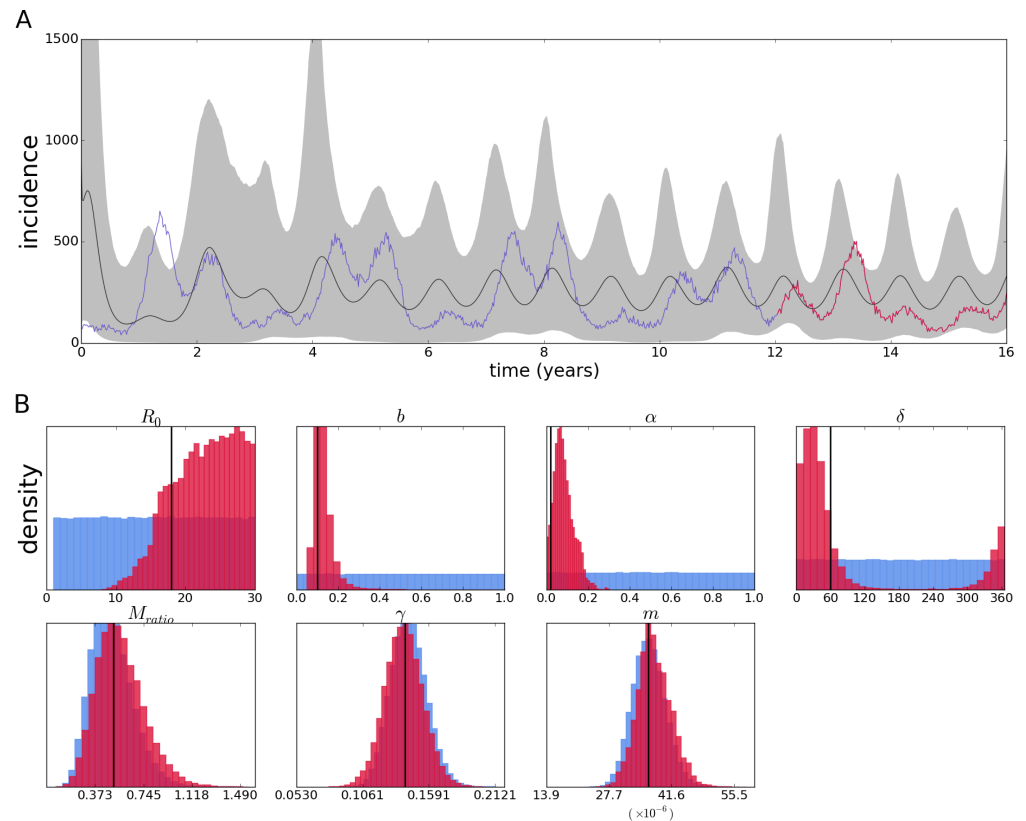
Figure 12. Reconstruction of the epidemic with a vector SIR model from sequences simulated by its corresponding individual-based model(A). Red portion of the time series denotes period from which samples were taken. Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B).

Inference using 119 type 1 sequences and 104 type 2 sequences sampled according to the above scheme from a population of simulated individuals is shown in figure 13.

The general *MCMC* settings are the same as with the time series wherever applicable, despite them being different implementations as explained in the methods section. Also, there is no way around the fact that some parameters in this estimation are absent and do not apply to the previous case – for instance the origin time of the epidemic (which for a time series is trivially defined as the first time point), the mutation rate and the tree itself (both of which have no impact on the incidence series). Otherwise, the estimates are generally comparable.

Biases are not nearly as pronounced as in the inference with the time series; Besides, the variables with gamma-distributed priors seem not to get much information from the likelihood and stay almost unaltered compared to their priors. The origin parameter, the time before the present when the epidemics starts, is fixed at 5852 days; this choice is discussed in the following sections. On the other hand, parameters with uninformative priors such as the mosquito mortality rate are quite well estimated; moreover, the basic reproductive number is quite accurately estimated.
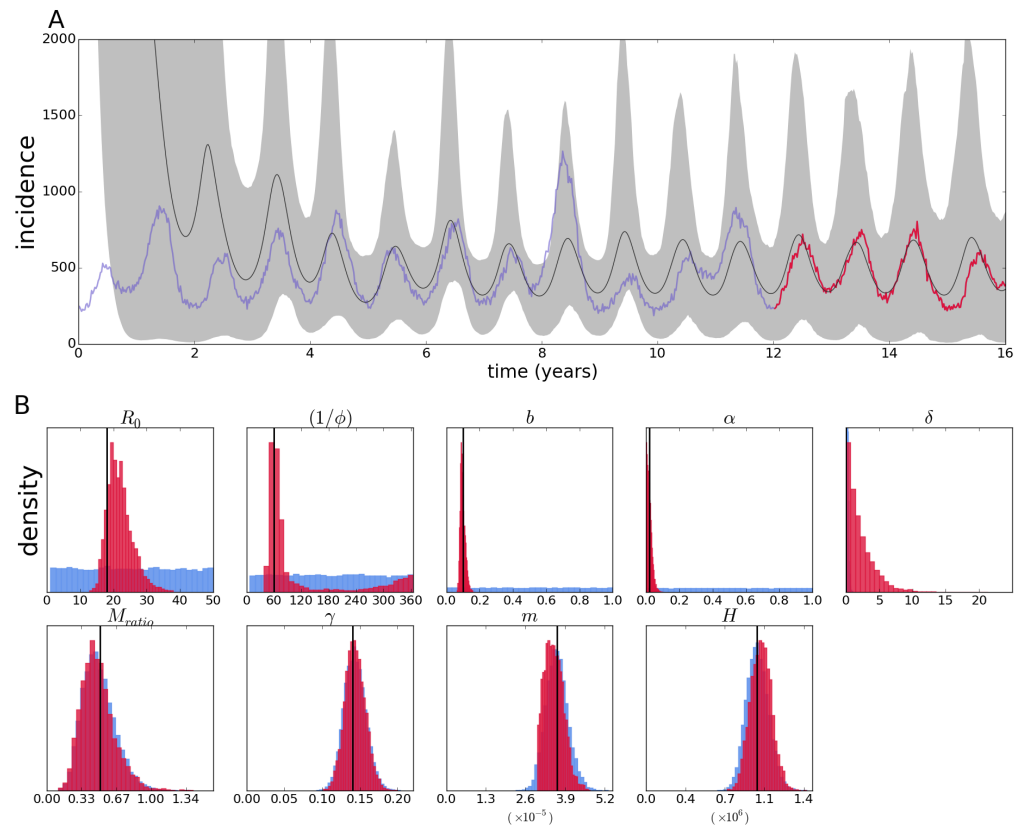
Figure 13. Reconstruction of the epidemic with a two-serotype model from sequences simulated by its corresponding individual-based model (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters (B).

3.2.3 Inference for epidemiological data from the city of Rio de Janeiro, Brazil

We apply the methods tested above to epidemiological data from the city of Rio de Janeiro, Brazil. Weekly incidence data for a period of approximately four years between the end of 2009 and that of 2013 was used to fit both the vector SIRx2 and the two-serotype models. Given the caveats observed for the simulated data, preliminary estimation was performed using the *MCMC* method for time series described above. In addition to the continuous model described above, a fixed scale (or reporting rate) parameter multiplies the incidence output to account for underreporting of cases in the recorded series – i.e. a parameter value lower than one means the actual epidemic is larger than the observed time series record.

The parameter estimates for the vector SIRx2 model are shown as posteriors (Figure 14B); the parameters with independent estimates (included in the inference method as prior probabilities) deviate from that expectation, notably the population size H, and human mortality rate m, both of which are grossly overestimated. It could be that there is a biological explanation for the departures from the independent estimates, but as seen in the previous subsection, bias and correlation between parameters is likely to affect the estimates. Another well accepted value is recovery happening in around a week $(1/\gamma)$; the posterior confidently places that at around twice that time. The biological parameters of interest without accepted values place the mosquito lifespan $(1/b)$ at around 22 days, and temporary immunity period $(1/\varphi)$ at 1 day, while R_0 is
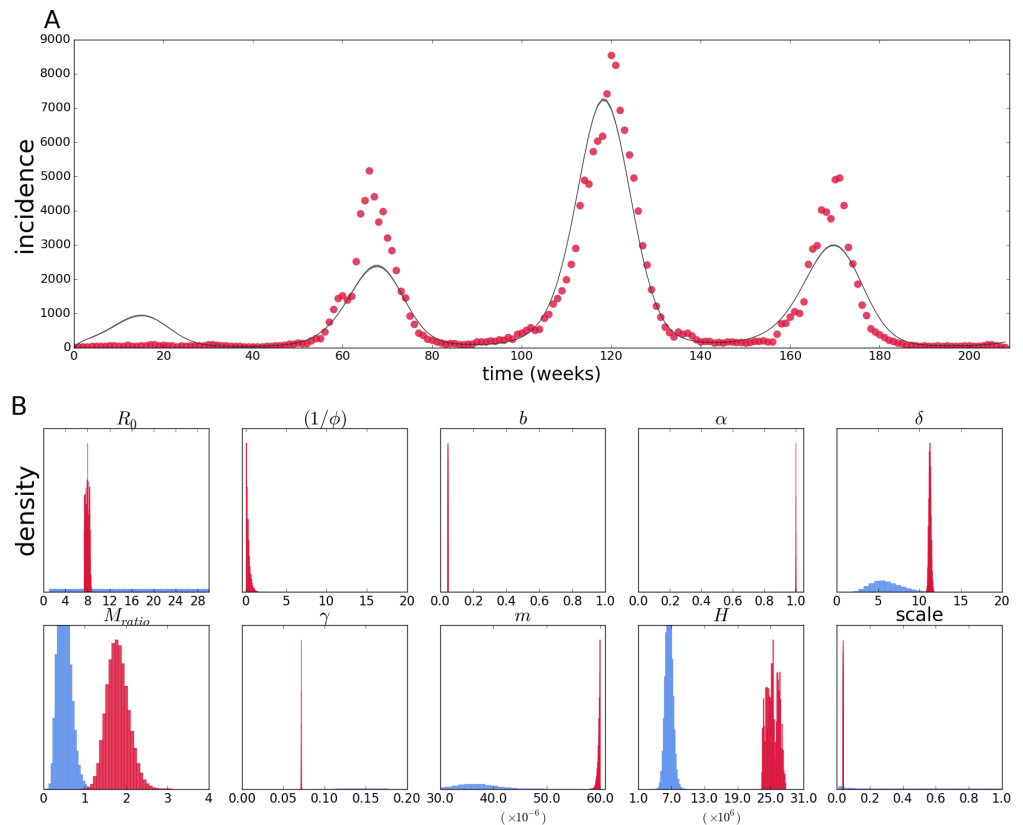
Figure 14. Fit of the vector SIRx2 continuous model with poisson likelihood to epidemiological data from Rio de Janeiro (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

estimated to be around 8. Parameters of more epidemiological interest are the reporting rate ("scale" parameter) close to only 4% and the vector to human ratio of close to 2 mosquitoes for every human host. The general fit (Figure 14A) is good, although the model predicts a lower incidence at the first and third peaks, as well as an early outbreak around the 20th epidemiological week, which is not present in the epidemiological data.

Next, the two-serotype model is fitted to the same data set; the results are shown in figure 15. The general fit is visibly improved. Interestingly, again some of the better known, or accepted, parameters deviate from expectation in the same direction as with the vector SIRx2 model; human lifespan $(1/m)$ is again underestimated, but population size H is inferred to be close to the census-recorded number. Recovery γ is again estimated to be twice as slow as the commonly accepted rate of a seven day long infection; mosquito mortality rate b also has a similar value to that in the one-serotype model.

The robustness of the estimates can be interpreted as a strong signal in the data for these parameters, although it is difficult to know how much they are affected by the deviation in the independently-estimated values, as well as the others with no accepted values such as the temporary immunity period $1/\varphi$. The robustness of some of the estimates can be further tested by fixing the parameters with well accepted values (γ, m, H). These results are in appendix A. Figure A.6 shows that while preventing parameter values very inconsistent with prior beliefs or independent estimates, most estimates are affected by the choice of free parameters estimated.
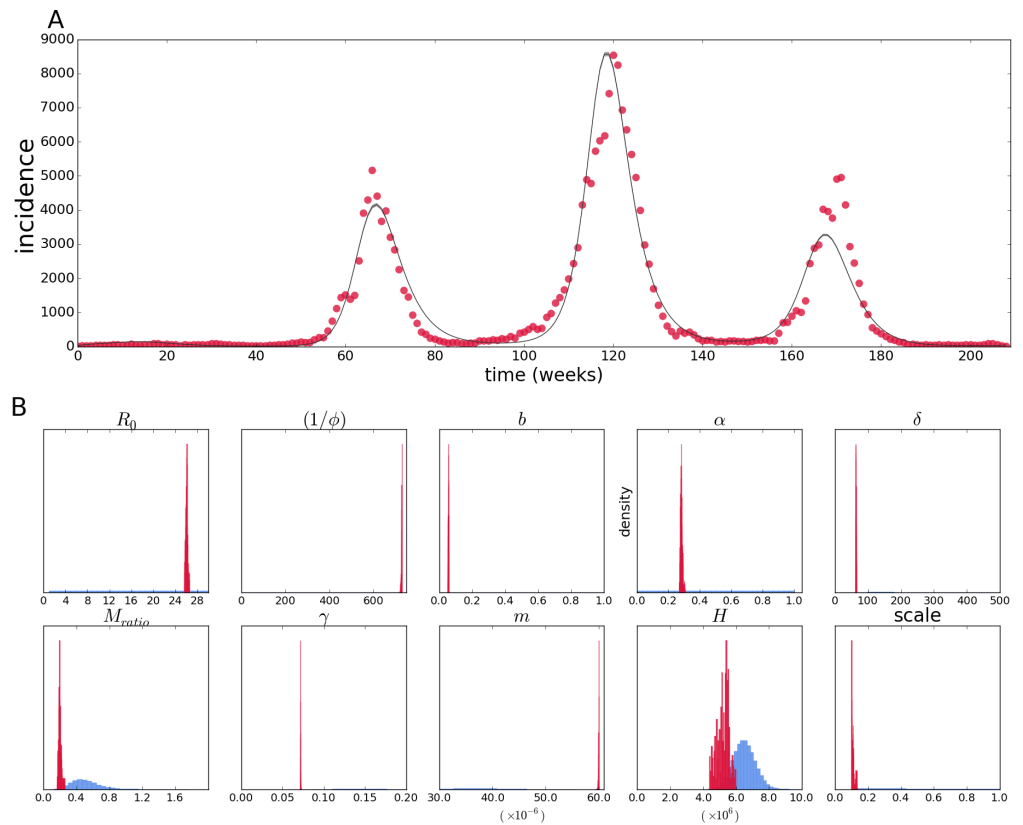
Figure 15. Fit of the two-serotype continuous model with poisson likelihood to epidemiological data from Rio de Janeiro (A). Posterior distributions (crimson color) and priors (light blue) of epidemiological parameters and initial conditions (B).

Contrary to the vector SIRx2 model, the estimate for R_0 is on the high end of the uniform prior probability assigned to the variable, around 26; the mosquito to human ratio is also opposite to the estimate of the previous model, being about twice as low as the mean of the prior distribution at 0.2, and the cross protection time is in the complete opposite side of the range at 729 days. Mosquito lifespan is close to the previous estimate, 18 days, and the reporting rate is on the order of 10%

The values inferred, however, must not be taken at face value, considering the limitations observed in the inference using the simulated data. We discuss that further in the next section.

4   Discussion

The issue of modeling dengue virus transmission is far from trivial, as is probably modeling most infectious diseases. Writing on a paper or programming a SIR-like model into a computer is indeed straightforward; that in its simplicity that alone can capture unintuitive features from disease transmission is quite surprising, but going beyond that is a long and winding road, and extending the basic models in the right direction even more so. It may also be reasonable to ignore variation and selection in some cases (even out of necessity), which greatly simplifies model formulation; still under these simplified models there is plenty to criticize in the existent body of work in modeling dengue transmission (Johansson et al. 2011). Basic simplifications whose consequences are

taken for granted may also rear up their ugly heads when the model needs to be   758
confronted to real data.   759

Inference is a logical next step, although it is an arduous task that may not be as   760
rewarding as straightforward simulation. Despite perfection being far out of sight, the   761
critical exercise can be a constructive one. The difficulties in the case of dengue are   762
illustrated by the issue of multiple serotype interactions, and more subtle effects that   763
are expected – like asymmetry between serotypes or antibody dependent enhancement.   764

From the forward simulation point of view, it is clear that the structure of the model   765
radically affects the observed incidence – in what can be understood as the crucial   766
factor of the availability of susceptibles – therefore interpreting epidemic series in terms   767
of a single preferred model and its associated parameters is guaranteed to be   768
problematic. The impact of structural differences for a given parameter set – as shown   769
by the forward simulations – is evident, others may not be so clear beforehand, but are   770
indisputable once we become aware of it; for instance, heterogeneity in any rates is more   771
than expected, it must be present, and can have dramatic consequences (Gomes et al.   772
2012, 2016). These effects are likely to be larger than those of small asymmetries.   773

A baseline problem of this order is likely not to be solved by adding parameters and   774
processes to the same model structure (Johansson et al. 2011), but instead the basic   775
structures and multiple extensions have to be systematically compared. In theory this   776
problem could be solved by formally comparing the performance of all available models   777
against real data; in practice, the task is a harder one: a record of a time series that   778
aggregates all kinds of infections may lack the information necessary to distinguish   779
between alternative models.   780

Identifiability analyses may uncover structural features of the model that may   781
prevent inference of particular parameters, and may point to reparametrization of   782
combinations that prevent structural identifiability issues (Bellu et al. 2007); however, it   783
can be cumbersome to implement for larger models (Eisenberg et al. 2013), and it is   784
uncommon (and possibly not feasible) that researchers in the different communities go   785
about all the methods that could improve their results. The analysis of the data   786
simulated by a continuous model strongly suggests that there are no severe structural   787
issues with the method, although I did not perform any of the above-mentioned   788
analyses. Alternative to structural identifiability analyses are more empirical   789
assessments of parameter inferability (Toni et al. 2009), and model comparison based on   790
information criteria (Gelman et al. 2013, chap. 7).   791

Formal model selection criteria rely on a reasonable fit by the different models;   792
conversely, there should be enough information in the data to grant support to the most   793
appropriate model – for instance, a more or less regular oscillating pattern produced by   794
a two-serotype model may be easily reproduced by a one-serotype model, and the latter   795
may be favored for having fewer parameters; what is more, a single cosine function (or a   796
couple of them) could fit the pattern with fewer of parameters, but that does not   797
change the fact that the data was produced by, and the correct model is still, a   798
two-serotype model.   799

An example of the difficulties mentioned above are the estimates obtained here from   800
incidence data of the city of Rio de Janeiro. Similar parameter values for different   801
models suggest robustness in the estimates; nevertheless, large deviations from   802
independent estimates may call into question these supposedly robust estimates.   803
Conversely, disparate values for different models may point to the inadequacy of one of   804
the models, and lack of robustness of estimates under model misspecification, but it does   805
not guarantee that either estimate is correct – it is not possible to check deviation from   806
the correct value if no independent estimate is available. Fixing the known parameters   807
can be an additional constraint to the parameter space; however, it may force the   808
remaining parameters into erroneous values due to the decreased flexibility in the model.   809

The more fine-grained the data is, the more precise can the comparison between $_{810}$ models be, but that assumes that some of the models can explain the data well enough. $_{811}$ It may, however, be the case that more detailed data causes the models only to fail more $_{812}$ miserably than before. In the end, all models are approximations (at best, reasonable $_{813}$ ones, but most likely rather crude ones), so that the multiple aspects of model and $_{814}$ inference framework must also improve as data becomes better and more plentiful. $_{815}$

The use of sequence data for model-based inference presents itself as both an exciting $_{816}$ perspective as well as a challenge: on the one hand it can make important distinctions $_{817}$ such as genotypes and serotypes of pathogens, and by design allow inference about the $_{818}$ entire population to be derived from a sample. It also carries, by default, information $_{819}$ about more than one time series at the same time – i.e. incidence, prevalence and, if $_{820}$ applicable, migration (Volz 2012). On the other hand fitting a model to sequences relies $_{821}$ on elaborate constructs that are difficult to visualize and evaluate – some of the $_{822}$ improvements that come with new kinds of data are therefore not without new issues. $_{823}$

One thing that seems to be unique to epidemiological models is that the data $_{824}$ associated to the pathogens can be acquired simultaneously in different formats, so $_{825}$ inference with one kind of data may be independently validated by another kind (e.g. $_{826}$ sequence-based reconstruction of the epidemic can be compared to incidence data, as $_{827}$ shown for pseudo data). Alternatively, these and other kinds of data (serological, vector $_{828}$ population data, etc) can be used in combination to improve estimation, if the problem $_{829}$ is scarcity of data. $_{830}$

This aspect is probably an important contrast to coalescent-based methods in fields $_{831}$ like conservation genetics, where great strides have been made to incorporate processes $_{832}$ like recombination (McVean & Cardin 2005), structured environments (as opposed to $_{833}$ change in effective population size) (Mazet et al. 2016) and even allow inference from a $_{834}$ single recombining genome (Li & Durbin 2011), but where very little validation of $_{835}$ alternative models exists, especially with alternative types of data, which are not $_{836}$ available for hundreds or thousands of years ago. Coalescent methods in epidemiology $_{837}$ offer the opportunity of trying the methods, and validating them with more stringent $_{838}$ criteria. $_{839}$

Increasingly it seems that scarcity is not the main problem (Pybus et al. 2013), but $_{840}$ rather the difficulty of inferring multi-dimensional parameter sets (or constraining their $_{841}$ space enough via independent estimation of individual parameters), and formally $_{842}$ comparing full models in a way that makes the estimates actually useful, in addition to $_{843}$ more subtle aspects of data collection (Frost et al. 2015). $_{844}$

Although its effects are clear in the individual-based simulations, the issue of $_{845}$ stochasticity in the system state was not directly tackled here; therefore, a deterministic $_{846}$ model can be forced to fit a different trajectory only by changing its parameters, even if $_{847}$ the deviation is caused by chance. Methods such as particle filtering (or Sequential $_{848}$ Monte Carlo) allow tracking of the stochastic system state along time (Ionides et al. $_{849}$ 2006); these have been incorporated into genealogy-based estimation (Rasmussen et al. $_{850}$ 2014a, 2011), potentially solving the issue of the effect of stochasticity in the population $_{851}$ immunological states. These methods apply straightforwardly to population $_{852}$ immunological states along time, which in the case on inference from time series is $_{853}$ directly correspondent to the likelihood; in the coalescent-based estimation that is one $_{854}$ of the components of the inference model, but it is not as easy to illustrate genetic drift $_{855}$ in a similar way. Implementing methods that account for stochasticity is beyond the $_{856}$ scope of this thesis, although the results shown strongly suggest that the simple fact $_{857}$ that stochasticity is present can hamper progress in an otherwise simple task. $_{858}$

It can be tempting to focus specifically on the fine-tuning mathematical models, $_{859}$ development new inference methods, and on extensive efforts to gather comprehensive $_{860}$ data sets, but it is important to take into account how the weaknesses of each of the $_{861}$

steps compound into a larger impediment. Concentrating particularly into just one of 862
these (or other even more particular) aspects may prevent a realistic use of data-driven, 863
model-based analysis. I have shown how model structure, assumptions about 864
stochasticity, prior information and data requirements all deserve specific treatments 865
lest the lack thereof introduces or amplifies biases and inaccuracies in the results, and 866
therefore hope to have contributed to the integration of model building, inference 867
frameworks, as well as future efforts to gather epidemiological data of different kinds. 868

Acknowledgements

References

1. Adams, B., Holmes, E. C., Zhang, C., Mammen, M. P., Nimmannitya, S., Kalayanarooj, S. & Boots, M. 2006 Cross-protective immunity can account for the alternating epidemic pattern of dengue virus serotypes circulating in Bangkok. Proc Natl Acad Sci U S A 103, 14234–14239. (doi:10.1073/pnas.0602768103)

2. Andrieu, C., Doucet, A. & Holenstein, R. 2010 Particle Markov chain Monte Carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 269–342. (doi:10.1111/j.1467-9868.2009.00736.x)

3. Araújo, J. M. G. de, Bello, G., Schatzmayr, H. G., Santos, F. B. D. & Nogueira, R. M. R. 2009 Dengue virus type 3 in Brazil: a phylogenetic perspective. Mem. Inst. Oswaldo Cruz 104, 526–529.

4. Bellu, G., Saccomani, M. P., Audoly, S. & D'Angiò, L. 2007 DAISY: A new software tool to test global identifiability of biological and physiological systems. Computer Methods and Programs in Biomedicine 88, 52–61. (doi:10.1016/j.cmpb.2007.07.002)

5. Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. 2015 GenBank. Nucleic Acids Research 43, D30–D35. (doi:10.1093/nar/gku1216)

6. Bhatt, S. et al. 2013 The global distribution and burden of dengue. Nature 496, 504–507. (doi:10.1038/nature12060)

7. Bianco, S., Shaw, L. B. & Schwartz, I. B. 2009 Epidemics with multistrain interactions: the interplay between cross immunity and antibody-dependent enhancement. Chaos 19, 043123. (doi:10.1063/1.3270261)

8. Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A. & Drummond, A. J. 2014 BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol 10, e1003537. (doi:10.1371/journal.pcbi.1003537)

9. de Bruycker-Nogueira, F., Nogueira, R. M. R., Faria, N. R. D. C., Simões, J. B. S., Nunes, P. C. G., de Filippis, A. M. B. & Santos, dos, F. B. 2015 Insights of the genetic diversity of DENV-1 detected in Brazil in 25years: Analysis of the envelope domain III allows lineages characterization. Infection, Genetics and Evolution 34, 126–136. (doi:10.1016/j.meegid.2015.07.007)

10. Castro, M. G., de Nogueira, F. B., Nogueira, R. M. R., Lourenço-de-Oliveira, R. & Santos, dos, F. B. 2013 Genetic variation in the 3' untranslated region of dengue virus serotype 3 strains isolated from mosquitoes and humans in Brazil. Virol. J. 10, 3. (doi:10.1186/1743-422X-10-3)

11. Castro, M. G. de et al. 2012 Dengue virus type 4 in Niterói, Rio de Janeiro: the role of molecular techniques in laboratory diagnosis and entomological surveillance. Mem. Inst. Oswaldo Cruz 107, 940–945. (doi:10.1590/S0074-02762012000700017)

12. Chikaki, E. & Ishikawa, H. 2009 A dengue transmission model in Thailand considering sequential infections with all four serotypes. J Infect Dev Ctries 3, 711–722.

13. Drummond, A. J. & Bouckaert, R. R. 2015 Bayesian evolutionary analysis with BEAST. Cambridge: Cambridge University Press. (doi:10.1017/cbo9781139095112)

14. Dushoff, J., Plotkin, J. B., Levin, S. A. & Earn, D. J. D. 2004 Dynamical resonance can account for seasonality of influenza epidemics. Proc Natl Acad Sci U S A 101, 16915–16916. (doi:10.1073/pnas.0407293101)

15. Eisen, L. & Moore, C. G. 2013 Aedes(Stegomyia) aegyptiin the Continental United States: A Vector at the Cool Margin of Its Geographic Range. Journal of Medical Entomology 50, 467–478. (doi:10.1603/ME12245)

16. Eisenberg, M. C., Robertson, S. L. & Tien, J. H. 2013 Identifiability and estimation of multiple transmission pathways in cholera and waterborne disease. Journal of Theoretical Biology 324, 84–102. (doi:10.1016/j.jtbi.2012.12.021)

17. Favier, C., Schmit, D., Müller-Graf, C. D. M., Cazelles, B., Degallier, N., Mondet, B. & Dubois, M. A. 2005 Influence of spatial heterogeneity on an emerging infectious disease: the case of dengue epidemics. Proceedings of the Royal Society B: Biological Sciences 272, 1171–1177. (doi:10.1098/rspb.2004.3020)

18. Frost, S. D. W., Pybus, O. G., Gog, J. R., Viboud, C., Bonhoeffer, S. & Bedford, T. 2015 Eight challenges in phylodynamic inference. Epidemics 10, 88–92. (doi:10.1016/j.epidem.2014.09.001)

19. Frost, S. D. W. & Volz, E. M. 2010 Viral phylodynamics and the search for an 'effective number of infections'. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365, 1879–1890. (doi:10.1098/rstb.2010.0060)

20. Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 2013 Bayesian data analysis, Third Edition. CRC Press. (doi:10.1080/01621459.2014.963405)

21. Gomes, M. G. M., Aguas, R., Lopes, J. S., Nunes, M. C., Rebelo, C., Rodrigues, P. & Struchiner, C. J. 2012 How host heterogeneity governs tuberculosis reinfection? Proceedings of the Royal Society B: Biological Sciences 279, 2473–2478. (doi:10.1098/rspb.2011.2712)

22. Gomes, M. G. M., Gjini, E., Lopes, J. S., Souto-Maior, C. & Rebelo, C. 2016 A theoretical framework to identify invariant thresholds in infectious disease epidemiology. Journal of Theoretical Biology 395, 97–102. (doi:10.1016/j.jtbi.2016.01.029)

23. Gordo, I., Gomes, M. G. M., Reis, D. G. & Campos, P. R. A. 2009 Genetic diversity in the SIR model of pathogen evolution. Plos One 4, e4876. (doi:10.1371/journal.pone.0004876)

24. Gordo, I. & Campos, P. R. A. 2007 Patterns of genetic variation in populations of infectious agents. BMC evolutionary biology 7, 116. (doi:10.1186/1471-2148-7-116)

25. Griffiths, R. C. & Tavaré, S. 1994 Sampling Theory for Neutral Alleles in a Varying Environment. Philosophical Transactions of the Royal Society of London B: Biological Sciences 344, 403–410. (doi:10.1098/rstb.1994.0079)

26. Guzman, M. G. & Vazquez, S. 2010 The complexity of antibody-dependent enhancement of dengue virus infection. Viruses 2, 2649–2662. (doi:10.3390/v2122649)

27. Hadinegoro, S. R. et al. 2015 Efficacy and Long-Term Safety of a Dengue Vaccine in Regions of Endemic Disease. N. Engl. J. Med. 373, 1195–1206. (doi:10.1056/NEJMoa1506223)

28. Halstead, S. B. 2007 Dengue. Lancet 370, 1644–1652. (doi:10.1016/S0140-6736(07)61687-0)

29. Ionides, E. L., Bretó, C. & King, A. A. 2006 Inference for nonlinear dynamical systems. Proc Natl Acad Sci U S A 103, 18438–18443. (doi:10.1073/pnas.0603181103)

30. Johansson, M. A., Hombach, J. & Cummings, D. A. T. 2011 Models of the impact of dengue vaccines: A review of current research and potential approaches. Vaccine 29, 5860–5868. (doi:10.1016/j.vaccine.2011.06.042)

31. Keeling, M. J. & Rohani, P. 2008 Modeling infectious diseases in humans and animals. Princeton University Press.

32. Kingman, J. F. C. 1982 The coalescent. Stochastic Processes and their Applications 13, 235–248. (doi:10.1016/0304-4149(82)90011-4)

33. Kliks, S. C., Nisalak, A., Brandt, W. E., Wahl, L. & Burke, D. S. 1989 Antibody-Dependent Enhancement of Dengue Virus Growth in Human Monocytes as a Risk Factor for Dengue Hemorrhagic Fever.

34. Koelle, K. & Rasmussen, D. A. 2012 Rates of coalescence for common epidemiological models at equilibrium. Journal of the Royal Society, Interface / the Royal Society 9, 997–1007. (doi:10.1098/rsif.2011.0495)

35. Kraemer, M. U. G. et al. 2015 The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. eLife 4, e08347. (doi:10.7554/eLife.08347)

36. Kuehnert D. Phylodynamics package. https://github.com/BEAST2-Dev/phylodynamics. Accessed: 2016-06-15

37. Li, H. & Durbin, R. 2011 Inference of human population history from individual whole-genome sequences. Nature 475, 493–496. (doi:10.1038/nature10231)

38. Mazet, O., Rodríguez, W., Grusea, S., Boitard, S. & Chikhi, L. 2016 On the importance of being structured: instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? Heredity (Edinb) 116, 362–371. (doi:10.1038/hdy.2015.104)

39. McVean, G. A. T. & Cardin, N. J. 2005 Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society of London B: Biological Sciences 360, 1387–1393. (doi:10.1098/rstb.2005.1673)

40. Miagostovich, M. P., Sequeira, P. C., Santos, Dos, F. B., Maia, A., Nogueira, R. M. R., Schatzmayr, H. G., Harris, E. & Riley, L. W. 2003 Molecular typing of dengue virus type 2 in Brazil. Rev. Inst. Med. Trop. Sao Paulo 45, 17–21.

41. Miagostovich, M. P., Santos, dos, F. B., Fumian, T. M., Guimarães, F. R., da Costa, E. V., Tavares, F. N., Coelho, J. O. & Nogueira, R. M. R. 2006 Complete genetic characterization of a Brazilian dengue virus type 3 strain isolated from a fatal outcome. Mem. Inst. Oswaldo Cruz 101, 307–313.

42. Nagao, Y. & Koelle, K. 2008 Decreases in dengue transmission may act to increase the incidence of dengue hemorrhagic fever. Proceedings of the National Academy of Sciences of the United States of America 105, 2238–2243. (doi:10.1073/pnas.0709029105)

43. Otero, M. & Solari, H. G. 2010 Stochastic eco-epidemiological model of dengue disease transmission by Aedes aegypti mosquito. Math Biosci 223, 32–46. (doi:10.1016/j.mbs.2009.10.005)

44. Paradis, E. coalescentMCMC R Package. http://colgem.r-forge.r-project.org/ Accessed: 2016-06-15

45. Patil, A., Huard, D. & Fonnesbeck, C. J. 2010 PyMC: Bayesian Stochastic Modelling in Python. Journal of statistical software 35, 1–81.

46. Pybus, O. G., Fraser, C. & Rambaut, A. 2013 Evolutionary epidemiology: preparing for an age of genomic plenty. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 368, 20120193–20120193. (doi:10.1098/rstb.2012.0193)

47. Rasmussen, D. A., Volz, E. M. & Koelle, K. 2014 Phylodynamic Inference for Structured Epidemiological Models. PLoS Comput Biol 10, e1003570. (doi:10.1371/journal.pcbi.1003570)

48. Rasmussen, D. A., Boni, M. F. & Koelle, K. 2014 Reconciling phylodynamics with epidemiology: the case of dengue virus in southern Vietnam. Mol Biol Evol 31, 258–271. (doi:10.1093/molbev/mst203)

49. Rasmussen, D. A., Ratmann, O. & Koelle, K. 2011 Inference for Nonlinear Epidemiological Models Using Genealogies and Time Series. PLoS Comput Biol 7, e1002136. (doi:10.1371/journal.pcbi.1002136)

50. Rico-Hesse, R. 2010 Dengue virus virulence and transmission determinants. Curr. Top. Microbiol. Immunol. 338, 45–55. (doi:10.1007/978-3-642-02215-9_4)

51. Rohani, P., Green, C. J., Mantilla-Beniers, N. B. & Grenfell, B. T. 2003 Ecological interference between fatal diseases. Nature 422, 885–888. (doi:10.1038/nature01542)

52. Ross, R. 1916 An Application of the Theory of Probabilities to the Study of a priori Pathometry. Part I. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 92, 204–230. (doi:10.1098/rspa.1916.0007)

53. Salje, H. et al. 2012 Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. Proceedings of the National Academy of Sciences of the United States of America 109, 9535–9538. (doi:10.1073/pnas.1120621109)

54. Santos, dos, F. B., Miagostovich, M. P., Nogueira, R. M. R., Edgil, D., Schatzmayr, H. G., Riley, L. W. & Harris, E. 2002 Complete nucleotide sequence analysis of a Brazilian dengue virus type 2 strain. Mem. Inst. Oswaldo Cruz 97, 991–995.

55. De Simone, T. S., Nogueira, R. M. R., Araújo, E. S. M., Guimarães, F. R., Santos, F. B., Schatzmayr, H. G., Souza, R. V., Teixeira Filho, G. & Miagostovich, M. P. 2004 Dengue virus surveillance: the co-circulation of DENV-1, DENV-2 and DENV-3 in the State of Rio de Janeiro, Brazil. Trans. R. Soc. Trop. Med. Hyg. 98, 553–562. (doi:10.1016/j.trstmh.2003.09.003)

56. Sistema de Informações de Agravos de Notificação : http://ces.ibge.gov.br/base-de-dados/metadados/ministerio-da-saude/sistema-de-informacoes-de-agravos-de-notificacao-sinan.html. Accessed: 2016-06-23

57. Smith, D. L., Battle, K. E., Hay, S. I., Barker, C. M., Scott, T. W. & McKenzie, F. E. 2012 Ross, macdonald, and a theory for the dynamics and control of mosquito-transmitted pathogens. Public Library of Science. (doi:10.1371/journal.ppat.1002588)

58. Stoddard, S. T. et al. 2013 House-to-house human movement drives dengue virus transmission. Proceedings of the National Academy of Sciences of the United States of America 110, 994–999. (doi:10.1073/pnas.1213349110)

59. Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. H. 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of The Royal Society Interface 6, 187–202. (doi:10.1098/rsif.2008.0172)

60. Vasilakis, N., Cardosa, J., Hanley, K. A., Holmes, E. C. & Weaver, S. C. 2011 Fever from the forest: prospects for the continued emergence of sylvatic dengue virus and its impact on public health. Nat. Rev. Microbiol. 9, 532–541. (doi:10.1038/nrmicro2595)

61. Vazquez-Prokopec, G. M., Kitron, U., Montgomery, B., Horne, P. & Ritchie, S. A. 2010 Quantifying the spatial dimension of dengue virus epidemic spread within a tropical urban environment. PLoS neglected tropical diseases 4, e920. (doi:10.1371/journal.pntd.0000920)

62. Villar, L. et al. 2015 Efficacy of a tetravalent dengue vaccine in children in Latin America. N. Engl. J. Med. 372, 113–123. (doi:10.1056/NEJMoa1411037)

63. Volz, E. M., Koelle, K. & Bedford, T. 2013 Viral phylodynamics. PLoS Comput Biol 9, e1002947. (doi:10.1371/journal.pcbi.1002947)

64. Volz, E. M. 2012 Complex Population Dynamics and the Coalescent Under Neutrality. Genetics 190, 187–201. (doi:10.1534/genetics.111.134627)

65. Volz, E. M. Ratmann, O., Severson, E. R. rcolgem R Package http://colgem.r-forge.r-project.org/ Accessed: 2016-06-15

59. Wakeley J. 2009 Coalescent Theory: An Introduction. Greenwood Village: Roberts & Company Publishers.

60. WHO 2016 Dengue vaccine: WHO position paper – July 2016. Wkly. Epidemiol. Rec. 91, 349–364.

61. Wikramaratna, P. S., Simmons, C. P., Gupta, S. & Recker, M. 2010 The effects of tertiary and quaternary infections on the epidemiology of dengue. Plos One 5, e12347. (doi:10.1371/journal.pone.0012347)