Epidemics 25 (2018) 80–88

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Kernel-density estimation and approximate Bayesian computation for flexible epidemiological model fitting in Python

Michael A. Irvine[a,*], T. Déirdre Hollingsworth[b]

[a] Institute of Applied Mathematics, University of British Columbia, Vancouver, Canada
[b] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK

ARTICLE INFO

Keywords:
Approximate Bayesian computation
Individual-based model
Lymphatic filariasis
Model fitting
Python library

ABSTRACT

Fitting complex models to epidemiological data is a challenging problem: methodologies can be inaccessible to all but specialists, there may be challenges in adequately describing uncertainty in model fitting, the complex models may take a long time to run, and it can be difficult to fully capture the heterogeneity in the data. We develop an adaptive approximate Bayesian computation scheme to fit a variety of epidemiologically relevant data with minimal hyper-parameter tuning by using an adaptive tolerance scheme. We implement a novel kernel density estimation scheme to capture both dispersed and multi-dimensional data, and directly compare this technique to standard Bayesian approaches. We then apply the procedure to a complex individual-based simulation of lymphatic filariasis, a human parasitic disease. The procedure and examples are released alongside this article as an open access library, with examples to aid researchers to rapidly fit models to data. This demonstrates that an adaptive ABC scheme with a general summary and distance metric is capable of performing model fitting for a variety of epidemiological data. It also does not require significant theoretical background to use and can be made accessible to the diverse epidemiological research community.

1. Introduction

There is a trend towards greater realism using individual-based models within the ecological and epidemiological modelling community (Grimm et al., 2006; Bansal et al., 2007; DeAngelis and Grimm, 2014; Heesterbeek et al., 2015). The strength of this approach lies in its ability to directly address policy-relevant questions, however properly estimating model parameters and measuring uncertainty in fits is often problematic/challenging (Deardon et al., 2010; Grimm and Railsback, 2013). In addition, the data will often be highly heterogeneous making model fitting difficult. Examples of this in epidemiology include both human and animal parasitic infections, such a soil-transmitted helminths and nematodes, where the variance in egg counts can be bigger than the mean (Shaw et al., 1998; Elkins et al., 1986; Grenfell et al., 1990). Data may also come in the form of multivariate time-series, such as number of diagnoses in different disease stages or different age-categories or age/risk/disease stage-stratified prevalence (Hollingsworth et al., 2008; Pullan et al., 2014). These data can be challenging to fit as it can be noisy and may not be easily modelled by simple distributions.

Complex individual-based models will often have computationally-intractable likelihoods, or likelihoods that are not easily defined or applied to data. In such cases, approximate Bayesian computation (ABC) has been proposed as a valid approach to model fitting (Csilléry et al., 2010). ABC has primarily been used to fit approximately Gaussian or Poisson-type data in the context of epidemiology (McKinley et al., 2009, 2014; Beaumont, 2010; Walker et al., 2010; Kypraios et al., 2016). Other data sources have been incorporated into model fitting using ABC, such a phylogenetic data (Tanaka et al., 2006; Luciani et al., 2009; Ratmann et al., 2012). It is often not clear what choice of summary statistic should be used and this is often domain specific, which can prevent these methods being applied elsewhere (Luciani et al., 2009; Marin et al., 2012).

Whilst these are general problems, they are of particular relevance in the calibration of complex individual-based models designed for policy-relevant questions. In this paper, we consider the case of lymphatic filariasis transmission. Lymphatic filariasis (LF) or elephantiasis is a neglected tropical disease, with over 40 million individuals displaying clinical manifestations of the disease, and with 53 countries requiring preventative chemotherapy. It is currently targeted for elimination as a public health problem by the World Health Organisation (WHO) by 2020 through the use of mass drug administration (MDA) (Rebollo and Bockarie., 2013; World Health Organization et al., 2011; Ottesen et al., 2008, 1997). As with many public health interventions, there is a certain amount of systematic non-adherence or heterogeneity in the use of interventions (Dyson et al., 2017). Coupled with this is the large amount of heterogeneity in

* Corresponding author.
E-mail addresses: m.irvine@math.ubc.ca (M.A. Irvine), Deirdre.Hollingsworth@bdi.ox.ac.uk (T.D. Hollingsworth).

https://doi.org/10.1016/j.epidem.2018.05.009
Received 21 February 2018; Received in revised form 5 May 2018; Accepted 24 May 2018
Available online 26 May 2018

exposure to infection across individuals. These complexities require that transmission models take into account the vector and parasite biology and human social factors (Irvine et al., 2015; Stolk et al., 2008; Chan et al., 1998). Due to the sparse nature of the data, parameter uncertainty in the fitted models must also be estimated if robust predictions are to be made (Singh and Michael, 2015). ABC then offers a strong alternative to other techniques for fitting complex individual-based models, which can also include uncertainty in the model parameters (Beaumont, 2010).

We developed a robust, adaptive ABC scheme for infectious disease epidemiological data. This approach incorporates a parameter-free method of estimating the distribution of the data and includes an adaptive scheme for selecting tolerance values. We have developed this scheme as an open-source python library with examples demonstrating its use. In the first section of this paper, we directly compare ABC to a more standard Bayesian fitting technique as an example of where the likelihood is known, by modelling counts drawn from a negative-binomial distribution. We vary the heterogeneity (shape parameter) in the distribution to investigate how the fitting performs for different degrees of heterogeneity. We compare how well fitting performs as the number of tolerance levels and number of particles (parameter sets) changes, showing how the automated tolerance selection procedure produces accurate model fits. In the next section we apply the technique to two simple individual-based models, which include overly-dispersed one-dimensional data and two-dimensional time-series data respectively. The results show that this technique is amenable to a wide range of models and data with little coding overhead or hyper-parameter tuning. Finally we demonstrate the technique on a complex individual-based model of LF and show how disparate forms of data can be included in the model fitting process, highlighting the ease of incorporating multiple data sources into the fitting (Smith et al., 2017).

2. Methods

2.1. Epidemiological count data

Count data such as number of diagnosed cases in one year or parasite/viral load per patient are abundant in epidemiology. Often these data will be treated as being drawn from a Poisson distribution (Wakefield, 2007; Pullan et al., 2012). This is where the data is drawn from a probability distribution of the form

$$P(X = x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}. \tag{1}$$

The Poisson distribution is special because the mean and the variance are equal. Whilst there is some theoretical justification for this, often sources of data can be more over-dispersed, where the variance of the distribution is greater than the mean. In this case the data can be described as a negative-binomial. The issue is then how to measure the amount of over-dispersion. Techniques will often focus on a particular distribution such as maximum likelihood or Bayesian Markov chain Monte Carlo (MCMC). These techniques have proved highly effective for models where the underlying rates (such as those produced from deterministic differential equation models) can be described. Individual-based and other stochastic complex models are not amenable to this technique, however, and so approximate fitting methods have been considered, such as ABC. It is not clear, however, how to incorporate an appropriate goodness of fit metric for over-dispersed data (for example comparing the means would not be able to capture the heterogeneity in the distribution). Here we propose the use of kernel density estimation in order to resolve this problem.

Kernel density estimation (KDE) is a non-parametric scheme for approximating a distribution using a series of kernels, or distributions (Bishop, 2006). The technique has previously been applied to approximating the likelihood of a summary statistic (Fearnhead and Prangle, 2012; Gutmann et al., 2016). However we use it here to directly compare between the modelled and real data. An important benefit of this approach is that, unlike with histograms, where placement of bins is important, kernels are centred on each data point and hence bins do not need to be selected. Often a Gaussian kernel is chosen to represent the data, this has the useful property of allowing the distribution to be defined everywhere in parameter space, thus making it possible to compare two empirical distributions. Without this property, the methodology would be unable to compare between two different empirical distributions if there was not significant overlap.

2.2. Overview of ABC methodology

ABC is a technique used to perform Bayesian inference when a likelihood is either computationally intractable or not feasible to define. As an alternative, a sufficient summary statistic is used for the model data and compared to the data to be fitted. A distance metric is used to define the error between the data drawn from the model and the real data. As the error between the summary statistics of the model-generated data and real data approaches zero, the posterior distribution is approximated with greater accuracy (Csilléry et al., 2010; Beaumont, 2010; Kypraios et al., 2016).

More precisely, the function f summarises the data D in some form, for example, the mean parasite load in certain age-groups. For particular model parameters, θ, the model produces output M_θ^*, where the star denotes this is a realisation of the model-data and is subsequently a random variable. We then define a distance metric, ρ, which compares the summary statistic from the data, $f(D)$ with that from the model, $f(M_\theta^*)$. The posterior is then approximated as the probability that the distance metric is below a threshold, ϵ, expressed as

$$P(\theta|D) \approx P(\rho(f(D), f(M_\theta^*)) < \epsilon). \tag{2}$$

The error in the approximation is assumed to decrease as the threshold, ϵ, decreases, with the method being exact when the threshold is zero (Rubin et al., 1984). This approximation is dependent on the choice of summary statistic f and distance metric ρ, which are often problem-specific. The approximation also requires an appropriate choice of ϵ to increase accuracy and decrease computation time. If ϵ is too large then the drawn samples are often a poor approximation of the posterior, and if ϵ is too low, then only very rarely would sampled M_θ^* meet the criterion leading to increased computation time.

One of the simplest conceptual algorithms for ABC is a partial rejection scheme where a particle (parameter set) θ is drawn from the prior distribution Θ. This particle is then used in the model M to produce some sample data M_θ^*. The sample data M_θ^* is then compared to the data D using the distance function ρ that gives a single-value for the discrepancy between the model data and the real data. This particle θ is then accepted if this discrepancy is below a pre-defined tolerance ϵ and rejected otherwise (Wilkinson, 2013) (e.g. for its first use see Pritchard et al., 1999, and see Blum and Tran (2010) for a smoothed rejection scheme applied to fitting an SIR model). In reality, this scheme can be inefficient if the prior is not similar to the posterior meaning that many particles are rejected. Also if the tolerance is too large then the sample of particles will be closer to the prior than the posterior. This means the scheme needs to be fine-tuned and may be impractical for most cases.

A way of overcoming the low particle acceptance rate issue is to start with a large tolerance ϵ and then to proceed as above until the desired number of particles are selected (Fig. 1). These particles can then be used to generate an empirical distribution that can then replace the prior in the algorithm. The tolerance can then be lowered and the rejection scheme can be repeated until the desired number of particles are sampled. This scheme provides a way of lowering the tolerance to increase the accuracy, whilst also overcoming the issue of a small acceptance rate (Walker et al., 2010). The distribution of tolerances will depend heavily on the number of particles used, here we explore how the number of particles affects the final distribution (see supplementary material).

The challenges with this scheme are to choose a set of tolerances, $\{\epsilon_t\}$, to efficiently reduce the error in the samples. Typically a set is chosen prior to fitting. We considered two schemes for tolerance selection. The first is to generate a set of tolerances by sampling the prior

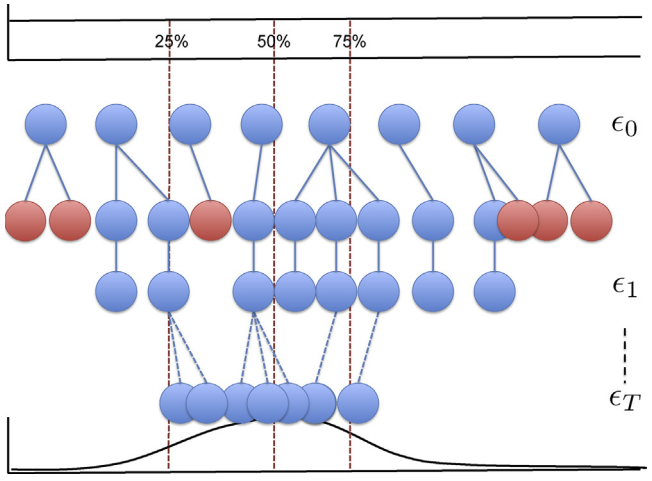

Fig. 1. An overview of the ABC partial rejection control technique fitting to one parameter and for three steps. The true underlying likelihood is shown at the bottom, with 25, 50 and 75 percentiles shown as red dotted lines. A number of particles (N) are fixed at the beginning, here $N = 8$. In the first step, these particles are drawn from a prior distribution, which is uniform between two values (top row, corresponds to steps 1–2 in Algorithm 1). For a given tolerance, a new particle is drawn for the updated tolerance ϵ_1 by choosing a particle at random, perturbing it slightly and then running a model evaluation (step 5–6 in Algorithm 1). It then checks if the tolerance of that particle is below ϵ_1, the particle is then either accepted (blue) or rejected (shown in red) (steps 7–8 in Algorithm 1). This procedure continues until all N particles are accepted at the new tolerance level (steps 9–10). As the tolerance decreases, the particles converge onto the target distribution (shown in the bottom row). (For interpretation of the references to color in text/this figure legend, the reader is referred to the web version of the article.)

distribution (Faisal et al., 2013). By drawing two sample particles $\theta_1, \theta_2 \sim \Theta$ and recording the error between them $\epsilon = \rho(f(M^*_{\theta_1}), f(M^*_{\theta_2}))$, a distribution of error values can be built up from the prior distribution. A range of tolerances then may be chosen by taking the 0th, 10th, 20th, …, 100th percentile values of the error value distribution.

An alternative way of selecting tolerances is to do it adaptively, based on the distribution of errors that were accepted in the previous iteration (Beaumont et al., 2002). This is accomplished by recording for particle i, the accepted error τ_i. The tolerance in the next iteration can then be chosen as some percentile of these values. Here, we adapted a scheme where the 50th percentile of these values was set as the new tolerance in order to keep the acceptance rate at reasonable levels. We found the adaptive scheme consistently outperformed the prior distribution scheme and as such we only consider the adaptive scheme here.

We considered data derived from both one-dimensional and two-dimensional distributions. The particular form of the summary statistic chosen for all examples was an empirical distribution derived from count data. Certain summary statistics and distance metrics such a the mean squared error between time-series data have underlying assumptions of normality and unimodality (Walker et al., 2010; Brown et al., 2018). We instead, adopt a scheme that is capable of incorporating a general distribution by using a non-parametric method to approximate the underlying probability density function f of the data. Note that here, as a simplification, discrete distributions are approximated by a continuous distribution. This was achieved using a Gaussian kernel-density estimator for the distribution. An empirical distribution $\hat{f}$ from count data $\{y_i\}$ was produced using a Gaussian kernel K by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=0}^{n-1} K(x - y_i). \tag{3}$$

Although each data-point is represented as a Gaussian with a small variance, the total distribution does not need to have the same properties and can for instance have higher variance or be multi-modal (Silverman, 1986). In order to compare between the two approximated distributions the non-symmetric Kullback–Leibler (KL) divergence was used. This measures the difference between the KDE-approximated probability distribution derived from the model data $\hat{p}$ and the KDE-approximated probability distribution derived from the real data $\hat{q}$. It is defined as

$$D_{\mathrm{KL}}(\hat{p}||\hat{q}) = \int_{-\infty}^{\infty} \hat{p}(x)\log\frac{\hat{p}(x)}{\hat{q}(x)}\mathrm{d}x, \tag{4}$$

where the divergence is greater than zero if the probability distributions differ and is zero if the distributions are equivalent. This method can also be easily adapted to a multivariate distribution, where an n-dimensional symmetric Gaussian with a fixed variance in each dimension can be used in the KDE step. The calculation of the KL divergence can also be extended by integrating over the entire support of the probability density function derived in the KDE step.

Our combined adaptive scheduling partial rejection control with kernel density estimation algorithm is as follows (Fig. 1). A number of particles (parameter sets) are drawn from the prior distribution $P(\theta)$ to produce a set of particles $\{\theta_i^1\}$. An initial tolerance value ϵ_1 is found by selecting the median value of the KDE KL divergence between the data and model derived data from the selected particles. A new set of particles is generated by randomly sampling from $\{\theta_i^1\}$ and perturbed using a zero-mean Gaussian random variable with small variance. The newly generated particle is accepted if $D_{\mathrm{KL}}\left(f(D)||f\left(M^*_{\theta_i^1}\right)\right) < \epsilon_1$, else it is rejected and another particle is generated according the procedure defined. Once the desired number of particles have been accepted the tolerance is lowered adaptively by selecting the median value of the accepted tolerances from the previous iteration. A new set of particles is then generated as before with the lowered tolerance ϵ. Once the particles are generated for the smallest tolerance, ϵ_T, the algorithm terminates and these are used as the sample for the posterior. A summary is given in Algorithm 1. We also show for a Gaussian likelihood that the minimization of the KL divergence with a KDE representation of the data is equivalent to maximising the likelihood (see supplementary material).

Algorithm 1. Adaptive ABC partial rejection control.

1: Set $t = 0$ } Initialise particles
2: draw N particles from prior distribution: $\theta_i^1 \sim \Theta$
3: Calculate the distribution of the data $f(D)$ using KDE.
4: Set $t = 1$
5: **while** $t < T$ **do**
6: draw particle from set $\{\theta_i^t\}$ with perturbation δ: $\phi = \theta_i^t + \delta$ } Generate new particle
7: Simulate using current parameters: $M^*_\phi \sim M(\phi)$
8: Calculate $f(M^*_\phi)$ using KDE.
9: **if** sufficiently similar to data $D_{KL}(f(M^*_\phi)||f(D)) < \epsilon_t$ **then** } Acceptance/rejection step
10: Add ϕ to list of new particles $\{\theta_i^{t+1}\}$.
11: **if** Number of new particles N **then**
12: Set ϵ_{t+1} to the median value of the particle errors at step t.
13: Increment t by one

2.3. Example applications of the method

2.3.1. Example one: negative binomial distribution

As a first example in order to compare how fitting using our ABC scheme compares to other fitting techniques, samples were drawn from a negative binomial distribution with varying mean and heterogeneity parameter k. When $k < 1$, the distribution is over-dispersed, with a greater variance to mean ratio than expected under a Poisson distribution. This means that the distribution is more heavy-tailed than for an equivalent Poisson distribution. When $k > 5$, the distribution is less over-dispersed and small samples more closely resemble a Poisson distribution. In order to test how the parameter fitting performs for increasing heterogeneity (decreasing k), a sample is drawn from a negative-binomial parameterised by the mean m and heterogeneity k,

$$P(X = x|m, k) = \frac{\Gamma(k+x)}{x!\Gamma(k)}\left(\frac{k}{k+m}\right)^k\left(\frac{m}{k+m}\right)^x. \tag{5}$$

The likelihood for an independent and identically distributed sample $X = (x_0, x_1, ...,x_{n-1})$ is then

$$P(X|m, k) = \prod_{i=0}^{n-1}\frac{\Gamma(k+x_i)}{x_i!\Gamma(k)}\left(\frac{k}{k+m}\right)^k\left(\frac{m}{k+m}\right)^{x_i}. \tag{6}$$

m was varied between 1 and 100 and k was varied between 0.1 and 5. In order to be consistent between the samples the prior used in ABC was fixed for all samples before observing the data. Exponential priors were used with means of the distributions chosen to be the average of the ranges explored for m and k.

A Metropolis–Hastings MCMC scheme was also implemented and fitted to the negative-binomial count data (Gilks et al., 1995). The same priors that were used for the ABC scheme were also used for the MCMC scheme to provide a faithful comparison.

The impact of number of particles and size of tolerance were also explored using this model. For fixed parameters ($m = 50$, $k = 3.0$) the derived distribution was estimated for tolerances from 1 to 25 and particle numbers 10 to 200. The resulting estimated posterior was then compared to the true posterior (derived from the MCMC scheme).

The previous example can be easily implemented in the developed Python library with code that sets up a function that outputs an array of samples drawn from a negative binomial distribution for inputs m and k (denoted ibm), defines the priors as a list of functions that generate a sample for each parameter (denoted priors), provides the fitting object with the individual-based model, the data (denoted xs), the priors, and sets the method and number of steps to iterate through (denoted by the method setup) (Listing 1). The method is then run with a specified number of particles (denoted by the method run).

Listing 1. Code for negative binomial distribution example.

```
import scipy.stats as stats
import abcprc as prc
#load data to be fitted
xs = load_data()
#define negative binomial distribution function
def ibm(m,k):
   # returns negative-binomial sample array for mean m and shape k

#load ABC fitting class
m = prc.ABC()
#define exponential priors for each parameter
priors = [stats.expon(scale=100.0).rvs,stats.expon(scale=1.0).rvs]
#set-up fitting process using the adaptive scheme with 25 tolerance levels
m.setup(modelFunc=ibm,xs=xs,priors=priors,method='Adaptive',toln=25)
#fit using 100 particles
m.run(100)
```

2.3.2. Example two: parasite model

As a simple epidemiological example, we propose an individual-based model where individuals acquire parasites at a constant rate that is drawn from a gamma-distribution with mean λ and shape parameter k. Each parasite within individuals are lost at a constant rate δ. When k is low the distribution of parasites is more heterogeneous with many individuals uninfected, but with a few highly infected, with very large parasite numbers. Schematically, the parasite dynamics within an individual P_i can be written for each individual i;

$$P_i \to P_i + 1 \text{ at rate } \lambda b_i, \tag{7}$$

$$P_i \to P_i - 1 \text{ at rate } \delta P_i, \tag{8}$$

where b_i is a random variable drawn from a gamma distribution with shape k and mean 1. This model could easily be extended by making the force of infection dependent on the current distribution of parasites as well as other factors such as environmental heterogeneity. It is however, meant to be instructive and as such the simplest form was used.

2.3.3. Example three: stochastic SIS model

The stochastic Susceptible-Infected-Susceptible (SIS) model was implemented as an example of time-series data that can be estimated using a two-dimensional distribution approach. The model can be described as a Markov Process with two events: an infection and a recovery. For a population of size n, with the number of infected I, the infection and recovery events occur according to

$$I \to I + 1 \text{ at rate } \beta(n - I)I, \tag{9}$$

$$I \to I - 1 \text{ at rate } \gamma I. \tag{10}$$

The parameters β and γ can be reparameterised using the basic reproduction number R_0 and the expected time to recovery γ^{-1} as $\beta = R_0/\gamma^{-1}$ and $\gamma = 1/\gamma^{-1}$. The model was simulated in discrete time-steps using a tau-leaping algorithm and the corresponding likelihood was calculated using the corresponding transition rate matrix for the Markov Process (see supplementary material). In order to utilize this data with the KDE approach described we may convert the one-dimensional time-series data into two-dimensional distribution data in the following way, where we can explicitly take advantage of the Markov property of the underlying model. For a time-series of number of infected individuals recorded at regular intervals $\mathbf{I} = (I_0, I_1, ...,I_T)$ the number of infected conditioned on the previous time-step $I_{t+1}|I_t$ can be represented as the matrix

$$I_{t+1}|I_t = \begin{pmatrix} I_0 & I_1 & ... & I_{T-1} \\ I_1 & I_2 & ... & I_T \end{pmatrix}^\top. \tag{11}$$

Each row in the matrix is a (I_{t+1}, I_t) pair which are points in 2D and can therefore be used to build up a two-dimensional probability density function (an example of this is shown in Fig. 3b).

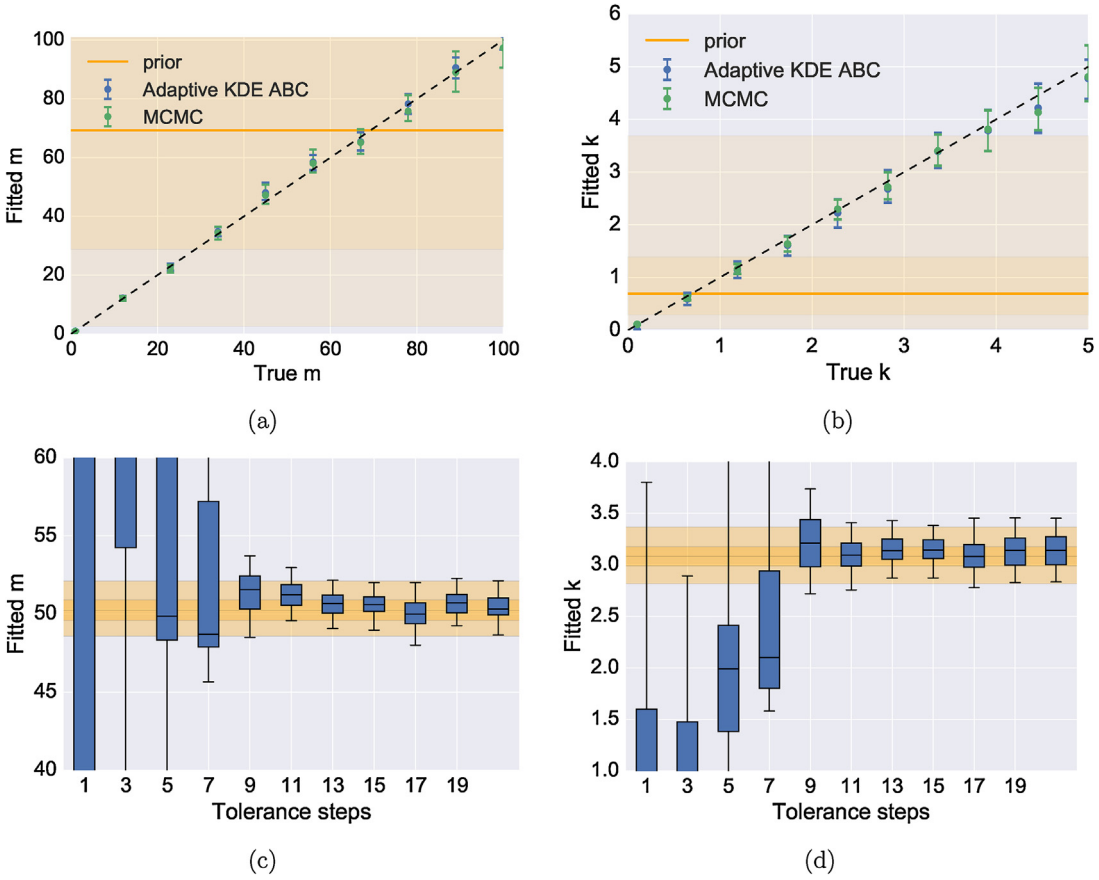

Fig. 2. Comparison between MCMC and ABC methods for fitting a negative binomial distribution for a range of mean, m and heterogeneity k. (a) Comparison between fits for different mean values m, the dashed line represents the true values and the shading represents the 95% and 50% percentile range of the prior distribution, with the median given as a solid line. The prior distribution was kept fixed for each fitting. The adaptive KDE scheme closely matches the MCMC scheme for all values considered. When the resulting fit is biased for ABC, it is also biased in the same way for MCMC providing confidence that the scheme is approximating the true posterior. (b) Comparison between MCMC and the adaptive ABC scheme for heterogeneity k. As $k > 3$ both the MCMC scheme and adaptive ABC scheme underestimate the true value in a consistent way due to the influence of the prior. Comparison between fitted distributions of the adaptive ABC scheme against the number of adaptive tolerance steps are shown for (c) m and (d) k. The true posterior calculated using MCMC is represented as a series of shaded regions with the 95% credible interval, 50% credible interval, and the median shown from lightest to darkest respectively. (For interpretation of the references to color in text/this figure legend, the reader is referred to the web version of the article.)

With the given data representation the methodology is implemented in the Python package in exactly the same way as for the one-dimensional negative-binomial example. The output of the model function (ibm) is used to determine the dimension of the data and the list of prior random variable generators are used to determine the size of the parameter space in the model code.

2.3.4. Example four: lymphatic filariasis

We used a stochastic individual-based model of lymphatic filariasis (Irvine et al., 2015). The model is a multi-scale stochastic simulation of individuals with worm burden, microfilaraemia (prevalence of the pre-larval stage of LF in the peripheral blood) and other demographic parameters relating to age and risk of exposure. Humans are modelled individually, with their own male and female worm burden denoted W_i^m and W_i^f. The density of microfilariae (mf) in the peripheral blood is also modelled for each individual and denoted M_i. The total mf density in the population contributes towards the current density of L3 larvae in the human-biting mosquito population. The model dynamics are divided into the individual human dynamics, including age and turnover; worm dynamics inside the host; microfilariae dynamics inside the host and larvae dynamics inside the mosquito.

Five villages in the East Sepik Province of Papua New Guinea have been the focus of extensive research into filariasis epidemiology and transmission (Bockarie et al., 2003, 1998; Michael and Singh, 2016; Irvine et al., 2018). These villages received annual mass drug administration from 1993 through 1998, with no further interventions until bed-nets (LLIN) were distributed in August 2009. Self reported LLIN use ranged from 75% to 90% (Reimer et al., 2013).

Microfilaria prevalence were measured in these communities in 2008 as part of the post-MDA evaluation (Reimer et al., 2013). This was done by a BinaxNow filariasis antigen test and by microscopic evaluation of 1 mL filtered venous blood, collected at night. The age of participants was also recorded.

The KDE ABC methodology was implemented on three geographically variable parameters, the vector to host ratio V/H, the heterogeneity of bites k and the probability of an infective bite leading to an establishment of an adult worm s_2. Each of these parameters were fit separately to the mf count distribution for each village. This was then compared to when the age-prevalence data was also included in the model fitting. Age-prevalence data was incorporated through the use of a mean squared distance function in addition to the KDE KL divergence function for the mf count data.

2.4. Implementation

The methodology and models were implemented in Python 2.7 (Python Software and Foundation, 2018), using the packages SciPy & NumPy (Van Der Walt et al., 2011) and seaborn for data visualisation

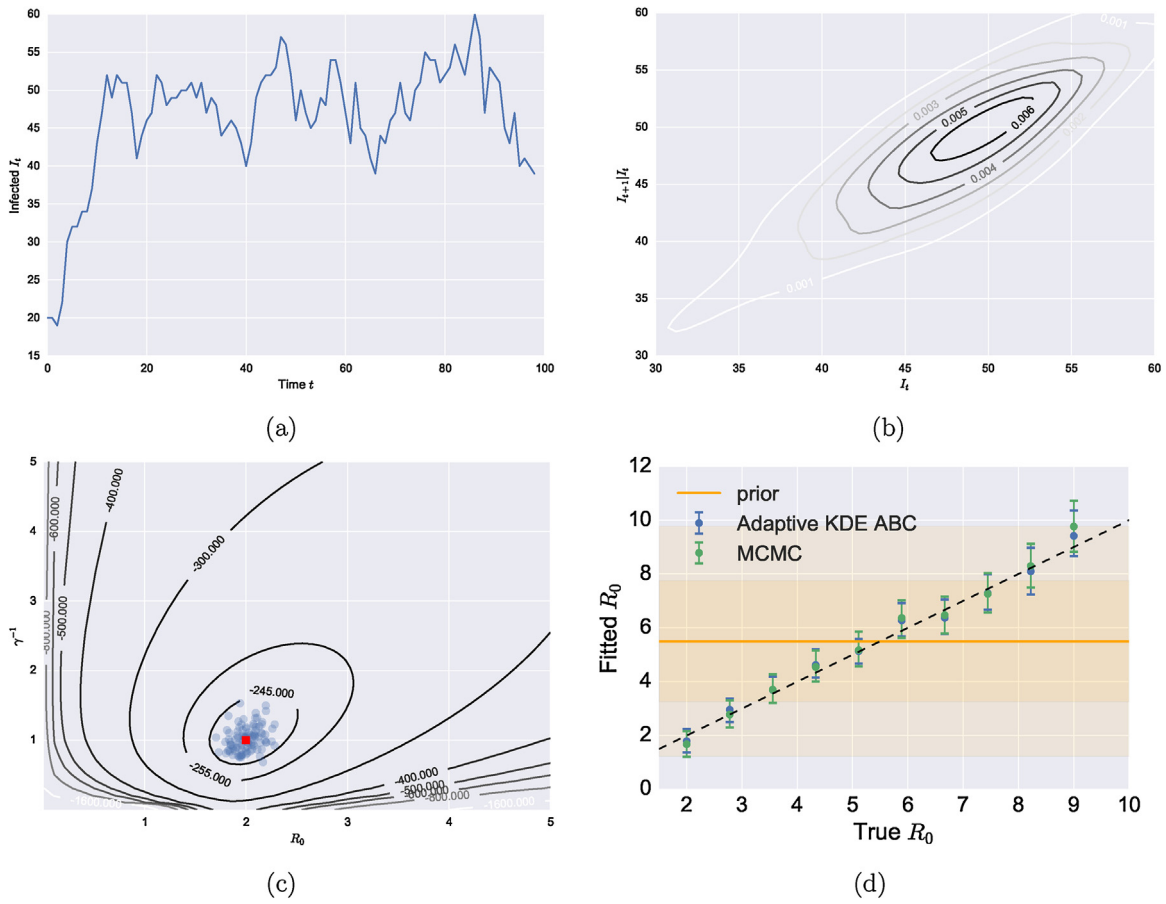

Fig. 3. The Adaptive KDE ABC scheme for stochastic time-series data. (a) A realisation of the SIS process with $R_0 = 2$ population size n = 100 and recovery time $\gamma^{-1} = 1$ (b) The corresponding time-series data converted into an empirical joint distribution using a two-dimensional KDE approach (level sets represent probability densities). (c) The estimated posterior distribution after 20 steps for 100 particles (in blue), with the true value in red. The true posterior is shown as level sets, with un-normalised log-values displayed. (d) Comparison of fitted distributions to R_0 between the adaptive KDE approach against the MCMC method for a range of R_0 values. The prior is represented as a series of shaded regions with the 95%, 50% and the median shown from lightest to darkest respectively. (For interpretation of the references to color in text/this figure legend, the reader is referred to the web version of the article.)

(Waskom et al., 2014). An open-source python library, including examples can be found at the following URL: https://github.com/sempwn/ABCPRC. This library has been tested for both Python version 2.7 and version 3.6.

3. Results

3.1. Drawing from a negative binomial distribution

MCMC was directly compared to the adaptive ABC method using samples drawn from a negative binomial distribution with a range of means m and heterogeneities k (Fig. 2). The ABC scheme was ran on 100 particles over 25 tolerance steps, while the MCMC scheme was ran for 10,000 steps with a burn-in period of 2000 steps and a fixed step-size. Visual inspection was used to determine the convergence of the MCMC chains. Exponential priors with rate 50 and 1 were used for m and k respectively. For small k, samples from the distribution are more over-dispersed and larger k values more closely approximate the Poisson distribution. For all mean m values considered both the MCMC method and ABC method closely match the true value (Fig. 2a). As the size of m grows so does the size of the 95% credible interval in both cases. Where the model fit is biased due to the data realisation producing more than expected lower probability samples (e.g. mean value 50), both MCMC and ABC are biased in a consistent way. This provides more confidence that the scheme is recovering the true posterior distribution. Further evidence of this can be seen in the fitting as heterogeneity k varies (Fig. 2b). Here the prior is stronger, with a smaller 95% interval relative to the parameter range considered. For small values of k, the estimated posterior distributions closely match the true values. As k increases above 3 the true value moves outside of the prior's 95% range and thus begins to have more influence on the posterior. This can be seen as the expected value of k estimated from both methods is consistently lower than the true value.

The number of adaptive tolerance steps strongly influences the estimated posterior for both the mean m (Fig. 2c) and the heterogeneity k (Fig. 2d). For a small number of steps (1-5), the estimate more closely resembles the prior distribution than the posterior distribution. From 7 to 9 steps the estimate is a combination of the prior and posterior distribution. For values above 10, the distributions closely match the true posterior. It should be noted that these values would likely change depending on the model and data, although we have found that 20 or more tolerance steps is sufficient for the estimate to converge to the posterior for the examples considered here.

The number of particles was also considered in how this hyper-parameter impacts the estimated posterior. Neither the expected value or the range were consistently effected by the particle size and even a small number of particles could approximate the true posterior reasonably well (see supplementary material). This suggests that if model evaluations are costly, then a small number of particles can be used to approximately determine the posterior before running the method on a large particle size.

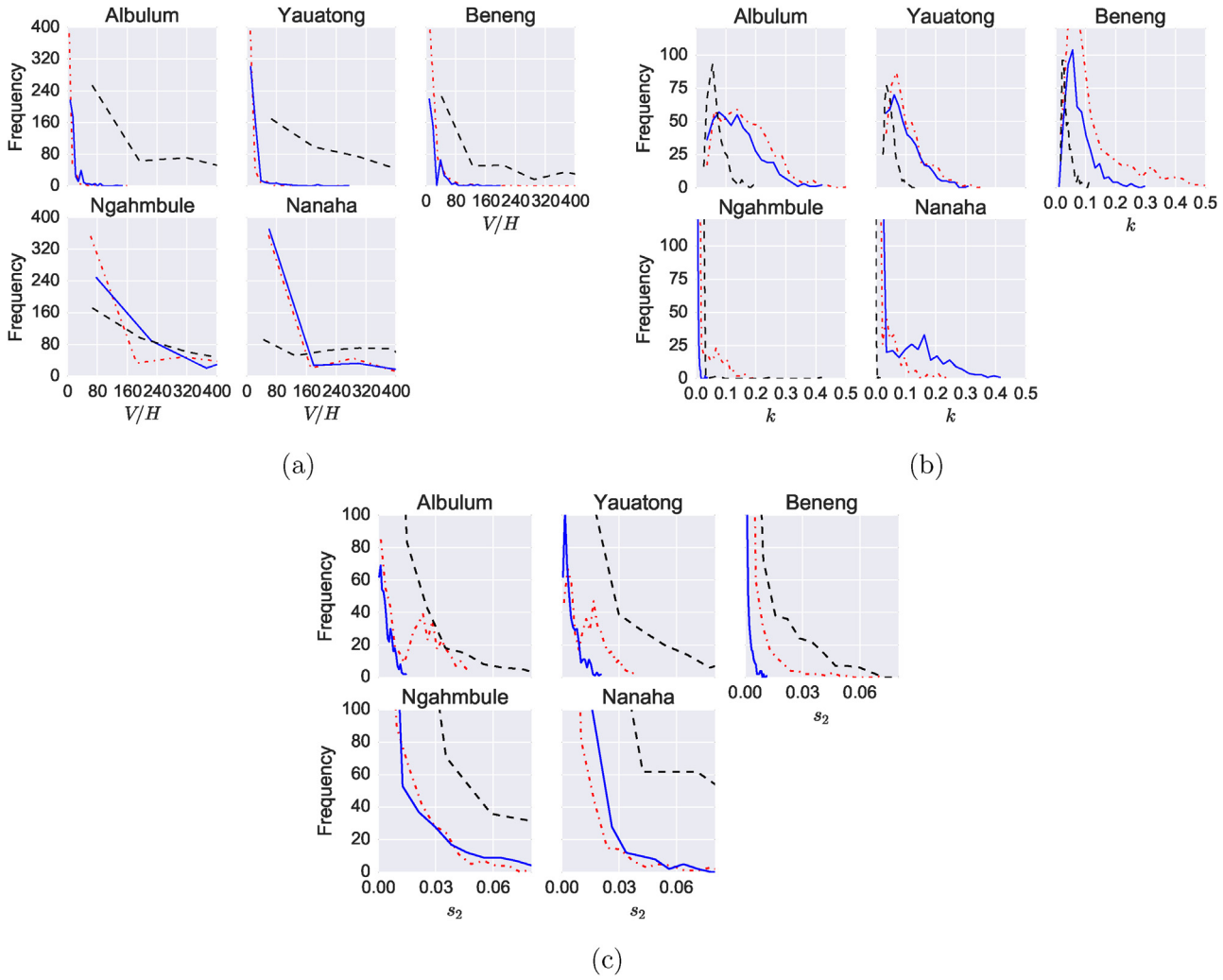

Fig. 4. Results from fitting an individual-based lymphatic filariasis infection model to PNG. The estimated marginal distributions for parameters (a) vector to host ratio, V/H; (b) heterogeneity of exposure, k and (c) probability of larvae developing to reproductive adult, s_2. Fitting just using count data in red; fitting using age prevalence data alone in black; and fitting using both count data and prevalence data in blue. (For interpretation of the references to color in text/this figure legend, the reader is referred to the web version of the article.)

3.2. Host–parasite model

The method was applied to the simple individual-based model of parasitic infection. A data sample was produced from the model (parameters: $\lambda = 10$, $\delta = 0.5$, $\gamma = 1.0$) and used in the fitting procedure. All parameters were given exponential prior distributions with mean rates broad enough to capture most dynamics. As the tolerance reduced, the variance in each of the marginal distributions lowered. The final distribution was unimodal for each parameter, with modal values close to the true underlying values. The final distribution also captures correlations between certain parameters such as between mortality and infection rate (see Fig. 2, supplementary material).

3.3. SIS model

A realisation of the SIS model was taken with parameters $R_0 = 2$, population size n = 100, and recovery time $\gamma^{-1} = 1$, for 100 time-steps (Fig. 3a). The corresponding joint distribution of the I_{t+1}, I_t data was approximated using a two-dimensional KDE (Fig. 3b). Here the joint distribution was approximately a bivariate correlated Gaussian, where the number of infected at time $t + 1$ was strongly dependent on the number of infected at time t. The empirical distribution also has a longer tail than expected for a Gaussian distribution due to the initial transient phase where the infected population is rapidly increasing from the initial conditions. The adaptive KDE ABC method was able to accurately determine the correct R_0 and γ^{-1} values and was consistent with the true posterior (Fig. 3c). For other R_0 values the adaptive ABC method was able to accurately approximate the true posterior and recover the true value (Fig. 3d).

3.4. LF in Papua New Guinea

Fitting was performed on five separate datasets of lymphatic filariasis infection including individuals' age and mf count. The summary statistic used was derived from mf count alone; mf age-prevalence and a combination of the two. Three parameters were fitted, where other parameters in the model were derived from literature estimates. Age-prevalence alone was unable to accurately determine the vector to host ratio and the probability s_2, with wide variances for the estimates of both (Fig. 4a and c). Using the mf count data only produced a smaller estimated range for these parameters, whilst giving a slightly wider range for the heterogeneity k (Fig. 4b). Combining both the mf count summary statistic and mf age-prevalence produces a more highly resolved marginal posterior for all three fitted parameters.

4. Discussion

Individual based models abound in epidemiology due to their intuitive description and greater ease of simulating many complex aspects of a system compared to deterministic models (Auchincloss and Diez Roux, 2008). These models increasingly involve processes that may not be easily captured by an ordinary differential equation or standard stochastic processes. This presents a great challenge, however, as standard fitting techniques have been developed for more traditional models, whereas ones for individual-based models have languished (Heesterbeek et al., 2015). Although for certain models it may be technically possible to write down a likelihood, there can be huge computational or technical barriers to do this. Whether this is due to a large number of hidden states or the sheer number of components in the model, this leads to having to resort to techniques such as visual inspection to perform fitting, introducing potential biases and not having a structured way to deal with the uncertainty in the fitted parameters. What is desirable is to have a technique where we can enjoy the benefits of Bayesian fitting, such as incorporating our prior knowledge and producing samples to estimate parameter uncertainty, without the often prohibitive procedure of conceiving of and calculating a likelihood.

Here we explored an ABC method as a solution for Bayesian model fitting. In particular we developed a technique that was amenable to a variety of data with minimal hyper-parameter tuning. The motivation is to provide a tool for model fitting with uncertainty quantification to a wide range of researchers, who may not have the necessary technical background to develop a full Bayesian approach with a developed likelihood. We performed model fitting using a summary statistic of the counts by approximating the distribution using a kernel density estimator. This allows fitting to be performed without explicit assumptions on the particular type of distribution the data takes as can be common with other model fitting techniques.

In order to compare the accuracy of ABC for increasingly heterogeneous count data, the procedure was carried out on various data generated from a negative-binomial distribution. For high heterogeneity, the procedure was able to accurately determine the shape parameter (k), as well as the mean parameter (m). This demonstrates that this technique is capable of handling a variety of heterogeneous data and can give similar results to standard Bayesian MCMC. the technique was also able to perform well on time-series data by transforming the data into a two-dimensional point representation. This technique would appear generally applicable to other time-series data including systems that may exhibit chaos (see supplementary material).

For many individual-based models a likelihood may be either computationally or analytically intractable. In these cases other methods have been proposed to overcome this issue. Using a partial rejection control scheme provides, at each iteration, a sample of particles (parameter sets) that are initially drawn from the prior, but as the tolerance decreases, these samples become more representative of the posterior. Although there are typically issues surrounding the choice of tolerances, such that the scheme is able to draw samples for the next iteration. Here, we overcome these issues by demonstrating two different schemes for choice a set of tolerances. This creates a much more efficient pipeline for fitting without the need to perform exploratory analysis of the error function beforehand (Walker et al., 2010).

One of the key issues with ABC is that it is an approximation method only. If the method does not sufficiently explore the space of parameters, the technique may produce spurious results. One possible diagnosis is to check the distribution of errors that were accept for each tolerance. If the errors are not significantly decreasing then this may indicate the procedure is stuck in a local minima and the variance of the priors may need increasing. The distribution of errors for the final tolerance can also indicate whether the procedure was halted prematurely or if lower tolerances can be accepted.

There is also an issue with the choice of summary statistics to be used and the number of parameters to fit to. It may be that some parameters can be estimated from independent studies, without the need to include them in the ABC procedure. It would then seem advisable to use these values either as a well-informed prior or as a point estimate as was done here. If the model is slow to evaluate, then this may also lead to practical fitting issues. Emulation methods may help to further increase the speed of fitting, by approximating the error manifold through the use of non-parametric fitting techniques such as Gaussian processes (Conti and O'Hagan, 2010; Drovandi et al., 2011).

One primary advantage of ABC over other techniques is the ability to utilize a range of data within model fitting. In the example of fitting an individual-based model of lymphatic filariasis infection to PNG data a combination of summary statistics was used. We explored fitting using just count data alone, constructing an empirical probability distribution and then comparing against the model count data using the KL divergence. This summary statistic was then combined with age-prevalence data, by constructing the prevalence in a defined set of age-categories and then using a weighted sum of squares in order to take into account the number of individuals in each age-category. We found that by adding in the extra information about the age-prevalence distribution the fitting was able to better resolve some of the parameters. ABC provides a way of incorporating many different types of data into the fitting and this suggests that the full number of pertinent summary statistics should be used.

5. Conclusion

The adaptive ABC method incorporating kernel density estimation and partial rejection control is a potentially powerful tool in model fitting for epidemiological data. We demonstrate that the same methodology can fit to both macro and micro-parasitic infectious diseases, one-dimensional or two-dimensional data, and can readily incorporate a wide array of data sources. In order for this tool to be readily-available to a wide-range of researchers we have developed this as an open-source python library, including example code to demonstrate its use.

Data availability

All code is packaged as a python library and can be found at the following GitHub repository: https://github.com/sempwn/ABCPRC. This includes all code for generating data used in example model fitting.

Competing interests

The authors declare no competing interests.

Author contributions

The study was conceived by MAI & TDH. The coding, implementation, analysis and drafting of the manuscript was performed by MAI. All authors reviewed and approved the final manuscript. LF data are from the Papua New Guinea 2008 study (Reimer et al., 2013). The authors of this study may be contacted at jxk14@case.edu.

Funding

The authors gratefully acknowledge funding of the NTD Modelling Consortium by the Bill and Melinda Gates Foundation in partnership with the Task Force for Global Health. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views, opinions, assumptions or any other information set out in this article should not be attributed to Bill & Melinda Gates Foundation and The Task Force for Global Health or any person connected with them.

Acknowledgements

The authors wish to thank Lisa J. Reimer for use of the lymphatic filariasis data and useful comments during manuscript preparation.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.epidem.2018.05.009.

References

Auchincloss, A.H., Diez Roux, A.V., 2008. A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health. Am. J. Epidemiol. 168 (1), 1–8.
Bansal, S., Grenfell, B.T., Meyers, L.A., 2007. When individual behaviour matters: homogeneous and network models in epidemiology. J. R. Soc. Interface 4 (16), 879–891.
Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. Genetics 162 (4), 2025–2035.
Beaumont, M.A., 2010. Approximate Bayesian computation in evolution and ecology. Ann. Rev. Ecol. Evol. Syst. 41, 379–406.
Bishop, C.M., 2006. Pattern Recognition and Machine Learning. Springer.
Blum, M.G.B., Tran, V.C., 2010. HIV with contact tracing: a case study in approximate Bayesian computation. Biostatistics 11 (4), 644–660.
Bockarie, M.J., Alexander, N.D.E., Hyun, P., Dimber, Z., Bockarie, F., Ibam, E., Alpers, M.P., Kazura, J.W., 1998. Randomised community-based trial of annual single-dose diethylcarbamazine with or without Ivermectin against Wuchereria bancrofti infection in human beings and mosquitoes. Lancet 351 (9097), 162–168.
Bockarie, M.J., Tisch, D.J., Kastens, W., Partono, F., Oemijati, S., Soewarta, A., Meyrowitsch, D.W., Simonsen, P.E., Makunde, W.H., Ciferri, F., et al., 2003. Mass treatment of filariasis in New Guinea. N. Engl. J. Med. 2003 (348), 1179–1181.
Brown, G.D., Porter, A.T., Oleson, J.J., Hinman, J.A., 2018. Approximate Bayesian computation for spatial SEIR (S) epidemic models. Spat. Spatio-temp. Epidemiol. 24, 27–37.
Chan, M.-S., Srividya, A., Norman, R.A., Pani, S.P., Ramaiah, K.D., Vanamail, P., Michael, E., Das, P.K., Bundy, D.A., 1998. Epifil: a dynamic model of infection and disease in lymphatic filariasis. Am. J. Trop. Med. Hygiene 59 (4), 606–614.
Conti, S., O'Hagan, A., 2010. Bayesian emulation of complex multi-output and dynamic computer models. J. Stat. Plan. Inference 140 (3), 640–651.
Csilléry, K., Blum, M.G.B., Gaggiotti, O.E., François, O., 2010. Approximate Bayesian computation (ABC) in practice. Trends Ecol. Evol. 25 (7), 410–418.
DeAngelis, D.L., Grimm, V., 2014. Individual-based models in ecology after four decades. F1000prime Rep. 6.
Deardon, R., Brooks, S.P., Grenfell, B.T., Keeling, M.J., Tildesley, M.J., Savill, N.J., Shaw, D.J., Woolhouse, M.E.J., 2010. Inference for individual-level models of infectious diseases in large populations. Stat. Sin. 20 (1), 239.
Drovandi, C.C., Pettitt, A.N., Faddy, M.J., 2011. Approximate Bayesian computation using indirect inference. J. R. Stat. Soc.: Ser. C (Appl. Stat.) 60 (3), 317–337.
Dyson, L., Stolk, W.A., Farrell, S.H., Hollingsworth, T.D., 2017. Measuring and modelling the effects of systematic non-adherence to mass drug administration. Epidemics 18, 56–66.
Elkins, D.B., Haswell-Elkins, M., Anderson, R.M., 1986. The epidemiology and control of intestinal helminths in the Pulicat Lake region of Southern India. I. Study design and pre-and post-treatment observations on ascaris lumbricoides infection. Trans. R. Soc. Trop. Med. Hygiene 80 (5), 774–792.
Faisal, M., Futschik, A., Hussain, I., 2013. A new approach to choose acceptance cutoff for approximate Bayesian computation. J. Appl. Stat. 40 (4), 862–869.
Fearnhead, P., Prangle, D., 2012. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) 74 (3), 419–474.
Gilks, W.R., Richardson, S., Spiegelhalter, D., 1995. Markov chain Monte Carlo in practice. CRC Press.
Grenfell, B.T., Das, P.K., Rajagopalan, P.K., Bundy, D.A.P., 1990. Frequency distribution of lymphatic filariasis microfilariae in human populations: population processes and statistical estimation. Parasitology 101 (03), 417–427.
Grimm, V., Railsback, S.F., 2013. Individual-based Modeling and Ecology. Princeton University Press.
Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S.K., Huse, G., et al., 2006. A standard protocol for describing individual-based and agent-based models. Ecol. Model. 198 (1), 115–126.
Gutmann, M.U., Corander, J., et al., 2016. Bayesian optimization for likelihood-free inference of simulator-based statistical models. J. Mach. Learn. Res.
Heesterbeek, H., Anderson, R.M., Andreasen, V., Bansal, S., De Angelis, D., Dye, C., Eames, K.T.D., John Edmunds, W., Frost, S.D.W., Funk, S., et al., 2015. Modeling infectious disease dynamics in the complex landscape of global health. Science 347 (6227), aaa4339.
Hollingsworth, T.D., Anderson, R.M., Fraser, C., 2008. HIV-1 transmission, by stage of infection. J. Infect. Dis. 198 (5), 687–693.
Irvine, M.A., Reimer, L.J., Njenga, S.M., Gunawardena, S., Kelly-Hope, L., Bockarie, M., Hollingsworth, T.D., 2015. Modelling strategies to break transmission of lymphatic filariasis-aggregation, adherence and vector competence greatly alter elimination. Parasit. Vect. 8 (1), 1.
Irvine, M.A., Kazura, J.W., Hollingsworth, T.D., Reimer, L.J., 2018. Understanding heterogeneities in mosquito-bite exposure and infection distributions for the elimination of lymphatic filariasis. In: Proc. R. Soc. B, vol. 285. The Royal Society. pp. 20172253.
Kypraios, T., Neal, P., Prangle, D., 2016. A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. Math. Biosci.
Luciani, F., Sisson, S.A., Jiang, H., Francis, A.R., Tanaka, M.M., 2009. The epidemiological fitness cost of drug resistance in mycobacterium tuberculosis. Proc. Natl. Acad. Sci. U.S.A. 106 (34), 14711–14715.
Marin, J.-M., Pudlo, P., Robert, C.P., Ryder, R.J., 2012. Approximate Bayesian computational methods. Stat. Comput. 22 (6), 1167–1180.
McKinley, T., Cook, A.R., Deardon, R., 2009. Inference in epidemic models without likelihoods. Int J. Biostat. 5 (1).
McKinley, T.J., Ross, J.V., Deardon, R., Cook, A.R., 2014. Simulation-based Bayesian inference for epidemic models. Comput. Stat. Data Anal. 71, 434–447.
Michael, E., Singh, B.K., 2016. Heterogeneous dynamics, robustness/fragility trade-offs, and the eradication of the macroparasitic disease, lymphatic filariasis. BMC Med. 14 (1), 1.
Ottesen, E.A., Duke, B.O., Karam, M., Behbehani, K., 1997. Strategies and tools for the control/elimination of lymphatic filariasis. Bull. World Health Organ. 75 (6), 491.
Ottesen, E.A., Hooper, P.J., Bradley, M., Biswas, G., 2008. The global programme to eliminate lymphatic filariasis: health impact after 8 years. PLoS Negl. Trop. Dis. 2 (10), e317.
Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W., 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molec. Biol. Evol. 16 (12), 1791–1798.
Pullan, R.L., Sturrock, H.J.W., Soares Magalhaes, R.J., Clements, A.C.A., Brooker, S.J., 2012. Spatial parasite ecology and epidemiology: a review of methods and applications. Parasitology 139 (14), 1870–1887.
Pullan, R.L., Smith, J.L., Jasrasaria, R., Brooker, S.J., 2014. Global numbers of infection and disease burden of soil transmitted helminth infections in 2010. Parasit. Vect. 7, 37.
Python Software Foundation. Python language reference, version 2.7.
Ratmann, O., Donker, G., Meijer, A., Fraser, C., Koelle, K., 2012. Phylodynamic inference and model assessment with approximate Bayesian computation: influenza as a case study. PLoS Comput. Biol. 8 (12), e1002835.
Rebollo, M.P., Bockarie, M.J., 2013. Toward the elimination of lymphatic filariasis by 2020: treatment update and impact assessment for the endgame. Expert Rev. Anti-infect. Ther. 11 (7), 723–731.
Reimer, L.J., Thomsen, E.K., Tisch, D.J., Henry-Halldin, C.N., Zimmerman, P.A., Baea, M.E., Dagoro, H., Susapu, M., Hetzel, M.W., Bockarie, M.J., et al., 2013. Insecticidal bed nets and filariasis transmission in Papua New Guinea. N. Engl. J. Med. 369 (8), 745–753.
Rubin, D.B., et al., 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann. Stat. 12 (4), 1151–1172.
Shaw, D.J., Grenfell, B.T., Dobson, A.P., 1998. Patterns of macroparasite aggregation in wildlife host populations. Parasitology 117 (06), 597–610.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Routledge.
Singh, B.K., Michael, E., 2015. Bayesian calibration of simulation models for supporting management of the elimination of the macroparasitic disease, lymphatic filariasis. Parasit. Vect. 8 (1), 1.
Smith, M.E., Singh, B.K., Irvine, M.A., Stolk, W.A., Subramanian, S., Hollingsworth, T.D., Michael, E., 2017. Predicting lymphatic filariasis transmission and elimination dynamics using a multi-model ensemble framework. Epidemics 18, 16–28.
Stolk, W.A., De Vlas, S.J., Borsboom, G.J.J.M., Dik, J., Habbema, F., 2008. LYMFASIM, a simulation model for predicting the impact of lymphatic filariasis control: quantification for African villages. Parasitology 135 (13), 1583–1598.
Tanaka, M.M., Francis, A.R., Luciani, F., Sisson, S.A., 2006. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173 (3), 1511–1520.
Van Der Walt, S., Chris Colbert, S., Varoquaux, G., 2011. The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. 13 (2), 22–30.
Wakefield, J., 2007. Disease mapping and spatial regression with count data. Biostatistics 8 (2), 158–183.
Walker, D.M., Allingham, D., Lee, H.W.J., Small, M., 2010. Parameter inference in small world network disease models with approximate Bayesian computational methods. Phys. A: Stat. Mech. Appl. 389 (3), 540–548.
Waskom, M., Botvinnik, O., Hobson, P., Cole, J.B., Halchenko, Y., Hoyer, S., Miles, A., Augspurger, T., Yarkoni, T., Megies, T., Coelho, L.P., Wehner, D., Ziegler, E., Zaytsev, Y.V., Hoppe, T., Seabold, S., Cloud, P., Koskinen, M., Meyer, K., Qalieh, A., Allan, D., 2014. Seaborn: v0.5.0 (November 2014).
Wilkinson, R.D., 2013. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Stat. Appl. Genet. Molec. Biol. 12 (2), 129–141.
World Health Organization, et al., 2011. Global Programme to Eliminate Lymphatic Filariasis: progress report on mass drug administration, 2010[nl]programme mondial pour l'élimination de la filariose lymphatique: rapport sur l'administration massive de médicaments, 2010. Wkly. Epidemiol. Rec. 86 (35), 377–387.