HIV with contact-tracing: a case study in Approximate Bayesian Computation

Michael Blum, Viet Chi Tran

Approximate Bayesian Computation for epidemiological models: Application to the Cuban HIV-AIDS epidemic with contact-tracing and unobserved infectious population

Michael G.B. Blum[1], Viet Chi Tran[2]

[1] CNRS, Laboratoire TIMC-IMAG, Université Joseph Fourier, Grenoble, France

[2] Laboratoire P. Painlevé, Université de Lille 1, France

3rd October 2008

Abstract

Statistical inference with missing data is a recurrent issue in epidemiology where the infection process is only partially observable. In this paper, Approximate Bayesian Computation, an alternative to data imputation methods such as Monte Carlo Markov chain integration, is proposed for making inference in epidemiological models. This method of inference is not based on the likelihood function and relies exclusively on numerical simulations of the model. We apply the Approximate Bayesian Computation framework to calibrate an epidemiological model dedicated to the analysis of the HIV contact-tracing program in Cuba. We first evaluate numerically, using synthetic data sets, the statistical properties of the estimated posterior distributions obtained with different variants of Approximate Bayesian Computation. Then, once the epidemiological model has been calibrated with the Cuban VIH database, we make predictions concerning the efficiency of the detection system, and the evolution of the disease in the forthcoming years.

Keywords: mathematical epidemiology, stochastic SIR model, contact-tracing, unobserved infectious population, Approximate Bayesian Computation, HIV/AIDS epidemics.
AMS Subject Classification: 92D30, 62F15, 62M05, 62N02, 60K99.

1 Introduction

Mathematical modelling in epidemiology plays an important role for understanding and predicting the spread of diseases, as well as for comparing and evaluating public health policies. It has been emphasized in the literature (*e.g.* [6, 29]) that although deterministic modelling can be a guide for describing epidemics, stochastic models have their importance in featuring realistic processes and in quantifying confidence in parameters estimates and prediction uncertainty. Standard mathematical models in epidemiology consist in compartmental models, in which the population is structured in different classes composed of

the susceptible, infectious and removed individuals (SIR models [27]). Parameter estimation for SIR models is usually a difficult task because of missing observations which is a recurrent issue in epidemiology (*e.g.* [6, 23, 31, 32, 38]). Indeed, the infected population may be partially observed and the infection times may be missing. The computation of the likelihood in this context is numerically infeasible because it involves integration over all unobserved infection events.

Markov Chain Monte Carlo (MCMC) methods, that treat the missing data as extra parameters, have thus become increasingly popular for calibrating stochastic epidemiological models with missing data (*e.g.* [10, 20, 31, 32]). However, for high-dimensional missing observations, fine tuning of the proposal distribution is required for efficient MCMC algorithm [24]. In this paper, we show that SIR models with missing observations can be calibrated with the Approximate Bayesian Computation (ABC) approach, an alternative to MCMC, originally proposed for making inference in population genetics [33, 5]. This approach is not based on the likelihood function and relies solely on numerical simulations of the model. The idea of using simulations in models for which the distribution theory is intractable has been pioneered by Diggle and Gratton [15] in a frequentist setting. Interestingly, Silverman [35] in the discussion following their paper anticipated that compartmental models might constitute applications of the likelihood free approach.

The current work is motivated by the study of the Cuban HIV/AIDS database ([13]) that contains the dates of detection of the 8662 individuals that have been found to be HIV positive in Cuba between 1986 and 2007. The database contains additional covariates including the manner by which an individual has been found to be HIV positive. The individuals can be detected either by the so-called *random screening* or *contact-tracing* methods. The latter is the mode of detection by which a person, that is found to be HIV positive, is invited to give the names of her/his sexual partners (*e.g.* [25]) so that they can in turn take a detection test. As usual for infectious disease data, the total number of infectious individuals as well as the infectious times are unknown. Only data relative to the detected individuals are indeed available. However, once an individual has been detected, it is sometimes possible to infer by medical methods the probable time at which he/she was infected. In the Cuban HIV database, these reconstructed times are available for the sixth first years of the epidemic.

The paper is organized as follows. In the first section, we introduce the stochastic SIR model with contact tracing that has been developed by Clémencon *et al.* [12]. The second section is devoted to the ABC methods. In this section, we present the standard ABC methodology and its regression-based variants that have been developed [5, 7] as well as an original extension to the case of path-valued summary statistics that is suited to our study. In the third section, we compare the statistical properties of the different ABC approaches using synthetic data sets. The last section concentrates on the analysis of the database for HIV-AIDS in Cuba and the prediction of the dynamic of this epidemics. We address several questions: what is the percentage of the epidemic that is known [14, 25], or equivalently what is the efficiency of the system of detection in Cuba [13]; how many new cases of HIV are expected in the forthcoming years; and what is the proportion of detections that will be obtained in the contact-tracing program.

2 A stochastic SIR model for HIV/AIDS epidemics with contact-tracing

In this work, we restrict our study to the sexually-transmitted epidemic of HIV in Cuba. It has been inferred that 90% of the seropositive individuals have contracted the disease by sexual contacts [25]. For modelling the dynamics of the number of known and unknown HIV cases, we consider the SIR-type model developed in [12]. The population is divided into three main classes S, I and R corresponding to the *susceptible*, *infectious*, and *detected* individuals considered as *removed* because we assume that they do not transmit the disease anymore (see Figure 1). The population of the susceptible individuals, of size S_t, at time $t > 0$, consists of the sexually active seronegative (healthy) individuals. Individuals immigrate into the class S with a rate λ_0 and leave it by dying/emigrating, with rate $\mu_0 S_t$, or by becoming infected. The class of infectious individuals, of size I_t at time $t > 0$, corresponds to the seropositive individuals who have not taken a detection test yet and may thus contaminate new susceptible individuals. We assume that each individual may transmit the disease to a susceptible individual at rate λ_1 so that the total rate of infection is equal to $\lambda_1 S_t I_t$. Individuals leave the class I when they die/emigrate with a total rate of $\mu_1 I_t$, or when they are detected to be HIV positive.

The class R of the detected individuals, of size R_t at time t, is subdivided into two subclasses whether their seropositivity has been revealed by random screening or contact tracing. As already mentioned, contact tracing in Cuba consists in testing the sexual contacts of detected individuals [25]. In the following, we denote by R_t^1 (resp. R_t^2) the size, at time $t > 0$, of the population of removed individuals detected by random screening (resp. contact-tracing). We assume that the total rate of detection by random screening is $\lambda_2 I_t$. Concerning contact tracing detection, the model shall capture the fact that the contribution of a removed individual to the rate of detection depends on the time elapsed since she/he has been found to be HIV positive. In the sequel, we will consider the two following expressions for the total rate of contact-tracing detection

$$\lambda_3 I_t \sum_{i \in R} \Psi(t - T_i) \quad \text{and} \quad \lambda_3 I_t \sum_{i \in R} \Psi(t - T_i)/(I_t + \sum_{i \in R} \Psi(t - T_i)), \tag{2.1}$$

where Ψ is a positive function and T_i denotes the time at which a removed individual i has been detected. The weight function Ψ determines the contribution of a removed individual i to the contact-tracing control according to the time $t - T_i$ she/he has been detected. In the following we will restrict our analysis to $\Psi(t) = e^{-ct}$, for $c > 0$, so that the contribution of a removed individual to the contact-tracing control decreases exponentially with the time elapsed since she/he has been detected. The first rate in (2.1) corresponds to a mass action principle, and is proportional to the weighted number of detected individuals. The second rate in (2.1) corresponds to a model with frequency dependence. Further details and examples of more general infection and detection rates can be found in [12] as well as a more elaborate mathematical definition of the model based on random measures.

In the following we denote by $\theta = (\mu_1, \lambda_1, \lambda_2, \lambda_3, c)$ the multivariate parameter of the model. When there is no missing observation, Clémencon *et al.* [12] studied the maximum likelihood estimators and proved their consistency and asymptotic normality.

2.1 Connection between the stochastic and the deterministic SIR model

In this section, we focus on the first form of detection rate by contact tracing proposed in equation (2.1) which corresponds to a mass action principle. Similar results can be obtained for the second rate of equation (2.1). The evolution of the SIR process described in Section 2 can be described by Stochastic Differential Equations (SDEs) driven by Poisson point measures and which describe the population at an individual level. These SDEs are given and studied in [12]. To link the stochastic SIR process with the classical equations of epidemiology [27], Clémençon *et al.* [12] show that in a large population renormalization, the individual-centered SIR process converges to the solution of the following system of PDEs

$$\left\{\begin{array}{lcl} \frac{ds_t}{dt} & = & \lambda_0 - \mu_0 s_t - \lambda_1 s_t i_t \\ \frac{di_t}{dt} & = & \lambda_1 s_t i_t - (\mu_1+\lambda_2) i_t - \lambda_3 i_t \int_{\mathbb{R}_+} \Psi(a)\rho_t(a)\,da \\ \frac{\partial \rho_t}{\partial t}(a) & = & -\frac{\partial \rho_t}{\partial a}(a) \\ \rho_t(0) & = & \lambda_2 i_t + \lambda_3 i_t \int_{\mathbb{R}_+} \Psi(a)\rho_t(a)\,da \end{array}\right. \tag{2.2}$$

where s_t, and i_t denote the size of the susceptible and infectious populations at time $t \geq 0$, and $\rho_t(a)$ denotes the density of individuals having been detected since a time a at time t ($0 \leq a \leq t$). This PDE system with age provides an alternative to delay equations (*e.g.* [28, 41, 42]) and discrete stage structured models (*e.g.* [26]). With the exponential form for Ψ, the PDE system reduces to the following ODE

$$\left\{\begin{array}{lcl} \frac{ds_t}{dt} & = & \lambda_0 - \mu_0 s_t - \lambda_1 s_t i_t \\ \frac{di_t}{dt} & = & \lambda_1 s_t i_t - (\mu_1+\lambda_2) i_t - \lambda_3 i_t r_t \\ \frac{dr_t}{dt} & = & \lambda_2 i_t + \lambda_3 i_t r_t - c r_t \end{array}\right. , \tag{2.3}$$

where $r_t = \int_0^t e^{-ca}\rho_t(a)\,da$, so that r_t measures, at time t, the contribution of the removed individuals to the rate of detection by contact tracing.

Apart from the inherent stochastic nature of epidemic propagation that has already been pointed out, and that may be particularly important for small populations (see *e.g.* [18]), considering a stochastic SIR model rather than its deterministic counterpart can present at least two important advantages for parameter calibration. First, it is quite straightforward to perform exact simulations from the stochastic model (see Section 2.2) and this is one motivation for considering ABC methods. Second, the individual-centered stochastic process suits the formalism of statistical methods, which are based on samples of individual data. Within this formalism, problems such as missing or noisy data can be tackled with the arsenal of statistical methods. Since the estimators constructed on the stochastic process converge to the parameters of the PDEs (see [12]), this provides new alternative approaches for calibrating the parameters of PDEs and ODEs (*e.g.* [2, 3, 9, 29]).

2.2 Exact path simulation of the SIR model with contact-tracing

The main difficulty for simulating the SIR model with contact tracing lies in the fact that the rate of detection by contact-tracing evolves with time. Here, we consider an acceptance-rejection technique for simulating the SIR-type process between time 0 and the end of the observation period at time $t = T$ (see [40, 17, 19] for similar algorithms).

In order to simplify the algorithm, we assume that the values of the parameter λ_0 and μ_0 are such that the size of the population of susceptible individuals remains constant during the observation period.

The algorithm can be described iteratively as follows:

1. Start with a population of $S = S_0$ susceptible individuals, $I = I_0$ infectious individuals, no detected individuals, and a vector of the detection times set to the null vector.

2. Assume that we have already simulate k events, and that the k^{th} event occurs at time t_k. We describe, in the following, how to simulate the $k+1^{\text{th}}$ event. Let $\tau = t_k$ be the current time.

   (a) Simulate an independent exponential random variable $\mathcal{E}$ with parameter

   $$C_k = \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} + \lambda_3 I_{t_k} R_{t_k}$$

   which is an upper bound of the sum of the rates of occurrence of all possible events. The time of the next putative event is defined as $\tau' = \tau + \mathcal{E}$.

   (b) Increment the ages by $\mathcal{E}$

   (c) If $\tau' > T$, then stop the simulation process.
   Else, simulate an independent uniform random variable U in $[0, C_k]$.
   If $0 \leq U < \lambda_1 S_{t_k} I_{t_k}$ then a susceptible individual is removed, and an infectious individual is added.
   If $\lambda_1 S_{t_k} I_{t_k} \leq U < \lambda_1 S_{t_k} I_{t_k} + \mu_1 I_{t_k}$ then an infectious individual is removed.
   If $\lambda_1 S_{t_k} I_{t_k} + \mu_1 I_{t_k} \leq U < \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k}$ then an infectious individual is removed, an individual detected by random screening is added, and a zero is added to the vector of the detection times.
   If $\lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} \leq U < \lambda_1 S_{t_k} I_{t_k} + (\mu_1 + \lambda_2) I_{t_k} + \lambda_3 I_{t_k} \sum_{i \in R} \psi(\tau' - T_i)$ then an infectious individual is removed, an individual detected by contact tracing is added, and a zero is added to the vector of the detection times.
   Else, nothing happens. Return to step 2a with the current time set equal to τ'.

The complexity of the algorithm scales with the total number of events, so that it depends on the parameter vector θ. When simulating the Cuban HIV epidemic, $t = 0$ corresponds to the beginning of the epidemic in 1986 and the simulations are performed until the end of the observation period at time $T = 21.5$, in July 2007.

3 Approximate Bayesian Computation for epidemic models

Bayesian approaches have been widely used in epidemiology (*e.g.* [11, 10, 22, 23, 32, 37, 38]) and the reasons to adopt a Bayesian approach are manyfold. First, in epidemiology, credibility intervals might be of greater interest than simple point estimates. The same reason was invoked in population genetics [39]. Whereas Bayesian algorithms usually provide samples from the posterior distribution from which it is straightforward to obtain

credibility intervals, frequentist confidence intervals are based on asymptotic derivations that might be poor approximations when the sample is too small or highly correlated. The second reason is that epidemiological models may contain nuisance parameters that should be integrated over when making inference. In the context of the SIR model with contact tracing, the parameters c and μ_1 are considered as the nuisance parameters whereas the parameters of interest are λ_1, λ_2, and λ_3. The third reason is that Bayesian methods offer a convenient way to handle missing observations that are treated as extra parameters. And, the last reason is that high dimensional integration, typically involved in Bayesian methods, might be more convenient than high dimensional optimization when the likelihood function is flat in certain directions. However, we lay absolutely no claim that Bayesian methods are the most relevant methods for SIR-type models and applications of the frequentist likelihood-free method of Diggle and Gratton [15], for instance, might be of great interest in this setting.

First, let us sum up the main principle of the ABC method. For simplicity, we deal here with densities, but the following description also holds when dealing with measures that are not absolutely continuous w.r.t. the Lebesgue measure. Let $\mathbf{x}$ be the available data and $\pi(\theta)$ be the prior density. Instead of focusing on the posterior density $p(\theta \,|\, \mathbf{x})$, ABC aims at a less informative *target* density $p(\theta \,|\, S(\mathbf{x}) = s_{obs}) \propto \Pr(s_{obs}|\theta)\pi(\theta)$ where S is a summary statistic with realization s_{obs} on the data, and which takes its values in a normed space. The summary statistic S can be a d-dimensional vector or an infinite-dimensional variable such as a L^1 function. Of course, if S is sufficient, then the two conditional densities are the same. The target distribution will also be coined as the *approximate posterior distribution* in the following. For obtaining the latter, the ABC method relies exclusively on simulations of the model and on comparisons between a vector of simulated and observed summary statistics.

Practically, ABC offers a convenient way for making inference since it can be applied to different variants of SIR-type model without modifications. Additionally, in the context of partially observed population and missing infectious times, MCMC methods require to reconstruct the unknown data which can be highly computationally intensive for large populations. For instance, [32] and [38] considered MCMC algorithms for populations consisting of about 100 individuals whereas the Cuban HIV-AIDS database contains almost 10,000 known HIV positive individuals, which makes the total (known and unknown) number of infectious individuals even larger.

3.1 Smooth rejection

Let us now describe how the ABC method with smooth rejection [5] generates random draws from the target distribution. We assume that a vector of summary statistics s_{obs} is calculated for the data. The algorithm can be described as follows

1. Generate N random draws (θ_i, s_i), $i = 1 \dots N$, where θ_i is generated from the prior distribution and where s_i is the vector of summary statistics calculated for the i^{th} synthetic data set, simulated from the generative model with parameter θ_i (see Section 2.2).

2. Associate to the i^{th} simulation the weight $W_i = K_\delta(\|s_i - s_{obs}\|)$, where $\|.\|$ is an appropriate metric and where $K_\delta(x) \propto K(x/\delta)$ in which $K : \mathbb{R} \mapsto \mathbb{R}_+$ is a Parzen-Rosenblatt kernel and $\delta > 0$ is a tolerance threshold.

3. Then, the distribution $(\sum_{i=1}^N W_i \delta_{\theta_i})/(\sum_{j=1}^N W_j)$ approximates the target distribution. In other words, the resulting weighted sample (θ_i, W_i) forms a sample from a distribution close to the target distribution.

In the case of a d-dimensional vector of summary statistics, because these statistics may span different scales, we consider a weighted Euclidean distance for $\|s_{obs} - s_i\|$, $i = 1 \dots N$, where the weights are equal to the inverse of the variance of each one-dimensional summary statistic [5]. When dealing with L^1-valued summary statistics, we consider the L^1-norm. Concerning the tolerance threshold δ, we rather set a tolerance rate P_δ that corresponds to the percentage of accepted simulations. This procedure amounts at choosing the P_δ quantile of the distances $\|s_i - s_{obs}\|$, $i = 1 \dots N$, for the tolerance threshold δ. When K is the indicator function $\mathbf{1}_{[0,\delta[}$, the ABC (non-smooth) rejection algorithm simply consists of keeping the simulations for which the distance between the vector of simulated and observed summary statistics is smaller than the prescribed tolerance δ. This was considered by Pritchard *et al.* [33] in the infancy of ABC applications in population genetics. In the sequel, an Epanechnikov kernel for K_δ is considered.

Once a sample from the target distribution has been obtained, as for classical Bayesian inference, several estimators may be considered for point estimation of each one-dimensional parameter λ_j, $j = 1 \dots 3$. We will consider here the means, medians and modes of the marginal posterior distributions.

- Using the weighted sample $(\lambda_{j,i}, W_i)$, $i = 1 \dots N$, the *means* of the target distributions $p(\lambda_j | s_{obs})$ are estimated as

$$\widehat{\lambda_j}(\delta) = \frac{\sum_{i=1}^N \lambda_{j,i} K_\delta(\|s_i - s_{obs}\|)}{\sum_{j=1}^N K_\delta(\|s_j - s_{obs}\|)}, \quad j = 1, \dots, 3 \tag{3.1}$$

  which is the well-known Nadaraya-Watson regression estimator of the conditional expectation $\mathbb{E}(\lambda_j \,|\, s_{obs})$ [30, 43].

- The *medians* of the marginal target distributions are estimated by computing the median of a non-weighted sample that has been obtained by sampling with replacement in the weighted sample $(\theta_i, W_i)_{i \in [\![1,N]\!]}$.

- The *modes* are estimated by maximizing the estimates $\widehat{p}(\lambda_j \,|\, s_{obs})$ of the marginal distributions of λ_j, $j = 1 \dots 3$, obtained as

$$\widehat{p}(\lambda_j \,|\, s_{obs}) = \frac{\sum_{i=1}^N K_\Delta(\lambda_{j,i} - \lambda) K_\delta(\|s_i - s_{obs}\|)}{\sum_{j=1}^N K_\delta(\|s_j - s_{obs}\|)}, \quad j = 1, \dots, 3 \tag{3.2}$$

  where Δ is a bandwith parameter for the density estimation.

For computing the 95% credibility intervals, we estimate the 97.5% and the 2.5% quantiles of the marginal target distributions in the same manner as the median.

Compared to standard likelihood approach, two important approximations are made in the ABC approach described above. First, and as mentioned before the target distribution is not the true posterior distribution. Consequently, choosing a vector of informative summary statistics is of considerable importance since a loose choice would lead to flat target distributions. Second, in the case a d-dimensional vector of summary statistics,

it is known (*e.g.* [8]) that when $N \to +\infty$, the estimators (3.1) and (3.2) converge if the tolerance rate satisfies $\lim_{N\to+\infty} \delta_N = 0$, so that their bias converge to 0, and $\lim_{N\to+\infty} N\delta_N^d = +\infty$, so that their variances converge to 0. This means that the tolerance rate has to be small, but large enough to retain a sufficient number of points for computations. As d increases, a larger tolerance threshold shall be chosen involving that the bias of the estimator might be large. This phenomenon known in statistics as the *curse of dimensionality* may be a serious issue for the ABC-rejection approach since large values of δ distort the approximation of the target distribution $p(\theta|s_{obs})$ by the approximate distribution obtained with the rejection method. The following paragraph presents the correction originally introduced by Beaumont *et al.* [5] and refined by Blum and Francois [7] to reduce the curse of dimensionality.

3.2 Regression adjustment for vector-valued summary statistics

For large thresholds δ, the rejection method may retain couples (θ_i, s_i) with summary statistics far from s_{obs} meaning heuristically that the associated θ_i's may not be considered as random draws from the distribution $p(\theta \,|\, s_{obs})$ anymore. To overcome this fact, Beaumont *et al.* [5] performed an adjustment by replacing in (3.1) and (3.2) the θ_i's by corrected values θ_i^*. Doing this, they found that the resulting target distributions were numerically insensitive to the tolerance threshold δ for a large range of small enough values of δ.

The idea of correction is the following. The (θ_i, s_i) are independent and identically distributed (i.i.d.) random variables (r.v) and it is possible to write $\theta_i = F(s_i, \varepsilon_i) =: F_{s_i}(\varepsilon_i)$ where F is a (possibly complicated) function and where ε_i are i.i.d r.v. independent of the s_i's. We assume that for all s, $F_s^{-1}(\theta)$ can be properly defined.

Proposition 3.1. *The r.v.* $(\theta_i^* := F_{s_{obs}} \circ F_{s_i}^{-1}(\theta_i))_{i\in[\![1,N]\!]}$ *are i.i.d. with density* $p(\theta \,|\, s_{obs})$.

The proof of this proposition stands in appendix. Of course, the function F_s is unknown and shall be approximated. The ABC algorithm with regression adjustment can be described as follows

1. Simulate, as in the Step 1 of the rejection algorithm, a sample $(\theta_i, s_i)_{i\in[\![1,N]\!]}$ of i.i.d. r.v.

2. By making use of this weighted sample, approximate the function F such that $\theta_i = F(s_i, \varepsilon_i)$.

3. Replace the θ_i by the corrected θ_i^*. The resulting weighted sample (θ_i^*, W_i), $i = 1 \dots N$, form a sample from the target distribution. The weights $W_i = K_\delta(\|s_i - s_{obs}\|)$ heuristically give less importance to values for which the adjustment has been more important.

Beaumont *et al.* local linear regressions (LOCL) The case where F is approximated by the linear model $F(s, \varepsilon) = \alpha + s^t\beta + \varepsilon$ was considered in [5]. The parameter α and β are inferred by minimizing the weighted least-square $\sum_{i=1}^N K_\delta(\|s_i - s_{obs}\|)(\|\theta_i - (\alpha + (s_i - s_{obs})^T\beta)\|_2^2$. The estimator $\hat{\alpha}$ corresponds to the estimation of $\mathbb{E}(\theta|s_{obs})$ obtained with standard local polynomial regression (see *e.g.* [16]). Using Proposition 3.1, the correction of [5] is derived as

$$\theta_i^* = \theta_i - (s_i - s_{obs})^T\hat{\beta},\ i = 1, \dots, N.$$

Blum and François' nonlinear conditional heteroscedastic regressions (NCH) To relax the assumptions of homoscedasticity and linearity inherent to the local linear regression model, Blum and Francois [7] approximated F by $F(s, \varepsilon) = m(s) + \sigma(s) \times \varepsilon$ where $m(s)$ denotes the conditional expectation $\mathbb{E}(\theta|s)$, and $\sigma^2(s)$ the conditional variance. The estimators $\hat{m}$ and $\log \hat{\sigma}^2$ of the conditional expectation and of the logarithm of the conditional variance are found by adjusting two feed-forward neural networks [34] using weighted least squares. To motivate the choice of neural networks, Blum and Francois [7] emphasized that the regression layer is not performed on the (possibly high dimensional) subspace of the summary statistics but on a subspace of lower dimension found via internal projections. For the NCH model, parameter adjustment is performed as follows

$$\theta_i^* = \hat{m}(s_{obs}) + (\theta_i - \hat{m}(s_i)) \times \frac{\hat{\sigma}(s_{obs})}{\hat{\sigma}(s_i)},\ i = 1, \ldots, N.$$

In practical applications of the NCH model, we train $L = 10$ neural networks for each conditional regression (expectation and variance) and we average the results of the L neural networks to provide the estimates $\hat{m}$ and $\log \hat{\sigma}^2$. This averaging procedure considerably reduces the variance of the posterior estimates (results not shown). It is coined as the predictive Bayes approach for neural networks in [34].

Reparameterization In both regression adjustment approaches, the regressions can be performed on transformations of the responses θ_i rather that on the responses themselves. Parameters whose prior distributions have finite supports are transformed via the logit function and non-negative parameters are transformed via the logarithm function. These transformations guarantee that the θ_i^*'s lie in the support of the prior distribution. The transformations have the additional advantage of stabilizing the variance [1].

3.3 Data and summary statistics

Data Starting at the time of the first detection in 1986, the Cuban AIDS data consists principally of the detection times at which the individuals are found to be HIV positive. At the time of the last detection event that is considered, in July 2007, there is a total of 8,662 individuals that have been found to be HIV positive. For each detection event, covariates are the way of detection (random screening and contact-tracing), and the possible time of infection when available (obtained by medical methods or discussions with the patient). For the sixth first years of the epidemic (1986-1992), the infection times have been reconstructed. We neglect here the noise arising from the reconstruction of the infectious times.

Summary statistics In the following, we consider two different ways of capturing the information contains in the Cuban AIDS database.

In the first approach, we consider three types of summary statistics. First, we compute the number of individuals detected, at the end of the observation period $T = 21.5$ years, by random screening $R^1_{T=21.5} = 6157$ and by contact tracing $R^2_{T=21.5} = 2145$. Second, for each year, we compute the number of individuals that have been found to be HIV positive $R^l_{j+1} - R^l_j$ for $j = 0, \ldots, 20$, and $l = 1, 2$. Last, we take profit of the reconstructed infectious times. We compute the number of new infectious for each of the the sixth first

years $I_{j+1} - I_j$ for $j = 0, \ldots, 5$, as well as the mean time during which an individual is infected but has not been detected yet. This mean time corresponds to the mean sojourn time in the class I for the sixth first years of the epidemic. When considering the total observation period, a total of 61 summary statistics are considered in this approach.

The second approach is based on a different set of summary statistics. We consider here the two (infinite-dimensional) statistics $(R^1_t, t \in [0,T])$ and $(R^2_t, t \in [0,T])$. The L^1 norm between the simulated paths R^l_i $(l = 1, 2,\ i = 1, \ldots N)$ and the observed ones R^l_{obs} $(l = 1, 2)$ is

$$\|R^l_{obs} - R^l_i\|_1 = \int_0^T |R^l_{obs,s} - R^l_{i,s}|\, ds \quad ,\ l = 1, 2,\ i = 1, \ldots, N.$$

For the weights W_i, we choose a product kernel so that $W_i = K_{\delta_1}(\|R^1_{obs} - R^1_i\|_1) K_{\delta_2}(\|R^2_{obs} - R^2_i\|_1)$ where δ_1, δ_2 are 2 possibly different tolerance thresholds. In practice, we set the same tolerance rate $P_{\delta_1} = P_{\delta_2}$ for the two summary statistics so that they contribute equally to the target distribution.

Let us emphasize a similarity that exists with deterministic approaches in which parameter calibration relies on the minimization of a *cost* function (*e.g.* [3, 9]). In the second approach based on the two trajectories the value of θ_i that maximizes the weights W_i, $i = 1 \ldots N$, maximizes $(1 - \|R^1_{obs} - R^1_i\|_1/\delta_1)^2 (1 - \|R^2_{obs} - R^2_i\|_1/\delta_2)^2$, when considering Epanechnikov kernels for K_{δ_1} and K_{δ_2}. For small differences between the simulated and observed paths, this amounts at minimizing the quadratic cost function $\|R^1_{obs} - R^1_i\|^2_1/\delta_1 + \|R^2_{obs} - R^2_i\|^2_1/\delta_2$.

4 Validation on synthetic data sets

To check the validity of the ABC algorithms, we simulate $M = 200$ synthetic data sets for a given value of the parameters. In order to work on a data similar to the Cuban database for the HIV/AIDS epidemic, we choose $\mu_1 = 2 \times 10^{-6}$, $\lambda_1 = 1.14 \times 10^{-7}$, $\lambda_2 = 3.75 \times 10^{-1}$, $\lambda_3 = 6.55 \times 10^{-5}$, and $c = 1$. The initial conditions are set to $S_0 = 6\,10^6$, the size of the Cuban population in the age-group 15-49 [13], $I_0 = 232$ and $R_0 = 0$. When analyzing the synthetic data sets, we simulate only 6 years of the epidemics.

In the following, we study four variants of ABC for estimating λ_1, λ_2, and λ_3. When using the finite dimensional vector of summary statistics, we perform the smooth rejection approach as well as the LOCL and NCH corrections with a total of 21 summary statistics: the 18 summary statistics corresponding to the yearly increase of R^1, R^2, and I; the final numbers of detected individuals R^1_6 and R^2_6; and the mean sojourn time in the class I. When considering the two trajectories $(R^1_t, t \in [0,T])$ and $(R^2_t, t \in [0,T])$ as the summary statistics, we perform the smooth rejection approach using the product kernel for the computation of the weights W_i. Each of the $M = 200$ estimations of the approximate posterior distributions are performed using a total of $N = 5000$ simulations of the SIR model with the mass action principle (first rate in equation (2.1)).

Prior distributions The prior distributions of the parameters $\mu_1, \lambda_1, \lambda_2$ and λ_3 are chosen to be uniform on a log scale. The prior distributions are defined on a log scale to reflect our uncertainty in the order of magnitude of the parameters. More specifically, the prior distribution for $\log(\mu_1)$ is $\mathcal{U}(-6, -4)$ where $\mathcal{U}(a, b)$ denotes the uniform distribution

on the interval (a, b). The prior distribution for $\log(\lambda_1)$ is $\mathcal{U}(-9, -6)$, the prior distribution for $\log(\lambda_2)$ is $\mathcal{U}(-4, 3)$ and the prior distribution for $\log(\lambda_3)$ is $\mathcal{U}(-8, 2)$. The bounds of the uniform distributions are set to keep the simulations from being degenerate. The prior for the parameter c is $\log(2)/\mathcal{U}(1/12, 5)$. This prior is chosen so that the half time of ψ is uniformly distributed between 1/12 and 5 years.

Point estimates of θ and credibility intervals Figure 2 displays the boxplots of the 200 estimated modes, medians, 2.5% and 97.5% quantiles of the posterior distribution for λ_1 as a function of the tolerance rate. The corresponding figures for λ_2 and λ_3 can be found in the Supplementary Material.

First, we find that medians and modes are equivalent for the NCH and LOCL variants, while the mode is less biased in the rejection methods.

For the lowest tolerance rates, the point estimates obtained with the four possible methods are close to the value $\lambda_1 = 1.14 \times 10^{-7}$ used in the simulations, with smaller credibility intervals for the LOCL and NCH variants.

When increasing the tolerance rate, the bias of the point estimates obtained with the rejection method with 21 summary statistics slightly increases. By contrast, up to tolerance rates smaller than 50%, the bias of the point estimates obtained with the three other methods remains small. As could be expected, the widths of the credibility intervals obtained with the rejection methods increase with the tolerance rate while it remains considerably less variable for the LOCL and NCH variants.

Mean square error To further investigate the differences between the statistical properties of the different methods, we compute the rescaled mean square errors (RMSEs). RMSEs are computed on a log scale and rescaled by the range of the prior distribution

$$\mathrm{RMSE}(\lambda_j) = \frac{1}{M}\sum_{k=1}^{M} \frac{(\log(\widehat{\lambda}_j^k) - \log(\lambda_j))^2}{\mathrm{Range}(\mathrm{prior}(\lambda_\mathrm{j}))^2}, \quad j = 1, \ldots, 3, \tag{4.1}$$

where $\widehat{\lambda}_j^k$ is a point estimate obtained with the k^{th} synthetic data set. Figure 3 displays the values of the RMSEs as functions of the tolerance rate. We find that the smallest values of the RMSE are usually reached for the lowest value of the tolerance rates (but see the RMSEs for λ_2 in Figure 3). For λ_1 and λ_2, the RMSEs of the point estimates obtained with the four different methods are comparable for the lowest tolerance rate. However, the smallest values of the RMSE's are always found when performing the rejection method with the two summary statistics R^1 and R^2. This finding is even more pronounced for the parameter λ_3 suggesting that the two trajectories R^1 and R^2 are the most informative summary statistics for estimating the λ_j, $j = 1 \ldots 3$.

Rescaled mean credibility intervals To compare the whole posterior distributions obtained with the four different methods, we additionally compare the different credibility intervals. As displayed by Figure 2 and the Figures 1 and 2 of the Supplementary Material, the credibility intervals obtained with the smooth rejection schemes increase importantly with the tolerance rate whereas such an important increase is not observed for the regression approaches. To further compare the 95% credibility intervals obtained

with the different methods, we computed the rescaled mean credibility intervals ($RMCI$) defined as follows

$$\text{RMCI} = \frac{1}{M}\sum_{k=1}^{M}\frac{|IC_j^k|}{\text{Range}(\text{prior}(\lambda_\text{j}))}, \quad j = 1, \ldots, 3, \tag{4.2}$$

where $|IC_j^k|$ is the length of the k^{th} estimated credibility interval for the parameter λ_j.

As displayed by Figure 4, credibility intervals obtained with the NCH method are clearly the thinnest, those obtained with the rejection methods are the widest and those obtained with the LOCL method have intermediate width. We additionally found that the RMCI's obtained with the regression methods also increase with the tolerance rate. This phenomenon is partly due to the increase of the variance of the extreme quantiles that is observed when the tolerance rate increases. In the following, we perform the NCH correction when considering the finite-dimensional vector of summary statistics. This choice is motivated by the small RMSEs and RMCIs obtained with the NCH method (Figure 3 and 4).

5 Application to the Cuban HIV epidemic

5.1 Parameter calibration and goodness of fit

Using the NCH method when considering the 61 summary statistics and smooth rejection with the 2 trajectories R^1 and R^2 as the summary statistics, we now fit the SIR model of Clemençon *et al.* [12] to the Cuban HIV data using a total of 100,000 simulations. We consider the two different rates of contact-tracing detection (2.1). We use the same initial conditions and priors as in Section 4.

To set the value of the tolerance rate P_δ for each of the two procedures, we consider the 21st years of the epidemic as the training data set and we choose the value of the tolerance rate P_δ that minimizes the prediction error at the end of the 21st year and a half of the epidemic. The prediction error is defined as

$$\text{Pred Error} = \mathbb{E}_{P_\delta}\left[\left(\frac{R^1_{21.5}(P_\delta) - R^1_{obs,21.5}}{R^1_{obs,21.5}}\right)^2 + \left(\frac{R^2_{21.5}(P_\delta) - R^2_{obs,21.5}}{R^2_{obs,21.5}}\right)^2\right] \tag{5.1}$$

where $\mathbb{E}_{P_\delta}$ denotes the expectation with respect to the approximate posterior distribution found with a tolerance rate set to P_δ.

Once the tolerance rate P_δ has been chosen and once the approximate posterior distribution has been obtained, we investigate the goodness of fit of the SIR-type model by simulating paths of the SIR model associated with parameters θ sampled from the approximate posterior distribution. In Figures 5 and 6, we display the Posterior Predictive Distributions (PPD) [21] of $R^1_{21.5}$, $R^2_{21.5}$, I_6, as well as the mean sojourn time in the class I. We investigate if the observed values fall within the supports of the PPDs.

The best predictions correspond to the model of frequency dependence (second rate in equation (2.1)) that has been fitted with the two trajectories R^1 and R^2 as the summary statistics. As displayed by Figure 6, $R^1_{obs,21.5}$ is close to the mode of the PPD and $R^2_{obs,21.5}$ is smaller than the mode but still contained in the PPD. By contrast, the mean sojourn

time in the class I is not contained in the PPD and the observed number of infectious individuals is in the lower tail of the PPD. An explanation might be that an age-structure has to be taken into account for the infection rate in order to capture the non-Markovian effects (*e.g.* [38]). A model with an increasing infectious rate could diminish the mean sojourn time in the class I and increase by compensation the number of infections to maintain the infection pressure constant.

Concerning the PPDs obtained with the NCH method, we find that they have extremely wide supports for both the model with a mass action principle (Figure 5), and the model with frequency dependence (results not shown). The PPD obtained with the mass action principle and the two trajectories R^1 and R^2 as the summary statistics are peaky but $R^1_{obs,21.5}$ is not contained in the PPD (see Supplementary Material). In this model, because the rate of contact tracing detection increases linearly with the total number of detected individuals, the number of individuals detected by contact tracing may increase too rapidly compared to the observed data.

In the following, we consider the model with frequency dependence that has been fitted with the two trajectories R^1 and R^2. For parameter estimation and estimation of credibility intervals, we choose a larger tolerance rate of $P_\delta = 5\%$ for each summary statistic so that enough data points (of the order of 1000) are available for estimating the credibility intervals. The point estimates of Table 1 have been obtained by maximizing the estimated marginal posterior distributions.

Parameter	Point estimates	Lower bound of the 95% credibility interval	Upper bound of the 95% credibility interval
λ_1	$5.4\,10^{-8}$	$3.9\,10^{-8}$	$2.3\,10^{-7}$
λ_2	0.13	0.007	1.17
λ_3	0.19	0.03	0.82

Table 1: *Point estimates and 95% credibility intervals for the infection rate λ_1 and the detection rates λ_2 and λ_3.*

The point estimate of the rate of infection λ_1 implies that the net rate of infection per infectious individuals $\lambda_1 S$ is equal to .32 (95%CI $= 0.23 - 1.37$). This means that the waiting time before an infectious individual, that has not been detected yet, infect an other individual is 3.1 years (95%CI $= 0.72 - 4.34$).

5.2 The dynamic of the Cuban HIV epidemic

Reconstruction of the cumulative numbers of detections Figure 7 displays the dynamics predicted by the SIR model for the number of individuals detected by random screening ($R^1_t, t \in [0,T]$), by contact tracing ($R^2_t, t \in [0,T]$) and for the number of unknown infectious individuals ($I_t, t \in [0,T]$). Interestingly, there is a really good fit between the real and predicted numbers of individuals detected by random screening except between 1992 and 1995. This period corresponds to the period of crisis that followed the collapse of the Soviet Union and during which the HIV detection system received less attention [12]. We also find that there is a slight discrepancy in the recent years (2000-2007) between the real and predicted numbers of individuals detected by contact tracing.

The SIR model predicts a larger number of contact tracing detections which may reveal a weakening in the contact tracing system.

Performance of the contact-tracing system When testing for the performance of the contact tracing system, Hsieh *et al.* [25] computed the coverage of the epidemic defined as the percentage of infectious individuals that have been detected $(R^1+R^2)/(I+R^1+R^2)$. As displayed by Figure 8, the SIR model predicts a coverage of 62% (95%CI = 36%−66%) in 2000 that is much lower than a coverage of 83% (75% − 87%) as inferred in [25] (the confidence interval is given in [14]). However, since the PPDs of Figure 6 shows that the SIR model predicts less infectious individuals than observed, the coverage might still be overestimated.

Predictions Additionally, simulations of the SIR model provide predictions for the evolution of the HIV dynamic in the forthcoming years. Obviously, predictions for the year 2015, for instance, should be interpreted with caution since they rely on the assumption that the different rates remain constant. The SIR model predicts that in 2015, 42,000 (95%CI = 29,000 − 107,000) individuals will be infected since the beginning of the epidemic in Cuba. Among these infected individuals, a proportion of 45% (95%CI = 29% − 46%) will be detected by random screening an a proportion of 21% (95%CI = 10% − 22%) will be detected by contact tracing. As displayed by Figure 7, the SIR-type model with contact tracing predicts that the total proportion, among the detected individuals, of individuals detected by contact tracing would reach an asymptote of 32% (95%CI = 25% − 33%) in 2015 whereas the counting data reveals a drop in the proportion of individuals detected by contact tracing. The total number of infected individuals in 2015 corresponds to 27,000 (95%CI = 19,000 − 80,000) new cases of HIV between July 2007 ($T = 21.5$) and January 2015 ($T = 29$). In the same period of time, the SIR model predicts that 12,000 individuals (95%CI = 9,000 − 24,000) will be detected by random screening and 6,000 individuals (95%CI = 4,000 − 8,000) will be detected by contact tracing.

6 Conclusions

In the context of temporal epidemiological data, we show that Approximate Bayesian Computation (ABC) techniques can provide reasonable estimates of the parameters of interest such as the infection and detection rates. ABC inference relies solely on the simulations of the model and can therefore be applied for various epidemiological models defined in terms of an implicit stochastic mechanism [15]. In its broad lines, ABC consists of rejecting the simulations that produce summary statistics too different from the observations. In this paper, we consider both finite dimensional summary statistics, as well as infinite dimensional ones, as the observations consists of the cumulative number of detected individuals as a function of time. The point estimates of the parameters λ_j, $j = 1 \dots N$, with the smallest quadratic errors were obtained with the rejection method based on the infinite-dimensional statistics. However, the 95% credibility intervals obtained with this method are large and critically depends on the tolerance rate. By contrast, regression-based adjutment methods, and the NCH method more particularly, considerable shorten the credibility intervals and are less sensitice to the tolerance rate.

Applications of regression-based ABC methods [5, 7] constitute therefore a solution for "stabilizing" the credibility intervals. However, no ABC-regression based methods have been developed so far for infinite-dimensional summary statistics.

In the last section of the paper, we calibrate the SIR model to the Cuban HIV data that contains the times at which the Cuban individuals have been found to be HIV positive. By comparing the posterior predictive distributions obtained with the fitted SIR model to the observed ones, we find that a model that accounts for a frequency-dependent rate of contact tracing detection provides a good fit to the data. We suggest here two possible improvements of the SIR model that could ameliorate the fit to the Cuban HIV data. First, because the mean time during which an infected individual has not been detected yet, is not well predicted by the SIR model, detection rates that depend on the time elapsed since infection could be considered. Second, the rates of infection and detection could vary in time so that changes in the dynamics of infection and detection would be captured. Such models would contain additional parameters and the simple rejection scheme might require a prohibitive number of simulations for targeting in the approximate posterior distribution. For high-dimensional models, adaptive ABC algorithms [36, 4], that use the simulations to modify the sampling distribution of the parameter θ, might constitute interesting ways to explore for the future of ABC in epidemiology.

Acknowledgment: The authors are grateful to Pr. H. De Arazoza of the University of La Havana and to Dr. J. Perez of the National Institute of Tropical Diseases in Cuba for granting them access to the HIV/AIDS database. This work has been supported by the French Agency for Research through the ANR MAEV and ANR Viroscopy, and by the Rhône-Alpes Institute of Complex Systems (IXXI).

References

[1] G.E.P Box an D.R. Cox. An analysis of transformations. Journal of the Royal Statistical Society: Series B, 26:211–246, 1964.

[2] R.M. Anderson and R.M.C. May. Infectious Diseases of Humans: dynamics and Control. Oxford University Press, Oxford, 1991.

[3] H.T. Banks, J.A. Burns, and E.M. Cliff. Parameter estimation and identification for systems with delays. SIAM Journal of Control and Optimization, 19(6):791–828, 1981.

[4] M.A. Beaumont, J.-M. Marin, J.-M. Cornuet, and C.P. Robert. Adaptivity for ABC algorithms: the ABC-PMC scheme. 2008. `arXiv:0805.2256v6`.

[5] M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian computation in population genetics. Genetics, 162:2025–2035, 2002.

[6] N.G. Becker and T. Britton. Statistical studies of infectious disease incidence. Journal of the Royal Statistical Society: Series B, 61(2):287–307, 1999.

[7] M.G.B. Blum and O. François. Likelihood-free Bayesian inference without rejection. arXiv:0809.4178, 2008.

[8] D. Bosq and J.-P. Lecoutre. Théorie de l'estimation fonctionnelle. Ecole Nationale de la Statistique et de l'Administration Economique et Centre d'Etudes des Programmes Economiques. Economica, 1987.

[9] J.A. Burns, E.M. Cliff, and S.E. Doughty. Sensitivity analysis and parameter estimation for a model of Chlamydia Trachomatis infection. Journal of Inverse Ill-Posed Problems, 15:19–32, 2007.

[10] S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boelle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Statistics in Medicine, 23:3469–3487, 2004.

[11] S. Cauchemez and N.M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in london. Journal of the royal society: interface, 5:885–897, 2004.

[12] S. Clemençon, V.C. Tran, and H. De Arazoza. A stochastic SIR model with contact-tracing: large population limit with statistical applications. Journal of Biological Dynamics, 2008. A paraître, last corrected version `http://fr.arxiv.org/abs/0807.3462`.

[13] H. de Arazoza, J. Joanes, R. Lounes, C. Legeai, S. Clémençon, J. Perez, and B. Auvert. The HIV/AIDS epidemic in Cuba: description and tentative explanation of its low prevalence. BMC Infectious Disease, 7:130, 2007.

[14] H. de Arazoza, R. Lounes, J. Perez, and T. Hong. What percentage of the cuban hiv-aids epidemic is known. Rev Cubana Med Trop, 55:30–37, 2003.

[15] P.J. Diggle and R.J. Gratton. Monte Carlo methods of inference for implicit statistical models. Journal of the Royal Society: Series B, 46:193–227, 1984.

[16] J. Fan. Design-adaptive nonparametric regression. Journal of the American Statistical Association, 87(420):998–1004, 1992.

[17] R. Ferrière and V.C. Tran. Stochastic and deterministic approaches for populations with age and trait-structure. 2008. arXiv:0809.3767.

[18] B.F Finkenstädt, O.N. Bjørnstad, and B.T. Grenfell. A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. Biostatistics, 3:493–510, 2002.

[19] N. Fournier and S. Méléard. A microscopic probabilistic description of a locally regulated population and macroscopic approximations. Ann. Appl. Probab., 14(4):1880–1919, 2004.

[20] C. A. Gilligan G. J. Gibson, A. Kleczkowski. Bayesian analysis of botanical epidemics using stochastic compartmental models. Proceedings of the national academy of sciences, 101:12120–12124, 2004.

[21] A. Gelman and X-L Meng. Model checking and model improvment. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, Markov chain Monte Carlo in practice. Chapman & Hall, 1996.

[22] G.J. Gibson. Markov chain Monte-Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. Applied Statistics, 46:215–233, 1997.

[23] G.J. Gibson and E. Renshaw. Likelihood estimation for stochastic compartmental models using Markov chain methods. Statistics and Computing, 11:347–358, 2001.

[24] W.R. Gilks and G.O. Roberts. Strategies for improving MCMC. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors, Markov chain Monte Carlo in practice. Chapman & Hall, 1996.

[25] Y.H. Hsieh, H. de Arazoza, S.M. Lee, and C.W. Chen. Estimating the number of Cubans infected sexually by human immunodeficiency virus using contact tracing data. Int. J. Epidemiol., 31(3):679–83, 2002.

[26] V. Isham. Stochastic models for epidemics with special reference to AIDS. The Annals of Applied Probability, 3(1):1–27, 1993.

[27] W.O. Kermack and A.G. McKendrick. Contributions to the mathematical theory of epidemics. Bulletin of Mathematical Biology, Springer New York, 53(1-2):33–55, 1991.

[28] Y. Kuang. Delay Differential Equations with Applications to Population Dynamics, volume 151 of Lecture Notes in Statistics. Springer, New York, 2000.

[29] D. Mollison. Epidemic models: their structure and relation to data. Cambridge University Press, 3 edition, 1995. Chap. 2: The structure of Epidemic Models.

[30] E.A. Nadaraya. On estimating regression. Theory of Probability and its Applications, 10:186–190, 1964.

[31] P.D. O'Neill. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markoc chain Monte Carlo methods. Mathematical Biosciences, 180:103–114, 2002.

[32] P.D. O'Neill and G.O. Roberts. Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A, 162:121–129, 1999.

[33] J.K. Pritchard, M.T. Seielstad, A. Perez-Lezaun, and M.W. Feldman. Population growth of human Y chromosomes: a study of y chromosome microsatellites. Molecular Biology and Evolution, 16:1791–1798, 1999.

[34] B.D. Ripley. Pattern recognition and neural networks. Cambridge University Press, Cambridge, 1996.

[35] B.W. Silverman. Discussion of the paper by Diggle and Gratton. Journal of the Royal Statistical Society: Series B, 46:212–214, 1984.

[36] S.A. Sisson, Y. Fan, and M. Tanaka. Sequential Monte Carlo without likelihoods. Proc. Nat. Acad. Sci. USA, 104:1760–1765, 2007.

[37] G. Streftaris and G.J. Gibson. Bayesian analysis of experimental epidemics of foot-and-mouth disease. Journal of the Royal Society: Series B, 271:1111–1117, 2004.

[38] G. Streftaris and G.J. Gibson. Bayesian inference for stochastic epidemics in closed population. Statistical Modelling, 4(1):63–75, 2004.

[39] S. Tavaré, D.J. Balding, R.C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. Genetics, 145:505–518, 1997.

[40] V.C. Tran. Modèles particulaires stochastiques pour des problèmes d'évolution adaptative et pour l'approximation de solutions statistiques. PhD thesis, Université Paris X - Nanterre. `http://tel.archives-ouvertes.fr/tel-00125100`.

[41] P. van den Driessche. Some epidemiological models with delays. In Differential Equations and Applications to Industry (Claremont, CA, 1994), pages 507–520, River Edge, NJ, 1996. World Sci. Publishing.

[42] P. van den Driessche. Time delay in epidemic models. In C. Castillo-Chavez, S. Blower, P. van den Driessche, D. Kirschner, and A.A. Yakubu, editors, Mathematical Approaches for Emerging and Reemerging Infectious Diseases: An Introduction, volume 125, pages 119–128, New York, 2002. IMA Volumes in Mathematics and its Applications, Springer.

[43] G.S. Watson. Smooth regression analysis. Shankya Series A, 26:359–372, 1964.

7 Appendix

Proof of Proposition 3.1. Let us denote by ν the common distribution of the ε_i's. Then the distribution of θ_i conditional on s_i is $\pi(d\theta \,|\, s_i) = \nu \circ F_{s_i}^{-1}$ the image distribution of ν by F_{s_i}. The aproximate posterior distribution can be expressed as

$$\pi(d\theta \,|\, s) = \nu \circ F_s^{-1} = \nu \circ F_{s_i}^{-1} \circ F_{s_i} \circ F_{s_i}^{-1} = \pi(d\theta \,|\, s_i) \circ \left(F_s \circ F_{s_i}^{-1}\right)^{-1}$$

which is the image distribution of $\pi(d\theta \,|\, s_i)$ by $F_s \circ F_{s_i}^{-1}$. The latter function is a *coupling* of $\pi(d\theta \,|\, s)$ and $\pi(d\theta \,|\, s_i)$. ■

Figure legends

Figure 1. Schematic description of the SIR model with contact tracing.

Figure 2. Boxplots of the $M = 200$ estimated modes and quantiles (2.5%, 50%, and 97.5%) of the approximate posterior distributions of λ_1. For each ABC method and each value of the tolerance rate, 200 posterior distributions are computed for each of the 200 synthetic data sets. The horizontal lines correspond to the true values $\lambda_1 = 1.14 \times 10^{-7}$ used when simulating the 200 synthetic data sets. The different tolerance rates are 0.01, 0.05, 0.10, 0.25, 0.50, 0.50, 0.75, and 1 for all the ABC methods except the rejection scheme with the two summary statistics. For the latter method, the tolerance rates are 0.007, 0.02, 0.06, 0.13, 0.27, 0.37, 0.45, 0.53, 0.66, 0.80, 1.

Figure 3. Rescaled mean squared error (RMSE) of the mode of the approximate posterior distributions when estimating the three parameters λ_1, λ_2, and λ_3. The RMSE's are plotted as a function of the tolerance rate.

Figure 4. Rescaled mean credibility interval (RMCI) of the approximate posterior distributions when estimating the three parameters λ_1, λ_2, and λ_3. The RMCI's are plotted as a function of the tolerance rate.

Figure 5. Bayesian posterior predictive distributions of $R^1_{21.5}$, $R^2_{21.5}$, I_6, and the mean sojourn time in the class I. The SIR model corresponds to the model with a mass action principle. The vertical lines correspond to the observations. The approximate posterior samples were obtained with the ABC-NCH method ($P_\delta = 0.5\%$) by making use of the 61 summary statistics (see text).

Figure 6. Bayesian posterior predictive distributions of $R^1_{21.5}$, $R^2_{21.5}$, I_6, and the mean sojourn time in the class I. The SIR model corresponds to the model with frequency dependence for contact tracing detection. The approximate posterior samples are obtained with the smooth rejection ABC algorithm by making use of the 2 infinite-dimensional summary statistics R^1 and R^2. A tolerance rate of $P_\delta = 1\%$ is considered for each summary statistic.

Figure 7. Median and 95% credibility intervals of the posterior predictive distributions of R^1_t, R^2_t, $R^1_t/(R^1_t + R^2_t)$, and I_t from $t = 0$ (1986) to $T = 29$ (2015). The posterior samples are generated by the rejection scheme with the two summary statistics. A tolerance rate of $P_\delta = 1\%$ is considered for each summary statistic.

Figure 8. Median and 95% credibility intervals of the posterior predictive distribution of the coverage of the epidemic from $t = 0$ (1986) to $T = 29$ (2015). The coverage is defined as the proportion of known HIV positive individuals. The posterior samples are generated by the rejection scheme with the two summary statistics R^1 and R^2. A tolerance rate of $P_\delta = 1\%$ is considered for each summary statistic.

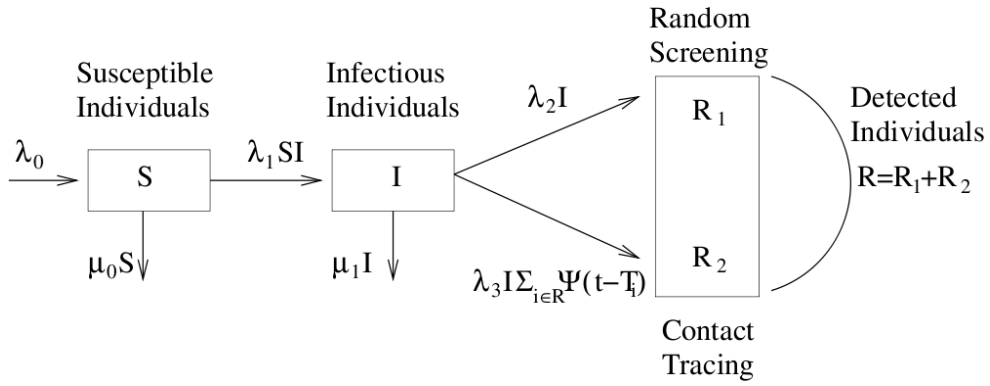
Random
Screening
Susceptible
Individuals
Infectious
Individuals
λ2I
R 1
Detected
Individuals
λ0
S
λ1SI
I
R=R1+R2
μ0S
μ1I
λ3IΣi∈RΨ(t−Ti)
R 2
Contact
Tracing

Figure 1:

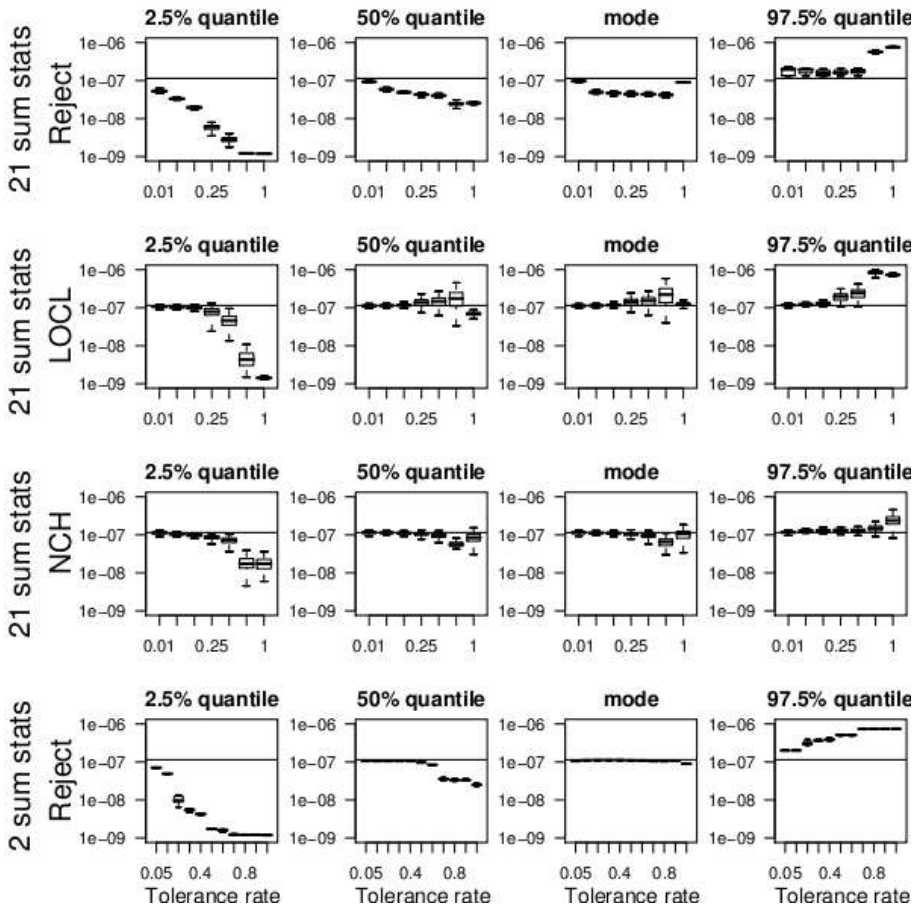
2.5% quantile
50% quantile
mode
97.5% quantile
21 sum stats
Reject
LOCL
NCH
2 sum stats
1e−06
1e−07
1e−08
1e−09
0.01
0.25
1
0.05
0.4
0.8
Tolerance rate

Figure 2:

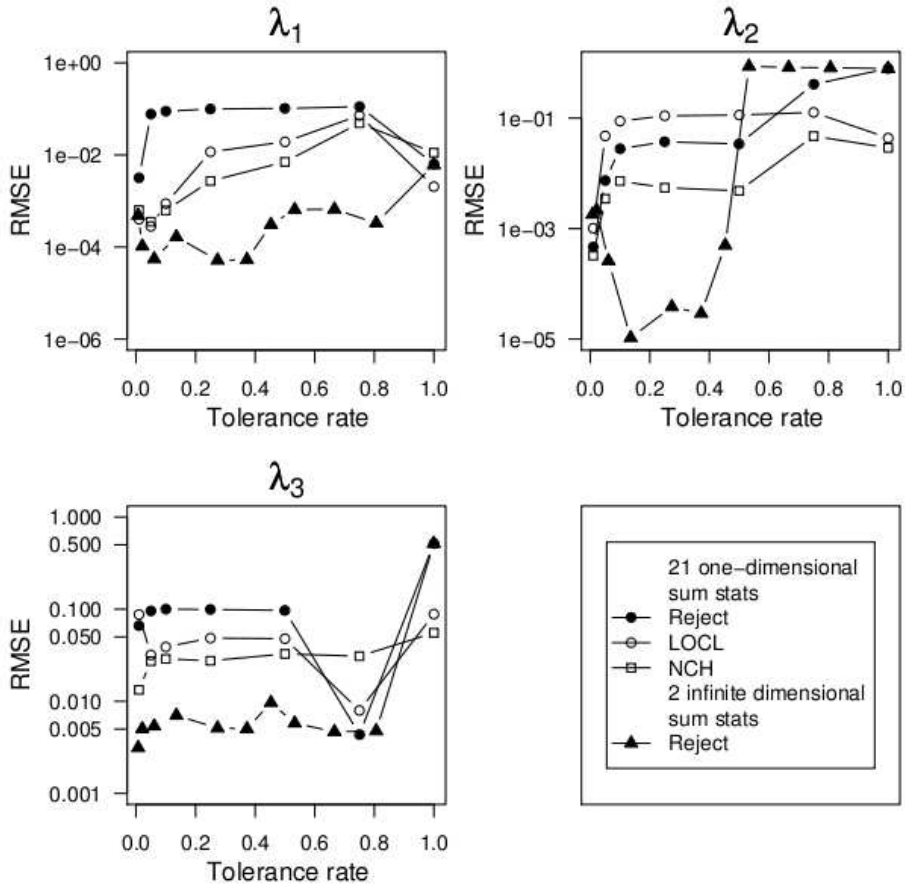
λ1
λ2
λ3
RMSE
Tolerance rate
1e+00
1e−02
1e−04
1e−06
1e−01
1e−03
1e−05
1.000
0.500
0.100
0.050
0.010
0.005
0.001
0.0
0.2
0.4
0.6
0.8
1.0
21 one−dimensional sum stats
Reject
LOCL
NCH
2 infinite dimensional sum stats
Reject

Figure 3:

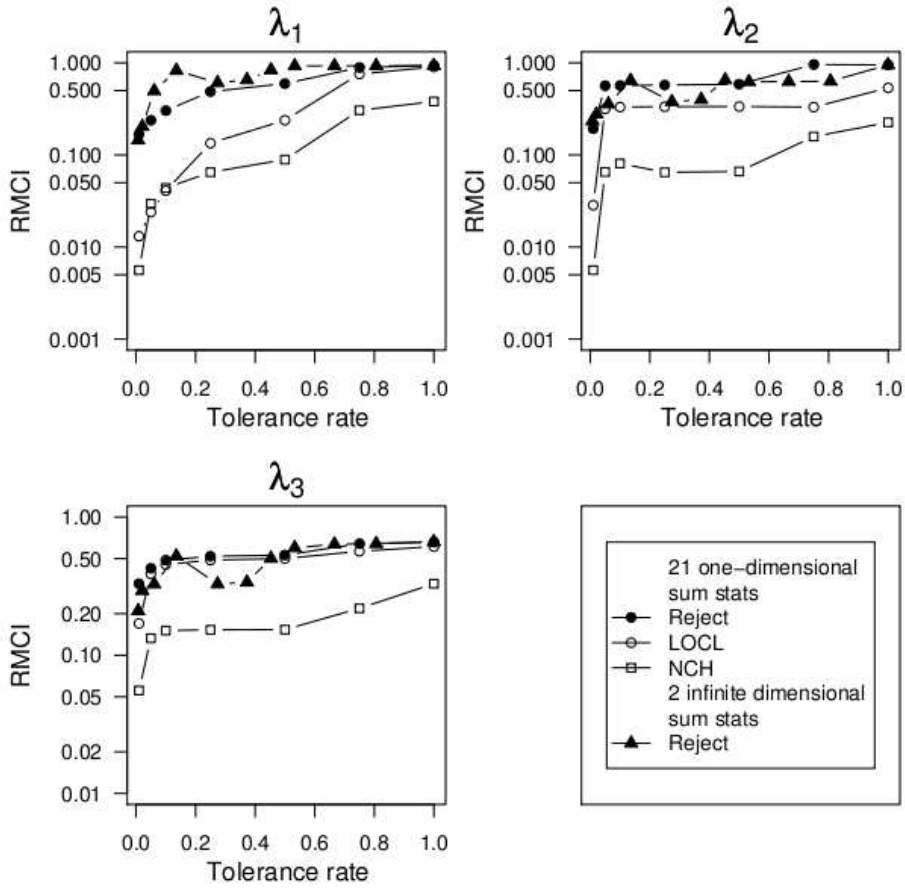
λ1
λ2
λ3
RMCI
Tolerance rate
1.000
0.500
0.100
0.050
0.010
0.005
0.001
1.00
0.50
0.20
0.10
0.05
0.02
0.01
0.0
0.2
0.4
0.6
0.8
1.0
21 one−dimensional sum stats
Reject
LOCL
NCH
2 infinite dimensional sum stats
Reject

Figure 4:

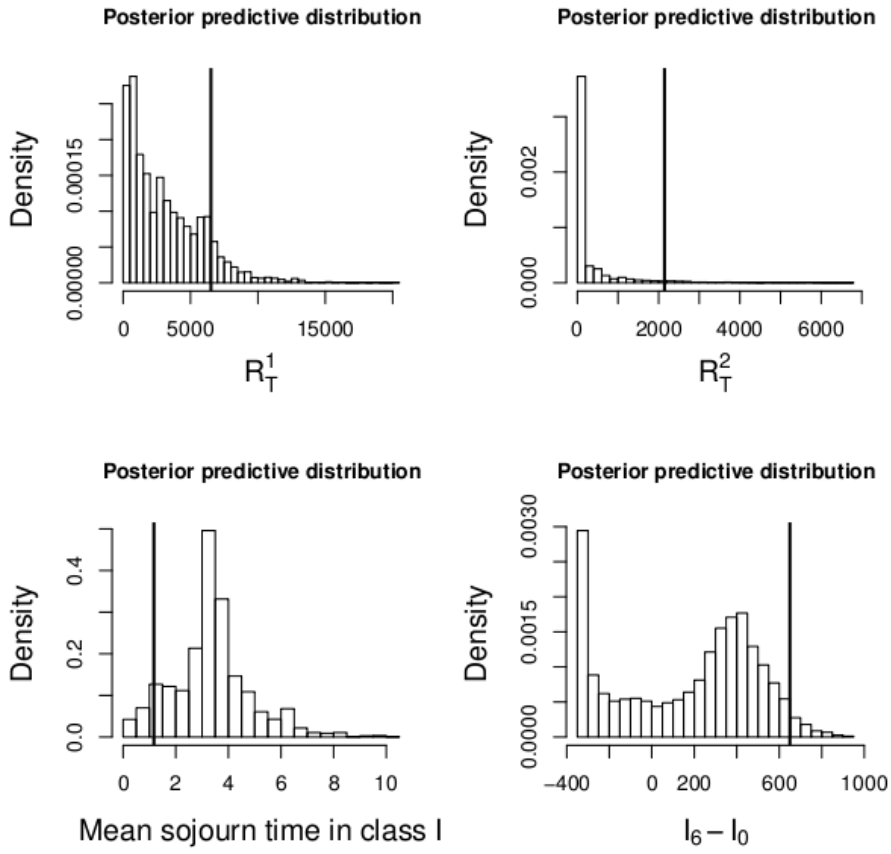
Posterior predictive distribution
Density
0.00000
0.00015
0
5000
15000
R_T^1
Posterior predictive distribution
Density
0.000
0.002
0
2000
4000
6000
R_T^2
Posterior predictive distribution
Density
0.0
0.2
0.4
0
2
4
6
8
10
Mean sojourn time in class I
Posterior predictive distribution
Density
0.0000
0.0015
0.0030
−400
0
200
600
1000
$I_6 - I_0$

Figure 5:

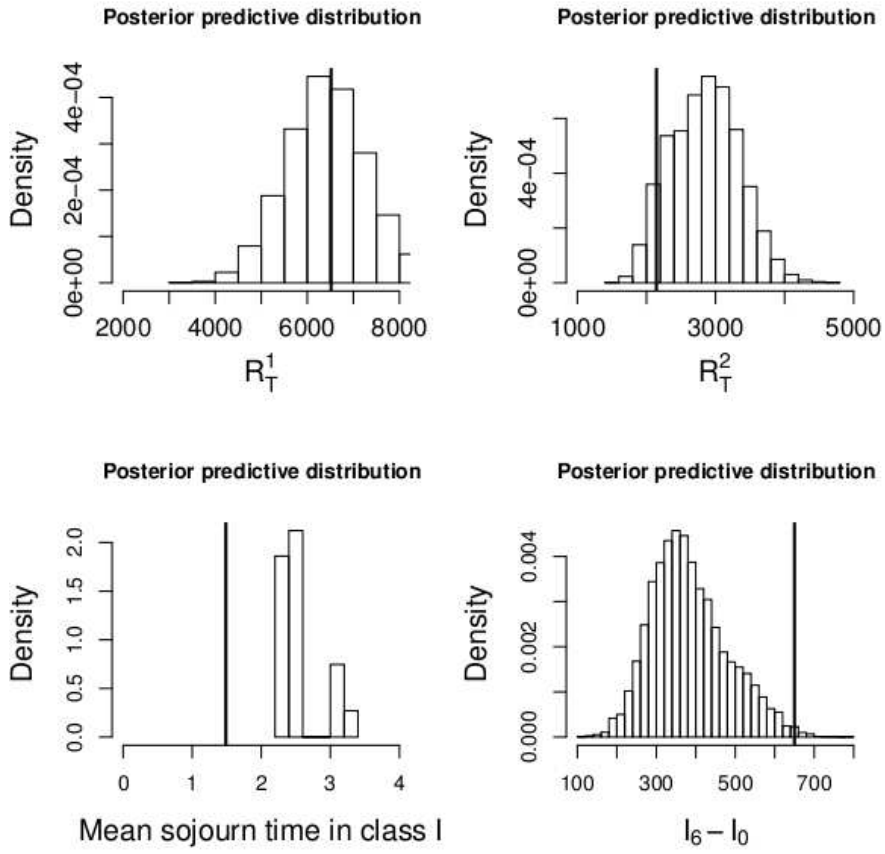
Posterior predictive distribution
Density
0e+00
2e−04
4e−04
2000
4000
6000
8000
R_T^1
Posterior predictive distribution
Density
0e+00
4e−04
1000
3000
5000
R_T^2
Posterior predictive distribution
Density
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
Mean sojourn time in class I
Posterior predictive distribution
Density
0.000
0.002
0.004
100
300
500
700
$l_6 - l_0$

Figure 6:

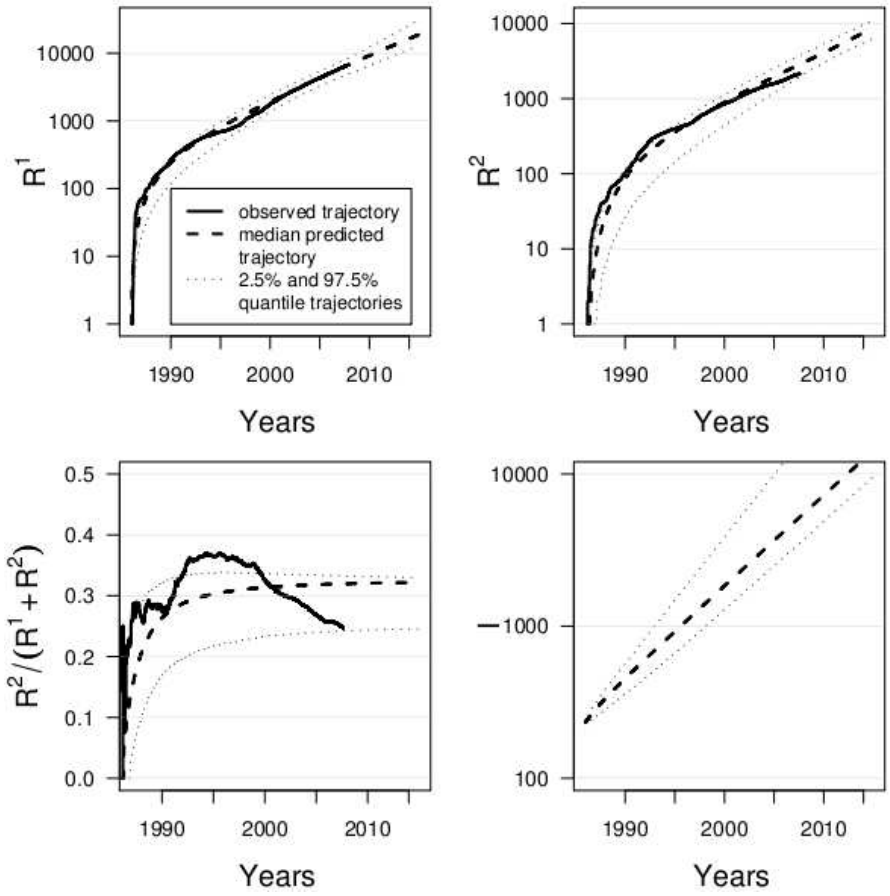
observed trajectory
median predicted trajectory
2.5% and 97.5% quantile trajectories
R1
R2
R2/(R1+R2)
Years
1990
2000
2010
10000
1000
100
10
1
0.5
0.4
0.3
0.2
0.1
0.0

Figure 7:

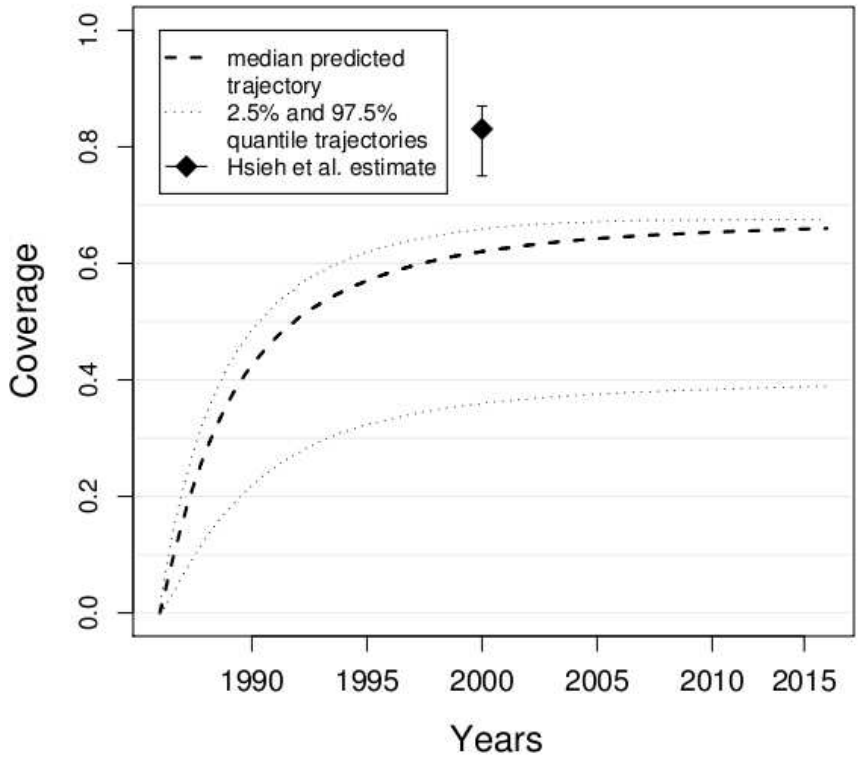
median predicted trajectory
2.5% and 97.5% quantile trajectories
Hsieh et al. estimate
Coverage
0.0
0.2
0.4
0.6
0.8
1.0
1990
1995
2000
2005
2010
2015
Years

Figure 8: