A Bayesian Framework for Parameter Estimation in Dynamical Models

Flávio Codeço Coelho[1,2]*, **Cláudia Torres Codeço**[3], **M. Gabriela M. Gomes**[1]

[1] Instituto Gulbenkian de Ciência, Oeiras, Portugal, [2] Escola de Matemática Aplicada, Fundação Getúlio Vargas, Rio de Janeiro, Brazil, [3] Programa de Computação Científica, Fundação Oswaldo Cruz, Rio de Janeiro, Rio de Janeiro, Brazil

Abstract

Mathematical models in biology are powerful tools for the study and exploration of complex dynamics. Nevertheless, bringing theoretical results to an agreement with experimental observations involves acknowledging a great deal of uncertainty intrinsic to our theoretical representation of a real system. Proper handling of such uncertainties is key to the successful usage of models to predict experimental or field observations. This problem has been addressed over the years by many tools for model calibration and parameter estimation. In this article we present a general framework for uncertainty analysis and parameter estimation that is designed to handle uncertainties associated with the modeling of dynamic biological systems while remaining agnostic as to the type of model used. We apply the framework to fit an SIR-like influenza transmission model to 7 years of incidence data in three European countries: Belgium, the Netherlands and Portugal.

Citation: Coelho FC, Codeço CT, Gomes MGM (2011) A Bayesian Framework for Parameter Estimation in Dynamical Models. PLoS ONE 6(5): e19616. doi:10.1371/journal.pone.0019616

Editor: Andrew Yates, Albert Einstein College of Medicine, United States of America

Received November 9, 2010; **Accepted** April 12, 2011; **Published** May 24, 2011

Funding: This research was funded by Fundação para a Ciência e a Tecnologia (FCT), the European Commission (grants MEXT-CT-2004-14338 and EC-ICT-231807), and the Brazilian Research Council (CNPq). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: fccoelho@fgv.br

Introduction

Mathematical models have long played a key role in understanding infectious disease epidemiology [1] as well as other biological dynamical systems. Their ability to combine established theory and data to predict empirical observation is unique and cannot be easily achieved by other methods [2]. In such models, data in the form of rate parameters and time-series, and theory in the form of the model formulation, interact to provide insight about each other. Parameter estimation and model selection techniques allow us to improve theory with the help of data (model selection) and estimate data which cannot be directly observed, with the help of theory (parameter estimation).

Proper representation of the intrinsic uncertainty associated with dynamic models of biological systems has been under increasing scrutiny through the development of a number of methods for parameter estimation and model calibration [3–10]. Such methods, to be effective, must strive to be as comprehensible as possible in the treatment of all identifiable sources of uncertainty related to a given mathematical representation of a biological system [5]. In practice, however, many uncertainty analysis methods fall short of this ideal. Some of the work in the recent literature focus on developing exact methods for parameter estimation, requiring, for instance, the derivation of the full likelihood function for the model at hand. Exact methods, however, tend to be closely coupled to a specific model or class of models, being less generally applicable [11–14].

In this paper we introduce a Bayesian framework for parameter estimation in dynamic models that is applicable to both deterministic and stochastic models [15]. The framework extends similar frameworks proposed for different types of models [4,6,16,17] and focuses of the analysis of dynamic models where full or partial time-series data are available for the model to be fit against. The fitting process estimates the posterior probability distributions for both the model's parameters and output series.

To ensure generality, the dynamic model, from the point of view of the inference machinery, is treated as a "black box" with inputs (parameters) and outputs (time-series), and the full uncertainty about each of these elements can be included in the form of prior distributions which will get updated based on observational data. Model comparison and selection analyses are facilitated by the pluggable nature of the model in the framework.

To illustrate the use of this framework, seven-years long time-series of influenza-like illness incidence data from Belgium, Netherlands and Portugal [18] were used to as a basis for parameter estimation of a deterministic influenza transmission model.

Methods

The core of the analytical framework proposed was inspired on the Bayesian Melding method [6] with modifications to make it work with dynamic models, that is, with time-series as model outputs. The Bayesian Melding method pioneered in providing a formal inferential framework that took into full account information available about a model's inputs and outputs. We proceed to give a brief description of the Melding method. For a complete description, see the original work. Let $\Theta = \{p_1, p_2, \ldots, p_n\}$ be the set of n parameters which are the inputs to the model M. The p_i are random variables with a joint probability prior distribution

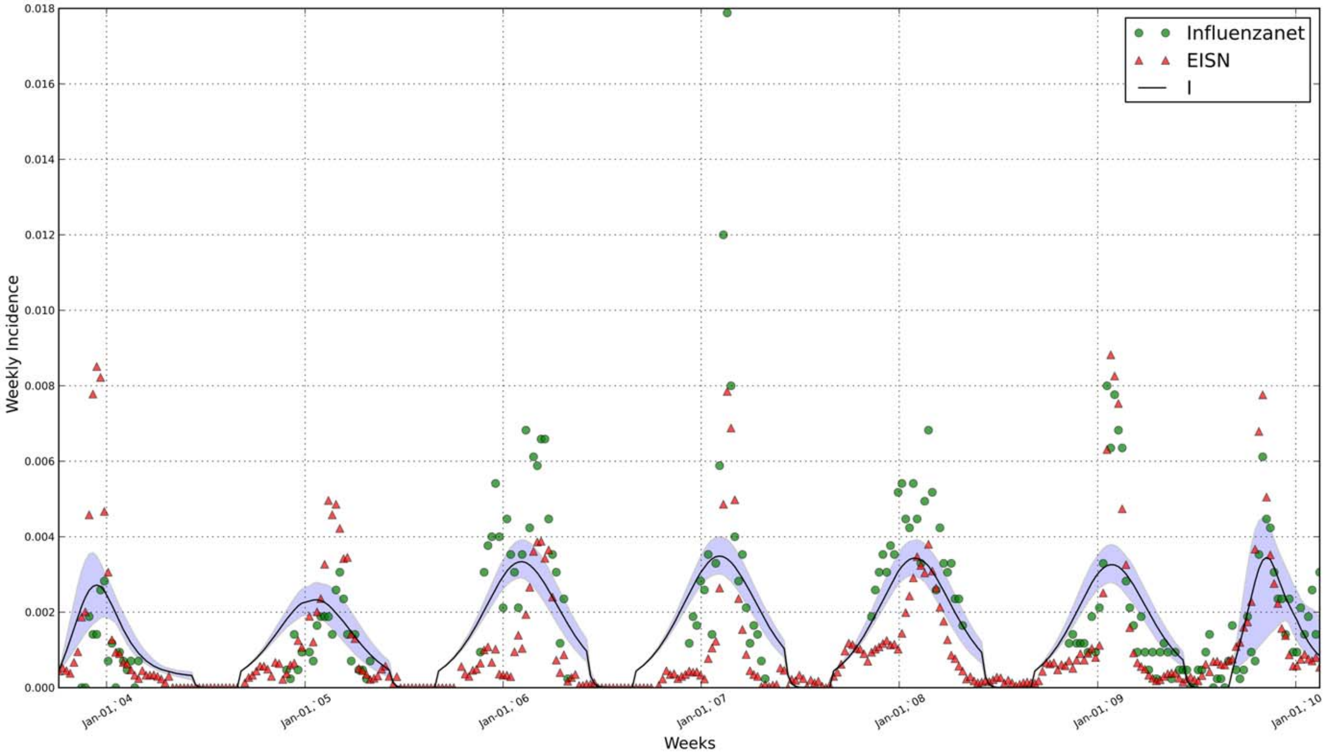

Figure 1. Belgian incidence data and model fit. Incidence median curve (black line) and 95% credible intervals (shaded area) for the model-generated incidence series. The model was fitted simultaneously to Influenzanet data (green circles) and EISN data (red triangles).
doi:10.1371/journal.pone.0019616.g001

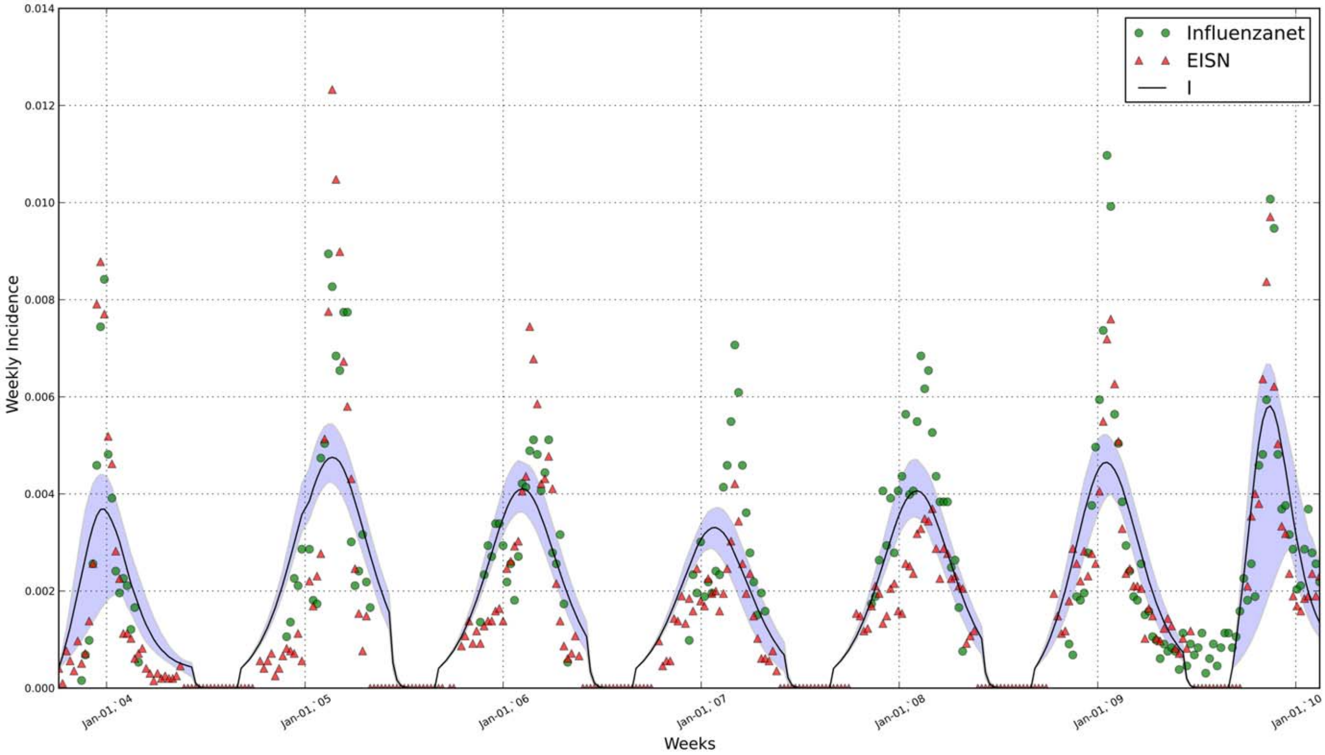

Figure 2. Incidence data from the netherlands and model fit. Incidence median curve (black line) and 95% credible intervals (shaded area) for the model-generated incidence series. The model was fitted simultaneously to Influenzanet data (green circles) and EISN data (red triangles).
doi:10.1371/journal.pone.0019616.g002

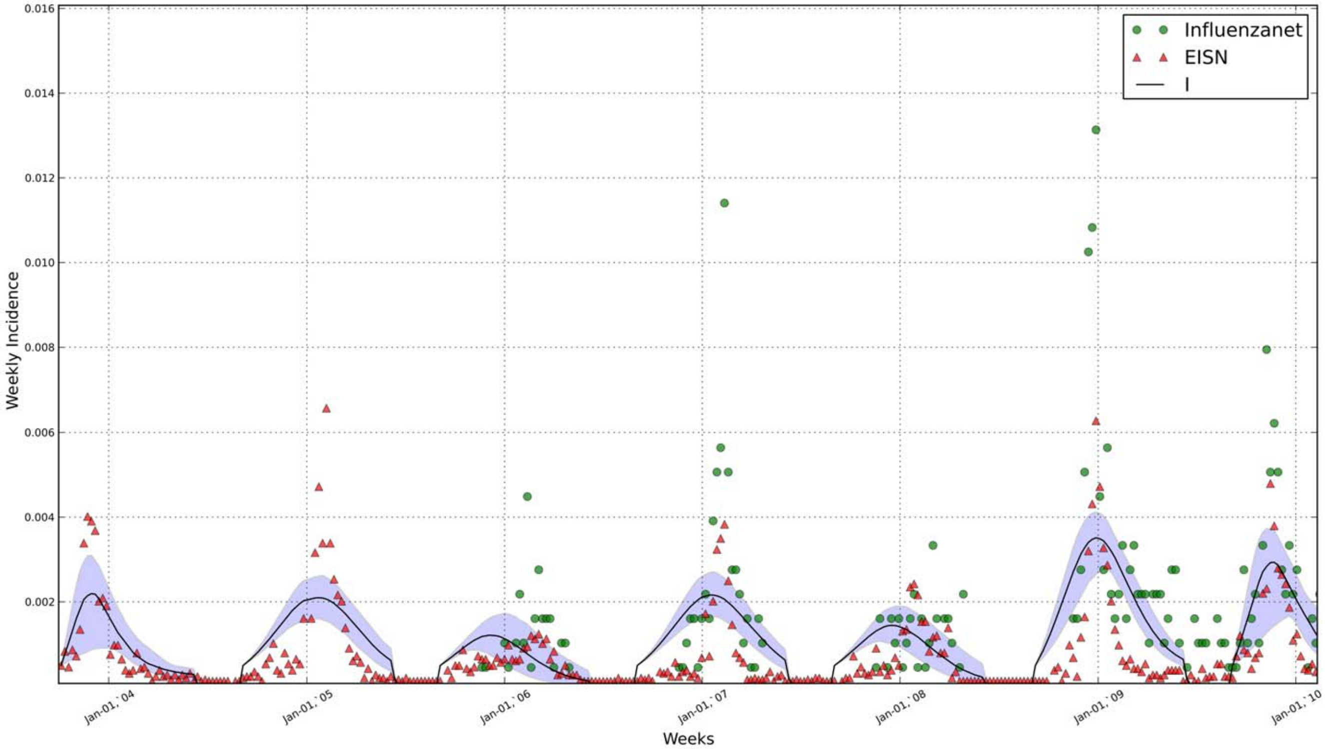

Figure 3. Portuguese incidence data and model fit. Incidence median curve (black line) and 95% credible intervals (shaded area) for the model-generated incidence series. The model was fitted simultaneously to Influenzanet data (green circles) and EISN data (red triangles).
doi:10.1371/journal.pone.0019616.g003

denoted by $q(\Theta)$. Therefore, $P(\Theta=\theta)\sim q(\Theta)$. Also let Φ be the set of m outputs of M, $\Phi=\{v_1(t),v_2(t),\ldots,v_m(t)\}$.

Since Φ is a function of Θ, the prior distribution of Θ, $q(\Theta)$ induces a prior probability on Φ, $q(\Phi)$:

$$\Phi=M(\Theta) \tag{1}$$

$$q(\Phi)=M(q(\Theta))$$

Let θ and ϕ be realizations of the model's inputs and outputs, respectively, such that $\phi=M(\Theta=\theta)$. The inferential problem consists in finding the joint posterior probability distribution of Θ, $\pi(\Theta)$, and that of Φ, $\pi(\Phi)$, given existing data (D). Data will enter the inference in the form of time-series corresponding to the models outputs. Data on the model's parameters can also be used to update Θ's joint prior probability distribution. The observed data used to fit the model may refer to only a subset of the model's outputs (Φ). The likelihood of the model's outputs is given by:

$$L(\Phi)=P(\mathfrak{D}|\Phi)=P(\mathfrak{D}|M(\Theta))=L(\Theta) \tag{2}$$

From equation 2, we see that data on the outputs will inform the likelihood of both Φ and Θ as they are connected by the model. In practice this means that the most likely sets of parameters (θ) will be the ones which generated the most likely outputs (ϕ). The dependency of the outputs on inputs is given by the model so the accuracy of the inference will depend of the model's identifiability, i.e. different θ generate different ϕ.

The posterior of Θ is updated according to equation 3.

$$\pi(\Theta)\propto q(\Theta)L(\Theta) \tag{3}$$

As already mentioned, this work introduces some extensions to the original Melding method. A couple of extensions stand out. One of them is the ability to use time-series data, the Bayesian Melding method made inferences based on data on single point in time. The second was the use of a multi-chain Markov-chain sampler to more efficiently tackle non-convex higher dimensional parameter-spaces.

Prior Information

Before starting the inference, prior probability distributions for the parameters in Θ, $q(\Theta)$, must be defined. The initial conditions for the model can be fixed or included as members of Θ. If prior information about the distribution of the outputs is available, it can be pooled with the induced prior on the outputs as described by Poole and Raftery [6]. In the particular application described below, we have used uninformative priors – $U(0,1)$ – for the outputs of the models since we had no expectations about them which could inform different prior distributions.

Likelihood Calculations

The exploration of the parameter space is done by Markov Chain Monte Carlo, as described below until K samples are accepted. For the application presented here, the error distribution of ϕ_i, where $i\in\{1,\ldots,K\}$, is assumed to be Normal, $N(\mu=\phi,\sigma^2)$. Thus $L(\phi_i)$ is a Normal likelihood function with fixed variance σ^2. Other parametric forms for the likelihood function can be

Table 1. Model Parameters; posterior estimates.

Name	**Belgium**	**Netherlands**	**Portugal**
	μ (95% **interval)**	μ (95% **interval)**	μ (95% **interval)**
$S_{0,2004}$	0.246 (0.202, 0.49)	0.337 (0.245, 0.5)	0.215 (0.126, 0.498)
$S_{0,2005}$	0.434 (0.302, 0.562)	0.805 (0.454, 0.93)	0.493 (0.363, 0.639)
$S_{0,2006}$	0.644 (0.453, 0.766)	0.685 (0.411, 0.815)	0.265 (0.122, 0.5)
$S_{0,2007}$	0.669 (0.423, 0.77)	0.543 (0.346, 0.657)	0.519 (0.374, 0.66)
$S_{0,2008}$	0.645 (0.404, 0.775)	0.67 (0.385, 0.789)	0.316 (0.144, 0.5)
$S_{0,2009}$	0.588 (0.416, 0.699)	0.664 (0.435, 0.764)	0.577 (0.43, 0.707)
$S_{0,2010}$	0.299 (0.205, 0.523)	0.43 (0.326, 0.592)	0.336 (0.222, 0.609)
α_{2004}	0.0186 (0.00148, 0.0901)	0.0776 (0.0032, 0.332)	0.152 (0.00227, 0.481)
α_{2005}	0.258 (0.0768, 0.394)	0.279 (0.12, 0.395)	0.236 (0.0689, 0.396)
α_{2006}	0.306 (0.0972, 0.444)	0.271 (0.112, 0.393)	0.245 (0.0409, 0.396)
α_{2007}	0.278 (0.116, 0.395)	0.301 (0.108, 0.398)	0.263 (0.0893, 0.391)
α_{2008}	0.228 (0.0447, 0.387)	0.258 (0.0909, 0.39)	0.249 (0.0698, 0.392)
α_{2009}	0.152 (0.0426, 0.289)	0.088 (0.012, 0.278)	0.0591 (0.0136, 0.259)
α_{2010}	0.107 (0.000647, 0.484)	0.0647 (0.00127, 0.424)	0.0929 (0.00248, 0.46)
R_e	1.1(1.09, 1.16)	1.11, (1.1, 1.18)	1.08, (1.06, 1.15)
m	1.78E-06 (1.35E-07, 2.95E-06)	1.98E-06 (1.05E-07, 2.97E-06)	2.84E-06 (8.98E-07, 3.92E-06)
τ	1.4	1.4	1.4

Parameters of the SIR model. Single numbers are values of fixed parameters. The rest are posterior means and their 95% band. $S_{0,*}$ are the initial fraction of susceptibles at each year; α_* are the fraction of symptomatics for each year; r_e is the effective reproductive number at the beginning of the season; m is the infectious immigration constant; τ is the recovery rate.

doi:10.1371/journal.pone.0019616.t001

adopted. Parameters values (θ) are retained with probability proportional to the likelihood of $M(\theta)$, as given by:

$$L(\phi) = \prod_{t=1}^{T} P(\mathfrak{D}[t]|\phi[t]) \tag{4}$$

Monte Carlo Simulations

A multi-chain differential evolution adaptive metropolis algorithm (DREAM) [19] was used to sample the joint posterior probability distribution of Θ, $\pi(\Theta)$. DREAM is a sophisticated algorithm where multiple adaptive chains are run in parallel with delayed rejection.

For the application presented, 16 chains (same as the dimensionality of the parameter space) were started from 16 randomly chosen points in parameter space and moved around with steps given by a gaussian proposal distribution centered at its current position with covariance being adapted every ten steps as described by Andrieu and Thoms [20]. Proposed θ_i are accepted proportionally to their posterior probability. The chains are run until the desired number of samples is reached after discarding a pre-determined number of burn-in samples. Convergence of the parallel chains was verified at every 100 iterations by the calculation of the Gelman-Rubins' R convergence diagnostic [21].

Application to Multi-Season Influenza Transmission

We used a deterministic model for influenza transmission, adapted from the Susceptible-Infected-Recovered (SIR) framework [1], to explain multi-season dynamics of influenza in Europe. The model was fitted to two sets of influenza-like illness incidence times-series (Influenzanet [18] and EISN [22]) collected between from 2004 and 2010 in Belgium, Netherlands and Portugal. The model differs from the standard SIR in that only a fraction, α, of the infected individuals is symptomatic and infectious, the remaining being asymptomatic and ineffective in passing on the virus. A small infectious immigration rate (m) is also added. The model is implemented as a set of ordinary differential equations:

$$\lambda = \beta(\alpha I + m)$$

$$\frac{dS}{dt} = -\lambda S$$

$$\frac{dI}{dt} = \lambda S - \tau I$$

$$\frac{dR}{dt} = \tau I$$

where the recovery rate (τ) is such that the infectious period last 5 days [23], and the migration parameter (m) is assumed to be proportional to the number of susceptibles, considering that infection is imported by susceptible individuals who acquire the virus while traveling abroad.

To model the seasonality of influenza epidemics in Europe, the transmission rate β is assumed to drop during the three summer months (June, July and August), thus virtually interrupting transmission of the disease, possibly due to school closure for summer vacations. For the rest of the year β is assumed to be large enough to allow for sustained transmission. During this period the effective reproduction number, R_e, is given by the expression:

$$R_e = \frac{\beta\alpha}{\tau} S_0, \quad (5)$$

where S_0 is the number of susceptibles at the beginning of each transmission season.

The model is parameterized in such a way that total population is normalized to 1 and S, I, and R are fractions of the total population. The initial fraction of susceptibles, S_0, was estimated along with other parameters of the model for each year while the initial fraction infected was set to match the prevalence of the first week of data. The remainder of the population was placed in the R compartment. The symptomatic fraction of I, denoted by α, was also estimated for each year. The output of the model, as represented by $\alpha * I(t)$ was fitted against the data.

For each country, we have estimated S_0 and α as season specific parameters, while R_e and m where fixed across the multiple seasons. From these 16 estimated quantities, β_h can be calculated by manipulating expression 5 if desired.

The model was fitted to the three countries' datasets. Uniform priors were attributed to all parameters: S_0 had $U(0,1)$ priors for all years; α had $U(0,0.4)$ priors for all years; R_e had $U(1,1.4)$ and m, $U(0,4e-6)$. The posterior distribution for parameters and series were obtained from 2000 samples generated by the DREAM algorithm after 2000 burn-in samples were discarded.

Results and Discussion

Figures 1, 2 and 3 show the fit of the model against data from both Influenzanet and EISN for the three countries. The model was able attain a good fit to the data, allowing for reasonably precise estimate of the parameters (table 1). We have performed some consistency checks on the estimates obtained (not shown). In particular we have found a positive correlation between the fraction of infections that are symptomatic in a given season (α) and the time of the epidemic peak (measured from September 1st), suggesting a role of weather factors in the performance of influenza surveillance systems, which is further explored in van Noort *et al.* [24] by combining data from other sources. Although here we chose the simplest model formulation for the purpose of illustration of the parameter estimation method, the results are compatible with other studies. Moreover, the procedure is readily applicable to more elaborate models.

The estimates of the basic reproductive number (R_0) for each season and country, can be obtained by dividing the R_e estimated for each country by the S_0 estimated for each year (table 1). Its values range from $1.64-4.58$ for Belgium, $1.38-3.26$ for the Netherlands, and $1.86-5.14$ for Portugal. These values, are in accordance to previously reported estimates of R_0 for influenza [25–27].

This work proposes a methodological framework to perform parameter estimation in dynamical models where time series data is available for the model to be fit against. The method described can be applied to a wide range of dynamical models, taking its utility beyond the application described in this paper. Currently, its applicability is limited in practice by the robustness of the MCMC samplers available in handling complex high-dimensional parametric spaces. This limitation can be reduced in the future by the development of more powerful posterior sampling methods.

The pluggable nature of the model, in the framework, allows for a simple way to compare multiple models and select which one fits best the available data. Goodness of fit statistics such as AIC [28], BIC [29] or DIC [30], provided by the framework, can be used for this. Model comparison and selection techniques are, however, not discussed in this paper but can be found in the literature [31].

For this work, an open-source software library [32] was developed which allows for the immediate application of the framework proposed here to other models by means of a simple Python script (as decribed in the library's documentation). The library can also be used from within a Sage worksheet [33], requiring little programming knowledge.

Acknowledgments

The authors acknowledge the helpful comments from the reviewers.

Author Contributions

Conceived and designed the experiments: FCC CTC MGMG. Performed the experiments: FCC. Analyzed the data: FCC CTC MGMG. Contributed reagents/materials/analysis tools: FCC MGMG. Wrote the paper: FCC CTC MGMG.

References

1. Anderson RM, May RM (1979) Population biology of infectious diseases: Part i. Nature 280: 361–367.
2. Ness RB, Koopman JS, Roberts MS (2007) Causal system modeling in chronic disease epidemiology: a proposal. Annals of Epidemiology 17: 564–8.
3. BretóC, He D, Ionides E, King A (2009) Time series analysis via mechanistic models. Annals of Applied Statistics 3: 319–348.
4. Alkema L, Raftery AE, Brown T (2008) Bayesian melding for estimating uncertainty in national HIV prevalence estimates. Sex Transm Infect 84: i11–16.
5. Coelho FC, Codeço CT, Struchiner CJ (2008) Complete treatment of uncertainties in a model for dengue r_0 estimation. Cadernos De Saúde Pública/Ministério Da Saúde, Fundação Oswaldo Cruz, Escola Nacional De Saúde Pública 24: 853–61.
6. Poole D, Raftery AE (2000) Inference for deterministic simulation models: The bayesian melding approach. Journal of the American Statistical Association 95: 1244–1255.
7. Bettencourt LMA, Ribeiro RM (2008) Real time bayesian estimation of the epidemic potential of emerging infectious diseases. PLoS ONE 3: e2185.
8. Calderhead B, Girolami M, Lawrence ND (2008) Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In: Neural Information Processing Systems. volume 22.
9. Girolami M (2008) Bayesian inference for differential equations. Theoretical Computer Science 408: 4–16.
10. Ramsay JO, Hooker G, Campbell D, Cao J (2007) Parameter estimation for differential equations: a generalized smoothing approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69: 741–796.
11. Vyshemirsky V, Girolami M (2008) BioBayes: a software package for bayesian inference in systems biology. Bioinformatics 24: 1933–1934.
12. Golightly A, Wilkinson DJ (2006) Bayesian sequential inference for stochastic kinetic biochemical network models. Journal of Computational Biology 13: 838851.
13. Golightly A, Wilkinson DJ (2008) Bayesian inference for nonlinear multivariate diffusion models observed with error. Computational Statistics and Data Analysis 52: 16741693.
14. Lecca P, Palmisano A, Ihekwaba A, Priami C (2009) Calibration of dynamic models of biological systems with KInfer. European Biophysics Journal 39: 1019–1039.
15. Coelho FC, Codeço CT (2009) A bayesian framework for parameter estimation in dynamical models with applications to forecasting. URL precedings.nature.com/documents/4044/version/1.
16. Ionides EL, Bret C, King AA (2006) Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences of the United States of America 103: 18438–43.
17. Ševčiková H, Raftery AE, Waddell PA (2007) Assessing uncertainty in urban simulations using bayesian melding. Transportation Research Part B 41: 652669.
18. Inuenzanet. website (acessed 2011).
19. Vrugt JA, ter Braak CJF, Diks CGH, Higdon D, Robinson B, et al. (2008) Accelerating markov chain monte carlo simulation by differential evolution with self-adaptive randomized subspace sampling. Technical report, Citeseer.
20. Andrieu C, Thoms J (2008) A tutorial on adaptive MCMC. Statistics and Computing 18: 343373.
21. Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7: 434–455.

22. ECDC (2009) European inuenza surveillance network (eisn). (accessed 2011). URL www.ecdc.europa.eu.
23. Lau LL, Cowling BJ, Fang VJ, Chan KH, Lau EH, et al. (2010) Viral shedding and clinical illness in naturally acquired inuenza virus infections.
24. van Noort SP, Aguas R, Ballesteros S, Gomes MGM (2011) The role of weather on the relation between inuenza and inuenza-like illness. Submitted.
25. Gran JM, Iversen B, Hungnes O, Aalen OO (2010) Estimating inuenza-related excess mortality and reproduction numbers for seasonal inuenza in norway, 1975–2004. Epidemiology and Infection 138: 1559–1568.
26. Chowell G, Viboud C, Simonsen L, Miller M, Alonso WJ (2010) The reproduction number of seasonal inuenza epidemics in brazil, 1996–2006. Proceedings of the Royal Society B 277: 1857.
27. Paterson B, Durrheim DN, Tuyl F (2009) Inuenza: H1N1 goes to school. Science 325: 1071.
28. Akaike H (2003) A new look at the statistical model identification. Automatic Control, IEEE Transactions on 19: 723716.
29. Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6: 461–464.
30. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. Journal Of The Royal Statistical Society Series B 64: 583–639.
31. Kass RE, Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90: 773–795.
32. Coelho FC (2009) bayesian-inference – project hosting on google code. (acessed 2011). URL http://code.google.com/p/bayesian-inference/.
33. Stein W, et al. (2009) Sage Mathematics Software (Version 4.1.1). The Sage Development Team. (accessed 2011). http://www.sagemath.org.