

Chapter 18

Bayesian Inference of Deterministic Population Growth Models

Luiz Max Carvalho, Claudio J. Struchiner and Leonardo S. Bastos

Abstract Deterministic mathematical models play an important role in our understanding of population growth dynamics. In particular, the effect of temperature on the growth of disease-carrying insects is of great interest. In this chapter we propose a modified Verhulst—logistic growth—equation with temperature-dependent parameters. Namely, the growth rate r and the carrying capacity K are given by thermodynamic functions of temperature T , $r(T)$ and $K(T)$. Our main concern is with the problem of learning about unknown parameters of these deterministic functions from observations of population time series $P(t, T)$. We propose a strategy to estimate the parameters of $r(T)$ and $K(T)$ by treating the model output $P(t, T)$ as a realization of a Gaussian process (GP) with fixed variance and mean function given by the analytic solution to the modified Verhulst equation. We use Hamiltonian Monte Carlo (HMC), implemented using the recently developed **rstan** package of the R statistical computing environment, to approximate the posterior distribution of the parameters of interest. In order to evaluate the performance of our algorithm, we perform a Monte Carlo study on a simulated example, calculating bias and nominal coverage of credibility intervals. We then proceed to apply this approach to laboratory data on the temperature-dependent growth of a Chagas disease arthropod vector, *Rhodnius prolixus*. Analysis of this data shows that the growth rate for the insect population under study achieves its maximum around 26 °C and the carrying capacity is maximum around 25 °C, suggesting that *R. prolixus* populations may thrive even in non-tropical climates.

L. M. Carvalho (✉) · C. J. Struchiner · L. S. Bastos
Program for Scientific Computing (PROCC), Oswaldo Cruz Foundation,
Rio de Janeiro, RJ, Brazil
e-mail: lmax.procc@gmail.com

C. J. Struchiner
e-mail: stru@fiocruz.br

L. S. Bastos
e-mail: lsbastos@fiocruz.br

18.1 Introduction

Deterministic models of population growth have played a major role in our understanding of biological populations [1, 2, 4, 7]. These mechanistic models allow us to draw a picturesque picture of the real world and capture the main features of the system(s) under study.

In many applications it is of interest to estimate parameters of these models using observed (output) data. This chapter proposes a Bayesian approach to this problem in the context of models for insect population growth. In what follows, we introduce the necessary notation and concepts to be used hereafter.

Consider a deterministic model $M(\cdot)$. Let $y \in \mathcal{Y} \subset \mathbb{R}^n$ be the set of model inputs and $x \in \mathcal{X} \subset \mathbb{R}^p$ be the model outputs. The deterministic model $M(x; \theta) = y$, where $\theta \in \Theta \subset \mathbb{R}^q$ is a q -dimensional parameter vector, completely specifies the relationship between x and y .

Here we are concerned with the problem of, having observed x and y , draw inference about θ . From a Bayesian perspective, we are concerned with obtaining the posterior distribution [7]:

$$p(\theta|x, y) \propto p(y, x|\theta)\pi(\theta) \quad (18.1)$$

$$\propto p(y|x, \theta)\pi(x|\theta)\pi(\theta) \quad (18.2)$$

$$\propto p(y|x, \theta)\pi(x)\pi(\theta), \quad (18.3)$$

where (18.3) follows from the assumption of a priori independence of the inputs and parameters.

We discuss several aspects of the uncertainty in such models and illustrate with a temperature-dependent Verhulst model, proposed in [11].

This chapter is organized as follows: in Sect. 18.1 we outline the necessary theory and notation. Sect. 18.2 details the model, likelihood, priors, and posteriors. A simulation study is presented in Sect. 18.2.4 and a discussion with our closing remarks is given in Sect. 18.3.

18.2 Logistic Growth with Temperature-Dependent Parameters

Global temperature change may be an important factor on infectious diseases emergence [6]. Arthropod-borne diseases are particularly influenced by climate change; their vectors are very sensitive to temperature variation. In this chapter we are interested in studying the temperature dependency of population growth of a Chagas disease vector, the insect *Rhodnius prolixus*.

To this end, we introduce a modified logistic growth equation also known as the Verhulst equation [10, 11], with temperature-dependent parameters. The ordinary non-linear differential equation

$$\frac{dP}{dt} = r \left(1 - \frac{P}{K} \right) P, \quad (18.4)$$

takes two parameters, the growth rate r and the carrying capacity K . For a given initial population condition N_0 , an analytic solution is available for the number $P(t)$ of individuals at time t :

$$P(t) = \frac{K}{1 + \left(\frac{K-N_0}{N_0}\right) e^{-rt}}. \quad (18.5)$$

In order to incorporate temperature-dependent behavior, we introduce temperature-dependent parameters, $r(T)$ and $K(T)$. The analytic solution in (18.5) is slightly modified to yield $P(t, T)$ for time t and temperature T :

$$P(t, T) = \frac{K(T)}{1 + \left(\frac{K(T)-N_0}{N_0}\right) e^{-r(T)t}}. \quad (18.6)$$

To complete the model we must specify $K(T)$ and $r(T)$ as smooth functions of T . We model $K(T)$ and $r(T)$ as Gaussian kernels over T :

$$K(T) = c_K \exp\left(-\frac{(T - a_K)^2}{b_K}\right) \quad (18.7)$$

$$r(T) = c_r \exp\left(-\frac{(T - a_r)^2}{b_r}\right). \quad (18.8)$$

18.2.1 Likelihood

Assume $P(t, T)$ to be a Gaussian process (GP) with fixed variance τ^2 . Recalling the notation in Sect. 18.1, we have $x = \{T, t, N_0\}$, $y = \{P(t, T)\}$ and $\theta = \{a_K, b_K, c_K, a_r, b_r, c_r, \tau^2\}$. Additionally, note that the parameter vector θ can be split into the disjoint sets $\theta_K = \{a_K, b_K, c_K\}$ and $\theta_r = \{a_r, b_r, c_r\}$, which parameterize $K(T)$ and $r(T)$, respectively.

Let $\mathbf{y} = \{y_1, y_2, \dots, y_N\}$ be an output vector with N measurements, which we observe directly. Moreover, let $\mathbf{t} = \{t_1, t_2, \dots, t_N\}$ be the vector which contains the observed times of the observations and \mathbf{T} the analogous vector for experimental temperatures. We then specify

$$y_i | t_i, T_i, N_0, \theta \sim \mathcal{N}(\mu(t_i, T_i, N_0; \theta), \tau^2) \quad (18.9)$$

$$\mu(t_i, T_i, \theta) = \frac{K(T_i; \theta_K)}{1 + \left(\frac{K(T_i; \theta_K) - N_0}{N_0}\right) e^{-r(T_i; \theta_r) t_i}}, \quad \forall i = 1, 2, \dots, N \quad (18.10)$$

which is equivalent to writing $y_i = M(t_i, T_i, N_0; \theta) + \epsilon$, $\epsilon \sim \mathcal{N}(0, \tau^2)$.

18.2.2 Priors

For $i = 1, 2, \dots, K$, let θ_i denote each of the $K = 7$ model parameters. Assuming a priori independence, we have

$$\pi(\theta) = \prod_{i=1}^K \pi(\theta_i). \quad (18.11)$$

We then specify prior distributions for the θ as follows¹:

$$a_K, a_r \sim \text{Normal}(20, 10) \quad (18.12)$$

$$b_K, b_r \sim \text{Gamma}(4, 5) \quad (18.13)$$

$$c_K \sim \text{Gamma}(1, 1000) \quad (18.14)$$

$$c_r \sim \text{Normal}(1/2, 2) \quad (18.15)$$

$$\tau^2 \sim \text{Gamma}(1/10, 10). \quad (18.16)$$

These priors were formulated to reflect both model restriction (e.g., positivity and concavity) and biological knowledge about model parameters.

The a_K and a_r parameters mimic the mean parameter in a Gaussian distribution, and control where the functions will achieve their maximum. We assume a normal distribution with moderate variance to provide a relatively uninformative accounting of placement in the positive side of the real line.

For b_K and b_r it is also necessary to ensure positive-definiteness, which we achieve using Gamma priors. Our prior for b allows very “stretched” forms of $K(T)$ and $r(T)$ (see Fig. 18.3), while ensuring concavity. Finally, c_K and c_r control curve height for $K(T)$ and $r(T)$ respectively.

The choice for c_K is essentially arbitrary, since little is known about natural populations of *R. prolixus* in terms of carrying capacity. We thus parameterize $\pi(b_K)$ to reflect rough projections taking into account the number of laid eggs. Since r in Eq. 18.4 can theoretically assume negative values, we used a prior for c_r that allows negative values.

Now, let θ^r and θ^K be partitions of the parameter vector, which in combination with the deterministic forms given in (18.7) and (18.8) induce prior distributions on r and K for each fixed temperature T , $\pi^*(r) = r(T; \pi(\theta^r))$, and $\pi^*(K) = K(T; \pi(\theta^K))$ respectively. These prior distributions can be easily approximated using Monte Carlo sampling and are presented in Fig. 18.3 (dashed lines).

¹ Please note that throughout this text we assume gamma distributions are parameterised in terms of shape and scale—as opposed to rate—and normal distributions are parameterised in terms of mean and standard deviation.

18.2.3 Posterior

Hence, from (18.3) and assuming a priori independence of \mathbf{t} and \mathbf{T}

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}, \mathbf{T}) \propto p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{t}, \mathbf{T})\pi(\mathbf{t})\pi(\mathbf{T})\pi(\boldsymbol{\theta}), \quad (18.17)$$

is the desired posterior.

This distribution can not be obtained in closed form, therefore we have resort to MCMC techniques to obtain an approximation. Hamiltonian Monte Carlo (HMC)[5] is an MCMC algorithm that replaces the random walk (RW) proposal of the Metropolis-Hastings algorithm by a “leapfrog” proposal based on the gradient of the log-posterior distribution. We then used the **stan** [9] package of the R Statistical Computing Environment [8] to approximate (18.17) through Hamiltonian Monte Carlo (HMC). The HMC implementation of **rstan** relies on the No-U-turn sampler (NUTS) [3] which automatically sets the step size (ϵ) and the number of steps, i.e., trajectory length (L), dropping the need for hand-tuning and making the algorithm efficient for a broad class of target distributions. Among the the advantages of HMC we would like to highlight the ability of making moves far away from the current value with higher acceptance probability than RW, thus providing quicker convergence even for highly correlated target posteriors. R code implementing the posterior approximation presented here is available at <https://github.com/maxbiostat/CODE/tree/master/BIDPGM>.

The MCMC was run for 50,000 iterations until convergence which was assessed by inspecting the trace- and autocorrelation plots and potential scale reduction factor. After discarding the burn-in we approximate the posterior distribution of $P(t, T)$ using Monte Carlo sampling as follows. Let Q be the number of samples.

1. Construct a grid of values for t and T
2. For each $q = 1, 2, \dots, Q$ draw a vector $\boldsymbol{\theta}^{(q)} = \{a_K^{(q)}, b_K^{(q)}, c_K^{(q)}, a_r^{(q)}, b_r^{(q)}, c_r^{(q)}\}$ from the posterior distribution of the parameters
3. Evaluate $M(t, T, N_0; \boldsymbol{\theta}^{(q)})$ to get $P(t, T)^{(q)}$

From these Q samples, we can compute several quantities of interest, for instance the posterior mean of the population for a particular temperature at a given time. Let t_j and T_j represent a particular pair of temperature and time. Then the posterior mean of $P(t_j, T_j)$ is

$$\mathbb{E}[P(t_j, T_j)] = \frac{1}{Q} \sum_{q=1}^Q P(t_j, T_j)^{(q)} = \frac{1}{Q} \sum_{q=1}^Q M(t_j, T_j, N_0; \boldsymbol{\theta}^{(q)}). \quad (18.18)$$

The posterior median and quantiles can be obtained in a similar fashion.

Table 18.1 Bias assessment using the simulated data set—posterior mean

Parameter	Value ^a	Posterior mean	Bias ^b	MSE	Coverage ^c
a_K	30.00	29.44	0.01	7.99	0.93
a_r	23.00	22.71	0.00	3.31	0.86
b_K	10.00	13.08	0.95	450.26	0.94
b_r	15.00	16.82	0.22	16.77	0.88
c_K	700.00	692.17	0.09	7203.06	0.96
c_r	0.40	0.43	0.00	0.04	0.85
τ	3.16	4.89	0.94	67.02	0.88

^a “True” value used for simulation
^b Bias divided by the true parameter value
^c Coverage of the 95 % credibility intervals

18.2.4 Simulation Study

We conducted a Monte Carlo simulation study in order to evaluate the approach proposed here. Simulation was carried out as follows: for each m in a total of M simulation steps,

1. Fix θ , τ and N_0
2. Construct a grid of values for t and T
3. For each point in the grid, sample from a normal distribution with mean $M(T, t, N_0; \theta)$ and variance τ , generating a data set $P^{(m)}$
4. Using 50,000 iterations of HMC, obtain an estimate $\hat{\theta}^{(m)}$ of model parameters

In this chapter, we use both the *a posteriori* mean and median as point estimates for θ . With these results at hand, we are able to compute the normalized squared bias and mean squared error (MSE) for each parameter, as well as nominal coverage for the 95 % credibility intervals. The normalized squared bias for each parameter is defined as

$$B(\theta_i) = \theta_i^{-1} \mathbb{E} \left[\hat{\theta}_i - \theta_i \right]^2.$$

(18.19)

Let $Z_i^{(m)}$ be the indicator variable that assumes 1 if the m th 95 % credibility interval contains the true value of θ_i and zero otherwise. Coverage is defined as

$$C(\theta_i) = \mathbb{E} \left[Z_i^{(m)} \right] = \frac{1}{M} \sum_{j=1}^M Z_{ij}.$$

(18.20)

Code for this step is also available at <https://github.com/maxbiostat/CODE/tree/master/BIDPGM>.

Table 18.1 shows the bias, MSE and coverage for each parameter using the posterior mean as a point estimate. Parameter values were chosen so as to reflect a biologically sound behavior for $P(t, T)$.

The results for the posterior median as point estimator were largely in agreement with those for the posterior mean and thus are omitted here.

In Fig. 18.1 we present prior and posterior distributions for model parameters used in this simulation study where we can notice the posteriors are substantially less diffuse than the priors.

18.2.5 *R. prolixus* data

We now turn our attention to a real-world data set. Chagas disease is an important tropical disease, transmitted by a blood sucking bug, *R. prolixus*. Temperature is key factor for both insect development and vector competence. We are thus interested in drawing inference on model parameters to understand the role of temperature in the insect's population dynamics.

In a laboratory experiment, $N_0 = 30$ females were observed in several temperatures, and the number of eggs laid was recorded for 35 days. We take the cumulative number of eclosed eggs for each temperature condition as a good approximation of $P(t, T)$, since all conditions (light, water, food, etc.) were optimal. The data thus consisted of $N = 350$ observations, from 10 temperature conditions for 35 days each.

In Fig. 18.2 we show prior and posterior distributions for each parameter using this data. Results are similar to that of the simulated data, with substantially concentrated posteriors in comparison to the priors. Interestingly, while the prior expectation for b_K was 25, we obtained a posterior mean around 106 (Table 18.2), indicating that the variability of the response of *Rhodnius* to temperature is much greater than we anticipated. See more on this at Sect. 18.3.

Figure 18.3 shows the prior and posterior (induced) distributions of the thermodynamic functions. These are easily obtained using Monte Carlo sampling from the prior and posterior distributions of θ , respectively.

Finally, we present the posterior distribution for $P(t, T)$ obtained using the approach described in Sect. 18.2.4. We also present the heat map of the original laboratory data described above and the posterior distribution obtained by Monte Carlo sampling of the marginal posteriors from MCMC (Fig. 18.4). It can be noticed that the posterior distribution allows a region of optimality for populational growth around $20 - 25^\circ\text{C}$. Each thermodynamic function, $r(T)$ and $K(T)$ can have its own point of maximum, however.

In Table 18.2 we provide posterior estimates obtained for the real data along with prior expectations and 95 % credibility intervals.

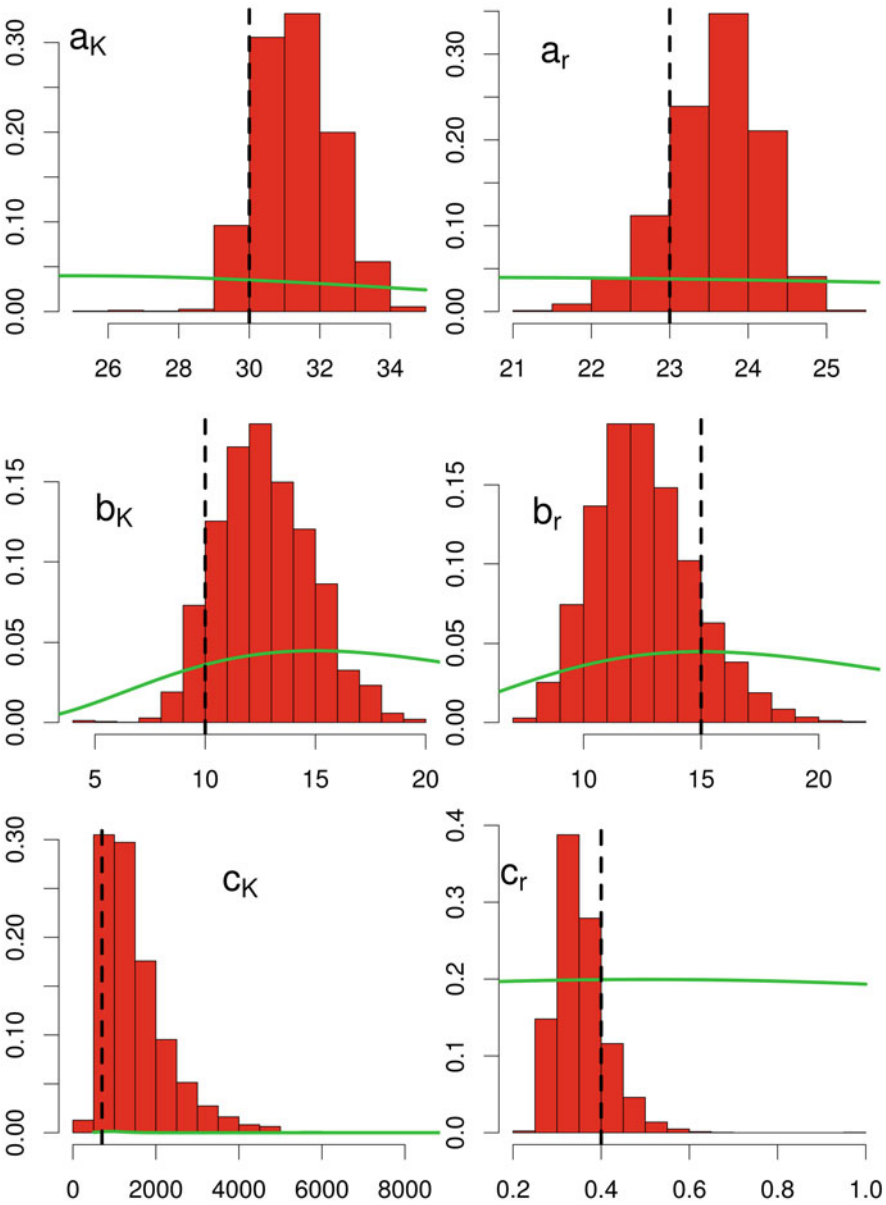


Fig. 18.1 Prior and posterior distributions of parameter for the simulated data. *Red bars* (histogram) show the marginal posterior distributions and *green lines* depict prior densities. *Dashed vertical lines* mark true parameter values. Please note that vertical axes differ

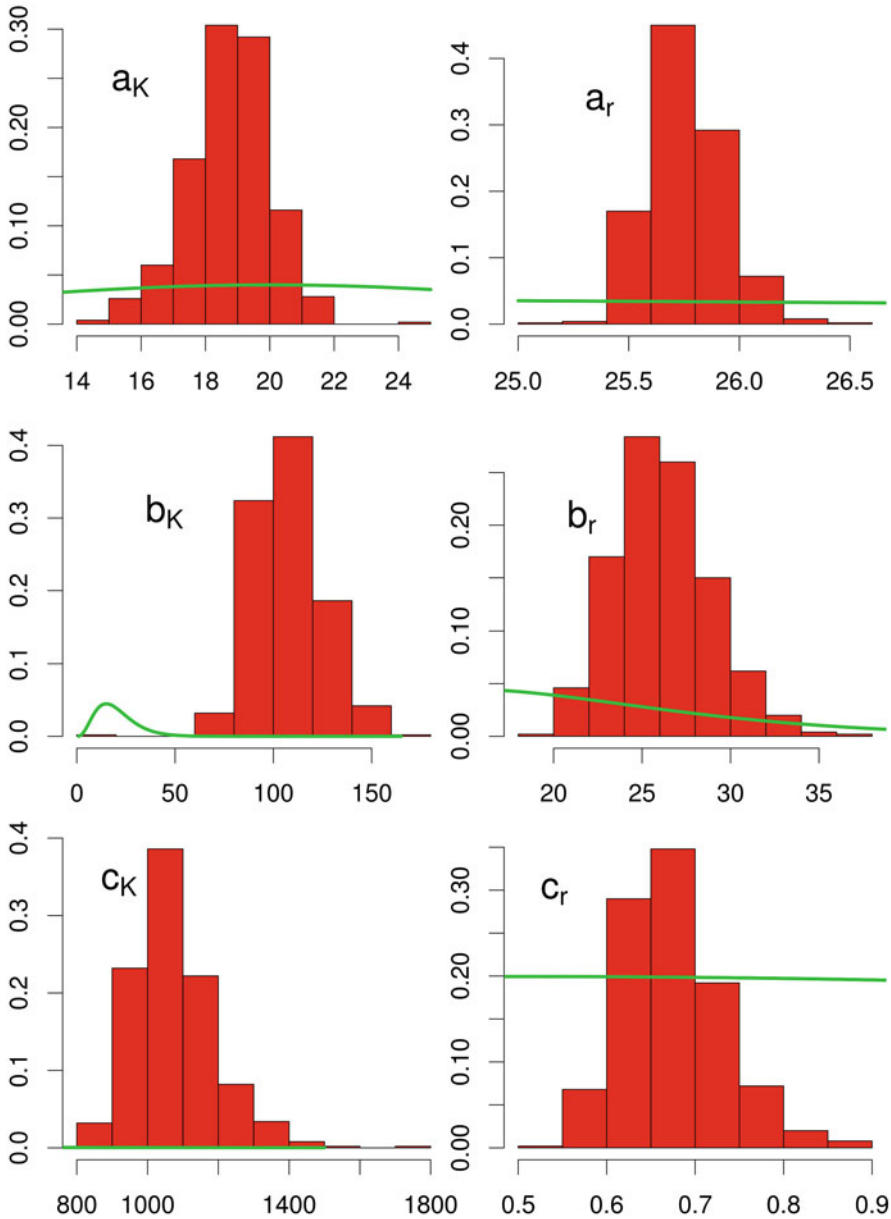


Fig. 18.2 Prior vs posterior distributions of parameter for the real data set. Please note that vertical axes differ

Table 18.2 Posterior inference results for the real set. We report posterior and prior expectations, along with the appropriate 95 % credibility intervals. Five independent chains were run for 50,000 iterations each with the first 25,000 discarded as burn-in. Convergence was assessed using trace plots and the potential scale reduction factor

	Posterior mean (95 % C.I.)	Prior mean (95 % C.I.)
a_K	19.23 (17.56 – 21.09)	25.00 (5.40–44.60)
a_r	25.73 (25.44–26.10)	25.00 (5.40–44.60)
b_K	106.17 (75.25–137.31)	20.00 (5.44–43.84)
b_r	26.77 (22.59–32.19)	20.00 (5.44–43.84)
c_K	1023.32 (898.28–1165.40)	1000.00 (25.31–3688.87)
c_r	0.66 (0.58–0.76)	0.50 (–3.41–4.41)
τ	177.33 (166.10–191.78)	1.00 (0.00–9.78)

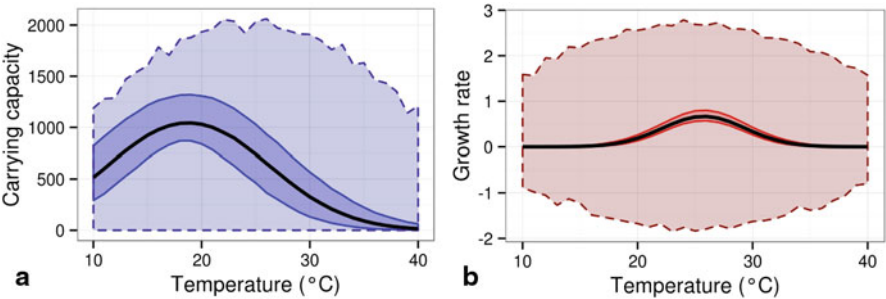


Fig. 18.3 Prior and posterior credibility intervals for the thermodynamic functions under for the real data set. We show $K(T)$ in (a) and $r(T)$ in (b). Dashed lines and lighter tones depict prior and solid lines and darker tones represent the posterior 95 % credibility intervals. Thick solid lines mark the medians

18.3 Discussion

Deterministic, differential-equation-based models are an important tool in Theoretical Biology, allowing us to study the behavior of biological systems using simple and easily interpretable equations. In this chapter we adapt a classical population growth model, the Verhulst logistic equation, initially proposed in 1838, to accommodate temperature-dependent parameters. We then proceed to develop a Bayesian approach to learn about model parameters when population time series are available.

Our approach is based on the idea that both the growth rate r and the carrying capacity K are smooth functions of temperature, which we model as Gaussian kernels of the form presented in (18.7). This parameterisation is very flexible, allowing us to model a variety of temperature-response patterns. It is biologically motivated, since the response of the insect populations to temperature change should have a maximum point and substantially decrease at extremely high and extremely low temperatures [11].

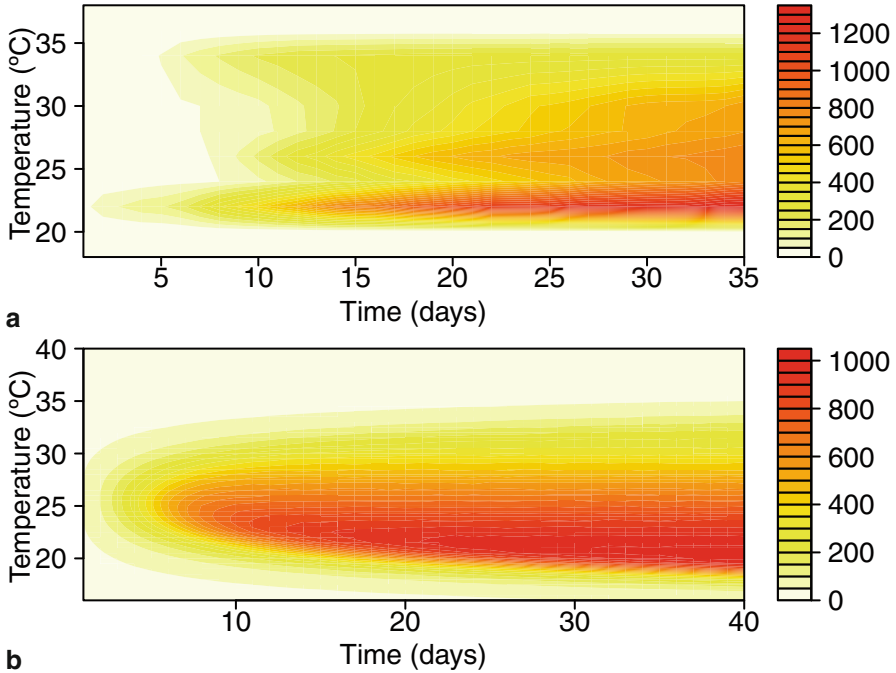


Fig. 18.4 Heat maps showing population through time and temperature. In **a** we show the laboratory data collected for *Rhodnius* and in **b** the posterior mean, obtained using Monte Carlo sampling from the posterior distribution of the parameters (Sect. 18.2.4). Please note that legend ranges differ slightly

The main idea is to model population through time as a GP with a deterministic mean function $M(\cdot)$ which is given by the solution to the differential equation (Eq. 18.6). The posteriors outlined in (18.3) and (18.17) allow for the incorporation of uncertainty regarding model inputs. Although this source of uncertainty is negligible in our setting due to carefully controlled experimental conditions, it could play an important role in studies dealing with field data.

It should also be noted that in this chapter we assume independence between model inputs. This assumption may be unrealistic, since temperature is likely to depend on time in general. In our particular conditions, all experiments were performed in controlled environments, where temperature was kept constant throughout time. In the case of population time series obtained from field data, this assumption is likely not to hold.

From Table 18.1 it can be noticed that our approach presents good frequentist properties, with high coverage probabilities of the credibility intervals for most parameters. The only exception is the GP variance, τ^2 , for which we could not recover the true value with good accuracy, albeit achieving good coverage. This result most likely stems from the restricting assumption of fixed variance (over time). Replacing

the fixed variance by a smooth function of time $\tau^2(t)$ could greatly improve model fit and is an important future direction of research.

In conclusion, in this chapter we provide insight on how to perform inference on the parameters of a complex, non-linear deterministic model of population growth when population time series are available. These parameters are directly interpretable and provide important information about the underlying biological dynamics. The framework proposed here can be adapted to a broad class of models, for example, to learn about the parameters of epidemiological models using data on disease incidence. Moreover, the Bayesian approach allows for a complete treatment of uncertainty on model inputs and outputs, thus providing more realistic predictions when data is subject to measurement uncertainty. The issue of how to incorporate and accommodate uncertainty about model structure is also an important one and constitutes an interesting avenue for future research.

Acknowledgements The authors would like to thank Professor Angela H. Lopes, Dr. Luciana Zimmermann and Luiz R. Vasconcellos (UFRJ) for providing the data set analysed in this chapter and for fruitful discussions. CJS was partially funded by CNPq and FAPERJ.

References

1. Costello, W.G., Taylor, H.M.: Deterministic population growth models. *Am. Math. Mon.* **78**(8), 841–855 (1971)
2. Gillespie, C.S., Golightly, A.: Bayesian inference for generalized stochastic population growth models with application to aphids. *J. R. Stat. Soc.: Ser. C (Appl. Stat.)* **59**(2), 341–357 (2010)
3. Hoffman, M.D., Gelman, A.: The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **15**, 1593–1623 (2014)
4. May, R.M., et al.: Simple mathematical models with very complicated dynamics. *Nature* **261**(5560), 459–467 (1976)
5. Neal, R.M.: MCMC using Hamiltonian dynamics. In: Brooks, S., Gelman, A., Jones, G., Meng, X.-L. (eds.) *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Press (2010)
6. Patz, J.A., Epstein, P.R., Burke, T.A., Balbus, J.M.: Global climate change and emerging infectious diseases. *J. Am. Med. Assoc.* **275**(3), 217–223 (1996)
7. Poole, D., Raftery, A.E.: Inference for deterministic simulation models: the Bayesian melding approach. *J. Am. Stat. Assoc.* **95**(452), 1244–1255 (2000)
8. R Core Team : R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2013). <http://www.R-project.org/>
9. Stan Development Team: Stan: a C++ library for probability and sampling, version 2.2 (2014). <http://mc-stan.org/>
10. Verhulst, P.F.: Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et physique publiée par a. Quetelet* **10**, 113–121 (1838)
11. Zimmermann, L.T., Carvalho, L.M.F., Vasconcellos, L.R., Bastos, L.S., Struchiner, C.J., Lopes, A.H.: Temperature-dependent oviposition and egg eclosion of Chagas disease vector *Rhodnius prolixus*. Submitted (2014)