bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Inference for epidemic models with time varying infection rates: tracking the dynamics of oak processionary moth in the UK

L E Wadkin[1,*], J Branson[2], A Hoppit[3], N G Parker[1], A Golightly[1] and A W Baggaley[1]

[1]School of Mathematics, Statistics and Physics, Newcastle University, UK, NE1 7RU

[2]GeoData, Geography and Environmental Science, University of Southampton, UK, SO17 1BJ

[3]Forestry Commission England, Nobel House, 17 Smith Square, London, UK, SW1P 3JR

*laura.wadkin@ncl.ac.uk

Abstract:

1. Invasive pests pose a great threat to forest, woodland and urban tree ecosystems. The oak processionary moth (OPM) is a destructive pest of oak trees, first reported in the UK in 2006. Despite great efforts to contain the outbreak within the original infested area of South-East England, OPM continues to spread.

2. Here we analyse data of the numbers of OPM nests removed each year from two parks in London between 2013 and 2020. Using a state-of-the-art Bayesian inference scheme we estimate the parameters for a stochastic compartmental SIR (susceptible, infested, removed) model with a time varying infestation rate to describe the spread of OPM.

3. We find that the infestation rate and subsequent basic reproduction number have remained constant since 2013 (with R_0 between one and two). This shows further controls must be taken to reduce R_0 below one and stop the advance of OPM into other areas of England.

4. *Synthesis.* Our findings demonstrate the applicability of the SIR model to describing OPM spread and show that further controls are needed to reduce the infestation rate. The proposed statistical methodology is a powerful tool to explore the nature of a time varying infestation rate, applicable to other partially observed time series epidemic data.

Keywords: Oak processionary moth, Bayesian inference, susceptible-infected-removed model, SIR model, stochastic differential equation, Markov chain Monte Carlo, epidemics

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

Invasive pests, such as non-native insects, pose a threat to forest, woodland and urban tree ecosystems by damaging and killing trees and reducing biodiversity [1–3]. This threat has increased in recent years due to growth in international travel and trade [4] coupled with a changing climate driving the migration of species into new ecosystems [5]. The loss of biodiversity has a profound economic impact, through short to long term control measures and the impact on ecosystem services [6–8].

The oak processionary moth (OPM), *Thaumetopoea processionea*, is an invasive and destructive pest of oak trees, causing defoliation and making trees vulnerable to other stressors and pathogens. The larvae of OPM have poisonous hairs, containing a urticating toxin (thaumetopoein) which is harmful to human and animal health [9–12].

OPM was introduced to the UK through accidental imports on live oak plants, first reported in 2006. Up to 2010, the governmental policy was one of eradication [13, 14]. However, in 2011 it was decided that OPM was fully established in the South-East England area and so the government moved to a containment strategy, aiming to contain the OPM infestations within this original outbreak area [14]. In 2018, legislation was introduced to curb continuing imports through the Plant Health Order [15]. Despite the containment strategies, the extent of OPM has continued to spread with recent analysis suggesting an expansion rate of 1.7 km/year for 2006–2014, with an increase to 6 km/year from 2015 onwards [16]. The regions surrounding the current infection area are particularly climatically suitable [17] and so being able to predict and control the future dynamics of the OPM population is crucial to protect these areas.

Mathematical models provide a powerful tool for describing and predicting the spread of tree disease and pest infections [18–20]. For OPM, previous work has included using models from electric network theory to predict high risk regions [21] along with species distribution models to examine the spatial distributions of OPM [22] and the effects of climate change on its expansion [17]. Bayesian inference can be used to inform and evaluate these ecological mathematical models [23]. Previously, Bayesian approaches have been used to estimate key parameters in the spatio-temporal invasion of alien species [24], however, the techniques have yet to be applied to data for the spread of OPM.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Nevertheless, the Bayesian paradigm provides a natural mechanism for quantifying and propagating uncertainty in the model parameters and dynamic components. Consequently, Bayesian inference techniques have been ubiquitously applied in the broad area of epidemiology [see e.g. 25–27, for an overview].

In this paper we use data tracking the numbers and locations of OPM nests removed from oak trees as part of a control program in two parks in south London. We illustrate the use of statistical inference techniques for estimating the parameters for a classic SIR compartmental model [28, 29] consisting of susceptible, infested and removed states. To allow for intrinsic stochasticity in the spread of OPM, we use an Itô stochastic differential equation [30] representation of the SIR model. This is further modified via the introduction of a time varying infestation rate, as is necessary to capture the effect of unknown influences such as preventative measures [31]. Bayesian inference for the resulting model is complicated by the intractability of the observed data likelihood, and subsequently, the joint posterior distribution of the key quantities of interest (model parameters and dynamic components). We overcome these difficulties via a linear Gaussian approximation of the stochastic SIR model, coupled with a Markov chain Monte Carlo scheme [32] for generating samples from the joint posterior. These methods are outlined in Section 2 and detailed in the Supplementary Information, Sections S1 and S2, for use as a toolbox to apply to other ecological datasets. We use the parameters from the compartmental model to estimate a yearly R_0 measure for OPM, analogous to the basic reproduction number for a pathogen [33], and estimate the OPM population in 2021.

2 Methods

In this section we present the observational time series data with a summary of the data collection methods (Section 2.1), the details of the stochastic SIR model (Section 2.2), and an outline of our statistical inference methods (Section 2.3). Further statistical details including the relevant algorithms are set out in the Supplementary Information (Sections S1 and S2).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.1 Data

The data used is part of the national Oak Processionary Moth Survey. The OPM survey is conducted by the Forestry Commission as part of the governmental OPM control programme [34]. The University of Southampton (GeoData) provide analysis, support and hold the data on behalf of the Forestry Commission. We note that for the wider OPM control programme, the data is collected for operational needs and therefore there are limitations for research purposes. For example, the surveillance strategy between 2013 and 2020 focussed on monitoring the expansion of the outer edge of the known infested area. Thus, although the data provides a sufficient estimate of the outer expansion, the presence of nests in the central infection zone is likely underestimated. However, in this paper we use a sub-set of this data from surveys in Richmond and Bushy Parks in which the whole park area was surveyed each year, and the number of nests recorded accurately.

The data used in this study was obtained through the recording of OPM presence in Richmond and Bushy Parks in South-West London. For each of the years 2013–2020 it contains i) the eastings and northings of infested tree and ii) the number of OPM nests removed from each tree. The dataset consists of 8470 unique infested trees, with 1767 in Bushy Park and 6703 in Richmond Park. The locations of infested trees are shown across the two parks in Figure 1.

The raw and cumulative time series of the numbers of removed nests are shown in Figure 2(a) and (b). We count each infested tree in the year it was first infested as one 'removal' in the SIR model (see Section 2.2), regardless of how many nests were recorded as removed from this location. The raw and cumulative time series for the number of these removals is shown in Figure 2(c) and (d). We use the latter cumulative time series, $R(t)$, as our observational data in the following sections.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

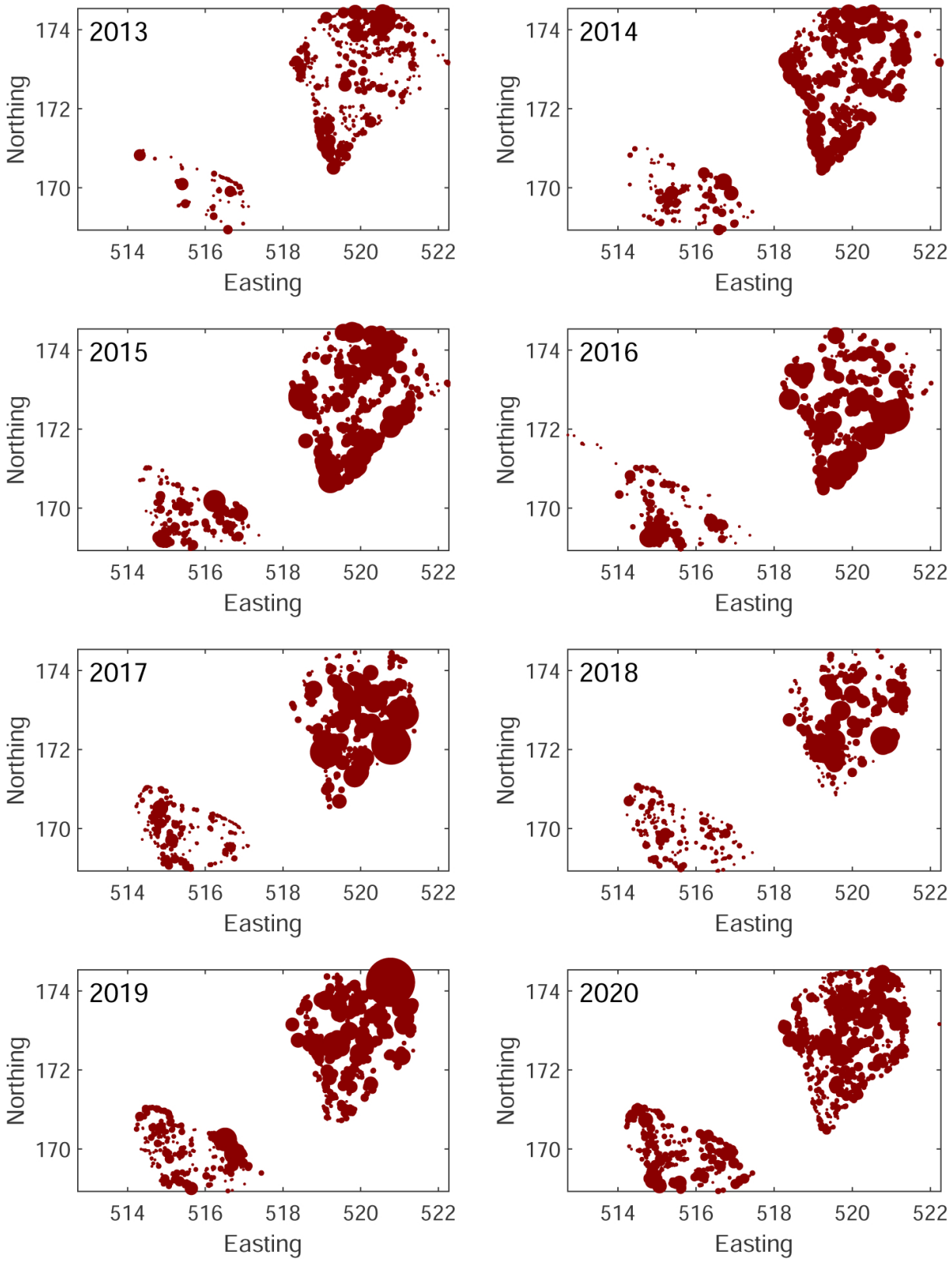


Figure 1: Map of the nests removed from Bushy (bottom left) and Richmond (top right) parks between 2013–2020. The area of the marker is proportional to the number of nests removed.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

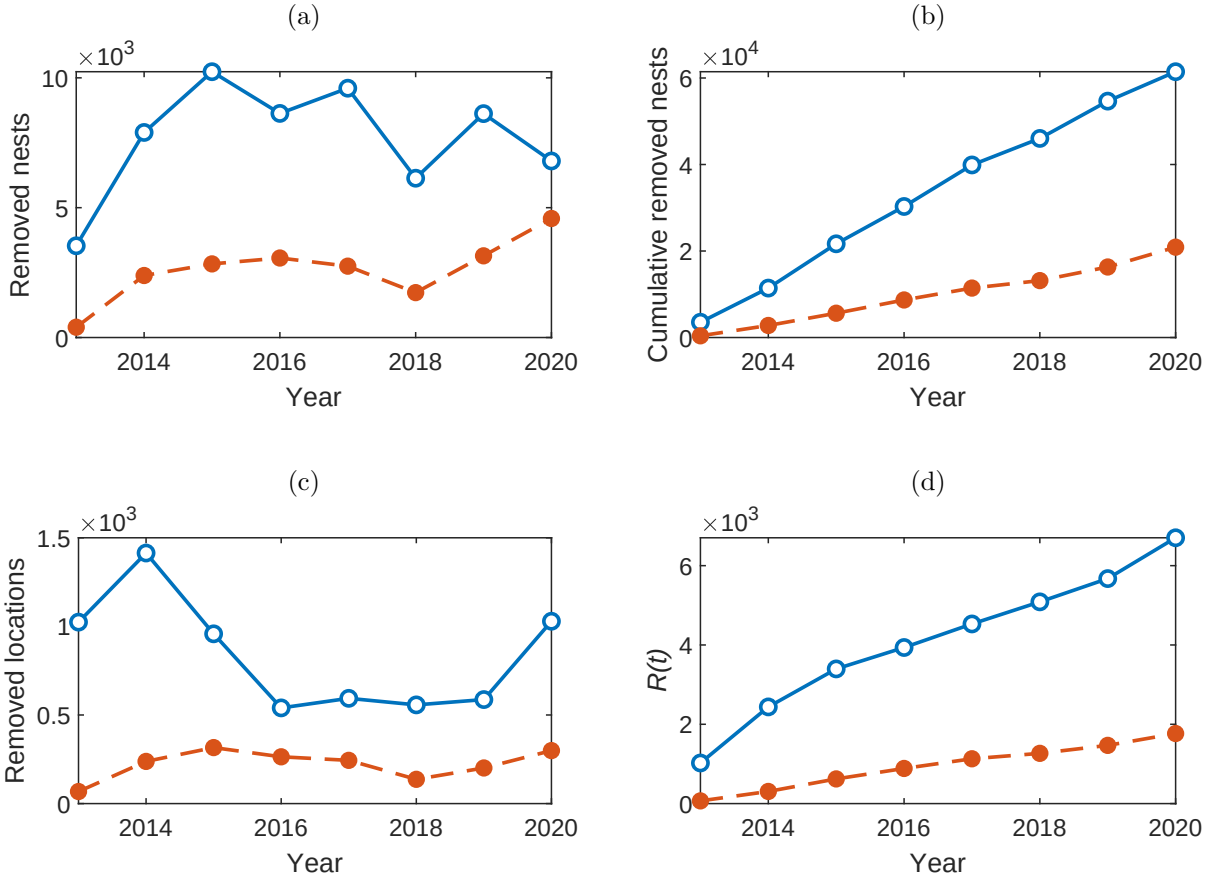


Figure 2: The (a) raw and (b) cumulative number of OPM nests removed from Richmond (blue) and Bushy (orange dashed) parks between 2013–2021. The number of (a) raw and (b) cumulative unique trees (described by their eastings and northings) which had nests removed between 2013–2021. The cumulative number of trees is the time series $R(t)$ corresponding to the 'removed' category in the SIR model (see Section 2.2).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2.2 Stochastic SIR model

We consider an SIR model [28, 35] in which a population of fixed size N is classified into compartments consisting of susceptible (S), infected (I) and removed (R) individuals. Here we refer to the I compartment as 'infested' due to the context of describing an invasive pest presence. Transitions between compartments can be summarised via two pseudo-reactions of the form

$$S + I \xrightarrow{\beta} 2I, \qquad I \xrightarrow{\gamma} R.$$

Hence, the first transition describes contact of an infective individual with a susceptible and with the net effect resulting in an additional infested individual and one fewer susceptible. The second transition accounts for removal (recovered with immunity, quarantined or dead) of an infested individual. The parameters β and γ govern the rate of infestation and removal, respectively. A fixed population of trees is appropriate as over the timescale of interest the number of trees born into the S compartment will be sufficiently small to be negligible. Our setting has individuals as trees and the contact process is understood to take place via the dispersal of OPM. It is clear that transitions should result in discrete changes to the numbers of trees in each state. This most naturally leads to a continuous time, discrete valued Markov jump process (MJP) description of disease dynamics, as detailed in the Supplementary Information, Section S1. We eschew the MJP formalism in favour of a continuous valued approximation, formulated as a stochastic differential equation (SDE). This is a pragmatic choice, since the SDE model ultimately leads to a computationally efficient inference scheme, and the model can be easily augmented with additional components, such as time varying parameters, which we now describe.

The SDE representation of the standard SIR model can be derived directly from the MJP (see Supplementary Information, Section S1). Here we extend this to include a time varying infection process. Let $X_t = (S_t, I_t, \tilde{\beta}_t)'$ where S_t and I_t denote the numbers in each of the states S and I at time $t \geq 0$ and $\tilde{\beta}_t = \log \beta_t$ is the (transformed) time varying infection process. Note that the fixed population size gives $R_t = N - S_t - I_t$ for all $t \geq 0$ so that the current state of the SIR model is

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

completely described by X_t. We model $\tilde{\beta}_t$ as a generalised Brownian motion process so that

$$d\tilde{\beta}_t = \sigma dW_t$$

and W_t is a standard Brownian motion process. Hence we assume that the log infection rate evolves according to a random walk in continuous time, with variability controlled by σ. Combining this process with component SDEs describing the dynamics of S_t and I_t gives the complete SDE description of the SIR model with time varying infection rate as

$$dX_t = a(x_t, \theta)dt + \sqrt{b(x_t, \theta)}dW_t. \tag{1}$$

Here, $x_t = (s_t, i_t, \beta_t)'$ is the state of the system at time t, $\theta = (\gamma, \sigma)'$ is the vector of static parameter values, $W_t = (W_{1,t}, W_{2,t}, W_{3,t})'$ is a vector of uncorrelated standard Brownian motion processes, and the drift function $a(x_t, \theta)$ and diffusion coefficient $b(x_t, \theta)$ are given by

$$a(x_t, \theta) = \begin{pmatrix} -\exp(\tilde{\beta}_t)s_t i_t \\ \exp(\tilde{\beta}_t)s_t i_t - \gamma i_t \\ 0 \end{pmatrix}, \qquad b(x_t, \theta) = \begin{pmatrix} \exp(\tilde{\beta}_t)s_t i_t & -\exp(\tilde{\beta}_t)s_t i_t & 0 \\ -\exp(\tilde{\beta}_t)s_t i_t & \exp(\tilde{\beta}_t)s_t i_t + \gamma i_t & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}. \tag{2}$$

Unfortunately, due to the nonlinear forms of $a(x_t, \theta)$ and $b(x_t, \theta)$, the SDE specified by (1)–(2) cannot be solved analytically. We therefore replace the intractable analytic solution with a tractable Gaussian process approximation, which is the subject of the next section. The resulting linear noise approximation is subsequently used as the inferential model.

Linear noise approximation

The linear noise approximation (LNA) provides a tractable approximation to the SDE given by (1)–(2). In what follows we give a brief derivation; formal details can be found in [36] [see also 37, 38].

Consider a partition of X_t as

$$X_t = \eta_t + R_t, \tag{3}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where $\{\eta_t, t \geq 0\}$ is a deterministic process satisfying the ODE

$$\frac{d\eta_t}{dt} = a(\eta_t, \theta), \qquad \eta_0 = x_0, \tag{4}$$

and $\{R_t, t \geq 0\}$ is a residual stochastic process. The residual process R_t satisfies

$$dR_t = \{a(x_t, \theta) - a(\eta_t, \theta)\}\, dt + \sqrt{b(x_t, \theta)}\, dW_t,$$

which will typically be intractable. Assumption that $||X_t - \eta_t||$ is "small" motivates a Taylor series expansion of $a(x_t, \theta)$ and $b(x_t, \theta)$ about η_t, with retention of the first two terms in the expansion of a and the first term in the expansion of b. This gives an approximate residual process $\{\hat{R}_t, t \geq 0\}$ satisfying

$$d\hat{R}_t = H_t \hat{r}_t\, dt + \sqrt{b(\eta_t, \theta)}\, dW_t,$$

where H_t is the Jacobian matrix with (i,j)th element

$$(H_t)_{i,j} = \frac{\partial a_i(\eta_t, \theta)}{\partial \eta_{j,t}}.$$

For the SIR model in (1)–(2) we therefore have

$$H_t = \begin{pmatrix} -\exp(\tilde{\beta}_t) i_t & -\exp(\tilde{\beta}_t) s_t & -\exp(\tilde{\beta}_t) s_t i_t \\ \exp(\tilde{\beta}_t) i_t & \exp(\tilde{\beta}_t) s_t - \gamma & \exp(\tilde{\beta}_t) s_t i_t \\ 0 & 0 & 0 \end{pmatrix}.$$

Given an initial condition $\hat{R}_0 \sim \mathrm{N}(\hat{r}_0, \hat{V}_0)$, it can be shown that $\hat{R}_t$ is a Gaussian random variable (see [32]). Consequently, the partition in (3) with R_t replaced by $\hat{R}_t$, and the initial conditions $\eta_0 = x_0$ and $\hat{R}_0 = 0$ give

$$X_t \sim \mathrm{N}\left(\eta_t, V_t\right), \tag{5}$$

where η_t satisfies (4) and V_t satisfies

$$\frac{dV_t}{dt} = V_t H_t' + b(\eta_t, \theta) + H_t V_t, \qquad V_0 = 0. \tag{6}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Hence, the linear noise approximation is characterised by the Gaussian distribution in (5), with mean and variance found by solving the ODE system given by (4) and (6). Although this ODE system will typically be intractable, a numerical scheme can be straightforwardly applied.

2.3 Bayesian inference

We consider the case in which not all components of the stochastic epidemic model are observed. Moreover, we assume that data points are subject to measurement error, which accounts for mismatch between the latent and observed process, due to, for example, the way in which the data are collected. Observations (on a regular grid) $y_t, t = 0, 1, \ldots n$ are assumed conditionally independent (given the latent process) with conditional probability distribution obtained via the observation equation,

$$Y_t \sim \mathrm{N}\left(P'x_t, \sigma_e^2 P'x_t P\right), \qquad t = 0, 1, \ldots, n \tag{7}$$

where $P = (1, 1, 0)'$. This choice of P is due to the data consisting of observations on the removed state R_t, which, for a known population size N, is equivalent (in information content) to observing the sum $S_t + I_t$. Note that the logarithm of the infection rate process, $\tilde{\beta}_t$ is completely unobserved. Our choice of observation model is motivated by a Gaussian approximation to a Poisson $Po(P'x_tP')$ distribution, with the role of σ_e^2 to allow a decoupling of the mean and variance. Moreover, the assumption of a Gaussian observation model admits a tractable observed data likelihood function, when combined with the LNA (see Section 2.2) as a model for the latent epidemic process X_t. Details on a method for the efficient evaluation of this likelihood function can be found in Section S2.3 of the Supplementary Information.

Given data $y = (y_0, y_1, \ldots, y_n)$ and upon ascribing a prior density $\pi(\theta)$ to the components of $\theta = (\gamma, \sigma, \sigma_e)'$ (augmented to include σ_e), Bayesian inference proceeds via the joint posterior for the static parameters θ and unobserved dynamic process $x = (x_0, x_1, \ldots, x_n)$. We have that

$$\pi(\theta, x|y) \propto \pi(\theta)\pi(y|\theta)\pi(x|y, \theta), \tag{8}$$

where $\pi(y|\theta)$ is the observed data likelihood and $\pi(x|y, \theta)$ is the conditional posterior density of the latent dynamic process. Although $\pi(y|\theta)$ and $\pi(x|y, \theta)$ can be obtained in closed form under

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the LNA, the joint posterior in (8) is intractable. In the Supplementary Information Section S2 we describe a Markov chain Monte Carlo scheme for generating (dependent) samples from (8). Briefly, this comprises two steps: i) the generation of samples $\theta^{(1)}, \ldots, \theta^{(M)}$ from the marginal parameter posterior $\pi(\theta|y) \propto \pi(\theta)\pi(y|\theta)$ and ii) the generation of samples $x^{(1)}, \ldots, x^{(M)}$ by drawing from the conditional posterior $\pi(x|y, \theta^{(i)})$, $i = 1, \ldots, M$.

Given inferences on the static parameters θ and the latent dynamic process x, we consider the following diagnostics for assessing model fit. The within sample predictive density is

$$\pi(\tilde{y}|y) = \int\int \pi(\tilde{y}|x, \theta)\pi(\theta, x|y)dxd\theta, \tag{9}$$

and the one step ahead out of sample predictive density is

$$\pi(y_{n+1}|y) = \int\int \pi(y_{n+1}|x_{n+1}, \theta)\pi(x_{n+1}|x_n, \theta)\pi(\theta, x|y)dx_{0:n+1}d\theta. \tag{10}$$

Hence, in both cases we properly account for parameter and latent process uncertainty. Although the densities in (9) and (10) will be intractable, we may generate samples via Monte Carlo, see Supplementary Information Section S2 for further details.

3 Results

We assume the epidemic time series (see Section 2.1) for the number of removed trees, $R(t)$, shown in Figure 2, can be described by the compartmental SIR model with a time-varying infestation rate (see Section 2.2). The aim is to estimate the key parameters through the Bayesian inference techniques described in Section 2.3. These are the time-varying infestation rate, $\beta(t)$, with corresponding stochastic noise parameter σ describing $d\tilde{\beta}_t = d\log(\beta_t) = \sigma W_{3,t}$, the removal rate γ, and the observation error σ_e.

Section S2.4 of the Supplementary Information provides details of the assumed population sizes for each site, initial numbers of infesteds, susceptibles, and infection rate, starting parameter values for the MCMC scheme and prior specification. Regarding the latter, we take an independent prior specification for the components of θ, so that $\pi(\theta) = \pi(\gamma)\pi(\sigma)\pi(\sigma_e)$. We then take lognormal

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$LN(1,1)$ distributions for σ and σ_e, and a lognormal $\mathrm{LN}(0, 0.5^2)$ distribution for γ. We assume that initial log infestation rate $\tilde{\beta}_0$ follows a Gaussian $\mathrm{N}(-8.5, 0.5^2)$ distribution. These choices are motivated by the assumption of a median removal time of around 1 year (95% credible interval: (0.38, 2.66)), and a basic reproduction number at time 0 of $R_0 = \beta_0 N/\gamma$ covering a wide range of plausible values. For example, with $N = 5 \times 10^3$ the prior distributions for γ and $\tilde{\beta}_0$ lead to a 95% credible interval for R_0 of $(0.25, 4.1)$. The initial conditions are chosen based on the increase in the removal category in the first available year, e.g., for Richmond Park there were 1414 new removals between 2013 and 2014 (new trees that had nests removed in 2014), so we assume this was approximately the number of infested locations in 2013. We investigated several choices for initial conditions and find our results robust to these variations.

3.1 Inference results

We ran the MCMC scheme for 10×10^3 iterations and monitored the resulting chains for convergence. Indicative trace plots can be found in the Supplementary Information, Figure S1, and suggest that the sampler has adequately explored the parameter space. Additional chains initialised at different starting values (not shown) further confirm convergence.

From the main MCMC run we obtain the posterior within-sample means (with 50% and 95% credible intervals) for $R(t)$, $S(t)$ and $I(t)$, shown in Figures 3 and 4(a-c) for Bushy and Richmond Park, respectively. The logarithmic time-dependent infestation rate, $\tilde{\beta}_t = \log(\beta)$, is shown in Figure 3 and 4(d). For Bushy Park, the logarithmic infestation rate is plausibly constant (given *a posteriori* variance) at $\tilde{\beta}_t \approx -8$, corresponding to an infestation rate of $\beta = 3.4 \times 10^{-4}$. Similarly, for Richmond Park, the infestation rate is plausibly constant with $\tilde{\beta}_t \approx -10$, corresponding to an infestation rate of $\beta = 4.5 \times 10^{-5}$. Reassuringly, samples from the within-sample predictive for $R(t)$ are consistent with the data used to fit the model (see panel (a) of Figures 3–4).

The posterior density plots of the parameters $\theta = (\gamma, \sigma, \sigma_e)$ are shown in Figures 3 and 4(e), for Bushy and Richmond park respectively. Pairwise joint posterior densities can be found in the Supplementary Information, Figure S2. The marginal posterior distribution of γ is centred around $\gamma \approx 1$ for both Bushy and Richmond. The marginal posterior for σ is centred around $\sigma \approx 0.75$ for Bushy and $\sigma \approx 0.5$ for Richmond. The observation error σ_e is centred around $\sigma_e \approx 1$ for Bushy and

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\sigma_e \approx 0.5$ for Richmond.

3.2 Estimation of R_0

From the posterior estimations of $\tilde{\beta}_t$ for each year and the parameter γ, we can estimate the basic reproduction number R_0. In a deterministic system, for an epidemic to die out, R_0 must be less than the threshold value of one. However, in the stochastic case it is possible for R_0 to be above one but the epidemic still die out as a result of the stochastic fluctuations. Therefore it is required that $R_0 < 1$ for the epidemic to shrink, upon averaging over the stochasticity. In an SIR model with a constant infection rate, β, the basic reproduction number is given by $R_0 = \beta N/\gamma$ (or simply β/γ if the total population number N has not been absorbed into the β constant in model formulation). Here we adapt this to use the time variant infection rate to get a reproduction number for each of the years between 2013 and 2020, $R_0(t) = \beta(t)N/\gamma$. Box plots showing the posterior distributions of R_0 for both parks are shown in Figure 5. For both parks R_0 has been stable, within errors, since 2013 (corresponding to the relatively constant β_t). However, this suggests that R_0 is still above one, and therefore the epidemic will continue to propagate in these areas, and potentially beyond.

3.3 Forward prediction

Predictions of the spread of OPM are needed to inform control strategies. To test the applicability of the SIR model with the inferred parameters and how well the model can capture future expansions in OPM, we can calculate a one year prediction. We remove the last data point, $R(2020)$, and re-infer the parameters for the new shortened observed time series. We then use these parameters to run the model forwards (10×10^3) simulations, matching the number of iterations in the MCMC) and obtain an estimate for $R(2020)$. The median predictions with upper and lower quartiles for 1000 runs are shown in Figure 6(a) and (c) for Bushy and Richmond, respectively. In both cases, the predictive interval captures the observed data. Realisations from 100 forward runs are shown in Figure 6(b) and (d) to show that results are mostly concentrated around the observed data, with some outliers over-estimating $R(t)$. One-step predictions for the whole time series are shown in the Supplementary Information, Figure S3.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Similarly we can produce predictions for the number of infested locations in 2021, $R(2021)$. The median predictions with upper and lower quartiles are shown in Figure 7(a) and (b) for Bushy and Richmond, respectively. This corresponds to an average (median) of 350 new infested locations (lower-upper quartile estimate range 150–800) in Bushy Park and 1100 (700–2000) in Richmond Park.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

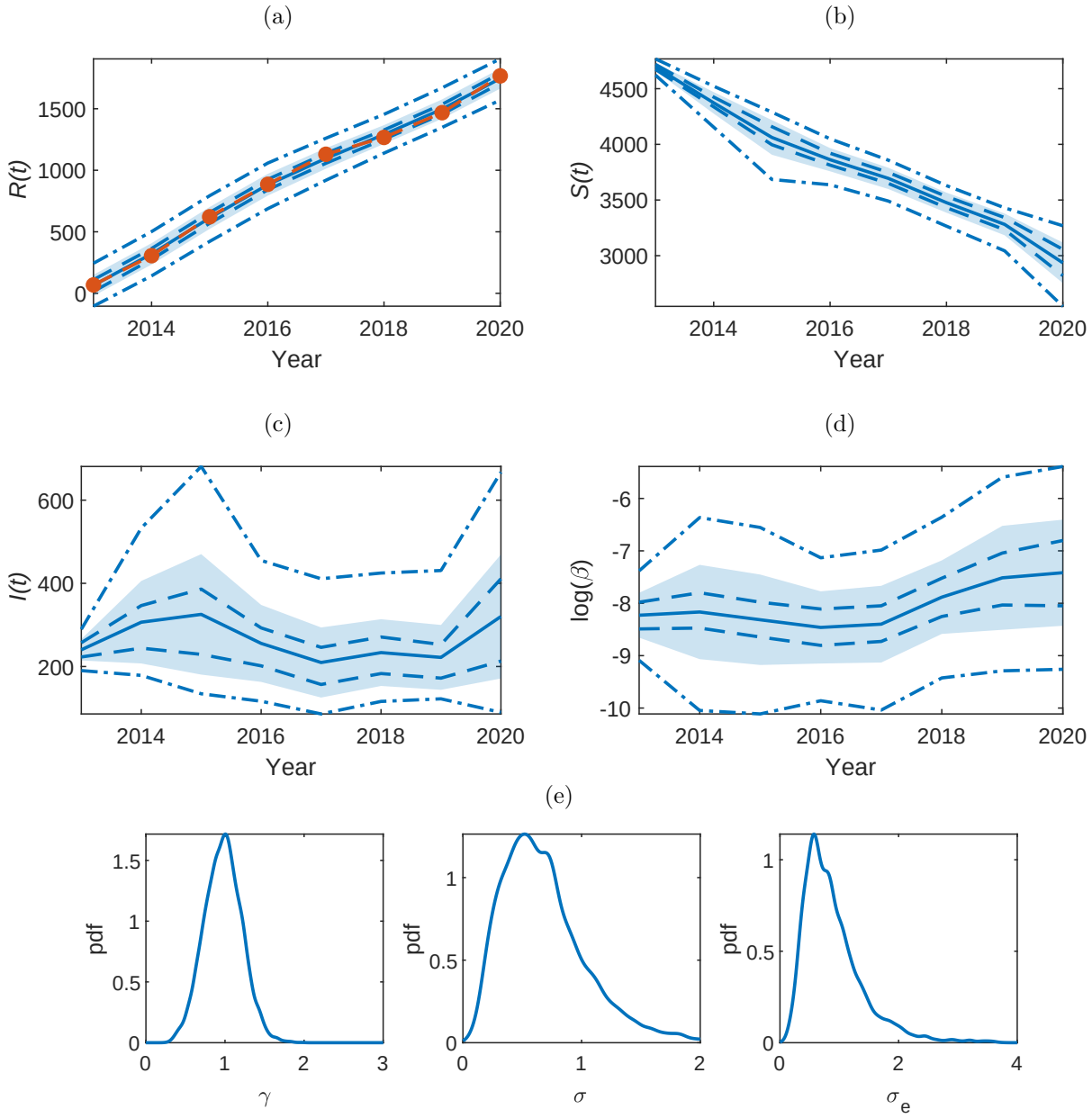


Figure 3: Bushy Park. The within-sample posteriors for (a) $R(t)$, (b) $S(t)$, (c) $I(t)$ and (d) $\log(\beta_t)$ with mean (blue solid line) $\pm$ one standard deviation (shaded region), the 50% (blue dashed), and the 95% (blue dot-dashed) credible regions. The observed time series for $R(t)$ is overlaid in (a) (orange dashed). The corresponding (e) posterior densities for the inferred parameters γ (removal rate), σ (noise on $\tilde{\beta}_t$) and σ_{e} (observation error).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

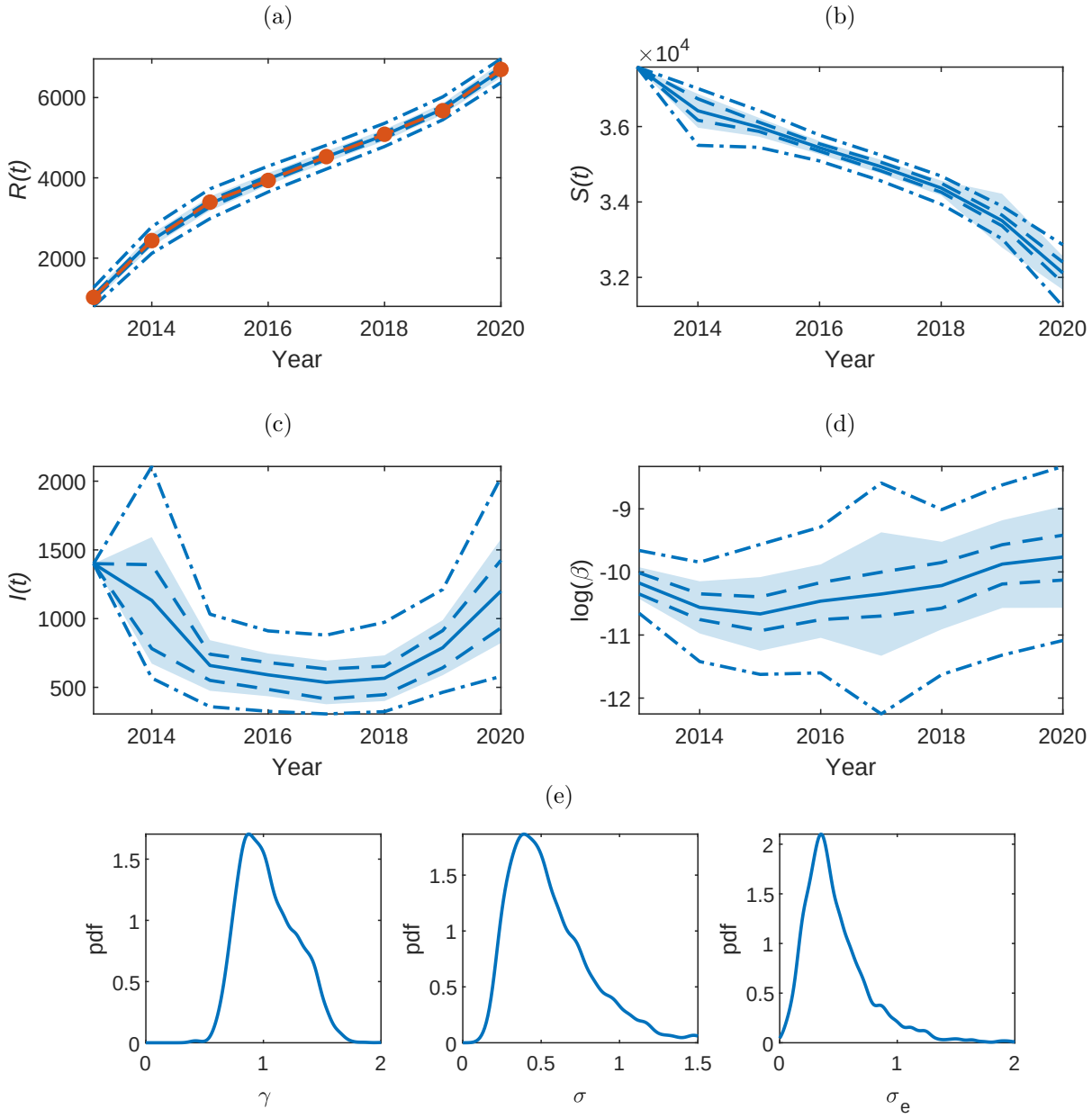
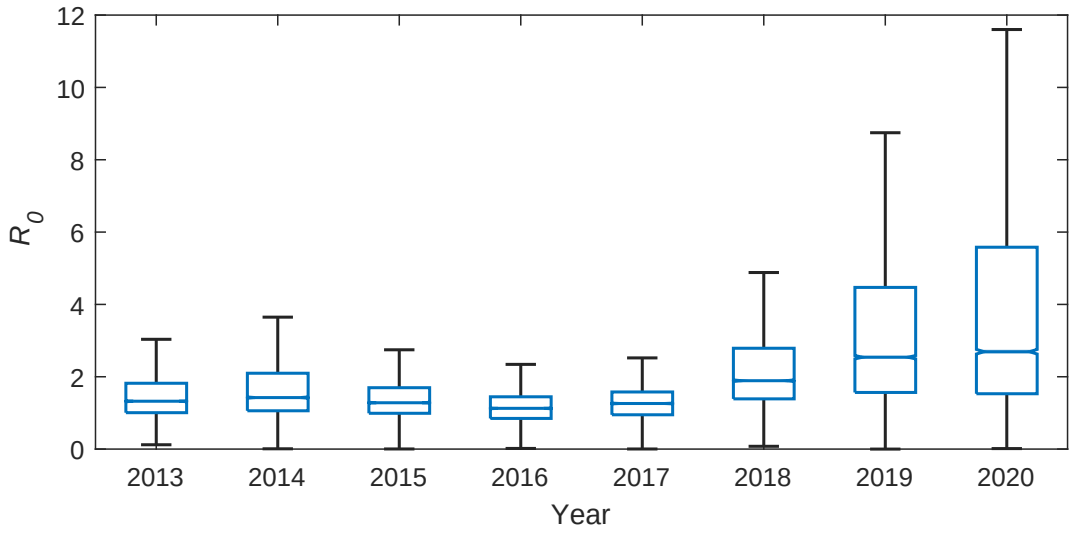


Figure 4: Richmond Park. The within-sample posteriors for (a) $R(t)$, (b) $S(t)$, (c) $I(t)$ and (d) $\log(\beta_t)$ with mean (blue solid line) $\pm$ one standard deviation (shaded region), the 50% (blue dashed), and the 95% (blue dot-dashed) credible regions. The observed time series for $R(t)$ is overlaid in (a) (orange dashed). The corresponding (e) posterior densities for the inferred parameters γ (removal rate), σ (noise on $\tilde{\beta}_t$) and σ_{e} (observation error).

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(a)



(b)

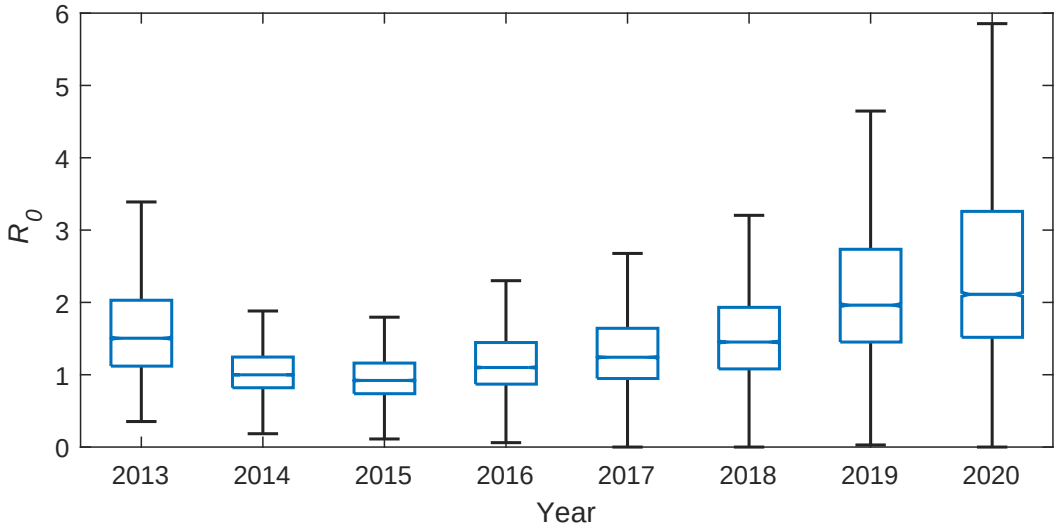


Figure 5: Posterior distributions of $R_0(t) = \beta_t N/\gamma$ for (a) Bushy and (b) Richmond Park. The central line indicates the median, with the bottom and top edges of the box showing the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, which are not shown here.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

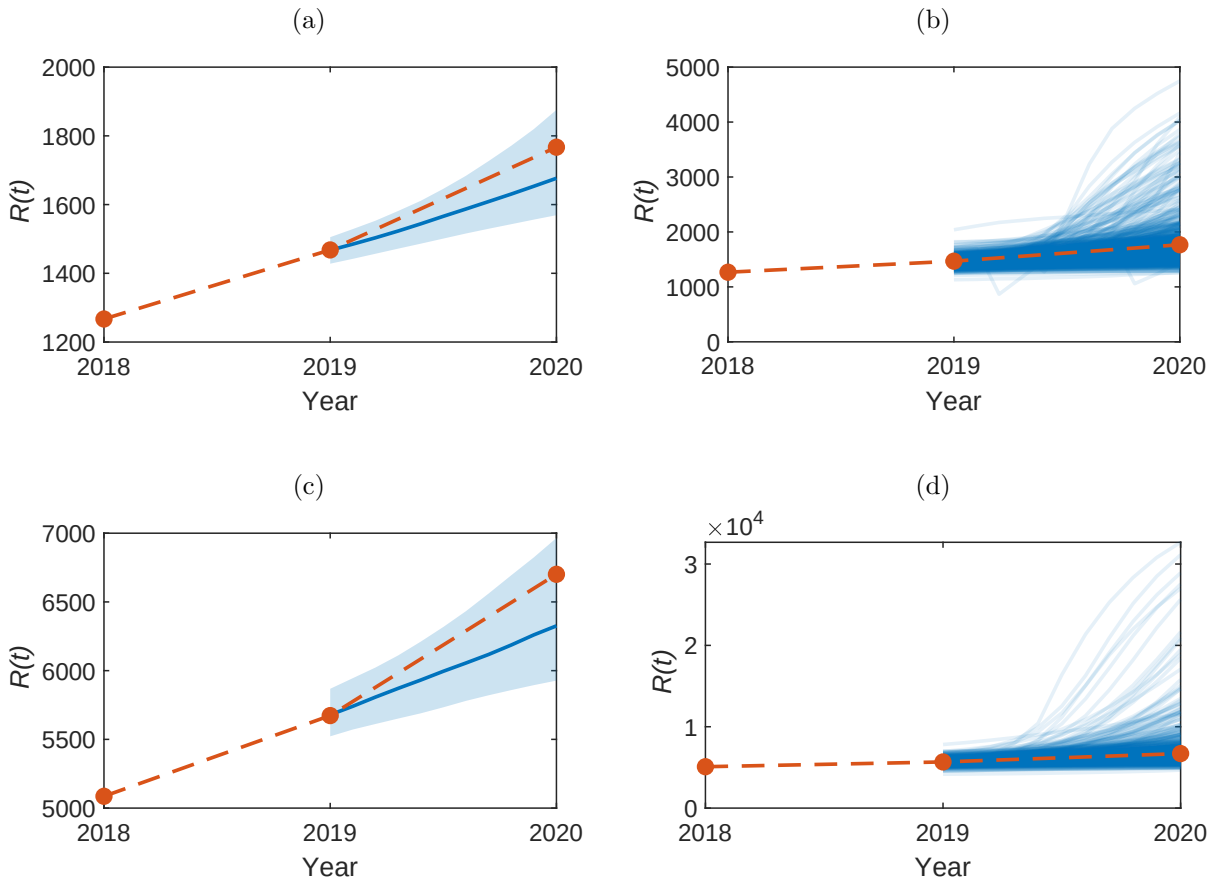


Figure 6: Model predictions for the number of removed nests in 2020, $R(2020)$, with median (blue line) for (a-b) Bushy and (c-d) Richmond. In (a) and (c) the shaded area shows the 50% credible region. In (b) and (d) 100 simulations are shown from the forward model. The orange line shows the observed data.

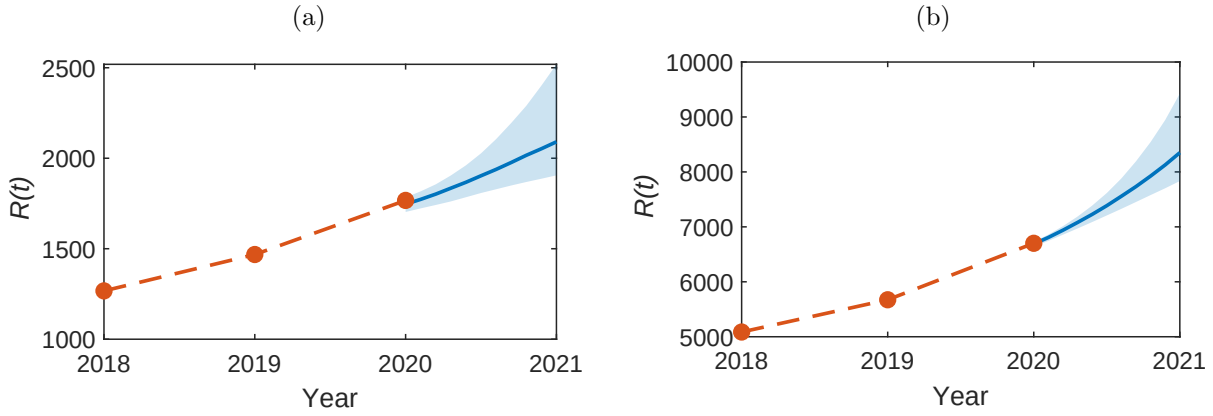


Figure 7: Model predictions for the number of removed nests in 2021, $R(2021)$, with median (blue line) for (a) Bushy and (b) Richmond park. The shaded area shows the 50% credible region. The orange line shows the observed data up to 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Discussion

Recent modelling work has suggested that the surroundings of the current OPM infection area in the UK are highly climatically suitable and therefore at very high risk from future infestations [17]. Since the government strategy for the containment of OPM relies on targetted control at the boundary of the current infested area, it is crucial to understand and be able to predict the future spread to optimize both the cost and efficacy of these control programmes [34].

We have shown the applicability of an SIR compartmental model with a time varying infection rate to describe the OPM epidemic in the UK between the years 2013 and 2020. The statistical methodology used is a powerful tool for inferring the parameters of such models from real data and is transferable to other epidemiological and ecological datasets. Previously, similar methodology has been used to describe the spread of infectious diseases (e.g., measles [39] and Ebola [40]) and the spatial expansion of non-native plants [24], but has not yet been applied to the study of invasive insects.

Our results show, along with previous analysis [16], that the spread of OPM is continuing at a stable rate despite the current intervention methods. Correspondingly, we show that the basic reproduction number R_0 has been above one since 2013. To see a reduction in the OPM population density and to protect the surrounding areas, a reduction of R_0 to below one would need to be seen. Although the basic reproduction number R_0 is typically used in the modelling of infectious diseases [41, 42], here it gives an analogous measure for the new infested locations caused by currently infested locations within the pest lifetime.

For simplicity and to be better described by an SIR model, we assumed that an infested location (tree) represented one removed tree regardless of how many nests were recorded as being removed from it. However, the defoliation effects and risks to human health from OPM are closely related to nest density (i.e., the numbers of nests per tree) [43]. In future work nest density could be taken into account through a nest density dependent infection rate.

A challenge of modelling OPM and other tree pests and diseases is the lack of a complete inventory oak trees in the UK, representing the susceptible population in our SIR model. This has been

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

previously noted and highlighted as a priority for future data collection by other modelling studies [21]. It is of particular importance for future spatial models of OPM, which require an estimate of the distribution of oak trees in the areas of interest.

It is also worth noting that many areas infested with OPM have been undergoing control measures [34] and so any inferred infestation rates represent the dynamics under these controls, rather than the inherent parameters of the uncontrolled pest spread. In Richmond and Bushy Parks, the yearly nest removal is a control measure. It would be interesting to conduct a similar analysis on a contained area that had undergone no (or different) control measures to assess the differences in the infestation rates and thus assess the efficacy of the controls. The effect of confounding factors such as the weather, difference in landscapes and presence of other pests and parasitoids, should also be investigated.

The results from this work can inform the development of future mathematical models for the spread of OPM. These models can be used to identify at-risk regions [21] and predict the distribution of OPM on a national scale. The development of these models will require further targetted data collection to obtain complete oak tree inventories, as well as data on the population numbers and locations of OPM (or indeed any other invasive insect or pathogen).

Author contributions LEW, NGP, AG and AWB conceptualised the project and designed the methodology. LEW analysed the data, implemented the computational models, produced the figures and led the writing of the manuscript. LEW and AG developed the inference code. JB and AH shared the OPM data and provided ecological expertise. NGP, AG and AWB acquired the funding. All authors contributed critically to the drafts and gave final approval for publication. The authors declare no conflicts of interest.

Acknowledgements We thank Gillian Jonusas from The Royal Parks for sharing the OPM data for Bushy and Richmond parks.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

[1] S. J. Manchester and J. M. Bullock. The impacts of non-native species on uk biodiversity and the effectiveness of control. *J. Appl. Ecol.*, 37(5):845–864, 2000.

[2] M. Kenis, M. Auger-Rozenberg, A. Roques, L. Timms, C. Péré, M. J. W. Cock, J. Settele, S. Augustin, and C. Lopez-Vaamonde. Ecological effects of invasive alien insects. *Biol. Invasions*, 11(1):21–45, 2009.

[3] P. H. Freer-Smith and J. F. Webber. Tree pests and diseases: the threat to biodiversity and the delivery of ecosystem services. *Biodivers. and Conserv.*, 26(13):3167–3181, 2017.

[4] B. A. Roy, H. M. Alexander, J. Davidson, F. T. Campbell, J. J. Burdon, R. Sniezko, and C. Brasier. Increasing forest loss worldwide from invasive pests requires new trade regulations. *Front. Ecol. Environ.*, 12(8):457–465, 2014.

[5] B. Lenzner, G. Latombe, C. Capinha, C. Bellard, F. Courchamp, C. Diagne, S. Dullinger, M. Golivets, S. D. H. Irl, I. Kühn, B. Leung, C. Liu, D. Moser, N. Roura-Pascual, H. Seebens, A. Turbelin, P. Weigelt, and F. Essl. What will the future bring for biological invasions on islands? An expert-based assessment. *Ecol. Evol*, 8:280, 2020.

[6] J. E. Aukema, B. Leung, K. Kovacs, C. Chivers, K. O. Britton, J. Englin, S. J. Frankel, R. G. Haight, T. P. Holmes, A. M. Liebhold, D. G. McCullough, and B. Von Holle. Economic impacts of non-native forest insects in the continental United States. *PLoS one*, 6(9):e24587, 2011.

[7] I. L. Boyd, P. H. Freer-Smith, C. A. Gilligan, and H. C. J. Godfray. The consequence of tree pests and diseases for ecosystem services. *Science*, 342(6160), 2013.

[8] C. J. A. Bradshaw, B. Leroy, C. Bellard, D. Roiz, C. Albert, A. Fournier, M. Barbet-Massin, J. Salles, F. Simard, and F. Courchamp. Massive yet grossly underestimated global costs of invasive insects. *Nat. Commun.*, 7(1):1–8, 2016.

[9] H. Maier, W. Spiegel, T. Kinaciyan, H. Krehan, A. Cabaj, A. Schopf, and H. Hönigsmann. The oak processionary caterpillar as the cause of an epidemic airborne disease: survey and analysis. *Br. J. of Dermatol.*, 149(5):990–997, 2003.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[10] H. Maier, W. Spiegel, T. Kinaciyan, and H. Hönigsmann. Caterpillar dermatitis in two siblings due to the larvae of Thaumetopoea processionea L., the oak processionary caterpillar. *Dermatology*, 208(1):70–73, 2004.

[11] S. Gottschling and S. Meyer. An epidemic airborne disease caused by the oak processionary caterpillar. *Pediatr. Dermatol.*, 23(1):64–66, 2006.

[12] S. Rahlenbeck and J. Utikal. The oak processionary moth: a new health hazard? *Br. J. Gen. Pract.*, 65(637):435–436, 2015.

[13] M. J. Mindlin, O. Le Polain de Waroux, S. Case, and B. Walsh. The arrival of oak processionary moth, a novel cause of itchy dermatitis, in the UK: experience, lessons and recommendations. *Public health*, 126(9):778–781, 2012.

[14] I. Tomlinson, C. Potter, and H. Bayliss. Managing tree pests and diseases in urban settings: the case of oak processionary moth in London, 2006–2012. *Urban For. Urban Green.*, 14(2):286–292, 2015.

[15] Plant Health (England) (Amendment) (no. 3) Order, 2018. [Online; accessed 06-December-2021].

[16] Y. F. Suprunenko, M. D. Castle, C. R. Webb, J. Branson, A. Hoppit, and C. A. Gilligan. Estimating expansion of the range of oak processionary moth (thaumetopoea processionea) in the UK from 2006 to 2019. *Agric. For. Entomol.*, 2021.

[17] M. Godefroid, N. Meurisse, F. Groenen, C. Kerdelhué, and J. P. Rossi. Current and future distribution of the invasive oak processionary moth. *Biol. Invasions*, 22(2):523–534, 2020.

[18] X. Wang and X. Song. Mathematical models for the control of a pest population by infected pest. *Comput. Math. with Appl.*, 56(1):266–278, 2008.

[19] V. I. Gertsev and V. V. Gertseva. Classification of mathematical models in ecology. *Ecol. Modell.*, 178(3-4):329–334, 2004.

[20] S. Orozco-Fuentes, G. Griffiths, M. J. Holmes, R. Ettelaie, J. Smith, A. W. Baggaley, and N. G. Parker. Early warning signals in plant disease outbreaks. *Ecol. Modell.*, 393:12–19, 2019.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[21] D. J. Cowley, O. Johnson, and M. J. O. Pocock. Using electric network theory to model the spread of oak processionary moth, thaumetopoea processionea, in urban woodland patches. *Landsc. Ecol*, 30(5):905–918, 2015.

[22] T. P. H. Scholtens. Predicting the oak processionary moth population using statistical modeling. Master's thesis, Eindhoven University of Technology, 2021.

[23] A. M. Ellison. Bayesian inference in ecology. *Ecol. Lett.*, 7(6):509–520, 2004.

[24] A. Cook, G. Marion, A. Butler, and G. Gibson. Bayesian inference for the spatio-temporal invasion of alien species. *Bull. Math.*, 69(6):2005–2025, 2007.

[25] C. Fuchs. *Inference for diffusion processes: with applications in life sciences.* Springer Science & Business Media, 2013.

[26] T. J. McKinley, J. V. Ross, R. Deardon, and A. R. Cook. Simulation-based Bayesian inference for epidemic models. *Comput. Stat. Data Anal.*, 71:434–447, 2014.

[27] T. Kypraios, P. Neal, and D. Prangle. A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. *Math. Biosci.*, 287:42–53, 2017.

[28] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. Maths. Phys*, 115(772):700–721, 1927.

[29] M. Bartlett. Some evolutionary stochastic processes. *J R Stat*, 11:211–229, 1949.

[30] B. Øksendal. *Stochastic differential equations: An introduction with applications.* Springer-Verlag, Berlin Heidelberg New York, 6th edition, 1995.

[31] J. Dureau, K. Kalogeropoulos, and M. Baguelin. Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, 14(3):541–555, 2013.

[32] P. Fearnhead, V. Giagos, and C. Sherlock. Inference for reaction networks using the Linear Noise Approximation. *Biometrics*, 70:457–456, 2014.

[33] J. A. P. Heesterbeek and K. Dietz. The concept of R0 in epidemic theory. *Stat. Neerl.*, 50(1):89–110, 1996.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[34] Forestry Commision, Oak Processionary Moth (Thaumetopoea processionea) Contingency Plan, 2021. [Online; accessed 28-October-2021].

[35] H. K. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*. Lect. Notes Stat. Springer-Verlag, New York, 2000.

[36] T. G. Kurtz. The relationship between stochastic and deterministic models for chemical reactions. *J. Chem. Phys.*, 57:2976–2978, 1972.

[37] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. North-Holland, 2001.

[38] M. Komorowski, B. Finkenstadt, C. Harper, and D. Rand. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinform.*, 10(1):343, 2009.

[39] S. Cauchemez and N. M. Ferguson. Likelihood-based estimation of continuous-time epidemic models from time-series data: application to measles transmission in London. *J. R. Soc. Interface*, 5(25):885–897, 2008.

[40] J. Fintzi, J. Wakefield, and V. N. Minin. A linear noise approximation for stochastic epidemic models fit to partially observed incidence counts. *arXiv preprint arXiv:2001.05099*, 2020.

[41] K. Dietz. The estimation of the basic reproduction number for infectious diseases. *Statistical methods in medical research*, 2(1):23–41, 1993.

[42] J. Ma. Estimating epidemic exponential growth rate and basic reproduction number. *Infect. Dis. Model.*, 5:129–141, 2020.

[43] H. Jactel, G. Birgersson, S. Andersson, and F. Schlyter. Non-host volatiles mediate associational resistance to the pine processionary moth. *Oecologia*, 166(3):703–711, 2011.

[44] L. S. T. Ho, J. Xu, F. W. Crawford, V. N. Minin, and M. A. Suchard. Birth/birth-death processes and their computable transition probabilities with biological applications. *J. Math. Biol.*, 76(4):911–944, 2018.

[45] A. Golightly, D. A. Henderson, and C. Sherlock. Delayed acceptance particle MCMC for exact inference in stochastic kinetic models. *Stat. Comput.*, 25:1039–1055, 2015.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[46] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Stat. Sci.*, 16:351–367, 2001.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary information

S1 SIR Model

S1.1 SIR as a Markov Jump Process

Consider an SIR model in which a population of fixed size N is classified into compartments consisting of susceptible (S), infected (I) and removed (R) individuals. Let $X_t = (S_t, I_t)'$ denote the numbers in each state at time $t \geq 0$ and note that $R_t = N - S_t - I_t$ for all $t \geq 0$. The dynamics of $\{X_t, t \geq 0\}$ can be described by a Markov jump process (MJP), that is, a continuous time, discrete valued Markov process. Assuming that at most one event can occur over an infinitesimal time interval $(t, t + \Delta t]$ and that the state of the system at time t is $x_t = (s_t, i_t)'$, the MJP is characterised by probabilities of the form

$$\mathbb{P}(X_{t+\Delta t} = (s_t - 1, i_t + 1)'|x_t, \theta) = h_1(x_t, \theta)\,\Delta t + o(\Delta t),$$
$$\mathbb{P}(X_{t+\Delta t} = (s_t, i_t - 1)'|x_t, \theta) = h_2(x_t, \theta)\,\Delta t + o(\Delta t),$$
$$\mathbb{P}(X_{t+\Delta t} = (s_t, i_t)'|x_t, \theta) = 1 - \sum_{i=1}^{2} h_i(x_t, \theta)\,\Delta t + o(\Delta t),$$

where $h(x_t, \theta) = (\beta s_t i_t, \gamma i_t)'$ is a hazard function, $\theta = (\beta, \gamma)'$ is a parameter vector containing infection and removal rates and $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$. The transition probability $\pi(x_t|x_0, \theta)$ governing the dynamics of the MJP over arbitrary time intervals of length t can be shown [37] to satisfy the chemical master equation (CME):

$$\frac{d}{dt}\pi(x_t|x_0, \theta) = \sum_{i=1}^{2} \left[h_i(x_t - S^i, \theta)\pi(x_t - S^i|x_0, \theta) - h_i(x_t, \theta)\pi(x_t|x_0, \theta)\right]$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where S^i denotes the ith column of the stoichiometry matrix

$$S = \begin{pmatrix} -1 & 0 \\ 1 & -1 \end{pmatrix}$$

which encodes the effect of each respective transition on the components of X_t. Although it is possible to evaluate $\pi(x_t|x_0, \theta)$ efficiently [44], we eschew the MJP formalism in favour of an approximation whereby X_t is modelled by a stochastic differential equation (SDE).

S1.2 SDE representation

Consider an infinitesimal time interval, $(t, t + dt]$, over which the hazard function $h(x_t, \theta)$ will remain constant almost surely. Let dN_t denote the counting process with components $dN_{1,t}$ and $dN_{2,t}$ containing the number of infections and removals over this interval. Hence $dN_{i,t}$, is Poisson distributed with rate $h_i(x_t, \theta)dt$. Upon noting that from $dX_t = SdN_t$, it should be clear that

$$\mathrm{E}(dX_t) = S\,h(x_t, \theta)dt, \qquad \mathrm{Var}(dX_t) = S\,\mathrm{diag}\{h(x_t, \theta)\}S'dt.$$

Hence, the Itô Stochastic differential equation (SDE) that best matches the MJP is given by

$$dX_t = S\,h(x_t, \theta)dt + \sqrt{S\,\mathrm{diag}\{h(x_t, \theta)\}S'}\,dW_t,$$

where $W_t = (W_{1,t}, W_{2,t})'$ is a 2-vector of standard Brownian motion and $\sqrt{S\,\mathrm{diag}\{h(x_t, \theta)\}S'}$ is a 2×2 matrix B such that $BB' = S\,\mathrm{diag}\{h(x_t, \theta)\}S'$. Explicitly, we have for the SIR model that

$$S\,h(x_t, \theta) = \begin{pmatrix} -\beta s_t i_t \\ \beta s_t i_t - \gamma i_t \end{pmatrix}, \qquad \sqrt{S\,\mathrm{diag}\{h(x_t, \theta)\}S'} = \begin{pmatrix} \beta s_t i_t & -\beta s_t i_t \\ -\beta s_t i_t & \beta s_t i_t + \gamma i_t \end{pmatrix}.$$

In Section 2.2 in the main text, we discuss extending the model to include a time varying infection rate.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

S2 Bayesian Inference under the LNA

We take the linear noise approximation (LNA) as described in Setion 2.2 of the main text as the inferential model. Although the observed data likelihood can be evaluated efficiently (with details below), the joint posterior is intractable. We therefore consider a Markov chain Monte Carlo scheme for generating samples from the posterior.

S2.1 Posterior exploration via MCMC

Our inference strategy comprises two steps:

1. Generate samples $\theta^{(1)}, \ldots, \theta^{(M)}$ from the marginal parameter posterior $\pi(\theta|y) \propto \pi(\theta)\pi(y|\theta)$.
2. Generate samples $x^{(1)}, \ldots, x^{(M)}$ by drawing from the conditional posterior $\pi(x|y, \theta^{(i)})$, $i = 1, \ldots, M$.

In step 1, we use a Metropolis-Hastings algorithm to draw from $\pi(\theta|y)$. This requires evaluation of the observed data likelihood $\pi(y|\theta)$. Note the factorisation

$$\pi(y|\theta) = \pi(y_0|\theta) \prod_{t=1}^{n} \pi(y_t|y_{0:t}, \theta). \tag{11}$$

Following [32] (see also [45]), we evaluate each constituent term in (11) via a forward filter. The forward filter requires the observation equation (main text (7)) to have a linear Gaussian structure, which is not the case as written, due to the variance being a function of the latent state process. Therefore, we run step t of the forward filter with $P'x_{t+1}P$ replaced by $P'\eta_{t+1}P$. That is, the unknown x_{t+1} is replaced by the LNA predictive mean η_{t+1}.

Since the parameters θ remain fixed throughout the calculation of $\pi(y|\theta)$, we drop them from the notation where possible. Define $y_{0:t} = (y_0, \ldots, y_t)'$. Now suppose that $X_0 \sim \mathrm{N}(\mathrm{a}_0, \mathrm{C}_0)$ *a priori*. Algorithm 1 (Supplementary Information Section S2.3) gives the forward filter. This can then be used inside Algorithm 2 (Section S2.3), which uses a random walk Metropolis algorithm to generate (dependent) draws from the marginal parameter posterior $\pi(\theta|y)$. The proposal mechanism requires an innovation variance Ω which can be chosen to maximise mixing efficiency, as measured by say

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

effective sample size per second. We take $\Omega = k\widehat{\mathrm{Var}}(\theta|y)$ where the posterior variance is estimated from a short pilot run and k is chosen subsequently to give an acceptance rate of around 20–30% [46].

Given samples $\theta^{(1)}, \ldots, \theta^{(M)}$ from $\pi(\theta|y)$, we generate samples of the latent process $x^{(i)} \sim \pi(x|y, \theta^{(i)})$, $i = 1, \ldots, M$ by noting that these samples can be efficiently generated using a backward sampling algorithm. This requires the covariance between X_t and X_{t+1} (since the former is drawn conditionally on a realisation of the latter) which depends on the 3×3 fundamental matrix G_t. This satisfies and ODE system

$$\frac{dG_t}{dt} = H_t G_t, \qquad G_0 = I_3, \tag{12}$$

where I_3 is the 3×3 identity matrix. The ODE in (12) can be time-stepped with the ODEs in (4) and (6) (main text) as part of the forward filter. Algorithm 3 (Section S2.3) then gives the backward sampler.

S2.2 Model checking

Given inferences on the static parameters θ and the latent dynamic process x, we consider the following diagnostics for assessing model fit. The within sample predictive density is

$$\pi(\tilde{y}|y) = \int\int \pi(\tilde{y}|x, \theta)\pi(\theta, x|y)dxd\theta \tag{13}$$

and the one step ahead out of sample predictive density is

$$\pi(y_{n+1}|y) = \int\int \pi(y_{n+1}|x_{n+1}, \theta)\pi(x_{n+1}|x_n, \theta)\pi(\theta, x|y)dx_{0:n+1}d\theta. \tag{14}$$

Hence, in both cases we properly account for parameter and latent process uncertainty. Although the densities in (13) and (14) will be intractable, we may generate samples via Monte Carlo. Recall that the inference algorithm described in Section S2.1 gives draws $\{(\theta^{(i)}, x^{(i)}), i = 1, \ldots, M\}$. We then generate $\{\tilde{y}^{(i)}, i = 1, \ldots, M\}$ via (7), by drawing $\tilde{y_t}^{(i)}$ from a $N(P'x_t^{(i)}, (\sigma_e^2)^{(i)}P'x_t^{(i)}P)$ distribution, independently for $t = 0, \ldots, n$ and $i = 1, \ldots, M$. Similarly, we generate samples from $\pi(y_{n+1}|y)$ by first drawing $x_{n+1}^{(i)}$ from $\pi(x_{n+1}|x_n^{(i)}, \theta^{(i)})$, followed by $y_{n+1}^{(i)}$ from $\pi(y_{n+1}|x_{n+1}^{(i)}, \theta^{(i)})$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

S2.3 Algorithms

Algorithm 1 LNA forward filter

1. Initialisation. Compute $\pi(y_0) = \mathrm{N}(y_0\,;\, P'a_0\,,\, P'C_0P + \Sigma_0)$ where $\Sigma_0 = \sigma_e^2 P'a_0P$. The posterior at time $t = 0$ is therefore $X_0|y_0 \sim \mathrm{N}(a_1, C_1)$, where

$$a_1 = a_0 + C_0P\left(P'C_0P + \Sigma_0\right)^{-1}\left(y_0 - P'a_0\right)$$
$$C_1 = C_0 - C_0P\left(P'C_0P + \Sigma_0\right)^{-1}P'C_0\,.$$

   Store the values of a_1 and C_1.

2. For $t = 0, 1, \ldots, n-1$,

   (a) Prior at $t+1$. Initialise the LNA with $\eta_t = a_t$ and $V_t = C_t$. Integrate the ODEs (4) and (6) forward to $t+1$ to obtain η_{t+1} and V_{t+1}.

   (b) One step forecast. Using the observation equation (7), we have that

$$Y_{t+1}|y_{0:t} \sim \mathrm{N}\left(P'\eta_{t+1}, P'V_{t+1}P + \Sigma_{t+1}\right)$$

   where $\Sigma_{t+1} = \sigma_e^2 P'\eta_{t+1}P$. Compute the updated marginal likelihood

$$\begin{aligned}\pi(y_{0:t+1}) &= \pi(y_{0:t})\pi(y_{t+1}|y_{0:t})\\ &= \pi(y_{0:t}) \times \mathrm{N}\left(y_{t+1}\,;\, P'\eta_{t+1}\,,\, P'V_{t+1}P + \Sigma_{t+1}\right).\end{aligned}$$

   (c) Posterior at $t+1$. Combining the distributions in (a) and (b) gives the joint distribution of X_{t+1} and Y_{t+1} (conditional on $y_{0:t}$) as

$$\begin{pmatrix} X_{t+1} \\ Y_{t+1} \end{pmatrix} \sim \mathrm{N}\left\{\begin{pmatrix} \eta_{t+1} \\ P'\eta_{t+1} \end{pmatrix}, \begin{pmatrix} V_{t+1} & V_{t+1}P \\ P'V_{t+1} & P'V_{t+1}P + \Sigma_{t+1} \end{pmatrix}\right\}$$

   and therefore $X_{t+1}|y_{0:t+1} \sim \mathrm{N}(a_{t+1}, C_{t+1})$, where

$$a_{t+1} = \eta_{t+1} + V_{t+1}P\left(P'V_{t+1}P + \Sigma_{t+1}\right)^{-1}\left(y_{t+1} - P'\eta_{t+1}\right)$$
$$C_{t+1} = V_{t+1} - V_{t+1}P\left(P'V_{t+1}P + \Sigma_{t+1}\right)^{-1}P'V_{t+1}\,.$$

   Store the values of a_{t+1}, C_{t+1}, η_{t+1} and V_{t+1}.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Algorithm 2 Random walk Metropolis algorithm

1. Initialise at $\theta^{(0)}$ in the support of $\pi(\theta|y)$. Set the iteration counter $i = 1$.
2. Propose $\theta^* = \theta^{(i-1)} + \epsilon_i$ where $\epsilon_i \sim \mathrm{N}(0, \Omega)$
3. With probability
$$\alpha(\theta^*|\theta^{(i-1)}) = \min\left\{1, \frac{\pi(\theta^*)\pi(y|\theta^*)}{\pi(\theta^{(i-1)})\pi(y|\theta^{(i-1)})}\right\}$$
put $\theta^{(i)} = \theta^*$ otherwise put $\theta^{(i)} = \theta^{(i-1)}$.
4. If $i = M$ stop otherwise put $i := i + 1$ and go to step 2.

Algorithm 3 LNA backward sampler

1. First draw x_n from $X_n|y \sim \mathrm{N}(a_n, C_n)$.
2. For $t = n-1, n-2, \ldots, 0$,
   (a) Joint distribution of X_t and X_{t+1}. Note that $X_t|y_{1:t} \sim \mathrm{N}(a_t, C_t)$. The joint distribution of X_t and X_{t+1} (conditional on $y_{1:t}$) is
$$\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix} \sim \mathrm{N}\left\{\begin{pmatrix} a_t \\ \eta_{t+1} \end{pmatrix}, \begin{pmatrix} C_t & C_t G'_{t+1} \\ G_{t+1} C_t & V_{t+1} \end{pmatrix}\right\}.$$
   (b) Backward distribution. The distribution of $X_t|X_{t+1}, y_{0:t}$ is $\mathrm{N}(\hat{a}_t, \hat{C}_t)$, where
$$\hat{a}_t = a_t + C_t G'_{t+1} V_{t+1}^{-1} (x_{t+1} - \eta_{t+1}),$$
$$\hat{C}_t = C_t - C_t G'_{t+1} V_{t+1}^{-1} G_{t+1} C_t.$$
   Draw x_t from $X_t|X_{t+1}, y_{0:t} \sim \mathrm{N}(\hat{a}_t, \hat{C}_t)$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

S2.4 Parameters for inference

		Bushy	Richmond
Total population	N	5000	40000
Initial infected	I_0	240	1400
Initial infection rate	$\tilde{\beta}_0$	-8.5	-10
Initial model parameters	$\theta_0 = (\gamma_0, \sigma_0, \sigma_e)'$	$(1, 0.5, 1)'$	$(2, 0.5, 1)'$
Observation matrix	P	(1, 1, 0)	
Prior distributions	$\pi(\theta)$	$\log\gamma \sim \mathrm{N}(0,\ 0.5^2)$ $\log\sigma \sim \mathrm{N}(1,\ 1)$ $\log\sigma_e \sim \mathrm{N}(1,\ 1)$	
Tuning parameter	Σ	$\begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	
Initial conditions mean	a_0	$(N - I_0 - R_0,\ I_0,\ \tilde{\beta}_0)$	
Initial conditions variance	C_0	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0.5 \end{pmatrix}$	

Table S1: The parameters used in the inference schemes described in Sections 2.3 and S2.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

S3 Additional results figures

The following section contains additional results figures referred to in the main text. We give the trace (Figure S1) and 2D contour plots (Figure S2) for the inferred model parameters. The one-step model predictions for all years are given in Figure S3

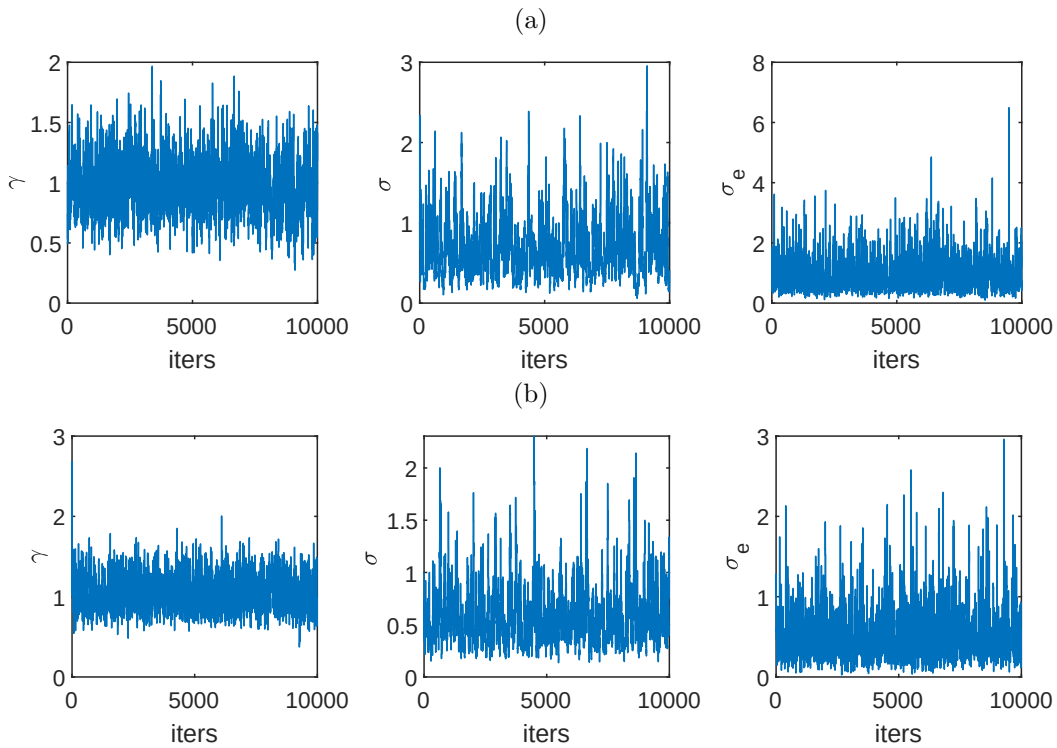


Figure S1: Trace plots for (a) Bushy and (b) Richmond Park.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

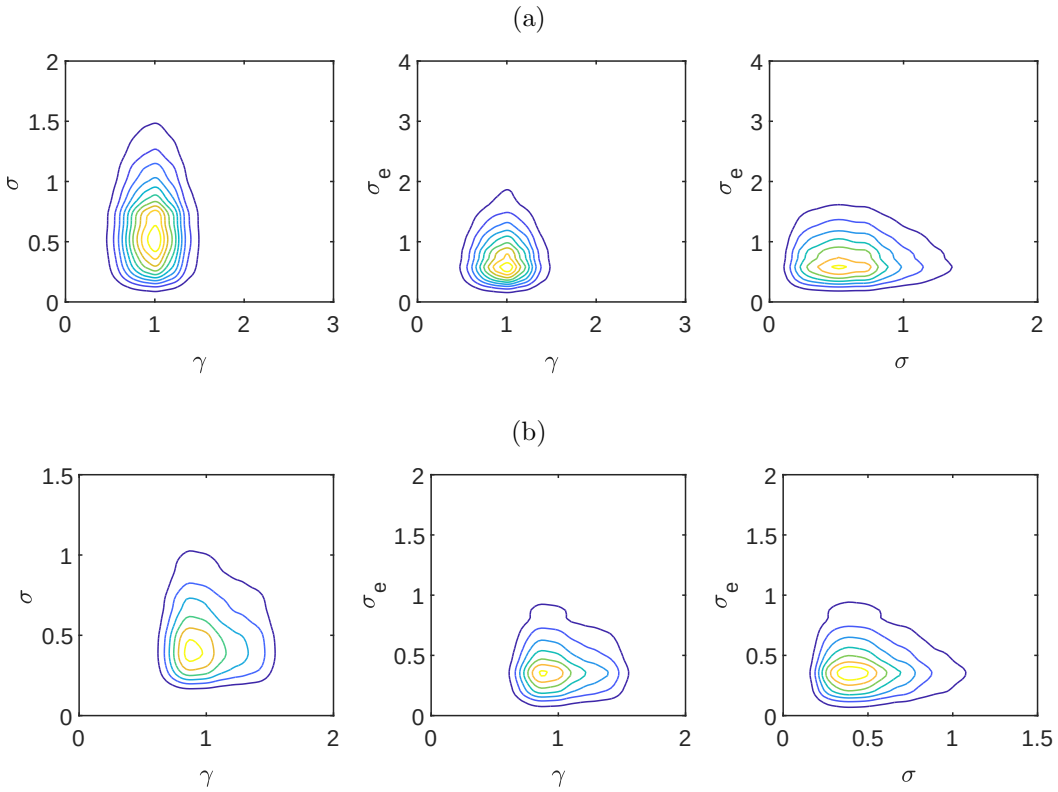


Figure S2: Pairwise joint posterior densities for the parameters γ, σ and σ_e for (a) Bushy and (b) Richmond Park.

bioRxiv preprint doi: https://doi.org/10.1101/2021.12.09.471950; this version posted December 12, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

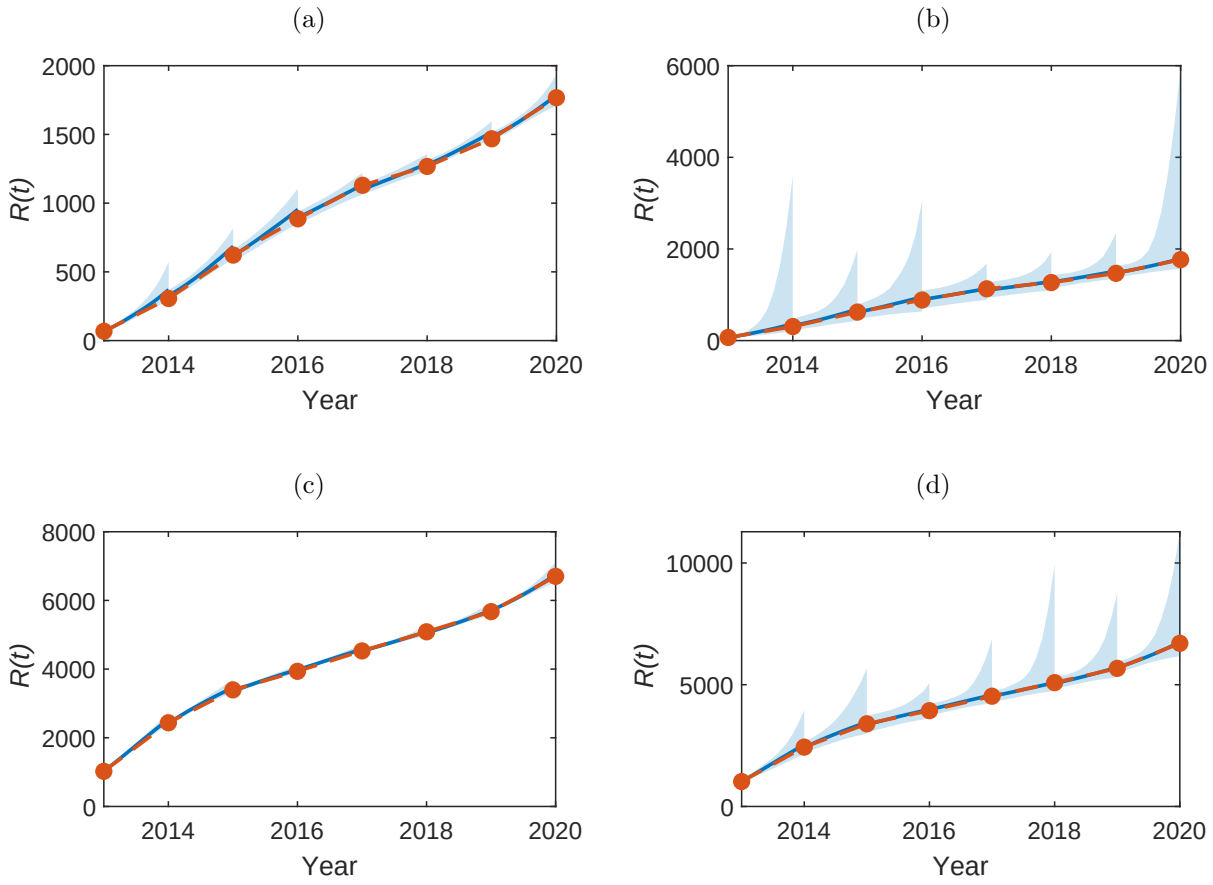


Figure S3: One-step model predictions for the number of removed nests, $R(t)$ for the years 2014–2020, with median (blue line) for (a-b) Bushy and (c-d) Richmond. In (a) and (c) the shaded area shows 50% credible region and in (b) and (d) the 95% credible region. The orange line shows the observed data.