Journal of Theoretical Biology 558 (2023) 111351

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

A Bayesian nonparametric method for detecting rapid changes in disease transmission

Richard Creswell [a,*], Martin Robinson [a], David Gavaghan [a], Kris V. Parag [b,c], Chon Lok Lei [d], Ben Lambert [e,*]

[a] Department of Computer Science, University of Oxford, Oxford, United Kingdom
[b] MRC Centre of Global Infectious Disease Analysis, Jameel Institute for Disease and Emergency Analytics, Imperial College London, London, United Kingdom
[c] NIHR Health Protection Research Unit in Behavioural Science and Evaluation, University of Bristol, Bristol, United Kingdom
[d] Institute of Translational Medicine, Faculty of Health Sciences, University of Macau, Macau, Macao Special Administrative Region of China
[e] College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, United Kingdom

ARTICLE INFO

Keywords:
Reproduction number
Bayesian nonparametrics
Outbreaks
Epidemiology
COVID-19
Changepoint detection

ABSTRACT

Whether an outbreak of infectious disease is likely to grow or dissipate is determined through the time-varying reproduction number, R_t. Real-time or retrospective identification of changes in R_t following the imposition or relaxation of interventions can thus contribute important evidence about disease transmission dynamics which can inform policymaking. Here, we present a method for estimating shifts in R_t within a renewal model framework. Our method, which we call EpiCluster, is a Bayesian nonparametric model based on the Pitman–Yor process. We assume that R_t is piecewise-constant, and the incidence data and priors determine when or whether R_t should change and how many times it should do so throughout the series. We also introduce a prior which induces sparsity over the number of changepoints. Being Bayesian, our approach yields a measure of uncertainty in R_t and its changepoints. EpiCluster is fast, straightforward to use, and we demonstrate that it provides automated detection of rapid changes in transmission, either in real-time or retrospectively, for synthetic data series where the R_t profile is known. We illustrate the practical utility of our method by fitting it to case data of outbreaks of COVID-19 in Australia and Hong Kong, where it finds changepoints coinciding with the imposition of non-pharmaceutical interventions. Bayesian nonparametric methods, such as ours, allow the volume and complexity of the data to dictate the number of parameters required to approximate the process and should find wide application in epidemiology. This manuscript was submitted as part of a theme issue on "Modelling COVID-19 and Preparedness for Future Pandemics".

1. Introduction

Throughout the SARS-CoV-2 pandemic, the *time-varying* reproduction number,[1] R_t, has been estimated and used to gauge the effectiveness of control measures (e.g. Flaxman et al., 2020; Li et al., 2021; Parag et al., 2021; Brauner et al., 2021; meta-analysis of such studies: Mendez-Brito et al., 2021). R_t represents the average number of secondary cases spawned by a single primary case. When $R_t > 1$, an outbreak is expected to grow exponentially; public health interventions often attempt to permanently shift $R_t < 1$ meaning an epidemic will, in the long run, die out.

A widely used approach for estimating R_t is through *renewal equations* which assume that future numbers of cases depend on the history of case counts, the generation times, representing the typical timescales between primary and secondary infections, and R_t (theory: Fraser (2007) and Nishiura and Chowell (2009); example software: Thompson et al. (2019)). These models are typically formulated in discrete time (usually at the daily resolution), and the dynamics are assumed stochastic. Here, we focus on the most popular version of these models which assume that the population is well-mixed and that there is no demographic heterogeneity.

A variety of approaches exist for estimating R_t using time series incidence data, either in real-time (i.e. using only information up until a current time t; Cori et al., 2013; Parag, 2021) or retrospectively (Wallinga and Teunis, 2004). These approaches make diverse assumptions about the continuous structure of R_t: that it is piecewise-constant within a sliding window of a given prespecified length (Wallinga and Teunis, 2004; Thompson et al., 2019); that it varies smoothly with the variation controlled by a Gaussian filter

* Corresponding authors.
E-mail addresses: richard.creswell@hertford.ox.ac.uk (R. Creswell), ben.c.lambert@gmail.com (B. Lambert).

[1] Also known as the *effective reproduction number*.

https://doi.org/10.1016/j.jtbi.2022.111351
Received 1 July 2022; Received in revised form 11 October 2022; Accepted 1 November 2022
Available online 13 November 2022

(Abbott et al., 2020; Parag, 2021); or that it is made up of an inferred number of pieces with a single, optimal, number of pieces inferred by considering a criterion derived from information theory (Parag and Donnelly, 2020).

Our approach also assumes that R_t is piecewise-constant, and that, within each piece, the epidemic follows a standard Poisson renewal process (Cori et al., 2013). We do not specify the number of pieces nor provide a limit on this number *a priori*. To do so, we use a Bayesian model with a *Pitman–Yor process* prior (Pitman and Yor, 1997) to represent the values of R_t across any feasible number of pieces. This process comes from the field of Bayesian nonparametrics—a broad class of models where the data are modelled by a (potentially) countably infinite set of parameters, where the complexity of the models, indexed by the number of parameters, increases in lockstep with the volume and complexity of the data (Ghahramani, 2013). Our approach, which we call *EpiCluster*, avoids the need to directly specify how often and how fast R_t need change to represent a given incidence curve. Instead, the data and a prior jointly determine how many pieces are needed to approximate the R_t curve, and, in Section 2, we introduce a default prior meant to find a parsimonious approximation of it with few changepoints. Our method, being Bayesian, provides a measure of uncertainty in both the number of pieces and R_t (see Fig. 1). We develop an efficient Markov chain Monte Carlo (MCMC) inference method for fitting our model to incidence data using collapsed Gibbs sampling (Lambert, 2018, Chapter 14), which efficiently steps between models of different dimensionalities (corresponding to different numbers of R_t pieces). We provide an open-source Python package implementing EpiCluster, which computes R_t profiles and runs in seconds to minutes (dependent on the length of data series and complexity of the R_t profile), which is available at github.com/SABS-R3-Epidemiology/epicluster.

By fitting our model to simulated data with known R_t profiles (in Section 3.1), we show that EpiCluster is adept at identifying times of rapid change in R_t as may occur following the imposition of major and broad-scale interventions (Dehning et al., 2020; Flaxman et al., 2020; Brauner et al., 2021)—either in real-time or retrospectively. It is less well suited to estimate R_t if it changes more gradually, and more appropriate methods exist for this purpose (e.g. Thompson et al., 2019; Parag, 2021). Unlike methods which directly model R_t as a function of known intervention timings and severities (e.g. Dehning et al., 2020; Flaxman et al., 2020; Brauner et al., 2021), our method is purely driven by the incidence series. Because of this, it provides a straightforward and intervention-agnostic initial step for assessing the impact of interventions, and similarly agnostic approaches have previously been used in retrospective analyses of COVID-19 transmission (Parag et al., 2021). Since it does not use additional information about interventions, our approach is likely to produce estimates with greater variability. But, it requires fewer assumptions to be made, which may be beneficial, since the assumptions around intervention timing (Soltesz et al., 2020) and modelling details (Sharma et al., 2020) may affect estimates and their interpretation. In Section 3.5, we apply our framework to data from the COVID-19 outbreaks in Australia and Hong Kong and show that it is able to find changepoints in R_t corresponding to the imposition of known interventions. Our method provides a tool for outbreak analysis complementary to existing methods and could form part of an analysis pipeline for associating interventions with changes in transmission.

2. Methods

2.1. Renewal process model

We estimate *instantaneous reproduction numbers* and mean this whenever we write R_t. Instantaneous reproduction numbers represent the average number of secondary cases that would be generated by an infected case at time t assuming that future transmission remains the same as at time t (Fraser, 2007). We assume that the data consist of

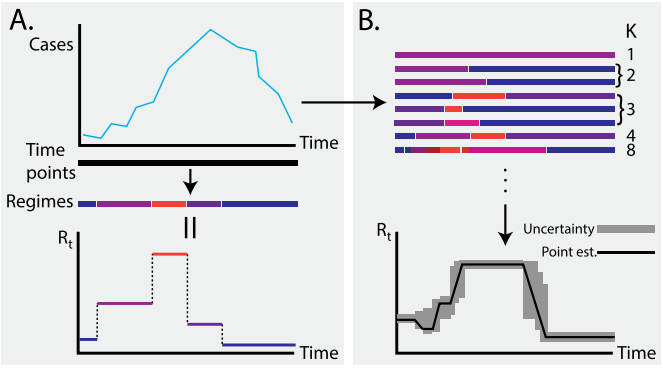

Fig. 1. Pitman–Yor based inference for the time-varying reproduction number. Panel A represents the intrinsic assumption underpinning our method: that R_t is piecewise-constant, and the pieces are shown as different coloured bars. Panel B shows how our nonparametric prior allows a decomposition of the time points into partitions comprising different numbers of pieces (K). Our MCMC sampler (see Algorithm 1) explores this space over partitions efficiently, resulting in posterior uncertainty in R_t. Whilst the individual samples of R_t trajectories are piecewise-constant, the average over many such trajectories is likely not to be, which is illustrated by the black point estimate line in panel B.

a series of daily case counts[2] for each day, t, from $t = 1$ to $t = T$: $\{I_t\}_{t=1}^{T}$ and that the case counts are perfectly known. Due to the within- and between-individual variability in rates of contact and infectivity, a *generation time distribution* is used to represent the duration between the time at which a parent case occurs and its offspring. We model the case count I_t as arising according to the Poisson renewal process:

$$I_t \sim \text{Pois}\left(R_t \Lambda_t\right), \text{ where } \Lambda_t = \sum_{s=1}^{t-1} w_s I_{t-s}, \tag{1}$$

where $R_t \geq 0$ is the time-varying reproduction number on day t, and $\Lambda_t \geq 0$ is the transmission potential. The w_s terms represent the generation time distribution: $0 \leq \omega_s \leq 1$ indicates the probability that a primary case takes between $s-1$ and s days to generate a secondary case, and $\sum_{s=1}^{\infty} \omega_s = 1$. Since it is not typically possible to know when individuals become infectious, generation times are not directly observed, making it difficult to estimate the generation time distribution. Here, we use the *serial interval distribution* in its place, which describes the time between the onset of symptoms between a primary and secondary cases. This is easier to estimate from infector–infectee pairs, since it is more directly observable and has a similar mean (Svensson, 2007), (although we recognize that it is possible to estimate a generation time distribution with the same data used to estimate serial intervals, by making assumptions about the duration of the incubation period; Hart et al., 2021).

2.2. Model of changing R_t

2.2.1. Exchangeable partition probability functions and the Pitman-Yor process

Here, we assume that the R_t profile can be decomposed into a number of regimes within which R_t is constant. Our goal is to avoid prespecifying the location of changepoints—representing the boundary between two different R_t regimes—nor their count, since these choices can bias analyses, but rather to learn an appropriate configuration of the time points into regimes using Bayesian inference. We develop a probabilistic model of the division of the time points into regimes.

[2] Technically, the renewal equation is formulated in terms of infections rather than cases, but, since we use the serial interval distribution in place of the generation time distribution, we keep with defining I_t as a case count.

To do so, we use a Pitman–Yor process (Pitman and Yor, 1997)[3] to account for a probabilistic decomposition of data points into clusters and, following (Martínez and Mena, 2014), we adjust this model to account for the time series nature of our data. The remainder of this subsection serves as a brief review of this model, starting with a treatment of the nonparametric clustering of unordered data points via *exchangeable partition probability functions* (EPPFs) and followed by appropriate modifications for the time series case (see Section 2.2.2).

In the standard clustering problem, we have a set $[T] = \{1, \dots, T\}$ (i.e., the labels of T data points), which we would like to divide into K mutually exclusive subsets $\{A_1, \dots, A_K\}$ such that $\cup_k A_k = [T]$ where none of the A_k are empty. We denote the set of all such groupings by $\mathcal{P}_{[T]}$; each element of $\mathcal{P}_{[T]}$ is called a *partition*. Random variables Π_T taking values in $\mathcal{P}_{[T]}$ are termed *random partitions* of $[T]$. A random partition has the property of *exchangeability* if its probability distribution can be written as a symmetric function p of the subset sizes, i.e.,

$$\text{Prob}(\Pi_T = \{A_1, \dots, A_K\}) = p(n_1, \dots, n_K)$$

where $n_k = |A_k|$ (i.e. n_k is the size of the subset, A_k).

Under these conditions p is known as an EPPF. A more complete treatment of the concept of EPPFs can be found in Pitman (2002) and Lijoi and Prunster (2010). A fairly general EPPF, which we will employ in this work, is derived from the Pitman–Yor process, a generalization of the Dirichlet process (Teh, 2010). This EPPF is given by (Pitman, 2002, eq. (3.6)):

$$p(n_1, \dots, n_K|\theta, \sigma) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{T-1\uparrow}} \prod_{j=1}^{K} (1-\sigma)_{n_j - 1\uparrow}, \tag{2}$$

where $x_{m\uparrow} := \prod_{j=0}^{m-1}(x+j)$, and $\sigma \in [0, 1)$ and $\theta > -\sigma$ are the two hyperparameters governing the process: σ is called the discount parameter, which essentially controls how the number of regimes, K, grows with the size of the dataset; θ is called the strength parameter with larger values giving greater weight to series with more regimes. In the limit $\sigma \to 0$, a Pitman–Yor process becomes a Dirichlet process which permits a slower growth (of order $\log T$ opposed to T^σ; Pitman, 2002, section 3.3) in the number of regimes with increases in data size.

2.2.2. Applicability of EPPFs to time series problems

Unlike the general clustering problem, in the time series case, the data points have an ordering which the clusters must respect. For example, consider an incidence series of length three: (I_1, I_2, I_3). For this series, allowable effective reproduction number allocations include: $\{\{I_1, I_2, I_3\}\}$, where all the data points are generated from a process with the same effective reproduction number: i.e. there is a single regime ($K = 1$); $\{\{I_1\}, \{I_2, I_3\}\}$, where the first data point was generated from a process with one effective reproduction number and the latter two data points from a process with a different one: i.e. there are two regimes ($K = 2$); $\{\{I_1, I_2\}, \{I_3\}\}$, where the first two points are grouped; and $\{\{I_1\}, \{I_2\}, \{I_3\}\}$, where each data point is generated from a process with a different reproduction number: i.e. there are three regimes ($K = 3$).

An allocation which would be disallowed is: $\{\{I_1, I_3\}, \{I_2\}\}$, where the first and third data points come from the same process which is distinct from that governing the second. Whilst, it is possible that transmission could return to a previous level, it is an assumption of our modelling process that only consecutive data points share the same R_t. By avoiding recurrence to historical regimes, we ensure that the changepoints identified are straightforward to interpret.

For a given EPPF, p', we can obtain a distribution p which is supported only on those partitions which respect an ordering of the labels using the following result (Martínez and Mena, 2014):

$$p(n_1, \dots, n_K) = \begin{cases} \frac{1}{K!}\binom{T}{n_1, \dots, n_K} p'(n_1, \dots, n_K), & \text{if allowable partitioning} \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where the large bracketed term indicates the multinomial coefficient.

Combining Eqs. (2) and (3), we obtain the following result for the prior distribution on the sequence of regime sizes in the time series case:

$$p(n_1, \dots, n_K|\theta, \sigma) = \frac{T!}{K!} \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{T-1\uparrow}} \prod_{j=1}^{K} \frac{(1-\sigma)_{n_j - 1\uparrow}}{n_j!}. \tag{4}$$

2.2.3. Hyperparameters of the process

In order to learn parsimonious assignments of the time points into regimes, our prior, given by Eq. (4), should favour configurations consisting of longer regimes. We favour longer regimes because they mitigate against overfitting—for typical data, the likelihood of the renewal process would be maximized by assigning each time point to its own cluster with an idiosyncratic value of R_t; the resulting profile of R_t values will tend to be jagged and exhibit spurious fluctuations. Additionally, longer regimes have the advantage of allowing more data to be leveraged in order to learn more precise estimates of R_t. However, by favouring longer regimes, it is possible that we miss shorter term fluctuations in R_t—this is akin to the issue of choosing window lengths for a number of existing methods (e.g. Thompson et al., 2019).

Eq. (4) induces a marginal distribution over the number of clusters whose mean has been derived as (Pitman, 2002, eq. (3.13)):

$$\mathbb{E}[K] = \frac{(\theta+\sigma)_{T\uparrow}}{\sigma(\theta+1)_{T-1\uparrow}} - \frac{\theta}{\sigma}, \tag{5}$$

for $\sigma \neq 0$. For small values of the hyperparameters θ and σ, $\mathbb{E}[K]$ is significantly smaller than the number of time points T (see Fig. S1), and the marginal distribution of K places little weight on values of K close to T, thus preferring sparsity in the number of clusters. For all results presented in this paper, we set $\theta = 0$ and choose σ as a function of T such that $E[K] = 1.5$ (with the appropriate value of σ selected by numerical optimization of Eq. (5)); this represents a prior belief that R_t is generally constant over the time series, but allows flexibility to add clusters when the data provides evidence that they are needed. For a time series of length $T = 100$, our choice of prior hyperparameters induces a marginal distribution over the number of clusters whose 2.5th percentile is 1 cluster and 97.5th percentile is 4 clusters.

2.3. Marginal likelihood of the data

In this subsection, we calculate the marginal likelihood of the data conditional on a particular arrangement of the time points into regimes, which involves integrating out R_t with respect to its prior distribution. This marginal likelihood enables efficient inference for the posterior distribution over regime configurations via collapsed Gibbs sampling (see Section 2.4).

The marginal likelihood for an incidence series conditional on a particular set of subset sizes $n_1, \dots, n_K$ (see Section 2.2) can be written as a product of marginal likelihoods for each regime:

$$p(I_1, \dots, I_T|n_1, \dots, n_K) = \prod_{k=1}^{K} \mathcal{L}_k(I_{k,1}, \dots, I_{k,n_k}|I_{-k}), \tag{6}$$

where $I_{k,j}$ denotes the jth data point in regime k, and $\mathcal{L}_k$ is the marginal likelihood of the data in the kth regime, which we assume is conditional on all cases observed prior to regime k (denoted by I_{-k}). We derive the regime-specific marginal likelihoods using the renewal model (Eq. (1)):

$$\mathcal{L}_k(I_{k,1}, \dots, I_{k,n_k}|I_{-k}) = \int_0^\infty p(R_k) \prod_{j=1}^{n_k} \text{Pois}(I_{k,j}|R_k \Lambda_{k,j}) dR_k, \tag{7}$$

[3] Also known as the two-parameter Poisson–Dirichlet process.

where $\Lambda_{k,j}$ is the transmission potential calculated for the jth time point in regime k, R_k is the value of the effective reproduction number for the kth regime, and $p(R_k)$ is the prior on R_k.

We choose a gamma distribution prior for R_k with shape parameter α and rate parameter β.[4] With this choice of prior, the integral in Eq. Eq. (7) can be evaluated analytically, resulting in:

$$\mathcal{L}_k(I_{k,1},\dots,I_{k,n_k}|I_{-k}) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\Gamma\left(\alpha+\sum_{j=1}^{n_k} I_{k,j}\right)\left(\beta+\sum_{j=1}^{n_k}\Lambda_{k,j}\right)^{-\alpha+\sum_{j=1}^{n_k} I_{k,j}} \times \prod_{j=1}^{n_k}\frac{\Lambda_{k,j}^{I_{k,j}}}{I_{k,j}!},$$

where $\Gamma(\cdot)$ is the gamma function.

Additionally, with the gamma prior on R_k, the posterior distribution of each R_k, conditional on the data assigned to regime k, is given by the conjugate gamma posterior (Creswell et al., 2022):

$$p(R_k|I_{k,1},\dots,I_{k,n_k},I_{-k}) = \text{gamma}(R_k|\text{shape}=\alpha+\sum_{j=1}^{n_k} I_{k,j},\text{rate}=\beta+\sum_{j=1}^{n_k}\Lambda_{k,j}). \tag{8}$$

As prior hyperparameters, we select $\alpha = 1$ and $\beta = 0.2$. With this choice, the prior mean and standard deviation are both equal to 5. The high standard deviation provides a relatively uninformative prior, and the high mean ensures that the outbreak is unlikely to be determined as under control (since >81% of prior probability is for $R_t > 1$) unless there is considerable evidence to suggest otherwise.

2.4. Inference

At particular values of the hyperparameters σ and θ, the target posterior of regime configurations is proportional to the product of and Eqs. (4) and (6):

$$p(n_1,\dots,n_K|I_1,\dots,I_T,\sigma,\theta) \propto p(I_1,\dots,I_T|n_1,\dots,n_K)\times p(n_1,\dots,n_K|\theta,\sigma).$$

For brevity, we suppress the dependence on cases and hyperparameters and denote the unnormalized posterior by $p(\gamma_K)$, where $\gamma_K := (n_1,\dots,n_K)$ indicates a particular configuration of the time points into K regimes.

Inference for this posterior is performed via Markov Chain Monte Carlo (MCMC) which provides a distribution over the number of regimes by jumping between models of different numbers of parameters. We use the same split-merge-shuffle structure as Martínez and Mena (2014). Each step of our MCMC algorithm is given in Algorithm 1, and we now describe it.

Different configurations of the time points into regimes are explored through the use of *split*, *merge*, and *shuffle* proposals. The split proposal takes an existing regime and proposes to split it into two regimes at some randomly located changepoint. The merge proposal takes two consecutive regimes and proposes to merge them into one. Both of these proposals consider an update to the total number of regimes, thus allowing the sampler to explore the marginal posterior distribution over the number of regimes. Additionally, the shuffle proposal shifts the boundary between two consecutive regimes, thus keeping the same number of regimes but efficiently exploring uncertainty in the location of a changepoint. At each iteration of the MCMC sampler, we make one shuffle proposal and randomly choose whether to make a split or merge proposal, with the MCMC tuning parameter q giving the probability of making the split proposal. For the results presented in this paper, we fix $q = 0.5$. The acceptance probabilities for the split, merge, and shuffle proposals are derived in Martínez and Mena (2014) and are given by $\min(1,\alpha_e)$, with $e \in \{\text{split},\text{merge},\text{shuffle}\}$.

[4] $p(R|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}R^{\alpha-1}e^{-\beta R}$.

α_{split} is calculated by:

$$\alpha_{\text{split}} = \begin{cases} (1-q)(T-1)\frac{p(\gamma_{K+1})}{p(\gamma_K)}, & \text{if } K=1,\\ \frac{1-q}{q}\frac{p(\gamma_{K+1})}{p(\gamma_K)}\frac{n_{\text{splittable}}(n_s-1)}{K}, & \text{if } K>1,\end{cases}$$

where $n_{\text{splittable}}$ is the number of splittable regimes (i.e., those with more than one time point assigned to them) in the original configuration, and n_s is the length of the regime selected for a split; γ_K is the current regime configuration, and γ_{K+1} is the split configuration.

The corresponding quantity for a merge move is given by:

$$\alpha_{\text{merge}} = \begin{cases} \frac{q}{1-q}\frac{p(\gamma_{K-1})}{p(\gamma_K)}\frac{K-1}{n^*_{\text{splittable}}(n_s+n_{s+1}-1)}, & \text{if } K<T,\\ q(T-1)\frac{p(\gamma_{K-1})}{p(\gamma_K)}, & \text{if } K=T,\end{cases}$$

where $n^*_{\text{splittable}}$ is the number of splittable regimes in the proposed configuration, and n_s and n_{s+1} are the sizes of the regimes which are proposed to be merged; γ_{K-1} is the merged regime configuration.

The equivalent quantity for a shuffle move is given by:

$$\alpha_{\text{shuffle}} = \frac{p(\gamma^*_K)}{p(\gamma_K)},$$

where γ^*_K is the shuffled configuration obtained from γ_K as described in Algorithm 1.

The values of R_t are updated using Gibbs steps conditional on the current regime configuration.

We run four separate MCMC chains, two initialized with all time points assigned to a single regime (i.e. $K = 1$) and the other two initialized with all time points assigned to their own singleton regime (i.e. with $K = T$). We assessed convergence of our MCMC algorithm (Algorithm 1) by monitoring convergence in K, the number of regimes. To do so, we computed the $\hat{R}$ statistic (Gelman and Rubin, 1992) and required $\hat{R} < 1.05$. Once convergence was determined, we discarded the first 50% of each of the MCMC chains as warm-up and combined the rest of the samples in order to calculate posterior percentiles and means.

Algorithm 1 One step of the MCMC sampler.

$q \leftarrow$ User specified value between 0 and 1 (MCMC tuning parameter)
$K \leftarrow$ Current number of regimes
for k in $1,\dots,K$ **do** ▷ *Update the R_t via Gibbs steps.*
  Draw a value for R_t in the kth regime from its conditional posterior, Eq. (8).
end for
if $K = 1$ **then**
  $q \leftarrow 1$
else if $K = T$ **then**
  $q \leftarrow 0$
end if
$Sp \sim$ Bernoulli(q) ▷ *Draw binary variable to allow random choice between split and merge proposals.*
if $Sp = 1$ **then** ▷ *Perform a split proposal.*
  Uniformly at random propose a regime to split.
  Uniformly at random propose an index within that regime at which to split.
  Accept the split regime configuration with probability α_{split}.
else ▷ *Perform a merge proposal.*
  Uniformly at random propose a regime (not the last) which will be merged with following regime.
  Accept the merged regime configuration with probability α_{merge}.
end if
$K \leftarrow$ Current number of regimes
if $K > 1$ **then** ▷ *Perform a shuffle proposal.*
  Uniformly at random propose a regime j (not the last) to shuffle.
  Uniformly at random propose an index within either regime j or $j+1$ to be the new changepoint between these two regimes.
  Accept the shuffled regime configuration with probability α_{shuffle}.
end if

2.5. Comparator methods

In Section 3, we compare the posterior distribution for R_t obtained by our nonparametric method to those yielded by two comparator methods. This first is the Cori sliding window method (Cori et al., 2013; Thompson et al., 2019), which assumes that R_t is constant over a sliding window of τ days looking backwards. The sliding window width has a significant effect on the posterior and the effective bias–variance trade-off. As a result, we consider two choices of τ (7 days and 28 days) when applying the method to synthetic data. The second comparator is the EpiFilter method (Parag, 2021), which applies sequential Bayesian smoothing and controls changes in R_t through a random walk prior.

2.6. Implementation and runtime

We implemented EpiCluster in Python 3. A Python package of the model, including the MCMC inference algorithm, is available at github.com/SABS-R3-Epidemiology/epicluster, while the notebooks and data for reproducing all results in this paper are available at github.com/SABS-R3-Epidemiology/epicluster-results. We ran the sliding window method using the **branchpro** Python package (Creswell et al., 2022). We ran the EpiFilter method through R code made available through Parag (2021). Using our software library and typical consumer hardware (3.6 GHz CPU), EpiCluster takes from several seconds to several minutes to learn the posterior, depending on the complexity of the R_t profile. By comparison, the sliding window method and EpiFilter methods are effectively instantaneous to compute on the time series studied here.

2.7. Handling imported cases

Some of the real data examples we consider (see Section 3.5) consist of case counts in locations where a substantial proportion of the case loads are due to imported cases. To account for this, we adapt our renewal model using the methods described in Creswell et al. (2022). In this approach, cases are classified as either *local* or *imported*. Local cases $\{I_t\}_{t=1}^{T}$ are those arising from local transmission in the spatial region under consideration, while imported cases $\{I_t^{\text{imp}}\}_{t=1}^{T}$ are those who were infected elsewhere before travelling to the region. Thus, imported cases contribute to local transmission, but did not arise from it. In outbreaks where a significant proportion of cases are imported, distinguishing local from imported cases is important for accurate estimation of R_t (Roberts and Nishiura, 2011; Thompson et al., 2019). We allow local and imported cases to have different risks of onwards transmission by weighting the imported cases by a number $\epsilon > 0$ (Creswell et al., 2022), and we set ϵ to appropriate values (see Section 2.8). The default choice of $\epsilon = 1$ corresponds to an equal risk of onwards transmission between local and imported cases. Note, any case and any subsequent lineages begot by an imported case are classified as local: it is only the rate at which newly imported cases infect others which is assumed to differ from purely local transmission.

We adapt Eq. (1) to model the dynamics of local cases I_t, resulting in:

$$I_t \sim \text{Pois}\left(R_t \sum_{s=1}^{t-1} w_s (I_{t-s} + \epsilon I_{t-s}^{\text{imp}})\right), \tag{9}$$

where R_t is the effective reproduction number that characterizes local transmission on day t. For problems where imported cases are not considered, we use Eq. (1).

2.8. Real incidence data

We fit to real case incidence data for local and imported COVID-19 cases for three regions: Victoria and Queensland in Australia and Hong Kong. In each of these three locations, we used cases with dates given by the date of symptom onset. We selected these regions as they exhibit a variety of different trends in R_t: a gradual decrease in Victoria, a more rapid decrease in Queensland, and a fall in R_t followed by the sudden appearance of a second wave in Hong Kong. Data for the Australian regions were obtained from the Australian national COVID-19 database (Price et al., 2020); data for Hong Kong were obtained from the Hong Kong Department of Health COVID-19 database (Hong Kong Department of Health, 2022). For the Australian states, cases of unknown origin were assumed to be local, and in Hong Kong, all cases other than those listed as "imported case confirmed" were treated as local.

The proportion of cases whose local or imported status is unknown varies substantially by region. For the time periods we considered, 57% of cases in Victoria, 8% of cases in Queensland, and 20% of cases in Hong Kong were not confirmed as either local or imported in the datasets, and we treated them as local. This assumption, if incorrect, would lead to upwards bias in our estimates for the reproduction number.

We assumed $\epsilon = 1$ in Eq. (9) for Victoria and Queensland; however, for Hong Kong, transmission networks suggest that imported cases were significantly less infective than local cases, so we set $\epsilon = 0.2$ (Liu et al., 2021; Creswell et al., 2022). In all three instances, we assumed that under-reporting and delays were negligible given the strong surveillance in these countries.

Generally, the relative transmissibility of imported versus domestic cases is unknown, although methods exist for estimating this (Creswell et al., 2022). And different assumptions made about ϵ affect the inferred R_t series: if ϵ is smaller, then a higher level of local transmission is necessary to sustain an epidemic (Creswell et al., 2022), typically shifting the inferred R_t series upwards. Since different assumptions for ϵ tend to shift rather than warp the inferred R_t series, we do not consider these here, since they are unlikely to affect the position of changepoints.

3. Results

3.1. EpiCluster reliably estimates sudden changes in R_t in retrospective analyses

To evaluate the performance of our model, we generated synthetic incidence data using Eq. (1) where the R_t profile was known (see Fig. 2). We considered three R_t profiles: one with a precipitous decline in R_t ("fast drop off"); another, with a decline in R_t followed by a later resurgence ("fast resurgence"; we included this profile since resurgences are harder to infer than declines in transmission strength; Parag and Donnelly, 2022); and another with a more gradual decline in R_t ("slow drop off"). The fast drop off and slow drop off time series were initialized with 5 cases on each of three days preceding the beginning of simulation, while the fast resurgence was initialized with 5 cases on each of fifty days preceding the beginning of simulation. Simulations for fast drop off and slow drop off used the COVID-19 serial interval (Nishiura et al., 2020), while the fast resurgence used the Ebola serial interval as estimated for the 2014 West African Outbreak (Van Kerkhove et al., 2015).

In Fig. 2, we compare R_t estimates from our method with those from two comparator methods: the sliding window method (Thompson et al., 2019) with two different choices of the sliding window width (7 days and 28 days), and the EpiFilter method (Parag, 2021).

Across the three R_t profiles considered, the estimates from the sliding window method lag behind the true values (Fig. 2B), since the windows are inherently backward-looking—the longer the window width, the longer the moving average and the slower it is to respond to changes in R_t; the estimates are also very variable. The EpiFilter method fares better and is able to reliably infer downward shifts in R_t (Fig. 2C), corresponding to suppression; this method overly smooths over the upward tick in transmission in the fast resurgence example.

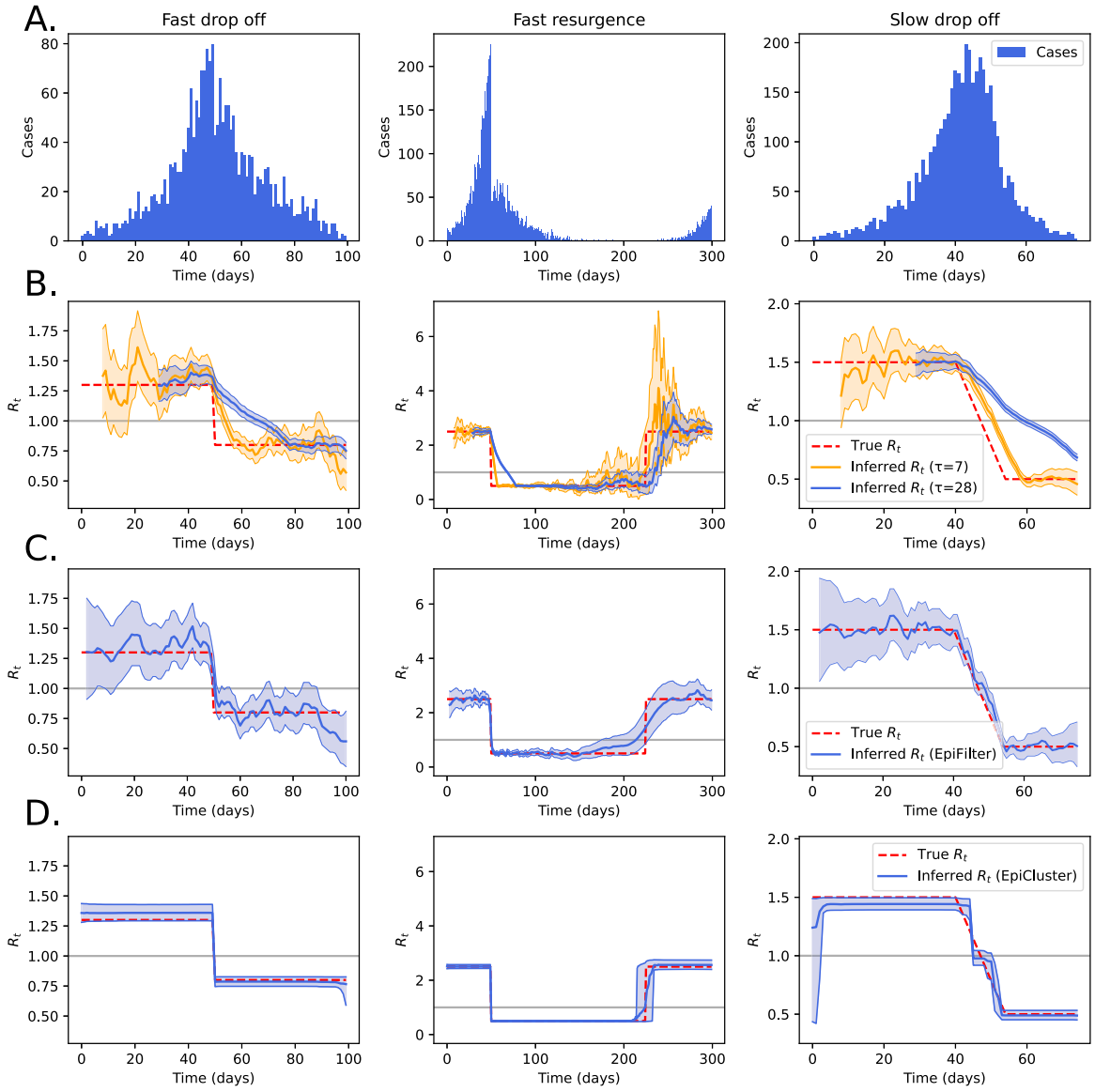

Fig. 2. Recovering synthetic R_t profiles in retrospective analyses. We generated synthetic case data (panel A) using the Poisson renewal model (Eq. (1)) with three prespecified profiles for R_t (dashed red lines in panels B/C/D). In panel B, we show the inferred R_t profile using a sliding window method (Thompson et al., 2019) for two different choices of the sliding window size ($\tau = 7$ and 28 days). In panel C, we show the inferred R_t profile using the EpiFilter method (Parag, 2021). In panel D, we show the inference results when using EpiCluster to recover R_t. In panels B, C and D, shaded regions indicate the central 90% of the posterior distribution of R_t, while the central line indicates the posterior mean, and the background grey line indicates $R_t = 1$.

Our method performs favourably in the two "fast" examples (Fig. 2D). Like the EpiFilter method, our approach is less able to infer resurgence than suppression (Parag and Donnelly, 2022). In the slow drop off example, our piecewise-constant method approximates the linear decline in R_t with a staircase-like profile, which is better estimated by EpiFilter. In Figure S2, we show the effect of changing the hyperparameters of our method on inference for the slow drop off example on the number and location of the regimes which are learned. As the two hyperparameters, θ and σ increase, more weight is given to a partitioning consisting of more regimes (see also Figure S1), and the staircase steps become finer.

To account for stochastic variation in the synthetic data generation, we repeated inference for the fast resurgence example 10 times (Fig. S3). For the three methods, the posterior means are qualitatively similar across all runs, suggesting that these results are consistent across different realizations of the renewal process.

In the fast drop off and fast resurgence examples, EpiCluster estimates R_t with low bias and high precision. This is because the R_t profiles in the simulated examples align well with the assumptions made in our modelling: namely, that the R_t profile is piecewise-constant. We now consider R_t profiles with notable deviations from this assumption. In Fig. 3, we compare the same methods on both noisy (left and middle columns) and oscillatory R_t profiles. When the magnitude of the noise is low (left column), the results mirror those from the previous example. When the noise level increases (middle column), all methods are late to predict the precipitous decline in R_t, and EpiFilter provides a better quantification of uncertainty than the nonparametric model. For the sinusoid example (right column), EpiFilter performs best, since the assumptions underpinning that method—that R_t follows a random walk—are closer to the reality of the generated data.

To evaluate the comparative inference performance of the methods quantitatively, for each R_t profile studied in Figs. 2 and 3, we repeated the generation of synthetic data 100 times and studied the distributions of mean squared error (MSE) between the inferred posterior mean of R_t and the true R_t profile for each method. These distributions of MSE values, estimated via kernel density estimation, are shown in Figure S7. For the majority of the examples, EpiCluster tended to

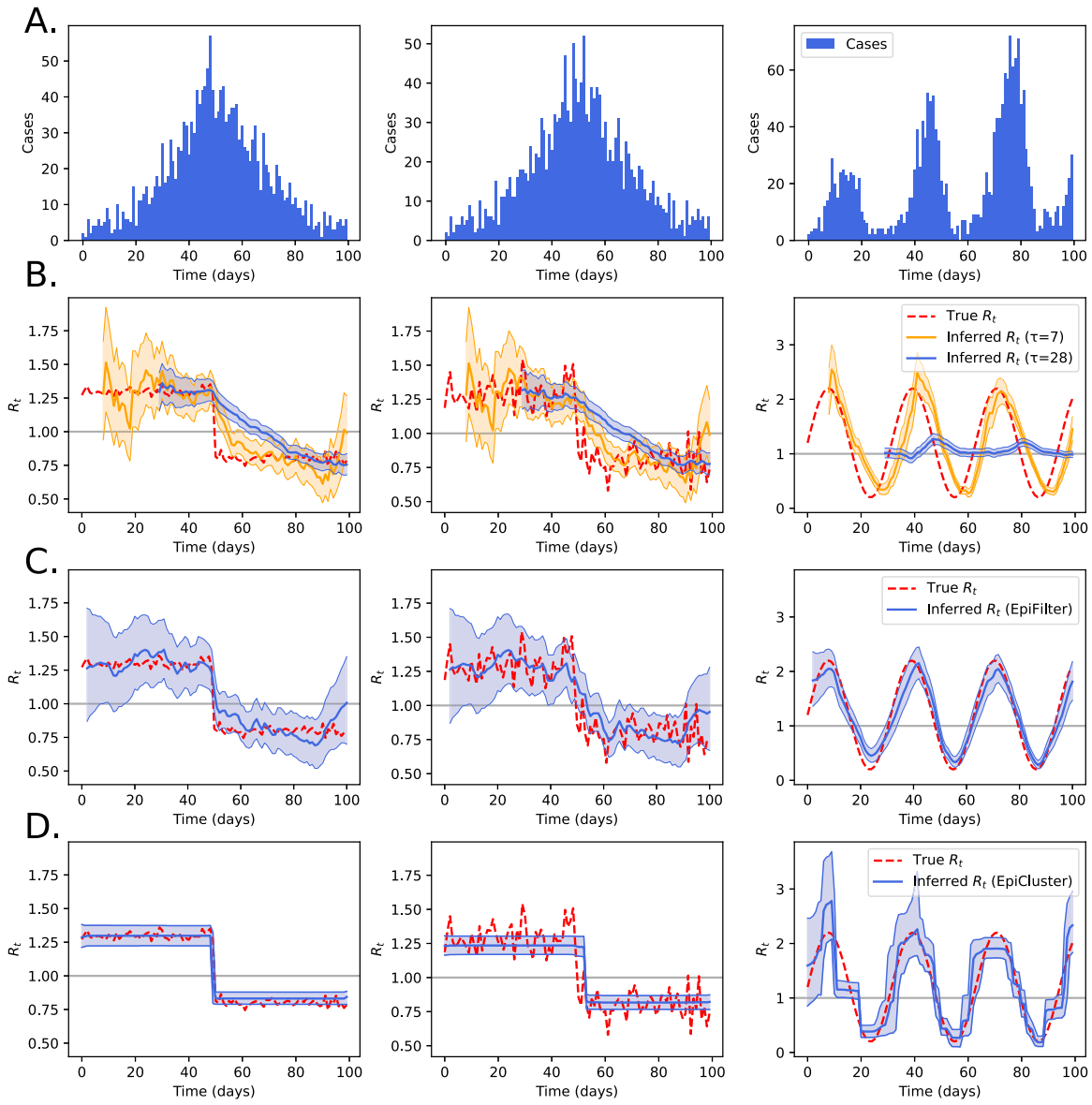

Fig. 3. Recovering noisy and oscillatory R_t profiles in retrospective analyses. We generated synthetic case data (panel A) using the Poisson renewal model with three prespecified profiles for R_t (dashed red lines in panels B/C/D). The R_t profiles were calculated using step functions with additive *i.i.d.* Gaussian noise of standard deviation 0.025 (left) and 0.1 (middle). In the right column, we show results when R_t follows a sine wave. In panel B, we show the inferred R_t profile using a sliding window method (Thompson et al., 2019) for two different choices of the sliding window size ($\tau = 7$ and 28 days). In panel C, we show the inferred R_t profile using the EpiFilter method (Parag, 2021). In panel D, we show the inference results when using EpiCluster to recover R_t. In panels B, C and D, shaded regions indicate the central 90% of the posterior distribution of R_t, while the central line indicates the posterior mean, and the grey line indicates $R_t = 1$.

produce R_t estimates with the lowest MSE values followed by EpiFilter, with the sliding window methods performing worse. On the sinusoid example (Fig. 3, right column), EpiFilter achieves lower MSE values than EpiCluster, presumably because the changes in R_t were more gradual in this case.

3.2. EpiCluster is effective at detecting sharp changes in transmission in real-time

The results thus far have considered retrospective analysis of outbreaks; these analyses are important for understanding the timing and impact of interventions following their imposition (e.g. Flaxman et al. (2020) and Brauner et al. (2021)). But, in unfolding epidemics of novel pathogens, it is crucial to know in as close to real time as data allows whether transmission changes rapidly either after an intervention is instituted or after it is discontinued. In this section, we compare how the three R_t estimation methods fared in inferring an epidemic resurgence in real-time: as new case data becomes available subsequent to a jump upwards in transmission. We used the same fast resurgence data as in Fig. 2 and fit each method for a series of datasets of different lengths. Each of these datasets began at the same point (at $t = 0$); the datasets ended at different points. The endpoints ranged from 5 days to 35 days post-resurgence with gaps of 5 days between them.

The posterior means of the inferred R_t series are shown in Fig. 4, while the full posteriors are shown in Fig. S4. The results illustrate that all three methods needed considerable data post resurgence to infer changes in transmission. For each series, EpiCluster generally fared best in inferring the timing and magnitude of resurgence, with the posterior uncertainty interval reliably including the true R_t profile.

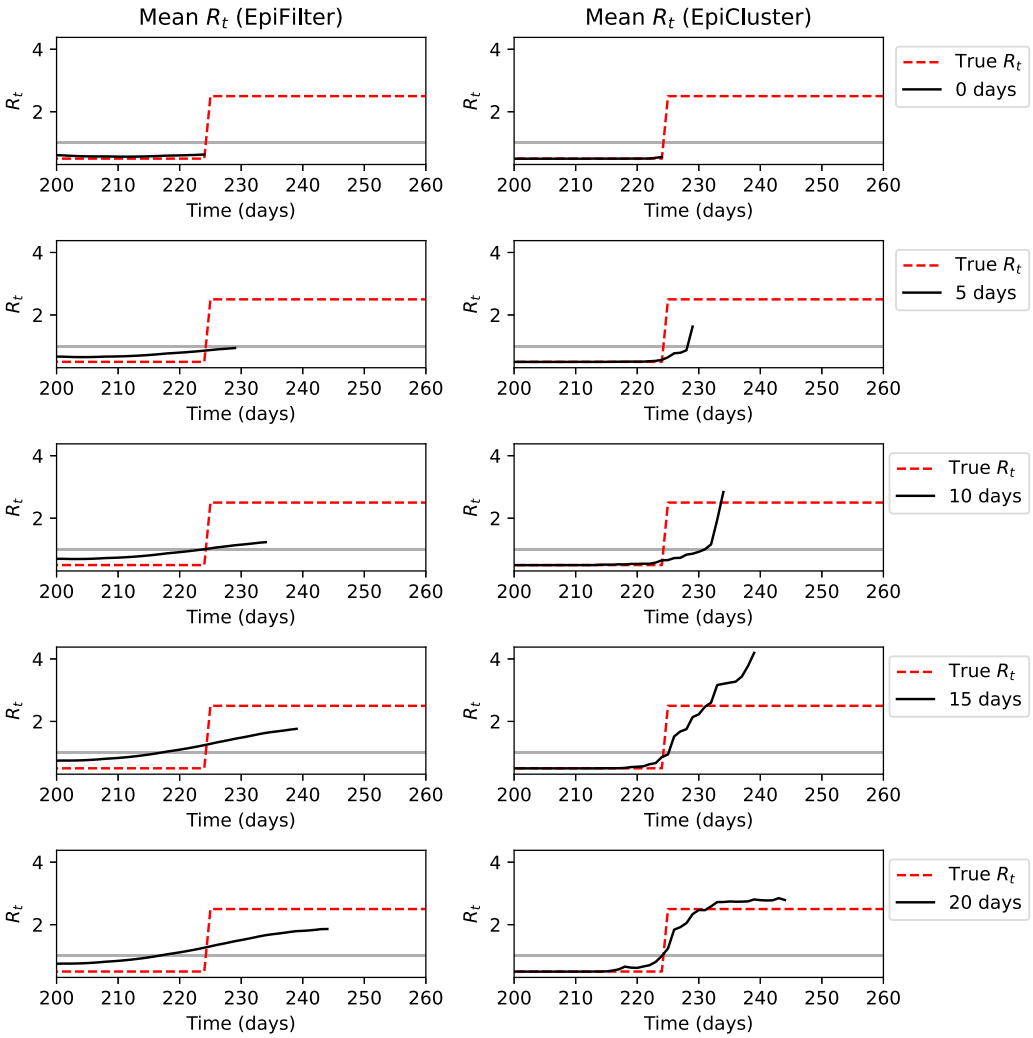

Fig. 4. Real-time estimation of a resurgence in R_t. We used the same fast resurgence synthetic data from Fig. 2 and performed inference for R_t based only on the time series up till the number of days after the resurgence indicated in the legend. In the left panel, we show the mean inferred R_t profile using the EpiFilter method (Parag, 2021). In the right panel, we show the results when using EpiCluster to recover the mean of R_t. The background grey line indicates $R_t = 1$.

3.3. Data generating processes with greater variability pose issues for all methods and EpiFilter generally performs best

Variation in transmissibility across different individuals within a population can lead to greater variation in cases than is accounted for by a Poisson renewal model, and each pathogen exists on a spectrum of dictating the degree of overdisperseness (Lloyd-Smith et al., 2005): SARS, for example, is prone to many superspreading events (Shen et al., 2004); whereas pneumonic plague exhibits less variation in offspring cases (Lloyd-Smith et al., 2005).

To study the robustness of EpiCluster under more variable data generating processes, we generated data using the fast drop off R_t profile and a negative binomial (NB) renewal model with inverse-dispersion parameter $\kappa > 0$: as $\kappa \to \infty$, the NB model approaches the Poisson. So low values of κ correspond to more overdispersed data. Using the fast drop off R_t profile, we generated case data under different values of κ, and, for each series, we fit the sliding window, EpiFilter and EpiCluster methods.

The results are shown in Fig. S5. When κ is large (i.e. the data are effectively generated from a Poisson distribution), the results match those observed in Fig. 3. As the data generating process exhibits more variation, all methods perform worse: generally failing to correctly identify the change in R_t and inferring a highly noisy R_t profile with many spurious fluctuations. However, the sliding window and EpiFilter methods generally produced more robust estimates in the presence of strong overdispersion.

3.4. EpiCluster estimates sharp changes in R_t for real COVID-19 incidence curves

Next, we performed retrospective inference of R_t for the early COVID-19 outbreaks in three selected regions: Victoria and Queensland, Australia, and Hong Kong (see 2.8), which were selected for the variety of transmission profiles they encompass. The R_t estimates for these regions are shown in Fig. 5, again comparing the sliding window approach (panel B) with the EpiFilter approach (panel C) and EpiCluster (panel D).

The first case of COVID-19 in Australia was reported in Victoria state on 25th January 2020 (Storen and Corrigan, 2020). Subsequently, Victoria quickly became a hub of transmission and declared a state of emergency on 16th March, including a ban on non-essential gatherings of over 500 people (Storen and Corrigan, 2020). On 18th March, more restrictions on movement followed with indoor public gatherings of more than 100 people banned and restrictions in aged care facilities introduced across Australia (Storen and Corrigan, 2020). On the 22nd March, the state Premier announced that Victoria would implement a shutdown of all non-essential activity across the state (Storen and Corrigan, 2020). The sliding window approach (Fig. 5B) and EpiFilter

(Fig. 5C) both estimated declines in transmission starting around 22nd March; EpiCluster infers a sharper decline around 25th March. All methods inferred that transmission subsequently remained below the level for sustained transmission, apart from an uptick in transmission estimated from EpiCluster coinciding with a burst of cases around 10th April, which likely reflects a violation of the assumptions of the model.

The first case of COVID-19 in Queensland, Australia occurred on 29th January 2020 (Storen and Corrigan, 2020), and the first wave began in early March. All three estimation methods inferred that, since imported cases were the dominant cause of the wave, there was relatively low community transmission, and the bulk of local R_t estimates were below 1 (Fig. 5). All methods inferred a decline in transmission beginning around the 16th March—the date when Victoria declared a state of emergency, and Australia introduced a self-isolation requirement for all international arrivals—and EpiCluster estimated a rapid decline on 17th March. To combat the insurgence of imported cases, the Queensland Premier announced that the state would restrict access to the border on 24th March: this included termination of all rail services and border road closures (Storen and Corrigan, 2020), and EpiCluster inferred a small decline occurring on this date.

Hong Kong, like Singapore and Taiwan, was quick to act on learning of the outbreak of COVID-19 in Wuhan, China, and the government enacted intensive surveillance campaigns and declared a state of emergency on 25th January, 2020 (Cowling et al., 2020, Fig. 1). On the 7th February 2020, Hong Kong introduced prison sentences for anyone breaching quarantine rules (OT&P Healthcare, 2022). This date broadly coincides with the decline in R_t detected across all three methods, and the decline detected by EpiCluster is especially rapid.

Hong Kong's second wave of COVID-19 began in March 2020 driven by imported cases from North America and Europe (Parag et al., 2021), and all three methods detect an increase in the local R_t shortly after 15th March. Policy responses to this wave by the Hong Kong government included a quarantine requirement on international arrivals (effective 19th March; Xinhua News Agency, 2020) a ban on foreign travellers (effective 25th March; OT&P Healthcare, 2022) and a ban on gatherings of more than four people (effective 27th March; OT&P Healthcare, 2022); a significant decrease in R_t is detected by all three methods around the times when these interventions were imposed. The EpiCluster results mirror the timing of this intervention most closely, suggesting that there was a short time lag between when the interventions were imposed and their effect.

To explore the sensitivity of our estimates for Hong Kong to the hyperparameters of the method, we performed a series of sensitivity analyses where these parameters were fixed at different values and inference was performed (Fig. S6). These experiments illustrate that, as either of the hyperparameters are increased, the R_t profile comprises a greater number of regimes, and there is greater uncertainty in the R_t estimates. The qualitative behaviour of the majority of estimates, however, remains the same, with a large decline in transmission around 7th February 2020 and a resurgence in mid March.

3.5. EpiCluster estimates sharp changes in R_t for other disease outbreaks

To study the applicability of EpiCluster to infectious diseases other than COVID-19, we applied the method to several outbreaks: the 1972 Smallpox outbreak in Yugoslavia, the 1861 Measles outbreak in Hagelloch, Germany, and the 2003 SARS outbreak; these datasets were obtained from the EpiEstim package (Cori et al., 2013). These results are shown in Fig. 6 and show R_t estimates excluding an initial period when cases are low, since these early data are more likely to be unreliable due to data limitations (the full inference results for these outbreaks, including time periods of very low incidence where EpiCluster learns R_t posteriors with high variance, are shown in Figure S8). In all three outbreaks, EpiFilter learns the smoothest R_t profile, while EpiCluster infers a more jagged R_t series with high uncertainty and larger, rapid fluctuations in the value of the effective reproduction number. For SARS, the large variation in R_t inferred by EpiCluster is almost certainly due to the model's assumptions being violated, and we return to this point in the discussion. For these outbreaks, the sliding window method (Thompson et al., 2019) learns an R_t which resembles the results from EpiCluster but typically with smaller and smoother fluctuations in the value of R_t.

4. Discussion

The time-varying reproduction number, R_t, is a threshold metric for facilitating decision making during epidemics. But, there is also value in using R_t estimates to retrospectively assess whether the imposition of interventions caused substantive and rapid reductions in transmission (e.g. Dehning et al., 2020; Flaxman et al., 2020). It is especially key to determine the timing of these reductions, since delays in imposition of interventions can substantially worsen outcomes, particularly during the growth of an epidemic (Pei et al., 2020). Here, we present a general Bayesian inference method using Pitman–Yor process priors which allows any feasible number of changepoints in transmission, and we provide a choice of hyperparameters (see Section 2.2.3), such that, *a priori*, R_t is assumed to remain relatively stable. Through simulated data examples, we show that the method is adept at estimating sharp changes in transmission: in both retrospective and real-time analyses. By fitting the model to real data from COVID-19 outbreaks, we infer discontinuous declines in transmission at times which broadly coincide with the imposition of interventions. The method allows effectively automated detection of changepoints in transmission and could be adapted to handle different types of models in epidemiology and, more generally, provides a framework for handling time-varying parameters.

The information available to estimate R_t changes throughout an epidemic: at the start, there is scant information, and estimates have high uncertainty; when an epidemic is brought under control, cases are initially higher, providing more information of changes in transmission; and resurgences qualitatively mirror the conditions at the start of an epidemic meaning R_t has greater uncertainty (Parag and Donnelly, 2022). Priors thus affect estimates differently at different stages during an epidemic and, by extension, variously for different types of epidemic. The fits of our model and the two comparator methods to COVID-19 case data demonstrate the strong information introduced by the priors. This makes sensitivity analyses particularly important, since no one prior choice satisfies all parties for all situations, and we recommend that, when using EpiCluster in practice, results from it be presented alongside those from existing approaches. We assume that R_t is piecewise-constant with transmission changing discontinuously with the number of pieces and location of breakpoints controlled through a Pitman–Yor process. If transmission changes more gradually, such as may occur during incremental relaxation of NPIs, these assumptions are inappropriate, and a model which allows a more gradual change in R_t will perform better (e.g. EpiFilter; Parag, 2021). Similarly, if the model mischaracterizes the data generating process, for example, by assuming that there are no substantial differences in transmissibility across individuals, estimates will also be poor (Fig. S5). Because of this, it is possible that the sharp changes in COVID-19 transmission identified by EpiCluster for the three locations considered (Fig. 5) reflected violations in the model's assumptions, and future work is to adapt our framework to handle such processes. The rapid fluctuations in R_t learned by EpiCluster for the non-COVID outbreaks (Fig. 6) similarly may reflect the inaccuracy of the Poisson distribution underlying our renewal model, rather than genuine changes in the value of R_t. Smallpox, measles, and SARS are all characterized by significant variation in the reproductive number from individual to individual, and this effect is most pronounced for SARS (Lloyd-Smith et al., 2005), which may partly explain why EpiCluster estimated particularly jagged R_t profiles for this disease outbreak.

Thus, an important extension to our framework would be to allow an overdispersed renewal model, such as that given by using

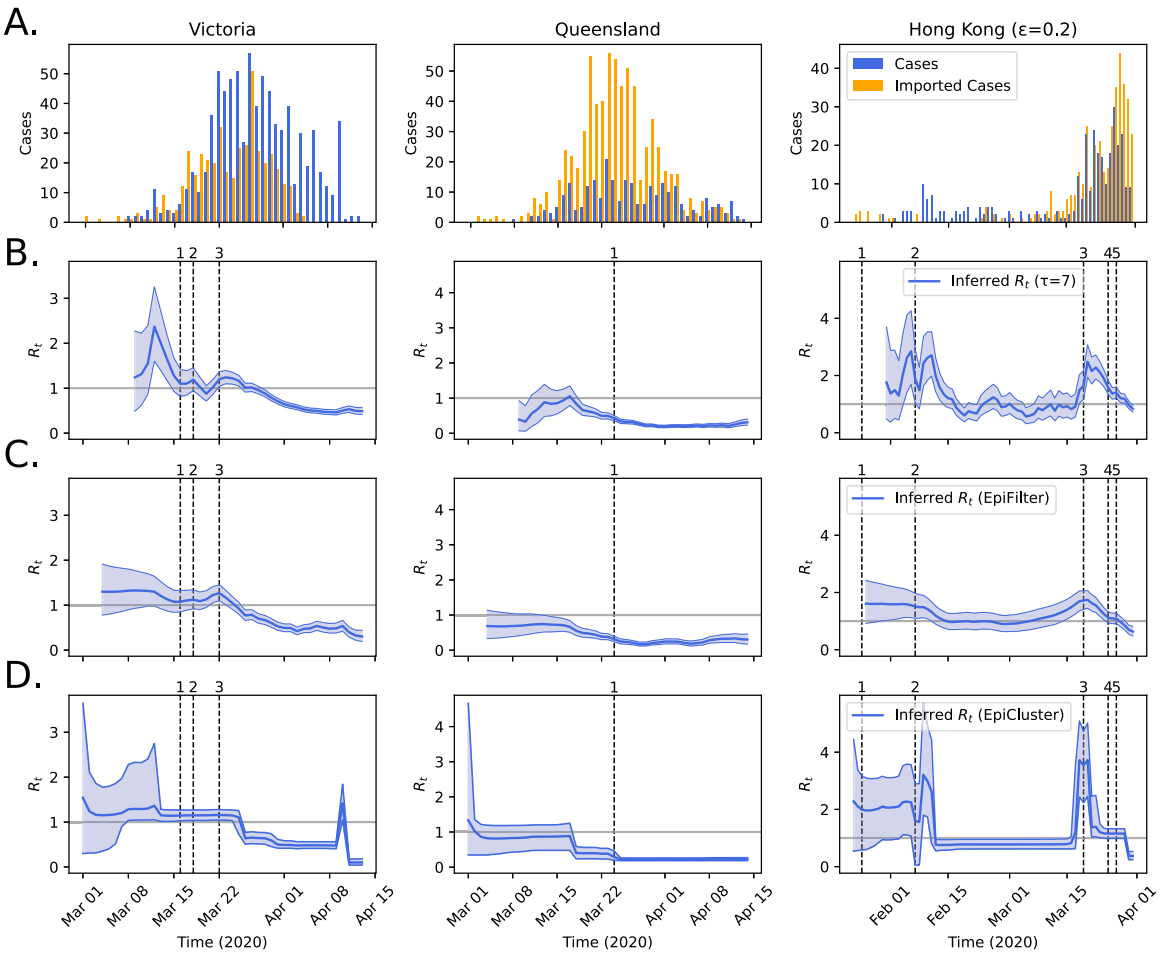

Fig. 5. Learning R_t from early COVID-19 epidemic incidence curves in three locations. Data on local and imported cases from the early COVID-19 pandemic in three selected regions is shown in panel A. In panel B, we show the inferred R_t profile using a sliding window method (Thompson et al., 2019) for two different choices of the sliding window size ($\tau = 7$ and 28 days). In panel C, we show the inferred R_t profile using the EpiFilter method (Parag, 2021). In panel D, we show the inference results when using EpiCluster to recover R_t. In panels B, C and D, shaded regions indicate the central 90% of the posterior distribution of R_t, while the central line indicates the posterior mean, and the background grey line indicates $R_t = 1$. Vertical dotted lines indicate policy-relevant dates. For Victoria: 1: ban on non-essential gatherings of over 500 people; 2: movement restrictions and ban on indoor gatherings of over 100 people; 3: shutdown of all non-essential activity. Queensland: 1: border restrictions and termination of rail services. Hong Kong: 1: state of emergency declared; 2: prison sentences introduced for those breaking quarantine; 3: compulsory quarantine of all arrivals; 4: ban on foreign travellers; 5: ban on gatherings over four people.

a negative binomial distribution in place of the Poisson. However, such an extension would make inference for our framework slower and more challenging, as the analytical integrability of the marginal likelihood, which enables fast collapsed Gibbs sampling for the cluster configurations (Section 2.3), depends on the Poisson likelihood. We did not consider reporting issues here, and these would likely also introduce biases (Gostic et al., 2020; Pitzer et al., 2021). In particular, the available incidence data indicates when cases were reported, but this is likely to be several days after the actual transmission event due to delays in symptom onset and reporting. Various techniques can be used to attempt to correct R_t estimates for these factors (Gostic et al., 2020), and these could be incorporated into our framework for R_t inference. Without such corrections, it is possible that our estimated changepoints in COVID-19 transmission may deviate by several days from the true dates when changes occurred.

The Pitman–Yor process is an example from a broad class of models from Bayesian nonparametrics where the complexity of the models grows along with the volume and complexity of the data (Ghahramani, 2013). Gaussian processes belong also to this class (Rasmussen, 2003) and have found wide application across epidemiology, notably for producing geostatistical maps of disease prevalence for illnesses such as malaria (Bhatt et al., 2015). More data and data of greater variety and complexity are being routinely collected in epidemiological surveillance, and there is a host of Bayesian nonparametric models (e.g. those described in Griffiths and Ghahramani, 2011; Ghahramani, 2013), which are well-placed for their analysis.

Across epidemiology, discretely sampled data are used to infer continuous-time parameters, such as the time-varying reproduction number, R_t in outbreak analysis, the effective population size in phylogenetic Skyline models (Pybus et al., 2000) and the historical force of infection in catalytic models (Muench, 2013). In any of these cases, transmission can change abruptly, and a model such as ours could be used to identify periods of rapid change.

CRediT authorship contribution statement

Richard Creswell: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Martin Robinson:** Supervision, Methodology, Writing – original draft, Writing – review & editing. **David Gavaghan:** Supervision, Methodology, Writing – original draft, Writing – review & editing. **Kris V. Parag:** Investigation, Methodology, Writing – original draft, Writing – review & editing. **Chon Lok Lei:** Supervision, Conceptualization, Methodology, Writing

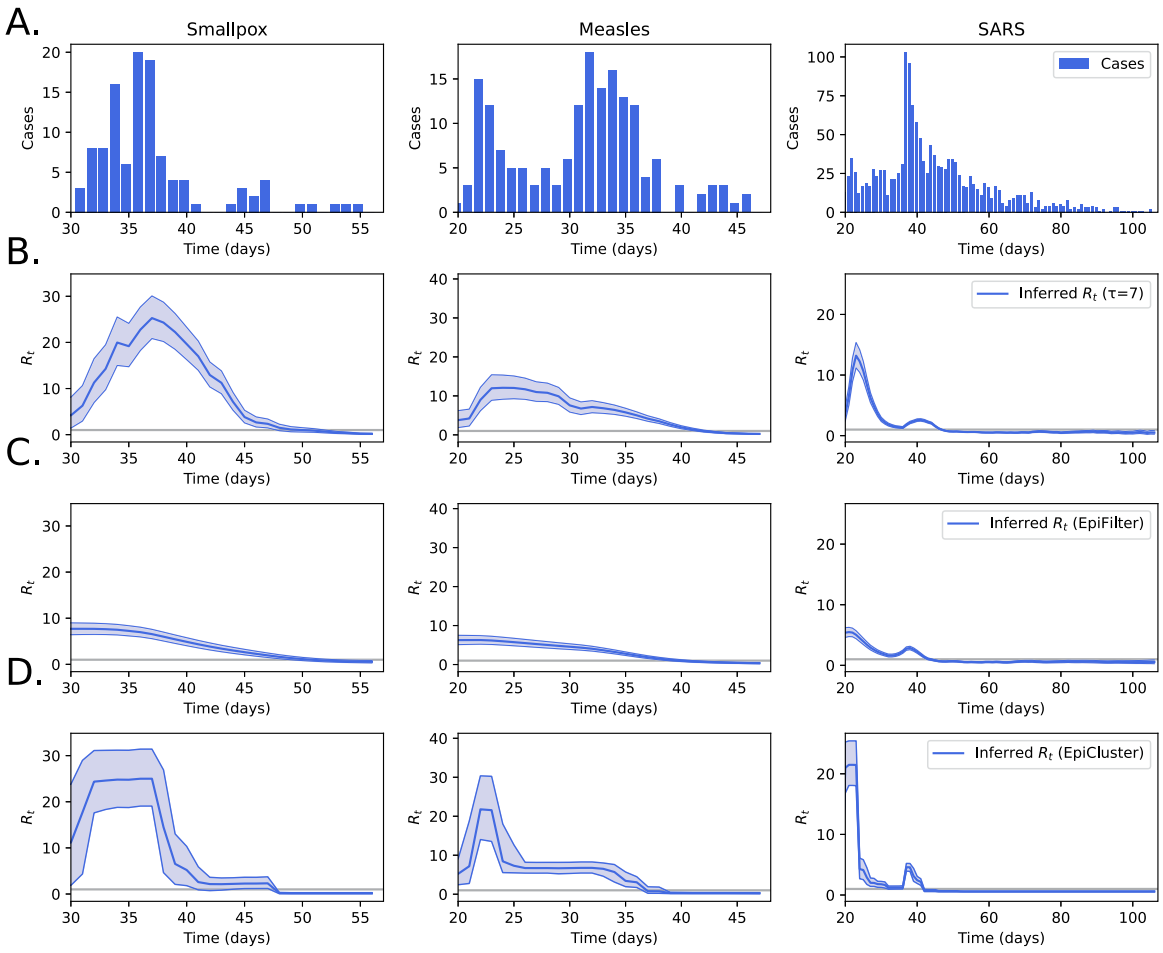

Fig. 6. Learning R_t profiles for non-COVID outbreaks. Data on local cases for three selected outbreaks are shown in A. In panel B, we show the inferred R_t profile using a sliding window method (Thompson et al., 2019) for two different choices of the sliding window size ($\tau = 7$ and 28 days). In panel C, we show the inferred R_t profile using the EpiFilter method (Parag, 2021). In panel D, we show the inference results when using EpiCluster to recover R_t. In panels B, C and D, shaded regions indicate the central 90% of the posterior distribution of R_t, while the central line indicates the posterior mean, and the background grey line indicates $R_t = 1$.

– original draft, Writing – review & editing. **Ben Lambert:** Supervision, Project administration, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

KVP acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/R015600/1), jointly funded by the UK Medical Research Council (MRC) and the UK Foreign, Commonwealth & Development Office (FCDO), under the MRC/FCDO Concordat agreement and is also part of the EDCTP2 programme supported by the European Union.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jtbi.2022.111351.

References

Abbott, S., Hellewell, J., Thompson, R.N., Sherratt, K., Gibbs, H.P., Bosse, N.I., Munday, J.D., Meakin, S., Doughty, E.L., Chun, J.Y., et al., 2020. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. Wellcome Open Res. 5 (112), 112.

Bhatt, S., Weiss, D., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C., Henry, A., Eckhoff, P., et al., 2015. The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. Nature 526 (7572), 207–211.

Brauner, J.M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčiak, T., Stephenson, A.B., Leech, G., Altman, G., Mikulik, V., et al., 2021. Inferring the effectiveness of government interventions against COVID-19. Science 371 (6531).

Cori, A., Ferguson, N.M., Fraser, C., Cauchemez, S., 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am. J. Epidemiol. 178 (9), 1505–1512.

Cowling, B.J., Ali, S.T., Ng, T.W., Tsang, T.K., Li, J.C., Fong, M.W., Liao, Q., Kwan, M.Y., Lee, S.L., Chiu, S.S., et al., 2020. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. Lancet Public Health 5 (5).

Creswell, R., Augustin, D., Bouros, I., Farm, H.J., Miao, S., Ahern, A., Robinson, M., Lemenuel-Diot, A., Gavaghan, D.J., Lambert, B., Thompson, R.N., 2022. Heterogeneity in the onwards transmission risk between local and imported cases affects practical estimates of the time-dependent reproduction number. Phil. Trans. R. Soc. A.

Dehning, J., Zierenberg, J., Spitzner, F.P., Wibral, M., Neto, J.P., Wilczek, M., Priesemann, V., 2020. Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. Science 369 (6500).

Flaxman, S., Mishra, S., Gandy, A., Unwin, H.J.T., Mellan, T.A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J.W., Molod, M., Imperial College COVID-19

Response Team, Ghani, A.C., Donnelly, C., Riley, S., Vollmer, M.A.C., Ferguson, N.M., Okell, L.C., Bhatt, S., 2020. Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature 584 (7820), 257–261.
Fraser, C., 2007. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One 2 (8).
Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statist. Sci. 457–472.
Ghahramani, Z., 2013. Bayesian non-parametrics and the probabilistic approach to modelling. Phil. Trans. R. Soc. A 371 (1984).
Gostic, K.M., McGough, L., Baskerville, E.B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J.A., De Salazar, P.M., et al., 2020. Practical considerations for measuring the effective reproductive number, R_t. PLoS Comput. Biol. 16 (12).
Griffiths, T.L., Ghahramani, Z., 2011. The Indian buffet process: An introduction and review. J. Mach. Learn. Res. 12 (4).
Hart, W.S., Maini, P.K., Thompson, R.N., 2021. High infectiousness immediately before COVID-19 symptom onset highlights the importance of continued contact tracing. ELife 10.
Hong Kong Department of Health, 2022. Latest local situation of COVID-19. https://data.gov.hk/en-data/dataset/hk-dh-chpsebcddr-novel-infectious-agent.
Lambert, B., 2018. A Student's Guide to Bayesian Statistics. Sage.
Li, Y., Campbell, H., Kulkarni, D., Harpur, A., Nundy, M., Wang, X., Nair, H., for COVID, U.N., et al., 2021. The temporal association of introducing and lifting non-pharmaceutical interventions with the time-varying reproduction number (R) of SARS-CoV-2: a modelling study across 131 countries. Lancet Infect. Dis. 21 (2), 193–202.
Lijoi, A., Prunster, I., 2010. Models beyond the Dirichlet process. In: Hjort, N.L., Holmes, C., Muller, P., Walker, S.G. (Eds.), Bayesian Nonparametrics. Cambridge University Press.
Liu, Y., Gu, Z., Liu, J., 2021. Uncovering transmission patterns of COVID-19 outbreaks: A region-wide comprehensive retrospective study in Hong Kong. EClinicalMedicine 36.
Lloyd-Smith, J.O., Schreiber, S.J., Kopp, P.E., Getz, W.M., 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438 (7066), 355–359.
Martínez, A.F., Mena, R.H., 2014. On a nonparametric change point detection model in Markovian regimes. Bayesian Anal. 9 (4), 823–858.
Mendez-Brito, A., El Bcheraoui, C., Pozo-Martin, F., 2021. Systematic review of empirical studies comparing the effectiveness of non-pharmaceutical interventions against COVID-19. J. Infect. 83 (3), 281–293.
Muench, H., 2013. Catalytic models in epidemiology. In: Catalytic Models in Epidemiology. Harvard University Press.
Nishiura, H., Chowell, G., 2009. The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In: Mathematical and Statistical Estimation Approaches in Epidemiology. Springer, pp. 103–121.
Nishiura, H., Linton, N.M., Akhmetzhanov, A.R., 2020. Serial interval of novel coronavirus (COVID-19) infections. Int. J. Infect. Dis. 93, 284–286.
OT&P Healthcare, 2022. COVID-19 timeline of events. https://www.otandp.com/covid-19-timeline, Accessed: 22 June 2020.
Parag, K.V., 2021. Improved estimation of time-varying reproduction numbers at low case incidence and between epidemic waves. PLoS Comput. Biol. 17 (9).
Parag, K.V., Cowling, B.J., Donnelly, C.A., 2021. Deciphering early-warning signals of SARS-CoV-2 elimination and resurgence from limited data at multiple scales. J. R. Soc. Interface 18 (185).
Parag, K.V., Donnelly, C.A., 2020. Adaptive estimation for epidemic renewal and phylogenetic skyline models. Syst. Biol. 69 (6), 1163–1179.
Parag, K.V., Donnelly, C.A., 2022. Fundamental limits on inferring epidemic resurgence in real time using effective reproduction numbers. PLoS Comput. Biol. 18 (4), e1010004.
Pei, S., Kandula, S., Shaman, J., 2020. Differential effects of intervention timing on COVID-19 spread in the United States. Sci. Adv. 6 (49).
Pitman, J., 2002. Combinatorial stochastic processes. Lect. Notes Math. 1875, 7–24.
Pitman, J., Yor, M., 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann. Probab. 855–900.
Pitzer, V.E., Chitwood, M., Havumaki, J., Menzies, N.A., Perniciaro, S., Warren, J.L., Weinberger, D.M., Cohen, T., 2021. The impact of changes in diagnostic testing practices on estimates of COVID-19 transmission in the United States. Am. J. Epidemiol. 190 (9), 1908–1917.
Price, D.J., Shearer, F.M., Meehan, M.T., McBryde, E., Moss, R., Golding, N., Conway, E.J., Dawson, P., Cromer, D., Wood, J., Abbott, S., McVernon, J., McCaw, J.M., 2020. Early analysis of the Australian COVID-19 epidemic. ELife 9.
Pybus, O.G., Rambaut, A., Harvey, P.H., 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155 (3), 1429–1437.
Rasmussen, C.E., 2003. Gaussian processes in machine learning. In: Summer School on Machine Learning. Springer, pp. 63–71.
Roberts, M.G., Nishiura, H., 2011. Early estimation of the reproduction number in the presence of imported cases: pandemic influenza H1N1-2009 in New Zealand. PLoS One 6 (5).
Sharma, M., Mindermann, S., Brauner, J., Leech, G., Stephenson, A., Gavenčiak, T., Kulveit, J., Teh, Y.W., Chindelevitch, L., Gal, Y., 2020. How robust are the estimated effects of nonpharmaceutical interventions against COVID-19? Adv. Neural Inf. Process. Syst. 33, 12175–12186.
Shen, Z., Ning, F., Zhou, W., He, X., Lin, C., Chin, D.P., Zhu, Z., Schuchat, A., 2004. Superspreading sars events, Beijing, 2003. Emerg. Infect. Diseases 10 (2), 256.
Soltesz, K., Gustafsson, F., Timpka, T., Jaldén, J., Jidling, C., Heimerson, A., Schön, T.B., Spreco, A., Ekberg, J., Dahlström, Ö., et al., 2020. On the sensitivity of non-pharmaceutical intervention models for SARS-CoV-2 spread estimation. MedRxiv.
Storen, R., Corrigan, N., 2020. COVID-19: a chronology of state and territory government announcements (up until 30 june 2020). https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/rp2021/Chronologies/COVID-19StateTerritoryGovernmentAnnouncements#_Toc52275800, Accessed: 22 June 2020.
Svensson, Å., 2007. A note on generation times in epidemic models. Math. Biosci. 208 (1), 300–311.
Teh, Y.W., 2010. Dirichlet process. Encyclopedia Mach. Learn. 1063, 280–287.
Thompson, R., Stockwin, J., van Gaalen, R.D., Polonsky, J., Kamvar, Z., Demarsh, P., Dahlqwist, E., Li, S., Miguel, E., Jombart, T., Lessler, J., Cauchemez, S., Cori, A., 2019. Improved inference of time-varying reproduction numbers during infectious disease outbreaks. Epidemics 29.
Van Kerkhove, M.D., Bento, A.I., Mills, H.L., Ferguson, N.M., Donnelly, C.A., 2015. A review of epidemiological parameters from ebola outbreaks to inform early public health decision-making. Sci. Data 2 (1), 1–10.
Wallinga, J., Teunis, P., 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am. J. Epidemiol. 160 (6), 509–516.
Xinhua News Agency, 2020. Carrie Lam: The Hong Kong SAR government will take compulsory quarantine measures to deal with the risk of foreign epidemic importation. Accessed: 13 Sept 2022.