Journal of the Royal Statistical Society Series C: Applied Statistics, 2023, **00**, 1–17
https://doi.org/10.1093/jrsssc/qlad055

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series C Applied Statistics

Original Article

Inference on extended-spectrum beta-lactamase *Escherichia coli* and *Klebsiella pneumoniae* data through SMC^2

L. Rimella[1], S. Alderton[2], M. Sammarro[3], B. Rowlingson[2], D. Cocker[3], N. Feasey[3], P. Fearnhead[1] and C. Jewell[1]

[1]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK
[2]Lancaster Medical School, Lancaster University, Lancaster, UK
[3]Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, UK

Address for correspondence: L. Rimella, Department of Mathematics and Statistics, Lancaster University, LA1 4YF, UK.
Email: l.rimella@lancaster.ac.uk

Abstract

We propose a novel stochastic model for the spread of antimicrobial-resistant bacteria in a population, together with an efficient algorithm for fitting such a model to sample data. We introduce an individual-based model for the epidemic, with the state of the model determining which individuals are colonised by the bacteria. The transmission rate of the epidemic takes into account both individuals' locations, individuals' covariates, seasonality, and environmental effects. The state of our model is only partially observed, with data consisting of test results from individuals from a sample of households. Fitting our model to data is challenging due to the large state space of our model. We develop an efficient SMC^2 algorithm to estimate parameters and compare models for the transmission rate. We implement this algorithm in a computationally efficient manner by using the scale invariance properties of the underlying epidemic model. Our motivating application focuses on the dynamics of community-acquired extended-spectrum beta-lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae*, using data collected as part of the Drivers of Resistance in Uganda and Malawi project. We infer the parameters of the model and learn key epidemic quantities such as the effective reproduction number, spatial distribution of prevalence, household cluster dynamics, and seasonality.

Keywords: antimicrobial-resistant bacteria, epidemiology, individual-based model, SMC^2

1 Introduction

Individual-based stochastic epidemic models offer a powerful approach to disentangling the complex nature of disease transmission in populations of interest and have been shown to provide unprecedented insight into the determinants of risk in outbreak settings in humans, livestock, and plants (Deardon et al., 2010; Jewell et al., 2009; Parry et al., 2014; Probert et al., 2018; Vlek et al., 2013). Typically, these models comprise a state-transition process, where individuals transition between a discrete set of epidemiological states; for example, the well-known SIR model assumes individuals start as *susceptible* to infection, before progressing sequentially to *infected*, and thereafter *removed* (either recovered with solid immunity or dead). The ability to model the transition rates as a function of time, incorporating both the configuration of the states, individual-level covariates, and known relationships between individuals, allows a detailed analysis of the importance of such features in a given outbreak setting.

In general, inference for epidemic models is complicated by the need to account for censored event data (e.g. unobserved susceptible to infected transitions) or risk-biased parameter estimates.

Received: August 24, 2022. **Revised:** March 2, 2023. **Accepted:** June 6, 2023

For well-characterised medium-sized populations where all individuals are observed—such as populations of farms, or patients within a hospital—a Bayesian approach employing Markov chain Monte Carlo data augmentation (daMCMC) represents the state of the art (Deardon et al., 2010; Jewell et al., 2009; Vlek et al., 2013). However, as the population and the number of censored transition events increase, or the fraction of the observable population decreases, these methods rapidly lose efficiency. Moreover, for cyclic state-transition models in which individuals can experience more than one instance of any transition event, exploring the space of the number of transition events, as well as when they occurred, presents a severe implementational challenge.

A popular alternative is approximate Bayesian computation (Fearnhead & Prangle, 2012; Kypraios et al., 2017; Sunnåker et al., 2013) which requires only a simulator from the model to give samples from an approximation to the true posterior distribution. However, the quality of this approximation requires the specification of informative, low-dimensional summary statistics, and these can be difficult to construct (Barnes et al., 2012; Prangle et al., 2014).

Another option is particle MCMC (PMCMC) (Andrieu et al., 2010), where the intractable likelihood is replaced by an estimate obtained using sequential Monte Carlo (SMC) techniques. The appealing aspect of PMCMC methods is their exactness, in the sense that they are proven to target the true posterior distribution of the parameters. However, they are computationally expensive as they require running an entire SMC for each MCMC step, and so it is unlikely to be computationally practicable in individual-based epidemic models where the population size is large. PMCMC algorithms are not sequential as they use SMC only to estimate the likelihood. The recent innovation, SMC^2 (Chopin et al., 2013), is an SMC algorithm that allows parameter inference, only requiring PMCMC steps when we need to overcome particle degeneracy of the parameters. SMC^2 appears to have multiple appealing features for individual-based epidemic models: it is a sequential algorithm, it does not require too many PMCMC steps, and it provides an estimate of the marginal likelihood of the model.

In this article, we apply the SMC^2 algorithm to an individual-level model of acquisition and loss of commensal antimicrobial resistance (AMR) carrying bacteria in three study communities in Malawi. As described in Section 1.1, the study represents a typical scenario in which a cyclic stochastic state-transition model is desired to investigate the drivers of transmission, and the observed data set represents a panel of individuals sampled sparsely from the population. We show the utility of SMC^2 for fitting a high-dimensional individual-based epidemic model like ours, identifying its advantages over other popular approaches for fitting such a model: it is easy to implement, it does not need any summary statistics, it is computationally feasible for large populations.

1.1 Transmission of ESBL *E. coli* and *K. pneumoniae* in Malawi

Our work is motivated by the challenge of fitting an individual-based epidemic model for the spread of bacterial infection. The data set consists of positive–negative sample results for colonisation with extended-spectrum β-lactamase (ESBL) producing *Escherichia coli* (*E. coli*) and *Klebsiella pneumoniae* (*K. pneumoniae*), individual ID, household ID, household location, individual-level variables: gender, income, and age, extracted from the complete data set (Cocker, Sammarro, et al., 2022). The samples were collected in three study areas in Malawi: Chikwawa, Chileka, and Ndirande, over a time span of about 1 year and 5 months (from 29 April 2019 to 24 September 2020) covering both the wet (November–April) and dry (May–October) seasons. Households involved in the study were sampled using an 'inhibitory with close pairs' design extended to allow for sampling within sites with spatially heterogeneous populations (Chipeta et al., 2017; Cocker, Sammarro, et al., 2022). The output of the collecting procedure is a time series with data appearing roughly twice a week (time sparsity) and some periods without samples (e.g. during the COVID outbreak).

To analyse this data, we introduce an individual-based epidemic model, where the state of the model determines which individuals are colonised on a given day. The dynamics of such models can be defined by specifying the rate at that any colonised individual colonises an uncolonised individual and the rate at which a colonised individual recovers. As we are modelling antimicrobial-resistant bacteria, a recovered individual is assumed to be susceptible to future colonisations. An individual-based model is flexible as it allows us to account for the different factors that affect the colonisation rate—and we consider and estimate the effect of time-of-year, distance between

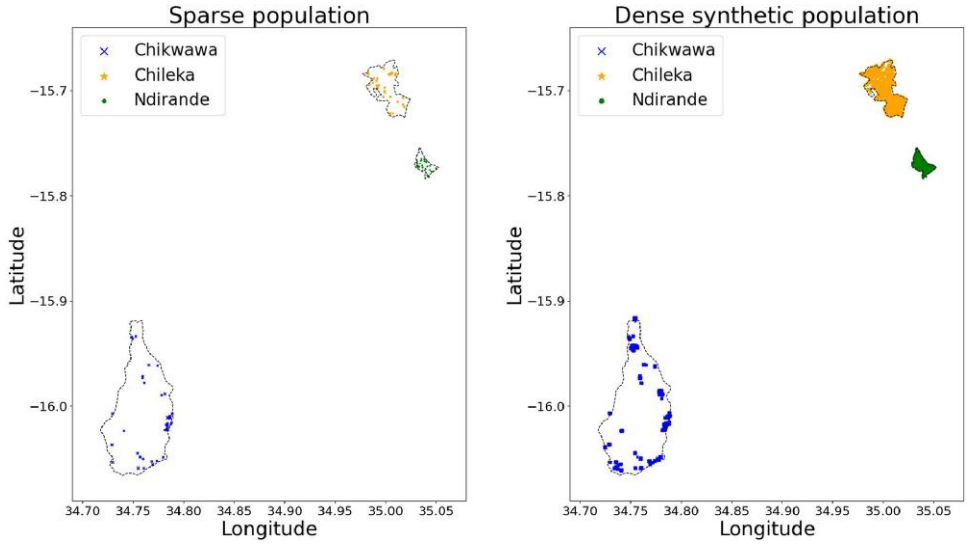

Figure 1. Households are represented with symbols whose size changes according to the household size. Sampled households are reported in the left plot (bottom left corner for Chikwawa, top right corner for Chileka and Ndirande). Synthetic households can be found in the right plot. Different symbols and different colours are associated with different areas.

individuals, whether individuals share the same home, and covariate information such as gender, income, and age on the rate at which one individual infects another.

To use such an approach, we need the state of our model to include not only the colonisation status of the individuals that we sample but also the infection status of individuals in the population at large. Due to the scale-invariance of epidemic models, and in order to make inference computationally feasible, we use a subsample of individuals from the population rather than all individuals and we checked that our results were robust to using such a subsample, by comparing results with different subsample sizes—see the supplementary material. The samples of individuals were obtained by creating a synthetic population based on sampling households but keeping all individuals within a household. We sampled household locations from the DRUM database household sample, the STRATAA census (Darton et al., 2017), OpenStreetMap (OSM) building data, or resampled from the DRUM database household sample itself with jitter, and individuals within each household were obtained from the DRUM database household sample with any missing member of the household sourced from the other households with the most similar characteristics. In total, we generated a synthetic population of 36,314 individuals for Ndirande distributed over 7,949 households, 13,337 individuals for Chileka distributed over 2,888 households, and 9,678 individuals for Chikwawa distributed over 2,416 households. The population sizes are selected both to ensure good posterior estimates, see Section 3, and to respect memory constraints on the GPU nodes of the high-end computing facility from Lancaster University. Figure 1 shows the data before and after the filling procedure.

2 Methodology

2.1 Agent-based UC model

Consider a population size n_I, which varies according to the area (e.g. $n_I = 36{,}314$ in Ndirande), and define an index set $\{1, \ldots, n_I\}$ with the notation $k \in \{1, \ldots, n_I\}$ identifying uniquely an individual in the population. Let $C_t \in \{0, 1\}^{n_I}$ be a vector representing the state of the population with respect to a single bacterium. For example, if we look at *E. coli*, $C_t^{(k)} = 1$ means the kth individual is colonised with *E. coli* at time t, and $C_t^{(k)} = 0$ means that they are uncolonised. A model of this nature is typically defined in continuous time, however, as the data used for analysis are collected in discrete-time intervals, we resort to using a discrete-time Markov chain. Initially, we consider a daily model, though more general discretisations are described in Section 3.

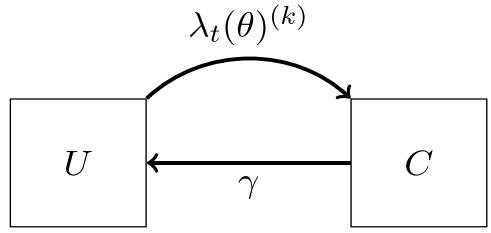

Figure 2. A graphical representation of the UC model dynamic for a general individual k described in equation (1). U stands for 'uncolonised', while C stands for 'colonised'.

For a daily discretisation, we model $(C_t)_{t\geq 0}$ as a discrete-time Markov chain, with a one-time unit corresponding to a day, where each component k evolves as:

$$C_0^{(k)} \sim \mathcal{B}(1-\mathrm{e}^{-\lambda_0}), \quad C_{t+1}^{(k)} \sim \begin{cases} \mathcal{B}\left(1-\mathrm{e}^{-\lambda_t(\theta)^{(k)}}\right) & \text{if } C_t^{(k)}=0 \\ \mathcal{B}(\mathrm{e}^{-\gamma}) & \text{if } C_t^{(k)}=1 \end{cases} \tag{1}$$

where $\mathcal{B}(\cdot)$ is the Bernoulli random variable, $1-\mathrm{e}^{-\lambda_0}$ is the initial probability of colonisation, $\lambda_t(\theta)^{(k)}$ is the transmission rate on individual k at time t, with these depending on unknown parameters θ, and γ is the recovery rate, which is common across individuals. The resulting discrete-time model can also be introduced directly, with parameters specifying the probability of recovery and infection for each day, but we believe that linking to an underlying continuous-time model aids with the interpretability of the parameters.

The above construction considers the colonisation process of the bacteria as a susceptible-infected-susceptible (SIS) model (Keeling & Rohani, 2011) because the nature of the bacteria does not allow the individuals to become immune. We refer to this model as the Uncolonised–Colonised model or the UC model for short. See Figure 2 for a graphical representation.

In (1), the recovery rate γ is assumed to be constant across individuals and over time, while the transmission rate $\lambda_t(\theta)^{(k)}$ is considered to be both time-varying and not homogeneous across individuals. We allow the transmission rate to take into account: within household transmission, between households transmission (and spatial distance), seasonality, the effect of individuals' covariates, and a fixed effect from the environment. We define these effects separately and we then combine them to formulate $\lambda_t(\theta)^{(k)}$.

Before listing the transmission rate components, we define the households as sets of individuals' indexes. Consider the household partition $\mathcal{H}$, which is a partition over the set $\{1, \ldots, n_I\}$, then $\mathrm{H} \in \mathcal{H}$ stands for a household and $k \in \mathrm{H}$ is an individual inside the household H. Throughout the manuscript, we use H^k to denote the household of individual k.

Firstly, consider the within household transmission. We consider two possible models for the within household rate, each defined as:

$$\lambda_t^w(\beta_1)^{(k)} := \beta_1 \frac{\sum_{k' \in \mathrm{H}^k} C_t^{(k')}}{\kappa_1(\mathrm{H}^k)},$$

but a different choice of $\kappa_1(\mathrm{H}^k)$. Here, β_1 is a positive parameter and $\kappa_1(\mathrm{H}^k)$ is either the number of individuals in household H^k, which we denote with $|\mathrm{H}^k|$, or 1. These choices correspond, respectively, to a model in which colonisation rate is diluted by, or constant with respect to, increasing household size.

Secondly, consider the between household transmission. We propose four models for the between household rate defined as:

$$\lambda_t^a(\beta_2, \phi)^{(k)} := \beta_2 \sum_{\mathrm{H} \in \mathcal{H}} D_\phi^{\mathrm{H}^k, \mathrm{H}} \frac{\sum_{k' \in \mathrm{H}} C_t^{(k')}}{\kappa_2(\mathrm{H})},$$

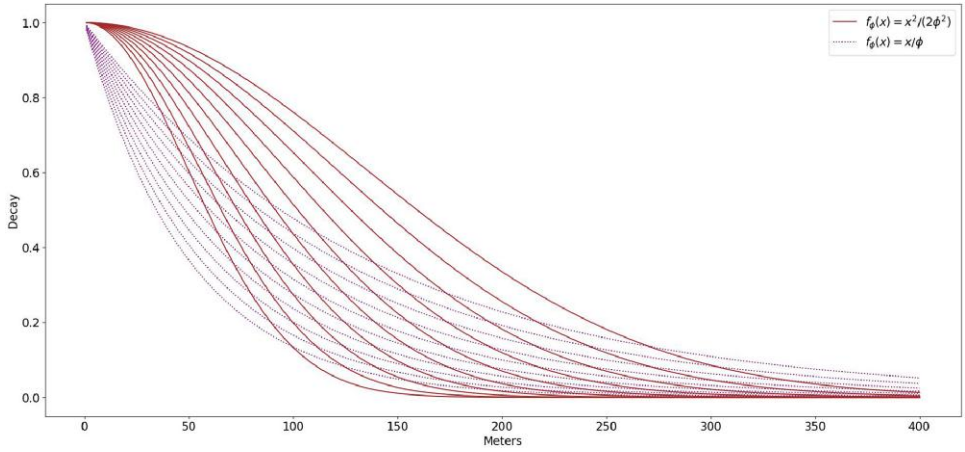

Figure 3. The considered decay functions, solid lines show the 'slow' (Gaussian) decay for $\phi \in (e^{-3}, e^{-2})$, while dotted lines show the 'fast' (exponential) decay for $\phi \in (e^{-3}, e^{-2})$.

but different choices of $\kappa_2(H)$ and $D_\phi^{H^k,H}$. Here, β_2, ϕ are positive parameters and the model formulation varies according to $\kappa_2(H)$, which is either 1 or $|H|$, and $D_\phi^{H^k,H}$, which is a spatial kernel defined as:

$$D_\phi^{H^k,H} := \begin{cases} e^{-f_\phi(d(H^k,H))} & \text{if } H^k \neq H \\ 0 & \text{otherwise,} \end{cases}$$

where $f_\phi(x)$ is either x/ϕ (exponential decay) or $x^2/(2\phi^2)$ (Gaussian decay) and $d(H^k, H)$ is the Euclidean distance between the rectangular coordinates of the households H^k and H in kilometres. Exploring the space of possible spatial kernels was outside the scope of this paper and we restricted our study to the Gaussian-exponential case. However, since our method allows straightforward estimation of the marginal likelihood, it would be possible to formally compare a number of competing spatial models. As already mentioned, $D_\phi^{H^k,H}$ is a spatial kernel that scales the transmission from each household with the distance, meaning that households that are far away from H^k are less likely to influence the colonisation process of k. In addition, $D_\phi^{H,H} = 0$ because the within household effect is modelled separately. This allows to decouple within household and between households transmissions and it improves identifiability. The form of f_ϕ distinguishes a fast $(f_\phi(x) = x/\phi)$ from a slow $(f_\phi(x) = x^2/(2\phi^2))$ decay at the origin, see Figure 3. In practice, $f_\phi(x) = x^2/(2\phi^2)$ implies a higher colonisation pressure from the neighbours.

Thirdly, we know that the prevalence of ESBL-producing *E. coli* and *K. pneumoniae* is higher during the wet season in Malawi (Lewis et al., 2019), so we additionally define a seasonal effect:

$$s_t(\alpha) := 1 + \alpha \cos(\text{frequency} \cdot t + \text{phase}),$$

where α is a parameter in (0, 1) and frequency, and phase are chosen such that the peak of the function is in the middle of the wet season. A graphical representation is available in the supplementary material. Seasonality should not influence the within household transmission, because we expect the household environment to be stable over time. For this reason, we use the seasonal effect as a multiplier of the between households transmission rate.

Next, the individuals' covariates might influence the transmission rate, hence we define an individual effect:

$$I(\delta)^{(k)} := e^{\langle (\text{covariates of } k), \delta \rangle},$$

where δ is a three-dimensional vector with each component referring to a different covariate (i.e. gender, income, and age), 'covariates of k' are the standardised covariates of individual k and $\langle \cdot, \cdot \rangle$ denotes the scalar product between vectors. In contrast with the seasonal effect, we assume the individual effect to impact both the within household and the between households transmissions, hence we employ it as a global multiplier.

Finally, we also assume the presence of a fixed effect ϵ, which is capturing the transmission that is not explained by the population dynamic and acts as a shift on the transmission rate.

The final formulation of the transmission rate combines these features and is defined as:

$$\lambda_t(\theta)^{(k)} = I(\delta)^{(k)}(\lambda_t^w(\beta_1)^{(k)} + s_t(\alpha)\lambda_t^a(\beta_2, \phi)^{(k)}) + \epsilon,$$

where $\theta = (\beta_1, \beta_2, \phi, \alpha, \delta, \epsilon)$.

We have defined eight different combinations of models, which vary according to $\kappa_1(|H|)$, $\kappa_2(|H|)$, f_ϕ, these are further combined with setting or learning δ, α, ϵ, for a total of fifty-five models.

2.2 Observation model

We use $Y_t \in \{0, 1, \text{NA}\}^{n_I}$ to indicate the test results of a specific bacterial species at time t, with NA standing for 'not available' (i.e. not tested at that time or not included in the study). For instance, if we look at ESBL *K. pneumoniae* then $Y_t^{(k)} = 0$ means that individual k has tested negative for colonisation with ESBL *K. pneumoniae* at time t and reported in our data set. We note that only a small subset of individuals is detected and it varies with time, hence we define the set $\mathcal{D}_t \subset \{1, \ldots, n_I\}$ to represent the detected individuals at time t. Additionally, regarding the specificity and sensitivity of the test, even though $C_t^{(k)} = 1$, there is a probability that we might get a false negative result (i.e. $Y_t^{(k)} = 0$). Keeping these in mind we define the conditional distribution of $Y_t^{(k)}$ given $C_t^{(k)}$ as:

$$Y_t^{(k)} \mid C_t^{(k)} \sim \begin{cases} \mathcal{B}(s_e) & \text{if } k \in \mathcal{D}_t, \ C_t^{(k)} = 1 \\ \mathcal{B}(1 - s_p) & \text{if } k \in \mathcal{D}_t, \ C_t^{(k)} = 0 \\ \text{NA} & \text{otherwise,} \end{cases} \tag{2}$$

where s_e, s_p are in (0, 1) and they represent the sensitivity and specificity of the test. As discussed in Section 1.1, the data are sparse in both time and space. This sparsity is treated in (2) through the evolving set $\mathcal{D}_t$, which can be directly extracted from the data.

Sensitivity s_e, specificity s_p along with the recovery rate γ, frequency and phase are treated as known.

3 Inference

By definition, $(C_t)_{t\geq 0}$ is an unobserved Markov chain and Y_t is conditionally independent of all the other variables in the model given C_t, hence $(C_t, Y_t)_{t\geq 0}$ is a hidden Markov model with finite state-space (Rabiner & Juang, 1986; Zucchini & MacDonald, 2009). Inference in a finite state-space hidden Markov model is naively pursued by computing the likelihood in closed form through the forward algorithm and then plugging it in a Markov chain Monte Carlo (MCMC) algorithm (Andrieu et al., 2003; Robert & Casella, 2004) to sample from the posterior distribution over the parameters of interest. However, in our case, this requires a marginalisation over the latent state-space and so operations of the order $\mathcal{D}(2^{n_I})$, which is infeasible for even moderate-size populations.

We implement the SMC^2 algorithm proposed by Chopin et al. (2013), which sequentially targets the posterior over both the parameters θ and the latent process $C_0, \ldots, C_t$.

3.1 SMC and SMC^2

SMC^2 can be intuitively seen as an SMC algorithm within an SMC algorithm, where the former controls the latent process C_t and the latter guide the parameters θ. The SMC algorithm for the latent process uses the auxiliary particle filter (APF) (Carpenter et al., 1999; Johansen &

Doucet, 2008; Pitt & Shephard, 1999) which proposes new states according to the distribution of $C_t \mid C_{t-1}, Y_t$. In our model, it is simple to check that $C_t^{(k)}$ are conditionally independent given C_{t-1}, Y_t. Furthermore, the distribution of $C_t^{(k)}$ will differ depending on whether we have data on individual k at time t. For individuals with data, the distribution of $C_t^{(k)}$ is $\mathcal{B}(p_t^{(k)})$ with:

$$p_t^{(k)} := \begin{cases} \dfrac{\left(1-\mathrm{e}^{-\lambda_t(\theta)^{(k)}}\right)\left[s_e^{Y_t^{(k)}}(1-s_e)^{1-Y_t^{(k)}}\right]}{\left(1-\mathrm{e}^{-\lambda_t(\theta)^{(k)}}\right)\left[s_e^{Y_t^{(k)}}(1-s_e)^{1-Y_t^{(k)}}\right]+\mathrm{e}^{-\lambda_t(\theta)^{(k)}}\left[s_p^{1-Y_t^{(k)}}\left(1-s_p\right)^{Y_t^{(k)}}\right]} & \text{if } C_{t-1}^{(k)}=0 \\ \dfrac{\mathrm{e}^{-\gamma}\left[s_e^{Y_t^{(k)}}(1-s_e)^{1-Y_t^{(k)}}\right]}{\mathrm{e}^{-\gamma}\left[s_e^{Y_t^{(k)}}(1-s_e)^{1-Y_t^{(k)}}\right]+(1-\mathrm{e}^{-\gamma})\left[s_p^{1-Y_t^{(k)}}\left(1-s_p\right)^{Y_t^{(k)}}\right]} & \text{if } C_{t-1}^{(k)}=1 \end{cases} \tag{3}$$

where the above is computed using (1) and (2) and with:

$$p_0^{(k)} := \frac{\left(1-\mathrm{e}^{-\lambda_0}\right)\left[s_e^{Y_0^{(k)}}(1-s_e)^{1-Y_0^{(k)}}\right]}{\left(1-\mathrm{e}^{-\lambda_0}\right)\left[s_e^{Y_0^{(k)}}(1-s_e)^{1-Y_0^{(k)}}\right]+\mathrm{e}^{-\lambda_0}\left[s_p^{1-Y_0^{(k)}}\left(1-s_p\right)^{Y_0^{(k)}}\right]}. \tag{4}$$

The APF for the UC-model is reported in Algorithm 1, where a key role is played by the denominators in (3)–(4):

$$\begin{aligned} w_0^{(k)} &:= \left(1-\mathrm{e}^{-\lambda_0}\right)\left[s_e^{Y_0^{(k)}}(1-s_e)^{1-Y_0^{(k)}}\right]+\mathrm{e}^{-\lambda_0}\left[s_p^{1-Y_0^{(k)}}\left(1-s_p\right)^{Y_0^{(k)}}\right], \\ w_t^{(k)} &:= \left\{\left(1-\mathrm{e}^{-\lambda_t(\theta)^{(k)}}\right)\left[s_e^{Y_t^{(k)}}(1-s_e)^{1-Y_t^{(k)}}\right]\right. \\ &\quad \left.+\mathrm{e}^{-\lambda_t(\theta)^{(k)}}\left[s_p^{1-Y_t^{(k)}}\left(1-s_p\right)^{Y_t^{(k)}}\right]\right\}\left(1-C_{t-1}^{(k)}\right) \\ &\quad +\left\{\mathrm{e}^{-\gamma}\left[s_e^{Y_t^{(k)}}(1-s_e)^{1-Y_t^{(k)}}\right]+(1-\mathrm{e}^{-\gamma})\left[s_p^{1-Y_t^{(k)}}\left(1-s_p\right)^{Y_t^{(k)}}\right]\right\}C_{t-1}^{(k)}. \end{aligned}$$

If $k \notin \mathcal{D}_t$ (the set of sampled individuals) then $C_t^{(k)} \mid C_{t-1}^{(k)}, Y_t$ is distributed as $C_t^{(k)} \mid C_{t-1}^{(k)}$ and follows (1), with $w_t^{(k)} = 1$. Both $p_t^{(k)}$ and $w_t^{(k)}$ depend on $C_{t-1}^{(k)}$, and we make this dependence explicit in Algorithm 1 by writing $p_t^{p,(k)}$ and $w_t^{p,(k)}$, where p is the particle index. Given the parameters θ, the APF allows us to build particle approximations of the distribution of $C_t \mid Y_0, \ldots, Y_t$ and estimates of the likelihood (i.e. the quantity $\mathcal{L}(\theta)$).

Algorithm 1 requires us to know θ, but it can be combined with another SMC algorithm to infer the parameters: resulting in the SMC^2 algorithm. This algorithm stores at iteration s a particle approximation to the joint posterior distribution of the parameters and latent state given the data up to the sth sample of households. These particle approximations are updated recursively from s to $s+1$ by simulating the dynamics of the latent state between the associated time-points (particles for the latent process), weighting by the likelihood of the data at time $s+1$, and, if needed, resampling of the parameters (particles for the parameters). A key component of SMC^2 is the use of a particle MCMC step at resampling events, which allows for new parameter values to be sampled from their correct conditional distribution. An advantage of SMC^2 is that we can monitor the particle weights to get an estimate of the marginal likelihood for our model (Chopin et al., 2013; Chopin & Papaspiliopoulos, 2020). Pseudocode for SMC^2 is reported in Algorithm 2, where from line 16 onward we briefly report the rejuvenation step and line 21 refers to a Metropolis–Hastings using the approximate likelihoods, more details are available in the supplementary

Algorithm 1 APF for UC-model

Require: P, θ, $Y_1, \ldots, Y_t$
for $p = 1, \ldots, P$ **do**
  for $k = 1, \ldots, n_I$ **do**
    Compute $p_0^{(k)}$, sample $C_0^{p,(k)} \sim \mathcal{B}(p_0^{(k)})$ and compute $w_0^{(k)}$
Set $w_0 \leftarrow \prod_{k=1}^{n_I} w_0^{(k)}$ and $\mathcal{L}_0(\theta) \leftarrow w_0$
for $s = 1, \ldots, t$ **do**
  for $p = 1, \ldots, P$ **do**
    for $k = 1, \ldots, n_I$ **do**
      Compute $p_s^{p,(k)}$, sample $C_s^{p,(k)} \sim \mathcal{B}(p_s^{p,(k)})$ and compute $w_s^{p,(k)}$
    Set $w_s^p \leftarrow \prod_{k=1}^{n_I} w_s^{p,(k)}$
  Set $\mathcal{L}_s(\theta) \leftarrow \mathcal{L}_{s-1}(\theta) \frac{1}{P} \sum_{p=1}^{P} w_s^p$
  Resample C_s^p proportionally to w_s^p

Algorithm 2 SMC^2 for inference in UC-model

Require P_θ,P, θ, $Y_1, \ldots, Y_t$
for $m = 1, \ldots, P_\theta$ **do**
  Sample θ^m from the prior and set $w_\theta^m \leftarrow 1$
  for $p = 1, \ldots, P$ **do**
    for $k = 1, \ldots, n_I$ **do**
      Compute $p_0^{(k)}$, sample $C_0^{m,p,(k)} \sim \mathcal{B}(p_0^{(k)})$ and compute $w_0^{(k)}$
  Compute $w_0 \leftarrow \prod_{k=1}^{n_I} w_0^{(k)}$, set $w_\theta^m \leftarrow w_0$ and $\mathcal{L}_0(\theta^m) \leftarrow w_0$
  Set the marginal likelihood $\mathcal{L}_0 \leftarrow w_0$
for $s = 1, \ldots, t$ **do**
  for $m = 1, \ldots, P_\theta$ **do**
    for $p = 1, \ldots, P$
      for $k = 1, \ldots, n_I$ **do**
        Compute $p_s^{m,p,(k)}$ depending on θ^m and $C_{t-1}^{m,p}$
        Sample $C_s^{m,p,(k)} \sim \mathcal{B}(p_s^{m,p,(k)})$ and compute $w_s^{m,p,(k)}$
      Set $w_s^{m,p} \leftarrow \prod_{k=1}^{n_I} w_s^{m,p,(k)}$
    Set $w_\theta^m \leftarrow w_\theta^m \frac{1}{P} \sum_{p=1}^{P} w_s^{m,p}$ and $\mathcal{L}_s(\theta^m) \leftarrow \mathcal{L}_{s-1}(\theta^m) \frac{1}{P} \sum_{p=1}^{P} w_s^{m,p}$
    Set the marginal likelihood $\mathcal{L}_s \leftarrow \mathcal{L}_{s-1} \frac{1}{P_\theta P} \sum_{m=1}^{P_\theta} \sum_{p=1}^{P} w_\theta^m w_s^{m,p}$
  If $ESS(w_\theta) \leq P_\theta/2$ **then**
    Resample θ^m proportionally to w_θ^m
    Propose $\tilde{\theta}^m$ given θ^m and run Algorithm 1 up to s:
      • get $\tilde{C}_s^{m,p}$ and $\mathcal{L}_s(\tilde{\theta}^m)$
    Keep $(\theta^m, C_s^{m,p}, \mathcal{L}_s(\theta^m))$ or replace with $(\tilde{\theta}^m, \tilde{C}_s^{m,p}, \mathcal{L}_s(\tilde{\theta}^m))$
    Set $w_\theta^m \leftarrow 1$

material and in Chopin et al. (2013). A key part of the rejuvenation step is the check on $ESS(w_\theta) := [\sum_{m=1}^{P_\theta} (w_\theta^m)^2]/(\sum_{m=1}^{P_\theta} w_\theta^m)^2$ being above a certain threshold, which, in our application, is set to $P_\theta/2$. The intuition behind this rejuvenation step is simple if our sample of parameters θ^m is not a good representation of our data we generate a new sample of parameters according to a Metropolis–Hastings kernel.

To get an efficient implementation of SMC^2, we combine it with APF (Johansen & Doucet, 2008) and we simulate the new latent states over a time-step of up to 7 days, see next paragraph. We also take advantage of the independence of our model over the two bacterial species and the three geographic regions so that we can parallelise the fitting procedure across species, regions, and models for the infection rate.

Time jumping APF within SMC^2

The time sparsity of the data makes the use of APF challenging for applications where $\mathcal{D}_t = \emptyset$ for most t's. Indeed, whenever $\mathcal{D}_t = \emptyset$ we are sampling from the transition kernel in (1), without correcting with observed data, which might lead to a low effective sample size of the weights and high-variance estimates of the likelihood (Ju et al., 2021; Rimella et al., 2022). We propose to use a coarser time discretisation and simulate new individuals' states every h days instead of every day. This can be done by generalising (1):

$$C_0^{(k)} \sim \mathcal{B}(1 - \mathrm{e}^{-\lambda_0}), \quad C_{t+h}^{(k)} \sim \begin{cases} \mathcal{B}\left(1 - \mathrm{e}^{-h\lambda_t(\theta)^{(k)}}\right) & \text{if } C_t^{(k)} = 0 \\ \mathcal{B}(\mathrm{e}^{-h\gamma}) & \text{if } C_t^{(k)} = 1 \end{cases} \tag{5}$$

and by using this new dynamic to compute a $p_{t+h}^{(k)}$ as in (3), with $h\lambda_t(\theta)^{(k)}$ and $h\gamma$ appearing instead of $\lambda_t(\theta)^{(k)}$ and γ. Our main results are based on a weekly discretisation, so $h = 7$. However, the DRUM data are not equally spaced in time, hence we define a simulation schedule between each pair of observations. The schedule is built by looking at a pair of times t_1, t_2 where $\mathcal{D}_{t_1} \neq \emptyset, \mathcal{D}_{t_2} \neq \emptyset$ and $\mathcal{D}_t = \emptyset$ for all $t \in [t_1, t_2]$, and by dividing the interval $[t_1, t_2]$ into subintervals of size 7 starting from t_2 and going backwards (with the final being less than 7 if $t_2 - t_1$ is not divisible exactly). Note that choosing a bigger h also affects the computational efficiency of the algorithm, by reducing the amount of simulation in the SMC^2 and so the computational cost. The validity of this procedure is checked empirically in Section 4.

4 Simulation study

As mentioned in the previous section, we run our experiments using an SMC^2 where the embedded APF is computing $C_{t+h} \mid C_t, Y_{t+h}$ by combining dynamic (5) with the emission distribution (2). The advantages of such an approach are mainly computational and we also find it to improve inference (e.g. smoother posterior distributions, higher effective sample size).

We test empirically the validity of this procedure on simulated data generated as follows:

- we set $\log(\beta_1) = -2.8$, $\log(\beta_2) = -4.4$, $\log(\phi) = -4.6$, $\delta = (0, 0, 0)$, $\alpha = 0.8$, $\log(\epsilon) = -6.1$, $\kappa_1(H) = \kappa_2(H) = |H|$, $f_\phi(x) = x/\phi$;
- we create a population as the one in Ndirande by merging the real data with the synthetic data;
- we simulate from (1) and report with (2), using an $\mathcal{D}_t$ as in the *K. pneumoniae* data from DRUM.

The above data are then analysed with an SMC^2 algorithm where $h = 1, 3, 7$. Each algorithm is run four times to check the reliability of the output. The run with the highest likelihood is then chosen and the corresponding posterior distributions are reported in Figure 4. From Figure 4, we can notice that we are able to recover almost exactly β_2, ϕ, and ϵ, while β_1 is underestimated, which is due to the multimodality of the model and the sparsity of the observations. Increasing h massively influences the computational cost which is around 316 min for $h = 1$, 101 min for $h = 3$, and 40 min for $h = 7$, and smooths the posterior distributions as shown in Figure 4, but seems to introduce little bias. We also noticed that a higher h is associated with a larger effective sample size of the parameters, indeed for $h = 1$, we run 19 rejuvenation steps, for $h = 3$, we run 15 rejuvenation steps, and for $h = 7$, we run 13 rejuvenation steps.

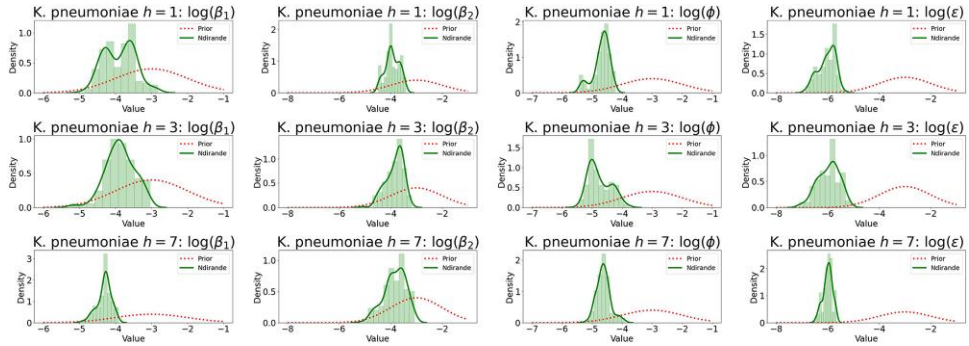

Figure 4. Posterior distribution for simulated data on Ndirande when $h = 1, 3, 7$. On the columns from left to right: posterior distribution for β_1, β_2, ϕ, ϵ. On the rows from top to bottom: $h = 1, 3, 7$.

5 Analysis of DRUM data

5.1 Model selection

As already mentioned in the previous section, SMC^2 outputs both a sample from the posterior distribution over the parameters and a marginal likelihood estimate. We use the latter for model selection and the former to estimate the parameters and interpret the results.

We perform inference in all the settings described in Section 2.1, with:

- $\lambda_0 = 0.13$, as the estimated percentage of the population affected by the bacteria was determined to be 13% in Sammarro et al. (2022);
- frequency $= 2\pi/365.25$ to ensure a period of 1 year, phase $= 0.55\pi$ to align with the wet–dry seasons in Malawi, and so match our underlying knowledge on the bacteria (Jewell & Brown, 2015);
- sensitivity $s_e = 0.8$ and specificity $s_p = 0.95$ as in Cocker, Chidziwisano, et al. (2022);
- $\gamma = 1/10$ as suggested in Lewis et al. (2019), to improve identifiability.

We note that because our observations are noisy versions of the state of the population, rather than the transition events themselves (e.g. Jewell et al., 2009), we cannot identify β_1 and β_2 from γ. Since in our analysis, we are more interested in the colonisation rates, we solve this by assuming a fixed value for γ according to a previous study (Lewis et al., 2019).

We run a total of 55 different models for each bacterial species–study area combination, with each SMC^2 run four times to ensure robustness of the output. For each bacteria and each model, we compute the posterior distribution over the models under a uniform prior. Finally, the marginal likelihoods are aggregated over the study areas, and we report the models with the five highest marginal likelihoods in Table 1. We find that:

- for *E. coli* it is better to estimate ϵ rather than setting it to 0, use an exponential decay in the spatial kernel rather than a Gaussian, set $\kappa_1(|H|) = 1$ rather than $|H|$, set $\kappa_2(|H|) = |H|$ rather than 1, set $\delta = (0, 0, 0)$ rather than estimate it and set the seasonality to 0.6 rather than estimate it;
- for *K. pneumoniae* it is better to estimate ϵ rather than setting it to 0, use a Gaussian decay in the spatial kernel rather than an exponential, set $\kappa_1(|H|) = |H|$ rather than 1, set $\kappa_2(|H|) = |H|$ rather than 1, set $\delta = (0, 0, 0)$ rather than estimate it and set the seasonality to 0.8 rather than estimate it.

For both *E. coli* and *K. pneumoniae*, we first try to learn δ and α: for the former, we find $\delta \approx (0, 0, 0)$ so we decide to set $\delta = (0, 0, 0)$; for the latter, we find a posterior distribution over α in the interval $(0.4, 0.6)$ for all the areas of study, see supplementary material, but this introduces a multimodal posterior distribution for the other parameters. Therefore, we take the pragmatic

Table 1. Table reporting the models with the highest posteriors under uniform priors

$\kappa_1(\lvert H\rvert)$	$f_\phi(x)$	α	Model posterior under uniform prior for *E. coli*	Model posterior under uniform prior for *K. pneu.*
1	x/ϕ	0.6	0.861	<0.005
$\lvert H\rvert$	$x^2/(2\phi^2)$	0.4	0.068	0.088
$\lvert H\rvert$	$x^2/(2\phi^2)$	0.8	0.021	0.579
$\lvert H\rvert$	x/ϕ	0.6	0.016	0.093
$\lvert H\rvert$	$x^2/(2\phi^2)$	0.2	0.012	0.015
$\lvert H\rvert$	x/ϕ	0.2	<0.005	0.068
$\lvert H\rvert$	x/ϕ	0.8	<0.005	0.078

Note. Model formulation changes according to $\kappa_1(|H|)$, $f_\phi(x)$, $\kappa_2(|H|)$, α, δ, ϵ, but $\kappa_2(|H|) = |H|$, $\delta = (0, 0, 0)$ and ϵ with $\mathcal{N}(-3, 1)$ are found to be the best. The best posterior scores are coloured in red.

decision to learn α over the grid (0.2, 0.4, 0.6, 0.8) to improve identifiability. We note that setting δ and α also reduces the computational cost and gives higher marginal likelihood estimates.

5.2 Parameter estimation

The posterior distributions over the parameters β_1, β_2, ϕ, ϵ from the model with the highest marginal likelihood are reported in Figure 5, showing significant departure from their corresponding prior distributions.

We notice that for *E. coli*, $\kappa_1(|H|) = 1$ suggesting a 'frequency dependent' behaviour where the within household transmission increases with the number of colonised individuals in the household. However, we find that $\kappa_1(|H|) = |H|$ for *K. pneumoniae* giving a 'density dependent' behaviour of the force of colonisation with respect to the household size, i.e. a dilutional effect on the force of colonisation as the household size increases (Cocker, Chidziwisano, et al., 2022; Sammarro et al., 2022). For the between households transmission rate, we find $\kappa_2(|H|) = |H|$ for both bacteria, which is plausible since we are modelling contacts with colonised households. Indeed, considering the transmission rate on individual k, we can assume that once a contact between k and household H happens, the contact is going to be successful (resulting in the colonisation of k) according to the probability of meeting a colonised individual in H, which is the proportion of colonised in H.

Another interesting aspect of the study is the comparison of the spatial decay parameter ϕ, which is shown in Figure 6. When comparing bacterial species, we observe that *K. pneumoniae* has a slow decay in space for closer households (Gaussian decay) compared to *E. coli*, *K. pneumoniae* then decays faster compared to *E. coli* for more distant households. This is most likely due to the different ways in which the bacteria transmit. *Escherichia coli* is frequently linked to the environment, especially faecal extraction by humans and animals, hence an individual is more likely to become colonised if living in a contaminated environment, hence we expect it to be more persistent with distance. Colonisation with *K. pneumoniae* typically occurs after direct contact hence it is restricted to the closest neighbours.

Given the sparsity of our data, seasonality (α) is difficult to identify. However, confirmed seasonality in other studies motivates its inclusion here Cocker, Chidziwisano, et al. (2022). Casting this as a model choice problem (Section 5.1), we find that setting $\alpha = 0.6$ or $\alpha = 0.8$ gives the largest marginal likelihood. This supports the existence of a strong seasonal effect on the household transmission rate and so a big variation between wet and dry seasons.

In this study, we find that $\delta = (0, 0, 0)$ gives the best marginal likelihood, from which we conclude that age, gender, and income do not play an important role in driving transmission, which is also consistent with Sammarro et al. (2022), Cocker, Chidziwisano, et al. (2022). In practice, setting $\delta = (0, 0, 0)$ implies that $I(\delta)^{(k)} = 1$, indicating homogeneous transmission rates within each household, consequently suggesting greater importance of the spatial interactions over the individuals' covariates.

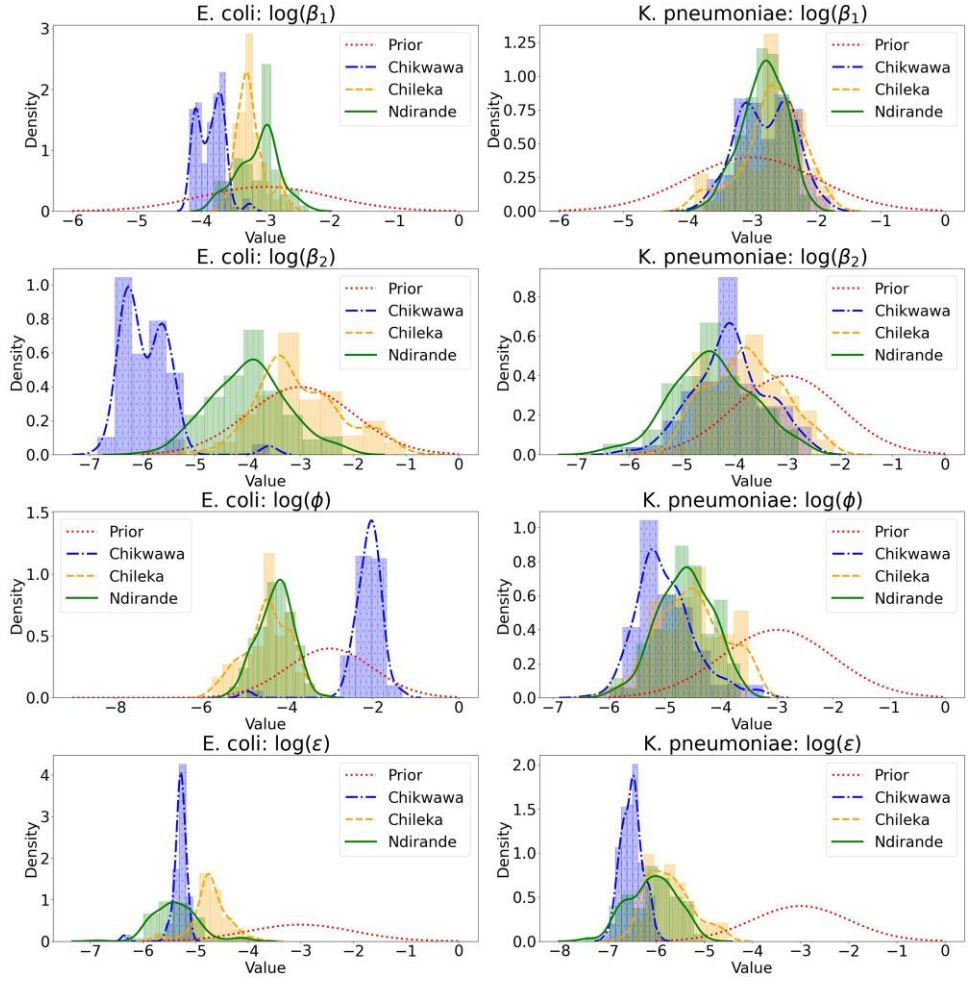

Figure 5. Estimated posterior distributions from the experiments' setting with the highest posteriors for each parameter, bacterial species, and area of study. On the left column *E. coli*, on the right column *K. pneumoniae*. On the rows from top to bottom histograms and KDEs of β_1, β_2, ϕ, ϵ in log scale. Different shapes, and colours refer to different cities and prior distributions.

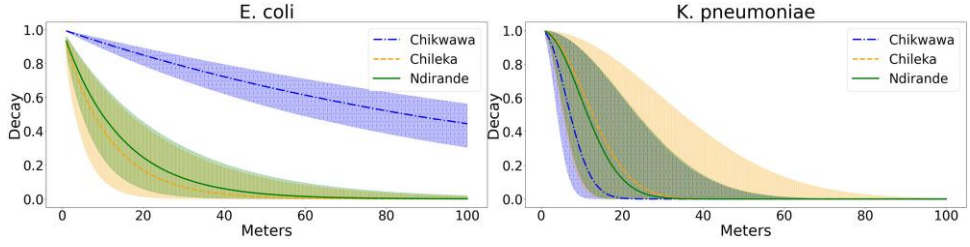

Figure 6. Spatial decay with distance (in metres). On the left *E. coli*, on the right *K. pneumoniae*. Different colours and lines' shapes show different areas. 90% credible intervals are reported in shaded regions, while lines show the medians.

To conclude, the fixed effect ϵ is stronger for *E. coli* than for *K. pneumoniae* (with the exception of Chileka). This is consistent with the archetypal nosocomial nature of *K. pneumoniae*, which spreads mainly through direct person-to-person contact (Podschun & Ullmann, 1998) and we expect the population dynamic to prevail, i.e. most of the infections are explained by the interactions within and between households.

5.3 Spatial and temporal incidence

As already mentioned, SMC^2 provides a sample from the posterior distribution over the parameters of interest, which can then be used to sample from the latent process and estimate how colonisation with the bacteria evolved over time and space.

The effective reproduction number, R_t, is a widely recognised measure of the evolution in time of the epidemic (Nishiura & Chowell, 2009). Ideally, we would like to compute the average number of colonisations arising from an average colonised at time t. However, there is no clear definition of 'average colonised' in a heterogeneous population where individuals colonise and become colonised at different rates. Given that R_t essentially measures the growth of the epidemic at a given time, we approximate the effective reproduction number by the expected number of new colonised over the expected number of new uncolonised. For our application, this alternative definition has several advantages:

- Per each particle of the SMC, we only need to sum the probabilities of susceptible individuals becoming infected and divide by the sum of the probabilities of infected individuals becoming susceptible, precisely:

$$R_t^p = \frac{\sum_{k=1}^{n_I} p_t^{p,(k)} \mathbb{I}\left(C_t^{p,(k)} = 0\right)}{\sum_{k=1}^{n_I} \left(1 - p_t^{p,(k)}\right) \mathbb{I}\left(C_t^{p,(k)} = 1\right)} \tag{6}$$

  where $p_t^{p,(k)}$ follows the definition in equation (3). From (6), we have a sample of R_t's which can be used to estimate the effective reproduction number (e.g. mean, mode) and quantify uncertainty (e.g. quantiles).
- It offers a straightforward interpretation of a growing epidemic whenever R_t is greater than 1 (i.e. more infected than recovered).
- It can be decomposed into an R_t within households and an R_t between households, which includes the fixed effect ϵ.

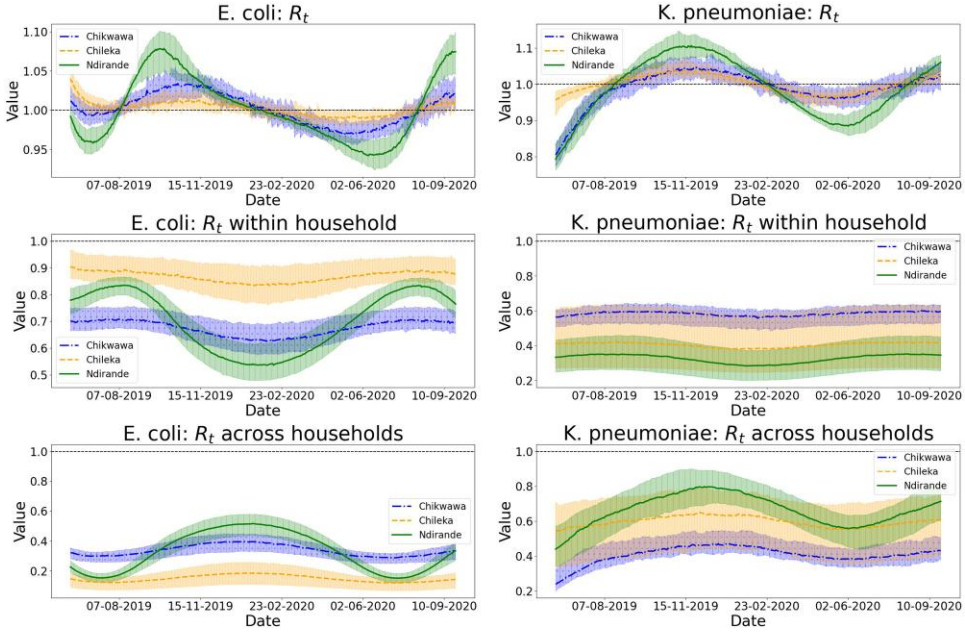

Figure 7. Effective R and its decomposition. *Escherichia coli* is reported in the first column, while *K. pneumoniae* is reported in the second one. The first row shows the effective R, the second row the effective R within households, and the third row is the effective R between households. Different colours are associated with different areas. 90% credible intervals are reported in shaded regions, while lines show the medians.

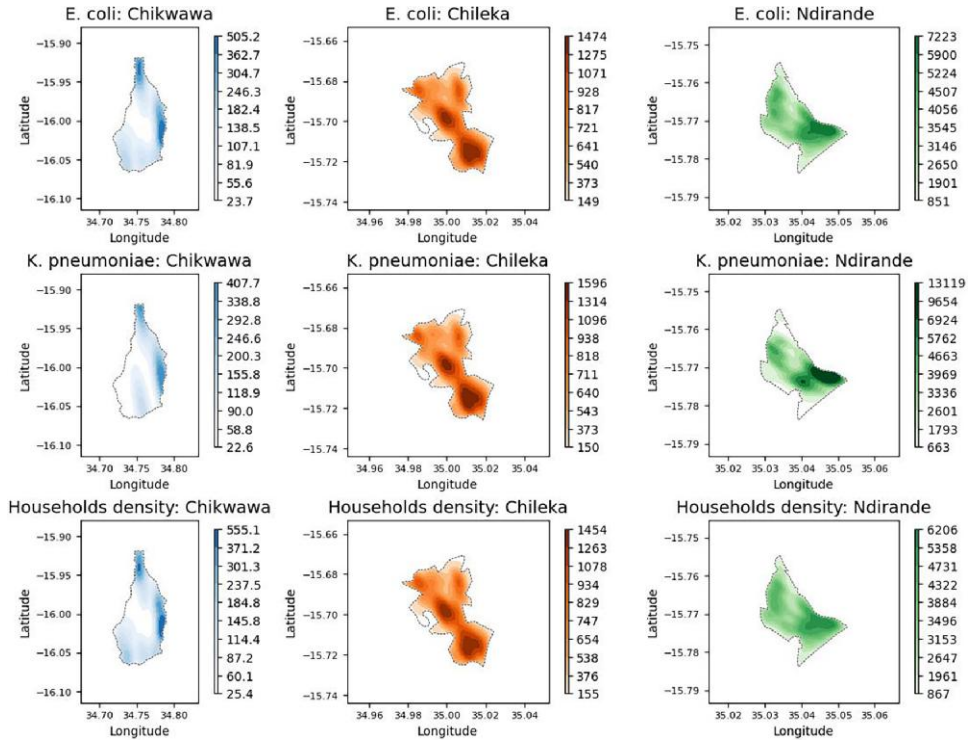

Figure 8. KDE of the spatial density with average colonisation prevalence over time and sampling dimension used as weights. On the columns from left to right: Chikwawa, Chileka, and Ndirande. On the rows from top to bottom: *E. coli, K. pneumoniae,* and the KDE with uniform weights. Different colours are associated with different areas of study and colour maps are the same in each area of study.

The approximate effective R is reported in Figure 7. We can observe that the effective R is fluctuating above and below 1, showing peaks during the wet season for both *E. coli* and *K. pneumoniae*. We notice a strong within household effective R for *E. coli*, which might indicate inadequate hygiene practices within the household. For *K. pneumoniae*, the between households effective R seems higher than the within household one, suggesting the interaction between households to be the highest source of colonisation, probably due to frequent interaction with neighbours and lack of social distancing.

We now turn our attention to the spatial dimension. In order to estimate the spatial density of colonisation, we employ a two-dimensional Kernel density estimation (KDE) on the households, utilising a carefully chosen set of KDE weights. We design a weighting system that quantifies the spread of the epidemic by estimating the average prevalence over time. In practice this is calculated by: (1) running the SMC; (2) computing the prevalence at t per each household; (3) averaging over time; (4) averaging over particles. More details are also available in the supplementary materials. Figure 8 shows the KDE estimates for average prevalence from *E. coli*, average prevalence from *K. pneumoniae* and uniform weighting (KDE estimate of the households' density). Given that areas with a high density of households would appear peaky in the KDE even with moderate weights, by comparing the results with a uniform weighting, we are able to determine if the spread of the bacteria is uniform in space or if it is concentrated in specific areas.

In Chikwawa, we observe that the density of *E. coli* is similar to the households' density, hence the bacteria spreads uniformly in space, while *K. pneumoniae* looks particularly intense in the east of the area. For Chileka, both bacteria's densities are close to the households' density, hence it seems that they spread uniformly in space. In Ndirande, *K. pneumoniae* has a strong prevalence in the southeast of the area, while *E. coli* looks uniform in space.

6 Conclusion

We propose to model the spread of AMR bacteria with a partially observed SIS model, called the UC model, where the transmission rate takes into account: within household contacts, between households contacts and spatial decay, seasonality, individuals' covariates, and environmental effect. We infer the parameters with the algorithm SMC^2, which also allows performing model selection according to the marginal likelihood of the data. The method is not case-specific and can be applied to any epidemics with spatial correlation. We present data on colonisation with ESBL-producing *E. coli* and ESBL-producing *K. pneumoniae* from three areas in Malawi: Chikwawa, Chileka, and Ndirande. As a first step, we impute missing data, by following previous studies (Darton et al., 2017) and then we apply our method to obtain a sample from the posterior distribution over the parameters of interest. From the study, we find *E. coli* to be more persistent in the environment (fixed effect) compared to *K. pneumoniae*, which is in concordance with our knowledge of the bacteria (Cocker, Chidziwisano, et al., 2022; Sammarro et al., 2022). We find that setting a high seasonal effect gives higher marginal likelihoods than smaller values, suggesting significant changes in transmission dynamics throughout the year. The effective R helps quantify the contributions of the within and between households contacts. We also argue that individuals' covariates are not influential in the colonisation process, or at least that our findings prefer models with transmission rates being homogeneous within the households. We also detect geographical hot-spots in the area of Chikwawa for *E. coli* and in the area of Ndirande for *K. pneumoniae*.

There are multiple appealing aspects of this approach. Posterior sampling in epidemiological modelling is a difficult task and it becomes even more challenging when dealing with sparse data. Our method provides an efficient way of performing Bayesian inference on the parameters of a SIS model that is both spatially and temporally sparse. Moreover, it is accompanied by a principled way of performing model selection and supported by strong mathematical results (Chopin et al., 2013). However, the pivotal point of our method is the interpretation, all the parameters and structures in the model have a direct connection with real-world data and it is particularly reassuring that our experimental results agree with the scientific knowledge that we have on the considered bacterial species (Cocker, Chidziwisano, et al., 2022; Sammarro et al., 2022). In addition, our approach provides simple ways of building useful tools for investigating outbreaks and tailoring public health interventions to contain pathogens.

To conclude, there are several strands of research that might follow from this work. One modelling assumption we have made is that the recovery time of each individual is exponential. This simplifies the model and its computation by making it Markov conditional just on the state of each individual: memory efficiency is achieved by only needing to store the current state at each iteration of the simulation. This essentially gives (in continuous time) an exponentially distributed sojourn in the state, though an interesting extension would be to relax this assumption by using a more general class of distribution with positive support. Inference under such a model is possible if we extend the state to include the entire epidemic time series, allowing the model to incorporate individuals' event history into the state update (Boguná et al., 2014; Feng et al., 2019). From a modelling perspective, following Smith et al. (2009), we could automatically detect the relation between transmission rates and host density (density-frequency dependence) by incorporating an additional parameter as the exponent of household size in the within household rate, and even a second parameter as exponent of the household size in the between household rate. This approach would significantly decrease the effort required for model selection by cutting the number of experiments in half, trading off with additional parameters to infer. There are numerous technical questions related to SMC^2 and determining the optimal selection of tuning parameters and proposal distributions. This study is limited to an SIS model, however, adapting it to more intricate compartmental models is straightforward and the methodology can be extended to any epidemiological model. Despite being utilised in the field of epidemiology, SMC^2 can be broadly applied to any data sets with spatial interactions, albeit the computational cost may become prohibitive.

Conflict of interest: The authors declare no conflict of interest.

Funding

This work is supported by MR/S004793/1 (Drivers of Resistance in Uganda and Malawi: The DRUM Consortium), EPSRC grants EP/R018561/1 (Bayes4Health) and EP/R034710/1 (CoSInES).

Data availability

The synthetic population and the cleaning procedure are available at: https://github.com/LorenzoRimella/SMC2-ILM. More details on how to generate the synthetic population can be found at: https://zenodo.org/record/7007232#.Yv5EZS6SmUk.

The above link contains an anonymised version of the STRATAA data set to avoid copyright issues. The anonymised version of the data provides the same synthetic population distribution. The authors can provide the real data, after authorisation from the data owners, if needed.

Supplementary material

Supplementary material are available at *Journal of the Royal Statistical Society: Series C* online.

References

Andrieu C., De Freitas N., Doucet A., & Jordan M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*(1), 5–43. https://doi.org/10.1023/A:1020281327116

Andrieu C., Doucet A., & Holenstein R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(3), 269–342. https://doi.org/10.1111/j.1467-9868.2009.00736.x

Barnes C. P., Filippi S., Stumpf M. P., & Thorne T. (2012). Considerate approaches to constructing summary statistics for ABC model selection. *Statistics and Computing*, *22*(6), 1181–1197. https://doi.org/10.1007/s11222-012-9335-7

Boguná M., Lafuerza L. F., Toral R., & Serrano M. Á. (2014). Simulating non-markovian stochastic processes. *Physical Review E*, *90*(4), 042108.

Carpenter J., Clifford P., & Fearnhead P. (1999). Improved particle filter for nonlinear problems. *IEE Proceedings-Radar, Sonar and Navigation*, *146*(1), 2–7. https://doi.org/10.1049/ip-rsn:19990255

Chipeta M., Terlouw D., Phiri K., & Diggle P. (2017). Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, *28*(1), e2425. https://doi.org/10.1002/env.2425

Chopin N., Jacob P. E., & Papaspiliopoulos O. (2013). SMC^2: An efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(3), 397–426. https://doi.org/10.1111/j.1467-9868.2012.01046.x

Chopin N., & Papaspiliopoulos O. (2020). *An introduction to sequential Monte Carlo*. Springer.

Cocker D., Chidziwisano K., Mphasa M., Mwapasa T., Lewis J. M., Rowlingson B., Sammarro M., Bakali W., Salifu C., Zuza A., Charles M., Mandula T., Maiden V., Amos S., Jacob S. T., Kajumbula H., Mugisha L., Musoke D., Byrne R. L., & Lester R. (2022). Investigating risks for human colonisation with extended spectrum beta-lactamase producing e. coli and k. pneumoniae in malawian households: A one health longitudinal cohort study.

Cocker D., Sammarro M., Chidziwisano K., Elviss N., Jacob S. T., Kajumbula H., Mugisha L., Musoke D., Musicha P., Roberts A. P., & Rowlingson B. (2022). Drivers of resistance in Uganda and Malawi (DRUM): A protocol for the evaluation of one-health drivers of extended spectrum beta lactamase (ESBL) resistance in low-middle income countries (LMICs). *Wellcome Open Research*, *7*(55), 55. https://doi.org/10.12688/wellcomeopenres.17581.1

Darton T. C., Meiring J. E., Tonks S., Khan M. A., Khanam F., Shakya M., Thindwa D., Baker S., Basnyat B., Clemens J. D., & Dougan G. (2017). The STRATAA study protocol: A programme to assess the burden of enteric fever in Bangladesh, Malawi and Nepal using prospective population census, passive surveillance, serological studies and healthcare utilisation surveys. *BMJ Open*, *7*(6), e016283. https://doi.org/10.1136/bmjopen-2017-016283

Deardon R., Brooks S. P., Grenfell B. T., Keeling M. J., Tildesley M. J., Savill N. J., Shaw D. J., & Woolhouse M. E. (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, *20*(1), 239.

Fearnhead P., & Prangle D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 419–474. https://doi.org/10.1111/j.1467-9868.2011.01010.x

Feng M., Cai S.-M., Tang M., & Lai Y.-C. (2019). Equivalence and its invalidation between non-markovian and markovian spreading dynamics on complex networks. *Nature Communications*, *10*(1), 3748. https://doi.org/10.1038/s41467-019-11763-z

Jewell C. P., & Brown R. G. (2015). Bayesian data assimilation provides rapid decision support for vector-borne diseases. *Journal of the Royal Society Interface*, *12*(108), 20150367. https://doi.org/10.1098/rsif.2015.0367

Jewell C. P., Kypraios T., Neal P., & Roberts G. O. (2009). Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, *4*(3), 465–496. https://doi.org/10.1214/09-BA417

Johansen A. M., & Doucet A. (2008). A note on auxiliary particle filters. *Statistics & Probability Letters*, *78*(12), 1498–1504.

Ju N., Heng J., & Jacob P. E. (2021). Sequential Monte Carlo algorithms for agent-based models of disease transmission.

Keeling M. J., & Rohani P. (2011). *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kypraios T., Neal P., & Prangle D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Mathematical Biosciences*, *287*, 42–53. https://doi.org/10.1016/j.mbs.2016.07.001

Lewis J. M., Lester R., Garner P., & Feasey N. A. (2019). Gut mucosal colonisation with extended-spectrum beta-lactamase producing enterobacteriaceae in sub-saharan africa: A systematic review and meta-analysis. *Wellcome Open Research*, *4*, 160.

Lewis J. M., Mphasa M., Banda R., Beale M. A., Heinz E., Mallewa J., Jewell C., Faragher B., Thomson N. R., & Feasey N. A. (2019). Dynamics of gut mucosal colonisation with extended spectrum beta-lactamase producing enterobacterales in Malawi. *PMC*, *4*(160), 31976380.

Nishiura H., & Chowell G. (2009). The effective reproduction number as a prelude to statistical estimation of time-dependent epidemic trends. In: Chowell, G., Hyman, J. M., Bettencourt, L. M. A., Castillo-Chavez, C. (eds.), *Mathematical and Statistical Estimation Approaches in Epidemiology*. Dordrecht: Springer. https://doi.org/10.1007/978-90-481-2313-1_5

Parry M., Gibson G. J., Parnell S., Gottwald T. R., Irey M. S., Gast T. C., & Gilligan C. A. (2014). Bayesian inference for an emerging arboreal epidemic in the presence of control. *Proceedings of the National Academy of Sciences*, *111*(17), 6258–6262. https://doi.org/10.1073/pnas.1310997111

Pitt M. K., & Shephard N. (1999). Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, *94*(446), 590–599. https://doi.org/10.1080/01621459.1999.10474153

Podschun R., & Ullmann U. (1998). Klebsiella spp. as nosocomial pathogens: Epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clinical Microbiology Reviews*, *11*(4), 589–603. https://doi.org/10.1128/CMR.11.4.589

Prangle D., Fearnhead P., Cox M. P., Biggs P. J., & French N. P. (2014). Semi-automatic selection of summary statistics for ABC model choice. *Statistical Applications in Genetics and Molecular Biology*, *13*(1), 67–82. https://doi.org/10.1515/sagmb-2013-0012

Probert W. J., Jewell C. P., Werkman M., Fonnesbeck C. J., Goto Y., Runge M. C., Sekiguchi S., Shea K., Keeling M. J., Ferrari M. J., & Tildesley M.J. (2018). Real-time decision-making during emergency disease outbreaks. *PLoS Computational Biology*, *14*(7), e1006202. https://doi.org/10.1371/journal.pcbi.1006202

Rabiner L., & Juang B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16. https://doi.org/10.1109/MASSP.1986.1165342

Rimella L., Jewell C., & Fearnhead P. (2022). Approximating optimal SMC proposal distributions in individual-based epidemic models.

Robert C. P., & Casella G. (2004). *Monte Carlo statistical methods* (*Vol. 2*). Springer.

Sammarro M., Rowlingson B., Cocker D., Chidziwisano K., Jacob S. T., Kajumbula H., Mugisha L., Musoke D., Lester R., Morse T., Feasey N., & Jewell C. (2022). Risk factors, temporal dependence, and seasonality of human ESBL-producing *E. coli* and *K. pneumoniae* colonisation in Malawi: A longitudinal model-based approach.

Smith M. J., Telfer S., Kallio E. R., Burthe S., Cook A. R., Lambin X., & Begon M. (2009). Host–pathogen time series data in wildlife support a transmission function between density and frequency dependence. *Proceedings of the National Academy of Sciences*, *106*(19), 7905–7909. https://doi.org/10.1073/pnas.0809145106

Sunnåker M., Busetto A. G., Numminen E., Corander J., Foll M., & Dessimoz C. (2013). Approximate Bayesian computation. *PLoS Computational Biology*, *9*(1), e1002803.

Vlek A. L., Cooper B. S., Kypraios T., Cox A., Edgeworth J. D., & Auguet O. T. (2013). Clustering of antimicrobial resistance outbreaks across bacterial species in the intensive care unit. *Clinical Infectious Diseases*, *57*(1), 65–76. https://doi.org/10.1093/cid/cit192

Zucchini W., & MacDonald I. L. (2009). *Hidden Markov models for time series: An introduction using R*. Chapman and Hall/CRC.