A mechanistic spatiotemporal model for drug resistant infections

Tamsin E. Lee[1,2,*]

1 Swiss Tropical and Public Health Institute, Basel, Switzerland.
2 University of Basel, Basel, Switzerland.

*** tamsin.e.lee@gmail.com**

Abstract

When drug resistance is suspected to be in a region, patients in the region are sampled and the suspicion is confirmed. This biased sampling limits our ability to capture underlying dynamics, meaning strategies to lengthen the lifespan of drugs are reactionary, not proactive.

Testing for drug resistant infections is becoming easier and cheaper, therefore we should revisit sampling decisions. We present a hierarchical mechanistic Bayesian model, and apply it to a simulated dataset, where we sample between 5% and 30% of the population in a biased and unbiased manner. We show that unbiased spatiotemporal data on the presence of drug resistant infections, combined with our model, highlights underlying dynamics.

Our mechanistic model is more accurate than a generalised additive model with space and time components. Moreover, highlighting underlying dynamics creates novel strategies that lengthen the lifespan of drugs. In low to middle income countries, generally, drug resistance emerges into a population from hotspots such as treatment centres (perhaps the use of sub-standard drugs), or major transport hubs, and then resistance spreads throughout the population. Using our model, we rank resistance hotspots, enabling resources to be targeted - such as verifying the quality of drugs at a particular health care centre.

Keywords: drug resistance hotspots, hierarchical mechanistic Bayesian model, generalised additive model.

Introduction

Pathogens such as bacteria, viruses and parasites are continually evolving. When a drug resistant mutation occurs by chance, it is given a survival advantage if selection pressure (from antimicrobials, such as antibiotics, antivirals and antimalarials) kills drug sensitive pathogens, leaving the drug resistant pathogens within a host. These pathogens can spread throughout a population, thereby leading to additional deaths and treatment costs. To safeguard the efficacy of drugs, capturing the underlying transmission dynamics at the population level is a priority, and mechanistic models are an important tools to achieve this [12].

In the early 2000s, hierarchical mechanistic models were developed for predicting the spread of ecological process [24, 25]. Based on careful assumptions which influence the spatiotemporal dynamics and data collection process, each model is tailored to the

specific problem and dataset, making them sophisticated and powerful models [3]. Using previous detailed documentation [9, 26] of such models, this paper demonstrates the potential of hierarchical mechanistic models to capture the underlying transmission dynamics of drug resistant pathogens. The authors believe this is the first time such a model is applied in this field.

Over 70% of the 273 modelling studies on antimicrobial resistance (AMR) focussed on five diseases: human immunodeficiency virus (HIV), influenza virus, *Plasmodium falciparum* (malaria), *Mycobacterium tuberculosis* (TB), and methicillin-resistant *Staphylococcus aureus* (MRSA) [16]. Three of these diseases, HIV, malaria, and TB, mostly affect people who live in countries where the health care received varies in factors that affect drug resistance, such as using low quality drugs or out of date recommendations, making identifying hotspots a critical part of the strategy.

Identifying hotspots for clinical investigation may have prevented the spread of artemisinin resistance in Africa [22]. Modelling the transmission dynamics at the population level requires spatiotemporal data. With regards to resistance to *Plasmodium falciparum*, the predominant malaria parasite, there is a solid understanding of the mutations responsible for resistance to different antimalarials. For example, resistance to artemisinin, the latest antimalarial in use, occurs when there is a mutation on the Kelch-13 gene. Such mutations are identified within infected patients by molecular marker studies.

WorldWide Antimalarial Resistance Network (WWARN) [27] has assembled a database of molecular marker studies for resistance to artemisinin, and another current antimalarial, sulphadoxine pyrimethamine (SP). This global spatiotemporal data, from 141 studies (for artemisisin) and 165 studies (for SP), is provided as an opensource spreadsheet and interactive map. By focussing on molecular marker studies, WWARN uses the most up to date antimalarial drug resistance surveillance as the prevalence of a mutation can be readily quantified from an analysis of fingerprick blood samples from infected humans [19].

Despite WWARN covering antimalarial resistance surveillance, the data is still insufficient to capture underlying mechanisms because currently molecular marker studies are only carried out in a region when we suspect drug resistance is present. When drug resistant pathogens are present it is likely already too late to mitigate their spread [2]. Meaning that once drug resistance is detected, the course of the spread of drug resistant infections is largely determined already. Early detection and response to antimalarial drug resistance is imperative to eliminate malaria [18]. Thus predicting ahead will support mitigation measures.

As the surveillance improves, we will be able to use mechanistic models to answer pertinent questions regarding the underlying transmission dynamics of drug resistant pathogens. Critical thought regarding the sampling approach can support surveillance and optimise resource use. Therefore to maximise the potential of molecular monitoring we should address the questions we can answer, and demonstrate the learning potential through modifying leading mechanistic models currently in use in other fields. Here we provide a specific question, and demonstrate how mechanistic models can answer such a question.

When a new drug is introduced into a region, it is necessary to monitor at the same time so that action is preemptive, not reactive. This stance is shared by many epidemiologists. When resistance to artemisinin, the latest antimalarial, was present in South East Asia but not yet in Africa, a paper called for early warning and detection systems in Africa that targeted hotspots for clinical investigations [22]. A hotspot of resistant pathogens can be defined as a location such as health care centre, a hospital, or a major transport hub, that largely contribute to the emergence of the resistant pathogen in a region.

We developed a hierarchical mechanistic Bayesian model that quantifies which hotspots are contributing the most resistant pathogens into the population, and the spread of these pathogens. The model accounts for spatial factors which affect the spread of the resistant pathogens, such as disease prevalence. The model is an adaptation of a model on the spread of chronic wasting disease in white tailed deer in the southwestern portion of Winconsin [7].

Ranking the hotspots can be used for strategy decisions with regards to checking the quality and procedures of a given health care centre, or, in the case of a transport hub being a major contributor of resistant pathogens, determining that drug resistant infections are being imported into the region. As well as quantifying the hotspots, the model can predict and forecast the true density of drug resistant infections in the region.

To investigate who is sampled for data collection, we investigated how different patient sampling approaches capture spatiotemporal dynamics of resistant pathogens. Namely, we compared an unbiased (random) approach with the current approach, which is biased towards sampling patients who are likely to have drug resistant infections. Unbiased sampling is inline with recent recommendations to mitigate antimalarial drug resistance, which recommends routine surveillance to complement the often sparse and outdated data from therapeutic efficacy studies [17].

Although this modelling study focussed on drug resistant *Plasmodium falciparum*, detecting hotspots is relevant more broadly. For example, consider that until 2015 the WHO recommended that a woman living with HIV takes antiretrovirals during pregnancy (to prevent their babies from becoming infected), but stops after delivery [14]. This recommendation was dropped because surveys conducted in nine countries in sub-Saharan Africa between 2012 and 2018 found that over half of the infants newly diagnosed with HIV carry a virus that is resistant to the standard class of drugs [23]. Clearly, a health care centre that is not up to date with these recommendations needs to be detected.

The methods presented here are not limited to a particular drug resistant pathogen. The model is general to any drug resistant pathogen where there is spatiotemporal data about its occurrence, whether the data is from molecular markers studies in the case of malaria, or by genome sequencing in the case of resistance to antivirals [13], or by analysing multiplication bacteria rates from samples [4].

Materials and Methods

We modify the hierarchical mechanistic model from [7], which modelled the spread of chronic wasting disease in white tailed deer in the southwestern portion of Winconsin. This model uses presence/absence data, with a single origin hotspot based on where the disease was first detected. This model assumes the hotspot only contributed the disease to the population at the first time interval. In our modified model, we allow for multiple hotspots, which continually contribute drug resistant pathogens into the population. Furthermore we use count data (not presence/absence), as in [3, 10].

The model uses an aggregate of resistance surveillance studies. For each study, there is the number of infected patients, and from these tested patients, the number who carry drug resistant pathogens (identified by means such as a carrying a mutation). For brevity, we refer to sampled patients who carried drug resistant pathogens as positive patients, and those who did not carry drug resistant pathogens as negative patients.

We assume that each location, $\mathbf{s} = (s_1, s_2)$, has a weight that depends on its distance from a hotspot. This weight is greater when there is a greater chance that an infected person develops a mutation that infers resistance to treatment. The weights, $\omega(\mathbf{s})$, are

defined by

$$\omega(\mathbf{s}) \quad = \quad \sum_{n=1}^{N} \frac{\theta_n e^{-(|\mathbf{s}-\mathbf{d_n}|^2)/\phi_n^2}}{\int_{\mathcal{S}} e^{-(|\mathbf{s}-\mathbf{d_n}|^2)/\phi_n^2}\, \mathrm{d}\mathbf{s}}, \tag{1}$$

where the magnitude of resistant pathogens contributed by hotspot n is θ_n, and these resistant pathogens disperse at rate ϕ_n from the hotspot. Eq. 1 is the sum, over n, of scaled bivariate Gaussian kernels with compact (truncated) support centred at a point with coordinate $\mathbf{d_n}$ where $|\mathbf{s} - \mathbf{d_n}|$ is the distance. The units of this distance depends on the size of the region being investigated, for example, 10 km by 10km. An example of $\omega(\mathbf{s})$, where there are five hotspots ($N = 5$) is shown in Fig. 1.

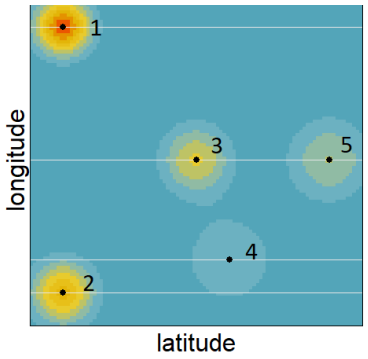

Figure 1. An example of the spatial weightings which are greater (red) where drug-resistant infections are more likely via five hotspots with magnitude $\theta_1 \geq \theta_2 \geq \theta_3 \geq \theta_4 \geq \theta_5$ and dispersal $\phi_4 \geq \phi_5 \geq \phi_3 \geq \phi_2 \geq \phi_1$ (see the definition of $\omega(\mathbf{s})$ from Eq. 1).

We now provide details of the hierarchical mechanistic model which has separate components to capture uncertainty in data collection, the spatiotemporal transmission dynamics, and uncertainty in the parameters. Following this, we provide details of the numerical implementation, beginning with the algorithm. Then, to thoroughly demonstrate our model, we simulate data over a region that is split into 10,000 grid points (a square that is 100 by 100).

For comparison, we applied a generalised additive model (GAM) to the same simulated data. Generalised additive models are a well-developed and sophisticated statistical tool, where spatial and temporal components can be explicitly included, and the effect of covariates are readily quantified. As with the hierarchical mechanistic model, the GAM produces an estimate for the density of positive patients over the whole region over different times. However, unlike the hierarchical mechanistic model, it cannot explicitly provide magnitude and dispersion measures for the resistance hotspots.

The hierarchical mechanistic Bayesian model

The model comprises three levels. First, the data level which states that the data is depends on the sampling probability and the density of positive patients. Second, the process level which captures the spatiotemporal mechanisms of the density of positive patients. Third, the parameter level which estimates the model parameters.

Data level

The data is an aggregate of M studies, which each have a unique location and time. The M studies are indexed by $i = 1, 2, \ldots, M$. We represented the data as $y_i \in \boldsymbol{Z}$,

which is the number of positive patients in study i. We model this as

$$y_i \sim \text{Poisson}(\lambda_i), \tag{2}$$

where $\lambda_i > 0$ is a latent spatiotemporal process defined by

$$\lambda_i = u(\mathbf{s}_i, t_i) e^{\mathbf{x}_i' \beta}, \tag{3}$$

where $u(\mathbf{s}_i, t_i)$ is the density of positive patients, at the location and time of study i (determined by the Process level). The total record of observed counts of positive patients is $\boldsymbol{Y}$. A patient at the location and time of the study i has a probability of being sampled. This probability depends on patient covariates, such as age, which are captured by $\mathbf{x}_i$. The β in Eq. 3 are the regression coefficients (determined by the Parameter level) for the individual covariates $\mathbf{x}_i$. Therefore, the data level, Eq. 2 and Eq. 3, states that the data y_i depends on the density of positive patients, and a probability of being sampled.

Process level

The density of positive patients is a dynamic process captured by the partial differential equation (PDE)

$$\frac{\partial}{\partial t} u(\mathbf{s}, t) \quad = \quad \left(\frac{\partial^2}{\partial s_1^2} + \frac{\partial^2}{\partial s_2^2} \right) \left[\mu(\mathbf{s}) u(\mathbf{s}, t) \right] + \rho(\mathbf{s}) u(\mathbf{s}, t) + \omega(\mathbf{s}), \tag{4}$$

where the diffusion rate $\log(\mu(\mathbf{s})) = \alpha_0 + \mathbf{a}(\mathbf{s})'\alpha$ depends on spatial covariates captured by $\mathbf{a}(\mathbf{s})$, and the growth rate $\rho(\mathbf{s}) = \gamma_0 + \mathbf{c}(\mathbf{s})'\gamma$ depends on spatial covariates captured by $\mathbf{c}(\mathbf{s})$. The $\alpha_0, \alpha, \gamma_0$, and γ are the regression coefficients (determined by the Parameter level) for the spatial covariates $\mathbf{a}(\mathbf{s})$ and $\mathbf{c}(\mathbf{s})$. The first two terms of Eq. 4, the diffusion and growth terms, make up the ecological diffusion equation, which is often used to model animal movement using abundance data [6, 8, 11].

For our application here, we no longer use the terminology 'diffusion' and 'growth' rate because they are misleading. In terms of the spread of drug resistant pathogens, $\mu(\mathbf{s})$ relates to transmission to neighbouring regions and the $\rho(\mathbf{s})$ relates to transmission within a local area, see Fig. 2. With regards to the choice of spatial covariates, the transmission of drug resistant pathogens depends on the prevalence of the disease, thus $\mathbf{a}(\mathbf{s})$ and $\mathbf{c}(\mathbf{s})$ include the disease prevalence.

The last term of Eq. 4, $\omega(\mathbf{s})$, is an additional term we added to account for the underlying time-independent component which assumes mutations associated with resistance originates from hotspots, see Eq. 1. We use zero boundary conditions, and initial conditions $u(\mathbf{s}, 0) = \omega(\mathbf{s})$, meaning that initially, the only influence on the emergence of drug resistant pathogens is the distance from a hotspot.

Parameter level

To complete the Bayesian specification of the spatiotemporal model, we describe the probability models for the parameters discussed in the data and process levels. The parameters which require prior distributions include the magnitude and dispersal of the resistance hotspots, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, which we assign priors $\theta_n \sim \text{TN}(0, 10^6)$ and $\phi_n \sim \text{TN}(0, 10^6)$ where $n \in [1, N]$ and TN refers to a normal distribution truncated below zero).

For regression coefficients $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ we used priors drawn from a normal distribution with mean 0 and variance 10, as in [7]: $\boldsymbol{\beta} \sim N(0, 10\mathbf{I})$, $\alpha_0 \sim N(0, 10)$, $\boldsymbol{\alpha} \sim N(0, 10\mathbf{I})$, $\gamma_0 \sim N(0, 10)$, and $\boldsymbol{\gamma} \sim N(0, 10\mathbf{I})$, where $\mathbf{I}$ is the identity matrix.

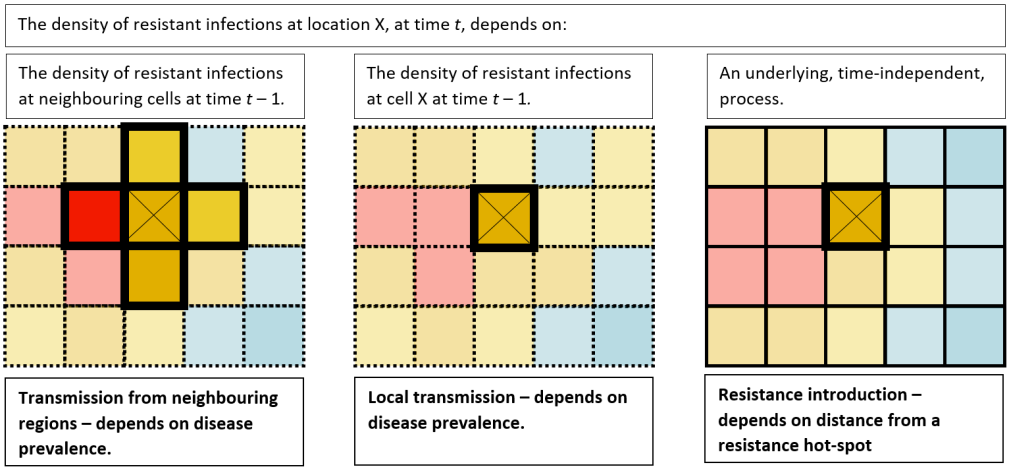

Figure 2. Visual description of each of three terms in Eq. 4 in relation to the spread of resistant pathogens. The highlighted boxes indicate the locations that affect the density at location X.

Numerical implementation

The parameters are determined using a Monte Carlo Markov Chain (MCMC), coded in R. From these parameter values, we can recover the spatiotemporal density of positive patients. The model algorithm is provided below, followed by details about the creation of the simulated data we used to demonstrate the model application. All codes are provided on GitHub. They are built upon the codes provided by [7]. We run the model for 250,000 iterations and remove the first 1,000 iterations due to burn-in.

The algorithm

For each MCMC iteration, the estimates for the parameters are updated using Metropolis Hastings. The algorithm is,

1. Set initial values for $\boldsymbol{\alpha}$, β, $\boldsymbol{\gamma}$, $\boldsymbol{\theta}$, $\boldsymbol{\phi}$
2. **while** $l < m$ **do**
3. update $u(\mathbf{s}, t)$
4. sample $[\boldsymbol{\theta}, \boldsymbol{\phi} | \boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}]$
5. update $u(\mathbf{s}, t)$
6. sample $[\boldsymbol{\alpha} | \beta, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}]$
7. update $u(\mathbf{s}, t)$
8. sample $[\beta | \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\phi}]$
9. sample $[\boldsymbol{\gamma} | \boldsymbol{\alpha}, \beta, \boldsymbol{\theta}, \boldsymbol{\phi}]$
10. **end while**

Updating $u(\mathbf{s}, t)$ requires solving a PDE (Eq. 4) over a fine-scale grid, which is very computationally expensive. We apply the homogenisation technique [6, 8, 20] which means that each time $u(\mathbf{s}, t)$ is updated (steps 3, 5, and 7 above), the process is

(a) Calculate $\mu(\mathbf{s})$, $\rho(\mathbf{s})$ and $e^{\mathbf{x}_i'\beta}$ using current estimates for the regression coefficients.

(b) Convert $\mu(\mathbf{s})$ and $\rho(\mathbf{s})$ to a coarser grid.

(c) Solve the PDE (Eq. 4) on the coarser grid.

(d) Convert the solution on the coarse grid to the original fine-scale grid.

Creating the simulated data

To demonstrate the utility of the model we created simulated data over a unit square with five resistance hotspots. We assume that the hotspot locations are known, but this is not a restriction of the model. The locations are $d_1 = (0.1, 0.9)$, $d_2 = (0.1, 0.1)$, $d_3 = (0.5, 0.5)$, $d_4 = (0.6, 0.2)$ and $d_5 = (0.9, 0.5)$.

We set the magnitudes of the hotspots to $\theta_1 = 80$, $\theta_2 = 70$, $\theta_3 = 65$, $\theta_4 = 60$ and $\theta_5 = 60$. We set the dispersal of each hotspot to $\phi_1 = 0.08$, $\phi_2 = 0.09$, $\phi_3 = 0.1$, $\phi_4 = 0.15$ and $\phi_5 = 0.12$. These values were chosen to include an example where the hotspot with the greatest magnitude (hotspot 1) is isolated so resistance does not disperse far, see Fig. 1.

In our demonstration, the covariates $\mathbf{a}(\mathbf{s})$ and $\mathbf{c}(\mathbf{s})$ are each a single covariate that varies in space. When estimating drug resistance, the main covariate should be the prevalence of the disease which has a non-random spatial pattern. To simply represent this, both $\mathbf{a}(\mathbf{s})$ and $\mathbf{c}(\mathbf{s})$ are set to values between -0.5 and 0.5 which are ordered so that the greater values are at the bottom of the square, graduating towards the top of square, see Fig. 3. Although we generated $\mathbf{a}(\mathbf{s})$ and $\mathbf{c}(\mathbf{s})$ in exactly the same manner for our demonstration, we continue with distinct notation for clarity, and to serve as a reminder that the model allows for two distinct spatial covariates which affect neighbouring and local transmission differently.

The transmission to neighbouring regions $\mu(\mathbf{s})$ depends on the spatial covariate $\mathbf{a}(\mathbf{s})$,

$$\log(\mu(\mathbf{s})) = \alpha_0 + \mathbf{a}(\mathbf{s})'\alpha, \tag{5}$$

where we set the coefficients to $\alpha_0 = -8$ and $\alpha_1 = 1$. The local transmission $\rho(\mathbf{s})$ depends on the spatial covariate $\mathbf{c}(\mathbf{s})$,

$$\rho(\mathbf{s}) = \gamma_0 + \mathbf{c}(\mathbf{s})'\gamma, \tag{6}$$

where we set the coefficients to $\gamma_0 = 0.2$ and $\gamma_1 = 0.1$. The simulated density of positive patients for 20 years, using Eq. 4, gives Fig. 4. Notice that although hotspot 1 (the top left hotspot) has the highest magnitude, the region surrounding hotspot 3 (the middle hotspot) has the highest density of positive patients. This demonstrates that simply observing the presence of positive patients can mask underlying spatial dynamics.

The probability that a positive patient is sampled is given by

$$e^{\mathbf{x}_i'\beta}, \tag{7}$$

where $\mathbf{x}_i$ contains patient information, such as the patient age. In our demonstration, $\mathbf{x}_i$ is a single covariate containing random values between -0.5 and 0.5, and $\beta = -10$ (note that there is not an intercept term in $\mathbf{x}_i$). With biased sampling, as is the current approach, we set the sampling probability to be one where resistance is present, and random otherwise, see the top plot of Fig. 5. With unbiased sampling, each location is equally likely to be sampled, see the bottom plot of Fig. 5. We assume that the same proportion of the region is sampled each year, either 5%, 10%, 15% 20% and 30%, however this is not a restriction of the model.

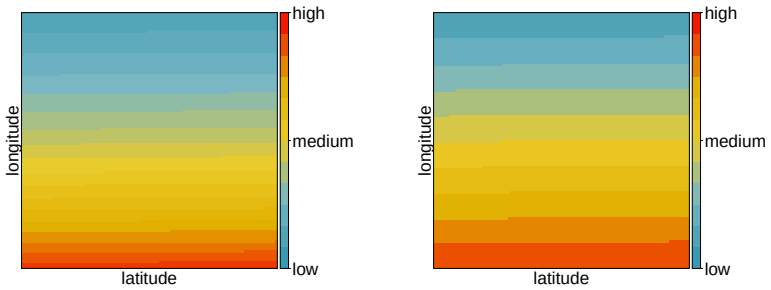

Figure 3. The transmission of drug resistant pathogens depends on spatial covariates $\mathbf{a}(\mathbf{s})$ and $\mathbf{c}(\mathbf{s})$, which represent covariates such as disease prevalence. Left: The transmission to neighbouring regions, $\mu(\mathbf{s})$, where $\log(\mu(\mathbf{s})) = \alpha_0 + \mathbf{a}(\mathbf{s})'\alpha$. Right: The local transmission, $\rho(\mathbf{s})$, where $\rho(\mathbf{s}) = \gamma_0 + \mathbf{c}(\mathbf{s})'\gamma$. Since this is a demonstration, the pattern is relevant (transmission of drug resistant pathogens is easier in the south than in the north), but the actual values are not, so they are omitted for clarity.

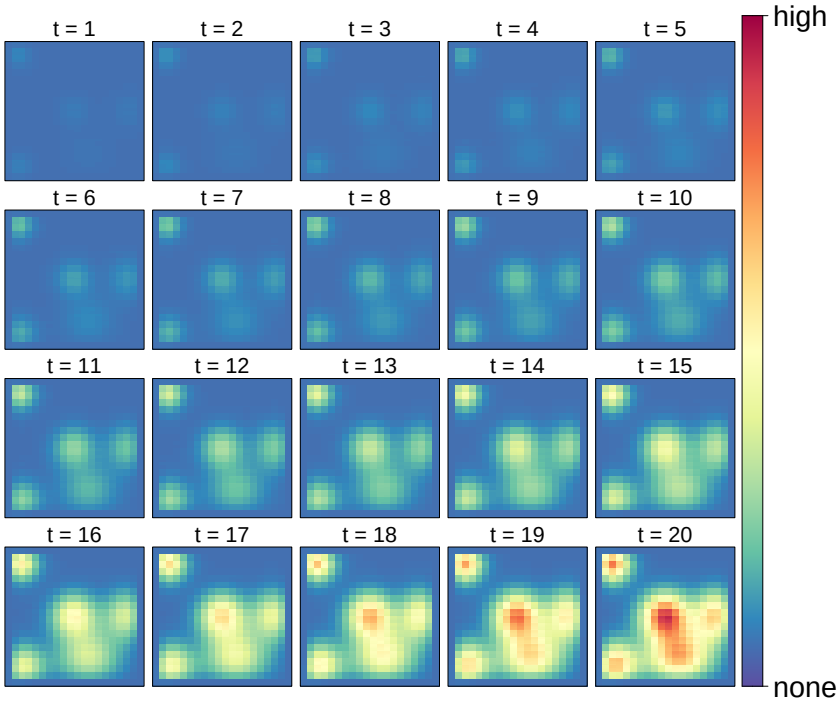

Figure 4. The simulated density of drug resistant infections. The values are not relevant for this simulated example so they are omitted for clarity.

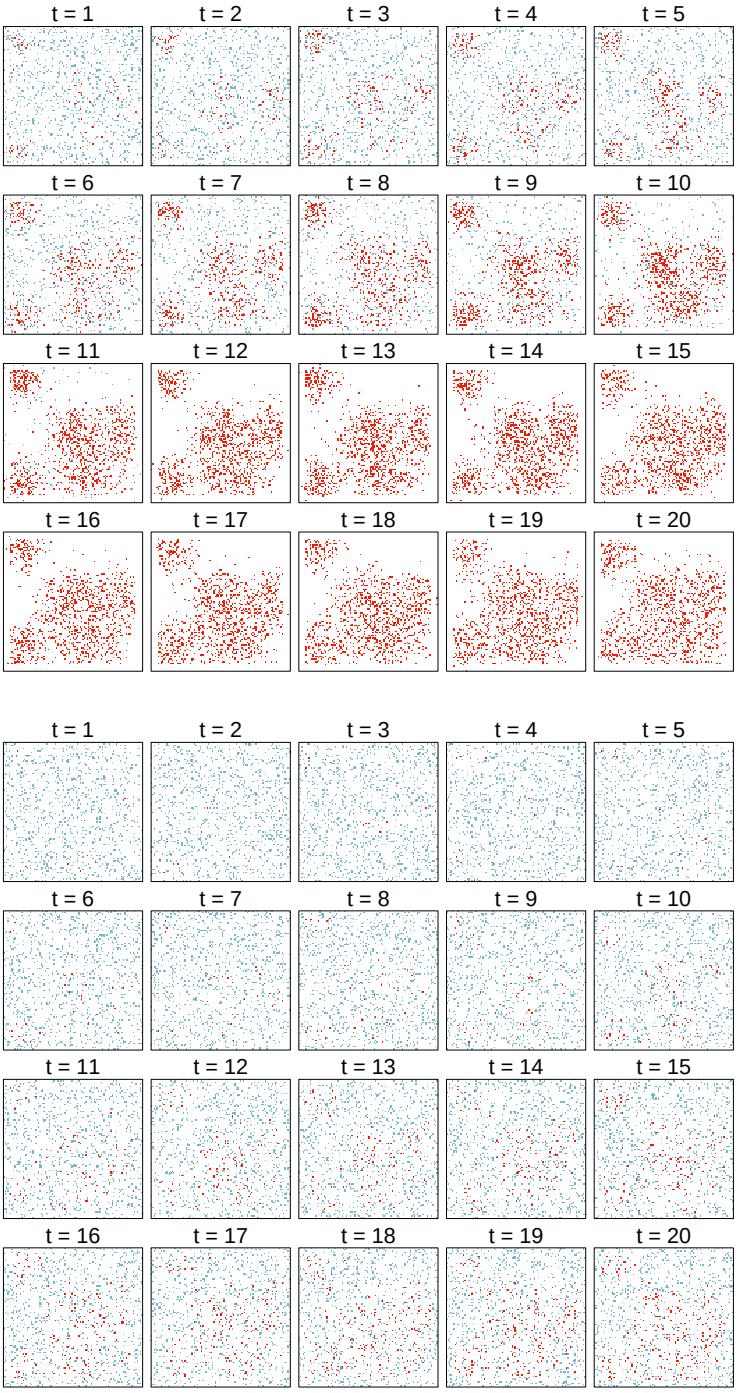

Figure 5. Bias and unbias sampling where 10% of region is sampling. For biased sampling, patients who are assumed to have a drug-resistant infection are more likely to be sampled. Red dots refer to a patient testing positive for a drug-resistant infection. Blue dots refer to a patient testing negative for a drug-resistant infection.

Comparison with generalised additive models

As with the hierarchical mechanistic model, a spatiotemporal generalised additive model (GAM) produces an estimate for the whole region over different times, but it cannot explicitly provide magnitude and dispersion measures for the resistance hotspots. Our GAM assumes that the data, $y_i \in \boldsymbol{Z}$ being the number of positive patients in study i, is modelled such that

$$y_i \sim \text{Poisson}(\lambda_i) \text{ where } g(\lambda_i) = \boldsymbol{x}_i\boldsymbol{\beta} + \eta_s + \eta_t. \tag{8}$$

The probability that a patient is positive, λ_i, is transformed using a link function $g(\cdot)$ and depends on both the individual data (such as the patient age), and spatial covariates (such as disease prevalence) which are included in the vector $\boldsymbol{x}_i$. Note that unlike the hierarchical mechanistic model, the individual and spatial covariates are treated the same. However, we can still explicitly include the effect of time, η_t, and spatial location, η_s, albeit they are modelled individually and do not depend on the covariates. To illustrate the importance of explicitly including the effect of time and space, we also used a GAM that did not explicitly state these components,

$$g(\lambda_i) = \boldsymbol{x}_i\boldsymbol{\beta}. \tag{9}$$

Model assessment

To evaluate the hierarchical mechanistic model and the GAM we use a standard measure of accuracy: the root mean squared error (RMSE) of the predicted true density of positive patients in space and time. We also examine the accuracy of the models by comparing the estimated regression coefficients with the actual values used when creating the simulated data.

Of most interest for drug resistance, with the hierarchical mechanistic model, we recover the magnitudes and dispersals of the hotspots. For strategic decisions, such as checking the quality and procedures of a particular health care centre, the actual magnitude of the hotspot is irrelevant. Only the ranking of this hotspot compared to the others is required. Therefore we focus on recovering the ranking of the magnitudes of the hotspots. For completeness, we also recover the ranking of the dispersals of the hotspots.

To highlight the ranking of hotspots, and not the actual magnitude, in Results we present the normalised values: $\hat{\theta}_1 = 1$, $\hat{\theta}_2 = 0.5$, $\hat{\theta}_3 = 0.25$, $\hat{\theta}_4 = 0$, $\hat{\theta}_5 = 0$, and $\hat{\phi}_1 = 0$, $\hat{\phi}_2 = 0.14$, $\hat{\phi}_3 = 0.29$, $\hat{\phi}_4 = 1$, $\hat{\phi}_5 = 0.57$, where the $\hat{(\cdot)}$ denotes the normalised value. See Creating the simulated data for the original magnitude and dispersal values. We do not recover the magnitudes and dispersals of the hotspots when applying the GAM because the GAM cannot recover this information.

Results

The predicted density of the positive patients, from our model, is compared to the 'true' density used to generate the simulated data. We calculated the RMSE for the total period, 20 years, thereby testing the overall prediction ability of the model, see Fig. 6. We found that with biased sampling, the RMSE is much greater. Even sampling 5% of the region with unbiased sampling, is more accurate than sampling 30% of the region with biased sampling. Moreover, with a small amount of samples collected in a biased manner, 5% of the region, our model was unable to converge. This raises concerns regarding the suitability of this method, even when more patients are sampled.

The estimates for the regression coefficients (β, α_1 and γ_1), from the MCMC, are recovered when unbiased sampling is used, but not when biased sampling is used, see

Fig. 7. This is especially true for β, the regression coefficient which quantifies the probability that an individual is sampled based on an individual covariate such as the age. This is expected since biased sampling ignores the individual covariates, and focuses only on whether it is believed that the patient carries drug resistant pathogens. With unbiased sampling, the accuracy of the regression coefficients is not greatly improved when more patients are sampled, although the interquartile range of the estimates is smaller, see Fig. 7. For the transmission regression coefficients, α_1 and γ_1, the median estimate is actually slightly closer to the actual value when 10% of the region is sampled. Nonetheless, the accuracy for all cases is so high that the difference is insignificant. This demonstrates that we can accurately capture the spatiotemporal transmission dynamics, separating for neighbouring transmission and local transmission, and quantify the relationship with spatial covariates such as disease prevalence.

The most novel application of our model is the ability to rank N resistance hotspots, see Fig. 8. Hotspot 1 has the greatest magnitude (normalised magnitude of one), meaning that it contributes the most resistant pathogens into the population. However, looking only at the density of positive patients (Fig. 4), the high disease prevalence in the lower part of the region (Fig. 3) incorrectly implies that hotspot 4 has the greatest magnitude, see Fig. 4. With unbiased sampling, the model recovers that hotspot 1 has the greatest magnitude, $\theta_1 = 80$. Even when only sampling 5% of the region, the model generally recovers the true ranking of hotspots. However, because the difference in magnitude between the second and third hotspots is small, $\theta_2 = 70$ and $\theta_3 = 65$), the model struggled to get this ordering correct and reversed their importance in the case of 5% of the region being sampled.

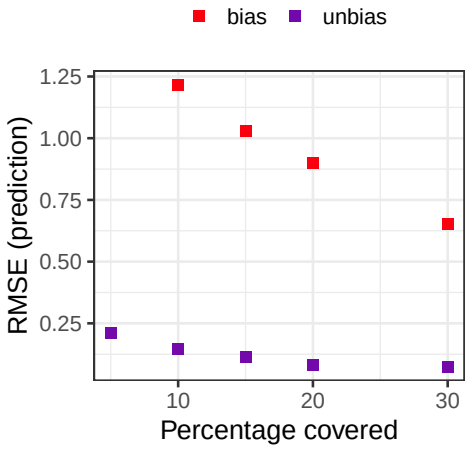

Figure 6. The root mean squared error (RMSE) over the 20 year period for different sampling techniques (bias, unbias), using the hierarchical mechanistic model.

Comparison with generalised additive models

The GAMs had a lower accuracy when compared to our hierarchical mechanistic model. We first discuss how the GAM which explicitly accounted for the effect of space and time (with η_s and η_t, Eq. 8), compared with the hierarchical mechanistic model, and then briefly discuss results from the GAM which did not explictly account for the effect of space and time, Eq. 9.

The GAM, Eq. 8, with unbiased sampling had RMSEs greater than the hierarchical mechanistic model with unbiased sampling. The GAM produced results where the RMSE is consistently around 0.48 irrespective of the proportion that was sampled (the

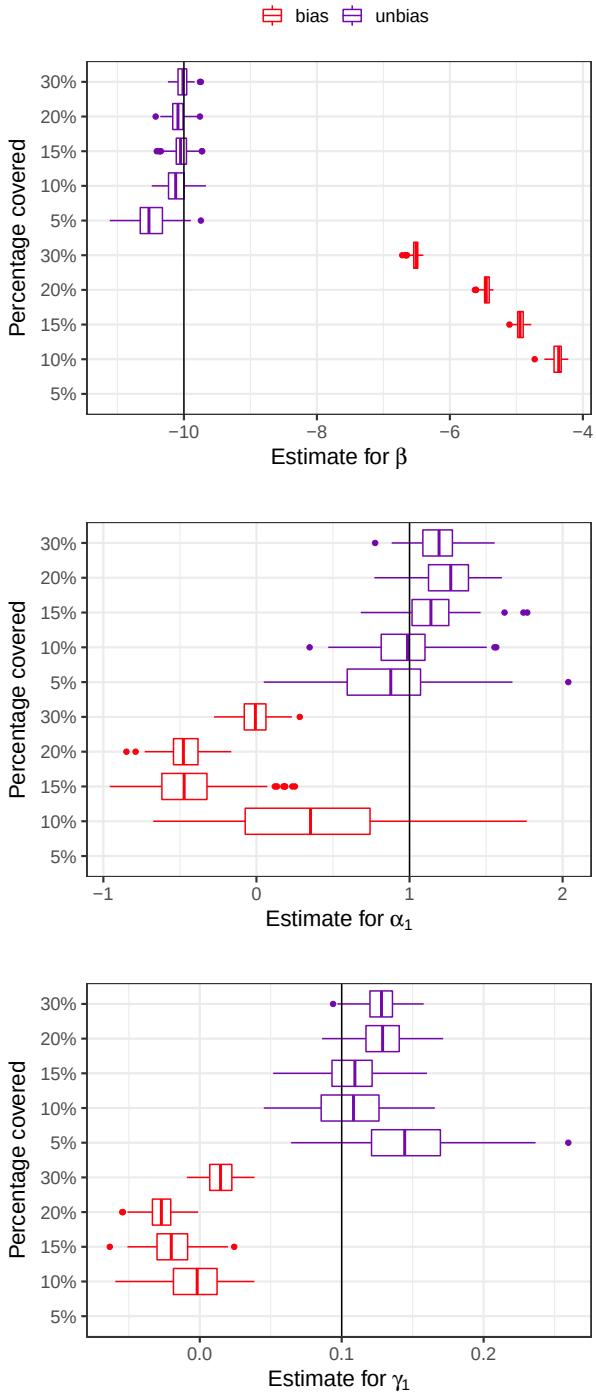

Figure 7. The estimates from the MCMC for the regression coefficients: β relating to the sampling probability, α_1 relating to transmission to neighbouring regions, and γ_1 relating to local transmission. The actual value is indicated by a solid line. The model fails when only sampling 5% of the region in a biased manner.

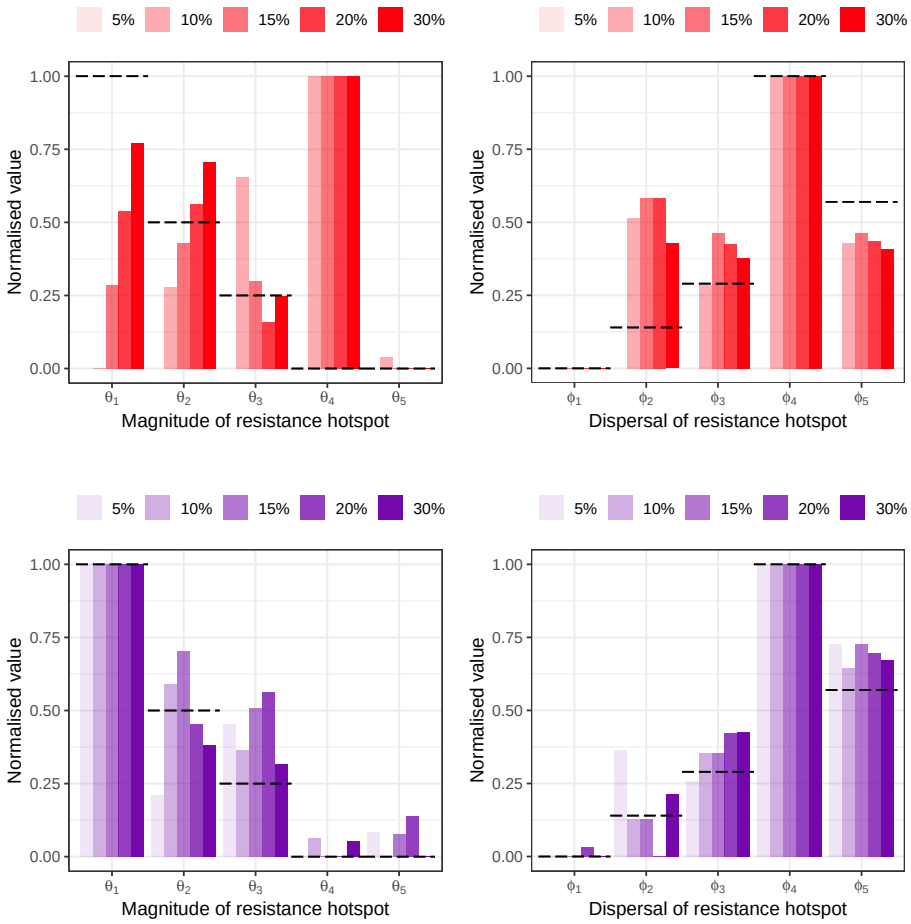

Figure 8. The normalised estimates from the MCMC for the magnitudes of the hotspots (left) and disperal of the hotspots (right) with biased sampling (red) and unbiased sampling (purple). The actual normalised value is indicated by the dashed line. The proportion of the region sampled is indicated by the intensity of the colour. The model fails when only sampling 5% of the region in a biased manner.

explained deviance is around 65.5%). Whereas, even when sampling only 5% of the region, the hierarchical mechanistic model, with unbiased sampling had a lower RMSE of 0.21, see Fig. 6. The GAM, Eq. 8, with biased sampling, had RMSEs that decreased slightly, from 0.77 to 0.53, as the proportion which is sampled increased from 5% to 30% (the explained deviance increased from 59% to 70.1%). This is actually better than the hierarchical mechanistic model, which had an RMSE that varied between 1.22 and 0.90 as the proportion which is sampled increased from 5% to 30%, see Fig. 6.

The regression coefficient estimate for the individual covariate (such as age), when using unbiased sampling, was approximately -10 irrespective of the proportion that was sampled, which corresponds exactly to the value used for the simulated data. This is unsurprising since the probability that a patient is sampled, given patient information (such as age), is given by Eq. 7 which corresponds exactly with the GAM, see Eq. 8. However, when using biased sampling, this coefficient was not recovered. Instead the coefficient estimate ranged between -3.6 and -6.6, dependent on the proportion that was sampled. Therefore, although the exact value is not recovered, the negative relationship is consistent.

The regression coefficient estimate for the local transmission, which depended on a spatial factor (such as disease prevalence), when using unbiased sampling, varied between 16.3 and 53.4, dependent on the proportion that was sampled. The regression coefficient estimate for neighbouring transmission, which also depended on a spatial factor (such as disease prevalence), when using unbiased sampling, varied between 2.4 and 23.5, dependent on the proportion that was sampled. This range is relatively larger than the corresponding range for local transmission, reflecting the model struggling more to capture this relationship (which was generated using a logarithm relationship, see Eq. 5 compared to Eq. 6). Generally, the coefficients for the spatial covariates increased as the proportion sampled increased, meaning that their influence on the spread of drug resistance is more appropriately accounted for when the unbiased sample size is increased.

When using biased sampling, the regression coefficient estimate for the local transmission, which depended on a spatial factor (such as disease prevalence), varied between 3.5 and 13.4, dependent on the proportion that was sampled. The regression coefficient estimate for neighbouring transmission, which also depended on a spatial factor (such as disease prevalence), when using biased sampling, varied between -1.4 and 8.3, dependent on the proportion that was sampled. When sampling 10% or 20%, the coefficient estimate is negative, and for 5%, 15% and 30% it is positive. This switch between positive and negative fundamentally alters our interpretation on the effect of disease prevalence on the spread of drug resistance to neighbouring regions. A model where such an important learning is sensitive to the sampled size is unsuitable, highlighting the dangers of biased sampling, especially with a GAM.

When we removed the spatial and temporal components from the GAM, Eq. 9, the RMSEs are greater and the explained deviance is less, indicating that including spatial and temporal effects increased the predictive ability of the model. When using unbiased sampling, the RMSE is around 0.71 irrespective of the proportion that was sampled (the explained deviance is around 26.3%). When using biased sampling, the RMSE ranges between 1.07 and 0.82 as the proportion which is sampled increased from 5% to 30% (the explained deviance increased from 19.8% to 30.3%). Since the GAM without spatial and temporal components was only included to demonstrate the importance of explicitly accounting for space and time, we do not discuss the regression coefficient estimates from these models.

Discussion

In the long term, to combat the emergence and spread of drug resistant diseases we must (i) use models that explicity include underlying dynamics, such as the distance from a resistance hotspot (which may be a health care centre of transport hub); and (ii) move beyond monitoring drug resistance, since monitoring involves only conducting studies when and where we suspect drug resistance is present.

Hierarchical mechanistic models are powerful tools in ecology, modelling species invasion and the spread of disease within a species [3, 7–11]. Although these models can support epidemiology, they have not yet been widely adopted in the field. We modified an existing model of disease spreading through a species [7], to model drug resistant pathogens spreading throughout a region. There are three levels to our model. First, the data level of our model accounted for individual factors, such as age, influencing the probability of being include in a study on drug resistance. Second, the process level of our model used a PDE to capture spatiotemporal mechanistic components, which quantified the relationship of transmission of resistance pathogens with spatial factors, such as the prevalence of the disease. To modify the PDE (from the previous hierarchical mechanistic model [7]) so that it is suited for modelling drug resistance, we added a new term which accounts for the possibility of resistance emerging within individuals based on their distance from resistance hotspots. Third, the parameter level included uncertainty in our model parameters. Our model accurately recovered the model parameters used to generate our simulated data.

Phenomenological regression models, such as GAMs, cannot include mechanistic components. Even recent sophisticated models, such as the logistic Gaussian processes in [5] and the stacked Gaussian process model in [1], cannot account for mechanistic causes of disease spread.

Another limitation of regression models, such as a GAM, is apparent when comparing the estimate for the density of positive patients. The GAM, with spatiotemporal components, provided estimates that are less accurate than the hierarchical mechanistic model, when sampling is unbiased. This difference in accuracy highlights the need to include mechanisitic components.

We proved that biased data cannot recover underlying dynamics. With biased sampling of 5% of the region, the hierarchical mechanistic model failed to converge. With larger sample sizes, biased sampling does not recover the ranking of hotspots, and could lead to inefficiently focusing resources on a wrong hotspot because of misleading factors. For example, assuming this wrong hotspot is a health care centre, our model demonstrated that the high density of positive patients is not necessarily due to low quality medicine or poor adherence, or other factors which enhance selection pressure, but could actually be due to the high disease prevalence in this region, and the close proximity of this hotspot to other hotspots. Consequently, investigating this hotspot would be a waste of resources.

Data collection is rapidly changing, and our model confirms that we need to reconsider how the data is gathered. Gathering data early, and unbiasedly (random), provides an opportunity to gain new understanding into the mechanisms of the spread of drug resistant pathogens, which ultimately leads to prolonging the life span of drugs. This is highlighted when comparing the predicted density of positive patients from using our model with using the GAM. When sampling is biased, the GAM is more accurate even though it fails to reliably capture the direction of the dependence of resistant pathogen presence and disease prevalence. Essentially, when the data is biased, our modelled relationships between spatiotemporal factors and the density of of the resistant pathogen are unreliable, whether using a hierarchical mechanistic model or a GAM.

The hierarchical mechanistic model is not limited to drug resistant pathogens. Hotspots are also a concern for diseases in general. For example, regression models have

been used to identify hotspots for lower respiratory infection morbidity and mortality in African children [21]. The demonstration of our model required spatiotemporal knowledge of the prevalence of the disease. If this is unknown, it can be estimated through current methods, such as multinomial regression [15]. Or alternatively, our model can be used to estimate the disease prevalence, and that outcome used as a covariate to estimate the spread of the drug resistant pathogen. Our model is flexible to features such as adding or removing hotspots over time (by changing the last term of Eq. 4). Or our model can determine hotspot locations by adding the coordinates as unknown model parameters in the parameter level. With this added flexibility, it could be interesting to compare the hotspots for the disease, and compare these with the hotspots for the corresponding drug resistant pathogens.

Acknowledgments

This research was funded by Tamsin Lee's Marie Curie Individual Fellowship 839121, Horizon 2020. Melissa Penny, who is funded by the Swiss National Science Foundation PP00P3_170702, provided epidemiological guidance and proofreading. Mevin Hooten provided early modelling guidance. Calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing center at University of Basel.

References

1. S. Bhatt, E. Cameron, S. R. Flaxman, D. J. Weiss, D. L. Smith, and P. W. Gething. Improved prediction accuracy for disease risk mapping using gaussian process stacked generalization. *Journal of The Royal Society Interface*, 14(134):20170520, 2017.

2. M. F. Boni, D. L. Smith, and R. Laxminarayan. Benefits of using multiple first-line therapies against malaria. *Proceedings of the National Academy of Sciences*, 105(37):14216–14221, 2008.

3. P. B. Conn, D. S. Johnson, J. M. V. Hoef, M. B. Hooten, J. M. London, and P. L. Boveng. Using spatiotemporal statistical models to estimate animal abundance and infer ecological dynamics from survey counts. *Ecological Monographs*, 85(2):235–252, 2015.

4. H. C. Davison, M. E. Woolhouse, and J. C. Low. What is antibiotic resistance and how can we measure it? *Trends in microbiology*, 8(12):554–559, 2000.

5. M. Deutsch-Feldman, O. Aydemir, M. Carrel, N. F. Brazeau, S. Bhatt, J. A. Bailey, M. Kashamuka, A. K. Tshefu, S. M. Taylor, J. J. Juliano, et al. The changing landscape of plasmodium falciparum drug resistance in the democratic republic of congo. *BMC infectious diseases*, 19(1):1–10, 2019.

6. M. J. Garlick, J. A. Powell, M. B. Hooten, and L. R. McFarlane. Homogenization of large-scale movement models in ecology. *Bulletin of Mathematical Biology*, 73(9):2088–2108, 2011.

7. T. J. Hefley, M. B. Hooten, R. E. Russell, D. P. Walsh, and J. A. Powell. When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecology Letters*, 20(5):640–650, 2017.

8. M. B. Hooten, M. J. Garlick, and J. A. Powell. Computationally efficient statistical differential equation modeling using homogenization. *Journal of agricultural, biological, and environmental statistics*, 18(3):405–428, 2013.

9. M. B. Hooten and T. J. Hefley. *Bringing Bayesian models to life*. CRC Press, 2019.

10. M. B. Hooten and C. K. Wikle. A hierarchical bayesian non-linear spatio-temporal model for the spread of invasive species with application to the eurasian collared-dove. *Environmental and Ecological Statistics*, 15(1):59–70, 2008.

11. M. B. Hooten and C. K. Wikle. Statistical agent-based models for discrete spatio-temporal systems. *Journal of the American Statistical Association*, 105(489):236–248, 2010.

12. G. M. Knight, N. G. Davies, C. Colijn, F. Coll, T. Donker, D. R. Gifford, R. E. Glover, M. Jit, E. Klemm, S. Lehtinen, et al. Mathematical modelling for antibiotic resistance control policy: do we know enough? *BMC infectious diseases*, 19(1):1–9, 2019.

13. A. Mari, T.-C. Roloff, M. Stange, K. K. Soegaard, E. Asllanaj, G. Tauriello, L. T. Alexander, M. Schweitzer, K. Leuzinger, A. Gensch, et al. Global surveillance of potential antiviral drug resistance in sars-cov-2: proof of concept focussing on the rna-dependent rna polymerase. *medRxiv*, pages 2020–12, 2021.

14. E. R. Mega. Alarming surge in drug-resistant hiv uncovered. *Nature*, 2019.

15. M. Nguyen, R. E. Howes, T. C. Lucas, K. E. Battle, E. Cameron, H. S. Gibson, J. Rozier, S. Keddie, E. Collins, R. Arambepola, et al. Mapping malaria seasonality in madagascar using health facility data. *BMC medicine*, 18(1):1–11, 2020.

16. A. M. Niewiadomska, B. Jayabalasingham, J. C. Seidman, L. Willem, B. Grenfell, D. Spiro, and C. Viboud. Population-level mathematical modeling of antimicrobial resistance: a systematic review. *BMC medicine*, 17(1):1–20, 2019.

17. C. Nsanzabana. Resistance to artemisinin combination therapies (acts): do not forget the partner drug! *Tropical medicine and infectious disease*, 4(1):26, 2019.

18. C. Nsanzabana, F. Ariey, H.-P. Beck, X. C. Ding, E. Kamau, S. Krishna, E. Legrand, N. Lucchi, O. Miotto, S. Nag, et al. Molecular assays for antimalarial drug resistance surveillance: a target product profile. *PloS one*, 13(9):e0204347, 2018.

19. L. C. Okell, J. T. Griffin, and C. Roper. Mapping sulphadoxine-pyrimethamine-resistant plasmodium falciparum malaria in infected humans and in parasite populations in africa. *Scientific reports*, 7(1):1–15, 2017.

20. J. A. Powell and N. E. Zimmermann. Multiscale analysis of active seed dispersal contributes to resolving reid's paradox. *Ecology*, 85(2):490–506, 2004.

21. R. C. Reiner, C. A. Welgan, D. C. Casey, C. E. Troeger, M. M. Baumann, Q. P. Nguyen, S. J. Swartz, B. F. Blacker, A. Deshpande, J. F. Mosser, et al. Identifying residual hotspots and mapping lower respiratory infection morbidity and mortality in african children from 2000 to 2017. *Nature microbiology*, 4(12):2310–2318, 2019.

22. A. O. Talisuna, C. Karema, B. Ogutu, E. Juma, J. Logedi, A. Nyandigisi, M. Mulenga, W. F. Mbacham, C. Roper, P. J. Guerin, et al. Mitigating the threat of artemisinin resistance in africa: improvement of drug-resistance surveillance and response systems. *The Lancet infectious diseases*, 12(11):888–896, 2012.

23. WHO. Hiv drug resistance, 2020. URL: `https://www.who.int/news-room/fact-sheets/detail/hiv-drug-resistance`.

24. C. K. Wikle. Hierarchical bayesian models for predicting the spread of ecological processes. *Ecology*, 84(6):1382–1394, 2003.

25. C. K. Wikle and M. B. Hooten. Hierarchical bayesian spatio-temporal models for population spread. *Applications of computational statistics in the environmental sciences: hierarchical Bayes and MCMC methods*, 145169, 2006.

26. C. K. Wikle, A. Zammit-Mangion, and N. Cressie. *Spatio-temporal Statistics with R*. Chapman and Hall/CRC, 2019.

27. WWARN. Worldwide antimalarial resistance network, 2021. URL: `https://www.wwarn.org/`.