Bayesian experimental design for control and surveillance in epidemiology

A Dissertation Presented

by

Bren Case

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Computer Science

October, 2023

Defense Date: August 18th, 2023
Dissertation Examination Committee:

Laurent Hébert-Dufresne, Ph.D., Advisor
Jean-Gabriel Young, Ph.D., Advisor
Lori Stevens, Ph.D., Chairperson
Jeff Buzas, Ph.D.
Donna M. Rizzo, Ph.D.
Holger Hoock, DPhil, Dean of the Graduate College

Abstract

Effective public health interventions must balance an array of interconnected challenges, and decisions must be made based on scientific evidence from existing information. Building evidence requires extrapolating from limited data using models. But when data are insufficient, it is important to recognize the limitations of model predictions and diagnose how they can be improved. This dissertation shows how principles from Bayesian Experimental Design can be applied to surveillance and control efforts to allow researchers to get more out of their data and direct limited resources to best effect. We argue a Bayesian perspective on data gathering, where design decisions are made to maximize utility on average over a joint distribution of beliefs and outcomes, is better suited to the epidemiological setting where observational studies are the norm. We illustrate these ideas using a range of models and topics across epidemiology.

We focus first on Chagas disease, where in Guatemala an endemic vector continues to cause a high rate of domiciliary infestation in rural communities, and shortages of insecticides and resources for critical house improvements hamper control efforts. Using an adaptive sampling and geospatial modeling framework, we show that interpolating from a traditional design goal of minimizing prediction uncertainty to targeting houses of high risk can satisfy competing objectives, namely, to efficiently identify houses in need of treatment while mitigating sampling bias. We next focus on tick surveillance in the southeastern United States. By framing tick collection surveys as a design problem over time and space, we show optimal survey design can yield greater information compared to random or convenience sampling. Finally, we shift attention from experimental design to the closely related concept of practical identifiability. We propose a novel method to quantify practical identifiability which reflects the average amount of posterior shrinkage that would occur in a Bayesian analysis, without requiring computationally expensive techniques like Markov Chain Monte Carlo. With this method, we demonstrate the limits of using epidemiological models to derive standard statistics such as the basic reproductive number early in an outbreak.

Citations

Material from this dissertation has been published in the following form:

Case, B. K. M., Young, J.-G., Penados, P., Monroy, C., Hébert-Dufresne, L., Stevens, L.. Spatial epidemiology and adaptive targeted sampling to manage the Chagas disease vector Triatoma dimidiata. *PLoS Neglected Tropical Diseases.* 16(6):e0010436, 2022. 10.1371/journal.pntd.0010436

Case, B. K. M., Young, J.-G., Hébert-Dufresne, L.. Accurately summarizing an outbreak using epidemiological models takes time. *Royal Society Open Science.* 10:230634, 2023. 10.1098/rsos.230634

Material from this dissertation has been submitted for publication to the American Journal of Public Health on September 20th, 2023 in the following form:

Case, B. K. M., Dye-Braumuller, K. C., Evans, E., Li, H., Rustin, L., Nolan, M. S.. Adapting vector surveillance using Bayesian Experimental Design: an application to an ongoing tick monitoring program in the southeastern United States.

Acknowledgements

First, I would like to give a huge thanks to all the collaborators I worked with throughout my PhD. Engaging with other researchers is a huge part of what makes science fulfilling for me, and I am grateful to have been able to work with so many bright people across a range of different disciplines, and to have formed relationships I hope will continue to grow over the years.

Thank you also to my advisors, Laurent Hébert-Dufresne and Jean-Gabriel Young, for your patient guidance and endless support. You have given me countless opportunities and ideas over the years.

I would also like to acknowledge Lori Stevens and Carlota Monroy for introducing me to Chagas disease. Thank you for taking me under your wings and into the jungles of Petén. I am also grateful to the leaders and other students of the QuEST traineeship program, especially the co-PIs Melissa Pespini and Lori, as the program was hugely influential for my growth as an interdisciplinary scientist, and to my defense committee for their time spent reviewing my work.

Finally, I am endlessly grateful to my family—my mom for encouraging me to go to grad school, my fiancée Ryan for braving the cold and building a life with me, and our dogs Sammie and Dot for being perfect.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Advancing wellbeing through the mitigation of infectious diseases is one of science's great contributions to humanity. Distilling scientific evidence into decisions is essential for coherent public health policy, and models are essential for this process. Models take data, extract patterns, and make predictions. Often, this process is treated as a one-way street. As more data come in, model predictions are updated and inform future action. It is less common to "complete the loop" and consider how actions based on models could impact future data [1, 2]. This dissertation explores model-based experimental design as a tool for directing surveillance and control efforts in a statistically principled way.

In epidemiology, reconciling data and theory is rarely possible with carefully controlled experiments. Noisy data are measured from heterogeneous populations and disparate sources, and building models incorporating all of the processes behind these data would be impossible. Instead, allowing uncertainty, i.e. a range of plausible behaviors, in a model frees us from the need to find a model with the perfect level of complexity [3]. Rather than matching the data exactly, a useful model should encode a range of dynamics that are constrained to a tractable level. Richard Levins

famously referred to this as "sacrificing precision to realism and generality" [4].

The Coronavirus Disease 2019 (COVID-19) pandemic has led to an explosion of activity applying models to data [5]. These models have often updated predictions in near real-time using multiple data streams [6], and have directly informed policy recommendations [7, 8]. Formal theories for strategically looping from predictions, decisions, and feedback on objectives have been proposed and successfully employed by collaborative teams across the world [9, 10]. However, COVID-19 also showed these efforts take time, and there have been many other examples where a lack of coordination within and between modelers and empiricists contributed to confusion and public distrust [11, 12]. Ecology also has a long history of feedback between theories and decision making, where a variety of models have been used to directly guide land management and agriculture, and to combat wildlife diseases [13, 14]. But here, too, successfully establishing prolonged feedback between modelers, data gatherers, and policy makers sometimes takes years of fine-tuning how modeling tools should be used for best effect [15].

Overall, direct feedback between modelers, empiricists, and policy makers remains somewhat rare [1, 16]. A 2016 survey of PubMed found a small fraction of biomedical papers employing ODEs inferred parameters using data, and just a fraction of these formally considered the limitations of available data for reliable inference [17]. Model-informed experimental design is one avenue where theoretical and computational scientists can integrate more closely with empiricists and feel closer to the systems they study. Employing modeling expertise also allows greater creativity in the design variables considered, which serves to enrich the research of empiricists as well. In particular, while experimental design studies typically use simple statistical models

for optimizing sample size or treatment groups, a model with more heterogeneous structure leads to more specificity in what can be manipulated or controlled.

In Bayesian experimental design (BED), the utility of design inputs is averaged over a joint distribution of beliefs and outcomes. This makes BED particularly applicable in fields like epidemiology, where observational studies are the norm and decisions must be robust with respect to uncertainty in the underlying system. Following an introduction covering a review of the modeling frameworks used in this work, the basics of Bayesian inference, and some standard BED theory, we illustrate these ideas using a range of models and topics across epidemiology.

1.1 Nonlinear modeling

It will be useful to conceptualize the models used throughout this dissertation in terms of two components, a *latent process* and *observation process*, which respectively measure the underlying, hypothesized dynamics of the system in question, and the noisy, perturbed data we obtain when measuring this latent system. We write these components as

$$\mathbf{y} \sim g(\boldsymbol{\mu}, \boldsymbol{\sigma}) \tag{1.1}$$

$$\boldsymbol{\mu} = f(\boldsymbol{\gamma}; \mathbf{d}), \tag{1.2}$$

where f is a deterministic function tying parameters $\boldsymbol{\gamma}$ and design points $\mathbf{d}$ to latent dynamics $\boldsymbol{\mu}$, and g is a stochastic function tying noisy data $\mathbf{y}$ to the latent dynamics,

optionally with additional noise parameters $\boldsymbol{\sigma}$.[1] The model inputs $\mathbf{d}$ are called the *design*, and represent the factors that may be controlled during data collection. In observational settings these may include measurement times and/or locations, while treatment groups or other covariates can be manipulated directly in experimental settings. The latent parameters $\boldsymbol{\gamma}$ in contrast represent those inputs which we do not know, or do not have control over. When fitting models to data, these values therefore must be estimated or fixed to settings well-established in the literature [19]. It is assumed we are generically interested in estimating a p-dimensional vector of unknown parameters $\boldsymbol{\theta} \subseteq (\boldsymbol{\gamma}^\top, \boldsymbol{\sigma}^\top)^\top$.

A model is said to be *nonlinear* if either g or f is nonlinear in $\boldsymbol{\theta}$. For mechanistic modeling, this usually arises from nonlinearity of f in $\boldsymbol{\gamma}$ (Figure 1.1), but this need not always be the case, as we will see in the following example. A key motivation for making the distinction between f and g is to acknowledge that g may be a source of model complexity in its own right, and that accounting for complex factors in g can interact with factors in f in unexpected ways [11, 20].

1.1.1 Example: mixed-effects models

Generalized linear models (GLMs) are a class of models in which $\boldsymbol{\mu}$ is a linear function of parameters $\boldsymbol{\beta}$. The latent output $\boldsymbol{\mu}$ is called the *linear predictor*, while $\boldsymbol{\beta}$ is a vector of *regression coefficients*. To transform $\boldsymbol{\mu}$ to an appropriate domain for the data, a nonlinear *link function* is used. The observation process g therefore transforms $\boldsymbol{\mu}$ through the link function before assigning an appropriate distribution to the data. For example, counts data are positive and integral, so if the domain of $\boldsymbol{\mu}$ is $\mathbb{R}$, a

[1] (1.1) and (1.2) can also usually be written in nonlinear regression format, $\mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a random variable drawn from a distribution parameterized by $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ [18].

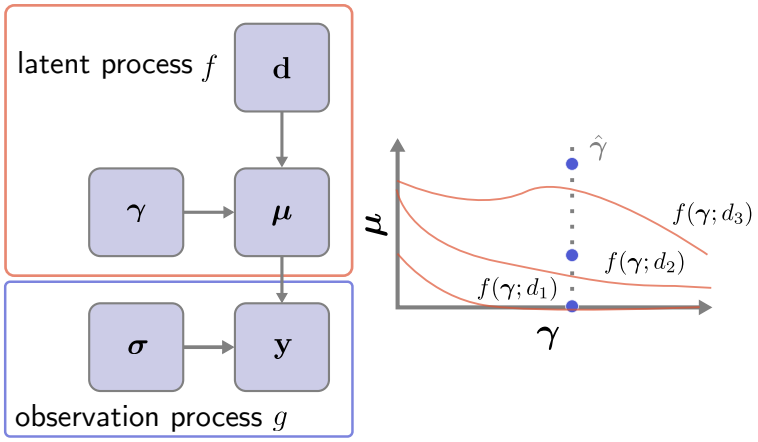

Figure 1.1: Nonlinear models separated into a latent process and observational process. Right: output from a latent process is shown as a function of $\boldsymbol{\gamma}$ at three design points $\mathbf{d} = (d_1, d_2, d_3)$. A model is "fit" to three data points (blue circles) by finding $\boldsymbol{\gamma}$ for which $f(\boldsymbol{\gamma}; d_i)$ and y_i are "close" for all i.

suitable observation process would be $g(\boldsymbol{\mu}) = \text{Poisson}(\exp(\boldsymbol{\mu}))$. Link functions are commonly defined moving from the data to $\boldsymbol{\mu}$ in the GLM literature, so the link function here would be the natural logarithm.

A more general family of models are Generalized Linear Mixed-effect Models (GLMMs), which may be written in standard form as

$$\mathbf{y} \sim g(\boldsymbol{\mu}) \tag{1.3}$$

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \tag{1.4}$$

where $\mathbf{X}$ and $\mathbf{Z}$ are design matrices and $\boldsymbol{\beta}$ and $\mathbf{u}$ are parameters to be estimated [21]. The parameters $\boldsymbol{\beta}$ are the so-called "fixed-effects" which are traditionally used to capture relationships between $\mathbf{y}$ and external covariates such as environmental variables, while $\mathbf{u}$ are the so-called "random-effects" which traditionally capture differences between assigned groups or other "clusters" determined by $\mathbf{Z}$. Statistically, the key

difference between the fixed and random effects is we are interested in understanding (i.e. modeling) the structure between and within groups of the random effects, as facilitated by some additional *hyperparameters* $\boldsymbol{\phi}$ which constrain the behavior of $\mathbf{u}$ [22].[2] In addition to giving a quantitative understanding of the variability between groups, modeling $\mathbf{u}$ hierarchically in this way can have beneficial effects, such as helping to eliminate heteroscedasticity in the response through pooling information between smaller groups [24], and allowing predictions for unmeasured groups which were not in $\mathbf{Z}$ [25]. GLMMs will appear in Chapters 2 and 3.

1.1.2 Example: the Ross-MacDonald model for Malaria transmission

While the flexibility of GLMMs allows for a surprisingly diverse range of complex dynamics, they are rarely used to try and capture *why* these dynamics may occur. Another important class of nonlinear models are *compartmental models*, which assign mechanisms at a population level through a series of ordinary differential equations (ODEs) [18]. By assigning system constituents to compartments assumed to have the same behavior on average, epidemiologically meaningful processes can be formally expressed while remaining at a tractable level of abstraction. Models developed this way tend to be more interpretable, and make it easy to consider the effect of different control measures through manipulating specific parameters [26].[3] Compartmental

[2]Because both $\boldsymbol{\beta}$ and $\mathbf{u}$ are "modeled" in Bayesian statistics by placing priors on them, the only difference between fixed and random effects is in how their priors are defined. This has lead some Bayesians to label the fixed vs. random dichotomy as misleading, preferring terms for $\mathbf{u}$ like "hierarchical effects" [23] or "richly parameterized" [21].

[3]Such mechanistic models are also more efficient in achieving a given level of complexity: the GLMMs used in this work have 100s of parameters, that are highly constrained by $\boldsymbol{\phi}$ to avoid overfitting. The SIR model of Chapter 4 has 2-4.

models will appear in Chapter 4.

As an example, consider the canonical compartmental model for disease transmission between hosts and mosquitoes, the Ross-MacDonald model [27]. We divide hosts into two compartments, S_H and I_H, which represent individuals who are *susceptible* to the disease or *infectious*. We similarly divide vectors into compartments S_V and I_V. We further assume the system is *closed*, so that both hosts and vectors have constant population sizes $N_H = I_H + S_H$ and $N_V = I_V + S_V$. This means we only need to track changes to the infectious compartments.

Changes to these compartments are defined with ODEs as a function of time t. The key mechanisms assumed by the model are that susceptible hosts may become infectious when bitten by an infected mosquito (the "probability" of biting a susceptible host is S_H/N_H), while mosquitoes become infected when biting infectious hosts (probability I_H/N_H). The full model equations are

$$\frac{d}{dt}I_H = kpI_V\frac{N_H - I_H}{N_H} - \alpha I_H \tag{1.5}$$

$$\frac{d}{dt}I_V = kq(N_V - I_V)\frac{I_H}{N_H} - \delta I_V, \tag{1.6}$$

where k is the biting rate of mosquitoes, p and q are respectively the probabilities of infection for hosts/vectors following exposure, α is the rate at which hosts recover from illness, and δ is the death rate for vectors.

Integrating (1.5) and (1.6) over t, together with some *initial conditions* $I_H(0)$ and $I_V(0)$, gives a solution to the system over time, $I_H(t)$ and $I_V(t)$. These values are referred to as the *state variables*, and are the latent dynamics we wish to learn about. Assuming data are only collected for infectious hosts at some discrete timepoints,

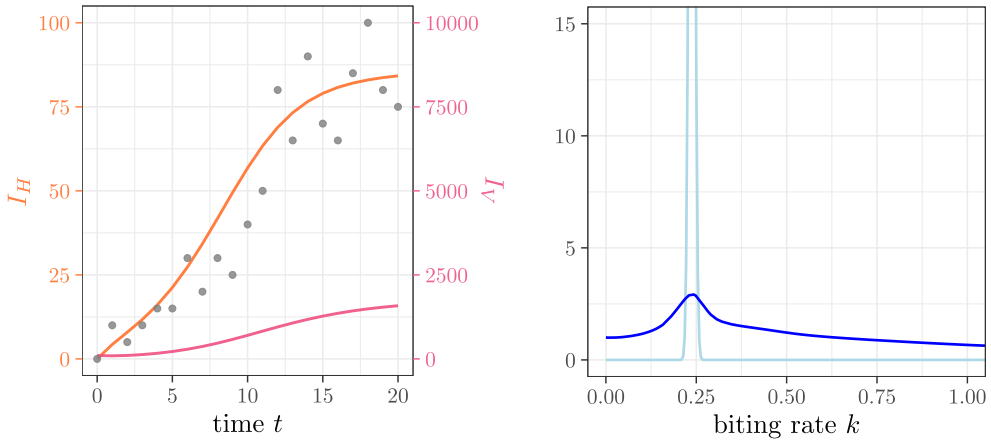

Figure 1.2: Example output from the Ross-MacDonald model equations with noisy data. Left: the number of infectious hosts are in orange and infectious vectors are pink. Parameter settings were $k = 1/4$, $p = q = 1/5$, $\alpha = 1/7$, $\delta = 1/5$, $N_H = 100$, and $N_V = 10^4$. Right: inference for the biting rate k according the information matrix (light blue) and the posterior distribution under prior $P(k) \sim$ Exponential(1) (dark blue).

adopting the notation of (1.1) and (1.2), we have $\mathbf{d} = (t_1, \ldots, t_n)^\top$ and $\mu_i = I_H(t_i)$. For the observation process, we may assume for example that m individuals are selected randomly for testing each day, so that y_i is the number of hosts who tested positive at day t_i. A simple model for this would be

$$y_i \sim \text{Binomial}(m, \mu_i/N_H).$$

Some example output from a typical parametrization of the model along with noisy data is shown in Figure 1.2.

1.2 Likelihood-based inference

For a particular design input and instance of parameters hypothesized to have generated the data, the latent and observation processes together produce a probability

density $P(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d})$ called a *likelihood*, which gives the probability that a particular dataset $\mathbf{y}$ arose from f and g under a particular parameterization $\boldsymbol{\theta}$ and design $\mathbf{d}$.[4] Given data, we naturally are interested in inferring the plausibility of different parameter combinations—we wish to find parameters with high likelihood. Even though $P(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d})$ is a density in $\mathbf{y}$, for parameter inference both $\mathbf{y}$ and $\mathbf{d}$ are of secondary interest and so a convention is to write $\mathcal{L}(\boldsymbol{\theta}) = P(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d})$. The quantity maximizing the likelihood,

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\Theta}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{\theta}), \tag{1.7}$$

is the *maximum likelihood estimator*. Note that, because the likelihood depends on the random variable $\mathbf{y}$, $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is itself a random variable from which we can derive uncertainty in our estimates. In the frequentist tradition, the distribution of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ reflects the *long-run frequency* from repeating our experiment: supposing we repeatedly sampled datasets under identical conditions and maximized the likelihood function of each, a histogram of these estimates will converge to match the density of our estimator.

In practice, of course, we do not have the ability to infinitely repeat our experiment, and so the distribution of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ must be approximated from the particular dataset we have at hand. One option here is to use some form of *bootstrapping*, where the data are repeatedly resampled as a surrogate for the actual sampling distribution, and the parameters are fit to each. However, this becomes complicated or impossible

[4] Letting Θ be some established domain of viable parameter settings, the space of all likelihoods $\{P(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d}) : \boldsymbol{\theta} \in \Theta\}$ forms a topological structure called a *model manifold*. Although rarely referenced directly, the global structure of this manifold is the fundamental property under consideration in this work; namely, experimental design seeks to maximize the geodesic distance between points on the manifold though manipulating $\mathbf{d}$, while identifiability measures the length of these distances, pulled back to the intrinsic geometry of Θ, for a given $\mathbf{d}$ [28].

when observations are correlated, like in a time series. Another option is to proceed using the so-called *asymptotic theory*, which, under certain conditions,[5] guarantees certain properties of the maximum likelihood estimator as the amount and/or precision of data grows to infinity. Here we emphasize two such properties, *consistency* and *efficiency*, which respectively imply that with sufficient data, the maximum likelihood estimator converges to the (unknown) value $\boldsymbol{\theta}^*$ which created the data, and that it becomes asymptotically normally distributed,[6]

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} \xrightarrow{d} N(\boldsymbol{\theta}^*, \mathcal{I}(\boldsymbol{\theta}^*)^{-1}). \tag{1.8}$$

In short, to obtain an approximate distribution for $\hat{\boldsymbol{\theta}}_{\text{MLE}}$, we need 1) the value $\boldsymbol{\theta}_{\text{MLE}}$ maximizing $\mathcal{L}$ for our particular dataset, and 2) an expression for the *information matrix* $\mathcal{I}$, into which we may plug $\boldsymbol{\theta}_{\text{MLE}}$ as a proxy for $\boldsymbol{\theta}^*$. The information matrix is a local measure of *curvature*, i.e. sensitivity of $\mathcal{L}$ to changes in $\boldsymbol{\theta}$, and has typical element

$$[\mathcal{I}(\boldsymbol{\theta})]_{ij} = -\,\mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log \mathcal{L}(\boldsymbol{\theta})\right]. \tag{1.9}$$

Importantly for this work, the information matrix is also the key quantity in classical approaches to both experimental design and practical identifiability. It will be discussed more in Chapter 4.

An advantage of the asymptotic theory is that the information matrix is relatively easy to derive, which can allow a more formal understanding of how uncertainty arises

[5]Notably for mechanistic modeling, these conditions include that the latent process f is injective in $\boldsymbol{\gamma}$, and that the likelihood function is compact, that is, creating a simpler model by taking a parameter to its boundary limit must have lower likelihood than the full model.

[6]In some disciplines it is more common to define the covariance as $(\sqrt{n_{\text{rep}}}\mathcal{I}(\boldsymbol{\theta}^*))^{-1}$, where n_{rep} is the number of independent replications of the data, assuming these replications are not already accounted for in the likelihood.

in the model. However, a limitation is that with nonlinear models, (1.8) only provides a lower bound on the true level of uncertainty. When models are highly nonlinear and data are limited, this bound can severely underestimate the actual variability of the estimator [29]. An alternative which is well-suited to such settings is Bayesian inference.

In the Bayesian paradigm, inferences about $\boldsymbol{\theta}$ given $\mathbf{y}$ are not expressed in the language of estimators. Instead, uncertainty is encoded though the *posterior distribution* $P(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{d}) \propto \mathcal{L}(\boldsymbol{\theta})P(\boldsymbol{\theta})$, which expresses movement away from a distribution of prior beliefs $P(\boldsymbol{\theta})$ towards the likelihood. There are several aspects of Bayesian inference which makes it philosophically appealing in an epidemiological setting. First, the frequentist concept of uncertainty as long-run variability feels incongruous when modelings things like an epidemic, which cannot be repeated under identical conditions. Bayesian posteriors are not based on this concept of long-run frequency [19]. Second, building up complex hierarchical structures is made very natural through prior distributions. For example, in GLMMs it is relatively easy to write down the distribution of random effects given their hyperparameters as $P(\mathbf{u} \mid \boldsymbol{\phi}) = N(\mathbf{0}, \Sigma(\boldsymbol{\phi}))$, so that the conceptually tricky "full" likelihood $P(\mathbf{y} \mid \mathbf{u}, \boldsymbol{\phi})$ can be factorized to give a posterior proportional to $P(\mathbf{y} \mid \mathbf{u})P(\mathbf{u} \mid \boldsymbol{\phi})P(\boldsymbol{\phi})$. Third, priors may be used to incorporate outside information about mechanistic parameters without having to commit to fixing the parameter to a single value. For example, there may be an established range of recovery times in the clinical literature which can be used to place informative priors on the host recovery rate α in the Ross-MacDonald model.

Estimates for the biting rate k using the information matrix and Bayesian methods are compared in Figure 1.2. As the posterior $P(k \mid \mathbf{y})$ makes clear, the biting rate is

rather poorly constrained by the data: there is a clear peak around the true value, but fat tails indicate the data can be similarly replicated with a large range of k. However, the local curvature measured from the information matrix does not capture this effect, severely underestimating uncertainty from the full likelihood.

1.3 Bayesian experimental design

We are now prepared to introduce how the design of experiments can be based on Bayesian principles. Experimental design is a form of decision analysis in which investigators seek an effective choice of design $\mathbf{d}$ among a space of available options $\mathcal{D}$. The return associated with a decision $\mathbf{d}$ is called the *utility*, which will depend on the (unknown) future data collected as a result of $\mathbf{d}$, as well as the (probably unknown) latent dynamics which create this data. The joint distribution of data and dynamics is $P(\mathbf{y}, \boldsymbol{\theta})$. Thus, a sensible way to proceed is to make the decision maximizing the average return over joint outcomes [30],

$$\begin{aligned} U(\mathbf{d}) &= \int_{\mathcal{Y}} \int_{\Theta} U(\mathbf{d}, \mathbf{y}, \boldsymbol{\theta}) P(\mathbf{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{y} \\ &= \int \int U(\mathbf{d}, \mathbf{y}, \boldsymbol{\theta}) P(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d}) P(\boldsymbol{\theta}) d\boldsymbol{\theta} d\mathbf{y}, \end{aligned} \tag{1.10}$$

where $U(\mathbf{d}, \mathbf{y}, \boldsymbol{\theta})$ is the utility provided by $\mathbf{y}$ created under the conditions provided by $\mathbf{d}$ and $\boldsymbol{\theta}$. In classical experimental design, utility is typically a function of the information matrix, and thus defined in terms of a function $U(\mathbf{d}, \boldsymbol{\theta})$ with the expectation over $P(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{d})$ done in (1.9). When $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is assumed known, the design maximizing $U(\mathbf{d}, \boldsymbol{\theta}^*)$ is called *locally optimal* [31]. When $\boldsymbol{\theta}$ is unknown, $U(\mathbf{d}, \boldsymbol{\theta})$ is instead averaged over something resembling a prior distribution, although this prior

is subsequently ignored when evaluating the information matrix. This is referred to as *pseudo-Bayesian* or *robust* experimental design [32].

In BED, utility is a function of the posterior distribution rather than the information matrix, and so integrating $\mathbf{y}$ out of $U(\mathbf{d}, \mathbf{y}, \boldsymbol{\theta})$ is usually impossible [33]. However, a common strategy is to instead express (1.10) as

$$U(\mathbf{d}) = \int U(\mathbf{d}, \mathbf{y}) P(\mathbf{y} \mid \mathbf{d}) d\mathbf{y}, \tag{1.11}$$

where $U(\mathbf{d}, \mathbf{y})$ instead involves an integral over the posterior. This can avoid the double integral provided a solution or approximation for $U(\mathbf{d}, \mathbf{y})$ exists, and there is an efficient way to directly obtain samples from the prior-predictive distribution $P(\mathbf{y} \mid \mathbf{d})$. This is an idea that will be expanded upon in Chapter 3.[7]

As an example, to measure the distance between the posterior mean having seen the data and true parameters which generated the data, the following utility function may be used in (1.10):

$$U(\mathbf{d}, \mathbf{y}, \boldsymbol{\theta}) = -\sum_p (\theta_p - \mathbf{E}\left[\theta_p \mid \mathbf{y}, \mathbf{d}\right])^2. \tag{1.12}$$

Letting $\hat{\boldsymbol{\theta}} = \mathbf{E}\left[\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{d}\right]$ and noting that $\sum_p (\theta_p - \hat{\theta}_p)^2 = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})$, (1.10) can

[7] There we will use recent developments in the software R-INLA to both efficiently sample from the predictive distribution and approximate $U(\mathbf{d}, \mathbf{y})$ [34].

then be simplified as

$$\begin{aligned} U(\mathbf{d}) &= \int\int U(\mathbf{d},\mathbf{y},\boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{d}) P(\mathbf{y} \mid \mathbf{d}) d\boldsymbol{\theta} d\mathbf{y} \\ &= -\int P(\mathbf{y} \mid \mathbf{d}) \int \underbrace{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}_{=\mathrm{tr}\left((\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top\right)} P(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{d}) d\boldsymbol{\theta} d\mathbf{y} \\ &= -\int P(\mathbf{y} \mid \mathbf{d}) \mathrm{tr}\left(\mathrm{cov}(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{d})\right) d\mathbf{y}, \end{aligned}$$

where for the last line we have used linearity of integration and definition of the covariance matrix [35]. Therefore, one viable way to proceed is to use something like Laplace's method to compute the posterior covariance, combined with an efficient integration method for (1.11) such as randomized quasi-Monte Carlo [36]. Note that this utility function $U(\mathbf{d},\mathbf{y}) = \mathrm{tr}\left(\mathrm{cov}(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{d})\right)$ is referred to as the Bayesian A-optimality criterion. An alternative that takes uncertainty from correlations between the parameters into account is Bayesian D-optimality, defined $U(\mathbf{d},\mathbf{y}) = -\log\det\mathrm{cov}(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{d})$. Both of these options will underestimate full posterior uncertainty if the posteriors are not roughly normal [33].

Perhaps the most widely-used Bayesian criteria to reflect the quality of a design is the Shannon information gain,

$$U(\mathbf{d},\mathbf{y},\boldsymbol{\theta}) = \log P(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{d}) - \log P(\boldsymbol{\theta}). \tag{1.13}$$

A design maximizing the expectation of (1.13) maximizes the mutual information between $P(\mathbf{y} \mid \mathbf{d})$ and $P(\boldsymbol{\theta})$, i.e. the information about the data that can be encoded in the parameters [37]. Integrating out $\boldsymbol{\theta}$ then gives $U(\mathbf{d},\mathbf{y}) = D_{\mathrm{KL}}(P(\boldsymbol{\theta} \mid \mathbf{y},\mathbf{d}) \mid\mid P(\boldsymbol{\theta}))$, the KL-divergence between the posterior and prior distributions [38, 33].

1.4 Outline

The rest of this dissertation is as follows. In Chapter 2, we develop a sequential design framework which adapts to local information and balances conflicting public health priorities. Because interventions targeting areas with highest perceived risk leads to preferential sampling bias, there is a tradeoff between sampling for accurate prediction and sampling to quickly reach those in need. Such biased sampling can therefore lead to misdiagnosing the remaining number of areas requiring treatment. This tradeoff is especially relevant for the control of Neglected Tropical Diseases (NTDs) like Chagas disease, where there is a high incentive to focus only on areas at risk while still ensuring sufficient disruption throughout the community as a whole. We solve this problem with an adaptive framework which transitions from prioritizing houses based on prediction uncertainty to targeting houses with a high risk of infestation, and test the framework in a simulation study using data from five villages in Guatemala. Due to the spatial nature of Triatomine infestations in the area, the method fits Bayesian geostatistical models, which include a random effect $\mathbf{u}$ to capture correlations between nearby houses and make spatially informed predictions. We find the method can accurately identify the necessary number of infested houses to meet the control target, while consistently using fewer samples than random designs.

Chapter 3 studies the design of tick surveillance schedules in the southeastern United States, where a lack of available data has lead to uncertainty in the spatiotemporal distribution of ticks throughout the region. By framing tick collection surveys as an experimental design problem over time and space, we show careful survey design can yield greater information compared to random or convenience sam-

pling. Optimal BED could therefore be used to maximize the efficiency of local vector control agencies throughout the region, which are currently constrained due to chronic under-funding and lack of infrastructure. We additionally show how recent advances in computational software can be used to deploy BED based on complex hierarchical models incorporating spatial, temporal, and species-level effects.

Chapter 4 shifts attention from BED to the closely related concept of *practical identifiability* (PI) issues in mechanistic models of epidemics. While traditionally PI has been studied using the variance-covariance matrix of an estimator using the information matrix (1.9), such second-order approximations underestimate uncertainty in limited data settings, where the distribution of plausible values may be incorrectly centered or highly skewed. Borrowing from computational and information theoretic ideas in BED, we propose a novel method of PI which reflects the average amount of posterior shrinkage that would occur in a Bayesian analysis, without requiring computationally expensive techniques such as Markov Chain Monte Carlo. Using this method, we revisit identifiability of the classic Susceptible-Infectious-Recovered compartmental model, and compare our ability to infer different summary statistics commonly derived from epidemiological models, such as the basic reproductive number of final outbreak size. Examining the rate of learning these quantities over time, we find identifiability of most statistics is limited until after the true underlying outbreak has peaked. We also discuss the relationship of our new method to other ways of measuring PI, and show the method has attractive properties in both limited and big data regimes.

Bibliography

[1] J. Lessler, W. J. Edmunds, M. E. Halloran, T. D. Hollingsworth, and A. L. Lloyd. Seven challenges for model-driven data collection in experimental and observational studies. *Epidemics*, 10:78–82, March 2015.

[2] Sereina A. Herzog, Stéphanie Blaizot, and Niel Hens. Mathematical models used to inform study design or surveillance systems in infectious diseases: A systematic review. *BMC Infectious Diseases*, 17(1):775, December 2017.

[3] Michael C. Abbott and Benjamin B. Machta. Far from Asymptopia. *Entropy*, 25(3):434, March 2023.

[4] Richard Levins. The Strategy of Model Building in Population Biology. *American Scientist*, 54(4):421–431, 1966.

[5] Qingyu Chen, Alexis Allot, and Zhiyong Lu. LitCovid: An open database of COVID-19 literature. *Nucleic Acids Research*, 49(D1):D1534–D1540, January 2021.

[6] Amani Alahmadi, Sarah Belet, Andrew Black, Deborah Cromer, Jennifer A. Flegg, Thomas House, Pavithra Jayasundara, Jonathan M. Keith, James M. McCaw, Robert Moss, Joshua V. Ross, Freya M. Shearer, Sai Thein Than Tun, James Walker, Lisa White, Jason M. Whyte, Ada W. C. Yan, and Alexander E. Zarebski. Influencing public health policy with data-informed mathematical models of infectious diseases: Recent developments and new challenges. *Epidemics*, 32:100393, September 2020.

[7] Chiara Poletto, Samuel V. Scarpino, and Erik M. Volz. Applications of predictive modelling early in the COVID-19 epidemic. *The Lancet Digital Health*, 2(10):e498–e499, October 2020.

[8] Nicholas G. Reich, Justin Lessler, Sebastian Funk, Cecile Viboud, Alessandro Vespignani, Ryan J. Tibshirani, Katriona Shea, Melanie Schienle, Michael C. Runge, Roni Rosenfeld, Evan L. Ray, Rene Niehus, Helen C. Johnson, Michael A. Johansson, Harry Hochheiser, Lauren Gardner, Johannes Bracher, Rebecca K. Borchering, and Matthew Biggerstaff. Collaborative Hubs: Making the Most of Predictive Epidemic Modeling. *American Journal of Public Health*, 112(6):839–842, June 2022.

[9] Katriona Shea, Michael C. Runge, David Pannell, William J. M. Probert, Shou-Li Li, Michael Tildesley, and Matthew Ferrari. Harnessing multiple models for outbreak management. *Science*, 368(6491):577–579, May 2020.

[10] Loïc Berger, Nicolas Berger, Valentina Bosetti, Itzhak Gilboa, Lars Peter Hansen, Christopher Jarvis, Massimo Marinacci, and Richard D. Smith. Rational policymaking during a pandemic. *Proceedings of the National Academy of Sciences*, 118(4):e2012704118, January 2021.

[11] Spencer J. Fox, Pratyush Potu, Michael Lachmann, Ravi Srinivasan, and Lau-

ren Ancel Meyers. The COVID-19 herd immunity threshold is not low: A re-analysis of European data from spring of 2020. *medRxiv:2020.12.01.20242289*, December 2020.

[12] Richard Horton. *The COVID-19 Catastrophe: What's Gone Wrong and How To Stop It Happening Again*. Polity Press, Cambridge, 2 edition, January 2021.

[13] Ivette Perfecto and John Vandermeer. Biodiversity Conservation in Tropical Agroecosystems. *Annals of the New York Academy of Sciences*, 1134(1):173–200, 2008.

[14] Olivier Restif, David T. S. Hayman, Juliet R. C. Pulliam, Raina K. Plowright, Dylan B. George, Angela D. Luis, Andrew A. Cunningham, Richard A. Bowen, Anthony R. Fooks, Thomas J. O'Shea, James L. N. Wood, and Colleen T. Webb. Model-guided fieldwork: Practical guidelines for multidisciplinary research on wildlife ecological and epidemiological dynamics. *Ecology Letters*, 15(10):1083–1094, 2012.

[15] Ian R Ball, Hugh P Possingham, and Matthew Watts. Marxan and relatives: software for spatial conservation prioritisation. *Spatial Conservation Prioritisation: Quantitative Methods and Computational Tools*, 14:185–196, 2009.

[16] Paul C. Cross, Diann J. Prosser, Andrew M. Ramey, Ephraim M. Hanks, and Kim M. Pepin. Confronting models with data: The challenges of estimating disease spillover. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1782):20180435, August 2019.

[17] Van Kinh Nguyen, Frank Klawonn, Rafael Mikolajczyk, and Esteban A. Hernandez-Vargas. Analysis of Practical Identifiability of a Viral Infection Model. *PLOS ONE*, 11(12):e0167568, December 2016.

[18] G. A. F Seber and C. J. Wild. *Nonlinear Regression*. John Wiley & Sons, Hoboken, 2 edition, 2003.

[19] Daniela De Angelis, Anne M. Presanis, Paul J. Birrell, Gianpaolo Scalia Tomba, and Thomas House. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, 10:83–87, March 2015.

[20] Madeline A. E. Peters, Megan A. Greischar, and Nicole Mideo. Challenges in forming inferences from limited data: A case study of malaria parasite maturation. *Journal of The Royal Society Interface*, 18(177):20210065, April 2021.

[21] James S. Hodges. *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. CRC Press, Boca Raton, 2016.

[22] Xavier A. Harrison, Lynda Donaldson, Maria Eugenia Correa-Cano, Julian Evans, David N. Fisher, Cecily E. D. Goodwin, Beth S. Robinson, David J. Hodgson, and Richard Inger. A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6:e4794, May 2018.

[23] Andrew Gelman, John Carlin, Hal Stern, Donald Rubin, David Dunson, and Aki Vehtari. *Bayesian Data Analysis*. CRC Press, Boca Raton, 3 edition, 2020.

[24] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

[25] Alain F Zuur and Elena N Ieno. A protocol for conducting and presenting results of regression-type analyses. *Methods in Ecology and Evolution*, 7(6):636–645, 2016.

[26] Maia Martcheva. *An Introduction to Mathematical Epidemiology*. Springer, New York, 2015.

[27] Krisztian Magori and John M Drake. The population dynamics of vector-borne diseases. *Nature Education Knowledge*, 4(4):14, 2013.

[28] Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, 83(3):036701, March 2011.

[29] M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering*, 8(5):447–455, September 2006.

[30] Christopher C Drovandi, Christopher Holmes, James M McGree, Kerrie Mengersen, Sylvia Richardson, and Elizabeth G Ryan. Principles of Experimental Design for Big Data Analysis. *Statistical science : a review journal of the Institute of Mathematical Statistics*, 32(3):385–404, August 2017.

[31] Herman Chernoff. Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, pages 586–602, 1953.

[32] Luc Pronzato and Eric Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1):103–120, July 1985.

[33] Elizabeth G. Ryan, Christopher C. Drovandi, James M. McGree, and Anthony N. Pettitt. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*, 84(1):128–154, 2016.

[34] Cristian Chiuchiolo, Janet van Niekerk, and Haavard Rue. Joint Posterior Inference for Latent Gaussian Models with R-INLA. *arXiv:2112.02861 [stat]*, December 2021.

[35] Antony M. Overstall, David C. Woods, and Ben M. Parker. Bayesian Optimal Design for Ordinary Differential Equation Models With Application in Biological Science. *Journal of the American Statistical Association*, 115(530):583–598, 2020.

[36] Christopher C. Drovandi and Minh-Ngoc Tran. Improving the Efficiency of Fully Bayesian Optimal Design of Experiments Using Randomised Quasi-Monte Carlo. *Bayesian Analysis*, 13(1):139–162, March 2018.

[37] D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, December 1956.

[38] Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Institute of Mathematical Statistics*, 10(3):273–304, 1995.

Chapter 2

Spatial epidemiology and adaptive targeted sampling to manage the Chagas disease vector *Triatoma dimidiata*

Abstract

Widespread application of insecticide remains the primary form of control for Chagas disease in Central America, despite only temporarily reducing domestic levels of the endemic vector *Triatoma dimidiata* and having little long-term impact. Recently, an approach emphasizing community feedback and housing improvements has been shown to yield lasting results. However, the additional resources and personnel required by such an intervention likely hinders its widespread adoption. One solution to this problem would be to target only a subset of houses in a community while still eliminating enough infestations to interrupt disease transfer. Here we develop a sequential sampling framework that adapts to information specific to a community as more houses are visited, thereby allowing us to efficiently find homes with domicil-

iary vectors while minimizing sampling bias. The method fits Bayesian geostatistical models to make spatially informed predictions, while gradually transitioning from prioritizing houses based on prediction uncertainty to targeting houses with a high risk of infestation. A key feature of the method is the use of a single exploration parameter, α, to control the rate of transition between these two design targets. In a simulation study using empirical data from five villages in southeastern Guatemala, we test our method using a range of values for α, and find it can consistently select fewer homes than random sampling, while still bringing the village infestation rate below a given threshold. We further find that when additional socioeconomic information is available, much larger savings are possible, but that meeting the target infestation rate is less consistent, particularly among the less exploratory strategies. Our results suggest new options for implementing long-term *T. dimidiata* control.

Author summary

Effective public health interventions for the control and elimination of neglected tropical diseases require an efficient use of resources while still causing long-term disease reduction at the community level. To use resources to best effect, areas most in need of control efforts must be identified. However, strategies for correctly identifying these areas are rarely known due to the complex environmental, biological, and cultural factors shaping disease spread. In turn, incorrect prioritization of control targets can cause the intervention to have no lasting effect. We address this tradeoff between efficiency and efficacy by adapting control priorities throughout an intervention, targeting areas of high uncertainty during the initial stages while shifting to areas of greatest risk at later stages. In the context of controlling *Triatoma dimidiata*, the

primary vector of Chagas disease in several countries in Latin America, our methods provide a means of targeting only a subset of homes for insecticide and housing improvements, while still reducing a village's overall infestation rate below the critical threshold.

2.1 Introduction

Chagas disease is a vector-borne neglected tropical disease (NTD) endemic to all countries in Latin America [1]. It is the most serious parasitic disease in the region, with a 2005 estimate of disease burden 5 to 10 times greater than malaria [2], and is mainly a threat to people living in poverty [3, 4]. The disease, which can lead to potentially fatal cardiovascular or gastrointestinal issues, is caused by the parasite *Trypanosoma cruzi* and transmitted by insects in the Triatominae subfamily [5]. Control initiatives for Chagas primarily focus on disrupting the transmission pathway to humans by reducing domestic Triatomine infestation levels, which is the primary mode of infection [6, 7]. A common control target is to reduce the proportion of infested households in a community to below 5% [8, 9], and it has been shown that reduction past 8% is sufficient to eliminate *T. cruzi* seroprevalence in children aged 6 months to 15 years [10].

In Central America, the prevalence of Chagas disease has declined significantly since the 1980s, thanks in part to the near-elimination of the invasive vector *Rhodnius prolixus*. However, the species-complex *Triatoma dimidiata* still poses a significant health risk to millions of people in many areas [11]. Unlike the invasive *R. prolixus*, *T. dimidiata* is endemic to Central America, living in peridomestic and sylvatic as well as domestic environments [12]. Unfortunately, efforts to control domestic *T. dimidiata*

populations are complicated by the resilience of *T. dimidiata* to traditional methods of vector control. Domestic populations of the insect can rebound within several months of insecticide spraying [13, 14, 15], and continuing to spray houses which report reinfestation appears to have little long-term effect [16]. Control measures for *T. dimidiata* are further complicated due to its significant variation in habitat, morphology, feeding patterns, and genetics [17, 18, 19], all of which interact to cause variation in its vulnerability to insecticides [20] and the domiciliary risk factors associated with its presence [9, 21]. Further, the sustainability of a given control strategy depends critically on cultural practices in the area [22, 23, 24]. Thus, meeting the goal of long-term *T. dimidiata* reduction requires adaptive, locale-specific strategies for surveillance and control [25, 26].

The limitations of insecticide for control of *T. dimidiata* in Guatemala have led to the gradual adoption of additional measures, mostly in the departments of Jutiapa and Chiquimula in southeastern Guatemala, which represent the majority of the country's cases reported to the Ministry of Health [27, 11]. A promising multidisciplinary approach, often referred to as the EcoHealth approach, applies cost-effective, locally-tailored house and peridomestic improvements by collaborating with villagers and health personnel, in conjunction with initial insecticide application [22]. A pilot study of two villages found the method led to low ($< 5\%$) infestation rates 5 years after housing improvements [28, 21], and an expansion of the project to five villages in Chiquimula led to a sustained four-fold reduction in infestation [25]. This suggests house improvements following the EcoHealth approach can effectively prevent reinfestation in the long term.

Barriers to the widespread adoption of community engagement-based interven-

tions include frequent shortages in insecticides [25] and the need for research experts and trained personnel to work with residents and identify infested houses. One possible solution to these issues is to more efficiently select homes in need of insecticides and improvements while still meeting necessary control targets. For example, by treating only a sample of homes such that overall the *T. dimidiata* infestation rate in the village goes below 5%, residual dispersal and non-domestic migration will be limited to homes that were recently improved, or were already unlikely to be suitable for infestation. Further, the EcoHealth approach's emphasis on practical improvements and community participation could help ensure that the risk factors identified from this sample continue to be addressed throughout the entire village. This control strategy would in turn free up resources to be applied in other communities.

To be successful, such a strategy must balance the incentive to target houses that are believed to be infested, with the need to correctly identify the infestation status of unvisited homes. Selecting houses to treat based on perceived infestation risk will quickly bring the village's infestation rate closer to the 5% goal, at least in the short term. However, a diverse range of samples throughout the area and across combinations of possible risk factors is required to reliably find the remaining infested houses, and to correctly predict whether the remaining number of infested houses is below that required to meet the 5% threshold [29, 9]. This was the conclusion in King et al. (2011), where a subset of households was inspected for Triatomines in villages across Guatemala using either random sampling or sampling based on pre-defined risk factors, and was used to predict whether the village infestation rate was below or above the 5% threshold. The authors found random sampling to consistently have higher prediction accuracy, noting that sampling based on a fixed set of factors failed

to explore the ways factors associated with infestation vary among different villages and regions [9]. In short, the dual objectives of prediction and targeted sampling lead to a exploration vs. exploitation tradeoff, where the space of houses to target must be searched to find configurations that are both of minimal sample size, and which contain enough information to correctly predict that the 5% target has been met.

The quantity and quality of covariate information available for exploration is a second key factor in the success of an intervention strategy. A number of studies have identified various socioeconomic factors associated with infested houses, such as the material and condition of house walls [4, 21, 30]. Therefore, if additional dependent variables are available prior to selecting houses for treatment, fewer observations may be needed to make accurate estimates. Another option is to rely only on variables available remotely, such as elevation. While this sacrifices potentially useful information, it may ultimately be more cost effective, since collecting socioeconomic risk factors for inference and prediction requires additional labor and logistical planning [31].

Here we aim to address the problem of treating a subset of houses to reduce *T. dimidiata* infestation to a target threshold while minimizing the necessary resources. Using 5 villages of varying size and baseline infestation rates in Chiquimula, Guatemala as a case study, we employ *adaptive geostatistical design* strategies which sequentially select houses based on observations from previous iterations, and use inherent spatial autocorrelation in the observed data to improve prediction and inference.

While historically quite theoretically driven, the principles of geostatistical design have recently been applied to other problems of survey design and analysis in spatial

epidemiology [32]. Chipeta et al. (2016) developed an adaptive sampling method which targets locations with high spatial uncertainty, and applied the approach to a cross-sectional malaria survey [33, 34]. Adaptive sampling using a similar strategy based on prediction entropy and spatial exploration was also shown to be effective for identifying hotspots of lymphatic filariasis [35]. Fronterre et al. (2020) used a non-adaptive, lattice-like sampling design combined with close pairs of points (proposed first in [36]) to predict whether an area's disease prevalence exceeds a certain threshold, and found the method outperformed a current WHO assessment protocol on a simulated dataset [37].

In this work, we develop a class of adaptive strategies which transition from prioritizing houses based on prediction uncertainty to houses based on percieved risk of infestation. We compare these strategies to random sampling with empirical data, and assess their ability to efficiently locate infested houses while correctly predicting whether the current selection meets the reduction target. Additionally, we examine the effect of including socioeconomic covariates on the performance of each strategy. In the context of Chagas vector control with the EcoHealth approach, our methods address two key questions: 1) how can houses be more efficiently targeted for treatment to sufficiently reduce village-wide vector incidence? and 2) can further efficiency gains be made by collecting additional socioeconomic information? More generally, our methods provide a formal, statistical framework for targeted control strategies in a resource-limited setting, and hence are particularly relevant to the control of NTDs.

2.2 Methods

2.2.1 Data preparation

Our data come from the follow-up EcoHealth project discussed in the introduction, which was conducted in the countries Honduras, El Salvador, and Guatemala, between August and December 2011, and is described in detail in Bustamante et al. (2015) and Lima-Cordón et al. (2018) [21, 18]. We focus only on the five villages in Chiquimula, Guatemala, since infestation rates were low in the other countries. These villages lie along an altitudinal gradient, with a climate ranging from hot and humid to cooler cloud forest. Villages are surrounded by a mix of banana plantations, shade grown coffee, and patches of the original forest [18].

All houses with missing factors necessary for our analyses were removed, leaving between 72% and 83% of the total number of houses recorded in each village (Table 2.3). After processing, there were 172 housing structures in El Amatillo, 147 in El Cerrón, 251 in El Guayabo, 108 in El Paternito, and 207 in La Prensa, for a total of 885 observations. The village-wide infestation rate was between 15% and 39%.

Each data entry was obtained by two trained personnel using the following protocol. After the informed consent of the residents, houses were searched for 35-45 minutes by one team member with a flashlight and forceps, searching walls, behind furniture, and other suitable environments for Triatomine shelter, while another performed interviews and assessed aspects of tidiness in the home [28, 21]. The home's geocoordinates were also recorded. These surveys produce a binary response indicating Triatomine presence in the home, and 26 covariates. A positive response indicates that adult or juvenile insects, dead insects, or eggs were found. The covariates, listed

in Table 2.4, include items related to socioeconomic factors, the number and type of domestic animals, and the house's structure and cleanliness. We added two more covariates based on the house's geocoordinates. The *distance to perimeter* is the shortest distance of the home to the village's convex hull, plus 50m, while the *density* is the number of other houses within 100m. Covariates were checked for multicollinearity, and all continuous variables were centered and scaled. Additionally, for convenience in setting priors, the coordinates of the houses were scaled such that the diameter of the village (maximum distance between any two points) was one.

2.2.2 Hierarchical modeling for geostatistics

Geostatistics is a field which studies spatial autocorrelation in point-referenced data, and leverages this information for inference and prediction [38]. Geostatistical models incorporate a spatial phenomena $Z = \{z(\boldsymbol{s}) \in \mathbb{R} \mid \boldsymbol{s} \in \mathscr{D}\}$ over a domain of possible locations $\mathscr{D}$, where $n = |\mathscr{D}|$ when $\mathscr{D}$ is discrete. The closer two points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are to each other, the more similar the values $z(\boldsymbol{s}_i)$ and $z(\boldsymbol{s}_j)$ will tend to be. This spatial surface Z is itself a function of possibly unknown *spatial parameters*, which control how the covariance between points behaves. Rather than observing Z directly, for each location $\boldsymbol{s}$ there is typically a measurable response $y(\boldsymbol{s})$, which is assumed to be a function of Z and some covariates $\boldsymbol{x}(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s}))^\top$.

Following the general hierarchical framework first outlined in [39], we assume the response at each $\boldsymbol{s} \in \mathscr{D}$ follows a generalized linear model with spatially correlated random effects. In our setting, this amounts to $y(\boldsymbol{s})$ being a binary variable indicating the infestation status of a home at position $\boldsymbol{s}$, which has a probability $r(\boldsymbol{s})$ of being infested. The probability $r(\boldsymbol{s})$, or the *risk* of having an infestation at location $\boldsymbol{s}$, will

then depend on the home's covariates and the risks of other homes in close proximity. More formally, the likelihood of $y(\boldsymbol{s})$ is as follows:

$$y(\boldsymbol{s}) \mid r(\boldsymbol{s}) \sim \text{Bernoulli}(r(\boldsymbol{s})) \tag{2.1}$$

$$\text{logit}(r(\boldsymbol{s})) = \eta(\boldsymbol{s}) = \boldsymbol{\beta}^\top \boldsymbol{x}(\boldsymbol{s}) + z(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}), \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of fixed effect coefficients. The spatial surface Z follows a zero-centered Gaussian distribution with Matérn covariance function (defined in Section 2.5.1) with smoothness parameter $\nu = 1$ [40]. This spatial process has two parameters σ_s and ρ, which respectively control the variance and *effective range*, here defined as the distance at which the correlation between two points reaches 0.1. Finally, ε is an independent random effect representing non-spatial latent variability at each location, which helps avoid finding spurious spatial correlation [37]. We assume $\varepsilon \sim N(0, \sigma_e^2)$.

The equations above describe the probability a house is infested, given some parameters and covariate information. In other words, they specify an assumption about how our response is generated as a function of these parameters. However, we are interested in reversing this process: given some finite set $\mathscr{S} \subset \mathscr{D}$ of locations, we will use the household data from these locations to make inferences about possible values of the parameters. Following convention, we write $\boldsymbol{y} = \{y(\boldsymbol{s}) \mid \boldsymbol{s} \in \mathscr{S}\}$ as the observed response, and $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \rho, \sigma_s, \sigma_e)^\top$ as the parameters to be estimated. Under the Bayesian paradigm, one treats $\boldsymbol{\theta}$ as a random variable and assigns priors based on domain knowledge or on hypotheses formed prior to data collection. Bayes'

theorem then gives the posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{y})$, which represents our updated beliefs about $\boldsymbol{\theta}$ given the data we have observed at $\mathscr{S}$. The posterior distribution can further be used to make predictions at unmeasured locations.

For our analysis, we use weakly-informative $N(0, 3.3)$ priors for $\boldsymbol{\beta}$, while for the spatial range and standard deviation we use the penalized-complexity prior of [41], set to induce tail probabilities of $\Pr(\rho < 0.1) = 0.05$ and $\Pr(\sigma_s > 3) = 0.1$. The variance for the non-spatial random effects σ_e^2 follows an inverse-gamma distribution with location 1 and scale 0.01.

Data preparation and analysis was performed in R version 4.1.0 [42]. For computational speed and convenience, all statistical models were fit using Integrated Nested Laplace Approximation (INLA) with the Stochastic Partial Differential Equation (SPDE) representation for the spatial effects, available from the R-INLA package [43, 44]. All materials necessary for the analysis are publicly available online [45], including a brief tutorial on spatial modeling with INLA.

2.2.3 Model comparison and full-village analysis

To verify the suitability of the model outlined above, we compare its performance to two simpler alternatives. The first removes the correlated spatial random effects $Z(\boldsymbol{s})$ while the independent $\varepsilon(\boldsymbol{s})$ effects are removed from the other, but otherwise each model is the same. Models are evaluated based their *deviance information criterion* (DIC) and *marginal likelihood* (ML), formally defined in Section 2.5.1. Both measure a model's goodness-of-fit to the data while penalizing model complexity.

Each village is analyzed separately, using all available data within the village. Additionally, we consider two sets of covariates to be available. The *global* covariate

set contains only variables obtainable from the geocoordinates, namely, the location's density and distance to the perimeter, which represents a small amount of covariate information that is convenient to collect. The other set contains these variables along with the socio-economic covariates listed in Table 2.4.

2.2.4 Predicting out-of-sample infestation status

Let $\boldsymbol{y}_0 = \{y(\boldsymbol{s}_0) \mid \boldsymbol{s}_0 \in \mathscr{D} \backslash \mathscr{S}\}$ indicate the unknown infestation status at unvisited locations. Given a response $\boldsymbol{y}$ observed at $\mathscr{S}$, the joint posterior predictive distribution for $\boldsymbol{y}_0$ is then

$$\Pr\left(\hat{\boldsymbol{y}}_0 \mid \boldsymbol{y}\right) = \int \Pr\left(\hat{\boldsymbol{y}}_0 \mid \boldsymbol{y}, \boldsymbol{\theta}\right) P\left(\boldsymbol{\theta} | \boldsymbol{y}\right) d\boldsymbol{\theta}. \tag{2.3}$$

We are interested in the total count of unvisited locations which are infested, defined as $I_0 = \sum \boldsymbol{y}_0$. To estimate the distribution $\Pr\left(\hat{\boldsymbol{y}}_0 \mid \boldsymbol{y}\right)$, and hence $\Pr\left(\hat{I}_0 \mid \boldsymbol{y}\right)$, from the posterior, we generate 5,000 Monte Carlo samples from $P\left(\boldsymbol{\theta} | \boldsymbol{y}\right)$, then for each of these samples $\boldsymbol{\theta}^{(i)}$, we draw a sample from $\Pr\left(\hat{\boldsymbol{y}}_0 \mid \boldsymbol{y}, \boldsymbol{\theta}^{(i)}\right)$, resulting in 5,000 samples from (2.3).

2.2.5 An adaptive sampling strategy for infestation reduction

In the present study, our objective is not only to make predictions given data observed at $\mathscr{S} \subset \mathscr{D}$, but to choose the set $\mathscr{S}$ itself to best control infestation. In this context, $\mathscr{S}$ is referred to as the *sampling design*. To evaluate the quality of a sampling design, we specify a *design target*. In our context, this target is the number of houses selected for treatment (i.e. the size of $\mathscr{S}$), subject to the constraint that the true infestation

rate among unvisited houses is below 5% (i.e. $I_0/n < 0.05$).

One possible strategy for choosing effective sampling designs is adaptive (or sequential) sampling. Rather than specifying our set of houses to sample completely before collecting data at these houses, we sample houses in batches: a collection of houses is selected, the data is gathered at these houses, a model is fit to the data so far, and a new batch of houses is chosen based on the model's predictions. At the end of this process, the set of houses we have visited becomes our final sampling design.

Implementing our adaptive strategy requires the following [33]: 1) an initial design $\mathscr{S}_1$ from which to fit the first model, 2) a *batch size* b as the number of new houses to add to the existing observations each iteration, 3) a *utility function* U to rank unobserved houses to target. For our application, we additionally require 4) a *termination condition* to predict whether the current design meets the infestation target.

For the utility function, a natural choice might be to rank a location $\boldsymbol{s}_0 \in \mathscr{D} \setminus \mathscr{S}$ according to its posterior predicted risk $\Pr(\hat{y}(\boldsymbol{s}_0) = 1 \mid \boldsymbol{y}) = \mathbf{E}[r(\boldsymbol{s}_0) \mid \boldsymbol{y}]$. However, this strategy has high sampling bias and fails to prioritize locations with potentially new information, hence ignoring possibly large amounts of the geospatial or covariate search space.

To better explore this space, we therefore balance this strategy with prioritizing locations of high uncertainty. Since predictions are generally less reliable with smaller sample sizes, the proposed utility function transitions smoothly from prioritizing a location's variance $\tilde{\nu}(\boldsymbol{s}_0)$, to prioritizing its expected risk $\tilde{r}(\boldsymbol{s}_0)$, where $\tilde{\nu}(\boldsymbol{s}_0)$ and $\tilde{r}(\boldsymbol{s}_0)$ have each been centered and scaled over the unvisited locations. The marginal posterior risk and variance based on an example sample is shown in Figure 2.5. If

m_i is the current number of locations observed at iteration i of sampling, we rank locations according to the utility function

$$U(\boldsymbol{s}_0) = t(i;\alpha)\tilde{r}(\boldsymbol{s}_0) + (1 - t(i;\alpha))\,\tilde{\nu}(\boldsymbol{s}_0) \tag{2.4}$$

where $t(i;\alpha) = ((m_i - m_1)/(n - m_1))^\alpha$ is a weighting function interpolating between 0 and 1. The *exploration parameter* $\alpha > 0$ controls the speed at which the predicted risk is prioritized, with $\alpha < 1$ transitioning to risk-based targeting more quickly and $\alpha > 1$ favoring exploration for longer.

We propose a termination condition based on our confidence that the current design satisfies the reduction target:

$$\Pr\left(\hat{I}_0 < \kappa n\right) \geq \gamma, \tag{2.5}$$

where κ is the desired infestation rate, and γ the desired confidence level. If $\mathscr{S}_i$ is the current design, we can compute this probability using draws from $\Pr\left(\hat{I}_0 \mid \boldsymbol{y}_i\right)$ as described above. We fix $\kappa = 0.05$ and $\gamma = 0.95$ throughout the manuscript.

Finally, we set the batch size $b = 3$, and sample the initial design $\mathscr{S}_1$ uniformly at random.

In summary, the adaptive sampling algorithm is as follows.

1. Sample initial design $\mathscr{S}_1$ uniformly and set $i = 1$
2. Fit the posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{y}_i)$ as described in Section 2.2.2
3. If $\mathscr{D} \setminus \mathscr{S}_i$ satisfies (2.5), return $\mathscr{S}_i$. Otherwise, go to step 4
4. Assign each location $\boldsymbol{s}_0 \in \mathscr{D} \setminus \mathscr{S}_i$ a utility $U(\boldsymbol{s}_0)$ using (2.4)

5. Choose the b locations with highest utility and add them to $\mathscr{S}_i$

6. Set $i = i + 1$ and repeat steps 2-6.

2.2.6 Simulation study comparing adaptive and random sampling

We compare the performance of the adaptive sampling procedure from above using several values of the exploration parameter α, along with a random sampling procedure, on each of the five Guatemalan villages. To evaluate each resulting design, we record the size m of the final design, and the true infestation rate, defined as the number of truly infested houses not in the design divided by the total number of houses in the village. For all experiments, we used a batch size $b = 3$.

For each village, the following experiment is repeated 50 times. An initial set of 10 locations is chosen at random. Then, adaptive sampling using each of $\alpha = 0, 0.15, 0.3, 0.7, 1, 2$ is performed, as well as a random procedure which uniformly samples b new houses each iteration, resulting in 7 final designs to be evaluated. In this way, we apply each of the different procedures to the same set of randomized initial samples.

2.3 Results

2.3.1 Full-village analysis

A comparison of the model given in Section 2.2.2 to the two simpler alternatives is shown in Fig 2.1 for each village. The proposed model had a lower (better) DIC than

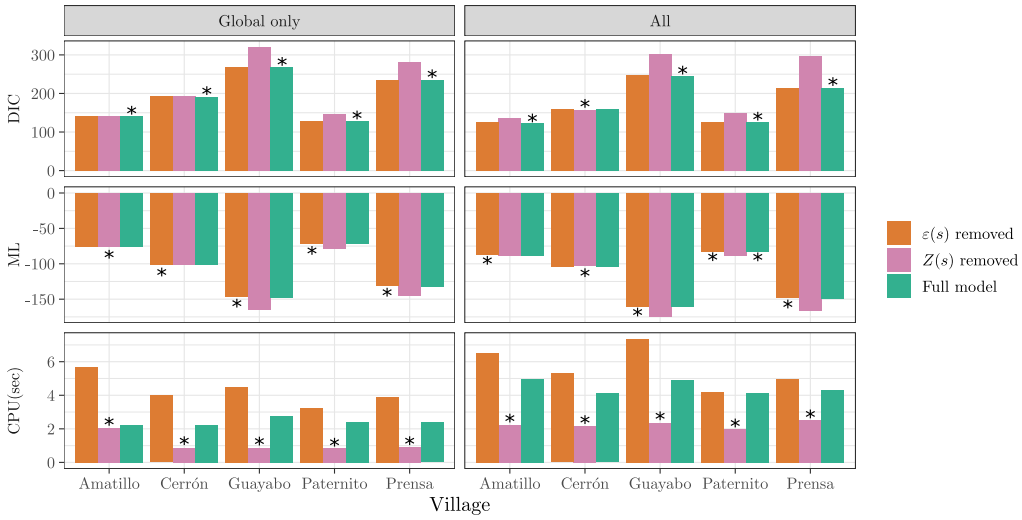

Figure 2.1: Model comparison based on two goodness-of-fit measures and computation time, by village and level of covariate information. The best-performing model is indicated with a star in each case. "Global only" means only covariates available from the coordinates are available, and "All" indicates all 28 covariates were used. "$Z(\boldsymbol{s})$ removed" refers to the model outlined in Sec. 2.2.2 but with no spatial effects, "$\varepsilon(\boldsymbol{s})$ removed" refers to the independent effects removed, and "Both" refers to the full model. DIC and ML are defined in Mathematical definitions.

the alternatives in all five villages, except for El Cerrón with the full covariate set, where the model with the spatial effects removed did slightly better. The spatial-only model ($\varepsilon(\boldsymbol{s})$ removed) had higher ML in all but 3 cases. However, the two models including a spatial effect had similar performance according to both measures. In contrast, in El Guayabo, El Paternito, and La Prensa, these two models had substantially better performance than the model with spatial effects removed. Fig 2.1 also shows the computation time of the proposed model was considerably faster than the spatial-only model, while the model with no spatial effects was fastest.

The proposed model, fit to all data in the village, also allowed us to draw inferences about the nature of *T. dimidiata* infestation at a local level. Table 2.1 summarizes the effective range after rescaling back to meters, while Figure 2.2 shows the full posterior

Village	Covariate set	Mode & 95% HPDI (meters)
El Amatillo	Global only	156 (37-931)
	All	150 (62-420)
El Cerrón	Global only	728 (4-8201)
	All	553 (6-7370)
El Guayabo	Global only	774 (334-1910)
	All	823 (358-2062)
El Paternito	Global only	651 (161-2860)
	All	874 (171-3751)
La Prensa	Global only	480 (212-1138)
	All	508 (205-1281)

Table 2.1: Posterior of the effective range from the full-village analysis. Each posterior is fit according to the model of Section 2.2.2, using two different sets of covariate information. The 95% highest posterior density interval (HPDI) is the smallest interval such that 95% of the posterior mass is contained within it.

distribution. The posterior mode was between 150m and 874m, or between 10% and 40% of the village diameter. Although there was notable variation in the effective range between the villages, it appears fairly similar between the two covariate sets for a given village, with a difference in mode between 6m and 223m. In the villages El Cerrón and El Paternito, the 95% highest posterior density interval (HPDI) for the effective range was quite large (spanning several kilometers), which suggests the spatial signal was weakest in these villages.

Examining the coefficients for the fixed effects, we found variation between the villages in the majority of the covariates in the full covariate set (Fig 2.3). The factors which consistently had a negative association with infestation, defined here as having a 50% HPDI fully below 0, were not having rats in the home (5 villages), having bedroom walls in good condition (4 villages), and not keeping construction materials around the home (4 villages). The factors with a consistent positive association were having a kitchen outside the home (4 villages), evidence of bird nests inside (3

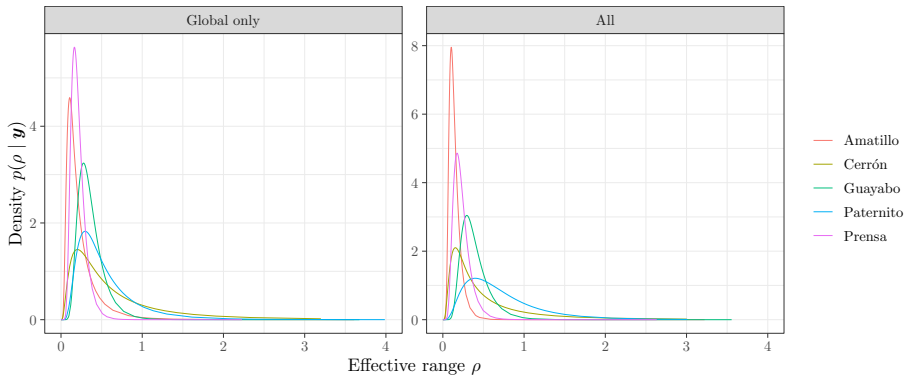

Figure 2.2: Effective range from full-village analysis. The posterior marginal $p(\rho \mid \boldsymbol{y})$ of the effective range, from fitting the model of Section 2.2.2 to all locations in each village. "Global only" means only covariates available from the coordinates are available, and "All" indicates all 28 covariates were used.

villages), and keeping chickens outside the home (3 villages).

2.3.2 Simulation study results

The results of the simulation experiment comparing adaptive and random sampling strategies are shown in Fig 2.4. From the 50 final designs obtained from each group of strategy × village × covariate set, we calculated the mean and 90% confidence interval (CI) for the percentage of houses in the final design, and for the true infestation rate remaining in the village. In all villages and both covariate sets, all adaptive strategies had a smaller mean design size (final number of sampled houses) than random. Moreover, when the exploration parameter was less than $\alpha = 2$, the 90% CI for the design size was entirely below (i.e. not overlapping) the CI for random sampling. Comparing the top and bottom rows of Fig 2.4, we further find that, except for when $\alpha = 2$ and random sampling was used, having the full covariate set tended to lead to a smaller design size compared to having the global covariate set.

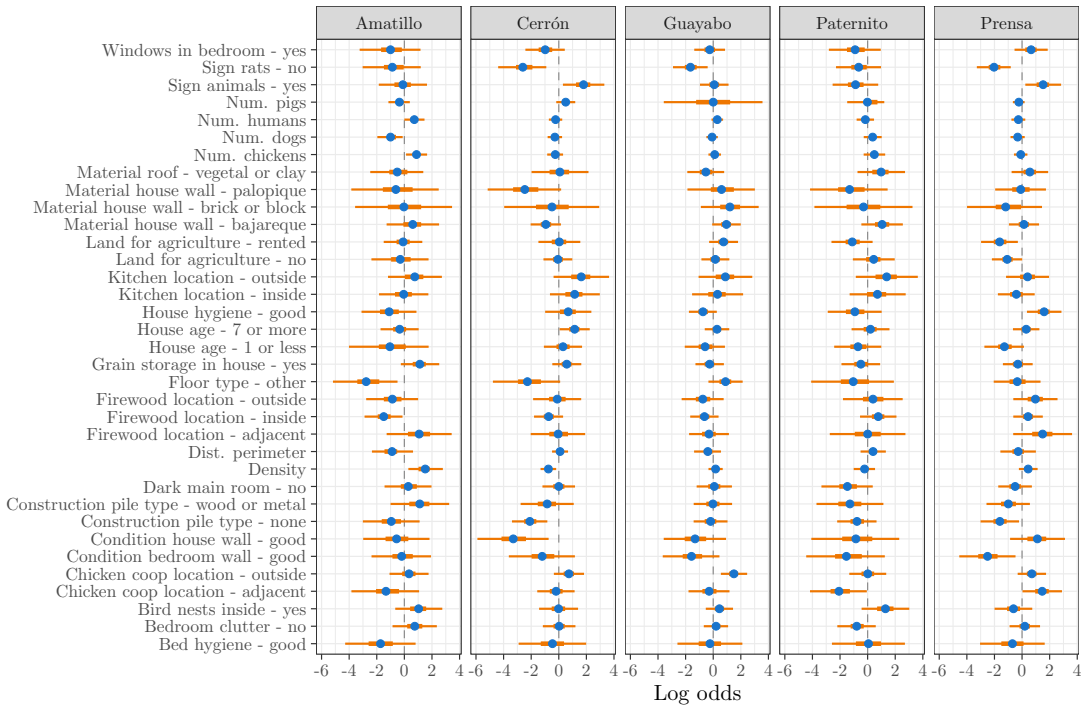

Figure 2.3: Fixed effect coefficients from fitting the model of Section 2.2.2 to all available data in each village. Coefficients are on log odds scale. The blue point is the posterior mean, the bold orange line is the 50% HPDI, and the thin orange line is the 95% HPDI.

Although the adaptive strategies tended to produce smaller designs, they were also more likely to produce designs which failed to satisfy the design target of reducing the true infestation rate below 5%, especially when the exploration parameter α was too low. This highlights the importance of exploring the design space during early stages of sampling, in order to mitigate prediction bias. With the global covariate set, in four villages all adaptive strategies had a 90% confidence interval below the target threshold, i.e. at least 95% of the designs had a true infestation rate below 5% in the final sample, while in La Prensa only $\alpha \geq 0.7$ gave a CI below the target. With the full covariate set, the strategies with lower α missed the target threshold more

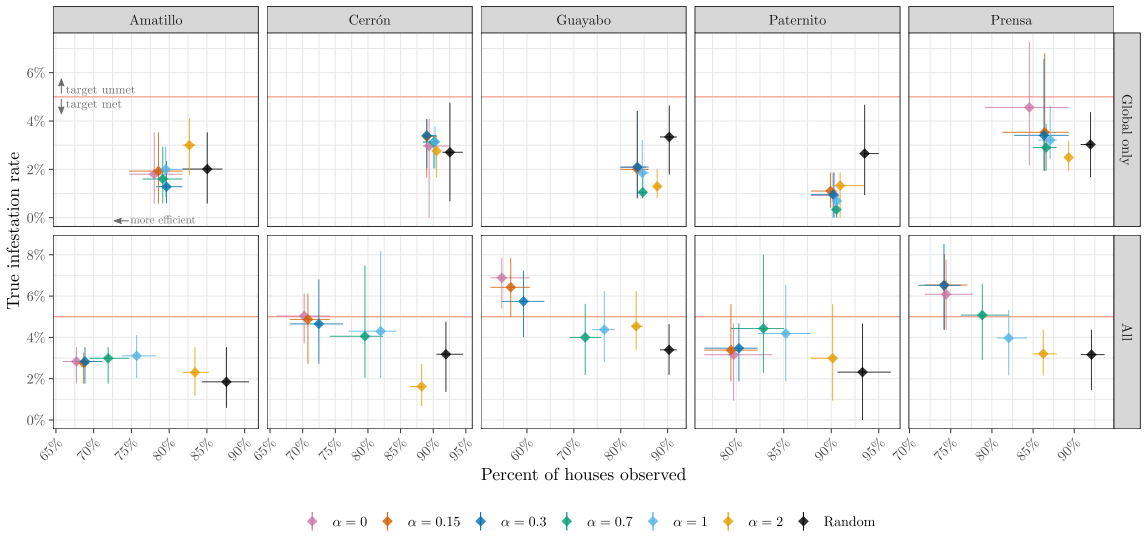

Figure 2.4: Results from the simulation experiment for each village and predictor set. The adaptive sampling procedure for varying α, along with random sampling, is performed for 50 initial sets of 10 houses. The x-axis is the percentage of houses in the final sampling design, and the y-axis is the true infestation rate remaining in the village. Diamonds show means, while cross-hairs show the 90% confidence intervals of the data corresponding to each axis. The red line indicates the 5% reduction target.

frequently. The mean infestation rate was above 5% for $\alpha \leq 0.3$ in two villages, and in one village for $\alpha = 0.7$. Moreover, the CI contained 5% in four villages for $\alpha \leq 1$, and one village for $\alpha = 2$. The CI was always below 5% with random sampling.

Table 2.2 provides a higher-level perspective of the relative performance of the adaptive strategies, with results from all villages pooled together. Again, we find that the accuracy (percentage of designs which met the 5% infestation target) was lowest for lower values of α, especially with the full covariate effect, and that accuracy was inversely proportional to the difference in design size compared to random.

2.4 Discussion

The resilience of NTDs, combined with the resource-limited context in which they inherently reside, presents unique challenges for their control and elimination. Interventions must not only provide effective and long-term disruption of the disease, but also be logistically feasible to allow widespread application [46]. This leads to a difficult tradeoff between preserving resources while actually making a difference at the community level. Because of the complex environmental, ecological, and cultural aspects of NTDs, a further challenge is the need for spatially and temporally adaptive solutions for efficient control [47, 26].

We have proposed an algorithm which balances this tradeoff between efficiency and efficacy in the context of Chagas disease vector control in southeastern Guatemala. With the goal of reducing a village's domiciliary Triatomine infestation rate below the government target of 5%, we consider an experimental design problem where only a subset of houses are treated for long-term vector elimination using locale-specific housing improvements. The algorithm uses a combination of adaptive sampling and Bayesian geostatistical modeling to minimize the number of treated houses while correctly predicting when the reduction target has been met.

The results from our simulation experiment show adaptive sampling strategies are universally more efficient than selecting houses at random, and are able to consistently predict whether the 5% infestation target has been achieved. In the case where only the spatial coordinates of houses are available, even the least exploratory strategy (i.e. $\alpha = 0$) was able to correctly predict the target was met with over 90% accuracy. This is surprising, as selecting houses based on current perceived risk will intuitively lead

Exploration parameter	Covariate set	Accuracy (%)		Difference (%)
		5% target	8% target	median (95% CI)
$\alpha = 0$	Global only	91.6	100	4.8 (1.6-11.7)
	All	54.4	97.6	19.4 (11.2-37.3)
$\alpha = 0.15$	Global only	97.6	100	4.8 (2.0-10.6)
	All	52.8	97.2	18.9 (11.2-36.1)
$\alpha = 0.3$	Global only	98.0	100	4.4 (2.0-9.6)
	All	57.2	97.2	18.4 (11.2-33.7)
$\alpha = 0.7$	Global only	99.6	100	4.4 (2.0-7.2)
	All	77.2	97.6	14.1 (8.4-19.4)
$\alpha = 1$	Global only	100	100	4.4 (2.0-7.3)
	All	81.2	97.6	10.6 (5.6-15.7)
$\alpha = 2$	Global only	100	100	2.8 (0-4.1)
	All	94.8	100	4.4 (1.8-8.4)
Random	Global only	99.2	100	-
	All	98.4	100	-

Table 2.2: Performance of adaptive sampling in the simulation study, compared to random sampling. Accuracy is the percentage of designs which met the control target, i.e. the out-of-sample infestation rate was below 5% (or 8%). Difference from random is the percentage of the village in the design, minus the corresponding percentage using random sampling, calculated for each initial design.

to biased predictions [32]. One possibility for this is that the spatial information alone leads to minimal shrinkage in the posterior during the initial stages of sampling, which would cause sampling to be nearly random even when $\alpha = 0$ (since unvisited locations will look very similar). A second explanation is that the termination condition we have used counteracts the effects of targeting perceived risk. In particular, sampling in this way can lead to higher variance at unvisited locations since we have neglected to gather relevant information about them, leading to a situation where both the expectation of $\hat{I}_0$ and the probability in (2.5) are low, and hence avoiding terminating prematurely.

Our second major finding is that including additional socioeconomic covariates

can significantly reduce the number of sampled houses, but only for the adaptive sampling strategies. However, this additional information also leads to a greater opportunity for bias in the adaptive strategies with lower exploration parameter, and even larger values of the parameter ($\alpha = 1, 2$) occasionally failed to meet the target, as did random sampling. While this is not ideal, it is reassuring that even with no exploration ($\alpha = 0$), the true infestation rate was below 8% in nearly all of the final designs (Table 2.2). Since this threshold has been shown to guarantee *T. cruzi* elimination throughout Central America [10], a viable strategy could be to set the target threshold slightly below the level actually desired, thus allowing maximum sampling efficiency while still meeting the actual target.

While the assumption that detailed socioeconomic information is available for the whole community beforehand is not necessarily realistic, the efficiency gained by including these covariates opens up new options for *T. dimidiata* control. For example, this information could be collected beforehand by community members or with a quick survey, thereby reducing the number of homes to search for infestation, which is more time-consuming to collect and requires trained personnel. Moreover, this could allow a more thorough search for Triatomine presence, which is a notoriously noisy measurement [48, 21].

On a practical level, our sampling strategy could compliment the EcoHealth approach in several ways. If local resources for housing improvements are readily available, then adaptive sampling could be used to conserve insecticide and the time of health personnel, while allowing house improvements for all residents who want them. If such resources are scarce, adaptive sampling could further indicate which homes are most in need of immediate improvement. Adaptive sampling could also be used to

more quickly develop a localized improvement strategy, by identifying socioeconomic variables most associated with infestation risk from a subsample of relevant houses.

It has been previously suggested that adaptive geostatistical sampling based on prediction variance, followed by targeting areas of higher risk in later stages, could be a way to accurately identify areas for public health interventions while conserving resources [33]. Altogether, our results support this idea, while rigorously comparing different balances of prioritizing variance and risk. Our experiments show that with α sufficiently high, this strategy is robust to different communities and levels of covariate information. The most preferable setting for α, however, will depend on the particularities of a given intervention. For example, in the initial stages of an intervention it may be preferable to visit as many communities as possible, with the understanding that the control target may fail to be reached in a few communities, while in later stages a smaller number of remaining areas can be targeted with a higher level of precision.

2.4.1 Full-village analysis

The results from the full-village analysis, where several models were fit to all available data in each of the villages, provide further insight into factors associated with *T. dimidiata* infestation in southeastern Guatemala. First, the superior explanatory power of the spatial models emphasizes the spatial nature of infestation, something not accounted for in previous studies of infestation risk [4, 21]. There are several possible explanations for the spatial autocorrelation present in this data. One possibility is a limited migratory range from existing domestic populations, leading to local clustering following the insect's dispersal season. Such a dynamic is supported

by several population genetics studies, which have found insects in nearby houses and adjacent villages are more related [49, 50], as well as insects from highly-infested houses and their neighbors following insecticide application [15]. Another explanation is that spatial patterns are due to unmeasured covariates, which are themselves spatially autocorrelated. While our results show spatial effects remain even after accounting for a number of socioeconomic variables, environmental factors such as altitude, temperature, and precipitation have been shown to be mildly correlated with *T. dimidiata* infestation [51, 52].

A second important finding from the full-village analysis is the variation in fixed-effect coefficients among the five villages. While several covariates, such as the condition of bedroom walls, consistently had a meaningful association with infestation, many others varied in their explanatory power, and even whether there was a positive or negative association. Important fixed-effects also varied compared to previous *T. dimidiata* studies in different areas. In particular, the distance from the village perimeter was found to lead to significantly higher infestation rates and vector abundance in Yucatan, Mexico [52]. However, not only was there no correlation ($\rho = 0.02$) between village perimeter and infestation overall in our data, but we found the importance of village perimeter as a covariate was negligible using either covariate set, which shows there is little association between these variables when accounting for spatial and various covariate effects as well. This could be due to the close proximity of several of the villages, or to deforestation leading to a disruption of sylvatic populations in the surrounding environment [53]. Ultimately, these results all add to the existing evidence that risk factors for *T. dimidiata* infestation must be considered in a local context [9, 23].

2.4.2 Limitations and future work

This study has a few limitations. First, our methods focus on bringing the infestation rate below 5% among untreated houses, with the logic that uninfested, unvisited homes already have factors not conducive to *T. dimidiata* infestation, and that the EcoHealth approach empowers communities to maintain these factors throughout the village [22]. However, it is certainly possible for the infestation rate to still rebound. For example, if a previous insecticide application was recent enough, some households may not be uninfested due to favorable conditions but rather that the vector population has had insufficient time to fully reestablish. Further, maintaining certain housing conditions alone may not be enough if changing environmental conditions alter the relationship between housing conditions and infestation risk. Additional research would therefore need to confirm the long-term effects of allowing a small fraction of houses to remain infested after an intervention.

We have also ignored any practical constraints when choosing locations during sampling, such as the travel time to selected points, or the logistics of centralizing incoming data after each iteration of sampling. It therefore may be more realistic to impose a penalty in (2.4) based on distance from current samples, or to increase the batch size b to allow a more natural sampling schedule. Finally, our decision to lump all signs of infestation into a single indicator ignores potentially useful information, including their different implications for disease risk. A recent study in Jutiapa found a large difference between adult and juvenile infestation rates [53], so it may be beneficial to separate these variables.

In the simulation study, several parameters of the adaptive sampling procedure were fixed throughout, such as the batch size b and size of the initial design, to limit

computational overhead. While our results demonstrate the settings we chose can be effective across multiple villages, an analysis involving the interaction between these parameters would provide a more thorough summary of when our procedure works best, and its robustness to different parameter settings.

The methods developed in this work can be applied in contexts other than Chagas vector control. In particular, adaptive geostatistical sampling using the utility function (2.4) and termination condition (2.5) can be applied nearly directly to other NTD control initiatives seeking to efficiently meet a reduction target. For example, schistosomiasis prevalence in schools has been shown to have spatial autocorrelation, and covariate information for schools can be acquired through teacher-given surveys [54]. Adaptive geostatistical sampling could therefore be applied to efficiently target schools at greater risk.

Our methods could also be applied to other Chagas endemic regions, and extended to account for more complex ecological data, such as jointly modeling several vector species and lifestages, or using spatiotemporal models [55] to improve the efficiency of follow-up surveys and adapt to seasonal effects. It would also be interesting to investigate the use of additional environmental variables as an alternative to the socioeconomic information, which may provide similar gains in sampling efficiency while being easier to collect. More generally, our framework could be used for monitoring multiple, possibly interacting, diseases [56]. This would allow sequential samples to take into account information both between and among pathogens. Finally, additional research should go into the development of tools for adaptive, targeted intervention strategies, and how such software can best be applied for the direct benefit of the community.

In conclusion, we have proposed an adaptive strategy for public health interventions, which transitions from prioritizing areas of greatest uncertainty to those perceived to be most at risk. This would allow control initiatives to use resources more efficiently, by targeting areas of greatest need while still benefiting the entire community. We believe the methods used in this work are well-suited to address the complex ecological, biological, and social factors inherent to disease spread, and hence are applicable to a wide range of epidemiological systems.

Acknowledgements

The authors are grateful to the researchers, Ministry of Health personnel, and people of Chiquimula who participated in data collection.

2.5 Supplemental information

Table 2.3: Additional information from the 2011 EcoHealth survey.

Village	Total number houses	Number houses after preparation	Infestation rate (after preparation)
El Amatillo	215	172	15.3%
El Cerrón	205	147	36.7%
El Guayabo	302	251	33.3%
El Paternito	138	108	38.3%
La Prensa	280	207	38.8%
Total	1,140	885	32.3%

Table 2.4: Socioeconomic variables used for model fitting.

House factors	**Values/range**
Bed hygiene	Good, poor
Sign of bird nests inside	Yes, no
Location of chicken coop	Inside or adjacent to house, outside house, none
Clutter in bedroom	Yes, no
Poorly lit bedroom	Yes, no
Floor material	Dirt, other
Piles of construction material	Adobe or clay, wood or metal, none
Firewood location	Inside house, directly outside, outside, none
Grain storage in house	Yes, no
House age	Less than one year, 2-6 years, more than 7 years
House hygiene	Good, poor
Kitchen location	Inside house, outside, shared or none
Land for agriculture	Rented, owned, none
Sign of rats	Yes, no
Sign of other small animals	Yes, no
Bedroom wall condition	Good, deteriorated
Wall condition throughout home	Good, deteriorated
Material house wall	Adobe, bajareque, palopique, brick or other
Material roof	Aluminum or cement, clay or vegetal material, nylon panels
Windows in bedroom	Yes, no
Number residents	1-15
Number chickens	0-60
Number dogs	0-12
Number pigs	0-12

2.5.1 Mathematical definitions

Matérn covariance function

For any pair of points at distance d from each other, the Matérn covariance between these points is

$$C(d) = \frac{\sigma_s}{2^{\nu-1}\Gamma(\nu)}(\kappa d)^{\nu} K_{\nu}(\kappa d),$$

where Γ is the Gamma function and K_ν the modified Bessel function. The parameters σ_s and ν are the spatial standard deviation and smoothness, respectively, while κ is implicitly defined via the effective range $\rho = \sqrt{8\nu}/\kappa$, which is the distance at which the correlation between points roughly becomes 0.1.

Deviance information criterion and marginal likelihood

Let $\mathcal{M}$ denote a statistical model of interest. The deviance information criterion of $\mathcal{M}$ is

$$D(\bar{\boldsymbol{\theta}}) + 2p_D,$$

where $D(\theta) = -2\log(p(\boldsymbol{y} \mid \theta))$ is the deviance, $\bar{\boldsymbol{\theta}}$ the posterior expectation of $\mathcal{M}$, and p_D the effective number of parameters [57].

The marginal likelihood is

$$p(\boldsymbol{y} \mid \mathcal{M}) = \int p(\boldsymbol{y} \mid \mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{M}) d\boldsymbol{\theta}.$$

Bibliography

[1] Peter J. Hotez, Maria Elena Bottazzi, Carlos Franco-Paredes, Steven K. Ault, and Mirta Roses Periago. The neglected tropical diseases of Latin America and the Caribbean: A review of disease burden and distribution and a roadmap for control and elimination. *PLoS neglected tropical diseases*, 2(9):e300, September 2008.

[2] César P Bouillon. The millennium development goals in Latin America and the Caribbean: progress, priorities and IDB support for their implementation. Available at SSRN: https://ssrn.com/abstract=1543858, 2005.

[3] Carlos Franco-Paredes, Anna Von, Alicia Hidron, Alfonso J. Rodríguez-Morales, Ildefonso Tellez, Maribel Barragán, Danielle Jones, Cesar G. Náquira, and Jorge Mendez. Chagas disease: An impediment in achieving the Millennium Development Goals in Latin America. *BMC International Health and Human Rights*, 7:7, August 2007.

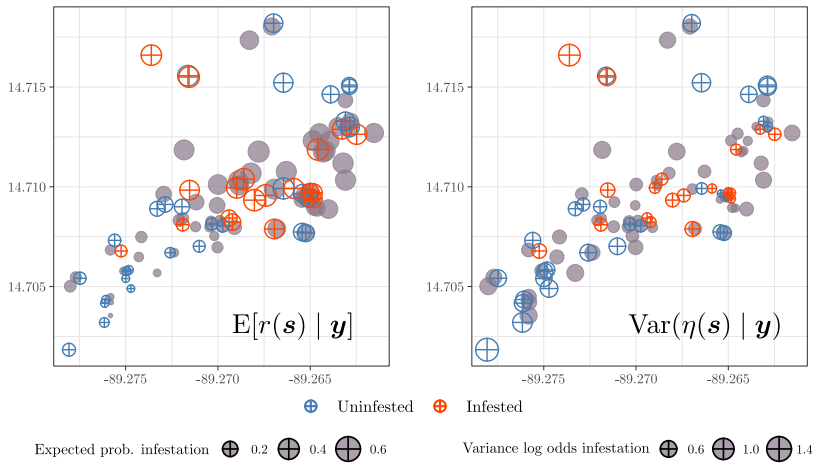

Figure 2.5: Example predicted risk from the full model in La Prensa. Posterior marginal risk (left) and variance of the linear predictor (right) for each house is shown, based on a random sample of 50% of the village. Out-of-sample predictions are shown in gray. The model was fit using the full socioeconomic covariate set.

[4] Dulce Maria Bustamante, Carlota Monroy, Sandy Pineda, Antonieta Rodas, Xochitl Castro, Virgilio Ayala, Javier Quiñónes, Bárbara Moguel, and Ranferi Trampe. Risk factors for intradomiciliary infestation by the Chagas disease vector Triatoma dimidiatain Jutiapa, Guatemala. *Cadernos De Saude Publica*, 25 Suppl 1:S83–92, 2009.

[5] World Health Organization. Chagas disease in Latin America: an epidemiological update based on 2010 estimates. *Weekly Epidemiological Record*, 90(06):33 – 44, 2015.

[6] World Health Organization. Chagas disease (American trypanosomiasis). Available from: https://www.who.int/health-topics/chagas-disease, 2021 [cited 20 September 2021].

[7] Christopher John Schofield. *Triatominae: biology & control*. Eurocommunica Publications, Bognor Regis, 1994.

[8] Yoichi Yamagata and Jun Nakagawa. Control of Chagas disease. *Advances in Parasitology*, 61:129–165, 2006.

[9] Raymond King, Celia Cordon-Rosales, Jonathan Cox, Clive Davies, and Uriel Kitron. Triatoma dimidiata Infestation in Chagas Disease Endemic Regions of Guatemala: Comparison of Random and Targeted Cross-Sectional Surveys. *PLoS Neglected Tropical Diseases*, 5(4):e1035, April 2011.

[10] Hirotsugu Aiga, Emi Sasagawa, Ken Hashimoto, Jiro Nakamura, Concepción Zúniga, José Chévez, Hector Hernández, Jun Nakagawa, and Yuichiro Tabaru. Chagas Disease : Assessing the Existence of a Threshold for Bug Infestation

Rate. *American Journal of Tropical Medicine and Hygiene*, 86(6):972–979, 2012.

[11] Jennifer K. Peterson, Kota Yoshioka, Ken Hashimoto, Angela Caranci, Nicole Gottdenker, Carlota Monroy, Azael Saldaña, Stanley Rodriguez, Patricia Dorn, and Concepción Zúniga. Chagas Disease Epidemiology in Central America: An Update. *Current Tropical Medicine Reports*, 6(2):92–105, 2019.

[12] Maria Carlota Monroy, Dulce Maria Bustamante, Antonieta Guadalupe Rodas, Maria Eunice Enriquez, and Regina Guadalupe Rosales. Habitats, dispersion and invasion of sylvatic Triatoma dimidiata (Hemiptera: Reduviidae: Triatominae) in Petén, Guatemala. *Journal of Medical Entomology*, 40(6):800–806, November 2003.

[13] Eric Dumonteil, Hugo Ruiz-Piña, Eugenia Rodriguez-Félix, Mario Barrera-Pérez, María Jesús Ramirez-Sierra, Jorge E. Rabinovich, and Frédéric Menu. Reinfestation of houses by Triatoma dimidiata after intra-domicile insecticide application in the Yucatán peninsula, Mexico. *Memorias Do Instituto Oswaldo Cruz*, 99(3):253–256, May 2004.

[14] Jennifer Manne, Jun Nakagawa, Yoichi Yamagata, Alexander Goehler, John S. Brownstein, and Marcia C. Castro. Triatomine infestation in Guatemala: Spatial assessment after two rounds of vector control. *American Journal of Tropical Medicine and Hygiene*, 86(3):446–454, 2012.

[15] Sara Helms Cahan, Lucia Orantes, Kimberly Wallin, John Hanley, Donna M. Rizzo, Lori Stevens, Patricia Dorn, Antonieta Rodas, and Carlota Monroy. Residual survival and local dispersal drive reinfestation by Triatoma dimidiata following insecticide application in Guatemala. *Infection, Genetics and Evolution*, 74:104000, October 2019.

[16] Kota Yoshioka, Ezequiel Provedor, and Jennifer Manne-Goehler. The resilience of triatoma dimidiata: An analysis of reinfestation in the nicaraguan chagas disease vector control program (2010–2016). *PLoS ONE*, 13(8):1–18, 2018.

[17] Andrés Gómez-Palacio, Sair Arboleda, Eric Dumonteil, and A. Townsend Peterson. Ecological niche and geographic distribution of the Chagas disease vector, Triatoma dimidiata (Reduviidae: Triatominae): Evidence for niche differentiation among cryptic species. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 36:15–22, December 2015.

[18] Raquel Asunción Lima-Cordón, Lori Stevens, Elizabeth Solórzano Ortíz, Gabriela Anaité Rodas, Salvador Castellanos, Antonieta Rodas, Vianney Abrego, Concepción Zúniga Valeriano, and María Carlota Monroy. Implementation science: Epidemiology and feeding profiles of the Chagas vector Triatoma dimidiata prior to Ecohealth intervention for three locations in Central America. *PLOS Neglected Tropical Diseases*, 12(11):e0006952, November 2018.

[19] Silvia Justi, Sara Cahan, Lori Stevens, Carlota Monroy, Raquel Lima-Cordón,

and Patricia Dorn. Vectors of diversity: Genome wide diversity across the geographic range of the Chagas disease vector Triatoma dimidiata sensu lato (Hemiptera: Reduviidae). *Molecular Phylogenetics and Evolution*, 120(December 2017):144–150, 2018.

[20] Lucila Traverso, Andrés Lavore, Ivana Sierra, Victorio Palacio, Jesús Martinez-Barnetche, José Manuel Latorre-Estivalis, Gaston Mougabure-Cueto, Flavio Francini, Marcelo G Lorenzo, Mario Henry Rodríguez, et al. Comparative and functional triatomine genomics reveals reductions and expansions in insecticide resistance-related gene families. *PLoS Neglected Tropical Diseases*, 11(2):e0005313, 2017.

[21] Dulce Bustamante, Marianela Menes Hernández, Nuria Torres, Concepción Zúniga, Wilfredo Sosa, Vianney De Abrego, and Carlota Monroy. Information to act: Household characteristics are predictors of domestic infestation with the Chagas vector Triatoma dimidiata in central America. *American Journal of Tropical Medicine and Hygiene*, 93(1):97–107, 2015.

[22] Carlota Monroy, Dulce Maria Bustamante, Sandy Pineda, Antonieta Rodas, Xochitl Castro, Virgilio Ayala, Javier Quiñónes, and Bárbara Moguel. House improvements and community participation in the control of Triatoma dimidiata re-infestation in Jutiapa, Guatemala. *Cadernos De Saude Publica*, 25 Suppl 1:S168–178, 2009.

[23] Dulce Bustamante, Sandra De Urioste-Stone, José Juárez, and Pamela Pennington. Ecological, social and biological risk factors for continued Trypanosoma cruzi transmission by Triatoma dimidiata in Guatemala. *PLoS ONE*, 9(8), 2014.

[24] Lori Stevens, Carlota Monroy, Antonieta Guadalupe Rodas, and Patricia Dorn. Hunting, Swimming, and Worshiping: Human Cultural Practices Illuminate the Blood Meal Sources of Cave Dwelling Chagas Vectors (Triatoma dimidiata) in Guatemala and Belize. *PLoS Neglected Tropical Diseases*, 8(9):e3047, 2014.

[25] Jennifer Peterson, Ken Hashimoto, Kota Yoshioka, Patricia Dorn, Nicole Gottdenker, Angela Caranci, Lori Stevens, Concepcion Zuniga, Azael Saldaña, Stanley Rodriguez, and Carlota Monroy. Chagas Disease in Central America: Recent Findings and Current Challenges in Vector Ecology and Control. *Current Tropical Medicine Reports*, 6(2):76–91, June 2019.

[26] Mark Booth and Archie Clements. Neglected Tropical Disease Control – The Case for Adaptive, Location-specific Solutions. *Trends in Parasitology*, 34(4):272–282, 2018.

[27] Guatemala Ministerio de Salud. Sistema de información gerencial de salud (SIGSA) - enfermedades transmitidas por vectores, años 2012 al 2019. Available from: https://sigsa.mspas.gob.gt/datos-de-salud/morbilidad/enfermedades-transmitidas-por-vectores, 2019 [cited 12 July 2021].

[28] David Lucero, Leslie Morrissey, Donna Rizzo, Antonieta Rodas, Roberto Gar-

nica, Lori Stevens, Dulce Bustamante, and Carlota Monroy. Ecohealth interventions limit triatomine reinfestation following insecticide spraying in La Brea, Guatemala. *American Journal of Tropical Medicine and Hygiene*, 88(4):630–637, 2013.

[29] Peter Diggle, Raquel Menezes, and Ting li Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, 59(2):191–232, 2010.

[30] John Hanley, Donna Rizzo, Lori Stevens, Sara Helms Cahan, Patricia Dorn, Leslie Morrissey, Antonieta Guadalupe Rodas, Lucia Orantes, and Carlota Monroy. Novel Evolutionary Algorithm Identifies Interactions Driving Infestation of Triatoma dimidiata, a Chagas Disease Vector. *The American Journal of Tropical Medicine and Hygiene*, 103(2):735–744, 2020.

[31] E. N. I. Weeks, C. Cordón-Rosales, C. Davies, S. Gezan, M. Yeo, and M. M. Cameron. Risk factors for domestic infestation by the Chagas disease vector, Triatoma dimidiata in Chiquimula, Guatemala. *Bulletin of Entomological Research*, 103(6):634–643, December 2013.

[32] Peter Diggle, Benjamin Amoah, Claudio Fronterre, Emanuele Giorgi, and Olatunji Johnson. Rethinking neglected tropical disease prevalence survey design and analysis: A geospatial paradigm. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 115(3):208–210, 2021.

[33] Michael Chipeta, Dianne Terlouw, Kamija Phiri, and Peter Diggle. Adaptive geostatistical design and analysis for prevalence surveys. *Spatial Statistics*, 15:70–84, 2016.

[34] Alinune N. Kabaghe, Michael G. Chipeta, Robert S. McCann, Kamija S. Phiri, Michèle Van Vugt, Willem Takken, Peter Diggle, and Anja D. Terlouw. Adaptive geostatistical sampling enables efficient identification of malaria hotspots in repeated cross-sectional surveys in rural Malawi. *PLoS ONE*, 12(2):1–14, 2017.

[35] Ricardo Andrade-Pacheco, Francois Rerolle, Jean Lemoine, Leda Hernandez, Aboulaye Meïté, Lazarus Juziwelo, Aurélien F. Bibaut, Mark J. van der Laan, Benjamin F. Arnold, and Hugh J.W. Sturrock. Finding hotspots: Development of an adaptive spatial sampling approach. *Scientific Reports*, 10(1):1–12, 2020.

[36] Michael Chipeta, Dianne Terlouw, Kamija Phiri, and Peter Diggle. Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics*, 28(1):1–11, 2017.

[37] Claudio Fronterre, Benjamin Amoah, Emanuele Giorgi, Michelle C Stanton, and Peter J Diggle. Design and Analysis of Elimination Surveys for Neglected Tropical Diseases. *The Journal of Infectious Diseases*, 221(Supplement_5):S554–S560, June 2020.

[38] Alan Gelfand, Peter Diggle, Montserrat Fuentes, and Peter Guttorp. *Handbook of Spatial Statistics*. Taylor \& Francis, Boca Raton, 2010.

[39] Peter Diggle, Jonathan Tawn, and Rana Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

[40] B. Matérn. Spatial variation. phd thesis, Stockholm university, 1960.

[41] Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Constructing Priors that Penalize the Complexity of Gaussian Random Fields. *Journal of the American Statistical Association*, 114(525):445–452, January 2019.

[42] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[43] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

[44] Finn Lindgren and Håvard Rue. Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, 63:1–25, February 2015.

[45] B. K. M. Case. Plos ntd publication materials, April 2022.

[46] World Health Organization. Ending the neglect to attain the sustainable development goals: a sustainability framework for action against neglected tropical diseases 2021-2030. *Geneva: World Health Organization*, Licence: CC BY-NC-SA 3.0 IGO, 2021.

[47] Joanne Cable, Iain Barber, Brian Boag, Amy R. Ellison, Eric R. Morgan, Kris Murray, Emily L. Pascoe, Steven M. Sait, Anthony J. Wilson, and Mark Booth. Global change, parasite transmission and disease control: Lessons from ecology. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 372(1719):20160088, May 2017.

[48] Carlota Monroy, Mildred Mejia, Antonieta Rodas, Regina Rosales, Masahiro Horio, and Yuichiro Tabaru. Cornparison of indoor searches with whole house demolition collections of the vectors of Chagas disease and their indoor distribution. *Medical Entomology and Zoology*, 49(3):195–200, 1998.

[49] Patricia L. Dorn, Sergio Melgar, Vanessa Rouzier, Astrid Gutierrez, Crescent Combe, Regina Rosales, Antonieta Rodas, Sarah Kott, Debra Salvia, and Carlota M. Monroy. The Chagas vector, Triatoma dimidiata (Hemiptera: Reduviidae), is panmictic within and among adjacent villages in Guatemala. *Journal of Medical Entomology*, 40(4):436–440, July 2003.

[50] Lori Stevens, Carlota Monroy, Antonieta Guadalupe Rodas, Robin Hicks, David Lucero, Leslie Lyons, and Patricia Dorn. Migration and gene flow among domestic populations of the chagas insect vector Triatoma dimidiata (hemiptera: Reduviidae) detected by microsatellite loci. *Journal of Medical Entomology*, 52(3):419–428, 2015.

[51] Dulce Maria Bustamante, Maria Carlota Monroy, Antonieta Guadalupe Rodas,

Jaime Abraham Juarez, and John B. Malone. Environmental determinants of the distribution of Chagas disease vectors in south-eastern Guatemala. *Geospatial Health*, 1(2):199–211, May 2007.

[52] Maria Jesus Ramirez-Sierra, Melba Herrera-Aguilar, Sébastien Gourbière, and Eric Dumonteil. Patterns of house infestation dynamics by non-domiciliated Triatoma dimidiata reveal a spatial gradient of infestation in rural villages and potential insect manipulation by Trypanosoma cruzi. *Tropical Medicine and International Health*, 15(1):77–86, 2010.

[53] Daniel Penados, José Pineda, Michelle Catalan, Miguel Avila, Lori Stevens, Emmanuel Agreda, and Carlota Monroy. Infestation dynamics of triatoma dimidiata in highly deforested tropical dry forest regions of Guatemala. *Memorias do Instituto Oswaldo Cruz*, 115(10):1–8, 2020.

[54] Hugh J. W. Sturrock, Pete W. Gething, Ruth A. Ashton, Jan H. Kolaczinski, Narcis B. Kabatereine, and Simon Brooker. Planning schistosomiasis control: Investigation of alternative sampling strategies for Schistosoma mansoni to target mass drug administration of praziquantel in East Africa. *International Health*, 3(3):165–175, September 2011.

[55] Sudipto Banerjee, Bradley P Carlin, and Alan Gelfand. *Hierarchical modeling and analysis for spatial data*. Boca Raton: CRC Press, 2 edition, 2014.

[56] Benjamin F. Arnold, Heather M. Scobie, Jeffrey W. Priest, and Patrick J. Lammie. Integrated Serologic Surveillance of Population Immunity and Disease Transmission. *Emerging Infectious Diseases*, 24(7):1188–1194, July 2018.

[57] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

Chapter 3

Adapting vector surveillance using Bayesian Experimental Design: an application to an ongoing tick monitoring program in the southeastern United States

Abstract

Objectives: To demonstrate the use of Bayesian Experimental Design (BED) in planning spatiotemporal surveillance of disease vectors, for maximizing information regarding environmental covariates and minimizing uncertainty in high-risk areas. We illustrate these principles using an ongoing tick surveillance study in South Carolina parks.

Methods: We implemented a BED workflow based on spatiotemporal models of tick presence. Following a model comparison study based on two years of initial data, several techniques for finding optimal future survey times and locations were compared

to random sampling.

Results: Two optimization algorithms found surveys better than all replications of random sampling, while a space-filling heuristic performed favorably as well. Further, optimal surveys of just 20 visits were more effective than repeating the schedule of 111 visits used in 2021.

Conclusions: BED shows promise as a flexible and rigorous means of survey design for vector control. Identifying sampling schedules with high expected utility can alleviate pressure on local agencies by limiting resources necessary for accurate information on arthropod distributions. For tick surveillance in South Carolina, optimal scheduling can improve critical public health information with just a handful of collection visits.

3.1 Introduction

Tickborne diseases now make up more than 75% of reported vector-borne infections in the United States [1]. This sharp increase in the last decade is likely due to an increased awareness in the public health importance of monitoring tickborne pathogens, as well as the continued geographic expansion of several medically important tick species [2, 3]. An accurate understanding of the spatial and temporal distribution of medically important ticks is a crucial first step to informing when and where people are at risk, and forms the basis of public health programs for the diagnosis and prevention of tickborne diseases [4]. However, maps of tick distributions throughout the US are lacking in spatial and temporal resolution, and often depend on outdated sources, disparate sampling techniques, or otherwise biased data [5, 6]. Statistical models are therefore important tools for explaining factors associated with tick presence and filling gaps in existing distribution maps. However, the reliability of model

predictions are critically dependent on the amount and quality of input data at an appropriate spatiotemporal scale [7]. An area with particularly limited knowledge of current tick distributions is the southeastern US, where local resources for monitoring and control are scarce and less than 10% of vector control agencies do tick surveillance of any kind [8, 9].

In addition to learning from existing data, a further use of tick distribution models is informing future surveillance and control efforts by anticipating the value of future sampling locations. For example, more fine-grained sampling might follow an initial surveillance effort focused on a subset of areas of potentially high risk [10]. Though usually such sampling decisions are made in an ad hoc manor, a decision-theoretic alternative is to assign a score to potential sampling times and locations and attempt to find visits maximizing some objective function [11]. Deciding future visits for surveillance or control efforts may then be framed as an optimal experimental design problem, where a calendar of sampling times and/or locations is chosen to maximize their information content or increase the impact of vector control measures [12, 13]. Because observational studies of tick distributions involve a complex web of interactions between variable environmental conditions, population dynamics, and uncertain measurement, it is important to employ design criteria which are robust to uncertainty in model parameters and experimental conditions [14].

Thanks to computational advances in recent decades, Bayesian inference has become popular for model fitting in ecology and epidemiology [15]. The advantages of Bayesian inference include a complete treatment of model uncertainty through providing posterior distributions of plausible model parameters and predicted outcomes, and the ability to incorporate entomological or medical knowledge through prior be-

liefs. For experimental design, the Bayesian paradigm provides a framework that is robust to system uncertainty and flexible for tailoring novel design criteria to specific challenges in public health [16, 17].

Implementing Bayesian Experimental Design (BED) involves three general steps (Figure 3.1). First, a statistical model is chosen based on its suitability to explain any existing survey data, resulting in a posterior distribution over model parameters given initial input. Next, a utility function is developed to score potential future survey schedules based on the quality of new information. Finally, the space of possible designs is searched for highly informative survey schedules using optimization. While an initial dataset is not required for BED, we frame the method for the situation where preliminary data are available from previous surveillance efforts. This information can help inform an appropriate choice of model, and more effectively differentiate designs of high quality [18]. The experimental goal is then to design a future survey which provides additional value beyond the initial dataset.

In this work, we outline principles for how BED can be incorporated in spatiotemporal surveys to maximize the value of vector surveillance and control efforts, and illustrate their use for an ongoing tick surveillance effort of South Carolina state parks and other public lands. We compare the ability of different search techniques to find survey schedules which maximize utility based on two design criteria tailored to different priorities of vector surveillance. In addition to informing future data collection efforts, we demonstrate how high utility designs can be further analyzed to provide novel insight into sources of uncertainty in tick distributions.

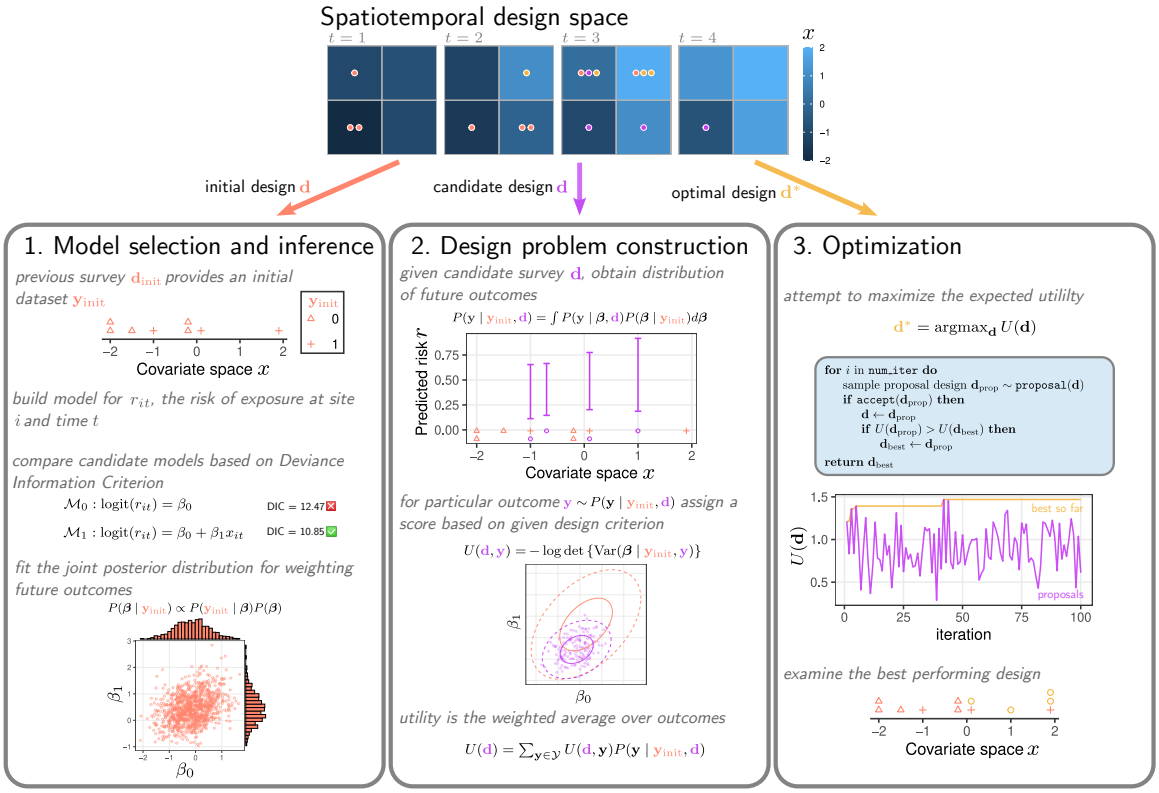

Figure 3.1: Implementing Bayesian Experimental Design in spatiotemporal surveillance. A motivating example with a single environmental covariate is shown, with the goal of establishing environmental factors associated with tick presence. Top: a small design space consisting of four possible survey locations (e.g. parks) and four timepoints (e.g. months). A surveillance schedule consists of arranging points in design space. The changing values of the environmental covariate x are shown for each survey point, and the values of x corresponding to a design $\mathbf{d}$ is mapped to the 1-dimensional covariate space, indicated by arrows for each $\mathbf{d}$. Bottom: in step 1, a response $\mathbf{y}_{\text{init}}$ and associated x values from an initial survey $\mathbf{y}_{\text{init}}$ are shown in covariate space, two candidate models are compared, and a posterior distribution for the selected model is fit. In step 2, the utility of some candidate design is defined as an average over future outcomes and the amount of new information which would be provided by each outcome. Here a Bayesian D-optimality criterion is used, which scores outcomes based on the volume of confidence ellipsoids approximating the updated posterior distribution. In step 3, finding an effective design is treated as an optimization problem over the space of candidate designs. A generic stepwise procedure is shown, and the best design found after 100 iterations is then examined. In accordance with BED theory, this design spreads additional points throughout the middle of covariate space, while putting special attention at the extreme $x \approx 2$ which was under sampled in the initial survey.

3.2 Methods

3.2.1 Data collection and preparation

Data were from a larger tick surveillance project of South Carolina city and state parks, beginning in March 2020 to present day. The project also included submissions from South Carolina animal shelters and citizen scientists, though these data were not used in this study. Here we use data from 2020 and 2021 state and city parks, with observations spanning 26 counties and 10 calendar months. A scientific collecting permit from the SC Department of Parks, Recreation, & Tourism was secured for both years, and written permission was granted from the appropriate municipal government for city parks. The coordinates for each site were selected near the entrance to each park in a forested area for consistency. Tick collections were performed following recommended CDC Ixodidae guidelines [19]. In brief, tick traps consisting of a 0.61m2 muslin cloth baited with 1.5lb dry ice each were placed in parks along hiking and nature trails and left in the park for 1.5-2 hours. Additionally, tick drags were performed along different hiking and nature trails. Tick drags were constructed with a 1.22m x 1.52m white duck canvas attached to a 1.22m wooden dowel, with zinc washers as weights on the bottom. Each collection at a park consisted of ten tick traps and a total of 30 minutes of tick dragging to ensure that the recommended surface area for host-seeking ticks was surveyed.

Ticks were brought to the Laboratory of Vector-Borne and Zoonotic Diseases at the University of South Carolina for processing, where they were identified to species, sex, and life stage. Morphological identifications were conducted with multiple dichotomous keys. The response outcome for each visit was then recorded as a binary

variable indicating presence/absence of nymphs for each species. *A. maculatum*, *I. brunneus*, and *I. affinis* were found in only 3, 5, and 12 visits, respectively. To improve the quality of model predictions, *A. maculatum* was therefore removed from all further analyses, and the Ixodes species were aggregated into a single *Ixodes spp.* group. In the original collections data, there were 59 distinct visits in 2020, and 111 visits in 2021. *A. americanum* was found in 45% of these 170 visits, *Ixodes spp.* was found in 24%, and *D. variabilis* in 8%.

3.2.2 Environmental risk factors

Several meteorological and geographic variables were selected as potential covariates of tick occurrence based on tick ecology and previous modeling studies [20, 21]. Land cover and forest canopy data were obtained from the USGS 2019 National Land Cover Database [22], while elevation and meteorological variables were obtained from PRISM [23]. Included meteorological variables were monthly total precipitation, monthly average of the daily maximum temperature, monthly average of the daily minimum temperature, average daily minimum temperature in January, and monthly average of the daily average humidity, calculated from average daily temperature and average dew point temperature [24]. Since we are interested in making predictions for future years, 30-year meteorological monthly normals were used, defined as the value for each month on average over the last 30 years. The continuous covariates were centered and scaled prior to all statistical analyses, and the monthly minimum temperature was removed due to multicollinearity.

3.2.3 Modeling tick distributions using Bayesian regression

Further information on the model specification and experimental design procedure is given in the Supplemental Methods (Section 3.6). We used a hierarchical, mixed-effects framework, where correlations between the observed tick distributions and various environmental, spatial, and temporal effects are captured [15]. The probability of encountering a tick of species j, in a visit to site i during month t, is therefore a function of the 7 environmental covariates, the survey location, and the month the visit took place.

To find a model parsimonious with the initial collections data, 28 candidates were constructed by simplifying different components of the full model. Each of the environmental, spatial, and temporal components were considered either shared or different between tick groups, and linear or spline-based functions were considered for the environmental effects. Models were compared based on the Deviance Information Criterion, which measures a model's goodness-of-fit to the data and robustness, while penalizing model complexity. The best performing model was used in all subsequent analyses.

3.2.4 Experimental designs for vector surveillance

Because of the limited capacity of vector control agencies and researchers, a reasonable surveillance strategy should consider the feasibility and convenience of sampling sites while allowing sufficient diversity to realistically be able to extrapolate to the region of interest. To maintain this balance, we restricted future sampling to a set of 57

sites on public land across South Carolina, including all 47 South Carolina state parks and historic sites, 6 locations within national parks and wildlife refuges, and 4 other natural areas present in the initial survey data. Collection visits were delineated monthly and could take place in any month, resulting in a space of 684 possible visits which may be added to a candidate survey.

In BED, potential outcomes resulting from a proposed design $\mathbf{d}$ are assigned a utility, or a score based on the quality of new information provided by that outcome. The utility of the design is then averaged over the posterior predictive distribution of possible outcomes (Figure 3.1, step 2). To score potential outcomes, two design criteria were considered. First was a form of Bayesian D-optimality, which quantifies uncertainty in the posterior covariance matrix for the environmental effects [25]. A second criterion was then designed to improve the reliability of prediction maps in regions where risk of exposure is highest. Here we assigned utility based on the maximum reduction in standard deviation of risk from the initial dataset among high-risk visits, where high-risk visits were defined as any point in the map with expected risk greater than 0.75 for at least one tick species. The maximum rather than mean reduction in uncertainty was chosen for greater variety compared to the first criterion, and represents an "all in" approach where a small number of visits are chosen to target uncertainty at a specific part of the prediction map.

Equipped with a predictive model and utility function, the space of possible designs can be searched to optimize utility (Figure 3.1, step 3). Because Bayesian design criteria are available in closed form only in the simplest cases, mathematical formulae are not available and numerical methods are typically used. This design space is far too vast to test the utility of most designs, therefore specialized search techniques

are needed to find designs of high quality [16]. We implemented two optimization algorithms – first was a Simulated Annealing algorithm, a popular stochastic search technique which incrementally adds and removes visits to a design while avoiding local minima, and second was an Exchange algorithm which rotates through neighboring months and sites until no neighbors improve utility [12]. We also considered two heuristics to choose designs based on characteristics suspected to lead to high utility, while requiring fewer computational resources than optimization. The first chose visits with the highest predicted variance given the existing collections data, and the second was a space-filling heuristic which spread visits evenly over time and prohibited visiting any two parks less than 25km away [26]. We then compared these four strategies to random sampling for an increasing sample size from 5 to 20 visits, as well as to several convenience sampling schedules which were constructed by repeating aspects of the initial collections data.

3.3 Results

In the model comparison study, the model with lowest DIC included linear environmental effects shared between species, and both spatial and temporal effects separate for each species, although several alternative models performed nearly as well (Figure 3.5). The posterior environmental and temporal effects from the best performing model are summarized in Figure 3.2. Daily maximum temperature and January minimum temperature both had a strong positive association with tick presence, while precipitation had a strong negative association, as coefficients for these variables had over 95% of the posterior mass above/below zero. Relative humidity, elevation, and a Mixed Forest land cover also appear negatively associated with risk. The coefficients

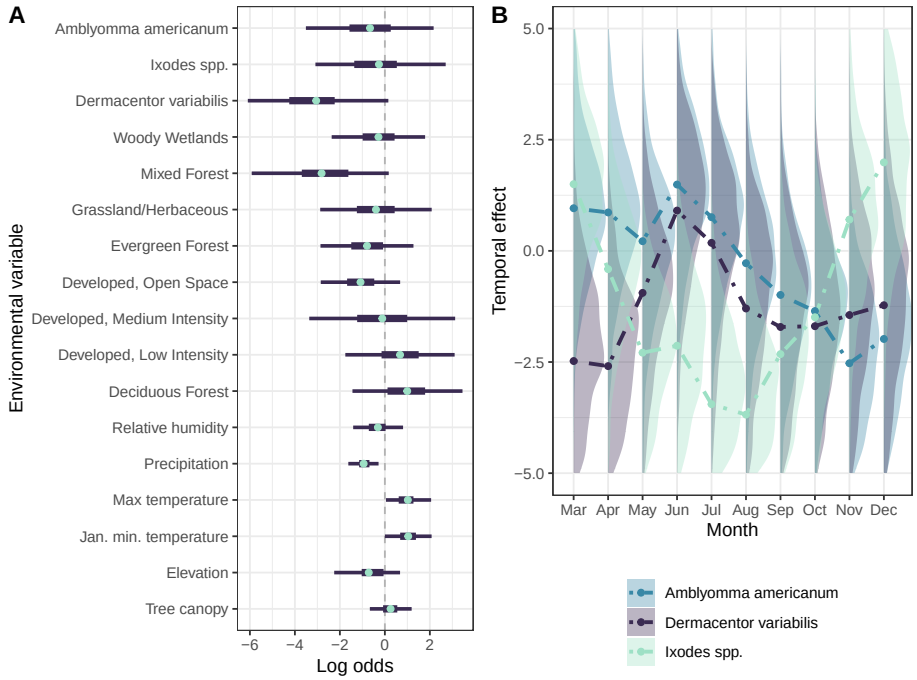

Figure 3.2: Posterior environmental and temporal effects from the initial survey. Results are shown for the best-performing model fit to the initial survey data from 2020-2021. All results are in log-odds scale. (A) Marginal posterior mean for each coefficient of the environmental variables are shown as points, while 50% and 95% Highest Posterior Density Intervals are shown as purple bars. (B) Mean temporal trend for each tick group is given by dashed lines, along with full marginal posterior densities for each month/tick group.

for land cover had particularly high variance, likely due to less data being available for any single land cover class. Spatiotemporal prediction maps throughout South Carolina for each tick group are shown in Figure 3.3. The average predicted risk for both *A. americanum* and *Ixodes spp.* was high in the southeast of the state, though in different months, while average risk was consistently low for all species in the northwest. Figure 3.3 also shows there remains considerable model uncertainty in tick presence throughout much of the state, illustrating the importance of continued surveillance.

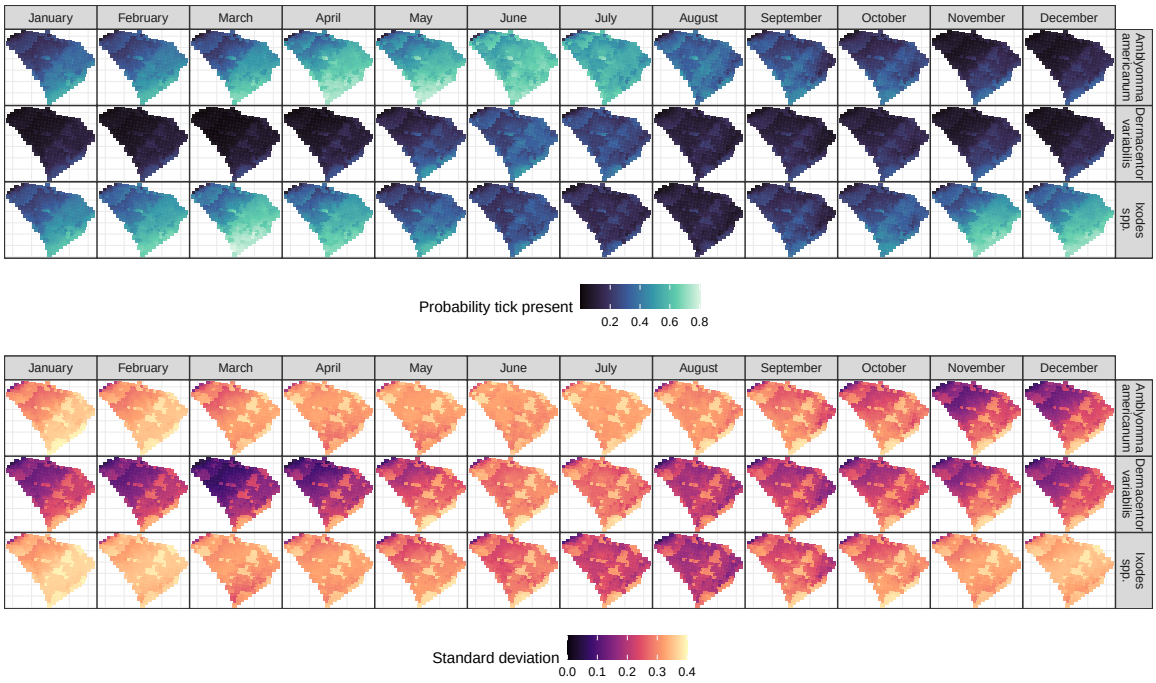

Figure 3.3: Spatiotemporal mean and standard deviation of risk. Results are shown for the best-performing model fit to the initial survey data. Posterior marginals for the probability of tick presence were computed along a 16km grid of locations across South Carolina, and summarized by the mean (top) and standard deviation (bottom).

Figure 3.4 summarizes the results from the simulation study, with utilities of the designs produced by each search method compared to random sampling. For the D-optimality criterion, both of the optimization algorithms and the space-filling heuristic were able to find designs better than any random sample, and had very similar performance for all sample sizes, although the space-filling strategy was less computationally expensive. The variance heuristic performed comparably or worse than random. The example designs based on the initial schedule also had much lower utility relative to their larger sampling budget – revisiting all 30 previously visited sites in June is worse than a schedule of 10 visits found using Exchange, and repeating all 111 visits from 2021 had lower utility than designs of 20 visits. Results were similar

for the second criterion of maximum variance reduction in high-risk visits, although space-filling was less effective than Exchange and Simulated Annealing. For this criterion, an optimal design of 20 locations was found with expected utility of 0.1, which amounts to a 30-50% reduction in uncertainty for a particular high-risk visit.

To obtain a sense of what makes certain surveys better than others, the environmental conditions of each possible visit were then projected using Factor Analysis for Mixed Data (FAMD), a dimension reduction technique for continuous and categorical data [27]. The associated covariates are embedded within the two most important dimensions in Figure 3.4C and D, and the third and forth dimensions in Figure 3.6. For the first criterion, the three successful strategies chose visits which spread covariates throughout the center of the two most important dimensions, while the Variance heuristic placed points on the edges and far away from the observed data. For the second criterion the successful designs instead placed many points in the bottom-right quadrant of space, showing there are particular survey conditions which were critical for maximizing information regarding certain high-risk sites.

3.4 Discussion

Accurate information regarding the time and place of probable tick encounters is an essential first step to reducing the burden of tick-borne pathogens. Statistical modeling allows extrapolating available information to a wider scale, which in turn enables local vector control agencies to better direct critical resources. However, the reliability of such model predictions are critically dependent on the nature of available data. Combining a Bayesian workflow and design of experiments is a principled approach to getting more out of data from existing surveillance efforts, and directing

future efforts for the greatest effect. Thanks to advances in software and computing throughout the last decade, optimal Bayesian survey design is feasible to implement for a diverse array of researchers throughout epidemiology.

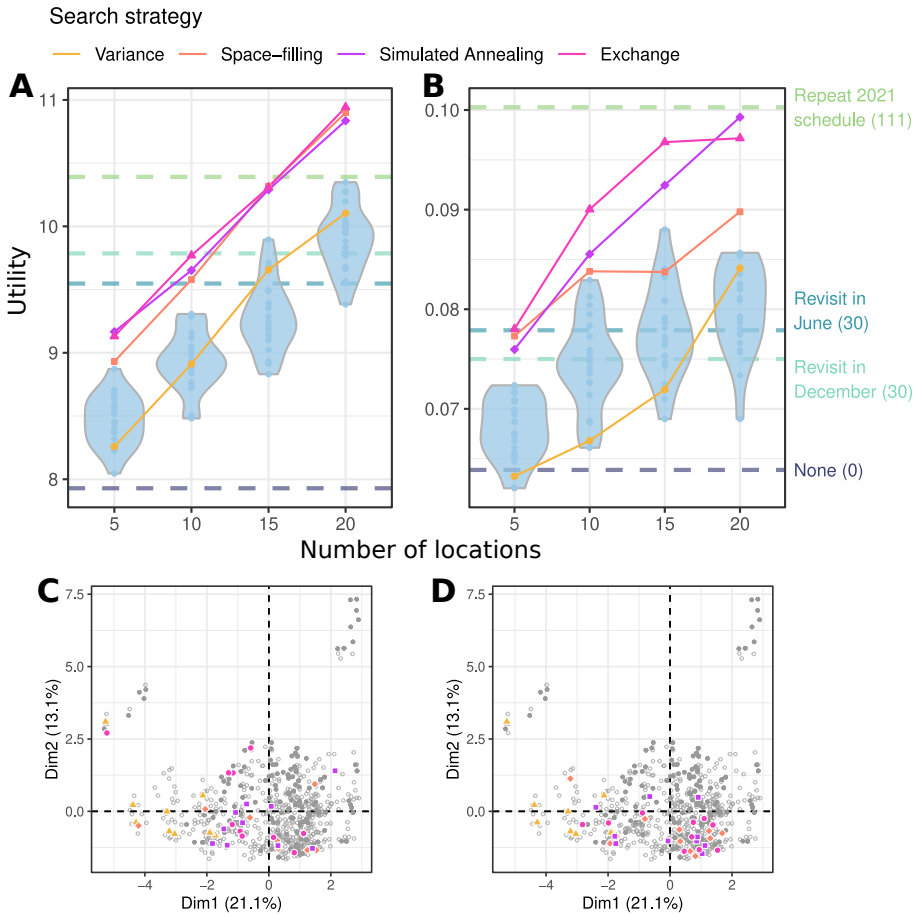

Figure 3.4: Comparing search methods for effective designs of tick surveillance. The utility of designs found using different search strategies are compared to 20 replications of random sampling (light blue distributions) for different sample sizes, and to different convenience sampling examples (dashed lines). (A) Results from optimizing the D-optimality criterion for the environmental covariates. (B) Results using the maximum variance reduction criterion among high-risk visits. (C-D) The full covariate space from all possible visits is projected to two dimensions using Factor Analysis for Mixed Data (FAMD) [27]. Filled gray points are visits which occurred in the initial data, while colored points are designs of size 10 found with each search method.

Our results for the application of scheduling monthly tick surveillance in public natural areas demonstrate large gains in information is possible through carefully chosen surveys. Even when restricting sampling to a limited number of locations, efficient survey design can make the difference for learning critical information and improving the reliability of tick distribution maps. Successful designs can also inform general practices for surveillance. For example, for the first design criterion based on environmental covariates, the success of a space-filling strategy shows that spreading future visits across time and space is more valuable than other intuitive options such as focusing sampling on specific months which were previously under-sampled. Because similar environmental conditions will tend to be clustered in time and space, spreading visits in this way will tend to spread design points across covariate space as well, which is generally D-optimal for simple logistic regression according to BED theory [28].

Analysis of the initial survey data during model comparison provides insight into the current tick patterns in natural areas throughout the southeastern US, while also demonstrating further data is needed. The top performing models all included a temporal trend for each tick species, and the posterior marginals for each trend show strong seasonal patterns (Figures 3.2B and 3.5). These residual trends could indicate contributions from variables not included in the analysis such as isothermality [20], or from seasonality in ecological factors such as host availability. Another explanation for this temporal trend is our use of 30 year normals data, which ignores climatic differences between the two years in the initial data that could lead to a temporal offset in risk between years.

The best suited models all included a term for spatial variability for each tick

species as well, which has previously been deemed important for modeling *I. scapularis* density [10]. Overall, the large variance in both the spatial and temporal effects among top models suggests uncertainty in tick presence is due to a combination of unmeasured environmental effects, population dynamics, and the variability inherent to all methods of tick collection in a natural environment [29]. It is also interesting that the top performing model did not include species-dependent nor non-linear effects with the environment, as the importance of such non-linear environmental effects are frequently stressed when modeling tick distributions [20, 30]. This is likely due in part to the relatively limited geographic range of our data.

Our prediction map (Figure 3.3) of expected probability of tick presence in South Carolina generally agrees with previously published results, although data at a similar spatial and temporal scale are limited. For *I. scapularis*, county level data show 30% of South Carolina counties had an established tick presence by 2015, most of which were in the southeast [31], while models calibrated to the same data predicted suitability in the center of the state as well [20]. County level establishment of *A. americanum* follows a similar pattern as *I. scapularis* [32], although predictions based on that data indicated all counties were highly suitable [33]. This previous prediction of *A. americanum* in the northwest is in contrast with our findings, as the species was never encountered during initial data collection in the region, and our model predicted low risk there in all months.

The application of BED for vector surveillance used in this work focused on establishing tick presence in public natural areas, although we note that the framework used here can be applied to other metrics such as abundance with minimal changes. While measuring tick presence or abundance in outdoor recreational areas such as

state parks is a widely used method for establishing human exposure risk [34, 35], and for detecting expanding ranges of ticks and tick-borne pathogens [36], the reliability of such data for predicting individualized risk of infection is unclear. For Lyme disease, it has been suggested that private property is the main source of exposure to host-seeking nymphs [37, 38]. The infection status of ticks is also a critical source of information, although the importance of measuring density of infected ticks compared to tick prevalence likely depends on the study area. At the county level, nymph density has been found ineffective for predicting Lyme disease incidence in low incidence areas, but is comparable to density of infected nymphs in high incidence counties [39]. Another limitation with our assumed data collection method is that patterns of tick presence and abundance will vary greatly over the span of a state park. Thus, establishing the distribution of ticks is ultimately just a single step to any comprehensive strategy for vector surveillance and control.

The BED procedure illustrated here suggests several avenues for future statistical and computational development. First, additional work is needed to better understand optimal designs for the types of mixed-effects models used in this and other studies of species distributions, as research combining BED and mixed-effects models is generally scarce [12, 18]. Second, specialized optimization strategies for finding optimal survey schedules should be developed, as spatiotemporal survey design presents distinct challenges such as incomplete control over the environmental conditions available among possible visits. Another possibility is to employ adaptive sampling, where locations are visited in smaller batches and the data collected from each iteration are able to inform sampling in future batches, although updating data sequentially leads to additional logistical constraints during surveillance [13, 40]. A

final avenue for future work is in the choice of design criteria, which may change depending on the specific goals of the analysis. While we simply restricted designs to a certain number of visits, researchers with different goals or resource limitations could employ criteria which account for distance traveled, availability of materials, or density of human traffic at a particular park and month. Additionally, a general procedure for design criteria which adapt to the current needs of local vector control agencies would allow widespread application of experimental design strategies at a fine-grained spatiotemporal scale.

3.5 Conclusions

In this work, we have outlined Bayesian Experimental Design as a formal approach to the surveillance of disease vectors. Compared to classical methods of experimental design, a Bayesian framework provides a natural way to incorporate initial survey data, while rigorously accounting for remaining uncertainty in model predictions. We applied a BED workflow to an ongoing tick surveillance study in South Carolina state parks, and found that surveys optimized to satisfy specific goals were universally more efficient than simple random sampling. These results demonstrate the promise of optimal survey design for researchers and vector control agencies to maximize the impact of the data they collect.

3.6 Supplemental information

3.6.1 Model specification

To model the distribution of different tick species simultaneously, we use a hierarchical framework analogous to a mixed-effects model, where environmental factors operate as "fixed" effects while residual variability within and between sites, months and tick species operate as "random" effects. Let y_{ijt} be a binary variable indicating the presence of a tick of species j during a visit to site i in month t, and r_{ijt} the the corresponding risk of tick encounter. There are K different covariates capturing the environment in the model, and x_{kit} indicates the value of each covariate during a visit. The full model specification is then

$$y_{ijt} \sim \text{Bernoulli}(r_{ijt}) \tag{3.1}$$

$$\text{logit}(r_{ijt}) = \eta_{ijt} = \sum_{k=1}^{K} f_{kj}(x_{kit}) + s_{ij} + m_{jt} \tag{3.2}$$

where for each tick species j, $\boldsymbol{s}_j$ and $\boldsymbol{m}_j$ are hierarchical effects for each visit site and month, and f_{kj} are (potentially nonlinear) functions of the covariates.

For environmental effects, we consider two possible forms for f. First is the linear case where $f_{kj}(x) = \beta_{kj}x$ for all k and j, and β_{kj} have $N(0, 5)$ priors. Second is a Bayesian analog to a spline model, where each $f_{kj}(x)$ is distributed as a random walk of order 1 over x with precision τ_f [41].

We assume site-level effects for each species are independently and identically distributed, so that $s_{ij} \sim N(0, \tau_s^{-1})$ with precision τ_s. For month-level effects, we assume temporal trends for each species are independently and identically distributed

AR(1) variables with marginal precision τ_m and lag correlation ρ [41]. Priors for τ_f, τ_s, and τ_m are set to logGamma$(1, 0.1)$, while ρ is distributed such that logit $\left(\frac{1+\rho}{1-\rho}\right) \sim N(0, 6.67)$.

3.6.2 Model comparison study

To find a model best supported by the existing collections data, we test different variations of the above full model by simplifying different components and testing all combinations. Each of the environmental, spatial, and temporal effects are considered as shared between between species (i.e. removing the j in (3.2)), as well as with the spatial and temporal components removed entirely. Finally, both the linear and spline forms for each f are considered. For example, a model with linear f, spatial effect shared between species, and no temporal effect would be $\eta_{ijt} = \sum_k \beta_{kj} x_{kit} + s_i$. Combining these simplifications results in 28 candidate models, and models are compared using DIC [42]. All models are fit in R version 4.2.2 using R-INLA version 23.02.27 [43].

Results from the model comparison study are shown in Figure 3.5. The top performing model is highlighted, and included linear f shared between species, and spatial and temporal effects for each species. Thus, the model chosen for this work has linear predictor

$$\eta_{ijt} = \sum_k \beta_k x_{kit} + s_{ij} + m_{ij}. \tag{3.3}$$

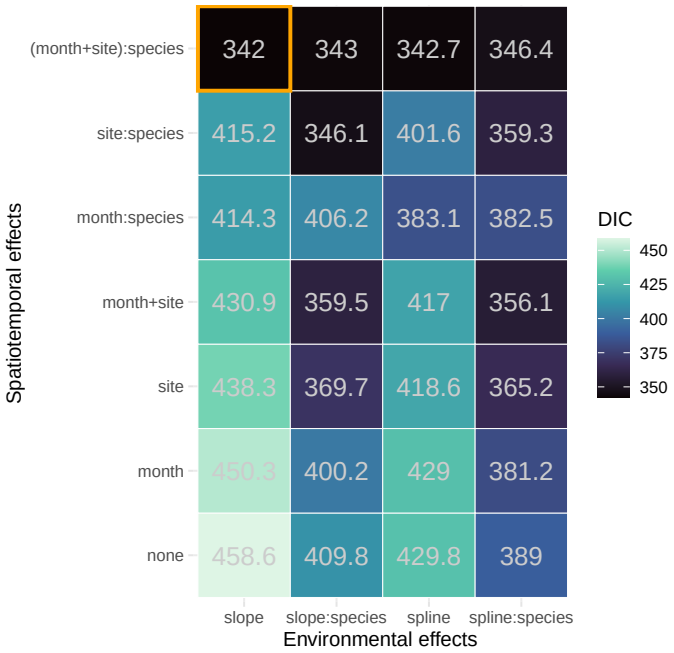

Figure 3.5: Deviance information criterion of different mixed-effects models, fit to the initial survey data. Each tile indicates a model comprised of the corresponding environmental and spatiotemporal effects, shared or independent between tick species. "Slope" indicates a linear f for each environmental variable, and "spline" indicates nonlinear f. The best-ranked model (lowest DIC) is highlighted in orange.

3.6.3 Bayesian Experimental Design

As covered in the main text, implementing BED involves specifying a utility function $U(\mathbf{d}, \mathbf{y})$, where in this work $\mathbf{d} = \{(i_1, t_1), \ldots, (i_m, t_m)\}$ is a spatiotemporal schedule of collection visits and $\mathbf{y}$ is potential future data for each tick species observed from the schedule $\mathbf{d}$. We consider two such functions, which represent the value of new data $\mathbf{y}$ for increasing some form of public health information. First is a form of Bayesian D-optimality,

$$U_1(\mathbf{d}, \mathbf{y}) = -\log\det \mathrm{cov}(\boldsymbol{\beta} \mid \mathbf{y}_{\text{init}}, \mathbf{y}),$$

where $\boldsymbol{\beta}$ are the linear coefficients of the environmental effects *a posteriori* fitted to $\mathbf{y}_{\text{init}}$ and $\mathbf{y}$.

A second criterion was then designed to improve the reliability of prediction maps in regions where risk of exposure is highest. We first extract covariates for a regular 4km grid spanning South Carolina and all 12 months. For each point (i,t) in this set $\mathcal{G}$, we define a subset of high-risk prediction points $\mathcal{H}$ containing (i,t) if $\mathbf{E}\left[r_{ijt} \mid \mathbf{y}_{\text{init}}\right] \geq 0.75$ for at least one species j. Utility is assigned based on the maximum reduction in standard deviation of risk from the initial dataset, among these high-risk points in $\mathcal{H}$,

$$U_2(\mathbf{d},\mathbf{y}) = \max_{(i,t)\in\mathcal{H}} \left\{\sigma(r_{ijt} \mid \mathbf{y}_{\text{init}}) - \sigma(r_{ijt} \mid \mathbf{y}_{\text{init}},\mathbf{y})\right\},$$

where $\sigma(X) = \sqrt{\text{Var}(X)}$.

For a utility function $U(\mathbf{d},\mathbf{y})$, the utility of $\mathbf{d}$ is then averaged over future outcomes. For discrete $\mathbf{y}$,

$$U(\mathbf{d}) = \sum_{\mathbf{y}} U(\mathbf{d},\mathbf{y}) P(\mathbf{y} \mid \mathbf{y}_{\text{init}},\mathbf{d}), \tag{3.4}$$

where

$$P(\mathbf{y} \mid \mathbf{y}_{\text{init}},\mathbf{d}) = \int P(\mathbf{y} \mid \boldsymbol{\theta},\mathbf{d}) P(\boldsymbol{\theta} \mid \mathbf{y}_{\text{init}}) d\boldsymbol{\theta}$$

is the posterior predictive distribution for $\mathbf{y}$ resulting from $\mathbf{d}$.

3.6.4 Description of search algorithms

Once a design criteria is chosen, the goal is to find some $\mathbf{d}$ with as close to optimal as possible. An optimal design for criteria U is defined

$$\mathbf{d}^* = \underset{\mathbf{d}}{\operatorname{argmax}}\, U(\mathbf{d}).$$

In experimental design, optimization over the utility surface U usually presents two broad challenges. First, calculating $U(\mathbf{d})$ is computationally expensive, requiring 10s of seconds or longer for a single evaluation, which limits the number of search iterations that are feasible to budget. Second, evaluations of the utility surface are subject to noise, since the expectation (3.4) must be approximated using Monte Carlo methods and samples from the posterior predictive distribution. Optimization algorithms therefore must be robust to noise, for example by having enough exploratory behaviour to avoid (potentially false) local maxima [44]. A third challenge particular to the surveys we consider is that the design space is discrete, which prohibits the use of gradient-based optimization methods.

With these constraints in mind, we consider 4 search strategies for finding good designs. The first two are optimization algorithms that begin with an initial design of visits chosen uniformly at random, then attempt to incrementally improve the design until $T = 150$ utility evaluations have occurred.

Simulated Annealing: this stepwise strategy proposes new designs by randomly selecting a new visit and randomly removing a current one. If the proposal is accepted, this design becomes the current one and the process repeats. To avoid local optima, new proposals with a lower utility are sometimes still accepted. If $s = 1, \ldots, T$ is the

current iteration, the probability of accepting a worse proposal is

$$p(s, \mathbf{d}_{\text{prop}}, \mathbf{d}) = \exp\left\{(\log_{10} U(\mathbf{d}_{\text{prop}}) - \log_{10} U(\mathbf{d}))/T(s)\right\},$$

where $T(s)$ is a decreasing function of s called the *cooling schedule*.

We use a cooling schedule of $T(s) = T_0 \left(1 - \frac{s-1}{T-1}\right)^{\alpha}$, where T_0 is the *cooling magnitude* and α the *curvature*. We set $T_0 = 0.2$ when optimizing the first criterion U_1, $T_0 = 0.02$ for U_2, and $\alpha = 1.3$ for both.

Exchange: the Exchange strategy attempts to search more systematically than SA by stepping through "nearby" design points until no steps improve utility [45, 46]. If $\mathbf{d}$ is some current design, the algorithm performs the following steps for each visit $(i, t) \in \mathbf{d}$: first the month t is incremented until U does not increase, then t is decremented until U does not increase, and then the 4 neighbor sites closest to i are checked. If none of these moves improve utility for any visit in $\mathbf{d}$, the algorithm terminates and returns $\mathbf{d}$, otherwise, this process continues until T utility evaluations have taken place.

Since this process is susceptible to terminating early in local optima, we consider a single run of the algorithm to be 3 independent replications with a different initial design. The best of these 3 designs is then chosen.

Variance heuristic: this strategy simply chooses visits based on their variance given the initial data, and is thus completely deterministic. Points are assigned a rank ν_{it} equal to the average of $\text{Var}(\eta_{ijt} \mid \mathbf{y}_{\text{init}})$ over species j, and then the m top ranked points are added to $\mathbf{d}$.

Space-filling heuristic: this strategy samples designs randomly, while ensuring visits are spread across time and space. First, the month of each visit is sampled

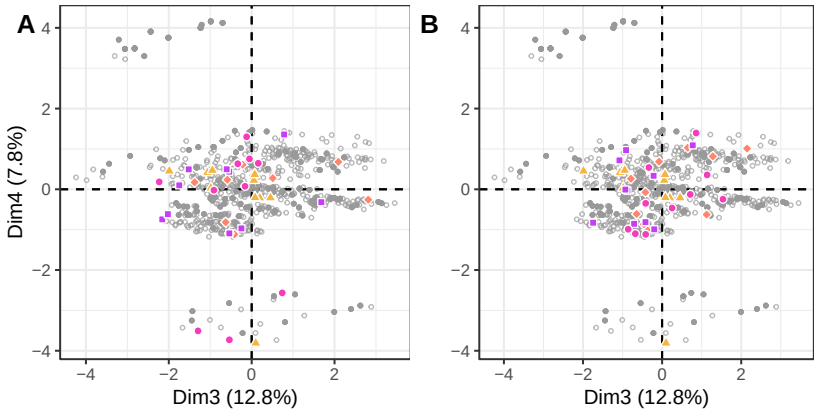

Figure 3.6: Third and fourth dimensions of successful survey designs projected using Factor Analysis for Mixed Data (FAMD).

without replacement, repeating as necessary if the sample size is greater than 12. Then, each site is assigned sequentially by sampling each site randomly, but only accepting sites which are at least 25km away from all sites chosen so far. As a stochastic strategy, 5 such designs are sampled, and the one with highest utility is returned.

In the main text, designs of increasing size are considered in increments of 5. To reduce computation time, the two optimization algorithms build their designs incrementally. Thus, only 5 visits are optimized at a time for these algorithms, and these new visits are added to the previous design when evaluating U.

Bibliography

[1] Ronald Rosenberg, Nicole P. Lindsey, Marc Fischer, Christopher J. Gregory, Alison F. Hinckley, Paul S. Mead, Gabriela Paz-Bailey, Stephen H. Waterman, Naomi A. Drexler, Gilbert J. Kersh, Holley Hooks, Susanna K. Partridge, Susanna N. Visser, Charles B. Beard, and Lyle R. Petersen. Vital Signs: Trends in Reported Vectorborne Disease Cases — United States and Territories, 2004–2016. *Morbidity and Mortality Weekly Report*, 67(17):496–501, May 2018.

[2] Kiersten J. Kugeler, Grace M. Farley, Joseph D. Forrester, and Paul S. Mead.

Geographic Distribution and Expansion of Human Lyme Disease, United States. *Emerging Infectious Diseases*, 21(8):1455–1457, August 2015.

[3] Daniel E. Sonenshine. Range Expansion of Tick Disease Vectors in North America: Implications for Spread of Tick-Borne Disease. *International Journal of Environmental Research and Public Health*, 15(3):478, March 2018.

[4] Rebecca J Eisen and Christopher D Paddock. Tick and Tickborne Pathogen Surveillance as a Public Health Tool in the United States. *Journal of Medical Entomology*, 58(4):1490–1502, July 2021.

[5] Terry L. Schulze, Robert A. Jordan, and Robert W. Hung. Biases Associated with Several Sampling Methods Used to Estimate Abundance of Ixodes scapularis and Amblyomma americanum (Acari: Ixodidae). *Journal of Medical Entomology*, 34(6):615–623, November 1997.

[6] Samantha M. Wisely and Gregory E. Glass. Advancing the Science of Tick and Tick-Borne Disease Surveillance in the United States. *Insects*, 10(10):361, October 2019.

[7] Kiersten J. Kugeler and Rebecca J. Eisen. Challenges in Predicting Lyme Disease Risk. *JAMA Network Open*, 3(3):e200328, March 2020.

[8] Kyndall C. Dye-Braumuller, Jennifer R. Gordon, Danielle Johnson, Josie Morrissey, Kaci McCoy, Rhoel R. Dinglasan, and Melissa S. Nolan. Needs assessment of southeastern United States vector control agencies: Capacity improvement is greatly needed to prevent the next vector-borne disease outbreak. *Tropical Medicine and Infectious Disease*, 7(5):73, May 2022.

[9] Emily M Mader, Claudia Ganser, Annie Geiger, Laura C Harrington, Janet Foley, Rebecca L Smith, Nohra Mateus-Pinilla, Pete D Teel, and Rebecca J Eisen. A Survey of Tick Surveillance and Control Practices in the United States. *Journal of Medical Entomology*, 58(4):1503–1512, July 2021.

[10] Maria A. Diuk-Wasser, Gwenaël Vourc'h, Paul Cislo, Anne Gatewood Hoen, Forrest Melton, Sarah A. Hamer, Michelle Rowland, Roberto Cortinas, Graham J. Hickling, Jean I. Tsao, Alan G. Barbour, Uriel Kitron, Joseph Piesman, and Durland Fish. Field and climate-based model for predicting the density of host-seeking nymphal Ixodes scapularis, an important vector of tick-borne disease agents in the eastern United States. *Global Ecology and Biogeography*, 19(4):504–514, 2010.

[11] Robert M. Dorazio and Fred A. Johnson. Bayesian Inference and Decision Theory—a Framework for Decision Making in Natural Resource Management. *Ecological Applications*, 13(2):556–563, 2003.

[12] Brian J. Reich, Krishna Pacifici, and Jonathan W. Stallings. Integrating auxiliary data in optimal spatial design for species distribution modelling. *Methods in Ecology and Evolution*, 9(6):1626–1637, 2018.

[13] B. K. M. Case, Jean-Gabriel Young, Daniel Penados, Carlota Monroy, Laurent

Hébert-Dufresne, and Lori Stevens. Spatial epidemiology and adaptive targeted sampling to manage the Chagas disease vector Triatoma dimidiata. *PLOS Neglected Tropical Diseases*, 16(6):e0010436, June 2022.

[14] Luc Pronzato and Eric Walter. Robust experiment design via stochastic approximation. *Mathematical Biosciences*, 75(1):103–120, July 1985.

[15] Eric Parent and Etienne Rivot. *Introduction to Hierarchical Bayesian Modeling for Ecological Data*. CRC Press, Boca Raton, August 2012.

[16] Elizabeth G. Ryan, Christopher C. Drovandi, James M. McGree, and Anthony N. Pettitt. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*, 84(1):128–154, 2016.

[17] Peter Diggle, Benjamin Amoah, Claudio Fronterre, Emanuele Giorgi, and Olatunji Johnson. Rethinking neglected tropical disease prevalence survey design and analysis: A geospatial paradigm. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 115(3):208–210, 2021.

[18] Elizabeth G. Ryan, Christopher C. Drovandi, and Anthony N. Pettitt. Simulation-based fully Bayesian experimental design for mixed effects models. *Computational Statistics & Data Analysis*, 92:26–39, December 2015.

[19] Centers for Disease Control and Prevention. Guide to the surveillance of metastriate ticks (Acari: Ixodidae) and their pathogens in the United States. Available from: https://www.cdc.gov/ticks/pdfs/Tick_surveillance-P.pdf, 2020.

[20] Micah B. Hahn, Catherine S. Jarnevich, Andrew J. Monaghan, and Rebecca J. Eisen. Modeling the Geographic Distribution of Ixodes scapularis and Ixodes pacificus (Acari: Ixodidae) in the Contiguous United States. *Journal of Medical Entomology*, 53(5):1176–1191, September 2016.

[21] Catherine A. Lippi, Holly D. Gaff, Alexis L. White, and Sadie J. Ryan. Scoping review of distribution models for selected Amblyomma ticks and rickettsial group pathogens. *PeerJ*, 9:e10596, February 2021.

[22] Jon Dewitz. National Land Cover Database 2019 products. Available from: https://www.sciencebase.gov/catalog/item/5f21cef582cef313ed940043, 2021.

[23] Oregan State University. PRISM Climate Group. Available from: https://prism.oregonstate.edu, December 2021 [accessed 5 July 2022].

[24] Oleg A. Alduchov and Robert E. Eskridge. Improved Magnus Form Approximation of Saturation Vapor Pressure. *Journal of Applied Meteorology and Climatology*, 35(4):601–609, April 1996.

[25] Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Institute of Mathematical Statistics*, 10(3):273–304, 1995.

[26] Michael Chipeta, Dianne Terlouw, Kamija Phiri, and Peter Diggle. Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance

structure. *Environmetrics*, 28(1):1–11, 2017.

[27] Jérôme Pagès. Analyse factorielle multiple de données mixtes: principe et exemple d'application. *Revue Statistique Appliquee*, 52(4):93–111, 2004.

[28] Kathryn Chaloner and Kinley Larntz. Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, 21(2):191–208, February 1989.

[29] Evelyn C. Rynkiewicz and Keith Clay. Tick community composition in Midwestern US habitats in relation to sampling method and environmental conditions. *Experimental and Applied Acarology*, 64(1):109–119, September 2014.

[30] Susan P Elias, Allison M Gardner, Kirk A Maasch, Sean D Birkel, Norman T Anderson, Peter W Rand, Charles B Lubelczyk, and Robert P Smith, Jr. A Generalized Additive Model Correlating Blacklegged Ticks With White-Tailed Deer Density, Temperature, and Humidity in Maine, USA, 1990–2013. *Journal of Medical Entomology*, 58(1):125–138, January 2021.

[31] Rebecca J. Eisen, Lars Eisen, and Charles B. Beard. County-Scale Distribution of *Ixodes scapularis* and *Ixodes pacificus* (Acari: Ixodidae) in the Continental United States. *Journal of Medical Entomology*, 53(2):349–386, March 2016.

[32] Yuri P. Springer, Lars Eisen, Lorenza Beati, Angela M. James, and Rebecca J. Eisen. Spatial Distribution of Counties in the Continental United States With Records of Occurrence of Amblyomma americanum (Ixodida: Ixodidae). *Journal of Medical Entomology*, 51(2):342–351, March 2014.

[33] Yuri P. Springer, Catherine S. Jarnevich, David T. Barnett, Andrew J. Monaghan, and Rebecca J. Eisen. Modeling the Present and Future Geographic Distribution of the Lone Star Tick, Amblyomma americanum (Ixodida: Ixodidae), in the Continental United States. *The American Journal of Tropical Medicine and Hygiene*, 93(4):875–890, October 2015.

[34] R C Falco and D Fish. Potential for exposure to tick bites in recreational parks in a Lyme disease endemic area. *American Journal of Public Health*, 79(1):12–15, January 1989.

[35] Erin Hassett, Maria Diuk-Wasser, Laura Harrington, and Pilar Fernandez. Integrating tick density and park visitor behaviors to assess the risk of tick exposure in urban parks on Staten Island, New York. *BMC public health*, 22(1):1602, August 2022.

[36] Tammi L. Johnson, Christine B. Graham, Karen A. Boegler, Cara C. Cherry, Sarah E. Maes, Mark A. Pilgard, Andrias Hojgaard, Danielle E. Buttke, and Rebecca J. Eisen. Prevalence and Diversity of Tick-Borne Pathogens in Nymphal Ixodes scapularis (Acari: Ixodidae) in Eastern National Parks. *Journal of Medical Entomology*, 54(3):742–751, May 2017.

[37] Lars Eisen and Rebecca J. Eisen. Critical Evaluation of the Linkage Between Tick-Based Risk Measures and the Occurrence of Lyme Disease Cases. *Journal*

of Medical Entomology, 53(5):1050–1062, September 2016.

[38] P. Mead, S. Hook, S. Niesobecki, J. Ray, J. Meek, M. Delorey, C. Prue, and A. Hinckley. Risk factors for tick exposure in suburban settings in the Northeastern United States. *Ticks and Tick-Borne Diseases*, 9(2):319–324, February 2018.

[39] Kim M. Pepin, Rebecca J. Eisen, Paul S. Mead, Joseph Piesman, Durland Fish, Anne G. Hoen, Alan G. Barbour, Sarah Hamer, and Maria A. Diuk-Wasser. Geographic Variation in the Relationship between Human Lyme Disease Incidence and Density of Infected Host-Seeking Ixodes scapularis Nymphs in the Eastern United States. *The American Journal of Tropical Medicine and Hygiene*, 86(6):1062–1071, June 2012.

[40] Michael Chipeta, Dianne Terlouw, Kamija Phiri, and Peter Diggle. Adaptive geostatistical design and analysis for prevalence surveys. *Spatial Statistics*, 15:70–84, 2016.

[41] Virgilio Gómez-Rubio. *Bayesian inference with INLA*. CRC Press, 2020.

[42] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002.

[43] Thiago G Martins, Daniel Simpson, Finn Lindgren, and Håvard Rue. Bayesian computing with inla: new features. *Computational Statistics & Data Analysis*, 67:68–83, 2013.

[44] Jürgen Branke, Stephan Meisel, and Christian Schmidt. Simulated annealing in the presence of noise. *Journal of Heuristics*, 14(6):627–654, December 2008.

[45] Ruth K. Meyer and Christopher J. Nachtsheim. The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs. *Technometrics*, 37(1):60–69, February 1995.

[46] J. A Royle. Exchange algorithms for constructing large spatial designs. *Journal of Statistical Planning and Inference*, 100(2):121–134, February 2002.

Chapter 4

Accurately summarizing an outbreak using epidemiological models takes time

Abstract

Recent outbreaks of Mpox and Ebola, and worrying waves of COVID-19, influenza and respiratory syncytial virus, have all led to a sharp increase in the use of epidemiological models to estimate key epidemiological parameters. The feasibility of this estimation task is known as the *practical identifiability* (PI) problem. Here, we investigate the PI of eight commonly reported statistics of the classic Susceptible-Infectious-Recovered model using a new measure that shows how much a researcher can expect to learn in a model-based Bayesian analysis of prevalence data. Our findings show that the basic reproductive number and final outbreak size are often poorly identified, with learning exceeding that of individual model parameters only in the early stages of an outbreak. The peak intensity, peak timing, and initial growth rate are better identified, being in expectation over 20 times more probable having seen the data by the time the underlying outbreak peaks. We then test PI for a

variety of true parameter combinations, and find that PI is especially problematic in slow-growing or less-severe outbreaks. These results add to the growing body of literature questioning the reliability of inferences from epidemiological models when limited data are available.

4.1 Introduction

Incredible efforts have been made in recent years to apply epidemiological models to the empirical data borne out of the COVID-19 pandemic. The LitCovid aggregator currently contains over 3,000 papers on "epidemic forecasting" and "modelling and estimating" trends of COVID-19 spread [1]. We are seeing similar waves of models and forecasts for recent outbreaks of mpox, Ebola, influenza and respiratory syncytial virus. However, the enormous variability in model predictions, even among works using the same model and similar data, erodes confidence when interpreting these efforts for policy decisions [2]. It is clear that uncertainty remains about what we can expect to learn from these models, and when.

Disease models tackle the difficult challenge of describing complex epidemic processes by relating mechanistic processes to population-level observations such as daily reported cases. Identifying combinations of parameters that plausibly replicate observed data can help summarize the epidemic dynamics. Common statistics include the basic reproductive number, the average number of new cases someone will cause in an entirely susceptible population, and the outbreak size, the fraction of the population who will eventually have had the disease. Because these indicators are the product of interacting social and biological phenomena, they are never available through direct observation. Fitting epidemiological models to data is one of the best options

for estimating these important quantities [3].

The classic Susceptible-Infectious-Recovered (SIR) model accounts for a minimal number of critical mechanisms of disease spread. Infectious individuals infect susceptible individuals at a rate β and recover at a rate α. These mechanisms can be tracked through time by a set of ordinary differential equations:

$$\frac{d}{dt}S = -\beta SI \ , \quad \frac{d}{dt}I = \beta SI - \alpha I \ , \quad \text{and} \quad \frac{d}{dt}R = \alpha I \ .$$

It is common to consider S, I and R as a fraction of the population in a given state such that $S + I + R = 1$ at all times. The initial state of the population might not be known—especially the susceptible pool $S_0 \equiv S(t = 0)$. Focusing on the second equation, we can see that the epidemic will grow exponentially at a rate $\beta S_0 - \alpha$ for initial small values of I, resulting in near-exchangeability of the parameters and causing large uncertainty in individual parameter values early on [4, 5]. Conversely, when I becomes small after the peak, the infectious population eventually decays exponentially at a rate α. These observations make clear that data regarding I will provide information about different parameters, or combinations thereof, at different points of an outbreak. In general, the amount of information that can be learned about a given quantity will in general depend on the structure of the model equations, the timing of observations, and the level of noise in the data [6].

Despite the model's simplicity, several authors have cautioned that the reliability of inferences drawn from the SIR model are questionable when based on prevalence data alone [7]. Due to the structural nature of the SIR equations, these issues are particularly acute during the early stage of an outbreak, when inferences are critical for informing timely public health response [8, 5]. Without careful incorporation of

additional data, these reliability problems can only grow with additional complexity in the model equations or observational structure [9, 2]. In order to draw meaningful conclusions, researchers are forced to rely on data from one or more epidemic waves [10], or make strong and potentially controversial assumptions about parameters governing disease spread [11]. A more general understanding of how properties of epidemiological models affect uncertainty in commonly reported summary statistics would help researchers quantify how much they can expect to learn in empirical studies, and establish sufficient criteria for reproducibility. Therefore, the goal of this report is to provide a comprehensive baseline for the reliability of estimates for a number of commonly reported statistics, with emphasis on the time necessary to predict these statistics in an emerging epidemic accurately and to illuminate the structural interactions between data, model dynamics, and summary statistics.

This question of whether quantities estimated from data are reliable, e.g. compatible with some hypothetical true parameters $\boldsymbol{\theta}^* = (\alpha^*, \beta^*, S_0^*)$ which generated the data, is termed the *practical identifiability* (PI) problem and has traditionally been studied using the variance-covariance matrix of an estimator for $\boldsymbol{\theta}^*$ [12]. However, such second-order approaches underestimate uncertainty in limited data settings, where the distribution of plausible parameters may be skewed [13, 14]. Here we propose a new measure that allows us to efficiently and directly measure our ability to learn various epidemiological quantities at all stages of an epidemic. If $u = \varphi(\boldsymbol{\theta})$ is an unknown variable to be estimated, our Bayesian interpretation of the identifiability of u is the expected logarithm of the ratio between posterior and prior probabilities,

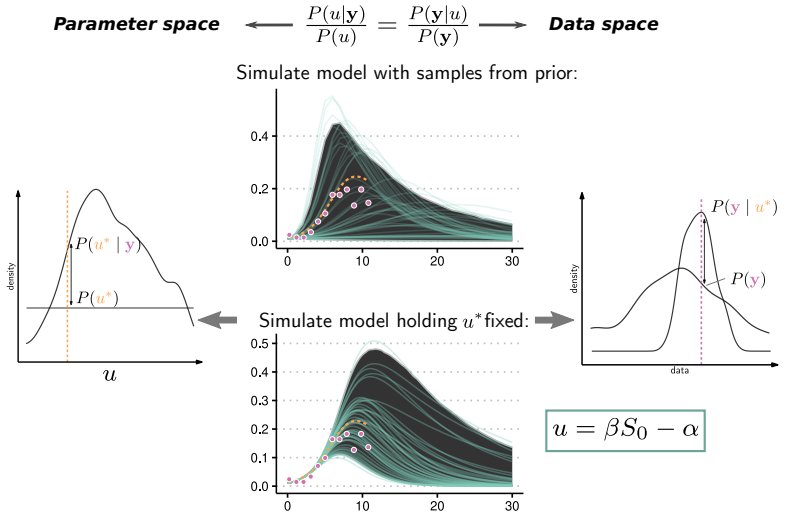

Figure 4.1: Intuition behind the proposed practical identifiability method. From Bayes' rule, δ_u may be written either as (4.1) or (4.5), and ultimately reflects the difference in information between two sets of model dynamics: first the set of dynamics encoded in the full prior $P(\boldsymbol{\theta})$, and second the dynamics compatible with u^* as encoded through a restricted prior $P(\boldsymbol{\theta} \mid u^*)$.

evaluated at $u^* = \varphi(\boldsymbol{\theta}^*)$:

$$\delta_u(\boldsymbol{\theta}^*) = \mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}^*}\left[\log P(u^* \mid \mathbf{y})\right] - \log P(u^*) \tag{4.1}$$

where $\mathbf{y} \mid \boldsymbol{\theta}^*$ are noisy observations of the underlying outbreak, e.g., daily case counts, and where the expectation is taken over realizations of the observation process. Since shrinkage in the posterior distribution is facilitated through the global behaviour of the model likelihood, (4.1) is able to capture uncertainty arising from complex model fits, such as bimodality in the likelihood surface. As with traditional approaches to PI, δ_u is a local measure of information gain, in the sense that changing the true dynamics $\boldsymbol{\theta}^*$ will in general give different answers [15]. This allows the effect of particular values of $\boldsymbol{\theta}^*$ to be studied. Intuitively, we can understand (4.1) as comparing two sets of model dynamics, the first as generated from the full prior $P(\boldsymbol{\theta}^*)$, and the

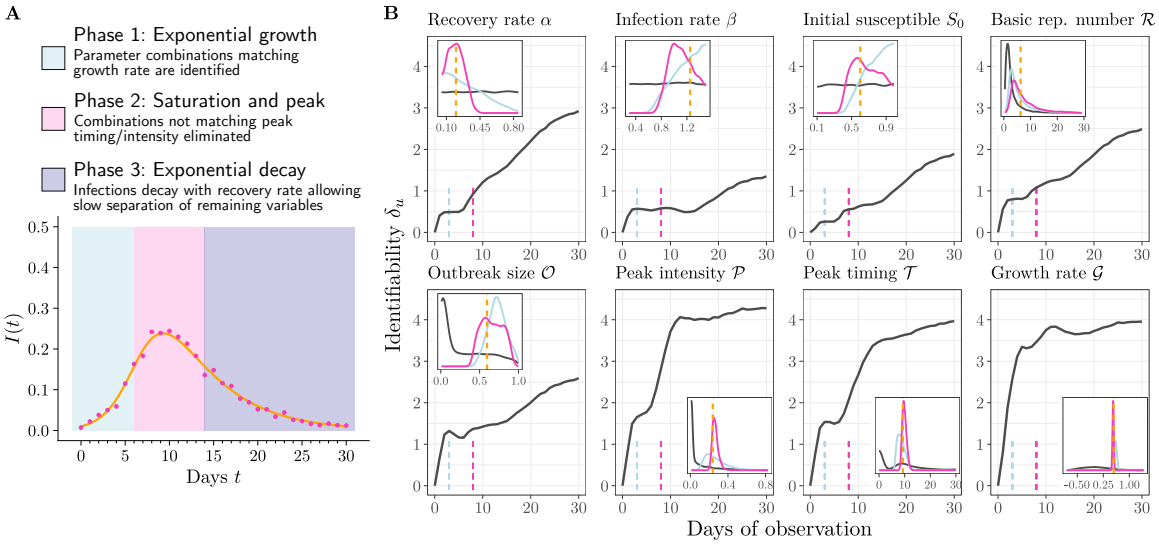

Figure 4.2: Practical identifiability (PI) of epidemiological summary statistics over time. (**A**) Unknown deterministic SIR process based on true parameters $\boldsymbol{\theta}^*$ (orange line), and single realization of observed data $\mathbf{y} \sim P(\mathbf{y} \mid \boldsymbol{\theta}^*)$ (pink dots). (**B**) Main panels show PI according to δ_u over an increasing observation window assuming daily observations. Insets give an example of how δ_u is interpreted, showing $P(u \mid \mathbf{y})$ and $P(u)$ for the single realization of $\mathbf{y}$ from (A), observed up to $T = 3$ (blue) and $T = 8$ (pink). The dashed orange line is the true value to be estimated. True parameters are $\alpha^* = 0.2$, $\beta^* = 1.25$, and $S_0^* = 0.6$, with $I_0 = 0.01$ assumed known. Prior beliefs are $\alpha \sim U(0.05, 0.85)$, $\beta \sim U(0.3, 1.5)$, $S_0 \sim U(0.1, 0.99)$.

second as if u^* were assumed known (Figure 4.1). Note the metric does not require computationally expensive Bayesian inference methods to compute—a simple Monte Carlo procedure for estimating (4.1) is provided in Section 4.4.3.

4.2 Results

Figure 4.2 shows the PI of the SIR model parameters, as well as five summary variables which are commonly calculated in terms of $\boldsymbol{\theta}$ (see Table 4.1 for mathematical definitions), for a typical parametrization $\boldsymbol{\theta}^*$ of the model. Infectious individuals are assumed to be independently tested at a fixed rate η at daily timepoints, giving

Table 4.1: Definitions of epidemiological summary statistics.

Name	Symbol	Formula
Reproductive number	$\mathcal{R}$	β/α
Outbreak size	$\mathcal{O}$	$1 - R(0) - S_0 \exp(-\mathcal{R}\mathcal{O})^*$
Peak intensity	$\mathcal{P}$	$I_0 + S_0 + [1 - \log(S_0/\mathcal{R})]/\mathcal{R}$
Peak timing	$\mathcal{T}$	Unknown
Growth rate	$\mathcal{G}$	$\beta S_0 - \alpha$

*Implicit equation

a likelihood $y_t \sim \text{Poisson}(\eta I(t; \boldsymbol{\theta}^*))$. We assumed $\eta = 1000$ is known throughout, which leads to limited observational noise to better study PI inherent to the SIR equations. δ_u is computed daily for the eight variables, up to a maximum of 30 days of observation.

The rate of learning for all variables is uneven over time, with each reaching plateaus of varying length before the peak. The infection rate β is the worst identified. Gaining information on α appears easier than β and S_0 and even exceeds learning for $\mathcal{R}$ and $\mathcal{O}$ after around $T = 20$ days of observation. PI of the peak intensity, peak timing, and growth rate increase more rapidly at first, with learning for growth rate happening particularly fast.

These findings illustrate the difficulty of learning key quantities early in an epidemic, under real-time conditions where the number of observations increases as the outbreak goes on. However, the question remains to what extent a lack of early learning may be attributed simply to a smaller sample size. Therefore we next examined the PI of several variables over an increasing observation window, but with the number of evenly distributed observations kept constant. Figure 4.3 shows that identifiability of β and $\mathcal{R}$ is lowest when observations are concentrated prior to the peak, confirming that the limits of early learning are indeed a structural property of

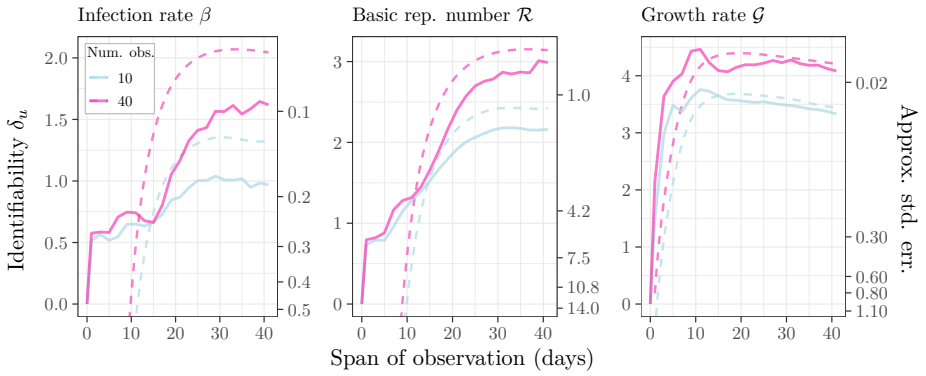

Figure 4.3: Practical identifiability of several variables as a function of testing frequency. Observations are evenly distributed over the interval $[0, T]$ for increasing days of observation T. Solid lines are PI calculated using Monte Carlo, while dashed lines indicate the approximation given by (4.18). The approximation also gives an asymptotic relationship between δ_u and a lower bound on the standard error, indicated with secondary axes. Priors and true parameters are the same as in Figure 4.2.

the SIR equations that cannot be overcome by allocating additional tests early on. Further, increasing the frequency of testing from 10 observations to 40 did little to increase PI during this period, but increased PI considerably for wider observation windows. Figure 4.3 also shows the functional relationship between the asymptotic limit of δ_u and the usual standard error for u, as given by (4.18), which can serve as an alternative interpretation of δ_u when there is sufficient data. For example, spreading 10 observations over 40 days gives $\delta_\beta \approx 1$, which for our chosen priors means we can expect the standard deviation for posteriors $P(\beta \mid \mathbf{y})$ to be at least 0.2 (but could be much larger in reality, since we are far from the asymptotic limit).

To test the sensitivity of these findings to $\boldsymbol{\theta}^*$, we then computed δ_u over a grid of values for β^* and S_0^* (Figure 4.4). Since slower-growing outbreaks will naturally contain less information per day [7], information gain was calculated using observations up until the first day after the epidemic peak. To investigate the factors of a true outbreak most associated with learning, for each true value of the eight variables

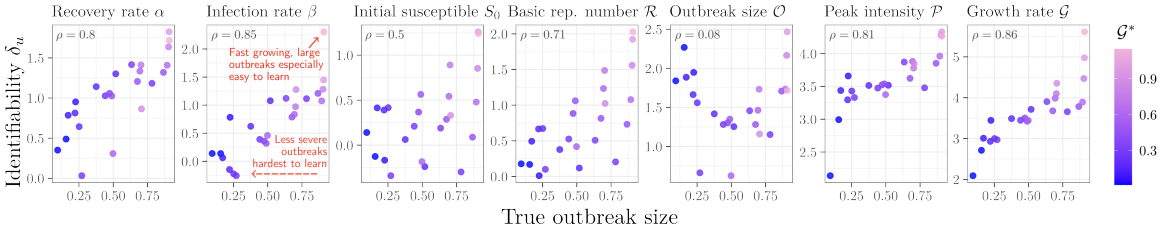

Figure 4.4: Change in identifiability when the true parameters $\boldsymbol{\theta}^*$ are varied. δ_u is calculated using daily observations up to the first day after the true (unobserved) outbreak has peaked. True parameters tested were all combinations of $\beta^* = 0.3, 0.5, \ldots, 1.5$ and $S_0^* = 0.1, 0.3, \ldots, 0.9$ with $\alpha^* = 0.2$ fixed. Pearson correlation between δ_u and true outbreak size is given in corners of each panel. Priors are the same as in Figure 4.2.

considered, the correlation between δ_u for each variable and the true value was computed. The outbreak size of the true epidemic was the most correlated with learning, followed by true growth rate, illustrating that less-severe outbreaks are harder to learn.

4.3 Discussion

The analysis presented here makes it clear that some epidemiological variables are easier to estimate through model dynamics than others, and emphasizes that most epidemiological summary statistics should be interpreted with caution when data are limited. Taken together, the rate of learning for all the variables suggests that learning takes place in three general phases. In phase 1, plausible parameter combinations quickly concentrate along the surface $\{\boldsymbol{\theta} : \beta S_0 - \alpha = \mathcal{G}^*\}$, as infections increase exponentially with the initial growth rate. This explains the sharp but modest gain in information of all variables except for $\mathcal{G}$ during this phase. In phase 2, infections begin to saturate and parameter combinations matching the true peak intensity and timing become more plausible. However, for β especially, saturating case counts do

little to further restrict the plausible parameter surface from phase 1. Finally, phase 3 is characterized by gradual information gain for the remaining variables. Since infections are slowly declining with α during this phase, this growth is explained by α^* gradually being identified, which propagates to allow some remaining combinations on the plausible surface to be eliminated.

Parameters describing the mechanisms of the model—β, α and S_0—take a particularly long time to learn on account of quickly reaching a plateau at low values of δ_u. As a result, the SIR model is more effective at forecasting short-term statistics of the dynamics such as peak timing and intensity, than it is at estimating mechanisms. This result shows how difficult it is to estimate parameters from early data in the hope of forecasting the impacts of mechanistic interventions such as reducing β with preventive measures or increasing α with treatment [16]. Importantly, even though a lack of identifiability implies a wide range of parameters lead to similar infectious dynamics early on, these plausible dynamics will still respond differently to interventions targeting specific mechanisms [17]. Thus, low PI simply means that an intervention's impact is difficult to forecast ahead of time.

Learning was nearly as difficult for the statistics $\mathcal{R}$ and $\mathcal{O}$ as for the individual model parameters, despite the fact that optimistically, these transformations would combine the information of each parameter they depend on. The failure of these statistics to resolve closely exchangeable parameter combinations limits their reliability for succinctly describing an epidemic. In contrast, the initial growth rate resolves such combinations to give rapid shrinkage to the correct value, despite encoding similar information as $\mathcal{R}$ about disease dynamics [18]. This suggests growth rates are a more reliable "first look" at an outbreak when using prevalence data under the SIR

model.

When varying the true values $\boldsymbol{\theta}^*$, see Figure 4.4, we find that less-severe outbreaks are generally harder to learn, despite having more daily observations available before their peak. The initial susceptible population S_0 appears the most poorly identified across values of $\boldsymbol{\theta}^*$ by the peak, and the expected posterior shrinkage is even slightly negative for 25% of the tested values. An interesting implication for control measures is that the more we reduce the severity of true infection dynamics, the harder it will be to accurately estimate the impacts of interventions. Further, the mode of intervention matters: variability along the y-axis in Figure 4.4 for similar values of $\mathcal{O}^*$ shows lowering S_0^* impacts learning differently than a reduction in β^*.

Previous investigations into the PI of the SIR model have mainly focused on the PI of α and β under the simplified model where $S_0 \approx 1$ is known. These works generally agree that PI of both α and β is limited during phase 1 [4, 5], but that the majority of information available has been learned by the time the disease has peaked [9, 19, 20]. Most comparably to the observational design in Figure 4.2, Capaldi et al. (2012) considered the asymptotic variance of $\hat{\beta}$ and $\hat{\alpha}$ over an increasing timespan, and found the variance of both estimators decreased rapidly and smoothly just before and after the peak, respectively [7]. In contrast, the delayed rate of learning of these parameters in Figure 4.2 paints a more pessimistic picture of PI when exact likelihoods and prior context is taken into account. This finding supports the idea that previous PI results based on approximation theory underestimate uncertainty, particularly during the early stages of an outbreak when the likelihood surface is highly nonlinear [13, 21].

In this work, we have proposed a novel means of assessing PI which measures the

expected posterior gain in density at the true value u^*. While comparing densities at a specific value may seem to ignore uncertainty in the posterior as a whole, we argue δ_u is better interpreted as a measure of *shrinkage*, rather than density, by marginalizing the global curvature of the likelihood onto a single dimension for u. If the projected span of high likelihood values is more narrow than the support of $P(u)$, shrinkage will occur and δ_u becomes positive. In this sense, (4.1) might be viewed as a quantitative alternative to the popular profile likelihood method, in which potential plateaus in the likelihood surface are projected to the space of some parameter θ_i and examined graphically [22]. Additionally, as shown in Section 4.4.4, δ_u may be interpreted in terms of standard measurements of uncertainty—in the limit of large data and under certain conditions, δ_u converges to a form of the usual standard error of the maximum likelihood estimator, penalized by the prior weight. Therefore, while our measure was specifically designed to give a more accurate picture of uncertainty in limited data regimes, it also has asymptotic behavior similar to the coefficient of variation for u.

The Bayesian nature of our method of assessing PI means that estimates of model parameters and any variables which depend on them are sensitive to prior beliefs. In this report, our choice of uniform priors represents modest assumptions about an emerging pathogen: *a priori*, just over 50% of scenarios result in an outbreak (i.e. have $\beta S_0/\alpha > 1$), and outbreaks range from modest to highly severe (70% of individuals infected at peak). However, for many pathogens, more informative prior information is frequently available, for example on the recovery rate of a disease [23]. Relative to more realistic prior settings, this may mean α is more difficult to gain information about than β and S_0.

While we have considered only noisy observation of the current infectious popula-

tion, real data may also come in the form of daily new infections or cumulative case counts, and may suffer from lags in reporting or preferential sampling [24, 25]. Learning epidemiological variables from such data will have their own distinct challenges [9]. PI of the SIR model should also be assessed with hierarchical models incorporating data from multiple sources, such as hospitalizations and isolated clinical experiments [26]. Yet, our work shows that even in its simplest form, learning parameters and statistics of SIR dynamics takes time, limiting which inferences, forecasts, and control policies can be made from early epidemic data.

4.4 Supplemental information

4.4.1 Data availability

Materials necessary to reproduce this analysis are available on GitHub and have been archived on Zenodo [27]. A Julia package MarginalDivergence.jl has been developed for efficient computation of our PI measure, including an interface for easy implementation of user-provided models. The package is currently unregistered.

4.4.2 Likelihood-based estimation of dynamical systems

While the methods considered here can be applied to any statistical process for which a likelihood exists, we are interested in processes of the form

$$y_i \sim g(\mathbf{x}(t_i), \boldsymbol{\sigma}) \tag{4.2}$$

$$\dot{\mathbf{x}} = h(\mathbf{x}(t), \boldsymbol{\tau}) \tag{4.3}$$

where $\mathbf{y} = (y_1, \ldots, y_n)$ are observations at discrete timepoints $t_1, \ldots, t_n$, and $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$ are parameters that are assumed known or are to be estimated. We are interested in our ability to estimate a p-dimensional vector of unknown parameters $\boldsymbol{\theta}^*$, comprised of some elements of $\boldsymbol{\sigma}^*$, $\boldsymbol{\tau}^*$, and $\mathbf{x}(0)^*$.

Given $\boldsymbol{\theta}$, (1.1) and (1.2) form a probability distribution $P(\mathbf{y} \mid \boldsymbol{\theta})$ called the *likelihood*. In the frequentist paradigm, an estimator for $\boldsymbol{\theta}^*$ can be obtained by maximizing $P(\mathbf{y} \mid \boldsymbol{\theta})$,

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \text{argmax}_{\boldsymbol{\theta}} P(\mathbf{y} \mid \boldsymbol{\theta}).$$

A popular way to assess issues of practical identifiability is through the variance-covariance matrix of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$, which can show marginal uncertainty in individual parameter estimators and correlations between pairs of estimators. The Cramer-Rao bound implies that in the limit of decreasing observation uncertainty (i.e. as the amount or precision of data increases), the variance of an unbiased estimator converges, given certain regularity conditions, to the inverse of the Fisher Information Matrix $\mathcal{I}(\boldsymbol{\theta}^*)$, where

$$[\mathcal{I}(\boldsymbol{\theta})]_{ij} = -\,\mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}} \left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log P(\mathbf{y} \mid \boldsymbol{\theta}) \right]. \tag{4.4}$$

This bound can underestimate variance when measurement noise is not infinitesimal [14, 21], leading some to question its applicability even for simple nonlinear models [13, 28]. An alternative is to estimate the distribution of $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ using Monte Carlo simulation, by sampling possible data sets $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots$ from $P(\mathbf{y} \mid \boldsymbol{\theta}^*)$ and finding the maximum of each likelihood $P(\mathbf{y}^{(j)} \mid \boldsymbol{\theta})$ using an optimization algorithm such as gradient descent. The resulting samples $\hat{\boldsymbol{\theta}}^{(j)}_{\text{MLE}}$ can then be inspected graphically or used to estimate the covariance matrix. This method has the convenience of

also working with estimates of transformations of the model parameters, without the need for further approximation [29].

4.4.3 Proposed method of assessing practical identifiability

While using Monte Carlo estimation of $\mathrm{Var}(\hat{\boldsymbol{\theta}}_{\mathrm{MLE}})$ to assess PI can alleviate the underestimation issues when using the information matrix, the use of optimization to obtain a sample of the estimator can lead to dependence on initial conditions or other hyperparameters of the optimization method used [12]. Again, the inaccuracy of this method will be most acute when the likelihood surface is flat or multi-modal, such as when limited data are available.

Rather than relying on optimization, we instead take a sampling-based Bayesian perspective. From the main text, we have for a variable of interest $u = \varphi(\boldsymbol{\theta})$, $\delta_u(\boldsymbol{\theta}^*) = \mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}^*}\left[\log P(u^* \mid \mathbf{y})\right] - \log P(u^*)$, which gives the average amount, over possible future outbreaks $P(\mathbf{y} \mid \boldsymbol{\theta}^*)$, a researcher can expect to learn about the true quantity u^* in a Bayesian analysis. A value of $\delta_u = c$ corresponds roughly to an expected gain in posterior probability e^c times greater than the prior.

Equation (4.1) can be rewritten by applying Bayes' rule, $P(u^* \mid \mathbf{y})/P(u^*) = P(\mathbf{y} \mid u^*)/P(\mathbf{y})$, where the margin $P(\mathbf{y} \mid u^*)$ equals $\int P(\mathbf{y} \mid \boldsymbol{\theta})P(\boldsymbol{\theta} \mid u^*)d\boldsymbol{\theta}$ and $P(\boldsymbol{\theta} \mid u^*)$ is the distribution of the epidemiological parameters compatible with a fixed variable of interest u^*—we give details below. This leads to

$$\delta_u(\boldsymbol{\theta}^*) = \mathbf{E}_{\mathbf{y}|\boldsymbol{\theta}^*}\left[\log \frac{P(\mathbf{y} \mid u^*)}{P(\mathbf{y})}\right], \tag{4.5}$$

which shows δ_u may be interpreted as the expected difference between two marginal likelihoods. The first, $P(\mathbf{y} \mid u^*)$, is the evidence under a reduced model where u^* is assumed known, and the second is the evidence of the full model.

We approximate $\delta_u(\boldsymbol{\theta}^*)$ by generating M paired Monte Carlo samples from $P(\boldsymbol{\theta} \mid u^*)$ and $P(\boldsymbol{\theta})$, and reusing these samples to obtain M samples from $P(\mathbf{y} \mid u^*)$ and $P(\mathbf{y})$ for each $\mathbf{y} \sim P(\mathbf{y} \mid \boldsymbol{\theta}^*)$, leading to

$$\delta_u(\boldsymbol{\theta}^*) \approx \frac{1}{N} \sum_{i=1}^{N} \left[\log \sum_{j=1}^{M} P(\mathbf{y}^{(i)} \mid \tilde{\boldsymbol{\theta}}^{(j)}) - \log \sum_{j=1}^{M} P(\mathbf{y}^{(i)} \mid \boldsymbol{\theta}^{(j)}) \right] \tag{4.6}$$

where $\tilde{\boldsymbol{\theta}}^{(j)} \sim P(\boldsymbol{\theta} \mid u^*)$, $\boldsymbol{\theta}^{(j)} \sim P(\boldsymbol{\theta})$, and $\mathbf{y}^{(i)} \sim P(\mathbf{y} \mid \boldsymbol{\theta}^*)$. $N = 3000$ and $M = 60,000$ were used for all computations in the main text.

Accuracy of Monte Carlo estimation of δ_u

The marginal likelihood $P(\mathbf{y})$ is notorious for being inefficient to estimate via Monte Carlo methods. One easy way to improve the reliability of estimates is to use resampling methods, where a single random number is used to seed an entire sequence of samples from the target distribution [30]. Here and in the main text, we use *systematic sampling* independently on each $\theta_i \in \boldsymbol{\theta}$, which amounts to generating independent random numbers for each i to generate M low-discrepancy samples from $P(\boldsymbol{\theta})$ (or $P(\boldsymbol{\theta} \mid u^*)$) [31]. To test our choice of M was large enough with this setup while still within a reasonable computational budget, we repeated calculations of $\log P(\mathbf{y})$ for increasing values of M, where a $\mathbf{y} \sim P(\mathbf{y} \mid \boldsymbol{\theta}^*)$ was sampled with 60 observations (every half day). $\boldsymbol{\theta}^*$ and $P(\boldsymbol{\theta})$ were the same as in Figure 1 of the main text. We concluded that even with 60 observations, which gives a likelihood sharper than the maximum 30 observations used in the main text, a choice of $M > 30,000$ was suf-

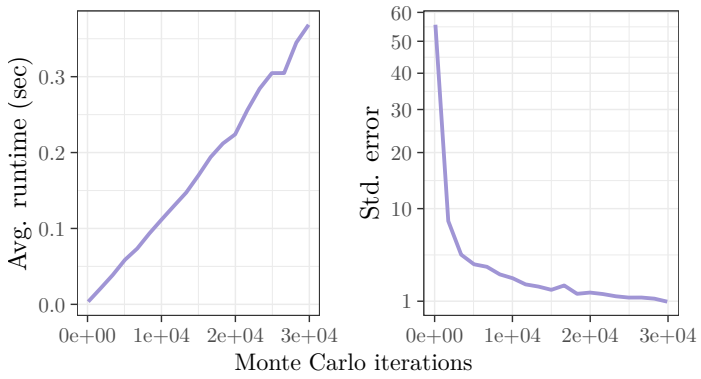

Figure 4.5: Speed and accuracy of approximating the log marginal likelihood $\log P(\mathbf{y})$ using M Monte Carlo simulations. The results are averaged over 100 repetitions of the sampling process for each M.

ficient to give a standard error less than 1, or less than 0.5% of the magnitude of $\log P(\mathbf{y})$. The runtime and standard errors from 100 independent computations of $P(\mathbf{y})$ are shown as a function of M in Figure 4.5.

Practical identifiability for a function of model parameters

We first justify our above claim that the density $P(\mathbf{y} \mid u)$ is the marginal likelihood for a reduced model with restricted priors $P(\boldsymbol{\theta} \mid u)$. Using the formula for a vector-to-scalar transformation of $\boldsymbol{\theta} \mid \mathbf{y} \mapsto u \mid \mathbf{y}$, we have

$$\begin{aligned} P(\mathbf{y} \mid u) &= P(\mathbf{y}) \frac{P(u \mid \mathbf{y})}{P(u)} \\ &= \frac{P(\mathbf{y})}{P(u)} \int P(\boldsymbol{\theta} \mid \mathbf{y}) \delta(\varphi(\boldsymbol{\theta}) = u) d\boldsymbol{\theta} \\ &= \frac{1}{P(u)} \int P(\mathbf{y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}, u) d\boldsymbol{\theta} \\ &= \int P(\mathbf{y} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid u) d\boldsymbol{\theta}. \end{aligned}$$

Thus, $P(\mathbf{y} \mid u^*)$ can be approximated using samples from $P(\boldsymbol{\theta} \mid u^*)$, as done in 4.6. However, the distribution function $P(\boldsymbol{\theta} \mid u^*)$ will generally not be available in closed form even when $P(\boldsymbol{\theta})$ is.

Simulating from $P(\boldsymbol{\theta} \mid u^*)$ can be accomplished with the following procedure: let $\theta_i \in \boldsymbol{\theta}$ be a chosen "pivot" parameter/index and define $\tilde{\varphi}(\theta_i \mid \boldsymbol{\theta}_{-i}) = u$ to be a univariate function conditional on $\boldsymbol{\theta}_{-i}$, where $\boldsymbol{\theta}_{-i}$ indicates the ith element of $\boldsymbol{\theta}$ has been removed. Assume $\tilde{\varphi}$ is invertible so that $\tilde{\varphi}^{-1}(u \mid \boldsymbol{\theta}_{-i}) = \theta_i$. Then, assuming independent priors on the elements of $\boldsymbol{\theta}$, using a change of variables and Bayes' rule we have

$$P(\boldsymbol{\theta}_{-i} \mid u^*) \propto \prod_{j \neq i} P(\theta_j) P(u^* \mid \boldsymbol{\theta}_{-i}), \tag{4.7}$$

where

$$P(u \mid \boldsymbol{\theta}_{-i}) = \left| \frac{d}{du} \tilde{\varphi}^{-1}(u \mid \boldsymbol{\theta}_{-i}) \right| P_{\theta_i}(\tilde{\varphi}^{-1}(u^* \mid \boldsymbol{\theta}_{-i})). \tag{4.8}$$

If the priors are not element-wise independent, note that (4.7) can be replaced with the more general $P(\boldsymbol{\theta}_{-i}) P(u^* \mid \boldsymbol{\theta}_{-i})$.

Because $\tilde{\varphi}$ is deterministic (i.e. θ_i can be uniquely determined given $\boldsymbol{\theta}_{-i}$ and u), samples from $P(\boldsymbol{\theta} \mid u^*)$ can therefore be obtained by first sampling $\boldsymbol{\theta}_{-i}^{(1)}, \ldots, \boldsymbol{\theta}_{-i}^{(n)}$ from (4.7) using a standard simulation technique such as Accept-Reject sampling, and then letting $\theta_i^{(j)} = \tilde{\varphi}^{-1}(u^*, \boldsymbol{\theta}_{-i}^{(j)})$.

The resulting densities for the five transformations in Table 4.1 are shown in Figure 4.6. For example, under the transformation $\varphi(\boldsymbol{\theta}) = \frac{\beta}{\alpha} =: \mathcal{R}$, we define $\tilde{\varphi}^{-1}(\alpha, S_0, \mathcal{R}) = \alpha\mathcal{R}$ and obtain

$$P(\alpha, S_0 \mid \mathcal{R}) \propto P_\alpha(\alpha) P_{S_0}(S_0) \alpha P_\beta(\mathcal{R}\alpha). \tag{4.9}$$

So we may sample $(\alpha^{(1)}, S_0^{(1)}), (\alpha^{(2)}, S_0^{(2)}), \ldots$ from (4.9), then let $\beta^{(i)} = \mathcal{R}^* \alpha^{(i)}$ to

obtain samples from $P(\alpha, \beta, S_0 \mid \mathcal{R}^*)$.

For the final outbreak size, we define $\mathcal{O} := R(\infty) - R(0)$ to be the total proportion of individuals who end up in the recovered compartment due to infection. For $R(\infty)$ we have from [32],

$$R(\infty) = 1 - S_0 \exp\left(-\mathcal{R}(R(\infty) - R(0))\right), \tag{4.10}$$

which we may use to solve for β and obtain the inverse function

$$\beta = \frac{-\alpha}{\mathcal{O}} \log \frac{1 - R(0) - \mathcal{O}}{S_0} \tag{4.11}$$

and derivative

$$\frac{d\beta}{d\mathcal{O}} = \frac{\alpha}{\mathcal{O}} \left(\frac{1}{\mathcal{O}} \log \frac{1 - R(0) - \mathcal{O}}{S_0} + \frac{1}{1 - R(0) - \mathcal{O}} \right). \tag{4.12}$$

For the peak intensity $\mathcal{P} := \max_t I(t)$, to obtain samples from (4.7) we may use the equation

$$\mathcal{P} = I_0 + S_0 - \frac{\alpha}{\beta} \log S_0 - \frac{\alpha}{\beta} \left(1 + \log \frac{\alpha}{\beta} \right). \tag{4.13}$$

Although (4.13) yields only implicit solutions for any θ_i, a closed-form solution for S_0 given $\mathcal{P}$ can be found using Lambert's W,

$$S_0 = -\mathcal{R}^{-1} W_{-1}(-B), \tag{4.14}$$

where $B = \exp\left(-\mathcal{R}(\mathcal{P} - I_0) - 1\right)$, and derivative

$$\frac{dS_0}{d\mathcal{P}} = \frac{1}{1 - Be^{W_{-1}(-B)}}. \tag{4.15}$$

Derivation of the necessary equations for the initial growth rate $\mathcal{G} := \beta S_0 + \alpha$ is straightforward.

Finally, the peak timing $\mathcal{T}$ does not have a known closed-form solution. Though more time-consuming, we can still approximate (4.7) by using univariate constrained optimization to evaluate the unknown φ^{-1}, and adjoint methods to obtain the corresponding derivative.

4.4.4 Asymptotic properties of practical identifiability

The above procedure for estimating δ_u using simple Monte Carlo becomes inefficient when the dimension of $\boldsymbol{\theta}$ or $\mathbf{y}$ become large. In the latter case, the proposed method of practical identifiability can instead be analyzed using the usual approximation theory in the limit of large data. The Bernstein-von Mises theorem gives

$$P(\boldsymbol{\theta} \mid \mathbf{y}) \approx \mathcal{N}(\hat{\boldsymbol{\theta}}_{\text{MLE}}, (n_t \mathcal{I}(\hat{\boldsymbol{\theta}}_{\text{MLE}}))^{-1}), \tag{4.16}$$

where n_t is the number of independent replications of the time series of observations, and $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Under certain regularity conditions, $\hat{\boldsymbol{\theta}}_{\text{MLE}} = \boldsymbol{\theta}^*$.

The information matrix of the model parameters can be separated in terms of the curvature of latent and observation processes. In the case where data from a single state variable x is observed, $\mathcal{I}(\boldsymbol{\theta}) = \mathbf{J}^\top \mathcal{O}(\boldsymbol{\theta}) \mathbf{J}$, where $\mathbf{J}$ is the Jacobian of x with

respect to $\boldsymbol{\theta}$, $\mathbf{J}_{ij} = \partial x(t_i)/\partial \boldsymbol{\theta}_j$, and $\mathcal{O}(\boldsymbol{\theta})$ is the information of $\mathbf{y}$ given $\mathbf{x}$ [6]. In the case of independent Poisson-distributed testing, $\mathcal{O}(\boldsymbol{\theta}) = \text{diag}(\eta I(t_1; \boldsymbol{\theta}), \ldots, \eta I(t_n; \boldsymbol{\theta}))$.

To obtain an approximation for the posterior of a transformation $u = \varphi(\boldsymbol{\theta})$, we may again choose a pivot parameter/index θ_i and introduce a change of variables $\mathbf{v} = (\theta_1, \ldots, \theta_{i-1}, u, \theta_{i+1}, \ldots, \theta_p)$, and define the vector-valued function $\tilde{\mathbf{f}}$ so that $\tilde{\mathbf{f}}(\mathbf{v}) = \boldsymbol{\theta}$. Letting $\mathbf{V}$ be the gradient of $\tilde{\mathbf{f}}$ with respect to $\mathbf{v}$, $\mathbf{V}_{kj} = \frac{\partial}{\partial v_j}\tilde{\mathbf{f}}_k = \frac{\partial \theta_k}{\partial v_j}$, we have

$$\mathcal{I}(\mathbf{v}) = \mathbf{V}^\top \mathbf{J}^\top \mathcal{O}(\boldsymbol{\theta}) \mathbf{J} \mathbf{V}. \tag{4.17}$$

Combining (4.16) and (4.17), therefore, gives the approximation

$$\delta_u(\boldsymbol{\theta}^*) \approx \frac{1}{2}\left(\log n_t - \log \mathcal{I}(\mathbf{v}^*)^{-1}_{ii} - \log(2\pi)\right) - \log P(u^*). \tag{4.18}$$

This reveals that, in the limit of sufficient data, δ_u is related to the local curvature of the latent and observational processes, just like traditional asymptotic approaches to PI. For individual model parameters, $\mathbf{V}$ is the identity matrix and δ_u becomes the logarithm of the usual standard error of the estimator $\hat{u}_{\text{MLE}}$, penalized by the prior log-probability of $\hat{u}_{\text{MLE}}$. For parameter transformations, the information of $\mathcal{I}(\boldsymbol{\theta})$ is then summarized further through the curvature induced by the transformation function φ. (4.18) also gives another way to see that penalizing by the prior density has a normalizing effect on δ_u, as transformations which increase the support of $P(u)$ will also have smaller prior densities, and therefore is analogous to using the coefficient of variation to allow comparing standard errors between variables.

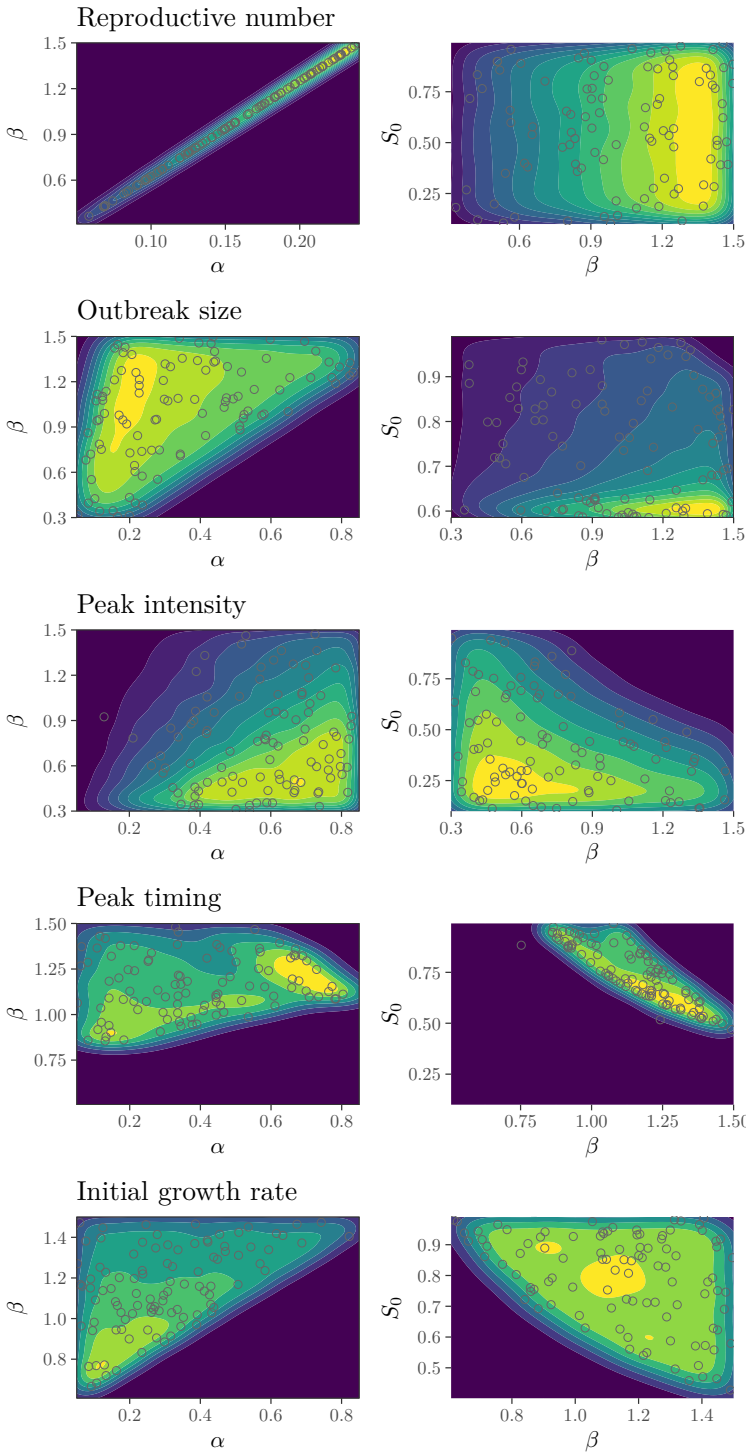

Figure 4.6: Density of 60,000 samples from $P(\alpha, \beta \mid u^*)$, and $P(\beta, S_0 \mid u^*)$ given different summary transformations. True values and priors are the same as in the main text.

Bibliography

[1] Qingyu Chen, Alexis Allot, and Zhiyong Lu. LitCovid: An open database of COVID-19 literature. *Nucleic Acids Research*, 49(D1):D1534–D1540, January 2021.

[2] Weston C. Roda, Marie B. Varughese, Donglin Han, and Michael Y. Li. Why is it difficult to accurately predict the COVID-19 epidemic? *Infectious Disease Modelling*, 5:271–281, 2020.

[3] Joseph T. Wu, Kathy Leung, and Gabriel M. Leung. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study. *Lancet (London, England)*, 395(10225):689–697, February 2020.

[4] Gerardo Chowell. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*, 2(3):379–398, August 2017.

[5] Omar Melikechi, Alexander L. Young, Tao Tang, Trevor Bowman, David Dunson, and James Johndrow. Limits of epidemic prediction using SIR models. *Journal of Mathematical Biology*, 85(4):36, September 2022.

[6] G. A. F Seber and C. J. Wild. *Nonlinear Regression*. John Wiley & Sons, Hoboken, 2 edition, 2003.

[7] Alex Capaldi, Samuel Behrend, Benjamin Berman, Jason Smith, Justin Wright, and Alun Lloyd. Parameter Estimation and Uncertainty Quantication for an Epidemic Model. *Mathematical Biosciences and Engineering*, 9(3):553–576, July 2012.

[8] Mario Castro, Saúl Ares, José A. Cuesta, and Susanna Manrubia. The turning point and end of an expanding epidemic cannot be precisely forecast. *Proceedings of the National Academy of Sciences*, 117(42):26190–26196, October 2020.

[9] Necibe Tuncer and Trang T. Le. Structural and practical identifiability analysis of outbreak models. *Mathematical Biosciences*, 299:1–18, May 2018.

[10] Rahul Subramanian, Qixin He, and Mercedes Pascual. Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proceedings of the National Academy of Sciences of the United States of America*, 118(9):e2019716118, March 2021.

[11] Tom Britton, Frank Ball, and Pieter Trapman. A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2. *Science*, 369(6505):846–849, August 2020.

[12] Nicholas N. Lam, Paul D. Docherty, and Rua Murray. Practical identifiability of parametrised models: A review of benefits and limitations of various approaches. *Mathematics and Computers in Simulation*, 199:202–216, September 2022.

[13] M. Joshi, A. Seidel-Morgenstern, and A. Kremling. Exploiting the bootstrap

method for quantifying parameter confidence intervals in dynamical systems. *Metabolic Engineering*, 8(5):447–455, September 2006.

[14] Dhruva V. Raman, James Anderson, and Antonis Papachristodoulou. Delineating parameter unidentifiabilities in complex models. *Physical Review E*, 95(3):032314, March 2017.

[15] Jesse Sharp, Alexander Browning, Kevin Burrage, and Matthew J. Simpson. Parameter estimation and uncertainty quantification using information geometry. *Journal of The Royal Society Interface*, 19:20210940, 2022.

[16] Michael Barnett, Greg Buchak, and Constantine Yannelis. Epidemic responses under uncertainty. *Proceedings of the National Academy of Sciences of the United States of America*, 120(2):e2208111120, January 2023.

[17] Amani Alahmadi, Sarah Belet, Andrew Black, Deborah Cromer, Jennifer A. Flegg, Thomas House, Pavithra Jayasundara, Jonathan M. Keith, James M. McCaw, Robert Moss, Joshua V. Ross, Freya M. Shearer, Sai Thein Than Tun, James Walker, Lisa White, Jason M. Whyte, Ada W. C. Yan, and Alexander E. Zarebski. Influencing public health policy with data-informed mathematical models of infectious diseases: Recent developments and new challenges. *Epidemics*, 32:100393, September 2020.

[18] Luís M. A. Bettencourt and Ruy M. Ribeiro. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. *PloS One*, 3(5):e2185, May 2008.

[19] Kimberlyn Roosa and Gerardo Chowell. Assessing parameter identifiability in compartmental dynamic models using a computational approach: Application to infectious disease transmission models. *Theoretical Biology and Medical Modelling*, 16(1):1, January 2019.

[20] Chiara Piazzola, Lorenzo Tamellini, and Raúl Tempone. A note on tools for prediction under uncertainty and identifiability of SIR-like dynamical systems for epidemiology. *Mathematical Biosciences*, 332:108514, February 2021.

[21] Keegan E. Hines, Thomas R. Middendorf, and Richard W. Aldrich. Determination of parameter identifiability in nonlinear biophysical models: A Bayesian approach. *Journal of General Physiology*, 143(3):401–416, March 2014.

[22] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, August 2009.

[23] Muge Cevik, Matthew Tate, Ollie Lloyd, Alberto Enrico Maraolo, Jenna Schafers, and Antonia Ho. SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: A systematic review and meta-analysis. *The Lancet. Microbe*, 2(1):e13–e22, January 2021.

[24] Weihsueh A. Chiu and Martial L. Ndeffo-Mbah. Using test positivity and re-

ported case rates to estimate state-level COVID-19 prevalence and seroprevalence in the United States. *PLOS Computational Biology*, 17(9):e1009374, September 2021.

[25] Sean L. Wu, Andrew N. Mertens, Yoshika S. Crider, Anna Nguyen, Nolan N. Pokpongkiat, Stephanie Djajadi, Anmol Seth, Michelle S. Hsiang, John M. Colford, Art Reingold, Benjamin F. Arnold, Alan Hubbard, and Jade Benjamin-Chung. Substantial underestimation of SARS-CoV-2 infection in the United States. *Nature Communications*, 11(1):4507, September 2020.

[26] Daniela De Angelis, Anne M. Presanis, Paul J. Birrell, Gianpaolo Scalia Tomba, and Thomas House. Four key challenges in infectious disease modelling using data from multiple sources. *Epidemics*, 10:83–87, March 2015.

[27] B. K. M. Case. brendandaisy/epi-summaries-over-time: RSOS Materials (Version v1), 2023. Zenodo. 10.5281/zenodo.8280113.

[28] Niels Krausch, Tilman Barz, Annina Sawatzki, Mathis Gruber, Sarah Kamel, Peter Neubauer, and Mariano Nicolas Cruz Bournazou. Monte Carlo Simulations for the Analysis of Non-linear Parameter Confidence Intervals in Optimal Experimental Design. *Frontiers in Bioengineering and Biotechnology*, 7, 2019.

[29] G. Chowell, N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore, and J. M. Hyman. The basic reproductive number of Ebola and the effects of public health measures: The cases of Congo and Uganda. *Journal of Theoretical Biology*, 229(1):119–126, July 2004.

[30] Randal Douc, Olivier Cappé, and Eric Moulines. Comparison of Resampling Schemes for Particle Filtering. *arXiv:cs/0507025 [cs]*, July 2005.

[31] Fredrik Bagge Carlson. MonteCarloMeasurements.jl: Nonlinear Propagation of Arbitrary Multivariate Distributions by means of Method Overloading. *arXiv:2001.07625 [cs, stat]*, January 2020.

[32] Howard Weiss. The SIR model and the Foundations of Public Health. *Materials Matemàtics*, 2013(3):17, 2013.

Conclusion

This work has presented Bayesian Experimental Design (BED) as a formal approach to surveillance and control efforts in epidemiology. Across several epidemiological domains spanning vector ecology, neglected tropical disease management, and infectious disease modeling, we demonstrated the promise of BED for directing decision making and allowing researchers to get more out of the data they collect. This work also demonstrates how BED is one avenue in which theoretical and computational scientists can integrate more closely with empiricists and incorporate additional context from the systems they study into their work.

Compared with classical experimental design frameworks, we have seen how BED can be highly flexible in terms of how utility is defined and how data are incorporated to inform beliefs. However, this flexibility can also contribute to BED being a daunting task to implement in practical applications. As concluded in Chapters 2 and 3, there is a need for software supporting BED, as the technical nature of modern advances in BED have left them largely inaccessible to a general audience [1]. Therefore, a broad challenge for the future of BED is the development of software that helps automate or guide each of the three key steps of BED—model definition and selection, derivation and approximation of a suitable utility, and optimization—which would greatly help a wider audience implement BED procedures.

A second broad challenge for the future of BED in epidemiology is the incorporation of more realistic observation processes into our models, and an improved understanding of how additional observational complexity alters optimal design theory. For infectious disease modeling, this might include a model for preferential sampling of symptomatic individuals based on prevalence and test availability [2]. This would contort the space of dynamics compatable with data, and interact with the existing identifiability issues of SIR models in interesting ways, which in turn will change the optimal allotment of tests over the course of an outbreak. For an example regarding Chagas disease, an important observational dynamic which we have ignored is from missing data, in particular households that do not consent to inspections or treatments [3]. While the fraction of these households was small enough that they could be reasonably ignored in our application, a more general solution would be to explicitly model the probability of allowing treatment and only allowing designs which include consenting households. A similar observational process to incorporate into future work is a statistical model for how houses are selected during adaptive sampling, which could help correct for the effects of preferential sampling and improve the accuracy of assessing whether the control target has been met [4].

There are some more technical points which seem more relevant to this work as a whole than to individual chapters. In Chapter 2, an adaptive design framework was proposed which weights future locations based on their current marginal risk and variance. One may have noticed, however, that this does not align with the canonical definition of utility as an average over the joint distribution $P(\mathbf{y}, \boldsymbol{\theta})$, given in (1.10). A more proper utility here would be to rank batches of new locations based on their *anticipated* variance reduction of unsampled locations, as was done in Chapter 3,

although this would involve an expectation over a posterior predictive distribution and an optimization step within each iteration of selecting batches, and thus would be much more computationally intensive [5]. An alternative would be to incorporate the design goal of meeting the reduction target directly into selecting batches of houses, for example by choosing new batches which maximize the tail probability given by the termination condition (2.5). Since this utility would theoretically balance the tradeoff between sampling bias and efficiency implicitly, it would be interesting to compare it to our explicit approach of interpolating between the two goals.

As a closing comment, note that for a dissertation on experimental design, a major limitation is that no data were collected in response to the theoretically optimal designs which were found in this work. While the designs found here were effective with respect to the range of possibilities predicted from the underlying model, whether this aligns with the value ultimately extracted from these designs assumes, among other things, correct model specification [6]. How optimal designs based on non-trivial models actually perform in the field is a topic that has been rarely discussed, and effective means for evaluating BED in practice is a major avenue for future work [7, 8]. However, the work in this dissertation has been a step towards a more complete synthesis between theory and data collection.

Bibliography

[1] Elizabeth G. Ryan, Christopher C. Drovandi, James M. McGree, and Anthony N. Pettitt. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*, 84(1):128–154, 2016.

[2] Weihsueh A. Chiu and Martial L. Ndeffo-Mbah. Using test positivity and reported case rates to estimate state-level COVID-19 prevalence and seroprevalence in the United States. *PLOS Computational Biology*, 17(9):e1009374, September 2021.

[3] Claudia Arevalo-Nieto, Justin Sheen, Gian Franco Condori-Luna, Carlos Condori-

Pino, Julianna Shinnick, Jennifer K. Peterson, Ricardo Castillo-Neyra, and Michael Z. Levy. Incentivizing optimal risk map use for Triatoma infestans surveillance in urban environments. *PLOS Global Public Health*, 2(8):e0000145, August 2022.

[4] Joe Watson, James V. Zidek, and Gavin Shaddick. A general theory for preferential sampling in environmental networks. *The Annals of Applied Statistics*, 13(4):2662–2700, December 2019.

[5] Steven Kleinegesse, Christopher Drovandi, and Michael U. Gutmann. Sequential Bayesian Experimental Design for Implicit Models via Mutual Information. *arXiv:2003.09379 [cs, stat]*, March 2020.

[6] Sabina J. Sloman, Daniel M. Oppenheimer, Stephen B. Broomell, and Cosma Rohilla Shalizi. Characterizing the robustness of Bayesian adaptive experimental designs to active learning bias. *arXiv:2205.13698 [stat]*, November 2022.

[7] Ruth Etzioni and Joseph B. Kadane. Optimal Experimental Design for Another's Analysis. *Journal of the American Statistical Association*, 88(424):1404–1411, December 1993.

[8] Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian Experimental Design. *arXiv:2302.14545 [cs, stat]*, February 2023.

Bibliography

Abbott, Michael C. and Benjamin B. Machta (2023). "Far from Asymptopia". *Entropy* 25.3, p. 434. arXiv: 2205.03343 [physics, stat].

Aguas, Ricardo et al. (2022). *Herd Immunity Thresholds for SARS-CoV-2 Estimated from Unfolding Epidemics*.

Aiga, Hirotsugu et al. (2012). "Chagas Disease : Assessing the Existence of a Threshold for Bug Infestation Rate". *American Journal of Tropical Medicine and Hygiene* 86.6, pp. 972–979.

Alahmadi, Amani et al. (2020). "Influencing Public Health Policy with Data-Informed Mathematical Models of Infectious Diseases: Recent Developments and New Challenges". *Epidemics* 32, p. 100393.

Alduchov, Oleg A. and Robert E. Eskridge (1996). "Improved Magnus Form Approximation of Saturation Vapor Pressure". *Journal of Applied Meteorology and Climatology* 35.4, pp. 601–609.

Althouse, Benjamin M. et al. (2020). "Superspreading Events in the Transmission Dynamics of SARS-CoV-2: Opportunities for Interventions and Control". *PLOS Biology* 18.11, e3000897.

Andrade-Pacheco, Ricardo et al. (2020). "Finding Hotspots: Development of an Adaptive Spatial Sampling Approach". *Scientific Reports* 10.1, pp. 1–12.

Arevalo-Nieto, Claudia et al. (2022). "Incentivizing Optimal Risk Map Use for Triatoma Infestans Surveillance in Urban Environments". *PLOS Global Public Health* 2.8, e0000145.

Arnold, Benjamin F. et al. (2018). "Integrated Serologic Surveillance of Population Immunity and Disease Transmission". *Emerging Infectious Diseases* 24.7, pp. 1188–1194.

Ball, Ian R, Hugh P Possingham, and Matthew Watts (2009). "Marxan and relatives: software for spatial conservation prioritisation". *Spatial Conservation Prioritisation: Quantitative Methods and Computational Tools* 14, pp. 185–196.

Banerjee, Sudipto, Bradley P Carlin, and Alan Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. 2nd ed. Boca Raton: CRC Press.

Barnett, Michael, Greg Buchak, and Constantine Yannelis (2023). "Epidemic Responses under Uncertainty". *Proceedings of the National Academy of Sciences of the United States of America* 120.2, e2208111120.

Berger, Loïc et al. (2021). "Rational Policymaking during a Pandemic". *Proceedings of the National Academy of Sciences* 118.4, e2012704118.

Bettencourt, Luís M. A. and Ruy M. Ribeiro (2008). "Real Time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases". *PloS One* 3.5, e2185.

Booth, Mark and Archie Clements (2018). "Neglected Tropical Disease Control – The Case for Adaptive, Location-specific Solutions". *Trends in Parasitology* 34.4, pp. 272–282.

Bouillon, César P (2005). *The Millennium Development Goals in Latin America and the Caribbean: progress, priorities and IDB support for their implementation.* Available at SSRN: https://ssrn.com/abstract=1543858.

Branke, Jürgen, Stephan Meisel, and Christian Schmidt (2008). "Simulated Annealing in the Presence of Noise". *Journal of Heuristics* 14.6, pp. 627–654.

Britton, Tom, Frank Ball, and Pieter Trapman (2020). "A Mathematical Model Reveals the Influence of Population Heterogeneity on Herd Immunity to SARS-CoV-2". *Science* 369.6505, pp. 846–849.

Bustamante, Dulce, Sandra De Urioste-Stone, et al. (2014). "Ecological, Social and Biological Risk Factors for Continued Trypanosoma Cruzi Transmission by Triatoma Dimidiata in Guatemala". *PLoS ONE* 9.8.

Bustamante, Dulce, Marianela Menes Hernández, et al. (2015). "Information to Act: Household Characteristics Are Predictors of Domestic Infestation with the Chagas Vector Triatoma Dimidiata in Central America". *American Journal of Tropical Medicine and Hygiene* 93.1, pp. 97–107.

Bustamante, Dulce Maria, Carlota Monroy, et al. (2009). "Risk Factors for Intradomiciliary Infestation by the Chagas Disease Vector Triatoma Dimidiatain Jutiapa, Guatemala". *Cadernos De Saude Publica* 25 Suppl 1, S83–92.

Bustamante, Dulce Maria, Maria Carlota Monroy, et al. (2007). "Environmental Determinants of the Distribution of Chagas Disease Vectors in South-Eastern Guatemala". *Geospatial Health* 1.2, pp. 199–211.

Cable, Joanne et al. (2017). "Global Change, Parasite Transmission and Disease Control: Lessons from Ecology". *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372.1719, p. 20160088.

Cahan, Sara Helms et al. (2019). "Residual Survival and Local Dispersal Drive Reinfestation by Triatoma Dimidiata Following Insecticide Application in Guatemala". *Infection, Genetics and Evolution* 74, p. 104000.

Capaldi, Alex et al. (2012). "Parameter Estimation and Uncertainty Quantication for an Epidemic Model". *Mathematical Biosciences and Engineering* 9.3, pp. 553–576.

Carlson, Fredrik Bagge (2020). "MonteCarloMeasurements.Jl: Nonlinear Propagation of Arbitrary Multivariate Distributions by Means of Method Overloading". *arXiv:2001.07625 [cs, stat]*. arXiv: 2001.07625 [cs, stat].

Case, B. K. M. (2022). *PLoS NTD Publication Materials*. Version v1.0.

— (2023). *brendandaisy/epi-summaries-over-time: RSOS Materials (Version v1)*. Zenodo. 10.5281/zenodo.8280113.

Case, B. K. M. et al. (2022). "Spatial Epidemiology and Adaptive Targeted Sampling to Manage the Chagas Disease Vector Triatoma Dimidiata". *PLOS Neglected Tropical Diseases* 16.6, e0010436.

Castro, Mario et al. (2020). "The Turning Point and End of an Expanding Epidemic Cannot Be Precisely Forecast". *Proceedings of the National Academy of Sciences* 117.42, pp. 26190–26196.

Cevik, Muge et al. (2021). "SARS-CoV-2, SARS-CoV, and MERS-CoV Viral Load Dynamics, Duration of Viral Shedding, and Infectiousness: A Systematic Review and Meta-Analysis". *The Lancet. Microbe* 2.1, e13–e22.

Chaloner, Kathryn and Kinley Larntz (1989). "Optimal Bayesian Design Applied to Logistic Regression Experiments". *Journal of Statistical Planning and Inference* 21.2, pp. 191–208.

Chaloner, Kathryn and Isabella Verdinelli (1995). "Bayesian Experimental Design: A Review". *Institute of Mathematical Statistics* 10.3, pp. 273–304.

Chen, Qingyu, Alexis Allot, and Zhiyong Lu (2021). "LitCovid: An Open Database of COVID-19 Literature". *Nucleic Acids Research* 49.D1, pp. D1534–D1540.

Chernoff, Herman (1953). "Locally optimal designs for estimating parameters". *The Annals of Mathematical Statistics*, pp. 586–602.

Chipeta, Michael et al. (2016). "Adaptive Geostatistical Design and Analysis for Prevalence Surveys". *Spatial Statistics* 15, pp. 70–84.

— (2017). "Inhibitory Geostatistical Designs for Spatial Prediction Taking Account of Uncertain Covariance Structure". *Environmetrics* 28.1, pp. 1–11. arXiv: 1605.00104.

Chiu, Weihsueh A. and Martial L. Ndeffo-Mbah (2021). "Using Test Positivity and Reported Case Rates to Estimate State-Level COVID-19 Prevalence and Seroprevalence in the United States". *PLOS Computational Biology* 17.9, e1009374.

Chiuchiolo, Cristian, Janet van Niekerk, and Haavard Rue (2021). "Joint Posterior Inference for Latent Gaussian Models with R-INLA". *arXiv:2112.02861 [stat]*. arXiv: 2112.02861 [stat].

Chowell, G. et al. (2004). "The Basic Reproductive Number of Ebola and the Effects of Public Health Measures: The Cases of Congo and Uganda". *Journal of Theoretical Biology* 229.1, pp. 119–126.

Chowell, Gerardo (2017). "Fitting Dynamic Models to Epidemic Outbreaks with Quantified Uncertainty: A Primer for Parameter Uncertainty, Identifiability, and Forecasts". *Infectious Disease Modelling* 2.3, pp. 379–398.

Cross, Paul C. et al. (2019). "Confronting Models with Data: The Challenges of Estimating Disease Spillover". *Philosophical Transactions of the Royal Society B: Biological Sciences* 374.1782, p. 20180435.

De Angelis, Daniela et al. (2015). "Four Key Challenges in Infectious Disease Modelling Using Data from Multiple Sources". *Epidemics*. Challenges in Modelling Infectious DIsease Dynamics 10, pp. 83–87.

Dewitz, Jon (2021). *National Land Cover Database 2019 products*. Available from: https://www.sciencebase.gov/catalog/item/5f21cef582cef313ed940043.

Diggle, Peter, Benjamin Amoah, et al. (2021). "Rethinking Neglected Tropical Disease Prevalence Survey Design and Analysis: A Geospatial Paradigm". *Transactions of the Royal Society of Tropical Medicine and Hygiene* 115.3, pp. 208–210.

Diggle, Peter and Søren Lophaven (2006). "Bayesian Geostatistical Design". *Scandinavian Journal of Statistics* 33.1, pp. 53–64.

Diggle, Peter, Raquel Menezes, and Ting li Su (2010). "Geostatistical Inference under Preferential Sampling". *Journal of the Royal Statistical Society. Series C: Applied Statistics* 59.2, pp. 191–232.

Diggle, Peter, Jonathan Tawn, and Rana Moyeed (1998). "Model-based geostatistics". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3, pp. 299–350.

Disease Control, Centers for and Prevention (2020). *Guide to the Surveillance of Metastriate Ticks (Acari: Ixodidae) and their Pathogens in the United States*. Available from: https://www.cdc.gov/ticks/pdfs/Tick_surveillance-P.pdf.

Diuk-Wasser, Maria A. et al. (2010). "Field and Climate-Based Model for Predicting the Density of Host-Seeking Nymphal Ixodes Scapularis, an Important Vector of Tick-Borne Disease Agents in the Eastern United States". *Global Ecology and Biogeography* 19.4, pp. 504–514.

Dorazio, Robert M. and Fred A. Johnson (2003). "Bayesian Inference and Decision Theory—a Framework for Decision Making in Natural Resource Management". *Ecological Applications* 13.2, pp. 556–563.

Dorn, Patricia L. et al. (2003). "The Chagas Vector, Triatoma Dimidiata (Hemiptera: Reduviidae), Is Panmictic within and among Adjacent Villages in Guatemala". *Journal of Medical Entomology* 40.4, pp. 436–440.

Douc, Randal, Olivier Cappé, and Eric Moulines (2005). "Comparison of Resampling Schemes for Particle Filtering". *arXiv:cs/0507025 [cs]*. arXiv: cs/0507025.

Drovandi, Christopher C et al. (2017). "Principles of Experimental Design for Big Data Analysis". *Statistical science : a review journal of the Institute of Mathematical Statistics* 32.3, pp. 385–404.

Drovandi, Christopher C. and Minh-Ngoc Tran (2018). "Improving the Efficiency of Fully Bayesian Optimal Design of Experiments Using Randomised Quasi-Monte Carlo". *Bayesian Analysis* 13.1, pp. 139–162.

Dumonteil, Eric et al. (2004). "Re-Infestation of Houses by Triatoma Dimidiata after Intra-Domicile Insecticide Application in the Yucatán Peninsula, Mexico". *Memorias Do Instituto Oswaldo Cruz* 99.3, pp. 253–256.

Dye-Braumuller, Kyndall C. et al. (2022). "Needs Assessment of Southeastern United States Vector Control Agencies: Capacity Improvement Is Greatly Needed to Prevent the next Vector-Borne Disease Outbreak". *Tropical Medicine and Infectious Disease* 7.5, p. 73.

Eisen, Lars and Rebecca J. Eisen (2016). "Critical Evaluation of the Linkage Between Tick-Based Risk Measures and the Occurrence of Lyme Disease Cases". *Journal of Medical Entomology* 53.5, pp. 1050–1062.

Eisen, Rebecca J, Kiersten J Kugeler, et al. (2017a). "Tick-Borne Zoonoses in the United States: Persistent and Emerging Threats to Human Health". *ILAR Journal* 58.3, pp. 319–335.

— (2017b). "Tick-Borne Zoonoses in the United States: Persistent and Emerging Threats to Human Health". *ILAR Journal* 58.3, pp. 319–335.

Eisen, Rebecca J and Christopher D Paddock (2021). "Tick and Tickborne Pathogen Surveillance as a Public Health Tool in the United States". *Journal of Medical Entomology* 58.4, pp. 1490–1502.

Eisen, Rebecca J., Lars Eisen, and Charles B. Beard (2016). "County-Scale Distribution of *Ixodes Scapularis* and *Ixodes Pacificus* (Acari: Ixodidae) in the Continental United States". *Journal of Medical Entomology* 53.2, pp. 349–386.

Elias, Susan P et al. (2021). "A Generalized Additive Model Correlating Blacklegged Ticks With White-Tailed Deer Density, Temperature, and Humidity in Maine, USA, 1990–2013". *Journal of Medical Entomology* 58.1, pp. 125–138.

Etzioni, Ruth and Joseph B. Kadane (1993). "Optimal Experimental Design for Another's Analysis". *Journal of the American Statistical Association* 88.424, pp. 1404–1411.

Falco, R C and D Fish (1989). "Potential for Exposure to Tick Bites in Recreational Parks in a Lyme Disease Endemic Area." *American Journal of Public Health* 79.1, pp. 12–15.

Fox, Spencer J. et al. (2020). "The COVID-19 Herd Immunity Threshold Is Not Low: A Re-Analysis of European Data from Spring of 2020". *medRxiv:2020.12.01.20242289*.

Franco-Paredes, Carlos et al. (2007). "Chagas Disease: An Impediment in Achieving the Millennium Development Goals in Latin America". *BMC International Health and Human Rights* 7, p. 7.

Fronterre, Claudio et al. (2020). "Design and Analysis of Elimination Surveys for Neglected Tropical Diseases". *The Journal of Infectious Diseases* 221.Supplement_5, S554–S560.

Fuglstad, Geir-Arne et al. (2019). "Constructing Priors That Penalize the Complexity of Gaussian Random Fields". *Journal of the American Statistical Association* 114.525, pp. 445–452.

Gelfand, Alan et al. (2010). *Handbook of Spatial Statistics*. Boca Raton: Taylor \& Francis. arXiv: 1011.1669v3.

Gelman, Andrew, John Carlin, et al. (2020). *Bayesian Data Analysis*. 3rd ed. Boca Raton: CRC Press.

Gelman, Andrew and Jennifer Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, Andrew and Xiao-Li Meng (1998). "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling". *Statistical Science* 13.2, pp. 163–185. JSTOR: 2676756.

Gelman, Andrew and Yuling Yao (2020). "Holes in Bayesian Statistics". *Journal of Physics G: Nuclear and Particle Physics* 48.1, p. 014002.

"Geostatistical Design" (2007). *Model-based Geostatistics*. New York, NY: Springer New York, pp. 199–212.

Gómez-Palacio, Andrés et al. (2015). "Ecological Niche and Geographic Distribution of the Chagas Disease Vector, Triatoma Dimidiata (Reduviidae: Triatominae): Evidence for Niche Differentiation among Cryptic Species". *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 36, pp. 15–22.

Gómez-Rubio, Virgilio (2020). *Bayesian inference with INLA*. CRC Press.

Guatemala Ministerio de Salud (2019 [cited 12 July 2021]). *Sistema de Información Gerencial de Salud (SIGSA) - enfermedades transmitidas por vectores, años 2012 al 2019*. Available from: https://sigsa.mspas.gob.gt/datos-de-salud/morbilidad/enfermedades-transmitidas-por-vectores.

Hahn, Micah B. et al. (2016). "Modeling the Geographic Distribution of Ixodes Scapularis and Ixodes Pacificus (Acari: Ixodidae) in the Contiguous United States". *Journal of Medical Entomology* 53.5, pp. 1176–1191.

Hanley, John et al. (2020). "Novel Evolutionary Algorithm Identifies Interactions Driving Infestation of Triatoma Dimidiata, a Chagas Disease Vector". *The American Journal of Tropical Medicine and Hygiene* 103.2, pp. 735–744.

Harrison, Xavier A. et al. (2018). "A Brief Introduction to Mixed Effects Modelling and Multi-Model Inference in Ecology". *PeerJ* 6, e4794.

Hassett, Erin et al. (2022). "Integrating Tick Density and Park Visitor Behaviors to Assess the Risk of Tick Exposure in Urban Parks on Staten Island, New York". *BMC public health* 22.1, p. 1602.

Herzog, Sereina A., Stéphanie Blaizot, and Niel Hens (2017). "Mathematical Models Used to Inform Study Design or Surveillance Systems in Infectious Diseases: A Systematic Review". *BMC Infectious Diseases* 17.1, p. 775.

Hines, Keegan E., Thomas R. Middendorf, and Richard W. Aldrich (2014). "Determination of Parameter Identifiability in Nonlinear Biophysical Models: A Bayesian Approach". *Journal of General Physiology* 143.3, pp. 401–416.

Hodges, James S. (2016). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. Boca Raton: CRC Press.

Horton, Richard (2021). *The COVID-19 Catastrophe: What's Gone Wrong and How To Stop It Happening Again*. 2nd ed. Cambridge: Polity Press.

Hotez, Peter J. et al. (2008). "The Neglected Tropical Diseases of Latin America and the Caribbean: A Review of Disease Burden and Distribution and a Roadmap for Control and Elimination". *PLoS neglected tropical diseases* 2.9, e300.

Johnson, Tammi L. et al. (2017). "Prevalence and Diversity of Tick-Borne Pathogens in Nymphal Ixodes Scapularis (Acari: Ixodidae) in Eastern National Parks". *Journal of Medical Entomology* 54.3, pp. 742–751.

Joshi, M., A. Seidel-Morgenstern, and A. Kremling (2006). "Exploiting the Bootstrap Method for Quantifying Parameter Confidence Intervals in Dynamical Systems". *Metabolic Engineering* 8.5, pp. 447–455.

Justi, Silvia et al. (2018). "Vectors of Diversity: Genome Wide Diversity across the Geographic Range of the Chagas Disease Vector Triatoma Dimidiata Sensu Lato (Hemiptera: Reduviidae)". *Molecular Phylogenetics and Evolution* 120.December 2017, pp. 144–150.

Kabaghe, Alinune N. et al. (2017). "Adaptive Geostatistical Sampling Enables Efficient Identification of Malaria Hotspots in Repeated Cross-Sectional Surveys in Rural Malawi". *PLoS ONE* 12.2, pp. 1–14.

King, Raymond et al. (2011). "Triatoma Dimidiata Infestation in Chagas Disease Endemic Regions of Guatemala: Comparison of Random and Targeted Cross-Sectional Surveys". *PLoS Neglected Tropical Diseases* 5.4. Ed. by Ricardo E. Gürtler, e1035.

Kleinegesse, Steven, Christopher Drovandi, and Michael U. Gutmann (2020). "Sequential Bayesian Experimental Design for Implicit Models via Mutual Information". *arXiv:2003.09379 [cs, stat]*. arXiv: 2003.09379 [cs, stat].

Kraemer, Moritz U. G. et al. (2021). "Data Curation during a Pandemic and Lessons Learned from COVID-19". *Nature Computational Science* 1.1, pp. 9–10.

Krausch, Niels et al. (2019). "Monte Carlo Simulations for the Analysis of Non-linear Parameter Confidence Intervals in Optimal Experimental Design". *Frontiers in Bioengineering and Biotechnology* 7.

Kugeler, Kiersten J. and Rebecca J. Eisen (2020). "Challenges in Predicting Lyme Disease Risk". *JAMA Network Open* 3.3, e200328.

Kugeler, Kiersten J., Grace M. Farley, et al. (2015). "Geographic Distribution and Expansion of Human Lyme Disease, United States". *Emerging Infectious Diseases* 21.8, pp. 1455–1457.

Lam, Nicholas N., Paul D. Docherty, and Rua Murray (2022). "Practical Identifiability of Parametrised Models: A Review of Benefits and Limitations of Various Approaches". *Mathematics and Computers in Simulation* 199, pp. 202–216.

Lessler, J. et al. (2015). "Seven Challenges for Model-Driven Data Collection in Experimental and Observational Studies". *Epidemics*. Challenges in Modelling Infectious DIsease Dynamics 10, pp. 78–82.

Levins, Richard (1966). "The Strategy of Model Building in Population Biology". *American Scientist* 54.4, pp. 421–431. JSTOR: 27836590.

Lima-Cordón, Raquel Asunción et al. (2018). "Implementation Science: Epidemiology and Feeding Profiles of the Chagas Vector Triatoma Dimidiata Prior to Ecohealth Intervention for Three Locations in Central America". *PLOS Neglected Tropical Diseases* 12.11. Ed. by Pamela Marie Pennington, e0006952.

Lindgren, Finn and Håvard Rue (2015). "Bayesian Spatial Modelling with R-INLA". *Journal of Statistical Software* 63, pp. 1–25.

Lindley, D. V. (1956). "On a Measure of the Information Provided by an Experiment". *The Annals of Mathematical Statistics* 27.4, pp. 986–1005.

Lippi, Catherine A. et al. (2021). "Scoping Review of Distribution Models for Selected Amblyomma Ticks and Rickettsial Group Pathogens". *PeerJ* 9, e10596.

Lucero, David et al. (2013). "Ecohealth Interventions Limit Triatomine Reinfestation Following Insecticide Spraying in La Brea, Guatemala". *American Journal of Tropical Medicine and Hygiene* 88.4, pp. 630–637.

Mader, Emily M et al. (2021). "A Survey of Tick Surveillance and Control Practices in the United States". *Journal of Medical Entomology* 58.4, pp. 1503–1512.

Magori, Krisztian and John M Drake (2013). "The population dynamics of vector-borne diseases". *Nature Education Knowledge* 4.4, p. 14.

Manne, Jennifer et al. (2012). "Triatomine Infestation in Guatemala: Spatial Assessment after Two Rounds of Vector Control". *American Journal of Tropical Medicine and Hygiene* 86.3, pp. 446–454.

Martcheva, Maia (2015). *An Introduction to Mathematical Epidemiology*. New York: Springer.

Martins, Thiago G et al. (2013). "Bayesian computing with INLA: new features". *Computational Statistics & Data Analysis* 67, pp. 68–83.

Matérn, B. (1960). *Spatial variation. PhD Thesis, Stockholm University.*
Mattingly, Henry H. et al. (2018). "Maximizing the Information Learned from Finite Data Selects a Simple Model". *Proceedings of the National Academy of Sciences* 115.8, pp. 1760–1765.
McCullagh, Peter (2002). "What Is a Statistical Model?" *The Annals of Statistics* 30.5.
Mead, P. et al. (2018). "Risk Factors for Tick Exposure in Suburban Settings in the Northeastern United States". *Ticks and Tick-Borne Diseases* 9.2, pp. 319–324.
Melikechi, Omar et al. (2022). "Limits of Epidemic Prediction Using SIR Models". *Journal of Mathematical Biology* 85.4, p. 36.
Meyer, Ruth K. and Christopher J. Nachtsheim (1995). "The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs". *Technometrics* 37.1, pp. 60–69.
Monroy, Carlota, Dulce Maria Bustamante, et al. (2009). "House Improvements and Community Participation in the Control of Triatoma Dimidiata Re-Infestation in Jutiapa, Guatemala". *Cadernos De Saude Publica* 25 Suppl 1, S168–178.
Monroy, Carlota, Mildred Mejia, et al. (1998). "Cornparison of Indoor Searches with Whole House Demolition Collections of the Vectors of Chagas Disease and Their Indoor Distribution". *Medical Entomology and Zoology* 49.3, pp. 195–200.
Monroy, Maria Carlota et al. (2003). "Habitats, Dispersion and Invasion of Sylvatic Triatoma Dimidiata (Hemiptera: Reduviidae: Triatominae) in Petén, Guatemala". *Journal of Medical Entomology* 40.6, pp. 800–806.
Nguyen, Van Kinh et al. (2016). "Analysis of Practical Identifiability of a Viral Infection Model". *PLOS ONE* 11.12, e0167568.
O'Neill, Philip D. and G. O. Roberts (1999). "Bayesian Inference for Partially Observed Stochastic Epidemics". *Journal of the Royal Statistical Society Series A: Statistics in Society* 162.1, pp. 121–129.
Overstall, Antony M., David C. Woods, and Ben M. Parker (2020). "Bayesian Optimal Design for Ordinary Differential Equation Models With Application in Biological Science". *Journal of the American Statistical Association* 115.530, pp. 583–598. arXiv: 1509.04099.
Pagès, Jérôme (2004). "Analyse factorielle multiple de données mixtes: principe et exemple d'application". *Revue Statistique Appliquee* 52.4, pp. 93–111.
Parent, Eric and Etienne Rivot (2012). *Introduction to Hierarchical Bayesian Modeling for Ecological Data.* Boca Raton: CRC Press.
Penados, Daniel et al. (2020). "Infestation Dynamics of Triatoma Dimidiata in Highly Deforested Tropical Dry Forest Regions of Guatemala". *Memorias do Instituto Oswaldo Cruz* 115.10, pp. 1–8.
Pepin, Kim M. et al. (2012). "Geographic Variation in the Relationship between Human Lyme Disease Incidence and Density of Infected Host-Seeking Ixodes Scapu-

laris Nymphs in the Eastern United States". *The American Journal of Tropical Medicine and Hygiene* 86.6, pp. 1062–1071.

Perfecto, Ivette and John Vandermeer (2008). "Biodiversity Conservation in Tropical Agroecosystems". *Annals of the New York Academy of Sciences* 1134.1, pp. 173–200.

Peters, Madeline A. E., Megan A. Greischar, and Nicole Mideo (2021). "Challenges in Forming Inferences from Limited Data: A Case Study of Malaria Parasite Maturation". *Journal of The Royal Society Interface* 18.177, p. 20210065.

Peterson, Jennifer et al. (2019). "Chagas Disease in Central America: Recent Findings and Current Challenges in Vector Ecology and Control". *Current Tropical Medicine Reports* 6.2, pp. 76–91.

Peterson, Jennifer K. et al. (2019). "Chagas Disease Epidemiology in Central America: An Update". *Current Tropical Medicine Reports* 6.2, pp. 92–105.

Piazzola, Chiara, Lorenzo Tamellini, and Raúl Tempone (2021). "A Note on Tools for Prediction under Uncertainty and Identifiability of SIR-like Dynamical Systems for Epidemiology". *Mathematical Biosciences* 332, p. 108514.

Poletto, Chiara, Samuel V. Scarpino, and Erik M. Volz (2020). "Applications of Predictive Modelling Early in the COVID-19 Epidemic". *The Lancet Digital Health* 2.10, e498–e499.

Pronzato, Luc and Eric Walter (1985). "Robust Experiment Design via Stochastic Approximation". *Mathematical Biosciences* 75.1, pp. 103–120.

Quinn, Katherine N. et al. (2021). "Information Geometry for Multiparameter Models: New Perspectives on the Origin of Simplicity". *arXiv:2111.07176 [cond-mat, physics:physics]*. arXiv: 2111.07176 [cond-mat, physics:physics].

R Core Team (2021). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rainforth, Tom et al. (2023). "Modern Bayesian Experimental Design". *arXiv:2302.14545 [cs, stat]*. arXiv: 2302.14545 [cs, stat].

Raman, Dhruva V., James Anderson, and Antonis Papachristodoulou (2017). "Delineating Parameter Unidentifiabilities in Complex Models". *Physical Review E* 95.3, p. 032314.

Ramirez-Sierra, Maria Jesus et al. (2010). "Patterns of House Infestation Dynamics by Non-Domiciliated Triatoma Dimidiata Reveal a Spatial Gradient of Infestation in Rural Villages and Potential Insect Manipulation by Trypanosoma Cruzi". *Tropical Medicine and International Health* 15.1, pp. 77–86.

Raue, A., V. Becker, et al. (2010). "Identifiability and Observability Analysis for Experimental Design in Nonlinear Dynamical Models". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 20.4, p. 045105.

Raue, A., C. Kreutz, et al. (2009). "Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood". *Bioinformatics* 25.15, pp. 1923–1929.

Reich, Brian J., Krishna Pacifici, and Jonathan W. Stallings (2018). "Integrating Auxiliary Data in Optimal Spatial Design for Species Distribution Modelling". *Methods in Ecology and Evolution* 9.6, pp. 1626–1637.

Reich, Nicholas G., Justin Lessler, et al. (2022). "Collaborative Hubs: Making the Most of Predictive Epidemic Modeling". *American Journal of Public Health* 112.6, pp. 839–842.

Reich, Nicholas G., Trish M. Perl, et al. (2011). "Visualizing Clinical Evidence: Citation Networks for the Incubation Periods of Respiratory Viral Infections". *PLOS ONE* 6.4, e19496.

Restif, Olivier et al. (2012). "Model-Guided Fieldwork: Practical Guidelines for Multidisciplinary Research on Wildlife Ecological and Epidemiological Dynamics". *Ecology Letters* 15.10, pp. 1083–1094.

Roda, Weston C. et al. (2020). "Why Is It Difficult to Accurately Predict the COVID-19 Epidemic?" *Infectious Disease Modelling* 5, pp. 271–281.

Roosa, Kimberlyn and Gerardo Chowell (2019). "Assessing Parameter Identifiability in Compartmental Dynamic Models Using a Computational Approach: Application to Infectious Disease Transmission Models". *Theoretical Biology and Medical Modelling* 16.1, p. 1.

Rosenberg, Ronald et al. (2018). "Vital Signs: Trends in Reported Vectorborne Disease Cases — United States and Territories, 2004–2016". *Morbidity and Mortality Weekly Report* 67.17, pp. 496–501.

Royle, J. A (2002). "Exchange Algorithms for Constructing Large Spatial Designs". *Journal of Statistical Planning and Inference* 100.2, pp. 121–134.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392.

Ryan, Elizabeth G., Christopher C. Drovandi, James M. McGree, et al. (2016). "A Review of Modern Computational Algorithms for Bayesian Optimal Design". *International Statistical Review* 84.1, pp. 128–154.

Ryan, Elizabeth G., Christopher C. Drovandi, and Anthony N. Pettitt (2015). "Simulation-Based Fully Bayesian Experimental Design for Mixed Effects Models". *Computational Statistics & Data Analysis* 92, pp. 26–39.

Rynkiewicz, Evelyn C. and Keith Clay (2014). "Tick Community Composition in Midwestern US Habitats in Relation to Sampling Method and Environmental Conditions". *Experimental and Applied Acarology* 64.1, pp. 109–119.

Schofield, Christopher John (1994). *Triatominae: biology & control*. Bognor Regis: Eurocommunica Publications.

Schulze, Terry L., Robert A. Jordan, and Robert W. Hung (1997). "Biases Associated with Several Sampling Methods Used to Estimate Abundance of Ixodes Scapularis and Amblyomma Americanum (Acari: Ixodidae)". *Journal of Medical Entomology* 34.6, pp. 615–623.

Seber, G. A. F and C. J. Wild (2003). *Nonlinear Regression*. 2nd ed. Hoboken: John Wiley & Sons.

Sharp, Jesse et al. (2022). "Parameter Estimation and Uncertainty Quantification Using Information Geometry". *Journal of The Royal Society Interface* 19, p. 20210940.

Shea, Katriona et al. (2020). "Harnessing Multiple Models for Outbreak Management". *Science* 368.6491, pp. 577–579.

Sloman, Sabina J. et al. (2022). "Characterizing the Robustness of Bayesian Adaptive Experimental Designs to Active Learning Bias". *arXiv:2205.13698 [stat]*. arXiv: 2205.13698 [stat].

Sonenshine, Daniel E. (2018). "Range Expansion of Tick Disease Vectors in North America: Implications for Spread of Tick-Borne Disease". *International Journal of Environmental Research and Public Health* 15.3, p. 478.

Spiegelhalter, David J. et al. (2002). "Bayesian Measures of Model Complexity and Fit". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4, pp. 583–639.

Springer, Yuri P., Lars Eisen, et al. (2014). "Spatial Distribution of Counties in the Continental United States With Records of Occurrence of Amblyomma Americanum (Ixodida: Ixodidae)". *Journal of Medical Entomology* 51.2, pp. 342–351.

Springer, Yuri P., Catherine S. Jarnevich, et al. (2015). "Modeling the Present and Future Geographic Distribution of the Lone Star Tick, Amblyomma Americanum (Ixodida: Ixodidae), in the Continental United States". *The American Journal of Tropical Medicine and Hygiene* 93.4, pp. 875–890.

Stevens, Lori, Carlota Monroy, Antonieta Guadalupe Rodas, and Patricia Dorn (2014). "Hunting, Swimming, and Worshiping: Human Cultural Practices Illuminate the Blood Meal Sources of Cave Dwelling Chagas Vectors (Triatoma Dimidiata) in Guatemala and Belize". *PLoS Neglected Tropical Diseases* 8.9, e3047.

Stevens, Lori, Carlota Monroy, Antonieta Guadalupe Rodas, Robin Hicks, et al. (2015). "Migration and Gene Flow among Domestic Populations of the Chagas Insect Vector Triatoma Dimidiata (Hemiptera: Reduviidae) Detected by Microsatellite Loci". *Journal of Medical Entomology* 52.3, pp. 419–428.

Sturrock, Hugh J. W. et al. (2011). "Planning Schistosomiasis Control: Investigation of Alternative Sampling Strategies for Schistosoma Mansoni to Target Mass Drug Administration of Praziquantel in East Africa". *International Health* 3.3, pp. 165–175.

Subramanian, Rahul, Qixin He, and Mercedes Pascual (2021). "Quantifying Asymptomatic Infection and Transmission of COVID-19 in New York City Using Observed Cases, Serology, and Testing Capacity". *Proceedings of the National Academy of Sciences of the United States of America* 118.9, e2019716118.

Tönsing, Christian, Jens Timmer, and Clemens Kreutz (2018). "Profile Likelihood-Based Analyses of Infectious Disease Models". *Statistical Methods in Medical Research* 27.7, pp. 1979–1998.

Transtrum, Mark K., Benjamin B. Machta, and James P. Sethna (2011). "Geometry of Nonlinear Least Squares with Applications to Sloppy Models and Optimization". *Physical Review E* 83.3, p. 036701.

Traverso, Lucila et al. (2017). "Comparative and functional triatomine genomics reveals reductions and expansions in insecticide resistance-related gene families". *PLoS Neglected Tropical Diseases* 11.2, e0005313.

Tuncer, Necibe and Trang T. Le (2018). "Structural and Practical Identifiability Analysis of Outbreak Models". *Mathematical Biosciences* 299, pp. 1–18.

University, Oregan State (2021 [accessed 5 July 2022]). *PRISM Climate Group*. Available from: https://prism.oregonstate.edu.

Watson, Joe, James V. Zidek, and Gavin Shaddick (2019). "A General Theory for Preferential Sampling in Environmental Networks". *The Annals of Applied Statistics* 13.4, pp. 2662–2700.

Weeks, E. N. I. et al. (2013). "Risk Factors for Domestic Infestation by the Chagas Disease Vector, Triatoma Dimidiata in Chiquimula, Guatemala". *Bulletin of Entomological Research* 103.6, pp. 634–643.

Weiss, Howard (2013). "The SIR Model and the Foundations of Public Health". *Materials Matemàtics* 2013.3, p. 17.

Wisely, Samantha M. and Gregory E. Glass (2019). "Advancing the Science of Tick and Tick-Borne Disease Surveillance in the United States". *Insects* 10.10, p. 361.

World Health Organization (2015). "Chagas disease in Latin America: an epidemiological update based on 2010 estimates". *Weekly Epidemiological Record* 90.06, pp. 33–44.

— (2021). "Ending the neglect to attain the sustainable development goals: a sustainability framework for action against neglected tropical diseases 2021-2030". *Geneva: World Health Organization* Licence: CC BY-NC-SA 3.0 IGO.

— (2021 [cited 20 September 2021]). *Chagas disease (American trypanosomiasis)*. Available from: https://www.who.int/health-topics/chagas-disease.

Wu, Joseph T., Kathy Leung, and Gabriel M. Leung (2020). "Nowcasting and Forecasting the Potential Domestic and International Spread of the 2019-nCoV Outbreak Originating in Wuhan, China: A Modelling Study". *Lancet (London, England)* 395.10225, pp. 689–697.

Wu, Sean L. et al. (2020). "Substantial Underestimation of SARS-CoV-2 Infection in the United States". *Nature Communications* 11.1, p. 4507.

Yamagata, Yoichi and Jun Nakagawa (2006). "Control of Chagas Disease". *Advances in Parasitology* 61, pp. 129–165.

Yoshioka, Kota, Ezequiel Provedor, and Jennifer Manne-Goehler (2018). "The Resilience of Triatoma Dimidiata: An Analysis of Reinfestation in the Nicaraguan Chagas Disease Vector Control Program (2010–2016)". *PLoS ONE* 13.8, pp. 1–18.

Zidek, James, Gavin Shaddick, and Carolyn Taylor (2014). "Reducing Estimation Bias in Adaptively Changing Monitoring Networks with Preferential Site Selection". *Annals of Applied Statistics* 8.3, pp. 1640–1670.

Zuur, Alain F and Elena N Ieno (2016). "A protocol for conducting and presenting results of regression-type analyses". *Methods in Ecology and Evolution* 7.6, pp. 636–645.