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

January 8, 2024

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

$$y_{ijt} \sim \text{Bernoulli}(r_{ijt}) \tag{1}$$

$$\text{logit}(r_{ijt}) = \eta_{ijt} = \mu_j + \sum_{k=1}^{K} f_{kj}(x_{kit}) + s_{ij} + m_{jt} \tag{2}$$

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

We assume $\mu_j \sim N(0, 5)$ priors for the intercepts. 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 [1].[1]

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

[1] For the categorical variable `land_cover`, only a linear form was considered using the standard dummy-variable approach.

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

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 (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 $\eta_{ijt} = \mu_j + \sum_k \beta_{kj} x_{kit} + s_i$. Combining these simplifications results in 28 candidate models, and models are compared using the Deviance Information Criterion. All models are fit in R version 4.2.2 using R-INLA version 23.02.27 [2].

Results from the model comparison study are shown in Figure S1. 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 the remainder of this work has linear predictor

$$\eta_{ijt} = \mu_j + \sum_k \beta_k x_{kit} + s_{ij} + m_{ij}. \tag{3}$$

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,

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

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

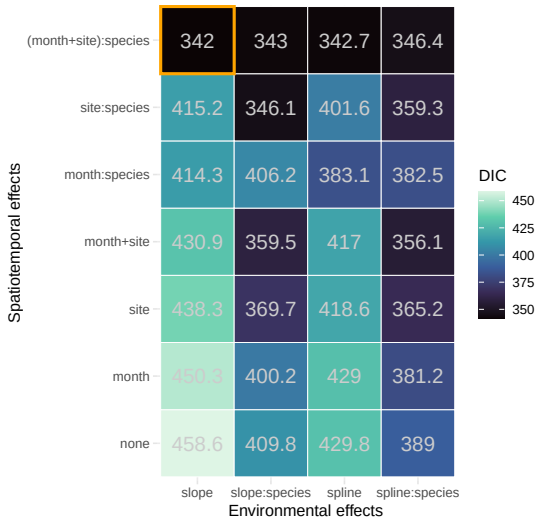

Figure S1: Deviance information criterion of different mixed-effects models, fit to the initial survey data. Each tile indicates a model comprised of the corresponding environmental and spatiotemporal effects, shared or independent between tick species. "Slope" indicates a linear f for each environmental variable, and "spline" indicates nonlinear f. The best-ranked model (lowest DIC) is highlighted in orange.

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 $\mathbf{y}$,

$$U(\mathbf{d}) = \sum_{\mathbf{y}} U(\mathbf{d}, \mathbf{y}) P(\mathbf{y} \mid \mathbf{y}_{\text{init}}, \mathbf{d}), \tag{4}$$

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 $\mathbf{d}$.

4 Description of search algorithms

Once a design criterion is chosen, the goal is to find some $\mathbf{d}$ with as close to optimal utility as possible. An optimal design for function 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 (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 [3]. 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*. T_0 controls the orders of magnitude worse a proposal can be to still have a good chance of acceptance at the beginning of the algorithm, and should be set relative to the magnitude of differences between values on the utility surface. A larger T_0 leads to more exploratory behavior. Curvature α controls how quickly the acceptance probability decreases from T_0, so that $\alpha < 1$ leads to more exploratory behavior. 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 [4, 5]. 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 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.

References

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

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

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

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

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