Journal of Data Science 23 (3), 542–559
July 2025
DOI: 10.6339/25-JDS1191
Statistical Data Science

Estimating Disease Prevalence from Preferentially Sampled, Pooled Data

Clinton P. Pollock[1,*], Andrew Hoegh[1], Kathryn M. Irvine[2], Luz A. de Wit[3], and Brian E. Reichert[4]

[1]*Department of Mathematical Sciences, Montana State University, Bozeman, MT, USA*
[2]*Northern Rocky Mountain Science Center, U.S. Geological Survey, Bozeman, MT, USA*
[3]*Bat Conservational International, Austin, TX, USA*
[4]*Fort Collins Science Center, U.S. Geological Survey, Fort Collins, CO, USA*

Abstract

After the onset of the COVID-19 pandemic, scientific interest in coronaviruses endemic in animal populations has increased dramatically. However, investigating the prevalence of disease in animal populations across the landscape, which requires finding and capturing animals can be difficult. Spatial random sampling over a grid could be extremely inefficient because animals can be hard to locate, and the total number of samples may be small. Alternatively, preferential sampling, using existing knowledge to inform sample location, can guarantee larger numbers of samples, but estimates derived from this sampling scheme may exhibit bias if there is a relationship between higher probability sampling locations and the disease prevalence. Sample specimens are commonly grouped and tested in pools which can also be an added challenge when combined with preferential sampling. Here we present a Bayesian method for estimating disease prevalence with preferential sampling in pooled presence-absence data motivated by estimating factors related to coronavirus infection among Mexican free-tailed bats (*Tadarida brasiliensis*) in California. We demonstrate the efficacy of our approach in a simulation study, where a naive model, not accounting for preferential sampling, returns biased estimates of parameter values; however, our model returns unbiased results regardless of the degree of preferential sampling. Our model framework is then applied to data from California to estimate factors related to coronavirus prevalence. After accounting for preferential sampling impacts, our model suggests small prevalence differences between male and female bats.

Keywords *Bayesian modeling; pooled testing; spatial sampling*

1 Introduction

Due to the COVID-19 pandemic there has been a growing interest in the potential for several bat species to be reservoirs of zoonotic pathogens (Rahman et al. 2020) and relatedly there is particular interest in the transmission dynamics of coronaviruses within bat populations (Meyer et al. 2024; Alison J. Peel 2025) and the potential for bat-to-human and human-to-bat transmission of these viruses (Wong et al. 2007). An initial step in addressing these questions involves exploring rates of coronavirus prevalence in bats. Mexican free-tailed bats (*Tadarida brasiliensis*) are known to be susceptible to exposure to SARS-CoV-2 (Hall et al. 2023). Here we aim to

*Corresponding author. Email: clintonpollock@msu.montana.edu.

Received October 31, 2024; Accepted May 30, 2025

understand what environmental or bat-level factors are associated with coronavirus prevalence in Mexican free-tailed bats.

Zoonotic spillover refers to the spreading of a pathogen from vertebrate animals to humans (Plowright et al. 2017). The species that harbors a pathogen such that it can be permanently maintained and also infect a target population is known as a "reservoir host" (Haydon et al. 2002). Several bat species are reservoir hosts of different pathogens, most notably rabies, but are also speculated to have played a role in the spread of the MERS-CoV coronavirus by spreading the virus to camels who, in turn, spread the virus to humans (Mackenzie et al. 2016). Spillover is a complex, multifaceted process, but understanding the pathogen prevalence within the host's population can be the first step to untangling that process.

Modeling landscape wide coronavirus prevalence in Mexican free-tailed bats presents a statistical challenge because there is a relative scarcity of their roosts across North America, making probabilistic sampling of locations highly inefficient. Valid inferences for ecological and epidemiological processes that manifest across space and time, such as coronavirus prevalence in bats, require careful consideration of spatial sampling and statistical models that properly address temporal and spatial processes. Although sampling designs exist to collect spatially balanced data, such as the Generalized Random Tessellation Stratified (GRTS) (Stevens Jr and Olsen 2004; Talbert and Reichert 2018) design, in situations like this with hard to reach populations, data can be opportunistically collected but require statistical models that explicitly account for preferential sampling.

To estimate coronavirus prevalence, viral samples are collected from individuals at bat roosts across California that are selected by preferential sampling. While preferential sampling was necessary to obtain a sufficient sample size for inference, it can result in biased estimates (Conroy et al. 2023; Moreira, Menezes, and Wise 2024) of prevalence or of factors impacting prevalence. In particular, it is plausible bats found at these preferentially sampled locations would have different coronavirus prevalence than probabilistically sampled locations. For example, bat samples collected at roosts with certain land-use characteristics might have higher rates of prevalence. Consequently, a naïve model to estimate prevalence across the landscape, using these preferentially sampled roosts, would exhibit some bias by overestimating prevalence in the greater population.

In addressing data that may exhibit response bias, such as bias induced by spatial preferential sampling, two common approaches are employed: sample weighting to align with the population or modeling the biasing mechanism (Vedensky, Parker, and Holan 2023). The former is often employed in surveys where robust census information allows for effective weighting. In many contexts, however, this census information does not exist, which can make the strategy less effective. Such is the case in many aspects of ecology where the bias needs to be addressed through estimation. This requires assessing the severity of preferential sampling and then modeling the preferential sampling impacts on the response of interest.

A comprehensive framework for using statistical modeling to estimate and account for response bias from spatial preferential sampling is presented by Diggle, Menezes, and Su (2010). The core concept involves using a Log-Gaussian Cox (Møller, Syversveen, and Waagepetersen 1998) process with a spatial correlation matrix to model the intensity surface at the sampled locations. This spatial process related to the preferential sampling intensity surface is in addition to a general spatial process related to measurements closer in space being correlated. The spatial intensity surface estimated from the preferential sampling process is then used as a covariate for the response. The method proposed by Diggle, Menezes, and Su (2010) has some limitations,

particularly regarding computational efficiency. In response, Pati, Reich, and Dunson (2011) introduced a Bayesian implementation of the model that implements traditional Markov Chain Monte Carlo (MCMC) techniques in place of Maximum Likelihood Estimation (MLE) Monte Carlo.

In addition to the challenge presented by preferential sampling, our scenario focused on modeling coronavirus prevalence in Mexican free-tailed bats also has another challenge due to grouping individual specimens into pools for testing. In general, pooling or group testing (Du and Hwang 1999) can be an efficient way to screen more samples than testing all samples individually. These pooling techniques found favor in screening humans for coronaviruses during the height of the COVID-19 pandemic (Mallapaty et al. 2020; Warasi, Hungerford, and Lahmers 2022). However, much of this work assumes that positive pools will be retested (Bilder, Tebbs, and Chen 2010) to identify positive individuals. When samples in positive pools are not retested, estimating the proportion of individual bats that are infected can be done but requires modeling to account for this uncertainty. Hoegh et al. (2021) showed that pooling without retesting can result in more efficient use of resources and more precise statistical inference in simple cases. However, accounting for pooling without retesting in more complicated scenarios can be challenging and requires modifying existing algorithms.

Due to the high cost of coronavirus testing – and given that individual prevalence is expected to be low – specimens are grouped into pools and a positive or negative result is recorded for each pool. As it is not important to identify exactly which bats were positive, positive pools are not retested and an individual bat's infection status remains unknown. However, while this enabled more individual specimens to be tested it also requires addressing this extra uncertainty related to which bat(s) in a pool are positive while estimating the proportion of individual bats that are infected. Models do not exist to esimate the proportion of individual bats that are infected from this scenario with pooled sampling and preferential sampling. We propose a method of accounting for preferential sampling and simultaneously estimating coefficients related to site-level, pool-level, and individual-level factors from pooled samples.

2 Data

Data were collected from known and probable bat roosts across 20 sites in California (Figure 1). Throughout the summer and fall of 2022, field researchers visited these predetermined sites. This study was approved by the Animal Care and Use Committee of Northern Arizona University with approval number 22-004, and for field work occurring in U.S. Forest Service lands, the study was concurrently approved by the Animal Care and Use Committee of the U.S. Forest Service Research and Development with approval number 2022-012. A total of 394 bats were collected and placed in bags after which the researchers took specimen samples and various measurements on the individual bats including mass, wingspan, age, and sex. The data that support the findings in this study are available from Wray et al. (2025).

Once collected, these samples were sent to a laboratory for testing. They were grouped into pools to optimize cost efficiency and enable a larger number of bats to be tested. Typically, each pool combined the samples of four bats from the same site, sample type, and species. A two-stage process was used for determining whether a specimen was positive, where all presumptive positive samples were sequenced to confirm positive results. For purposes of the analysis, only the final results, post-sequencing, are used. For most of the positive pools, individual specimens were not retested.

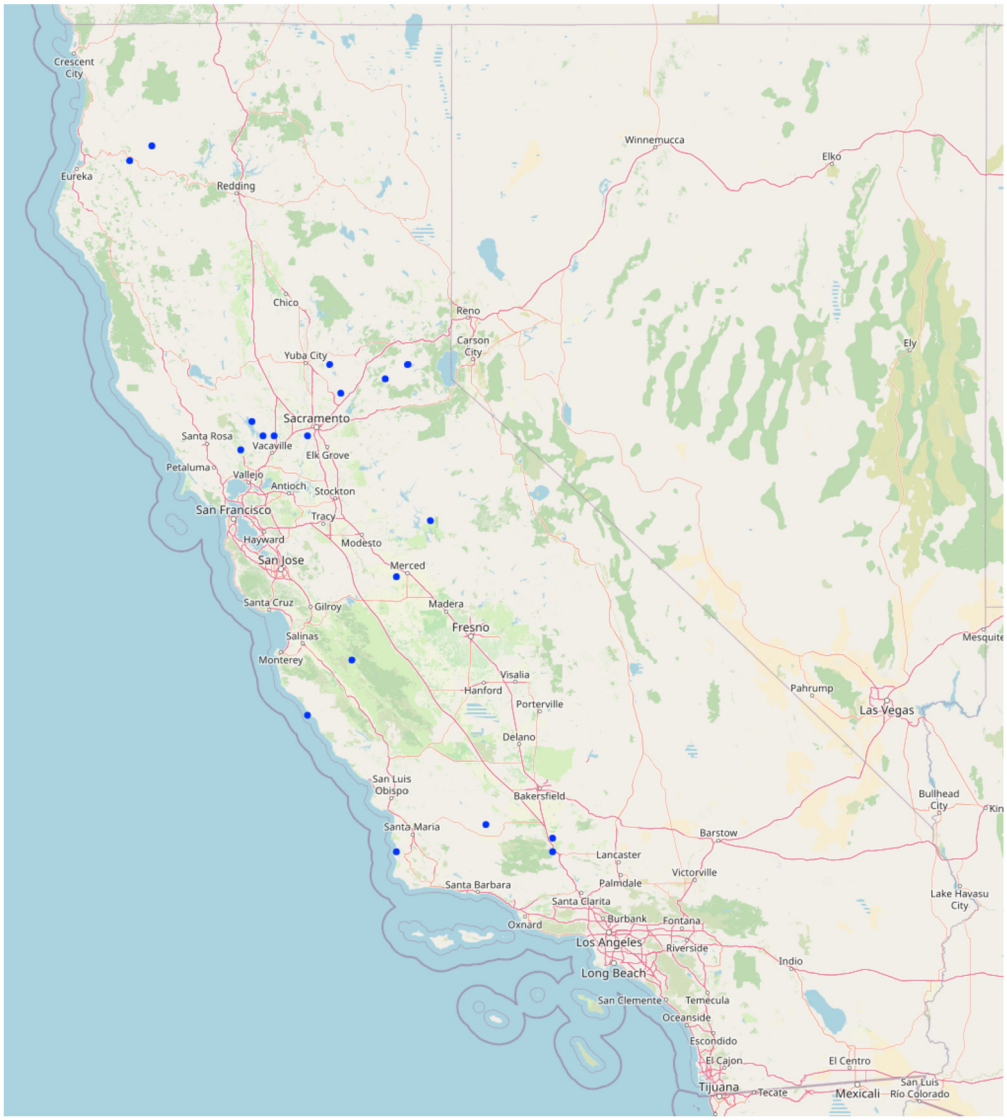


Figure 1: Locations of sites where bats were captured and samples were collected in California. Mexican free-tailed bats are migratory (seasonally). Spatial correlation between roosts is modeled using a spatial random effect. Map extracted from Open Street Map (OpenStreetMap contributors 2017) using leaflet (Cheng et al. 2025).

Consequently, for each bat sampled, we know the physical characteristics of the individual bat, the geographical coordinates of the sampling location, information on which other individuals are part of the same pool, and whether the pool to which it belongs tested positive or negative for an alpha-coronavirus. In addition to these individual-level data, site-level covariates related to land use and surrounding human population characteristics were extracted from the Google Earth Engine (Gorelick et al. 2017).

3 Methods

3.1 Gaussian and Shared Latent Processes

Modeling ecological and epidemiological processes, such as coronavirus prevalence in bats, requires spatial statistical models that properly account for the correlation structure of the observations collected near each other. The framework presented here relies on Gaussian processes to model this spatial correlation. A Gaussian process is a stochastic process which treats the multivariate normal distribution as dimensionally infinite such that it can be mapped to some smooth function in $\mathbb{R}^n$. For any pair of points s, s' in the function's domain, their covariance is given by the positive semi-definite function $C_\theta(s, s')$, and for any finite set of m points $\boldsymbol{S}$, the function's output is distributed as $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} \in \mathbb{R}^m$, $\Sigma = C_\theta(\boldsymbol{S})$. To distinguish between continuous functions and their finite realizations, bold typeface will be used for finite sets and vectors, and standard typeface will be used for scalars, arbitrary function inputs, vector or matrix elements, and matrices. For spatial problems, it is common to include spatial random effects that are modeled as a Gaussian process in $\mathbb{R}^2$, where $C_\theta(s, s')$ decays with Euclidean distance.

To address bias from preferential sampling, Diggle, Menezes, and Su (2010) proposed using a spatial covariate to jointly model the probability of a site being selected in addition to its influence on the response. More precisely, given the spatial Gaussian process $w(\boldsymbol{s})$, the sampling intensity of the preferentially sampled sites $\boldsymbol{s}$ is modeled as inhomogeneous Poisson point process with intensity

$$\lambda(s) = \exp\{\alpha + \beta w(\boldsymbol{s})\}$$

and jointly the response

$$\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu} + w(\boldsymbol{s}), \sigma^2 I_n),$$

where $\boldsymbol{Y}$ would be a continuous response. As a result of this joint estimation, wherever $w(\boldsymbol{s})$ induces a larger response it also increases its sampling intensity, leading to larger values of $w(\boldsymbol{s})$ being sampled more often. In our setting, $\boldsymbol{Y}$ would be coronavirus prevalence, albeit modeled with a binomial framework. Hence, if $w(\boldsymbol{s})$ was large it would relate to higher likelihood of location $\boldsymbol{s}$ being sampled and a higher expected prevalence value at location $\boldsymbol{s}$.

Following Diggle, Menezes, and Su (2010), Pati, Reich, and Dunson (2011) extended the model, adapted it to a Bayesian approach, and addressed some computational challenges. First, they split the spatial component into two separate Gaussian processes and applied the same approach with the covariate vector $X\boldsymbol{\beta}$. They then added a coefficient α to the preferential components such that

$$\begin{aligned} \boldsymbol{Y} &\sim \mathcal{N}(\alpha X\boldsymbol{\beta}_w + X\boldsymbol{\beta}_d + \alpha w(\boldsymbol{s}) + d(\boldsymbol{s}), \sigma^2) \\ \lambda(\boldsymbol{s}) &= \exp\{X\boldsymbol{\beta}_w + w(\boldsymbol{s})\}, \end{aligned}$$

where $\boldsymbol{\beta}_w$ is the vector of coefficients on site level covariates, $\boldsymbol{\beta}_d$ is the vector of coefficients on bat level covariates, and $d(\boldsymbol{s})$ is a second spatial effect that is independent of selection. This enabled the modeling of spatial effects unrelated to sampling preference, the influence of sampling preference on other covariates' impact on the mean, and direct inference into the extent of preferential sampling, as the sign and magnitude of α imply different relationships (e.g. $\alpha > 0$ implies preference for larger values of $w(\boldsymbol{s}) + X\boldsymbol{\beta}_w$).

To approximate a Poisson point process, Pati, Reich, and Dunson (2011) implemented a grid approximation. This increased the model's computation time dramatically as $w(\boldsymbol{s})$ must be sampled for each grid cell. Sampling from an n-dimensional multivariate normal distribution has a time complexity of $O(n^3)$. Thus, for applications with even moderately sized grids, the full model can be computationally prohibitive. To account for this, Pati, Reich, and Dunson (2011) implemented a low-rank predictive process approximation in which the Gaussian process is evaluated at a smaller selection of locations called knots and then the full-rank values are imputed as their best linear unbiased prediction values given the lower rank estimates. That is, $\tilde{w}(\boldsymbol{s}) = C_\theta(\boldsymbol{s}, \boldsymbol{s}_{knot})\,[C_\theta(\boldsymbol{s}_{knot}, \boldsymbol{s}_{knot})]^{-1}\, w(\boldsymbol{s}_{knot})$ for $w^* \sim \mathcal{N}(\mathbf{0}, C_\theta(\boldsymbol{s}_{knot}, \boldsymbol{s}_{knot}))$. Note that $\boldsymbol{s}_{knot}$ is a vector thus $C_\theta(\boldsymbol{s}_{knot}, \boldsymbol{s}_{knot})$ returns a square matrix with spatial covariance. For very low-rank approximations without additional corrections, this method can lead to reduced resolution in the Gaussian process, which manifests as an inflated nugget variance when estimating the parameters in $C_\theta(s, s')$ (Banerjee 2017). Because of this limitation, we have instead chosen to implement a Vecchia approximation, which considers each site sequentially and induces sparsity by only conditioning on the spatially nearest m previously sampled points (Euclidean distance) (Vecchia 1988). By only conditioning on a small subset of points, computation is much more efficient, as sampling from multivariate normal distributions scales as $O(n^3)$. At the same time, as long as the number of points is sufficient and the order in which the points are sampled leads prior points to capture surrounding influence, the approximation error diminishes rapidly (Katzfuss and Guinness 2021).

For simulation and real data fitting, we set $m = 10$, as this value provides a balance between computational efficiency and approximation accuracy. For generating the data in the simulation study, we used $m = 100$ to ensure a highly reliable approximation of the full Gaussian process, as simulating the full process was computationally infeasible. To maximize the accuracy of the approximation, a maximin ordering of sites was used (Jimenez and Katzfuss 2023), which ensures that each new site in the ordering is as far as possible from the previously ordered sites.

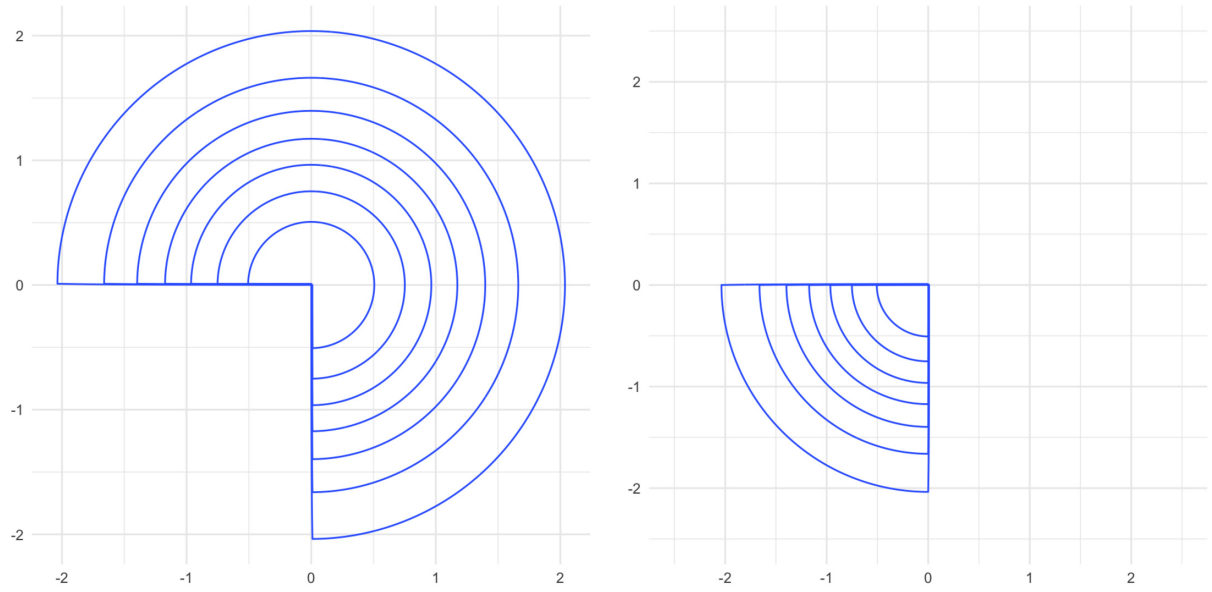
3.2 Bayesian Probit Regression and Latent Normal Random Variables

Albert and Chib (1993) observed that while the standard probit model,

$$\begin{aligned}\boldsymbol{Y} &\sim \text{Bin}(\boldsymbol{p})\\ \boldsymbol{p} &= \Phi(X\boldsymbol{\beta})\\ \boldsymbol{\beta} &\sim \mathcal{N}(\mu_\beta, \Sigma_\beta),\end{aligned}$$

yields a "largely intractable" posterior for $\boldsymbol{\beta}$, the same model could be rewritten as $Y_i = \mathbb{1}_{Z_i>0}$, $\boldsymbol{Z} \sim \mathcal{N}(X\boldsymbol{\beta}, I_n)$, where the full conditional distribution of $\boldsymbol{\beta}$ is a multivariate normal distribution and the full conditional of $\boldsymbol{Z}_i$ is a truncated normal distribution. Because both of these distributions can easily be sampled from, the exact posterior can be obtained through MCMC methods (Albert and Chib 1993). We propose a method of extending this computational convenience to pooled probit models that is, to our knowledge, novel. In addition to the computational efficiency obtained from the conjugacy in the Albert and Chib (1993) approach, our approach also enables individual-level covariates to be estimated, even though the positive or negative results are reported at the pool-level.

Suppose that instead of $Y_i = \mathbb{1}_{Z_i>0}$, $Y_i = \mathbb{1}_{\max(\boldsymbol{Z}_i)>0}$ for some pool $\boldsymbol{Z}_i$ of random variables Z_{ij}. As no other alterations are made, the full conditional for $\boldsymbol{\beta}$ is unchanged. Regarding $\boldsymbol{Z}$, it is useful to consider it partitioned by pool, such that $\boldsymbol{Z}_i \sim \mathcal{N}(X_i\boldsymbol{\beta}_i, 1)$ for a given pool i. In the unpooled



(a) $Y_i = 1$ (b) $Y_i = 0$

Figure 2: Full conditional contour plots for standard normal Z_{ij} with a pool size of 2.

case, the full conditional distribution for Z_i is truncated to the values that make Y_i possible, that is, $P(Z_i|\beta_i) \cdot (\mathbb{1}_{Z_i>0})^{Y_i} \cdot (\mathbb{1}_{Z_i\leqslant 0})^{1-Y_i}$. Similarly, the domain of the full conditional for the pooled $\boldsymbol{Z}_i$ is restricted to the values that make Y_i possible, $P(\boldsymbol{Z}_i|\boldsymbol{\beta}_i) \cdot (\mathbb{1}_{\max(\boldsymbol{Z}_i>0)})^{Y_i} \cdot (\mathbb{1}_{\max(\boldsymbol{Z}_i)\leqslant 0})^{1-Y_i}$. Note that $\max(\boldsymbol{Z}_i) \leqslant 0$ is equivalent to every Z_{ij} being negative or 0, and $\max(\boldsymbol{Z}_i) > 0$ is equivalent to at least one Z_{ij} being positive. Thus, in the pooled case, the full conditional for $\boldsymbol{Z}_i$ is restricted to the nonpositive orthant if $Y_i = 0$ (an n-dimensional generalization of quadrant III in $\mathbb{R}^2$) and to all other orthants if $Y_i = 1$. A visual depiction can be seen in Figure 2. For this application, our pool size is four or less. In general, the pool size is linked to the expected prevalence in the population (Warasi, Hungerford, and Lahmers 2022).

For numerical stability, the marginal probability that Z_{ij} is less than zero in the $Y_i = 1$ case can be written as

$$\frac{F_{Z_{i1}}(0)}{1 - \prod_{j=1}^{n_{pool}} F_{Z_{ij}}(0)} \cdot \frac{K}{K} = \exp\left(\log(K) + \log(F_{Z_{i1}}(0)) - \log\left(K - e^{\log(K) + \sum_{j=1}^{n_{pool}} \log(F_{Z_{ij}}(0))}\right)\right),$$

where $F_{Z_{ij}}(x)$ is the normal cumulative distribution function (cdf) with mean $\boldsymbol{X}'_{ij}\boldsymbol{\beta}_i$ and unit variance and K is some large constant. The full process is outlined in Algorithm 1. Modeling the pools with this framework allows for the model specified by Pati, Reich, and Dunson (2011) to be adapted to a pooled binary response with minimal changes.

3.3 Latent Probit Sampling

A consequence of the use of a Log-Gaussian Cox Process is the difficulty of obtaining posterior draws for the shared latent Gaussian process, $w(\boldsymbol{s})$. Pati, Reich, and Dunson (2011) implemented a block Metropolis-Hastings sampler, requiring a large amount of additional computation to repeatedly calculate the likelihood of the proposed draws. While the pooled probit framework conveniently adapts to a normal response, pooled binary data introduce additional uncertainty,

Algorithm 1 Pooled probit full conditional sampling.

Let $TN(\mu, \sigma^2, a, b)$ describe a truncated normal distribution with mean μ, variance σ^2, lower bound a, and upper bound b
$j \leftarrow 1$
repeat
  $r \leftarrow F_{Z_j}(0) \div \left(1 - \prod_{k=j}^{n} F_{Z_k}(0)\right)$
  $U \sim Unif(0, 1)$
  if $r < U$ **then**
    $Z \sim TN(\mu_j, 1, -\infty, 0)$
  else
    $Z \sim TN(\mu_j, 1, 0, \infty)$
  end if
  $Z_j \leftarrow Z$
  $j \leftarrow j + 1$
until $Z_{j-1} < 0$
while $j \leqslant n$ **do**
  $Z_j \leftarrow Z \sim N(\mu_j, 1)$
end while

as the exact outcomes of multiple normal random variables are replaced with a single Bernoulli random variable for the pool. This additional uncertainty can lead to highly correlated posterior draws, slowing exploration of the sample space because this model requires a large number of draws and a large amount of computation per draw.

To address this, we replaced the inhomogeneous Poisson point process instead modeling sampled points as coming from a fine grid of Bernoulli trials. Similar to the response, initial selection probabilities are given by a probit link such that $v(s_m) \sim N(X_{vm}^{\top}\boldsymbol{\beta}_v + \alpha_v w(s_m), 1)$ for some set of site-level covariates X_v and coefficients β_v and α_v. As with the inhomogeneous Poisson point process models, this allows the model to correct for spatial sampling bias as $w(\boldsymbol{s})$ influences both the response and the probability of selection. However, unlike the inhomogeneous Poisson point process models, this produces a full conditional for $w(\boldsymbol{s})$ that can be sampled via Gibbs sampling (see Equation 3).

A weakness of this strategy is that it can be sensitive to the choice of grid resolution. A well-known feature of the Poisson distribution is that, in the homogeneous case, if two adjacent grid cells each have a Poisson distribution with mean $\lambda/2$, then the mean for the combined plot (the two subplots together) is λ. In contrast, if we consider two sites, s_1, s_2 with some feature $\boldsymbol{x}$ such that $P(s_k = 1) = E(s_k) = \Phi(x_k)$, the same property does not hold as in the Poisson distribution. To remedy this, a second stage in selection was added. For the sites that were selected in the first stage, a second Bernoulli trial is performed, each with the same probability, π. Then, for $\pi = 0.5$, $E(s_k)/2 = \Phi(x_k) \cdot \pi$.

3.4 Complete Bat Model

The complete model, including the Vecchia approximations $\tilde{w}(s)$ and $\tilde{d}(s)$, is specified as

$$\begin{aligned} Y_{ij}|\boldsymbol{Z}_{ij} &= \mathbb{1}_{\max(Z_{ijk})>0} \\ \boldsymbol{Z}|X, \boldsymbol{\beta}, \alpha, U, V, \tilde{w}(\boldsymbol{s}), \tilde{d}(\boldsymbol{s}) &\sim \mathcal{N}(X\boldsymbol{\beta} + \alpha UV\tilde{w}(\boldsymbol{s}) + UV\tilde{d}(\boldsymbol{s}), I) \end{aligned} \tag{1}$$

$$
\begin{aligned}
U &\text{ is a replication matrix for the bats in each site} \\
V &\text{ is a selection matrix for the nonzero entries of } \boldsymbol{r} \\
r_i &= \mathbb{1}_{0<v_k, p_k=1} \\
p_i|\pi &\sim \text{Bernoulli}(\pi) \\
\boldsymbol{v}|X_v, \boldsymbol{\beta}_v, \alpha_v, \tilde{w}(\boldsymbol{s}) &\sim \mathcal{N}(X_v\boldsymbol{\beta}_v + \alpha_v\tilde{w}(\boldsymbol{s}), I) \\
\tilde{d}(s_k)|\boldsymbol{B}_{dk}, \boldsymbol{s}_{g_d(k)}, \sigma_d^2, H_{dk} &\sim N(\boldsymbol{B}_{dk}\boldsymbol{s}_{g_d(k)}, \sigma_d^2 H_{dk}) \\
\tilde{w}(s_k)|\boldsymbol{B}_{wk}, \boldsymbol{s}_{g_w(k)}, H_{dk} &\sim N(\boldsymbol{B}_{wk}\boldsymbol{s}_{g_w(k)}, H_{wk}) \\
\boldsymbol{B}_{qk}|s_k, \boldsymbol{s}_{g_q(k)} &= C_{\theta_q}(s_k, \boldsymbol{s}_{g_q(k)})C_{\theta_q}(\boldsymbol{s}_{g_q(k)}, \boldsymbol{s}_{g_q(k)})^{-1} \\
H_{qk}|s_k, \boldsymbol{s}_{g_q(k)}, \boldsymbol{B}_{qk} &= C_{\theta_q}(s_k, s_k) - \boldsymbol{B}_{qk}C_{\theta_q}(\boldsymbol{s}_{g_q(k)}, s_k),
\end{aligned}
$$

where X is the design matrix of all bat and site level covariates for each bat, X_v is the design matrix of only site-level covariates, and ijk indexes the kth bat in the jth pool from the ith site. $\boldsymbol{B}_{qk}$ and H_{qk} are the block matrices $\boldsymbol{B}_i$ and $\boldsymbol{D}_i$ outlined in Katzfuss and Guinness (2021) for the Vecchia approximation for $q \in \{w, s\}$. Each block corresponds to a location s_k given a conditioning vector $\boldsymbol{s}_{g_q(k)}$. Because, in this case, each block corresponds to a single location, each H_{qk} represents a scalar value and each $\boldsymbol{B}_{qk}$ represents a transposed vector of interpolation weights.

Because the Gaussian processes $w(\boldsymbol{s})$ and $d(\boldsymbol{s})$, along with their approximations, are based on site-level covariates, their entries need to be replicated for each bat at a single site. Note that more complicated spatial structures, such as anisotropy, could be encoded in $d(\boldsymbol{s})$. This is achieved using the replication matrix U, which uses basis vectors corresponding to sites and repeats them once for each bat at those corresponding sites. Additionally, as most of the sites are unobserved, a second selection matrix V removes values of $\tilde{w}(s)$ and $\tilde{d}(s)$ that correspond to those unobserved sites. As an example, if there were three total sites, the first and third sites were selected, and there were three samples collected at the first site and two at the second, then

$$
U = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, V = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tilde{d}(\boldsymbol{s}) = \begin{bmatrix} \tilde{d}(s_1) \\ \tilde{d}(s_2) \\ \tilde{d}(s_3) \end{bmatrix}
$$

$$
V\tilde{d}(\boldsymbol{s}) = \begin{bmatrix} \tilde{d}(s_1) \\ \tilde{d}(s_3) \end{bmatrix}, \text{ and } UV\tilde{d}(\boldsymbol{s}) = \begin{bmatrix} \tilde{d}(s_1) \\ \tilde{d}(s_1) \\ \tilde{d}(s_1) \\ \tilde{d}(s_3) \\ \tilde{d}(s_3) \end{bmatrix}.
$$

For a covariance function for both $w(\boldsymbol{s})$ and $d(\boldsymbol{s})$ we use

$$
C_{\theta_q}(s, s') = \exp\left\{\frac{-D(s, s')}{\phi_q}\right\}, \tag{2}
$$

where $D(s, s')$ is the Euclidean distance between the points s, s'. Although if an application warranted a more complex spatial random effect, that could be specified with a different covariance function. Care should be taken when specifying the priors for the covariance functions of $w(\boldsymbol{s})$

and $d(\boldsymbol{s})$ as these parameters are generally weakly identified. In addition, the prior on α_v may need to constrain the parameter space to positive values to avoid bimodality in the posterior. However, this is not necessary if α, α_v and $\tilde{w}(\boldsymbol{s})$ are individually treated as nuisance parameters. In general, interest would be in $\alpha \times w(\boldsymbol{s})$ and $\alpha_v \times w(\boldsymbol{s})$, rather than the α terms directly, and the products do not exhibit bimodality.

In the model, $\tilde{w}(\boldsymbol{s})$ appears in the conditional likelihoods of both the latent response vector $\boldsymbol{Z}$ and the latent selection vector $\boldsymbol{v}$. Katzfuss and Guinness (2021) demonstrated that the individual block matrices can be used to construct a vector $\boldsymbol{\mu}_{\tilde{w}}$ and sparse matrix $H_{\tilde{w}}^{-1}$ such that

$$\tilde{w}(\boldsymbol{s})|\boldsymbol{\mu}_{\tilde{w}}, H_{\tilde{w}}^{-1} \sim \mathcal{N}(\boldsymbol{\mu}_{\tilde{w}}, H_{\tilde{w}}).$$

In this case it follows that

$$\begin{aligned}
p(\tilde{w}(\boldsymbol{s})|\cdot) &\propto p(\boldsymbol{Z}|\tilde{w}(\boldsymbol{s}), \cdot)p(\boldsymbol{v}|\tilde{w}(\boldsymbol{s}), \cdot)p(\tilde{w}(\boldsymbol{s})) \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{Z} - (X\boldsymbol{\beta} + \alpha UV\tilde{w}(\boldsymbol{s}) + UV\tilde{d}(\boldsymbol{s}))'\right. \\
&\qquad \left.\times (\boldsymbol{Z} - (X\boldsymbol{\beta} + \alpha UV\tilde{w}(\boldsymbol{s}) + UV\tilde{d}(\boldsymbol{s})))\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}(\boldsymbol{v} - (X_v\boldsymbol{\beta}_v + \alpha_v\tilde{w}(\boldsymbol{s})))'\right. \\
&\qquad \left.\times (\boldsymbol{v} - (X_v\boldsymbol{\beta}_v + \alpha_v\tilde{w}(\boldsymbol{s})))\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}(\tilde{w}(\boldsymbol{s}) - \boldsymbol{\mu}_{\tilde{w}})'H_{\tilde{w}}^{-1}(\tilde{w}(\boldsymbol{s}) - \boldsymbol{\mu}_{\tilde{w}})\right\} \\
&\propto \exp\left\{((\boldsymbol{Z} - X\boldsymbol{\beta} - UV\tilde{d}(\boldsymbol{s}))'\alpha UV + (\boldsymbol{v} - (X_v\boldsymbol{\beta}_v))'\alpha_v + \boldsymbol{\mu}_{\tilde{w}}'H_{\tilde{w}}^{-1})\tilde{w}(\boldsymbol{s})\right. \\
&\quad \left.-\frac{1}{2}\tilde{w}(\boldsymbol{s})'\left(\alpha^2(UV)'UV + \alpha_v^2 + H_{\tilde{w}}^{-1}\right)\tilde{w}(\boldsymbol{s})\right\}.
\end{aligned} \tag{3}$$

This implies that

$$\begin{aligned}
&\Rightarrow \tilde{w}(\boldsymbol{s})|\cdot \sim \mathcal{N}(\boldsymbol{\mu}_{\tilde{w}}^*, H_{\tilde{w}}^*) \\
&\quad \text{for: } H_{\tilde{w}}^* = \left(\alpha^2(UV)'UV + \alpha_v^2 + H_{\tilde{w}}^{-1}\right)^{-1}, \\
&\qquad \boldsymbol{\mu}_{\tilde{w}}^* = H_{\tilde{w}}^*((\alpha UV)'(\boldsymbol{Z} - X\boldsymbol{\beta} - UV\tilde{d}(\boldsymbol{s})) + (\boldsymbol{v} - \alpha_v(X_v\boldsymbol{\beta}_v)) + H_{\tilde{w}}^{-1}\boldsymbol{\mu}_{\tilde{w}}),
\end{aligned}$$

which allows for Gibb's sampling of $\tilde{w}(\boldsymbol{s})$. Note that this holds for any multivariate normal approximation of $w(\boldsymbol{s})$.

4 Simulation Study

As this work is motivated by understanding coronavirus prevalence in bat populations, our simulation study mimics this setting. Nine scenarios were considered by permuting three values each for α_v (the effect of $w(\boldsymbol{s})$ on selection) and β_{vPop} (the effect of "Developed, High Intensity" land use on a $\log+1$ scale). When either α_v or β_{vPop} is zero, "no impact" is expected. The other combinations of non-zero α_v and β_{vPop} are expected to lead to "nominal impact" or "large impact." Across these scenarios we evaluate our pooled, preferential sampling paradigm relative to a naive approach that does not account for preferential sampling and treats the sampling probabilities of all sites as equal. The complete naive model is given as

$$Y_{ij}|\boldsymbol{Z}_{ij} = \mathbb{1}_{0<\max(Z_{ijk})} \tag{4}$$

$$\boldsymbol{Z}|X, \boldsymbol{\beta}, U, V, \tilde{d}(\boldsymbol{s}) \sim \mathcal{N}(X\boldsymbol{\beta} + UV\tilde{d}(\boldsymbol{s}), I) \tag{5}$$

$$\tilde{d}(s_i)|\boldsymbol{B}_{di}, \boldsymbol{s}_{g_d(i)}, \sigma_d^2, H_{di} \sim N(\boldsymbol{B}_{di}\boldsymbol{s}_{g_d(i)}, \sigma_d^2 H_{di}) \tag{6}$$

$$\boldsymbol{B}_{di}|s_i, \boldsymbol{s}_{g_d(i)} = C_{\theta_d}(s_i, \boldsymbol{s}_{g_d(i)})C_{\theta_d}(\boldsymbol{s}_{g_d(i)}, \boldsymbol{s}_{g_d(i)})^{-1} \tag{7}$$

$$H_{di}|s_i, \boldsymbol{B}_{di}, \boldsymbol{s}_{g_d(i)} = C_{\theta_d}(s_i, s_i) - \boldsymbol{B}_{di}C_{\theta_d}(\boldsymbol{s}_{g_d(i)}, s_i), \tag{8}$$

where X is the design matrix of site-level covariates, and X_b is the design matrix of bat-level covariates. For covariance functions we chose the covariance function specified in Equation 2.

The models are compared on median bias in estimating a site-level covariate, β_{Pop}, which is the effect for a site's "Developed, High Intensity" value from the U.S. Geological Survey's National Land Cover Database (Yang et al. 2018) within 10 km, transformed using a $\log+1$ scale. For the simulations, real data from California were used; based on the NLCD database resolution, this results in 2,813 unique pixels. β_{Pop} was fixed at 0.1 and the intercept was chosen to fix the proportion of positive pools to be near 0.5.

4.1 Simulation Study Results

Figure 3 contains the estimated values of the coefficient associated with human population in the proximity of the roost for each replication in our simulation study. If either coefficient is near zero, the bias in the naive model is negligible and the results are similar to our model

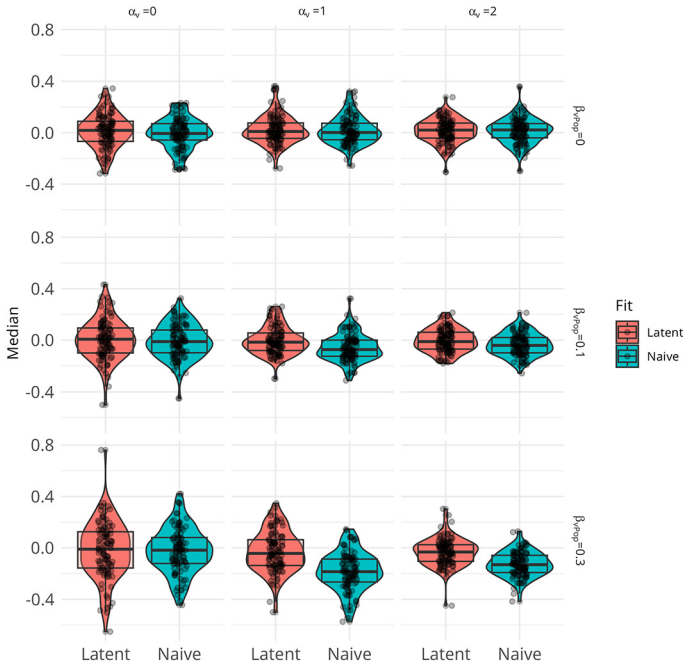


Figure 3: Comparison of bias in estimated medians for β_{Pop} from simulation study. The top row corresponds to no impact of population density on selection. The first column corresponds to no impact of the estimated spatial effect on selection. If either case is true, then a naive model appears to produce unbiased results. However, if both are false, then the bias can be seen to grow. For instance, if higher population increases both the probability of selection and the prevalence, then we'd expect to see increasingly large bias in estimates of prevalence that are based on samples selected from higher population sites.

Table 1: Root Mean Square Error (RMSE) from simulation scenarios.

α_v	β_{vPop}	Latent RMSE	Naive RMSE
0	0	0.014	0.000
0	0.1	0.010	0.010
0	0.3	0.020	0.022
1	0	0.022	0.017
1	0.1	0.000	0.057
1	0.3	0.028	0.058
2	0	0.014	0.014
2	0.1	0.000	0.035
2	0.3	0.033	0.127

accounting for preferential sampling. This makes intuitive sense, as bias is not expected to be present in these situations and there are no discernible differences between our approach and the simpler naive approach. However, as both coefficients grow, the naive model can clearly be seen to produce more biased estimates. In these cases, two things occur: first, preferential sampling, with respect to nearby human population, occurs and, second, nearby human population impacts the response – simulated coronavirus prevalence. In these scenarios the spatial correlation in $w(\boldsymbol{s})$, the spatial effect associated with preferential sampling is relatively high, it is likely that sampled sites would not be representative of all sites and hence biased estimates occur using a naive model that does not account for preferential sampling. When the magnitude of both these factors increase (bottom right corner of Figure 3), the bias in the naive model increases. It should be noted that the log population values range from about 0 to 11, so a difference of 0.1 in the coefficient represents a large difference. Julia code (Bezanson et al. 2017) for the simulation study is available at (https://anonymous.4open.science/r/pooled_preferential-B17C/README.md).

The Root Mean Squared Error (RMSE) for the posterior median was also calculated and is shared in the Table 1. Similar to the results for bias in Figure 3, we see similar performance between the two methods when either α_v or β_{vPop} are zero; however, in other cases the latent model that accounts for the preferential sampling results in improvements over the naive model.

Although Figure 3 and Table 1 show improved performance for the latent model, those improvement do come with a computing cost. For one simulation each under the latent and naive scenarios, computation time and effective sample size for β_{Pop} were calculated. For the naive scenario, the number of milliseconds per ESS was 32.403. For the latent scenario, the number of milliseconds per ESS was $10,893.734$. These calculations were conducted using an Intel i5-11600k processor at 4.9 GHz.

5 Data Analysis

We analyzed coronavirus prevalence data collected from Mexican free-tailed bats in California in 2022 (Figure 1) and explored factors related to coronavirus prevalence while controlling for preferential site selection. For each sampling location, site-level covariates related to land use and surrounding population characteristics were gathered via the Google Earth Engine (Gorelick

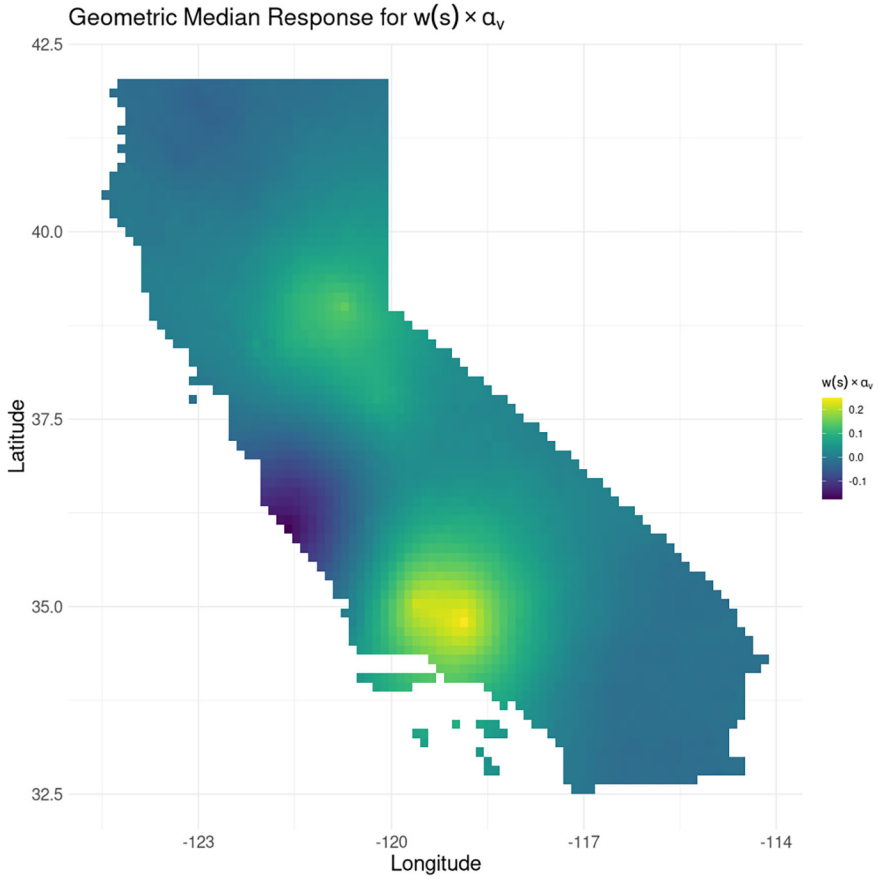


Figure 4: $w(\boldsymbol{s}) \times \boldsymbol{\alpha}$ for each pixel in California. $w(\boldsymbol{s})$ is a random effect associated with preferential sampling intensity and $\boldsymbol{\alpha}$ is a covariate for that random effect. Together, along with other covariates, they inform the prevalence of coronavirus prevalence in Mexican free-tailed bats in California.

et al. 2017). In particular, the following variables were calculated in a 10 km radius around the site: human population and the proportion of land use classified as barren, herbaceous, evergreen, grasslands, mixed forests, pasture, and shrubs. Additional covariates for individual bats were collected during capture. For each bat, mass, forearm length, sex, reproductive status, and age were collected. If the research goal was to predict prevalence across the landscape, the information from Figure 4 and Figure 5 could be combined to create a prediction across the study area in California. However, our research goal is more focused on estimating the relationship between site and bat-level covariates and coronavirus prevalence while controlling for potential bias from preferential sampling.

The results are summarized by two components in Equation 1. The first component, $\boldsymbol{\alpha} \times w(\boldsymbol{s})$, represents the contribution of the response on the latent probit scale that corresponds to the preferential sampling intensity. Recall that $w(\boldsymbol{s})$ is a spatial process related to sampling intensity and $\boldsymbol{\alpha}$ can be interpreted as the coefficient on that process. Figure 4 shows the product of $\boldsymbol{\alpha} \times w(\boldsymbol{s})$ across the State of California, based on a gridded approach, using human population and land cover characteristics within a 10 km radius. The yellower areas indicate higher preferential sampling intensity, while the darker areas indicate lower preferential sampling intensity. The

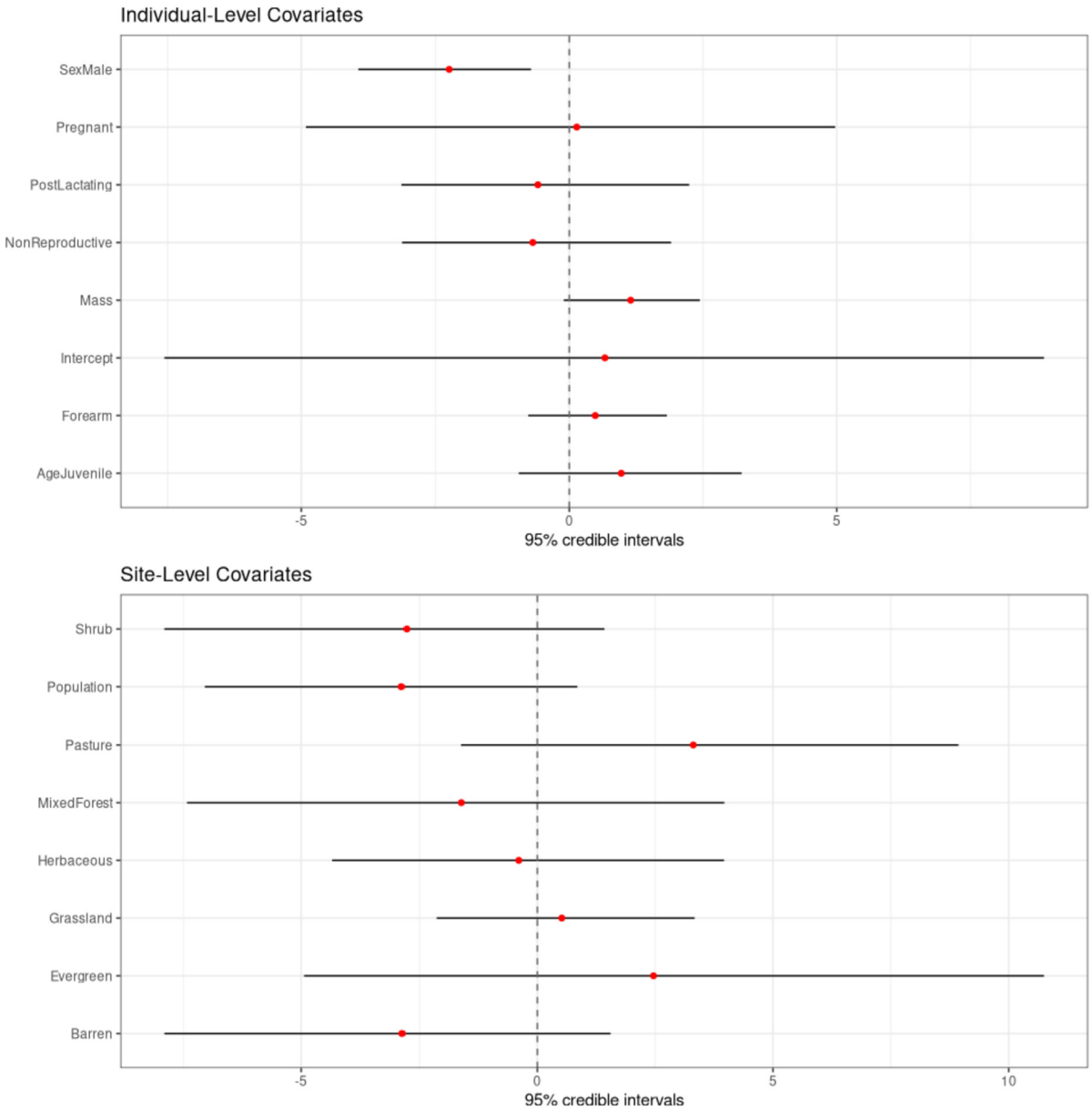


Figure 5: 95% credible intervals for individual and site-level covariates. The black bars represent the credible intervals and the red points are posterior means. With the exception of sex, most intervals are relatively wide and include zero.

higher intensity regions in Figure 4 are near major population centers in both northern and southern California. This suggests that locations close to these population centers are more likely to be sampled. However, the magnitude of the $\alpha \times w(\boldsymbol{s})$, which corresponds to the degree of preferential sampling, ranging from approximately .25 to $-.25$ is relatively small when considered on the probit scale implied with Equation 1. Recall the naive model specified in Equations 4–8 lacks the $\alpha \times w(\boldsymbol{s})$ term, hence without explicitly controlling for preferential sampling intensity, that unexplained variation can induce bias in the β values.

The second component is the $X\beta$ term. Controlling for the traditional spatial random effect, $d(\boldsymbol{s})$, and the preferential sampling intensity, $\alpha \times w(\boldsymbol{s})$, $X\beta$ is the signal of disease prevalance, on the latent scale, attributable to the covariates. The matrix X includes both the site and bat-level covariates. Most of the 95% credible intervals for the site-level covariates are fairly wide and include zero (Figure 5). Although these human density and land-use covariates were potentially expected to contribute to the preferential sampling bias, it would have been somewhat unexpected for these covariates to contribute to elevated or decreased prevalence levels. Figure 5 also contains 95% credible intervals for the individual bat-level covariates. Many of these intervals are also wide and include zero, but the exception would be sex where male bats have lower prevalence. Reproductive cycles are known to increase stress and have been shown to increase viral shedding in female bats (Ruiz-Aravena et al. 2022), which is a likely cause of differences in prevalence between male and females bats as much of the sampling during this period occurred during the later stages of maternity. Overall, the wide credible intervals result from limited information in our pooled dataset, which includes only 20 total sites.

6 Discussion

We have extended existing preferential sampling methodology to account for scenarios where samples are tested in pools without retesting individuals. Inspired the Albert and Chib (1993) latent probit approach and the Pati, Reich, and Dunson (2011) preferential sampling algorithm, we present a data augmentation algorithm that accounts for uncertainty in the pooled responses that can also address preferential sampling issues. Our approach also enables inferences from individual-level covariates within pools. Even in problematic preferential sampling settings, when variables related to site selection are also related to disease prevalence, our method produced accurate covariates estimates.

Furthermore, in scenarios with no preferential sampling or no relationship between preferentially sampled variables and the response of interest, our model performs similarly to a naive model with both producing unbiased results. Although the model requires additional computation, the results are similar in these cases. A limitation is that the additional computation can be substantial. Despite this limitation, our model can be used as a tool for identifying issues stemming from preferential sampling and in certain cases simpler models could follow.

While our extended research team has developed an algorithm for predicting bat roosts (Oram et al. 2025), we do not have a complete sampling frame of all bat roosts in California or even probable bat roosts. Rather than making predictions across the entire state of California, similar to Figure 4, a use-availability approach motivated by Johnson, Williams, and Riordan (2021) could be used to make predictions at background locations that are similar to our sampled points. With sets of preferentially and probabilistically sampled roosts, this would mimic case-control study and methods presented in Savitsky et al. (2023) and could be used for drawing inferences.

Pooling can be a useful tool in certain situations (Hoegh et al. 2021), namely when prevalence is expected to be low and individual specimens can be collected cheaply relative to the cost of testing samples. However, pooling, without retesting, can also be detrimental due to the loss of information as the disease status of individuals are not directly known. Ultimately sophisticated models like the one we have detailed can only recover so much information. Without cost or time constraints, collecting and testing as many individual samples as possible would give the most precise results. In this case, our framework would simplify but still produce the same

results, relative to a naive model. In addition, our model framework can control for potential bias induced by preferential sampling and infer relationships between site-level, pool-level, and individual-level covariates.

Supplementary Material

Code for the simulation studies and data analysis is available at https://anonymous.4open.science/r/pooled_preferential-B17C/README.md. The raw data is available at https://doi.org/10.5066/P14HVQHW.

Acknowledgements

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This material is based upon work supported by the U.S. Geological Survey under Grant/Cooperative Agreement number G21AC10748-00.

References

Albert JH, Chib S (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422): 669–679. https://doi.org/10.1080/01621459.1993.10476321

Banerjee S (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12(2): 583. https://doi.org/10.1214/17-BA1056R

Bezanson J, Edelman A, Karpinski S, Shah VB (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1): 65–98. https://doi.org/10.1137/141000671

Bilder CR, Tebbs JM, Chen P (2010). Informative retesting. *Journal of the American Statistical Association*, 105(491): 942–955. https://doi.org/10.1198/jasa.2010.ap09231

Cheng J, Schloerke B, Karambelkar B, Xie Y (2025). *Leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library*. https://rstudio.github.io/leaflet/.

Conroy B, Waller LA, Buller ID, Hacker GM, Tucker JR, Novak MG (2023). A shared latent process model to correct for preferential sampling in disease surveillance systems. *Journal of Agricultural, Biological, and Environmental Statistics*, 28(3): 483–501. https://doi.org/10.1007/s13253-023-00535-4

Diggle PJ, Menezes R, Su T-l (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 59(2): 191–232. https://doi.org/10.1111/j.1467-9876.2009.00701.x

Du D-Z, Hwang FK-m (1999). *Combinatorial Group Testing and Its Applications*, volume 12. World Scientific.

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017). Google Earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202: 18–27. https://doi.org/10.1016/j.rse.2017.06.031

Hall JS, Hofmeister E, Ip HS, Nashold SW, Leon AE, Malave CM, et al. (2023). Experimental infection of Mexican free-tailed bats (tadarida brasiliensis) with SARS-CoV-2. *Msphere*, 8(1): e00263–22.

Haydon DT, Cleaveland S, Taylor LH, Laurenson MK (2002). Identifying reservoirs of infection: A conceptual and practical challenge. *Emerging Infectious Diseases*, 8(12): 1468–1473. https://doi.org/10.3201/eid0812.010317

Hoegh A, Peel AJ, Madden W, Ruiz Aravena M, Morris A, Washburne A, et al. (2021). Estimating viral prevalence with data fusion for adaptive two-phase pooled sampling. *Ecology and Evolution*, 11(20): 14012–14023. https://doi.org/10.1002/ece3.8107

Jimenez F, Katzfuss M (2023). Scalable Bayesian optimization using Vecchia approximations of Gaussian processes. In: *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics* (F Ruiz, J Dy, J-W van de Meent, eds.), volume 206 of Proceedings of Machine Learning Research. PMLR, 1492–1512. https://proceedings.mlr.press/v206/jimenez23a.html.

Johnson NG, Williams MR, Riordan EC (2021). Generalized nonlinear models can solve the prediction problem for data from species-stratified use-availability designs. *Diversity and Distributions*, 27(11): 2077–2092. https://doi.org/10.1111/ddi.13384

Katzfuss M, Guinness J (2021). A general framework for Vecchia approximations of Gaussian processes. *Statistical Science*, 36(1): 124–141. https://doi.org/10.1214/19-STS755

Mackenzie JS, Childs JE, Field HE, Wang L-F, Breed AC (2016). The role of bats as reservoir hosts of emerging neuroviruses. In: *Neurotropic Viral Infections* (CS Reiss, ed.), 403–454. https://doi.org/10.1007/978-3-319-33189-8_12.

Mallapaty S, et al. (2020). The mathematical strategy that could transform coronavirus testing. *Nature*, 583(7817): 504–505. https://doi.org/10.1038/d41586-020-02053-6

Meyer M, Melville DW, Baldwin HJ, Wilhelm K, Nkrumah EE, Badu EK, et al. (2024). Bat species assemblage predicts coronavirus prevalence. *Nature Communications*, 15(1): 2887. https://doi.org/10.1038/s41467-024-46979-1

Møller J, Syversveen AR, Waagepetersen RP (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3): 451–482. https://doi.org/10.1111/1467-9469.00115

Moreira GA, Menezes R, Wise L (2024). Presence-only for marked point process under preferential sampling. *Journal of Agricultural, Biological, and Environmental Statistics*, 29(1): 92–109. https://doi.org/10.1007/s13253-023-00558-x

OpenStreetMap contributors. (2017). Planet dump retrieved from https://planet.osm.org. https://www.openstreetmap.org.

Oram J, Wray AK, Davis HT, de Wit LA, Frick WF, Hoegh A, et al. (2025). Predicting Bat Roosts in Bridges Using Bayesian Additive Regression Trees. *Global Ecology and Conservation*. 60(e03551). https://doi.org/10.1016/j.gecco.2025.e03551

Pati D, Reich BJ, Dunson DB (2011). Bayesian geostatistical modelling with informative sampling locations. *Biometrika*, 98(1): 35–48. https://doi.org/10.1093/biomet/asq067

Peel AJ, Ruiz-Aravena M, Kim K, (2025). Synchronized seasonal excretion of multiple coronaviruses coincides with high rates of coinfection in immature bats. *Accepted*, *Nature Communications*.

Plowright RK, Parrish CR, McCallum H, Hudson PJ, Ko AI, Graham AL, et al. (2017). Pathways to zoonotic spillover. *Nature Reviews. Microbiology*, 15(8): 502–510. https://doi.org/10.1038/nrmicro.2017.45

Rahman MT, Sobur MA, Islam MS, Ievy S, Hossain MJ, et al. (2020). Zoonotic diseases: Etiology, impact, and control. *Microorganisms*, 8(9): 1405. https://doi.org/10.3390/microorganisms8091405

Ruiz-Aravena M, McKee C, Gamble A, Lunn T, Morris A, Snedden CE, et al. (2022). Ecol-

ogy, evolution and spillover of coronaviruses from bats. *Nature Reviews. Microbiology*, 20(5): 299–314. https://doi.org/10.1038/s41579-021-00652-2

Savitsky TD, Williams MR, Gershunskaya J, Beresovsky V, Johnson NG (2023). Methods for combining probability and nonprobability samples under unknown overlaps. *Statistics in Transition*, 24(4): 1–34.

Stevens Jr DL, Olsen AR (2004). Spatially balanced sampling of natural resources. *Journal of the American Statistical Association*, 99(465): 262–278. https://doi.org/10.1198/016214504000000250

Talbert C, Reichert BE (2018). *North American Bat Monitoring Program (NABat) Master Sample and Grid-Based Sampling Frame*. U.S. Geological Survey. https://doi.org/10.5066/P9O75YDV.

Vecchia AV (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 50(2): 297–312. https://doi.org/10.1111/j.2517-6161.1988.tb01729.x

Vedensky D, Parker PA, Holan SH (2023). A look into the problem of preferential sampling through the lens of survey statistics. *American Statistician*, 77(3): 313–322. https://doi.org/10.1080/00031305.2022.2143898

Warasi MS, Hungerford LL, Lahmers K (2022). Optimizing pooled testing for estimating the prevalence of multiple diseases. *Journal of Agricultural, Biological, and Environmental Statistics*, 27(4): 713–727. https://doi.org/10.1007/s13253-022-00511-4

Wong S, Lau S, Woo P, Yuen K-Y (2007). Bats as a continuing source of emerging infections in humans. *Reviews in Medical Virology*, 17(2): 67–91. https://doi.org/10.1002/rmv.520

Wray A, de Wit L, Banner K, Foster J, Frick W, Gibson A, et al. (2025). OneHealth: U.S. geological survey data release. *North American Bat Monitoring Program (NABat)*. https://doi.org/10.5066/P14HVQHW.

Yang L, Jin S, Danielson P, Homer C, Gass L, Bender SM, et al. (2018). A new generation of the United States national land cover database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146: 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006