

# Modeling the Marked Presence-only Data: A Case Study of Estimating the Female Sex Worker Size in Malawi

Ian Laga, Xiaoyue Niu, and Le Bao \*

Department of Statistics, Pennsylvania State University

## Abstract

Certain subpopulations like female sex workers (FSW), men who have sex with men (MSM), and people who inject drugs (PWID) often have higher prevalence of HIV/AIDS and are difficult to map directly due to stigma, discrimination, and criminalization. Fine-scale mapping of those populations contributes to the progress towards reducing the inequalities and ending the AIDS epidemic. In 2016 and 2017, the PLACE surveys were conducted at 3,290 venues in 20 out of the total 28 districts in Malawi to estimate the FSW sizes. These venues represent a presence-only data set where, instead of knowing both where people live and do not live (presence-absence data), only information about visited locations is available. In this study, we develop a Bayesian model for presence-only data and utilize the PLACE data to estimate the FSW size and uncertainty interval at a  $1.5 \times 1.5$ -km resolution for all of Malawi. The estimates can also be aggregated to any desirable level (city/district/region) for implementing targeted HIV prevention and treatment programs in FSW communities, which have been successful in lowering the incidence of HIV and other sexually transmitted infections.

*Keywords:* HIV; Key population; Bayesian modeling; Small area estimation; Prevalence mapping.

---

\*This work was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award number R01AI136664. Correspondence to: lebao@psu.edu

# 1 Introduction

The work in this paper is motivated by the study of the HIV/AIDS epidemic, which is especially a public health threat in sub-Saharan African countries. Eastern and Southern Africa account for 45% of HIV infections and 53% of individuals living with HIV globally. The 90-90-90 treatment target set by UNAIDS aims to achieve three goals by the end of 2020: 90% of all people living with HIV know their HIV status, 90% of all people with diagnosed HIV infection will receive sustained antiretroviral therapy, and 90% of all people receiving antiretroviral therapy will have viral suppression (Joint United Nations Programme on HIV/AIDS, 2014a). While saving lives has been very successful, the number of new HIV infections was not falling fast enough to meet the 2020 goal (Joint United Nations Programme on HIV/AIDS, 2018). New HIV infections were reduced by only 18% from 2010 to 2017, much lower than the 2020 target of 75%. In order to decrease new HIV infections, it is first necessary to understand where those with AIDS live. A very large portion of new HIV infections occur in key populations like sex workers, men who have sex with men, and intravenous drug users. Female sex workers (FSW) are one of the largest groups and are 12 times more likely to have HIV than the general population (Joint United Nations Programme on HIV/AIDS, 2014b). One study of 16 sub-Saharan African countries found the HIV prevalence for FSW to be above 37% (Joint United Nations Programme on HIV/AIDS, 2014b). It is clear that treating HIV positive individuals is not enough to meet the 90-90-90 goal, and success depends on reducing the spread of HIV in these key populations.

FSW populations are difficult to map directly because they represent a relatively small part of the population and have an incentive to hide their status. Small area estimates often suffer from large uncertainty due to small sample sizes. The University of North Carolina

performed on-site surveying of venues believed to house FSW to estimate district-level FSW size in Malawi, which is by far the most detailed venue level data (UNC et al., 2018). The venue is the unit of observation in the study. Their team published the Priorities for Local AIDS Control Efforts (PLACE) report, which documents how venues were discovered and visited and how FSW size estimates were calculated using probability weights. Only a subset of the reported venues were visited. Prior to surveying venues, the PLACE team obtained accurate counts of the total number of venues in each district. By mapping venues instead of FSW directly, the PLACE team can reach a much larger portion of the FSW population.

In an ideal case, regression models to map population sizes are built on records of both where people live and where no people live. Records that include both pieces of information are known as “presence-absence” or “used-unused” data (Pearce and Boyce, 2006). Regression methods can typically be applied directly to presence-absence data without much trouble. However, recording absences can be difficult, leading to presence-only data, which is data where only coordinates of observed sites are recorded, offering no information of presence for other locations. Most of the original presence-only literature used a “pseudo-sampling” approach, which samples background observations and assumes they are absences. Combining the pseudo-absences with the observed presences created a pseudo-presence-absence data set that could be modeled with more traditional techniques. We recommend Pearce and Boyce (2006) for a nice introduction. Maxent also emerged as a popular choice for modeling presence-only data (Phillips et al., 2004, 2006).

Given the lack of information about absences, modeling abundance given presence-only data is challenging, so traditional models limit inference to binary presence-absence estimates, rather than any measure of cell-level abundance. To the best of their knowledge,

Pearce and Boyce (2006) knew of no application modeling abundance given presence only. Ward et al. (2009) later developed an EM approach which can estimate abundance given presence only when an estimate of population prevalence is available, but it can only handle binary presence-absence observations. Later researchers developed point process models for the data which overcame many of the existing problems. See Banerjee et al. (2014) and Renner et al. (2015) for some examples.

Additionally, the observations may be marked points. For marked point processes, the observations contain additional information about the observation. For example, the locations of trees constitute a point process, and the height or width of the trees are “marks.” Banerjee et al. (2014) offers a nice chapter on marked point processes.

The PLACE data are an example of the marked presence-only data, where locations of visited venues constitute the point process, the visited venue sizes represent the marks, and the locations of unvisited venues are unknown. The PLACE report also provides the total number of venues in each district. In this article, we propose a Bayesian model-based approach that can map abundance given both presence-only data and the total number of observations. We utilize the PLACE data to estimate Malawi FSW sizes at fine-scale cells. Our proposed model is flexible and can account a wide range of data properties. Our model can handle cells with more than one observation, estimate abundance given the total number of observations, include random effects to reduce heterogeneity and account for systematic differences between areas, handle marked observations, incorporate uneven sampling efforts, and calculate credible intervals for domains that include multiple cells.

The paper is organized as follows. Section 2 introduces the data and their special features and how to handle the spatial misalignment issue. In Section 3, we introduce our choices of distributions and a calibrated Bayesian presence-only approach for modeling

the FSW size at the cell-level. We validate our approach on a complete presence-absence subset of the PLACE data and via simulations in Section 4. We apply the method on the entire PLACE data in Section 5, providing cell-level and district level estimates of FSW size, including credible intervals. Finally, Section 6 covers concluding remarks and future work.

## 2 Data and Special Features

We combine multiple data sources in order to estimate the FSW size at a fine-scale grid. Our responses (FSW size at visited venues and the venue locations) come from PLACE, while the predictors come from a combination of environmental, demographic, and health data. Subsection 2.1 further introduces the PLACE data. In subsection 2.2, we explore the auxiliary variables we use as predictors of FSW size. Data pre-processing and spatial alignment is handled in subsection 2.3.

### 2.1 PLACE Data

The Priorities for Local AIDS Control Efforts (PLACE) is funded by the United States Agency for International Development (USAID) and the United States President’s Emergency Plan for AIDS Relief (PEPFAR) to understand the HIV epidemic and help reach the 90-90-90 target (Measure Evaluation, 2018). PLACE was developed to address the local behavior of HIV transmission. In 2016, The University of North Carolina implemented the PLACE I study in Malawi, surveying five districts (Lilongwe, Blantyre, Mangochi, Machinga, and Zomba) and one city (Mzuzu). Through additional funding, the PLACE II study was implemented in 15 additional districts, including the rest of Mzimba, the district

for Mzuzu. The four main objectives enumerated in the PLACE report (2018) are: (1) to conduct programmatic mapping in selected district to identify venues where key populations can be reached, (2) to estimate the size of key population in each district who can be reached at venues, (3) to characterize HIV service coverage indicators for HIV programs reaching key populations, (4) in a subset of districts, to survey and test members of key population groups. Figure 1 includes a map of the visited venues in the Lilongwe district (UNC et al., 2018). We can see that venues are highly clustered and typically lie along major roads or near the city center.

In order to achieve the first goal of PLACE, the team interviewed community informants to form a list of venues which were believed to host FSW, cleaned the list, and verified that the list was complete. The district-level venue counts were obtained for the 20 districts in PLACE I and II. The PLACE team visited a subset of venues. GPS locations were only recorded for these visited venues and thus formed the presence-only data. Interviews with knowledgeable personnel at each venue provided estimates of the number of FSW present at each venue. In PLACE I, the original plan was to visit all identified venues. However, due to time constraints, a convenience sample of venues was visited. In PLACE II, venues were divided into groups based on priority, and venues were sampled from the high priority groups via a predetermined structure. In the PLACE Report, FSW estimates for district where no data was collected were obtained by combining expert knowledge with prevalence estimates from other districts. For example, the FSW prevalence in Thyolo was assumed to be equal to the average prevalence of all district in the Southern Region. The complete methodology for both PLACE reports is recorded in the PLACE Report Malawi (UNC et al., 2018).

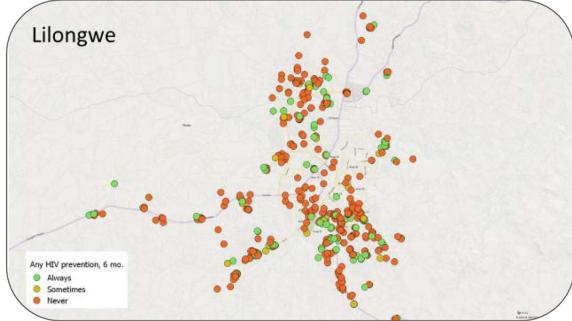


Figure 1: An example of visited sites in Lilongwe, reprinted with permission from the PLACE report. Colors indicate whether any HIV prevention was available at the site in the last 6 months. HIV prevention activities include condom availability, HIV testing, and similar activities.

## 2.2 Auxiliary Variables

Using the auxiliary variables enables us to map a more complete picture of Malawi key populations. In the PLACE report, venue weights were used to extrapolate the FSW size to the district level (UNC et al., 2018). The proportion of venues visited and operational were used to calculate the venue weights. This venue weighting assumes that unvisited venues behave similarly to visited venues, ignoring the location and local characteristics of unvisited venues. We can improve FSW size estimates by employing a regression-based approach with auxiliary data from multiple sources which accounts for the local characteristics of all venues. The auxiliary variables that we use are the demographic and health survey (DHS) data, night-time lights (nightlight), and WorldPop.

### 2.2.1 DHS Data

The DHS data are the results of a survey of 26,361 household in Malawi where respondents were asked questions like annual income, age at first sex, and number of sexual partners.

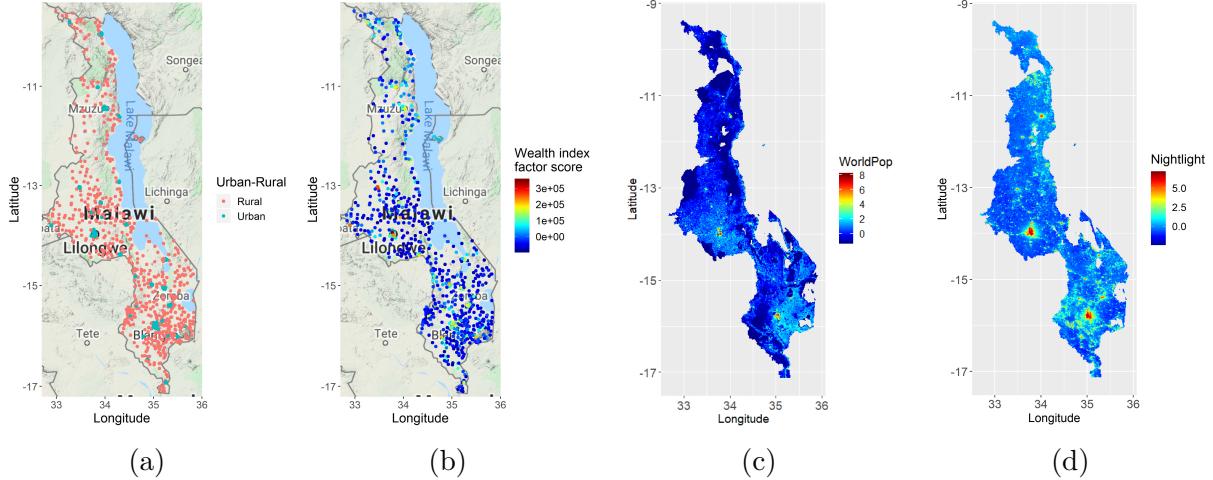


Figure 2: Illustrative maps of the auxiliary covariates. (a) shows which clusters are urban and rural. (b) shows the distribution of the mv191 (combined wealth index factor score). (c) and (d) show the scaled WorldPop and nightlight cell estimates, respectively.

Respondents were surveyed between October 2015 and February 2016. DHS divided households into 850 “clusters,” which represent groups of household that are close in proximity. DHS perturbed the cluster locations to ensure the privacy of respondents. We treat the recorded cluster locations as the truth since the effect of the perturbations will have a much smaller effect on the estimates than the variability of the venue counts and the FSW sizes. As the data contains thousands of questions, after consulting with relevant experts, we created an initial list of covariates that we believed would be useful in modeling either the venue count or venue size. DHS clusters (as well as WorldPop and nightlight data discussed later) are shown in Figure 2.

### **2.2.2 Nightlight**

The most popular FSW venues likely produce light during the night, so a measure of nightlight intensity should be an important predictor of both venue count and FSW size. We use nightlight data recorded by the Earth Observations Group at the National Oceanic and Atmospheric Administration and the National Centers for Environmental Information (NOAA/NCEI) (NOAA, 2016). The nightlight data measures the low levels of visible-near infrared radiance via satellites. The data was recorded at a roughly 450 meter resolution, resulting in 567,284 locations of recordings from polar orbiting satellites twice a day. We use 2016 data, aggregated by month. Month averages are then averaged, yielding annual average nighttime values across Malawi. The cell-level nightlight values are extremely right-skewed, so the values were taken on the log-scale after alignment. The cell-level estimates after centering and scaling are shown in Figure 2.

### **2.2.3 WorldPop**

Areas with higher populations should intuitively have more venues, and potentially more FSW per venue. In order to capture this relationship, we use the 2015 adjusted WorldPop data set, which adjusts the population estimates to match UN national estimates (WorldPop, 2015). WorldPop estimates the number of persons per grid square at approximately 100 by 100 meter cells. Cells with zero WorldPop values were removed and the remaining values were taken on the log-scale. These cells typically represent inhabitable zones like forests and lakes, so we limit inference to cells that are habitable. The cell-level estimates after centering and scaling are also shown in Figure 2.

## 2.3 Combining the Data

The venues in Malawi represent a presence-only marked point process since the observations contain only the longitude, latitude, and venue size estimate of *visited* venues in Malawi. However, the responses and predictors exist on different spatial levels. This issue is known as spatial misalignment, and has been discussed in detail (see Mugglin et al. (2000), Gelfand et al. (2001), Young et al. (2009), Carlin et al. (1999), Banerjee et al. (2014).) Since we have a two-part model, we also have two spatial resolutions: cell-level data for the venue count model and venue-level data for the venue size model.

For the venue count model, we first divide Malawi into cells of approximately 1.5x1.5-km. We discuss the choice of cell size in Section 4. The areas of the cells vary between 2.318 km<sup>2</sup> and 2.392 km<sup>2</sup>, which has a negligible effect on the estimates. The small differences in area arise because the distance between each degree of latitude depends on the distance to the equator. For the venue count response, we count the number of observed venues in each cell. For the DHS predictors, we perform universal kriging to estimate the value in each cell. We impose a Matérn covariance function on the Gaussian Process in the kriging model, and parameters are estimated using restricted maximum likelihood via the **fields** package (Nychka et al., 2017). The WorldPop and nightlight data are recorded at a much finer scale than our Malawi cells, so cell-level values are calculated by averaging the WorldPop and nightlight points which lie inside the cell. After removing the cells with WorldPop equal to zero, there are 39,312 cells for all of Malawi and 20,751 cells in PLACE II districts, of which 385 contained observed PLACE II venues.

For the venue size model, our spatial resolution is the observed venue locations. Thus, only the predictors need to be aligned since the venue size response is already known at observed venues. The procedure for the DHS predictors is the same as for the cell-

level, where estimates are now obtained at venue locations via kriging. The WorldPop and nightlight values at each venue are obtained by finding the closest WorldPop and nightlight recording.

## 3 Methodology

In this section, we introduce the approach for modeling the venue size (number of FSW at the venue-level) and venue count (number of venue at the cell-level). The venue locations comprise a presence-only data set. Our model assumes that the true cell-level venue counts follow a zero-inflated negative binomial (ZINB) distribution, and we observe a thinned version of these counts. The details of the thinning are discussed in Section 3.3. While analyzing spatial point patterns by dividing the domain into cells and modeling the number of occurrences in the cells is not new, we are unaware of its application to presence-only data. In this section, we first introduce the venue size and venue count models. Following the modelling choices, we discuss how the models are used to predict the number of FSW in each cell which includes a calibration step.

### 3.1 Models

Here we summarize the modeling choices and further specify our choices of priors in the Bayesian framework. Both the venue counts and the venue sizes seem to follow zero-inflated distributions. We assume that given the covariates for a cell, the venue count and the venue size are conditionally independent and we assess whether or not this assumption is reasonable in subsection 4. We use the zero-inflated model described in Lambert (1992) to model both the venue counts and the FSW size. In this formulation, zeros can come from

either excess zeros or from the conditional distribution,  $f_c(y)$ . The zero-inflated distribution is given by

$$\begin{aligned} P(Y = 0) &= p + (1 - p)f_c(0) && \text{for } y = 0 \\ P(Y = y) &= (1 - p)f_c(y) && \text{for } y > 0, \end{aligned} \tag{1}$$

where  $p$  is the probability of excess zeros and the choices of  $f_c(y)$  are motivated by predictive performance and are specified in Sections 3.2 and 3.3.

### 3.2 Venue Size

We can directly model the average number of FSW at the venue-level (venue size) using both PLACE I and PLACE II data. We choose a hurdle log-normal model for the venue size model, i.e.  $f_c(y)$  in Equation (1) is given by  $\log(Y_{\text{size}}) \sim N(\mu, \sigma_{\text{size}}^2)$ , where  $Y_{\text{size}}$  represents the observed venue sizes with corresponding covariate matrix  $\mathbf{X}_{\text{size}}$ . We use a logit link for  $p$ . Given the covariates, we assume the linear regression models:

$$\text{logit}(p_{i,\text{size}}) = \alpha_{0,\text{size}} + \mathbf{X}_{i,\text{size},p}\boldsymbol{\alpha}_{1,\text{size}} + \gamma_{d[i],\text{size},p} \tag{2}$$

$$\log(\mu_{i,\text{size}}) = \beta_{0,\text{size}} + \mathbf{X}_{i,\text{size},\mu}\boldsymbol{\beta}_{1,\text{size}} + \gamma_{d[i],\text{size},\mu} \tag{3}$$

where  $i$  denotes the cell number,  $d[i]$  denotes the district for cell  $i$ ,  $\mathbf{X}_{i,\text{size},p}$  and  $\mathbf{X}_{i,\text{size},\mu}$  are subsets of  $\mathbf{X}_{\text{size}}$ , and  $\gamma$  are the random district effects. We impose weak priors on the regression coefficients, where  $\alpha_0 \propto \text{logistic}(0, 1)$ ,  $\boldsymbol{\alpha}_1 \propto \mathbf{1}$ ,  $\beta_0 \propto t_3(-2, 10)$ ,  $\boldsymbol{\beta}_1 \propto \mathbf{1}$ ,  $\sigma_{\text{size}}^2 \propto t_3(0, 10)$ . For the district random effects, we impose normal priors, i.e.  $\gamma_{d[i],\text{size},\cdot} \sim N(0, \sigma_{\text{district},\text{size},\cdot}^2)$ , where the  $\cdot$  argument is either for  $p$  or  $\mu$ . Additionally,  $\sigma_{\text{district},\text{size},\cdot}^2 \propto \mathbf{1}$ .

We choose to model venue size via a hurdle log-normal distribution based on both visual diagnostics, posterior predictive p-values, and leave-one-out (loo) cross-validation.

The hurdle log-normal distribution offers much better results for these metrics than a ZINB distribution, despite the venue sizes being counts. Because of the large number of zeros, non-zero-inflated models do not perform well.

### 3.3 Venue Count

We first assume that the true cell-level venue counts,  $\tilde{\mathbf{Y}}$ , follow a ZINB distribution and the observed venues  $\mathbf{Y}$  follow a  $\pi$ -thinning of  $\tilde{\mathbf{Y}}$ , i.e.

$$\begin{aligned} Y_i &= \sum_{k=1}^{\tilde{Y}_i} I_{i,k}, \\ \tilde{Y}_i &\sim ZINB(p_i, \mu_i, \phi) \\ I_{i,k} &\sim Bern(\pi_{d[i]}) \end{aligned} \tag{4}$$

where  $I_{i,k}$  indicates whether venue  $k$  in cell  $i$  was sampled and  $\pi_{d[i]}$  is the sampling probability for the district for cell  $i$ . Here, we use  $ZINB(p_i, \mu_i, \phi)$  to denote the zero-inflated negative binomial distribution with probability of excess zero  $p_i$ , conditional mean  $\mu_i$ , and conditional over-dispersion  $\phi$ , where the negative binomial parameterization we consider is

$$f_c(\tilde{y}_i) = \binom{\tilde{y}_i + \phi - 1}{\tilde{y}_i} \left( \frac{\mu_i}{\mu_i + \phi} \right)^{\tilde{y}_i} \left( \frac{\phi}{\mu_i + \phi} \right)^\phi. \tag{5}$$

This  $\pi$ -thinning, also known as binomial subsampling, assumes that each venue is sampled independently according to the sampling effort in each district (Puig and Valero, 2007).

In order to show that the distribution of  $\mathbf{Y}$  also follows a ZINB distribution where only the mean of the negative binomial component is multiplied by  $\pi_{d[i]}$ , we present the following two propositions.

**Proposition 1.** Let a random variable  $X \sim \text{NegBin}(\text{mean} = \mu, \text{dispersion} = \phi)$ , as defined in (4). Let  $Y = \sum_{k=1}^X I_k$ , where  $I_k \sim \text{Bern}(\pi)$ , the  $\pi$ -thinning of  $X$ . Then,  $Y \sim \text{NegBin}(\text{mean} = \pi\mu, \text{dispersion} = \phi)$ .

**Proposition 2.** Let a random variable  $\tilde{Y} = (1 - Z)X$  be any zero-inflated discrete random variable, where  $Z \sim \text{Bern}(p)$  and  $f_c(x)$  represents the p.m.f. of  $X$ . Let  $X^* = \sum_{i=1}^X I_{x,i}$  denote the  $\pi$ -thinning of the count distribution, where  $I_{x,i} \sim \text{Bern}(\pi)$ . If  $Y = \sum_{i=1}^{\tilde{Y}} I_{y,i}$ , where  $I_{y,i} \sim \text{Bern}(\pi)$ , then  $Y \stackrel{D}{=} (1 - Z)X^*$ .

Another way of stating Proposition 2 is that the thinning operator ignores the zero-inflated component. With these two propositions, we arrive at the main proposition for our model.

**Proposition 3.** Let  $\tilde{Y}$  be a zero-inflated negative binomial random variable, with negative binomial mean  $\mu$  and dispersion  $\phi$ , i.e.  $\tilde{Y} \sim \text{ZINB}(p, \mu, \phi)$ . If  $Y = \sum_{k=1}^{\tilde{Y}} I_k$ , where  $I_k \sim \text{Bern}(\pi)$ , then  $Y \sim \text{ZINB}(p, \pi\mu, \phi)$ .

All proofs are shown in the Appendix. Using Proposition 3,  $Y_i \sim \text{ZINB}(p_i, \pi_{d[i]}\mu_{i,\text{count}}, \phi)$ , where  $Y_i$  is our observed venue count for cell  $i$ . In standard generalized linear model terms, this is equivalent to adding an offset of  $\log(\pi_{d[i]})$  to the linear predictor, providing an easy way to fit the model using common statistical software. Because the sampling probability in each district,  $\pi_{d[i]}$ , are known, we can estimate all parameters in the original model directly by fitting an offset model. Thus, our model for the observed venue counts, probability of

excess zeros  $\mathbf{p}_{\text{count}}$ , and mean of the conditional distribution  $\boldsymbol{\mu}_{\text{count}}$  are given by

$$\begin{aligned} Y_i &\sim ZINB(p_{i,\text{count}}, \mu_{i,\text{count}}, \phi) \\ \text{logit}(p_{i,\text{count}}) &= \alpha_{0,\text{count}} + \mathbf{X}_{i,\text{count},p} \boldsymbol{\alpha}_{1,\text{count}} + \gamma_{d[i],\text{count},p} + \eta_{i,\text{count},p} \\ \log(\mu_{i,\text{count}}) &= \beta_{0,\text{count}} + \mathbf{X}_{i,\text{count},\mu} \boldsymbol{\beta}_{1,\text{count}} + \gamma_{d[i],\text{count},\mu} + \log(\pi_{d[i]}), \end{aligned} \quad (6)$$

where  $\boldsymbol{\eta} \sim MVN(0, k_p(x))$ . Based on the data and loo cross-validation, a spatial error term  $\boldsymbol{\eta}$  is placed on the Bernoulli model but not on the negative binomial model. We impose the same priors on the regression parameters and the random effects. For the shape of the negative binomial distribution, we impose a gamma prior, i.e.  $\phi \sim gamma(0.01, 0.01)$ . For the spatial term  $\boldsymbol{\eta}$ , we choose the exponentiated-quadratic kernel  $k(x_i, x_j) = \sigma_{gp}^2 \exp(-||x_i - x_j||^2/(2l_{\text{scale}}^2))$  and impose a flat prior on  $\sigma_{gp}^2$  and an informative inverse-gamma prior on  $l_{\text{scale}}$  as  $l_{\text{scale}} \propto inv - gamma(0.976289, 0.008892)$ , where the hyperparameters were estimated from the data.

One key advantage of our model is that traditional GLM regression diagnostics can be used, a significant limitation of other presence-only models. Furthermore, our cell-level approach can be extended to other distributions, like the zero-inflated Poisson, or non-zero inflated Poisson and negative binomial distributions, for example. The method requires only knowing the distribution of the thinned process and being able to estimate the parameters of the original process using only the thinned process.

### 3.4 Prediction and Calibration

To predict the number of FSW in each cell, we fit a venue size model and a venue count model independently. The venue size model estimates the number of FSW at a venue via

a hurdle log-normal distribution. This model is fit using the observed venue sizes. Using cell-level covariates, we predict the venue size for venues in all cells. The calibrated venue count model estimates the number of venues in a cell in three steps: (1) fit a ZINB model using observed venue counts, accounting for the thinning process in each district via an offset, (2) predict the venue count for all cells using the fitted model, and (3) calibrate the predicted venue counts in each district to match the known district-level venue counts. After modeling venue size and count, we predict the cell-level FSW sizes by multiplying the calibrated venue counts by the predicted venue sizes.

Predictions of the true venue counts can be made by setting  $\pi_{d[i]} = 1$  for all cells. For districts where district-level venue counts are known, we calibrate the predictions from each posterior sample of the venue count model so that the predicted venue count equals the known venue count at the district-level. For each district, this is done in two steps for each set of posterior samples: (1) obtain district-level venue estimates for each posterior sample by summing the predicted values for each cell with  $\pi_{d[i]} = 1$ , and (2) multiply each predicted value for the posterior sample by the calculated scaling factor  $\lambda$ , so that the sum of predicted values equals the known count. However, we perform no calibration for the remaining districts, as no information is available. Note that we simulate district random effects when they were not estimated from the model. These random effects are the PLACE I districts and remaining unvisited districts in the venue count model and the unvisited districts in the venue count and venue size models.

Note that step (1) already assumes the expected number of venues in each district is equal to the district-level venue counts provided by the PLACE report for districts included in the venue count model (only PLACE II districts). Step (3) then calibrates the venue counts for the remaining districts where district-level venue counts are known. To remain

consistent across districts, we calibrate the venue counts for all districts where the counts are known. If the district level venue counts are believed to have measurement errors, then the exact matching is not needed and step (3) could be omitted entirely or relaxed. For Malawi, the PLACE team interviewed thousands of community informants to find all possible venues, so we believe the district-level venue counts very accurate, and thus adopt the exact matching approach.

## 4 Model Validation

In this section, we check our model assumptions for the PLACE data, and where that is impossible, we use a simulation study to explore the behavior of the models when those assumptions are violated. We have four main assumptions which we check in the following order: (1) the true venue counts follow a zero-inflated negative binomial distribution, (2) venue size and venue count are conditionally independent, (3) the cell resolution provides accurate results, and (4) venues were sampled uniformly within each district. We can check the first two assumptions directly for our PLACE data, but we rely on a simulation study to evaluate the effect of cell-resolution and non-uniform sampling.

### 4.1 Real Data Validation

Checking the performance of presence-only models using real data is difficult because the truth is typically unknown. However, the PLACE team visited almost every venue in three districts, Mchinji, Mwanza, and Neno, which provides us with a validation data set where almost all true venue counts are observed. The number of observed venues against the number of operational venues for the complete districts are 98/102, 75/78, and 64/65,

respectively. We refer to these three districts as the complete districts, even though a few venues were not visited. Note that even though almost all venues were visited, the venue sizes are not available at all venues, so the district-level FSW sizes are still unknown. We use these complete districts to first show the appropriateness of our assumed ZINB distribution for the venue counts, and then to show the conditional independence between the venue counts and the venue sizes.

#### 4.1.1 Distribution of True Venue Counts

We first check that the true venue counts follow a ZINB distribution by fitting a non-thinned ZINB model with the venue counts from the complete district. We rely on posterior predictive p-values to verify the goodness of fit, and these results can be found in Table 3 in the Appendix. More details about posterior predictive p-values and test statistics can be find in Section 5 where we introduce the p-values for all of Malawi. Based on these values, we find the ZINB distribution provides a very reasonable goodness of fit for the complete districts. Visual diagnostics similar to those in Figure 8 were also considered for the complete districts, but are not included here as there is no clear evidence that the models do not adequately fit the data.

#### 4.1.2 Model Independence

Next, we provide evidence for the independence between the venue counts and venue sizes by examining the cell-level residuals. For the venue count model, the residuals are defined as the observed number of venues in cell  $i$  minus the predicted number of venues. For the venue size, the residuals are the average venue size in cell  $i$  minus the predicted venue size. However, since true-data is generally unknown, we limit our analysis to only the

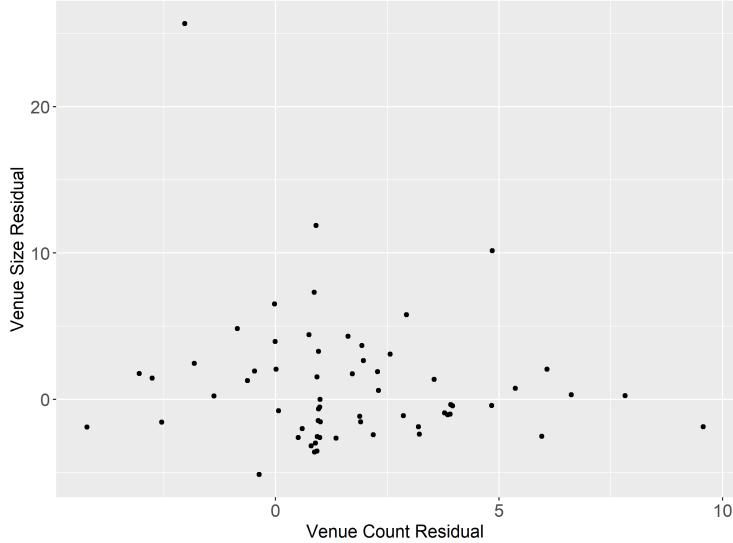


Figure 3: A scatter plot of venue size residuals and venue count residuals.

three complete districts and only the cells with at least one venue. Cells with zero venues are excluded because there is no average venue size. The scatter plot of the venue size residuals and venue count residuals is shown in Figure 3. The residuals do not appear to have any strong correlation structure, suggesting the conditional independence assumption is reasonable.

## 4.2 Simulation Study

To address the final two assumptions, we rely on a simulation study. We first show that cell resolution does affect total FSW size estimates, but it is difficult to determine the optimal cell resolution using residual diagnostics. We then show that if venues are sampled proportionally to a covariate, total FSW size is again biased, but including the relevant covariate helps mitigate most of the bias. We now introduce our simulation design. For

our simulation study, we control three factors: cell resolution, (non)uniform sampling, and whether all covariates related to nonuniform sampling are included in the regression model.

*Design of cell resolutions:* Our spatial domain for the simulations is the unit square. Specifically, we first simulate a spatial grid of covariates, divide the domain into a grid of 45 by 45 equal sized cells, yielding 2025 cells. In each cell, we simulate the counts, locations, and sizes of venues according to a ZINB distribution for the counts and a hurdle log-normal distribution for the sizes. Simulation parameters are chosen to resemble the PLACE data. These 2025 cells represent the “true resolution” since the venue counts at this resolution are simulated from a ZINB distribution. We then divide the domain into either 484 (large) cells, 5041 (small) cells, or use the true cells, and find the covariates in each cell. The large cells correspond to a grid of 22 by 22 equal sized cells and the small cells to 71 by 71 equal sized cells (relating these cell sizes back to the Malawi study, these are roughly equivalent to considering  $3 \times 3$ -km or  $1 \times 1$ -km cells instead of our  $1.5 \times 1.5$ -km cells).

*Design of (non)uniform sampling:* We thin the venues either uniformly (half of all venues are sampled with equal probability) or nonuniformly (half of all venues are sampled where the probability of sampling a venue is strongly correlated with one of the covariates). The observed venue counts are found for each cell resolution, our thinned ZINB model is fit to the counts and calibrated, and a hurdle log-normal model is fit to the sizes. For the fitting, we either fit the model with all covariates (all covariates) or exclude the covariate correlated to the sampling probabilities (missing covariate).

Total FSW size is found by predicting the count and size to each cell for each posterior sample, calibrating the counts, summing the cell-level FSW estimates across all cells, and finding the mean of each predicted total FSW size. This process is repeated 1000 times. The results for the simulations are shown in Figure 4, where the different cell sizes are

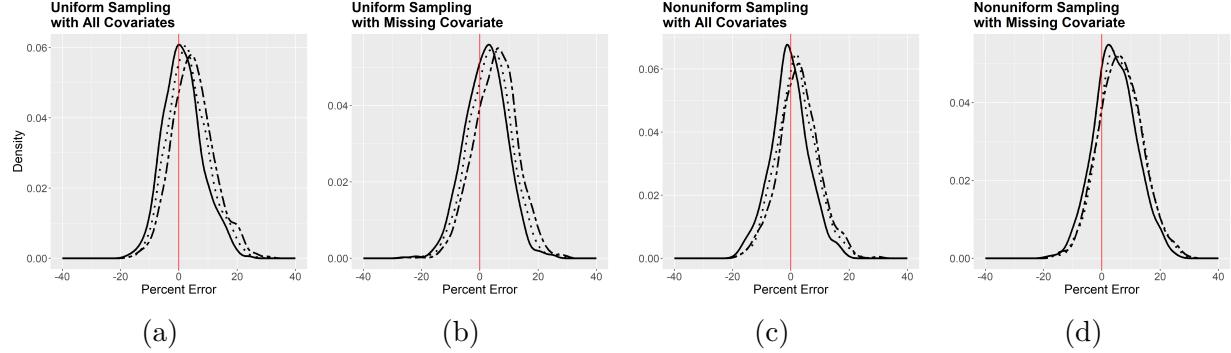


Figure 4: Distribution of total FSW estimates under different model assumptions and cell resolutions. We consider uniform sampling of venues with all covariates in the regression model (a), uniform sampling with a missing covariate (b), nonuniform sampling with all covariates (c), and nonuniform sampling with a missing covariate (d). The solid line corresponds to the true cell resolution, the dotted line to small cells, and the dot-dash line to large cells. Percent error is defined as  $(Y - \hat{Y})/Y$ , where  $Y$  is the true total FSW size.

shown as different line types, the different model assumptions are provided in the titles, and the percent error is defined as  $(Y - \hat{Y})/Y$ , where  $Y$  is the true total FSW size

#### 4.2.1 Cell Resolution

We first show that cell resolution affects total FSW size estimates. We can see in subplot (a) that under the ideal scenario when sampling is uniform and all covariates are included in the regression model, the percent error is centered around zero. Under all scenarios, the cell resolution does influence the results, but the distribution of posterior-means are mostly overlapping. Based on these simulations, there also does not appear to be any systematic bias based on cell-size. For scenario (a), the average percent errors for the same, small, and large cell resolutions are 1.39%, 3.19%, and 4.91%, respectively. The difference between the average percent errors is similar for the other scenarios. In all cases, the residual diagnostics of all cell resolutions appear to provide good fit, with no model obviously performing better

than another, suggesting that it is very difficult to choose the best cell resolution based only on the data.

#### 4.2.2 Uniform Sampling

Finally, we conclude from Figure 4 that while nonuniform venue sampling biases total FSW estimates, including the covariate related to sampling helps mitigate this bias. Comparing subplots (a) and (c), we can see that nonuniform sampling does introduce slight bias. For subplot (d), excluding the covariate related to sampling further biases the results. While results are best when sampling is uniform and all covariates are included in the model, the errors introduced by violations of these assumptions are still relatively small. For the “same” cell resolution, the average percent errors for each assumption scenario are 1.39% (a), 1.89% (b), -0.45% (c), and 4.56% (d), clearly showing the larger bias introduced by excluding the sampling covariate, but the differences between the average errors are relatively small compared to the variation introduced by observing only half of the venues.

## 5 Estimation of FSW in Malawi

Here we apply our model to the PLACE data and map the distribution of FSW in Malawi. We first divided Malawi into cells. The cell size should be small enough such that the behavior inside the cells are homogeneous while large enough that computation is feasible. To balance these two factors, we divided Malawi into cells of approximately 1.5x1.5-km. For the venue size model, we used all visited venues from both PLACE I and II studies. Due to the convenience sampling for venue locations in PLACE I, only PLACE II venues are used to model venue count, excluding Mzimba. Mzimba was removed because the largest city of

Table 1: Posterior predictive p-values for Malawi.

Test Statistics	Venue Count	Venue Size
Positive Mean	0.487	0.482
Positive Standard Deviation	0.680	0.523
Over-dispersion index	0.725	0.548
Maximum	0.855	0.777
Proportion of Zeros	0.521	0.507

Mzimba was visited in PLACE I, while only the rest of Mzimba was visited in PLACE II. Additionally, 15 venues with more than 100 recorded FSW were excluded from the venue size model because in all cases, the recorded number of women present at a busy-time was less than 100, clearly showing the very large counts are unreliable. This process leaves us with 20,751 cells for the venue count model and 2,541 venues for the venue size model.

All analysis was done using R (R Core Team, 2019). Covariates were standardized prior to fitting. Spatial kriging was done via the **fields** package (Nychka et al., 2017). Bayesian modeling was implemented using the Stan probabilistic programming language (Stan Development Team, 2019) and the **brms** package (Bürkner, 2017). Since the number of samples was fairly large, we fit an approximate Gaussian Process with 5 basis functions and a multiplicative constant of  $5/4$  for  $\eta_{i,\text{count},p}$ , as suggested in the **brms** package. The multiplicative constant defines the range over which predictions from the Gaussian Process should be computed, and complete details of the approximate Gaussian Process used are found in Riutort-Mayol et al. (2020).

We again investigate model fit via posterior predictive p-values and visual diagnostics. These p-values are shown in Table 1 for both the venue count and venue size model and visual posterior checks are shown in Figure 8 in the Appendix. For a given test statistic  $T(y)$  for data  $y$ , the posterior predictive p-value is defined as  $P(T(y^{\text{rep}}) > T(y)|y)$ , where

Table 2: Table of predictors in the final models. Districts are included as random effects in all models. The – and + indicate the direction of significant coefficients for each model at the significance level 0.05, • indicates the predictor was included for predictions but its coefficient was not significant, and a blank space indicates the predictor was not included in the model.

Predictor Label	Meaning	$\mu_{\text{size}}$	$p_{\text{size}}$	$\mu_{\text{count}}$	$p_{\text{count}}$
v155	Literacy			•	•
mv201	Total children ever born		-		
mv167	Times away from home in last 12 months	•			•
Built	Built-up index. Higher values roughly correspond to more buildings		+	•	+
hivpos	Percent HIV positive	•			
WorldPop	2015 population estimate	•	-	+	-
nightlight	Night-time light activity	+	-		-

$y^{\text{rep}}$  represents the potential replication of the data (Meng et al., 1994; Gelman et al., 2013). Extremely large or small posterior predictive p-values might indicate violations of model assumptions. The test statistics we consider are the positive mean ( $T(y) = E(y|y > 0)$ ), positive standard deviation ( $T(y) = SD(y|y > 0)$ ), the over-dispersion index ( $T(y) = Var(y)/E(y)$ ), the maximum ( $T(y) = \max(y)$ ), and the proportion of zeros ( $T(y) = P(y = 0)$ ). These test statistics were chosen because they check the main aspects of our zero-inflated distributions. Based on these metrics, the observed venue counts closely followed a ZINB distribution. Similarly, the hurdle log-normal distribution outperformed other standard models we considered like the ZINB and the values indicate no clear lack of fit.

Table 2 summarizes the variables selected in each model, where  $\mu_{\text{size}}$  represents the mean of the log-normal distribution on the log scale for the venue size,  $\mu_{\text{count}}$  the mean

of the negative binomial distribution for the venue count, and  $\mathbf{p}_{\text{size}}$  and  $\mathbf{p}_{\text{count}}$  for mean of the Bernoulli distribution for the venue size and count models, respectively. To avoid overfitting, variables were selected by removing the least significant variable sequentially from saturated occurrence and abundance models. The collection of nested models were compared using posterior predictive p-values, leave-one-out predictive error, and visual diagnostics. The selected model performed the best with respect to leave-one-out predictive error and the visual diagnostics and posterior predictive p-values did not indicate any poor model fit. Note that our goal is not testing the significance of associations but providing the spatial distribution of the venues and FSWs. The coefficients represent the conditional relationship given other covariates instead of the marginal relationship.

The district-level FSW size estimates are shown in Figure 5 as posterior densities. Stars, triangles, and circles indicate the PLACE Report estimates for PLACE I districts, PLACE II districts, and districts in neither PLACE study, respectively. There is no FSW estimate from the PLACE Report for Likoma. The population of Likoma is extremely small compared to the other districts, so we do not include credible intervals for Likoma in Figure 5. Not all 95% credible intervals cover the PLACE report district estimates, although most estimates are very close. Note, however, that the PLACE estimates are not gold standards for the district-level FSW estimates. We expect the PLACE district estimates to be larger than our estimates because the PLACE method likely oversamples from large venues in populated areas. We find this to be the case for many, but not all districts. The random effects were not estimated for PLACE I districts in the venue count model or for districts in neither PLACE study in the venue size and count models, so these districts have larger credible intervals due to simulating the random effects.

We also present the estimated scaling factors  $\lambda$  for each district in Figure 6. The  $\lambda$

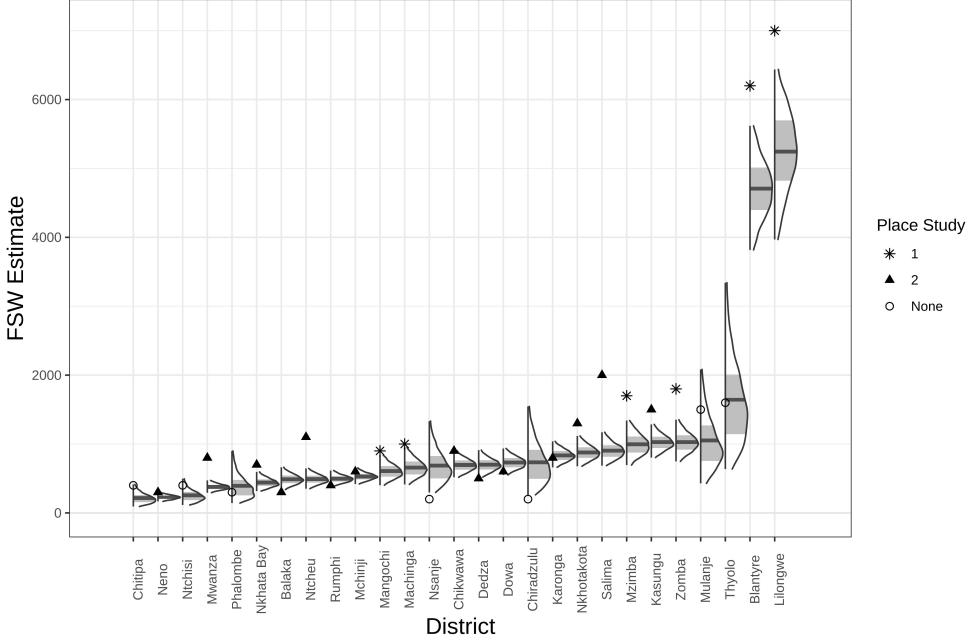


Figure 5: District-level FSW size 95% credible intervals for all districts in Malawi. The light gray area indicates the 50% credible interval and the dark gray lines indicate the posterior mean. Stars, triangles, and circles indicate the PLACE Report estimates for PLACE I districts, PLACE II districts, and districts in neither PLACE study, respectively.

represents the discrepancy between the known total number of venues and the estimated total number of venues in each district, where  $\lambda = 1$  corresponds to perfectly estimated venue counts and  $\lambda > 1$  corresponds to underestimated venue counts. For districts where venue count data was included in the model (PLACE II) and the data informs the district random effects, the scaling factors are all very close to  $\lambda = 1$ . For PLACE I districts, district-level venue counts are still reasonably close to the truth, although the total number of venues is consistently overestimated in Mangochi and slightly underestimated in Blantyre and Lilongwe. One possible explanation is that there are large district random effects that are not explained by the covariates. The random effect could be related to social,

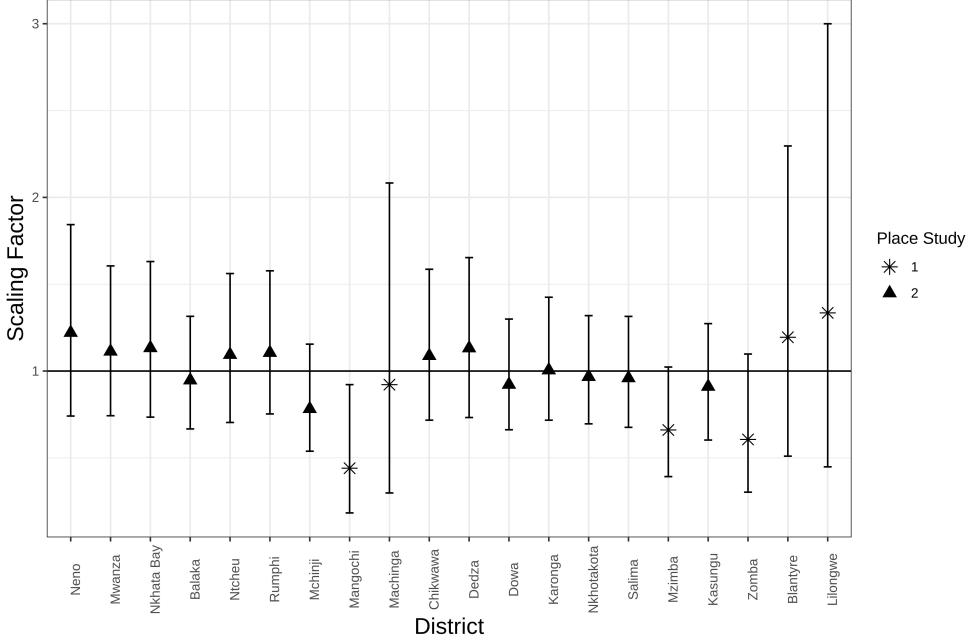


Figure 6: Scaling factors  $\lambda$  for each districts. Stars, and triangles indicate the mean  $\lambda$  for PLACE I and PLACE II districts, respectively (no scaling factors can be estimated for other districts). The error bars indicate the lower and upper 95% quantiles of the estimated  $\lambda$ . A flat line at  $\lambda = 1$  is shown for reference.

environmental, or political factors that the covariates cannot capture. The difference could also be a result of different efforts in identifying venues in the data collection stage. In particular, Mangochi is the only district where the target number of community informants was not met, potentially leading to an underestimation of the number of venues in the PLACE report. Overall, the model reasonably estimates the total number of venues in each district, even when venue count data is not used, as is the case for PLACE I districts.

For privacy reasons, we omit a cell-level mapping of the FSW in Malawi at the resolution used in the model fitting. Instead, we aggregate to much larger cells, approximately  $4.8 \times 4.8\text{-km}$ . Maps of predicted FSW size, FSW size on the log-scale, and the log of the variance

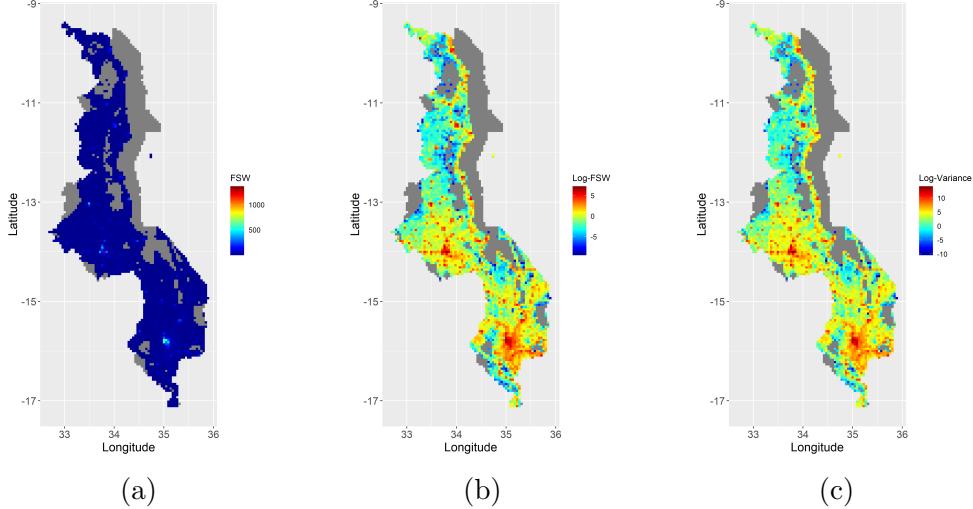


Figure 7: Maps of FSW counts in each cell in Malawi (a) and on the log-scale (b). The log of the variance estimates for FSW counts are shown in (c). Cells with estimated zero FSW are shown as gray.

estimates of FSW size are shown in Figure 7.

## 6 Conclusion

In this paper, we have introduced a fully model-based approach for modeling marked presence-only data. The Bayesian implementation allows us to easily incorporate other data-related issues aside from presence-only, like spatial dependency, while also seamlessly producing credible intervals. The zero-inflated model can, however, be fit using frequentist methods, although confidence intervals and other considerations may be more difficult to calculate.

The results in this paper are critical for both informing broad country and district level targets set by UNAIDS and others as well as local community based outreach programs.

National organizations use the larger scale FSW size estimates to properly fund programs to implement HIV prevention and treatment services. The cell-level FSW size estimates are useful for implementing community-based programs like short 30-minute session to teach FSW to negotiate condom use (Kerrigan et al., 2015). Furthermore, fine-scale mapping can further improve the distribution of HIV services. Access to these services is often unequal, resulting in immediate access for some but long wait times, up to years, for others (Joint United Nations Programme on HIV/AIDS, 2021).

It is important to note that FSW may also work at multiple venues and thus be counted at multiple venues, and FSW who do not work at venues are not counted. Because of this, our goal was not to map the location of all FSW in Malawi, but to map the FSW who attend venues. It is realistic that one may encounter the same FSW at two different venues on two different days due to the mobility. The Malawi PLACE study minimized this impact by asking for the number of FSW who visit the site between 11 p.m. and 2 a.m. on Saturday night. By modeling this data directly, our results also reflect this possibility. However, our analysis does not address FSW net migration across multiple years. FSW do not remain in one location, and major changes can take place due to key events, like the introduction of new policy measures which affect the operation of venues. Thus, our analysis and results provide a reliable map for 2016 and 2017, the time frame of the PLACE study, but do not accurately portray the uncertainty in estimates for other years.

The sampling probabilities  $\pi_{d[i]}$  could be refined if a more sophisticated sampling process was used and recorded. Our model can also solve the problem of nonuniform sampling in presence-only studies. Consider a typical presence-only data set where observations come from respondent reports. In these cases, more frequently traveled areas generate more responses. If some measure of “popularity” can be created for each cell, then our ZINB model

naturally accounts for the varying levels of exposure across the domain. This approach is more natural and theoretically justified given the assumed model than other presence-only approaches, especially those that require choosing pseudo-absences in complicated ways.

The model could also be extended to model venue size and venue count jointly. However, there does appear to be very little correlation between venue size and venue count after accounting for the covariates and this would greatly increase the computation time. Another question that arises is whether interviewing more venues or recording the locations of unvisited venues offers better predictive power. For example, if the survey team can skip an interview at one venue in order to record the location of ten more unvisited venues, would the size of prediction confidence intervals increase or decrease? This is an interesting follow-up question that has real-world application for future sampling procedures.

## References

- Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.
- Carlin, B. P., Xia, H., Devine, O., Tolbert, P., and Mulholland, J. (1999). Spatio-temporal hierarchical models for analyzing atlanta pediatric asthma er visit rates. In *Case studies in Bayesian statistics*, pages 303–320. Springer.
- Gelfand, A. E., Zhu, L., and Carlin, B. P. (2001). On the change of support problem for spatio-temporal data. *Biostatistics*, 2(1):31–45.

Gelman, A. et al. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, 7:2595–2602.

Joint United Nations Programme on HIV/AIDS (2014a). 90-90-90: an ambitious treatment target to help end the aids epidemic. [https://www.unaids.org/sites/default/files/media\\_asset/90-90-90\\_en.pdf](https://www.unaids.org/sites/default/files/media_asset/90-90-90_en.pdf). Geneva: Unaids.

Joint United Nations Programme on HIV/AIDS (2014b). The Gap Report. [http://files.unaids.org/en/media/unaids/contentassets/documents/unaidspublication/2014/UANAIDS\\_Gap\\_report\\_en.pdf](http://files.unaids.org/en/media/unaids/contentassets/documents/unaidspublication/2014/UANAIDS_Gap_report_en.pdf). Geneva, Switzerland.

Joint United Nations Programme on HIV/AIDS (2018). UNAIDS Data 2018. [https://www.unaids.org/sites/default/files/media\\_asset/unaids-data-2018\\_en.pdf](https://www.unaids.org/sites/default/files/media_asset/unaids-data-2018_en.pdf). Geneva, Switzerland.

Joint United Nations Programme on HIV/AIDS (2021). End Inequalities . End Aids . Global Aids Strategy 2021-2026.

Kerrigan, D., Kennedy, C. E., Morgan-Thomas, R., Reza-Paul, S., Mwangi, P., Win, K. T., McFall, A., Fonner, V. A., and Butler, J. (2015). A community empowerment approach to the hiv response among sex workers: effectiveness, challenges, and considerations for implementation and scale-up. *The Lancet*, 385(9963):172–185.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14.

Measure Evaluation (2018). How the priorities for local aids control efforts (place) method helps prevent hiv transmission and reach the 90-90-90 targets. <https://www.measureevaluation.org/resources/publications/fs-15-143>.

Meng, X.-L. et al. (1994). Posterior predictive  $p$ -values. *The Annals of Statistics*, 22(3):1142–1160.

Mugglin, A. S., Carlin, B. P., and Gelfand, A. E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, 95(451):877–887.

NOAA (2016). Version 1 viirs day/night band nighttime lights. [https://www.ngdc.noaa.gov/eog/viirs/download\\_dnb\\_composites.html](https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html).

Nychka, D., Furrer, R., Paige, J., and Sain, S. (2017). fields: Tools for spatial data. R package version 9.8-6.

Pearce, J. L. and Boyce, M. S. (2006). Modelling distribution and abundance with presence-only data. *Journal of applied ecology*, 43(3):405–412.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259.

Phillips, S. J., Dudík, M., and Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83.

Puig, P. and Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli*, pages 544–555.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379.

Riutort-Mayol, G., Bürkner, P.-C., Andersen, M. R., Solin, A., and Vehtari, A. (2020). Practical hilbert space approximate bayesian gaussian processes for probabilistic programming. *arXiv preprint arXiv:2004.11408*.

Stan Development Team (2019). RStan: the R interface to Stan. R package version 2.19.2.

UNC, CEDEP, NAC, and FHI360 (2018). Place report malawi: September 2018.

Ward, G., Hastie, T., Barry, S., Elith, J., and Leathwick, J. R. (2009). Presence-only data and the em algorithm. *Biometrics*, 65(2):554–563.

WorldPop (2015). Malawi 100m population. <https://www.worldpop.org/geodata/summary?id=123>.

Young, L. J., Gotway, C. A., Yang, J., Kearney, G., and DuClos, C. (2009). Linking health and environmental data in geographical analysis: it's so much more than centroids. *Spatial and spatio-temporal epidemiology*, 1(1):73–84.

# Appendices

## A Proofs

### A.1 Proof of Proposition 1

*Proof.* Let  $Y$  and  $X$  be as described in Proposition 1. Then the pmf of  $Y$ , the thinned process, is given by,

$$\begin{aligned}
P(Y = y) &= \sum_{x=0}^{\infty} P(Y = y|X = x)P(X = x) \\
&= \sum_{x=y}^{\infty} \binom{x}{y} \pi^y (1 - \pi)^{x-y} \binom{x + \phi - 1}{x} \left(\frac{\mu}{\mu + \phi}\right)^x \left(\frac{\phi}{\mu + \phi}\right)^{\phi} \\
&= \left(\frac{\pi}{1 - \pi}\right)^y \left(\frac{\phi}{\mu + \phi}\right)^{\phi} \frac{1}{y!(\phi - 1)!} \sum_{x=y}^{\infty} \frac{(x + \phi - 1)!}{(x - y)!} \left(\frac{(1 - \pi)\mu}{\mu + \phi}\right)^x \\
&= \left(\frac{\pi}{1 - \pi}\right)^y \left(\frac{\phi}{\mu + \phi}\right)^{\phi} \frac{1}{y!(\phi - 1)!} (y + \phi - 1)! \left(\frac{\mu(1 - \pi)}{\mu + \phi}\right)^y \left(\frac{\mu + \phi}{\mu\pi + \phi}\right)^{y+\phi} \\
&= \binom{y + \phi - 1}{y} \left(\frac{\pi\mu}{\pi\mu + \phi}\right)^y \left(\frac{\phi}{\pi\mu + \phi}\right)^{\phi},
\end{aligned}$$

which is exactly the negative binomial pdf with mean  $\pi\mu$  and dispersion  $\phi$ .  $\square$

### A.2 Proof of Proposition 2

*Proof.* Let  $\tilde{Y}$ ,  $Z$ ,  $X$ ,  $X^*$ , and  $Y$  be as described in Proposition 2. First, let us construct a zero-inflated distribution with the  $\pi$ -thinning of  $X$ ,  $Y^* = (1 - Z)X^* = (1 - Z)\sum_{i=1}^X I_{y^*,i}$ . Second, consider the  $\pi$ -thinning of the original zero-inflated distribution,  $Y = \sum_{i=1}^{\tilde{Y}} I_{y,i} =$

$\sum_{i=1}^{(1-Z)X} I_{y,i}$ . Conditioning on  $Z$  for both distributions, we find

$$P(Y^* = y|Z = 0) = 1 - p$$

$$P(Y^* = y|Z = 0) = P\left(\sum_{i=1}^X I_{y^*,i} = y\right) p$$

and

$$P(Y = y|Z = 0) = 1 - p$$

$$P(Y = y|Z = 0) = P\left(\sum_{i=1}^X I_{y,i} = y\right) p$$

Since  $I_{y^*,i} \stackrel{D}{=} I_{y,i}$ , it follows that  $Y^* \stackrel{D}{=} Y$ .  $\square$

### A.3 Proof of Proposition 3

*Proof.* Let  $\tilde{Y}$ ,  $Y$ , and  $I_k$  be as defined in Theorem 3. From Proposition 2,  $Y$  is also a zero-inflated random variable,  $Y = (1 - Z)X^*$ , where  $X^* = \sum_{i=1}^X I_{x,i}$ ,  $I_{x,i} \sim Bern(\pi)$ . From Proposition 1,  $X^* \sim NegBin(\text{mean} = \pi\mu, \text{dispersion} = \phi)$ . Thus,  $Y$  is a zero-inflated negative binomial negative binomial, with negative binomial mean  $\pi\mu$  and dispersion  $\phi$ .  $\square$

## B Additional Diagnostics

Table 3: Posterior predictive p-values for the complete districts.

Test Statistics	Venue Count	Venue Size
Positive Mean	0.537	0.546
Positive Standard Deviation	0.560	0.514
Over-dispersion index	0.566	0.508
Maximum	0.496	0.584
Proportion of Zeros	0.482	0.483

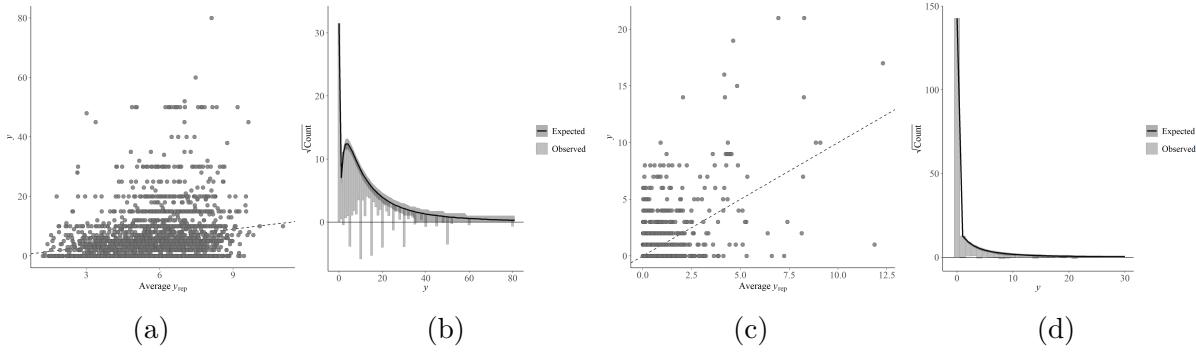


Figure 8: Diagnostic plots for venue size and count. Subplots (a) and (b) concern the venue size while subplots (c) and (d) concern the venue count. (a) and (c) are scatter plots of average predicted values against the observed predicted counts. (b) and (c) are hanging rootograms and plot the square roots of the counts against the response variable. The observed counts are shown as light gray bars while the expected counts and corresponding uncertainty are shown as a black line with dark gray shadow.