Distribution Agreement

Signature:

Shannon K. McClintock Date

Model-Based Statistical Methods for Public Health Surveillance Subject to Imperfect Observations

By

Shannon K. McClintock
Doctor of Philosophy

Biostatistics

Lance A. Waller, Ph.D.
Advisor

Andrew Hill, Ph.D.
Committee Member

Qi Long, Ph.D.
Committee Member

Matthew Strickland, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

Date

Model-Based Statistical Methods for Public Health Surveillance Subject to Imperfect Observations

By

Shannon K. McClintock
M.S., Emory University, 2011
B.A., East Carolina University, 2005

Adviser: Lance A. Waller, Ph.D.

An Abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2012

Abstract

Model-Based Statistical Methods for Public Health Surveillance Subject to Imperfect Observations

By Shannon K. McClintock

We examine statistical modeling issues in three areas of public health surveillance: estimation of vaccination coverage, linking local observations and remotely sensed covariates, and adjustment for zero inflation due to underreporting.

When the proportion of the vaccinated population is an unknown value less than 100%, we explore application of logistic growth models, namely the standard logistic growth model and a reparameterization naturally constraining vaccination coverage parameter estimates. We compare the performance of three methods of estimation for each model (nonlinear least squares, maximum likelihood estimation, and Bayesian estimation).

Buruli ulcer is a neglected tropical disease affecting Australia and West Africa. We examine both on-site local water characteristics and broad scale remotely sensed environmental attributes with respect to the presence of the causative pathogen, *Mycobacterium ulcerans*. Our findings support hypotheses regarding conditions suitable for *M. ulcerans* growth, but diverge from other published results regarding the distribution of and factors related to Buruli ulcer disease. In addition, our findings suggest locations of reported cases and pathogen presence need not coincide, supporting the notion that human interaction with the environment plays a role in transmission.

In Buruli ulcer surveillance, districts which do not report cases are programmatically treated as districts without cases but are not actually confirmed as disease-free districts. Moreover, there is substantial reason to believe that some non-reporting districts actually have cases; consequently, our data are subject to 'false' zeros. We evaluate the performance of the zero inflated Poisson model in the presence of false zeros, as well as propose a hierarchical zero inflated Poisson model with the ability to estimate an observation's conditional probability of being a false zero given that a zero was observed.

Model-Based Statistical Methods for Public Health Surveillance Subject to Imperfect Observations

By

Shannon K. McClintock
M.S., Emory University, 2011
B.A., East Carolina University, 2005

Adviser: Lance A. Waller, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2012

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The World Health Organization (WHO) defines public health surveillance as "an ongoing, systematic collection, analysis and interpretation of health-related data essential to the planning, implementation, and evaluation of public health practice,"[1]. This scientific discipline is necessary to reduce the spread of disease as well as to maintain healthy populations. By its nature and broad scope, surveillance data can be challenging to collect; furthermore, the data available may not always be able to directly answer the question of public health importance for which it was collected. Motivated by challenges in modeling vaccination coverage and monitoring of neglected tropical diseases, we present a general framework for statistical methodologies applicable to unique niches in public health surveillance.

1.2 Vaccination Coverage

Vaccination has a dual role of protecting an individual from vaccine preventable diseases as well as reducing rates of vaccine preventable diseases in a community. In addition, vaccination status is often used as a marker of a child's health status and adequacy of care [2]. Estimation of coverage is useful for monitoring and evaluation of vaccination programs, determining if the population coverage necessary for disease elimination has been achieved, and assessing the health services available to children in a community. Much of the vaccination literature assesses vaccination coverage at specific age intervals [2, 3]. However, simple point estimation of the proportion of children covered at specific age intervals does not provide a flexible framework for additional inference that may be useful for the public health researcher. The proposed research seeks to develop methods that provide accurate and reliable estimates of vaccination coverage as a function of age.

1.3 Neglected Tropical Diseases

Neglected tropical diseases (NTDs) are diseases that have been mostly eliminated from developed nations yet still persist in developing countries, often affecting the poorest populations. Of the 14 NTDs identified by the World Health Organization, most can be treated, eliminated, and prevented [4]. Research on NTDs frequently involves the intersection of a wide range of technologies: from high-tech remote sensing, to mid-tech on the ground surveillance with mobile phones or surveys with PDAs, to low-tech paper data collection forms. Linking all of the data through statistical models is important, challenging, and offers an opportunity to aid in surveillance, treatment, and control efforts. Buruli ulcer is a neglected tropical skin disease caused by *Mycobacterium ulcerans* (MU) and is highly endemic in West Africa. We focus on evaluating associations with environmental predictors specific to the presence of the pathogen that causes Buruli ulcer, *M. uclerans.*

1.4 Surveillance of NTDs

Surveillance data on NTDs from developing nations can have several limitations. Often within disease notification and reporting systems, case reports are submitted and confirmation that non-reporting areas are actually disease-free is not obtained. While this may be a reliable system in developed nations monitoring high profile diseases, for monitoring of NTDs in developing nations there may be substantial reason to believe that cases are present in non-reporting areas. This creates 'false zeros' in surveillance data. We seek to determine the impact of false zeros on estimation of the mean rate of cases, and to determine if it is possible to distinguish false zeros from true zeros.

Chapter 2

Constraining Parameter Estimates in a Logistic Growth Model

2.1 Overview

The United States Agency for International Development (USAID) began implementing Demographic and Health Surveys (DHS) to collect data in the areas of population, health, and nutrition from women aged 15-49 in 1984. To date, the nationally administered household surveys have been performed in more than 50 countries [5]. One specific focus of the DHS is the timing, completeness, and drop out rates for children's immunizations [5]. Nine childhood vaccinations are recommended by the World Health Organization (WHO) for African countries: one for tuberculosis at birth, four for polio (birth, 6, 10, and 14 weeks), three for DPT (6, 10, and 14 weeks), and one for measles at 9 months [6]. While timely administration of vaccines is paramount to the efficacy of vaccination, often vaccinations do not adhere to the recommended schedule. We are interested in assessing the coverage of the the combined diphtheria, pertussis, and tetanus vaccine (DPT) via the 2003 Kenya Demographic and Health Survey (DHS).

The DHS asks mothers for vaccination information for all children under the age of 5. Whether or not a child was vaccinated and date of vaccination can be gathered by either the child's official vaccination card or maternal recall [7]. The preferred source of data is the vaccination card; however, often the card is not available (lost or never obtained), or information on the card is difficult to decipher [3]. Maternal recall is thought to be reliable for younger children, though as children get older mothers may have forgotten the child's vaccination history. An assessment of DHS surveys conducted worldwide from 1993-2003 found that 50.1% of children had a health card and showed it to the interviewer [6], and a summary report of 28 DHS studies from 1990 to 1994 found that card retention rates varied from 35.1% (Bolivia) to 87.8% (Rwanda) [3]. The 2003 Kenya DHS reports that 60% of children had available vaccination cards [7].

2.2 Introduction

Vaccination has a dual role of protecting an individual from diseases as well as reducing rates of vaccine preventable diseases in a community. In addition, vaccination status is often used as a marker of a child's health status and adequacy of care. Estimation of vaccination coverage, i.e. the proportion of individuals vaccinated, is useful for monitoring and evaluation of vaccination programs, determining if the population coverage necessary for disease elimination has been achieved, and assessing the health services available to children in a community. Much of vaccination literature assesses vaccination coverage at specific age intervals [2, 3]. However, simple point estimation of the proportion of children covered at specific age intervals does not provide a flexible framework for additional inference that may be useful for the public health researcher.

Several researchers have proposed the use of survival analysis techniques to model time to vaccination in order to assess the timeliness of vaccination as well as the proportion of a population vaccinated [8, 9, 10, 11, 12, 13]. Such studies model the cumulative probability of vaccination as one minus the Kaplan-Meier survival function, and children included in the study who had not yet been vaccinated at the time of the interview are considered to be right-censored observations. Furthermore, comparisons of vaccination rates in different subgroups of the population can be implemented through either the log-rank test for Kaplan-Meier survival curves or through a Cox proportional hazards model.

While survival analysis methods are useful for estimating the timeliness of vaccination, in many cases they are not suited to accurately estimate vaccination coverage because this is generally provided by empirical estimates of the tail end of the "inverse" Kaplan-Meier curve. Some authors note that this tail end of the curve is generally estimated by fewer observations and can be statistically unstable. Hence, this is an indirect approach to estimating vaccination coverage at a certain age, and caution is recommended when interpreting this estimate [8].

Another limitation of existing survival analysis methods to estimate vaccination coverage is that they utilize exact information on the date of birth and date of vaccination of the child. While developed nations are often able to collect such data through vaccination cards, card retention rates can vary greatly. A summary report of 28 Demographic and Health Surveys (DHS) studies from 1990 to 1994 found card retention rates varying from 35.1% (Bolivia) to 87.8% (Rwanda) [3]. In such surveys, whether or not a child was vaccinated and date of vaccination can be gathered by either the child's official vaccination card or maternal recall [7]. If a vaccination card is not available (lost or never obtained), mothers may verbally indicate if a child has been vaccinated or not. Restricting analysis to only children who retain health cards could severely bias vaccination coverage estimates. Overall vaccination coverage may be underestimated as children who do not have health cards but were nevertheless vaccinated would be excluded from the analysis. On the other hand, limiting analysis to children who retain health cards could also overestimate vaccination coverage as this could exclude individuals who did not have access to a health clinic from the analysis.

Therefore, instead of modeling time to vaccination, another option is to model the probability of a child receiving vaccination (indicated either by maternal recall or vaccination card) as a function of the child's age at the time of the interview. This allows inclusion of all children regardless of whether or not they retained a vaccination card. However, standard models for binary outcomes, such as logistic or probit regression, estimate the probability of response on the full range from 0-100% [14]. Without modification, these models cannot be used directly to estimate a probability of response with an asymptote less than one, which is the case with vaccination coverage.

We propose two versions of the three parameter nonlinear logistic growth model to estimate vaccination coverage. In addition to directly estimating coverage, this model also provides estimates of median vaccination age and characterizes the time elapsed for vaccination uptake in a population. Section 2.3 introduces the model, Section 2.4 discusses various methods of estimation, and Section 2.5 provides simulation results comparing approaches. Lastly, Section 2.6 applies the model to the Kenya 2003 DHS to estimate coverage of the combined diphtheria, pertussis, and tetanus (DPT) vaccination. We conclude with a discussion in Section 2.7.

2.3 The Logistic Model

Throughout the 20^{th} century a wide range of scientific disciplines have embraced various versions of the logistic model to estimate sigmoidal non-linear functions. Specific applications of the logistic growth model include ecologic population growth models, bioassay (quantal or quantitative), and epidemiologic risk models. These applications share the characteristic that an upper bound for specific quantities may be unknown.

Academic literature on population growth models dates back to the early 19^{th} century. Discussion in this realm began when the Reverend Thomas Robert Malthus (1766-1834) introduced the notion that human population growth may be limited by its natural resources. In 1838 the Belgian professor of mathematics Pierre Verhulst introduced a differential equation to model this population growth with a carrying capacity

$$\frac{\mathrm{d}\,N}{\mathrm{d}\,t} = rN\left(\frac{K-N}{K}\right) \tag{2.1}$$

where N is the population, r is the growth rate, and K is the carry capacity [15]. This original formulation is usually considered too simple to model real life processes, and nowadays partial differential equations are often used in the field of ecology for population growth models.

In 1920,"vital" statisticians Pearl and Reed presented a population growth model for studying the population of the United States [16]. They analyzed census data recorded from 1790 to 1910, comparing the results from polynomial models to the population growth model. They described a general model

$$y = \frac{be^{ax}}{1+ce^{ax}} \tag{2.2}$$

and chose to use a form of that model

$$y = \frac{b}{e^{-ax}+c} \tag{2.3}$$

to estimate the US population for a given future year as well as the "carrying capacity" of the United States, or asymptotic population ceiling as time goes to infinity. Of (2.2), the authors note that the curve starts at 0 when $x = -\infty$, asymptotes to a constant k when $x = +\infty$, has a point of inflection, and varies continuously from 0 to the asymptotic constant k when $x \in (-\infty, +\infty)$. Pearl and Reed used an ad-hoc method to estimate parameters, and suggested that further publications would focus on parameter estimation. In 1922, they published a follow-up article with properties of the growth model and its relationship to the differential equation presented by Verhulst [17]. Pearl (1927) formally dubbed models of form (2.2) as "logistic" in a tribute to the name Verhulst originally prescribed. In this publication, Pearl describes in depth how the logistic model can be used to describe growth in populations, from bacteria to human.

Joseph Berkson suggested a related yet slightly different form of the logistic function in 1944 for the analysis of quantal response bioassay data where the response is the proportion p affected out of n exposed [18]. In 1953, Berkson presented the the logistic function as the logistic regression model statisticians recognize today [19].

$$P = 1 - Q = \frac{1}{1 + e^{-(\alpha+\beta x)}}, \quad \text{logit}(P) = \ln\left(\frac{P}{Q}\right) = \alpha + \beta x \tag{2.4}$$

These bioassay problems typically involved exposing animals to a variety of doses and observing a dichotomous response. Berkson proposed obtaining parameter estimates by minimizing the "logit χ^2" quantity

$$\chi^2(\text{logit}) = \sum npq(l - \hat{l})^2 \tag{2.5}$$

where $l = \ln(p/q)$ and $\hat{l} = \ln(\hat{p}/\hat{q})$, due to the limited computing abilities at the time. Nowadays, parameter estimates are typically obtained by the iteratively re-weighted least squares, or otherwise known as Fisher's scoring algorithm [14].

In reviewing applications of the logistic regression model, Berkson discusses that some assumptions may be unreasonable. Specifically, he notes that it is necessary to have an infinitely large dose x in order for P to achieve 100% response; similarly, a dose of zero is necessary for $P = 0$. Berkson states that these assumptions are unrealistic because in reality an animal would not need an infinitely large dose to achieve 100% response; rather, there would be a threshold dose at which 100% response would be achieved.

Similar to Berkson's logistic function (2.4), Oliver (1964) discusses methods of estimating the logistic growth function, parameterized as

$$y = \frac{k}{1 + b\exp(-at)} \tag{2.6}$$

Of Berkson's logistic function (2.4), Oliver notes "the implication is that, over time, almost every member of the population eventually receives the characteristic," [20]. Oliver's parameterization allows for a limiting value of the response of interest, k, and he advocates use of a least squares method for parameter estimation.

Rodbard and Frazier (1975) discuss various models appropriate for data from radioimmunoassays [21]. From the Fourth International Biometrics Congress in Hannover 1970, they use Finney's proposed four parameter logistic model to analyze such data. Radioimmunoassay (RIA) is a procedure used to measure existing antigens in a system without invoking responses from an actual living organism or tissue. It is carried out by mixing known quantities of radioactive antigen to antibodies, adding unlabeled antigen to the solution, and measuring the displaced labeled antigen. RIA essentially creates a dose-response problem in which the concentration of bound antigen is a nonlinear function of the initial quantities of antigen and antibody in the system.

$$y = \frac{a - d}{1 + (X/c)^b} + d \tag{2.7}$$

In this equation, the response y is the count of bound antigens and the predictor X is the dose, or quantity of unlabeled antigen. The parameters a, b, c, and d have the following interpretations: a is the estimated response when $X = 0$, d is the response when $X = \infty$, c is the value of dose that gives a response at $(a+d)/2$ (also known as effective dose 50 or ED50), and b shapes the slope at the center of the curve. Rodbard and Frazier note that Newton-Rhapson, Gauss-Newton, and Marquardt-Levenberg are all acceptable methods by which to obtain parameter esimtates. However, they also note that due to the nonlinearity of the model and the interdependence among the parameters convergence may be difficult to achieve. Various parameterizations of the logistic growth model have been considered for radiogland and other related assays, and Ratkowsky and Reedy (1986) discuss these parameterizations as well as guidelines for choosing the appropriate parameterization [22].

While all of the previously described models take slightly different forms, they can all be re-parameterized by changing the asymptote from a constant (one) to an estimable parameter or by changing the explanatory variable x to $\log(x)$ in order to relate back to one another. In the applications of Verlhust, Pearl and Reed, and Rodbard and Frazier, the upper bound of the logistic curve is an unknown yet estimable quantity where the outcome of interest y takes on continuous values greater than zero, and the explanatory variable is often time. In the applications of Berkson, the upper bound of the logistic curve is fixed and known at 1, the outcome of interest takes on either dichotomous 0/1 outcomes (or count outcomes of responses out of number of trials), and the explanatory variable is dosage. To analyze vaccination data, we would like to estimate an unknown upper bound less than one (the asymptotic vaccination coverage) where the outcome of interest (whether or not vaccinated) is dichotomous, as a function of age.

For our application, we begin with a version of the three parameter logistic growth model presented by Pinheiro and Bates [23]

$$\Pr(Y_i = 1) = \frac{\phi_1}{1 + \exp\left(-\frac{(x_i - \phi_2)}{\phi_3}\right)}. \tag{2.8}$$

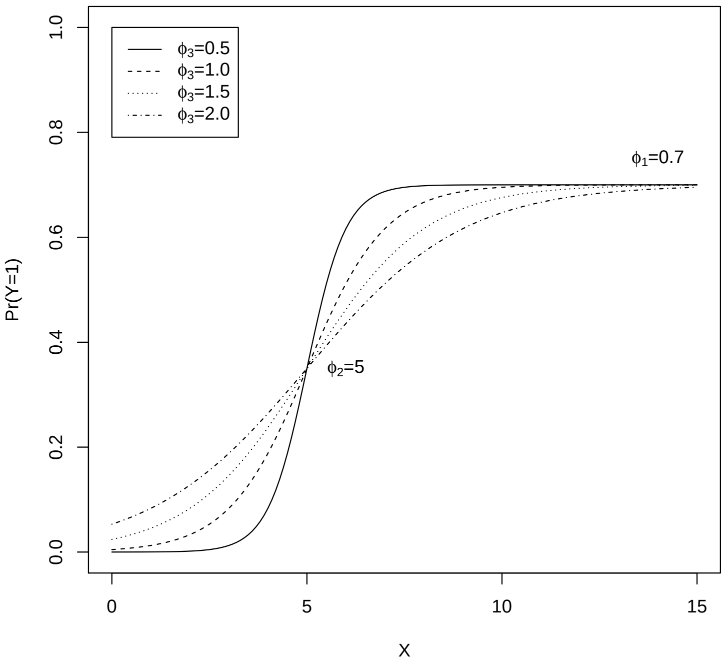


Figure 2.1: Shape of the nonlinear logistic model given $\phi_1 = 0.7$, $\phi_2 = 5$, and varying ϕ_3.

The outcome Y_i is dichotomous, and $Y_i = 1$ indicates that the child was vaccinated by the time of the interview; the covariate x_i is age. The parameter ϕ_1 represents the limiting proportion of the population vaccinated, ϕ_2 is the age at which probability of vaccination reaches $\frac{1}{2}\phi_1$, and ϕ_3 characterizes the rate of vaccination and estimates the age elapsed between the probability reaching $\frac{1}{2}\phi_1$ and $\approx \frac{3}{4}\phi_1$. Figure 2.1 displays the shape of the nonlinear logistic curve.

As ϕ_1 represents a proportion, it should logically be constrained within the interval $(0, 1)$. However, imposing constraints on a parameter estimate is a challenging task. Therefore, we also examine a reparameterized version of Model (2.8) inherently constraining the numerator, i.e.

$$\Pr(Y_i = 1) = \frac{\frac{1}{1+\exp(-\lambda)}}{1 + \exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)}. \tag{2.9}$$

While λ can take on any values in $(-\infty, +\infty)$, the numerator is coerced to stay in $(0,1)$ through reparameterization. From Model (2.9), estimates of $\hat{\phi}_1$ can be recovered by $\hat{\phi}_1 = \frac{1}{1+\exp(-\hat{\lambda})}$. Asymptotic confidence intervals for ϕ_1 can be created by first calculating asymptotic confidence intervals for λ via $\hat{\lambda} \pm z_{1-\frac{\alpha}{2}} * SE(\hat{\lambda})$ resulting in the interval $(\hat{\lambda}_L, \hat{\lambda}_U)$. Then apply the transformation $\left(\frac{1}{1+\exp(-\hat{\lambda}_L)}, \frac{1}{1+\exp(-\hat{\lambda}_U)}\right)$ to create a confidence interval for ϕ_1.

The similarities between the logistic function proposed by Berkson and the three parameter logistic growth model proposed should be noted. In regular logistic regression with one predictor we have

$$Pr(Y=1) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \tag{2.10}$$

and in the three parameter logistic growth model we have

$$Pr(Y=1) = \frac{\phi_1}{1+e^{\frac{-(x-\phi_2)}{\phi_3}}} = \frac{\phi_1}{1+e^{-\left(-\frac{\phi_2}{\phi_3}+\frac{x}{\phi_3}\right)}}$$

Therefore, these models are equivalent when $\phi_1 = 1$ if we let

$$\beta_0 = -\frac{\phi_2}{\phi_3} \text{ and } \beta_1 = \frac{1}{\phi_3}, \text{ or } \phi_2 = -\frac{\beta_0}{\beta_1} \text{ and } \phi_3 = \frac{1}{\beta_1}.$$

2.4 Methods of Estimation

Due to the difficulty of estimating constrained parameters we consider several methods of estimation for Models (2.8) and (2.9). We evaluate these in terms of bias, coverage, ability to enforce parameter constraints, and ability to accommodate additional data artifacts such as survey weights or random effects.

2.4.1 Nonlinear least squares

To obtain parameter estimates in Models (2.8) or (2.9), one approach is to use nonlinear least squares, i.e., minimizing $\sum_i \left(Y_i - \frac{\phi_1}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^2$ or $\sum_i \left(Y_i - \frac{\frac{1}{1+\exp(-\lambda)}}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^2$. Even though Y_i is binary in our models and is not normally distributed, the nonlinear least-squares estimates are consistent as long as Models (1) and (2) are correctly specified. Estimation by nonlinear least squares is attractive as it can be performed by either the `nls` or `nlme` functions in R, which can incorporate survey weights or random effects.

2.4.2 Maximum Likelihood Estimation

Maximum likelihood estimation allows for appropriate treatment of the binary outcome. Here, we consider $Y_i \sim \text{Bern}(p_i)$, where $p_i = \frac{\phi_1}{1+\exp\left(\frac{-(x_i-\phi_2)}{\phi_3}\right)}$ under Model (2.8) or $p_i = \frac{\frac{1}{1+\exp(-\lambda)}}{1+\exp\left(\frac{-(x_i-\phi_2)}{\phi_3}\right)}$ under Model (2.9). Letting $\theta = (\phi_1, \phi_2, \phi_3)$, the log likelihood function for Model (1) is:

$$\begin{aligned}
\log \mathrm{L}(\theta) &= \sum_{i=1}^{n} \log \left[p_i^{y_i} (1-p_i)^{1-y_i} \right] \\
&= \sum_{i=1}^{n} \log \left[\left(\frac{\phi_1}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^{y_i} \left(1 - \frac{\phi_1}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^{1-y_i} \right] \\
&= \sum_{i=1}^{n} y_i \log \phi_1 - \log \left\{ 1 + \exp\left(-\frac{(x_i-\phi_2)}{\phi_3} \right) \right\} \\
&\quad + (1-y_i) \log \left\{ 1 + \exp\left(-\frac{(x_i-\phi_2)}{\phi_3} \right) - \phi_1 \right\}. \qquad (2.11)
\end{aligned}$$

Appendix A contains details of the likelihood for Model and (2) and the asymptotic distribution of parameter estimates for both Models (2.8) and (2.9). The `optim` function in R can be used to obtain maximum likelihood estimates by different estimation algorithms. We explore the default algorithm, Nelder-Mead, as well as L-BFGS-B, a version of the Broyden-Fletcher-Goldfarb-Shanno algorithm which allows box constraints to restrict parameter estimates [24, 25]. Hence, for Model (2.8) the estimate of ϕ_1 can be analytically constrained in $(0, 1)$. Both algorithms require user input of the negative log likelihood as well as starting values for the estimation routine. Both also obtain standard error estimates of parameters by the square root of the diagonal of the inverse of the Hessian matrix. For the L-BFGS-B algorithm, users must specify a lower bound and upper bound for the box constraints for all estimated parameters.

2.4.3 Bayesian Estimation

We also consider a Bayesian framework for estimating parameters in Models (2.8) and (2.9) based on the likelihoods in Equation (2.11) and in Appendix A [26]. Inference is obtained by sampling from the joint posterior distribution of the parameters using Markov Chain Monte Carlo implemented in WinBUGS 1.4. We define parameter estimates as the posterior median and credible sets from associated 2.5 and 97.5 percentiles. Estimates and credible sets for ϕ_1 from Model (2.9) can be obtained directly from MCMC samples by transformation of the sampled λ's, with $\phi_1 = \frac{1}{1+\exp(-\lambda)}$.

Care must be taken in choosing appropriate prior distributions for the parameters. In general, the priors should conform to the the plausible range of values which the parameter may take. Similarly, prior distributions of parameters may be specifically chosen to impose constraints on parameter estimates and credible sets. For Model (2.8) it is clear that $0 < \phi_1 < 1$, and an example of an appropriate uninformative prior is $\phi_1 \sim \text{Unif}(0, 1)$. For Model (2.9), λ can reasonably take on values in $(-\infty, +\infty)$; however, an uninformative prior for λ does not necessarily correlate to an uninformative distribution for the parameter of interest ϕ_1. For example, a uniform distribution for λ implies a heavy-tailed U-shaped distribution for ϕ_1, whereas a standard normal logistic distribution for λ corresponds to a uniform distribution for ϕ_1. Lastly, for either Model (2.8) or (2.9) the plausible values for ϕ_2 and ϕ_3 will vary depending on the application. Nevertheless, the prior distributions should still reflect that these are strictly positive quantities in our application of vaccination studies.

2.5 Simulation

To assess and compare performance of the different models and estimation approaches, we perform a simulation study.

2.5.1 Details

We set the true values of the parameters as $\phi_1 = 0.70$, $\phi_2 = 5.0$, and $\phi_3 = 1.5$; for Model (2.9), this yields $\lambda = 0.85$. We generate 500 simulations of sample size 350 where age is $\mathbf{X} \sim$ Unif(0.1,15). For each of the nonlinear least squares, Nelder-Mead, and L-BFGS-S algortihms the starting values for the algorithm are the true parameter values. The results from the L-BFGS-S algorithm are only presented for Model (2.8) since Model (2.9) results are nearly identical to those from the Nelder-Mead algorithm. The box constraints imposed on the L-BFGS-S algorithm are $0.01 \leq \phi_1 \leq 0.99$, $0.10 \leq \phi_2 \leq 100$, and $0.10 \leq \phi_3 \leq 100$.

For Bayesian estimation, we implement MCMC for each simulation using 3 chains, each with different starting values, for 5,000 iterations. The first 1,000 iterations were discarded for burn-in. For Model (2.8) the prior distributions for the parameters were $\phi_1 \sim \text{Unif}(0.01, 0.99)$, $\phi_2 \sim \text{Unif}(0.1, 20)$, and $\phi_3 \sim \text{Unif}(0.1, 7)$. For Model (2.9), the prior distribution for λ is standard logistic, and priors for ϕ_2 and ϕ_3 are the same as for Model (2.8). Convergence of each iteration for the Bayesian method was verified by the Gelman and Rubin statistic $\hat{R}$, which compares the variance of the between- and within-variances of each chain [26]. Each scalar estimator is said to have converged

	Bias				Coverage				Mean Length			
	λ	ϕ_1	ϕ_2	ϕ_3	λ	ϕ_1	ϕ_2	ϕ_3	λ	ϕ_1	ϕ_2	ϕ_3
Model (1)												
NLS	-	0.003	0.059	-0.005	-	94.4	92.8	94.0	-	0.17	1.89	1.63
NELDER-MEAD	-	0.002	0.032	0.006	-	94.8	94.4	93.2	-	0.18	2.04	1.35
L-BFGS-S	-	0.002	0.032	0.006	-	94.8	94.4	93.0	-	0.18	2.04	1.35
BAYES	-	0.013	0.139	0.175	-	95.4	95.8	91.8	-	0.20	2.39	1.76
Model (2)												
NLS	0.030	0.003	0.059	-0.005	93.4	93.4	92.8	94.0	0.98	0.17	1.89	1.63
NELDER-MEAD	0.022	0.002	0.032	0.006	94.4	94.4	94.4	93.0	0.89	0.18	2.04	1.35
BAYES	0.074	0.013	0.140	0.175	95.0	95.0	95.6	91.4	1.21	0.20	2.40	1.76

Table 2.1: Summary of simulation results displaying the bias of the parameter estimates, the coverage of the confidence intervals/credible sets, and mean length of the confidence intervals/credible sets.

if $\hat{R} \approx 1$.

2.5.2 Results

Results from the 500 simulations of two different models with three different methods of estimation are presented in Table 2.1 and Appendix B (Figures B.1 - B.4). Using Model (2.8), no point estimates of $\phi_1 > 1$ are obtained under this simulation design; as a consequence, Nelder-Mead and L-BFGS-S results are nearly identical. The Bayesian approach exhibits the greatest bias for all parameters, but also provides slightly better coverage for ϕ_1 and ϕ_2, though not for ϕ_3. The Bayesian method has slightly longer mean credible set length compared the confidence interval length of the other methods. However, NLS, Nelder-Mead, and L-BFGS-S each generate confidence intervals whose upper bounds exceed 1.

Model (2.9) results are very similar to Model (2.8) results. However, with the re-parameterization of the numerator in (2.9) confidence intervals based on NLS or MLE no longer fall outside (0, 1). While bias and mean confidence interval length are nearly identical between Model (2.8) and Model (2.9), the coverage of Model (2.9) is marginally worse than Model (2.8).

The simulations reveal rare yet plausible data patterns which result in NLS and MLE estimates far removed from the true values and extremely wide confidence inter-

	$\hat{\lambda}$		$\hat{\phi}_1$		$\hat{\phi}_2$		$\hat{\phi}_3$	
Model (1)								
NLS	-	-	0.99	(0.56, 1.41)	8.89	(5.16, 12.62)	3.45	(1.54, 5.36)
NELDER-MEAD	-	-	0.93	(0.61, 1.24)	8.39	(5.41, 11.36)	3.04	(1.59, 4 4.49)
L-BFGS-S	-	-	0.93	(0.61, 1.24)	8.39	(5.41, 11.36)	3.04	(1.59, 4.49)
BAYES	-	-	0.89	(0.70, 0.98)	7.99	(6.07, 9.29)	2.94	(1.98, 4.07)
Model (2)								
NLS	4.29	(-27.31, 35.89)	0.99	(0.00, 1.00)	8.89	(5.16, 12.62)	3.45	(1.54, 5.36)
NELDER-MEAD	2.54	(-2.10, 7.19)	0.93	(0.11, 1.00)	8.39	(5.42, 11.35)	3.04	(1.60, 4.48)
BAYES	2.05	(0.84, 4.78)	0.89	(0.70, 0.99)	7.98	(6.14, 9.50)	2.96	(2.00, 4.08)

Table 2.2: Summary of erratic simulation: point estimates and 95% confidence intervals/credible sets.

vals. Such situations appear to be due to the pattern of reported outcomes observed in older ages. While all realizations are in accord with the underlying model, at times too few outcomes occur at given age ranges to allow reliable estimation of the upper asymptote corresponding to vaccination coverage (our parameter of primary interest). The Bayesian approach exhibits considerably more stability and is less sensitive to these situations. To see this in more detail, Table 2.2 provides results from such a simulation. In this case, Model (2.8) yields estimates of ϕ_1 much greater than the true value of 0.7, and confidence intervals for NLS, Nelder-Mead, and L-BFGS-S exceed one. The Nelder-Mead and L-BFGS-S algorithms produce the same parameter estimates and confidence intervals because the parameter estimates were well within the specified box constraints of the L-BFGS-S algorithm. In Model (2.9), again estimates of ϕ_1 are much greater than the true value of 0.7, and the confidence intervals for NLS and Nelder-Mead are quite large. This implies that the reparameterization of Model (2.8) could affect the stability of parameter estimates in Model (2.9). On the other hand, Bayesian estimates of all parameters for both Models (2.8) and (2.9) are closer to the true parameter value with tighter credible sets compared to point estimates and confidence intervals from nonlinear least squares and maximum likelihood estimation.

		DPT1		DPT2		DPT3	
Entry	Value	N	(%)	N	(%)	N	(%)
No	0	881	(16.2)	1323	(24.4)	1930	(35.6)
Vacc. date on card	1	2580	(47.5)	2396	(44.1)	2157	(39.7)
Vacc. marked on card	1	20	(0.4)	20	(0.4)	18	(0.3)
Reported by mother	1	1949	(35.9)	1689	(31.1)	1323	(24.4)

Table 2.3: Summary of 2003 Kenya DHS DPT outcomes.

2.6 Application to Kenya 2003 DHS

Both Models (2.8) and (2.9) are applied to the 2003 Kenya DHS data set. The DHS asks mothers for vaccination information for all children under the age of 5. Our main outcomes of interest are whether or not a child was vaccinated with the diphtheria, pertussis, and tetanus vaccine series denoted by DPT1, DPT2, and DPT3, respectively, which are recommended at 6, 10, and 14 weeks [6]. Vaccination can be verified by the vaccination card or by maternal recall. The 2003 Kenya DHS reports that 60% of children had vaccination cards available [7]. Table 2.3 displays the unweighted sample frequencies of the response to the questionnaire.

All children who have a vaccination date on card, marked on card without a date, or reported by mother are considered to have been vaccinated by the date of the interview. Children for whom “No” is reported are considered unvaccinated by the date of the interview. For the independent variable we use the age of the child in months at the time of the interview as all dates recorded in the DHS data set are recorded in Century Month Code (CMC). It should be noted that this data set does not provide mother or household ID’s, therefore, we are possibly ignoring correlated outcomes in children raised by the same mother or in the same household. Lastly, the DHS data are survey data based on a cluster design with associated sampling weights. For this analysis, we ignore the sampling weights and use this data set as an illustration of the methods described; therefore, the vaccination coverage estimates presented may not be representative of true population coverage estimates.

We analyze coverage of DPT1, DPT2, and DPT3 from the 2003 Kenya DHS data set by nonlinear least squares, maximum likelihood estimation, and Bayesian estimation for both Models (2.8) and (2.9) (Table 3). With the exception of nonlinear least squares in estimating ϕ_3, all produce very similar point estimates as well as similar confidence intervals/credible sets. Confidence intervals are slightly wider for DPT3 parameter estimates compared to DPT1 and DPT2, which reflects greater uncertainty associated with DPT3 due to fewer individuals being vaccinated compared to the other two vaccinations. Parameter estimates show highest coverage for DPT1, then DPT2, and lastly DPT3. The median vaccination age varies, as expected, but are all slightly higher than the WHO targeted vaccination ages of 1.5, 2.5, and 3.5 months. The time elapsed between which 50% and 75% of the vaccination coverage has been achieved is greatest for DPT3.

2.7 Discussion

Due to difficulties in estimating constrained parameters in logistic growth models, we compared two model formulations with three methods of estimation. In general, Model (2.9) is preferred to Model (2.8) since the reparameterization ensures that estimates of ϕ_1 are within its logical constraints. Nonlinear least squares provides estimates based on an inappropriate error model and may not be robust to all situations, and yields parameter estimates in the Kenya data application which differ slightly from both maximum likelihood and Bayesian estimates. In our simulations, all methods exhibited similar parameter coverage rates, with the exception of ϕ_3 for which the Bayesian setting yielded slightly lower coverage.

Although both maximum likelihood and Bayesian estimation appropriately account for the binary nature of the outcome, MLE inference is based on the asymptotic distribution of parameter estimates whereas Bayesian inference is based on samples

	$\hat{\lambda}$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
Model 1, DPT1				
NLS	-	0.8723 (0.8629,0.8817)	1.7666 (1.6356,1.8975)	0.5445 (0.4236,0.6654)
NELDER-MEAD	-	0.8725 (0.8632,0.8819)	1.8109 (1.6320,1.9897)	0.5337 (0.3842,0.6832)
L-BFGS-S	-	0.8725 (0.8632,0.8819)	1.8111 (1.6323,1.9900)	0.5338 (0.3843,0.6832)
BAYES	-	0.8730 (0.8633,0.8820)	1.8125 (1.6389,2.0040)	0.5518 (0.4132,0.7304)
Model 2, DPT1				
NLS	1.9212 (1.8370,2.0054)	0.8723 (0.8626,0.8814)	1.7666 (1.6356,1.8975)	0.5445 (0.4236,0.6654)
NELDER-MEAD	1.9233 (1.8393,2.0072)	0.8725 (0.8629,0.8816)	1.8114 (1.6325,1.9904)	0.5339 (0.3844,0.6835)
BAYES	1.9270 (1.8380,2.0055)	0.8729 (0.8627,0.8814)	1.8150 (1.6425,2.0030)	0.5531 (0.4159,0.7330)
Model 1, DPT2				
NLS	-	0.8068 (0.7958,0.8178)	2.8911 (2.7152,3.0670)	0.6086 (0.4497,0.7675)
NELDER-MEAD	-	0.8060 (0.7949,0.8172)	2.8971 (2.6913,3.1028)	0.5019 (0.3543,0.6494)
L-BFGS-S	-	0.8060 (0.7949,0.8172)	2.8974 (2.6916,3.1031)	0.5016 (0.3542,0.6491)
BAYES	-	0.8060 (0.7947,0.8172)	2.9080 (2.6969,3.1416)	0.5200 (0.3877,0.6947)
Model 2, DPT2				
NLS	1.4292 (1.3587,1.4998)	0.8068 (0.7955,0.8175)	2.8911 (2.7152,3.0670)	0.6086 (0.4497,0.7675)
NELDER-MEAD	1.4244 (1.3531,1.4958)	0.8060 (0.7946,0.8169)	2.8980 (2.6921,3.1038)	0.5019 (0.3543,0.6496)
BAYES	1.4260 (1.3560,1.5010)	0.8063 (0.7951,0.8178)	2.9125 (2.7125,3.1360)	0.5206 (0.3916,0.7125)
Model 1, DPT3				
NLS	-	0.7089 (0.6961,0.7218)	4.3102 (4.0079,4.6125)	1.0150 (0.7544,1.2755)
NELDER-MEAD	-	0.7072 (0.6939,0.7205)	4.3106 (3.9952,4.6260)	0.8017 (0.5880,1.0154)
L-BFGS-S	-	0.7072 (0.6939,0.7205)	4.3115 (3.9960,4.6270)	0.8018 (0.5881,1.0155)
BAYES	-	0.7077 (0.6942,0.7198)	4.3250 (4.0289,4.6350)	0.8218 (0.6291,1.0640)
Model 2, DPT3				
NLS	0.8901 (0.8278,0.9525)	0.7089 (0.6959,0.7216)	4.3102 (4.0079,4.6125)	1.0150 (0.7544,1.2755)
NELDER-MEAD	0.8819 (0.8177,0.9461)	0.7072 (0.6938,0.7203)	4.3120 (3.9965,4.6275)	0.8016 (0.5880,1.0153)
BAYES	0.8839 (0.8191,0.9458)	0.7076 (0.6940,0.7203)	4.3230 (4.0150,4.6485)	0.8189 (0.6346,1.0545)

Table 2.4: Point estimates and 95% confidence intervals/credible sets for DPT1, DPT2, and DPT3 coverage from the 2003 Kenya DHS based on Models (2.8) and (2.9) with four methods of estimation: (1) nonlinear least squares, (2) Nelder-Mead, (3) L-BFGS-S, (4) Bayesian estimates.

from the posterior distribution of the parameters. In general, Bayesian estimates exhibited greater bias and longer credible set length compared to confidence interval length. For the main parameter of interest, ϕ_1, Bayesian estimation maintained good coverage across simulations, and though the bias was noticeably greater than the other methods, it is still virtually negligible on the scale of which the parameter is being estimated (1 percentage point). The tradeoff between bias and precision is a common consideration in statistical estimation, and especially in terms of Bayesian inference [27]. We view the Bayesian framework as attractive in its ability to naturally restrict parameter estimates through prior distributions, its inference which does not depend on asymptotic rates of convergence, and its stability in the infrequent but not entirely rare data settings yielding unstable MLEs observed in our simulation despite the greater bias and longer credible set length.

Each of the methods of parameter estimation considered above encountered simulated data sets which resulted in numerical convergence issues. For Model (2.8), nonlinear least squares and maximum likelihood optimization by the Nelder-Mead algorithm do not inherently constrain ϕ_1 such that $0 < \phi_1 < 1$, and numerical computational errors in these estimation routines are encountered frequently resulting in wide confidence intervals for these data sets. While the BFGS box-constrained algorithm can appropriately constrain parameter estimates, numerical errors during estimation are still encountered; furthermore, the estimation algorithm can remain on the edge of the user specified bounds, resulting in multiple simulations with similar and unrealistic parameter estimates on the edge of the bounds. Lastly, WinBUGS occasionally encountered numerical instability in the MCMC algorithm. All estimation methods discussed require certain user based input such as starting points for estimation routines, box constraint boundaries on parameters, and prior distributions for parameters. Changing such inputs could occasionally overcome errors in estimation routines and facilitate convergence, although such adjustments are difficult to automate to guarantee convergence.

Both models explored have the flexibility to accommodate extensions of interest. Certain covariates, such as maternal education or rural versus urban, are thought to directly impact vaccination coverage. To accommodate these effects, the analysis could be performed stratified on such covariates, or they could be taken into account directly by introducing new parameters in the three parameter logistic growth model. Covariates can be easily included in the model to affect the asymptote, inflection point, or slope.

While the DHS data represent survey data based on a cluster design with weights, we estimate the unweighted sample vaccination coverage rather than the weighted population representative coverage. Of the methods of estimation explored, only nonlinear least squares can incorporate the survey weights in a straightforward manner within the analysis. In order to appropriately take sampling into account with maximum likelihood estimation, the negative log likelihood needs to be optimized adjusting for the survey weights, which falls outside the scope of many general purpose optimization routines (e.g., `optim` function in R). To account for the survey design in the maximum likelihood or Bayesian framework, future work will explore both model-based and design-based approaches. In a model-based approach, one could take into account the clusters of individuals by estimating a random effect for each cluster. In a design-based approach, one could utilize resampling techniques. This would provide more accurate estimates of the population vaccination coverage from data obtained by a complex sampling design, as opposed to a simple random sample.

We defined the outcome of the model as whether or not a child was vaccinated as indicated by either vaccination card or maternal recall. As maternal recall is imperfect, it is possible that outcomes indicated by maternal recall could be misclassified. Various studies have shown mixed results on the validity of maternal recall; while some show that maternal recall tends to be in agreement with vaccination history

[11, 28, 29, 30, 31], others suggest substantial discordance [32, 33, 34]. These results are country and vaccination specific, and therefore efforts should be made to verify the validity of maternal recall before utilizing the proposed modeling method. Future research should assess the effect of misclassification of the outcome on parameter estimates.

The results demonstrate that the nonlinear logistic model can be used to estimate a logistic asymptote less than 1.0 when the outcome of interest is binary. Specifically, we used this model to estimate vaccination coverage, an application where the primary interest is this unknown asymptote. Moreover, this model also estimates two other meaningful parameters in this setting, which represent the median vaccination age among those vaccinated and characterize the rate of vaccination uptake. This model is most applicable to vaccination research in which respondents are unable to estimate age at the time of vaccination, and utilizes data more generally available than that required for standard approaches based on survival analysis methods. Some additional analytic complications remain including incorporation of survey weights, adjustment for mismeasured covariates, and further reducing sensitivity to particular data patterns. Even with these challenges, the logistic model enables researchers to base inference regarding vaccination coverage on all respondents regardless of whether or not they retained their vaccination cards, hereby eliminating possible bias due to only analyzing complete data cases.

Chapter 3

Linking Remotely Sensed Data to Local Observations

3.1 Introduction

Buruli ulcer (BU) is a potentially debilitating skin disease caused by infection with *Mycobacterium ulcerans*. The disease often begins as a painless nodule and if left untreated can ulcerate, resulting in permanent scarring and disability [35]. Details of clinical symptoms, diagnosis, and treatment are reviewed elsewhere [35, 36, 37]. BU cases have been reported in at least 31 countries spread across Africa, Asia, Australia, and Latin America, demonstrating increasing prevalence and expanding geographic distribution during the past century [37]. Endemic in areas of sub-Saharan Africa, countries in West Africa including Benin, Côte d'Ivoire, and Ghana have the highest burden of the disease [38]. Accurate surveillance data reflecting the true disease incidence in West Africa remains elusive due to local variations in aspects such as case confirmation, access to care, diagnosis, and reporting practices.

While it is known that *M. ulcerans* causes BU, the exact mode of transmission is still unknown. Although many vectors and reservoirs of the disease have been hypothesized, none have been conclusively identified [39, 40, 41, 42, 43, 44, 45, 37, 46, 47]. An extensive review of the ecology and transmission of BU is provided by Merritt et al. (2010) [47]. Epidemiological studies have found BU incidence to be associated with water exposure through swimming, domestic water-related activities, and proximity to water [48, 49, 39, 50, 35, 51, 37, 52, 46]. Such studies generally support transmission via environmental contact through an open skin lesion or scratching such a lesion [37, 47]. Furthermore, many studies suggest that BU is associated with disturbed environments, such as deforested areas and farmlands [37, 53, 54, 55]. An environmental pathogen with a distribution in nature thought to be greater than that of the disease, *M. ulcerans*' existence appears independent of human interaction [56]. Studies have verified the presence of *M. ulcerans* DNA in many areas of aquatic systems including suspended material in water, detritus, biofilm, and aquatic insects [57, 41, 58, 45, 46, 59, 56]. *M. ulcerans* DNA has also been found among many aquatic invertebrates collected from 27 aquatic habitats of both endemic and non-endemic communities of Ghana [46, 56].

A small number of studies undertaken in Ghana, Côte d'Ivoire, and Benin have examined the geographical patterns of BU disease endemic areas taking into account landscape and environmental factors [53, 60, 55, 54]. Positive associations with BU incidence were found with mean arsenic content of soil, proximity to gold mining sites, irrigated rice crops area, agriculture, forest, and wetness index variability. Negative associations with BU were found with dam surface area, urban land cover, and mean elevation. One study also reported geographic clusters of communities with higher than expected and lower than expected disease prevalence [54].

Although these studies of BU incidence and prevalence have provided important information for understanding geographic and environmental associations with human disease, similar studies evaluating the factors driving pathogen distribution in the environment have not been conducted. This study sought to investigate the spatial distribution of *M. ulcerans* among aquatic habitats in Ghana and identify environmental characteristics associated with the presence of *M. ulcerans*. We hypothesized that the presence of *M. ulcerans* was associated with both broad scale environmental features as well as highly localized characteristics of aquatic systems. At the most localized level, we measured several physical and chemical properties of the aquatic systems themselves. For broad scale environmental features, we used remotely sensed observations to infer landscape and land use/land cover properties among the same aquatic systems. Finally, as an initial test of local similarity between the geographic distribution of the pathogen and disease, we assessed the association of *M. ulcerans* presence and the Buruli ulcer disease reporting history of the local community.

3.2 Methods

3.2.1 Study Area

A total of 98 aquatic habitats were sampled from water bodies routinely used by communities in five regions of Ghana: Greater Accra (N = 24), Eastern (N = 5), Ashanti (N = 34), Central (N = 5), and Volta (N = 30) (Figure 3.1). Due to geographic proximity, the sites sampled in the Eastern and Central regions of Ghana are henceforth classified with the Greater Accra and Ashanti sites, respectively. Aquatic environmental sampling was performed in Greater Accra and Ashanti in the summers of 2005-2007 with 9-11 sites sampled each year, with the exception of Ashanti in 2006 during which 20 sites were sampled; sampling took place on a single date in each village. All community sites were randomly selected within region, and the water bodies were located within or near (<200m) the villages. Sampled water bodies were selected based on discussions with community leaders with regards to daily and frequent domestic water use. Different aquatic habitats were sampled to include streams, rivers, wetlands, ponds, and reservoirs, and were classified as lentic (still) or lotic (flowing).

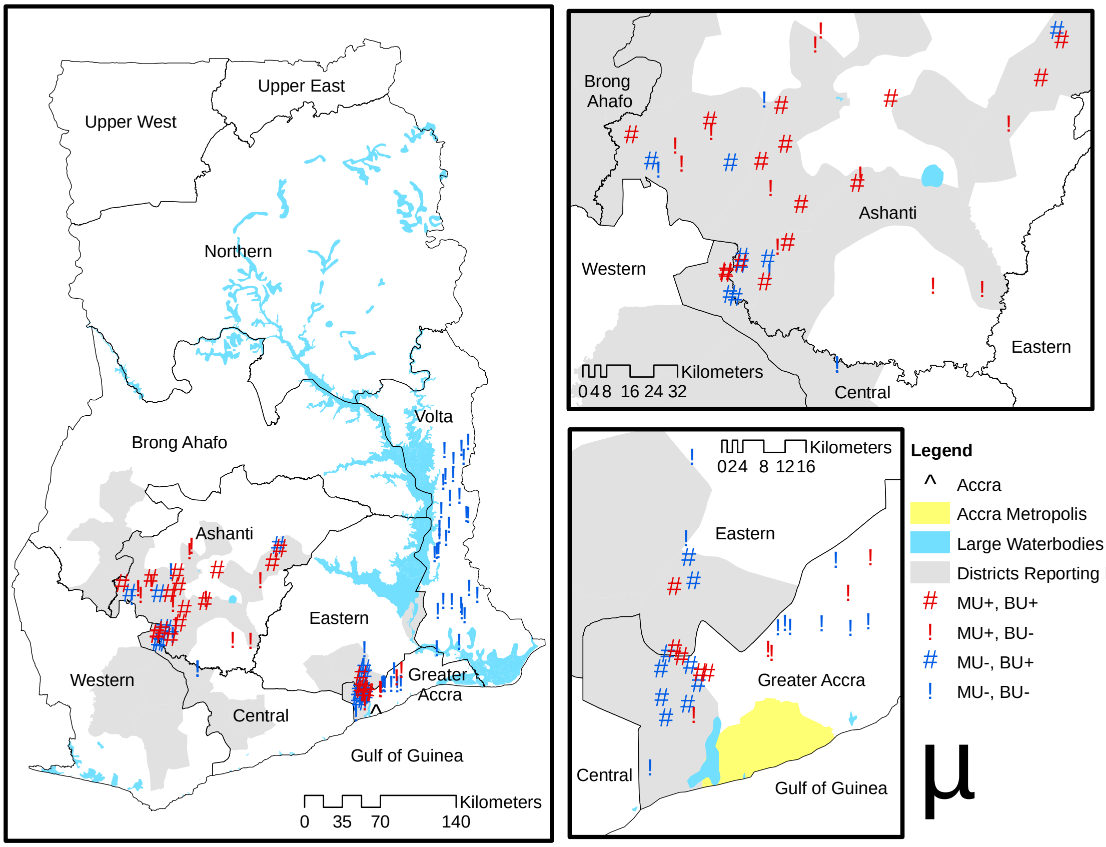


Figure 3.1: Locations of sampled aquatic habitats in the Ashanti, Greater Accra, and Volta regions of Ghana. Administrative districts that reported cases from 2004-07 are shaded in gray. Red symbols indicate *M. ulcerans* positive sites, whereas blue symbols indicate *M. ulcerans* negative sites. Triangles represent sites that reported cases 2004-07, and circles represent sites with no reported cases 2004-07.

The BU surveillance data of the sampled communities were obtained from the Ghana Health Services National Buruli ulcer Control Programme. As validation surveys have found BU cases in communities where no case reporting occurred [47], it is possible (indeed probable) that at least some non-reporting communities have cases. Therefore, we consider disease reporting history at two levels. Community level reporting is defined as the presence of reported BU cases in the sampled communities from 2004-2007. In contrast, district level reporting is defined based on whether the sampled community is located within a district that reported BU cases in the same years.

3.2.2 Detection of *M. ulcerans*

Sample collection and subsequent laboratory processing for detection of *M. ulcerans* followed a strict protocol. Macrophyte (i.e., aquatic plants) biofilm and suspended material in the water were collected via standardized environmental sampling to identify the presence of *M. ulcerans* in the aquatic habitats. At each water body, one detrital biofilm sample as well as macrophyte biofilm samples from the two most dominant living macrophyte type were collected (N=3). Macrophyte biofilm samples consisted of 1-5 specimens depending on the plant type. The submerged portion of each plant sample was placed into a Ziploc bag with 100 ml of pre-filtered, bottled water. Once sealed, plants were rubbed within the bag to dislodge and suspend any epiphytic material. Approximately 10 ml of the resulting liquid suspension and a small portion of plant material were preserved with 100% ethanol for polymerase chain reaction (PCR) analysis. To assess suspended material in the open water column for *M. ulcerans*, a composite water sample (∼12 L) was collected from random open water areas within the water body at the mid-water column depth. From the composite, five 100-200 ml sub-samples were filtered through a 1.6 micron fiberglass filter followed by a 0.2 micron nitrocellulose filter (Whatman Inc). Filters were sealed in aluminum foil packets for subsequent laboratory analysis.

All samples were processed at the University of Tennessee, Knoxville, Tennessee. *Mycobacterium ulcerans* DNA was detected by a tiered PCR based detection method in which DNA was subjected first to amplification of the enoyl reductase (ER) domain. ER-PCR positive samples were further evaluated by variable number tandem repeat (VNTR) analysis to differentiate *M. ulcerans* from other mycolactone producing mycobacteria. Detailed methods of DNA detection are discussed by Williamson et al. (2008) [56]. If any of the environmental samples contained *M. ulcerans* DNA then the corresponding aquatic site was identified as *M. ulcerans* positive.

3.2.3 Water Characteristics

A variety of physical and chemical properties were evaluated for each aquatic habitat. One-liter water samples were collected to evaluate the water's physicochemical characteristics (Table 3.1). Several parameters (e.g., dissolved oxygen, temperature, conductivity) were measured *in situ* using a YSI 6600 Data Sonde (Yellow Springs Instruments, Inc., OH). Water samples were stored on ice and then frozen until analysis at the Environmental Chemistry Division of the Water Research Institute, Ghana using established and standard water quality methods [56].

3.2.4 Remotely Sensed Covariates

Land use/land cover (LULC) was inferred from dry season 2000 and 2002 Landsat ETM+ imagery with 30 m resolution (Figure 3.2), obtained from the University of Maryland Global Land Cover Facility [61]. Details of the geoprocessing techniques appear in Wagner et al. 2008 [54]. Easily distinguished LULC categories including agriculture, forest, shrubland, urban, water, and wetlands were summarized in 0.1, 0.5, 1.0, and 5 km circular buffers around the water bodies. LULC covariates included the percent of pixels characterized by an LULC type within the specified buffer distance, as well as by a presence/absence indicator corresponding to specific LULC types within the buffer.

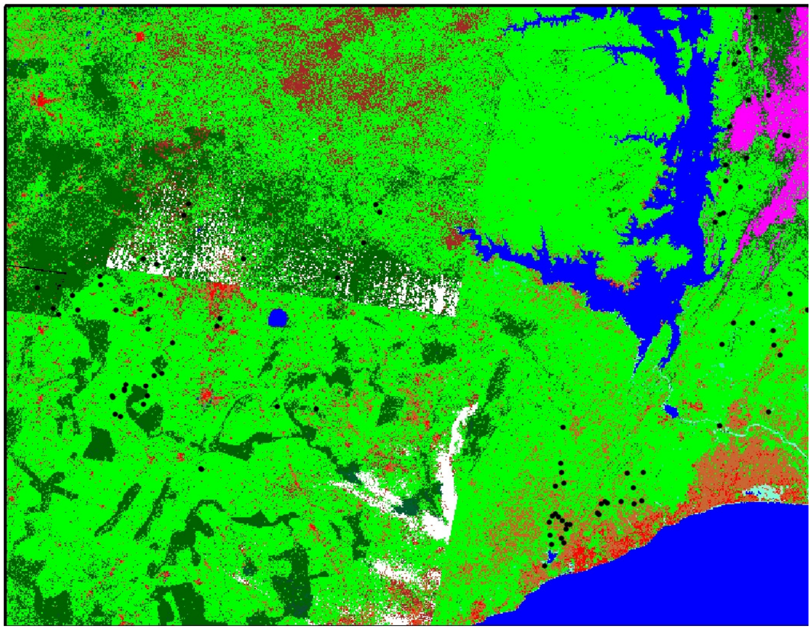


Figure 3.2: Land use/land cover derived from LandsatETM+ imagery. Land use/land cover was summarized in 0.1, 0.5, 1.0, and 5 km circular buffers around the sampled sites, represented by black dots.

A digital elevation model (DEM) was derived from NASA Shuttle Radar Topographic Mission (SRTM) (2000) data with a 3 arc second (90m) resolution at a WRS-2 unfilled finished A processing level (Figure 3.3), obtained from the University of Maryland Global Land cover Facility. The DEM gaps were filled and compound topographic index, or wetness index, was calculated using the following equation [62]:

$$\text{WI} = \frac{\ln((FA + 1.0) * 90\text{m})}{\text{slope} + 0.0001}. \tag{3.1}$$

The minimum, maximum, mean, and standard deviation of the wetness index for buffer diameter sizes of 0.5, 1.0, and 5 km were also calculated as an approximate measure of potential land surface moisture content and its spatial variability. Lastly, site-specific values of elevation were extracted. All environmental covariates were calculated and extracted in ArcGIS 9.3.1 (ESRI Inc., Redlands, CA).

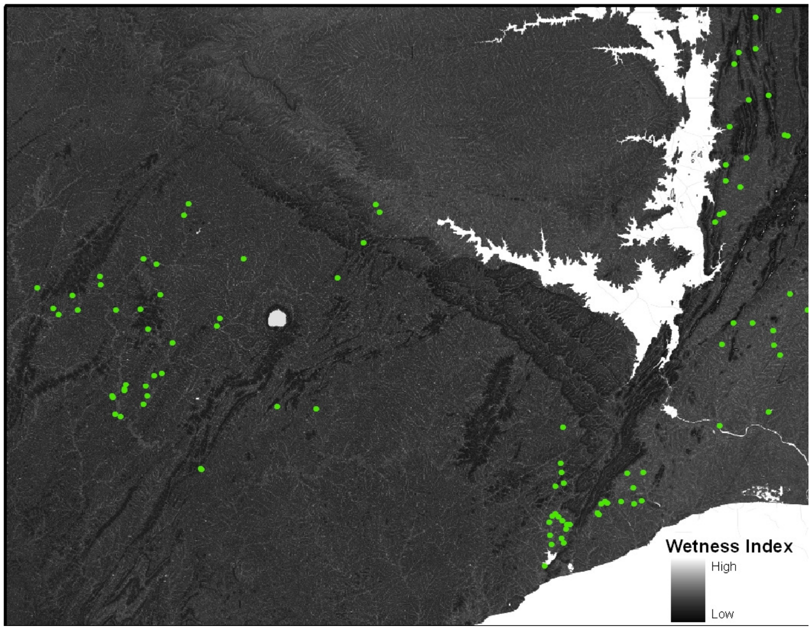


Figure 3.3: Wetness index provided by the digital elevation model derived from NASA SRTM data. Green dots represent sampled sites.

3.3 Statistical Analysis

Ripley's K function [63] was used to test for the presence and scale of any patterns of spatial clustering of *M. ulcerans* positive sites relative to negative sites using the `splancs` package in `R` [64]. Geographic regions were assessed separately for clustering patterns, and the distances at which clustering patterns were evaluated were approximately one-third of the distances separating the two furthest sites in each region. The square-root transformation of the K function was employed to linearize the function and stabilize the variance. The expectation of the transformed function minus the distance at which it is evaluated was 0 under the null hypothesis of complete spatial randomness. Estimated functions greater than zero implied global spatial clustering, whereas estimated functions less than 0 implied global spatial dispersion. We calculated the differences between the transformed K functions that summarized the spatial distribution of *M. ulcerans* positive and negative sites (case-control K function) and assessed significance via Monte Carlo simulation (999 simulations).

To investigate location and significance of individual clusters of *M. ulcerans* positive sites, we calculated Kulldorff's Bernoulli spatial scan statistic using circular windows for the Ashanti and Accra areas separately [65]. The statistical significance of potential clusters was evaluated through Monte Carlo hypothesis testing (999 simulations) in SaTScan 8.0 [66] at the 0.05 significance level.

Logistic regression was used to identify variables associated with the presence of *M. ulcerans* among the aquatic sites using SAS software version 9.2 (SAS Institute Inc., Cary, NC). Most of the physicochemical water characteristics were log transformed due to highly skewed distributions. Both percent LULC within buffers as well as indicators of LULC classes' presence were considered for model selection. Model selection was based on Akaike's Information Criterion (AIC), an information-theoretic approach for selecting a best model from a set of candidate models [67]. AIC was adjusted for small sample size (AICc) because of the low ratio of the sample size to the number of parameters. The model with the lowest AICc was regarded as the best fitting model of those considered. Due to the high degree of correlation among LULC variables as well as water variables, a set of candidate variables were identified for model building. Only one LULC covariate at each buffer distance which resulted in the greatest reduction of AICc was considered in model selection, and only uncorrelated water quality covariates ($\rho < 0.3$) with the greatest reduction in AICc were considered in model selection. Region interactions were considered with each candidate covariate. We focused on five best-fitting models constructed using five sets of candidate covariates, namely: (1) water, (2) LULC, (3) DEM (terrain), (4) LULC+DEM (landscape), (5) all sets of covariates (overall).

Model fit was assessed by a variety of methods for the final model constructed from all sets of covariates. Standardized residuals were used to check for the presence of outliers, and observations were evaluated to identify those with high leverage [14].

In addition, the residuals from the Greater Accra and Ashanti regions were assessed separately for spatial autocorrelation by the empirical semivariogram using the `geoR` package in `R`. The semivariogram estimated variability between distinct pairs of sites as a function of the distance h between them. Simulation envelopes (999) were constructed by Monte Carlo simulation in order to test the null hypothesis of spatial independence among the residuals: if all points fall within the envelopes then the null hypothesis of spatial independence is not rejected [68].

In addition to assessments of associations between *M. ulcerans* presence and environmental attributes, we also explored the association between reported BU cases and *M. ulcerans* presence. The unadjusted association was tested by Pearson's chi-square test at the 0.05 significance level. The association adjusted for environmental covariates was assessed by entering the BU reporting history variables individually into the best fitting logistic regression model. We required one of two conditions to include BU reporting history in our model. First, if the updated model's AICc was at least 2 units smaller than that of the best fitting model then the BU reporting history variable improved model fit and was included. Second, if the AICc of the updated model remained within 2 units from the AICc of the best fitting model then the updated model was considered competitive with the best fitting model [67].

3.4 Results

No cases of BU have been reported from the Volta region, and environmental sampling did not detect *M. ulcerans* in this region ($N = 30$). For the present study, data from the Volta region were excluded from further analyses in order to investigate factors relating to variation in *M. ulcerans* presence.

A higher proportion of aquatic habitats tested positive for *M. ulcerans* in the

Characteristic	**Greater Accra** ($N = 29$)		**Ashanti** ($N = 39$)	
General n(%)				
M. ulcerans present	10	(34%)	26	(67%)
Community level reporting	14	(48%)	23	(59%)
District level reporting	20	(69%)	37	(95%)
Lentic aquatic system (still)	22	(76%)	13	(33%)
Water Covariates Median (Min, Max)				
Calcium (mg/L)*	19.6	(2.0, 94.2)	11.6	(3.6, 24.4)
Calcium hardness as CaCO3 (mg/l)*	44.1	(0.0, 231.0)	24.0	(1.3, 56.1)
Carbon trioxide (mg/L)*	109.0	(1.2, 449.0)	41.5	(12.2, 151.0)
Chlorine (mg/L)*	28.8	(3.0, 647.0)	9.9	(3.0, 69.5)
Chlorophyll (mg/L)*	11.0	(4.5, 76.1)	8.7	(0.3, 125.7)
Color apparent (Hz)*	35.0	(1.2, 500.0)	30.0	(5.0, 180.0)
Dissolved oxygen percent saturation	35.0	(0.2, 134.8)	56.6	(0.3, 90.3)
Field specific conductivity (μS/cm)*	266.2	(67.0, 2601.4)	116.0	(39.0, 696.3)
Field temperature (Celcius)	26.5	(24.3, 32.0)	24.2	(23.2, 27.8)
Field turbidity (NTU)*	17.1	(0.1, 331.5)	34.4	(0.0, 353.7)
Iron (mg/L)*	1.2	(0.0, 7.2)	2.5	(0.2, 7.4)
Magnesium (mg/L)*	7.8	(2.0, 86.9)	5.9	(3.0, 15.1)
Manganese (mg/L)*	0.1	(0.0, 3.0)	0.0	(0.0, 0.3)
Nitrate (mg/L)*	0.4	(0.0, 2.8)	1.0	(0.0, 21.0)
Nitrogen dioxide (mg/L)*	0.0	(0.0, 0.6)	0.0	(0.0, 0.1)
Nitrogen/phosphate ratio*	7.2	(6.3, 8.7)	6.7	(5.5, 7.6)
Oxidation-reduction potential	84.8	(-184.5, 146.8)	64.5	(-188.1, 168.9)
pH	7.2	(6.3, 8.7)	6.7	(5.5, 7.6)
Phosphate (mg/L)*	0.1	(0.0, 1.0)	0.1	(0.0, 0.5)
Sulfate (mg/L)*	10.0	(0.2, 68.6)	5.2	(0.5, 29.2)
Suspended solids (mg/L)*	33.0	(8.0, 451.0)	12.0	(2.0, 64.0)
Total alkalinity as $CaCO_3$ (mg/l)*	90.0	(1.0, 368.0)	34.0	(10.0, 124.0)
Total dissolved solids (mg/L)*	138.0	(37.2, 1172.0)	62.0	(22.0, 188.0)
Total nitrogen*	0.4	(0.0, 2.8)	1.0	(0.0, 21.0)
Terrain Covariates Median (Min, Max)				
Elevation	46.0	(15.0, 152.0)	175.0	(100.0, 378.0)
Wetness at site	4.7	(3.0, 12.1)	4.9	(2.7, 10.8)
Average wetness (500m)	5.7	(4.8, 6.3)	4.8	(4.0, 5.6)
Average wetness (1km)	5.4	(4.5, 6.3)	4.7	(4.0, 5.4)
Average wetness (5km)	5.2	(4.5, 6.7)	4.5	(3.8, 5.4)
Standard deviation wetness (500m)	1.9	(1.0, 3.1)	1.9	(0.8, 2.8)
Standard deviation wetness (1km)	1.9	(1.4, 3.3)	1.8	(1.2, 2.9)
Standard deviation wetness (5km)	1.8	(1.7, 4.3)	1.8	(1.6, 1.9)

Table 3.1: Descriptive statistics of environmental attributes in regions exhibiting *M. ulcerans* presence. Starred variables were natural log transformed for statistical analysis.

Ashanti sites (66%, $N = 39$) than in the Greater Accra region (34%, $N = 29$) (Table 3.1). Physical and chemical properties of water from sites varied by region. Ashanti sites were at a higher elevation than Accra sites, which were located closer to the coast (median elevation for Accra=46 meters above sea level, Ashanti=175 m). Both Accra and Ashanti had similar at-site median wetness index, however the average wetness index within buffers around the sites was greater in Accra than Ashanti. Both regions exhibit similar average wetness index variability within buffers around the site.

The six predominant LULC types observed included: agriculture, urban, water, forest, wetland, and shrubland. The majority of sites in the Greater Accra region had agriculture and shrubland present within the buffers, whereas fewer Ashanti sites contained these land cover types (Figure 3.4). Wetlands were present within buffers around the Greater Accra sites, but not at all in Ashanti sites. Both regions exhibited similar patterns for presence of urban and forested areas. At 30 m resolution, the percentages of water and wetlands in all buffers around aquatic sites were very low for both Greater Accra and Ashanti and therefore were not considered for the model selection process.

The spatial distribution of *M. ulcerans* positive and negative aquatic sites was assessed for both clustering patterns and individual clusters of positive sites. The difference between the transformed K functions showed no significant global clustering of *M. ulcerans* positive sites relative to negative ones for either the Greater Accra or Ashanti regions (Figure 3.5). Moreover, Kulldorff's spatial scan statistic found no significant spatial clusters of *M. ulcerans* positive aquatic habitats.

Logistic regression identified factors associated with *M. ulcerans* from the distinct sets of covariates (Table 3.2). The AICc of the models ranged from 55.0 to 96.1, with the best fit achieved by combining covariates from all sets. This model with the lowest AICc contained seven main effects: (1) region, (2) elevation, (3) wetness index at the site, (4) standard deviation of wetness index within 500 m of the site,

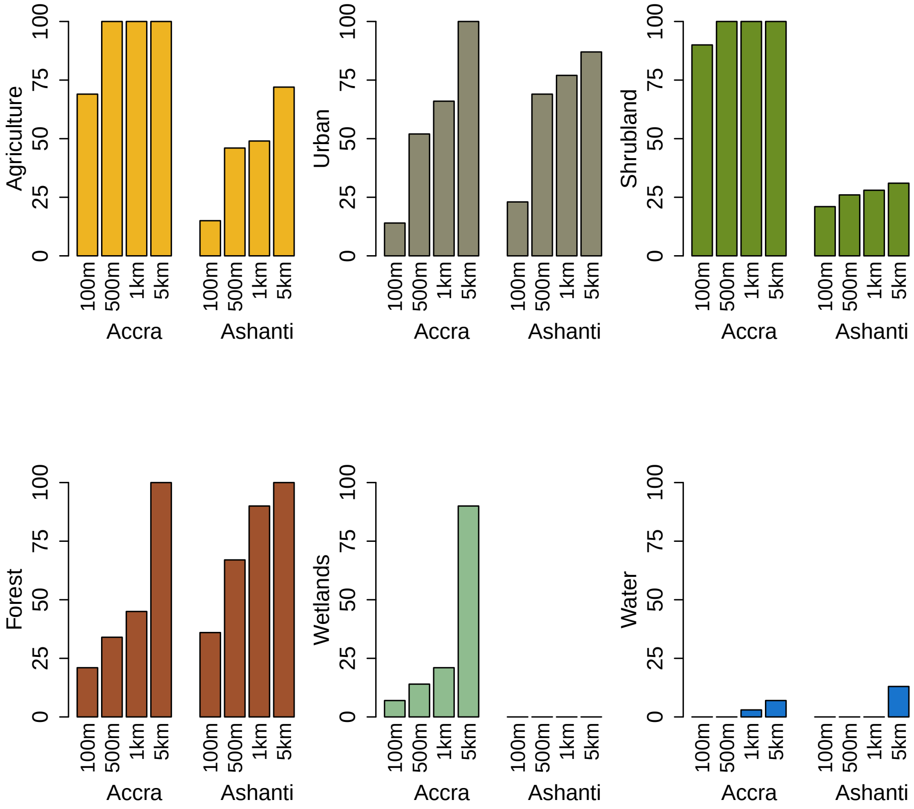
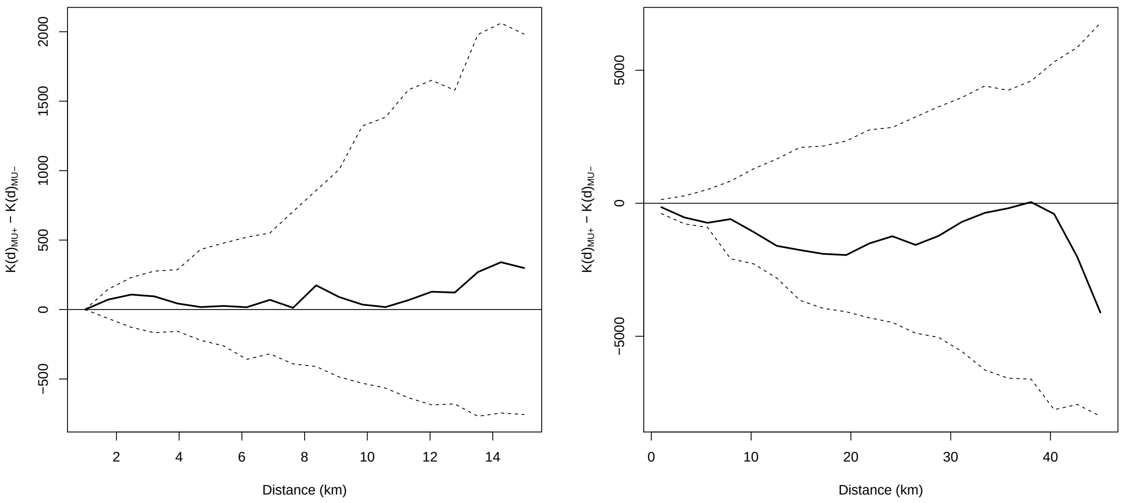


Figure 3.4: Land cover within buffers around aquatic habitats in the Greater Accra and Ashanti regions. Bars display the percent of sites with the feature present within each buffer.



(a) Greater Accra

(b) Ashanti

Figure 3.5: Ripley's case-control K-function for Greater Accra (a) and Ashanti (b). Clustering patterns for Greater Accra were assessed up to 15 km, and for Ashanti 45 km. Shown in the solid horizontal line, the expected value of this function is 0 under the assumption of complete spatial randomness. The bold line shows the observed difference between the transformed K functions of the positive and negative sites, and the dashed lines show the theoretical confidence bounds calculated by Monte Carlo simulation.

	Water	**LULC**	**Terrain**	**Landscape**	**All**
AICc (df)	86.4 (6)	84.1 (4)	72.1 (7)	58.4 (9)	55.0 (11)
Parameter estimate (SE)					
General					
Intercept	-2.30 (1.65)	2.15 (0.94)	-9.28 (3.75)	-19.04 (8.03)	-32.97 (14.72)
Accra	3.28 (2.29)	-2.37 (0.84)	15.91 (4.60)	27.19 (9.53)	44.17 (17.47)
Terrain					
Elevation			0.05 (0.02)	0.10 (0.04)	0.12 (0.05)
Accra × Elevation			-0.05 (0.02)	-0.09 (0.04)	-0.10 (0.05)
Wetness			-0.35 (0.16)	-0.52 (0.23)	-0.91 (0.41)
STD(Wetness$_{500m}$)			2.36 (1.25)	3.65 (2.13)	4.77 (3.75)
Accra×STD(Wetness$_{500m}$)			-5.13 (1.71)	-7.00 (2.76)	-8.67 (4.67)
LULC					
I(Urban$_{100m}$)		1.72 (0.88)		5.53 (2.01)	6.25 (2.47)
I(Forest$_{1km}$)		-1.86 (0.90)		-2.06 (1.44)	-3.01 (1.63)
Water					
Log(CA hardness)	1.38 (0.56)				4.33 (1.97)
Accra × Log(CA hardness)	-1.51 (0.70)				-4.41 (2.05)
DO	-0.02 (0.01)				
Log(NO_3)	0.21 (0.14)				

Table 3.2: Best fitting model results from five categories of covariates presented in columns sorted by descending AICc, with the smallest AICc indicating the best fit.

(5) indicator for urban land cover within 100 m of the site, (6) indicator for forest land cover within 1 km of the site, and (7) log of calcium water hardness. This model also contained three interaction terms with region: elevation, standard deviation of wetness index within 500m of the site, and calcium water hardness.

The best fitting model showed that the odds of *M. ulcerans* presence increased as elevation increased (within the relatively modest range of elevations considered), with a more pronounced elevation effect in Ashanti than Greater Accra. As the wetness index at the site increased, the odds of *M. ulcerans* presence decreased. As the standard deviation of the wetness index within 500 m of the site increased, the odds of *M. ulcerans* decreased in Accra but increased in Ashanti. Sites that had urban land cover present within 100 m but did not exhibit forest within 1 km had the highest odds of *M. ulcerans* presence. This is followed by (in order of highest to lowest) urban/forested, and then non-urban/non-forested, and lastly non-urban/forested areas had the lowest odds of *M. ulcerans* presence (Figure 3.6). *Mycobacterium ulcerans*

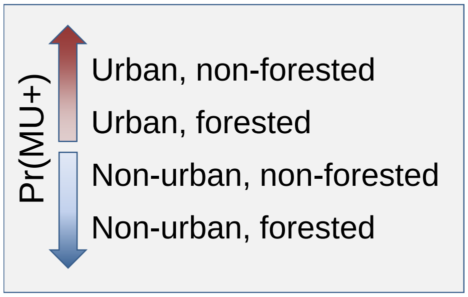


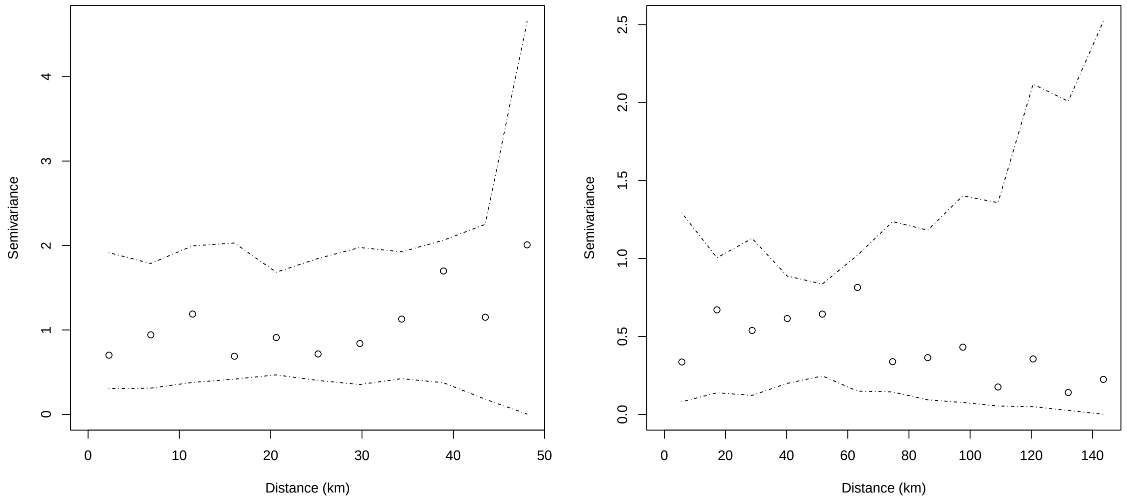
Figure 3.6: Depiction of results from best fitting model for land use / land cover. Land cover representing most disturbed areas are associated with a higher probability of *M. ulcerans* presence compared to land cover representing least disturbed areas.

presence was weakly negatively associated with water calcium hardness in Accra and strongly positively associated with calcium water hardness in Ashanti.

The best fitting land use/land cover, terrain, and landscape models reflected similar results to the model described above. The landscape model which included all remotely sensed covariates improved model fit when compared to the LULC or terrain model alone. The best fitting water model contained two additional variables that did not improve the fit of the final model: dissolved oxygen percent saturation (DO) and log of nitrate (NO_3). For both Greater Accra and Ashanti, the presence of *M. ulcerans* had a negative association with DO and a positive association with NO_3.

Diagnostics of the overall model best fitting model revealed no outliers in the standardized deviance residuals, and no observations with high leverage were identified. In addition, the empirical semivariogram of the residuals showed no evidence of significant spatial autocorrelation in either Greater Accra or Ashanti after adjusting for environmental covariates (Figure 3.7). Ashanti had lower semivariance estimates than Accra, reflecting the lower overall variance in the residuals. Three sites in Ashanti and six sites in Greater Accra indicate locations of poor model fit such that the observed outcome (*M. ulcerans* presence/absence) was not in accord with the predicted probability from the best fitting model (Figure 3.8).

Lastly, we observed variation in case reporting among the two regions. Ninety-five percent of sites in Ashanti were located within a district that reported cases,



(a) Greater Accra

(b) Ashanti

Figure 3.7: Empirical semivariogram of residuals for the Greater Accra (a) and Ashanti (b) regions. Circles display the observed semivariance at distance h and dashed lines indicate Monte Carlo simulation envelopes.

whereas 59% of the Ashanti communities actually reported cases. Sixty-nine percent of sites in Greater Accra were located within a district that reported cases, and 48% of communities actually reported cases. There was no significant association between *M. ulcerans* presence and district level reporting (p=0.06) or community level reporting (p=0.80). When added to the overall best fitting model, both community level reporting and district level reporting increased AICc by approximately 3 units. Therefore, neither summary of case reporting appreciably improved model fit nor explained variation in the presence of *M. ulcerans* after adjusting for environmental covariates.

3.5 Discussion

This is the first study to evaluate environmental factors associated with *M. ulcerans* in its natural habitat on such a broad scale. In the Greater Accra and Ashanti

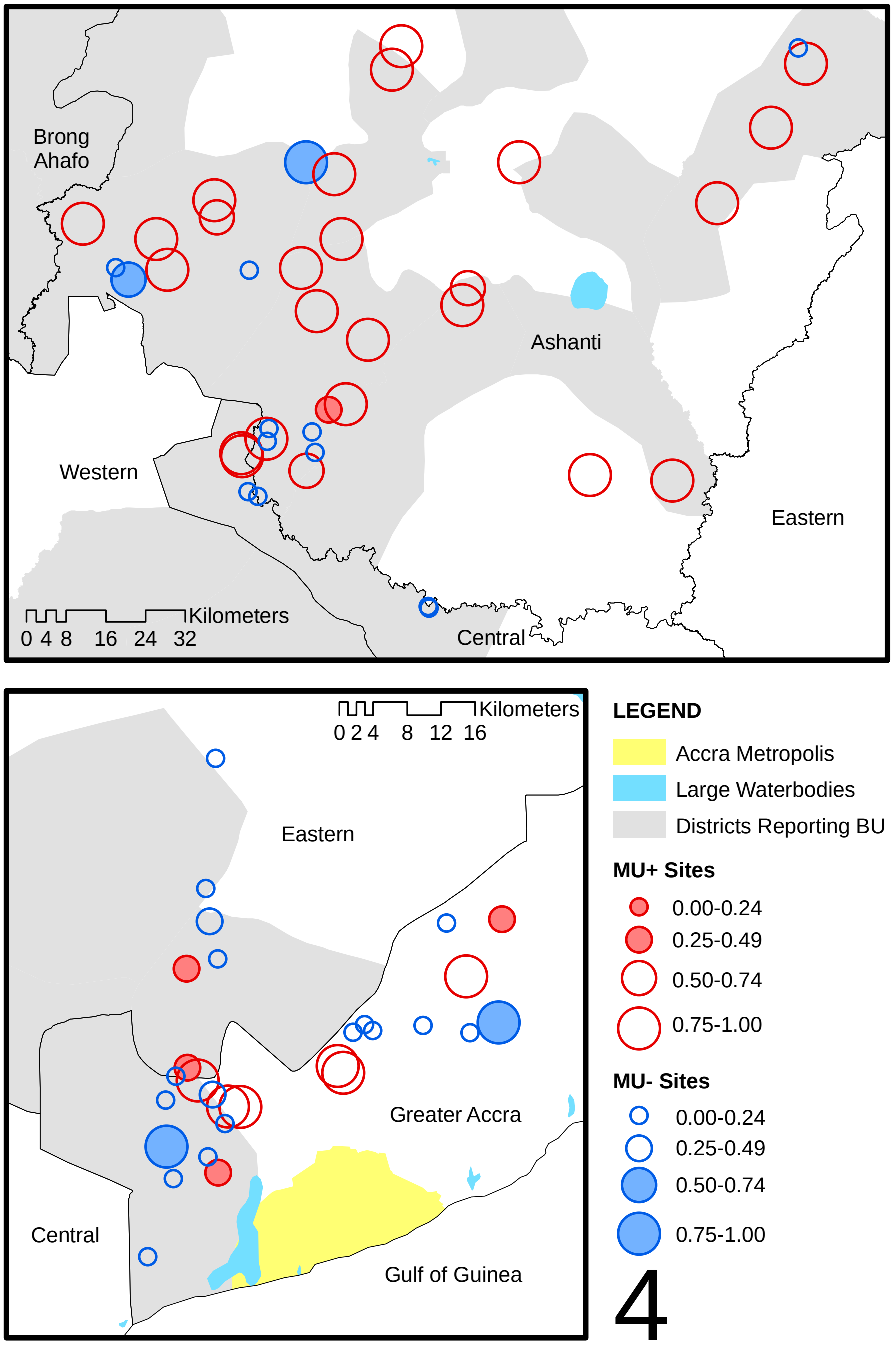


Figure 3.8: Map of the predicted probability of *M. ulcerans* positive based on the best fitting model. Blue circles represent sites that were actually *M. ulcerans* negative and red circles *M. ulcerans* positive. The size of the circle indicates the predicted probability, and shaded circles indicate poor fit (MU+ sites with low predicted probability or MU- sites with high predicted probability).

regions, no significant evidence of local or global clustering of aquatic habitats with *M. ulcerans* was present, suggesting the growth of *M. ulcerans* may be dependent on the local environment and may exist in isolated pockets. The best fitting model of those considered included elements from both on-the-ground highly localized measurements and broad scale remotely sensed features, indicating that characteristics of local aquatic systems, general land use/land cover, and topographic features were all associated with the presence of *M. ulcerans*. Some of these results concur with laboratory results or speculation on *M. ulcerans* growth, whereas other results diverge from published literature. We explore these agreements and discrepancies for the distinct models below.

3.5.1 Water Variables

The best fitting model relating the presence of *M. ulcerans* to physical and chemical properties of water contains dissolved oxygen percent saturation, nitrate, and calcium water hardness. Low oxygen and increased nutrients are known indicators of eutrophic aquatic conditions that were hypothesized to be related to *M. ulcerans* populations dynamics [37, 69, 47], which was later confirmed through laboratory studies [70, 45].

The relationship between water hardness and *M. ulcerans* has not been discussed previously in published literature. Water hardness quantifies the mineral content in water, and can be measured in terms of calcium or magnesium (primary components), or total hardness (which includes other ions). The hardness of an aquatic system is influenced naturally by the underlying geology of the system: as water passes through soil and rock it collects minerals which are deposited in the aquatic system. However, human activity on the watershed can also influence hardness. For example, drainage from mining sites can contribute a variety of minerals to an aquatic system, increasing its hardness.

Our model indicated a weak negative association between *M. ulcerans* and water hardness in Greater Accra sites, and a strong positive association with water hardness sites located in the Ashanti region. The interaction effect of water hardness and region could possibly be explained by the distinct underlying geological processes in the two regions as well as by differences in human activities. The sites sampled in the Greater Accra region are near the coast in which agriculture dominates the economy; sites sampled in the Ashanti region practice agriculture as well but are also located in prospective gold mining regions [71]. While Ashanti sites tend to have lower water hardness than sites in Accra, high values of water hardness in Ashanti could be a result of run-off from mines reflecting a disturbed environment and underlying water chemistry conditions potentially creating an environment which is conducive to *M. ulcerans*. Greater Accra sites, which tend to have higher values of water hardness, could inhibit *M. ulcerans* growth due to more complex and indirect effects of water hardness on other components of water quality such as pH and ion balance. Water hardness tends to be positively correlated with pH; sites with high water hardness could have pH values outside the optimal range for growth and survival of aquatic organisms. Furthermore, the Greater Accra region was characterized by higher conductivity, related to higher salt concentrations of those sites near the coast. The interactions between underlying geology, proximity to the ocean, and water table exchange with surface waters is complex. Evaluation of specific water quality conditions that may enhance the presence and size of *M. ulcerans* populations in different regions should be studied.

Certain aquatic factors commonly discussed in the literature with *M. ulcerans* such as temperature and waterbody flow did not contribute to our final model. All but one of the sampled aquatic habitats were below the optimal laboratory growing temperature of 30-33°C [72], suggesting environmental temperatures for population survival or growth may differ from laboratory conditions. Furthermore, BU disease occurrence has been associated with both still and moving waterbodies [49, 39, 73, 74, 50, 35, 51, 52]. However, the site classification of lentic versus lotic waterbodies did not improve model fit and therefore provided no insights into suitable aquatic conditions for *M. ulcerans* across the sites.

3.5.2 Land Use/Land Cover Variables

We chose to consider indicators for the presence of specific LULC categories within a buffer in addition to the percentage observed for two reasons. First, the percent LULC may not have a linear relationship with the log odds of *M. ulcerans* presence and the true relationship may be difficult to ascertain. Second, given the short buffer distances examined combined with the relatively coarse resolution of the satellite data, indicators of LULC presence provide a more robust measure than class percentages, reducing the potential impact of misclassification. Taken together, we find the LULC presence indicators provide additional flexibility in estimation and interpretation of observed associations within the data than the use of LULC percentages alone.

We identified two fine scale (<1km) LULC variables (as indicators of presence/absence) associated with *M. ulcerans*. Sites with more disturbed environments (urbanized, non-forested) were more likely to have *M. ulcerans* present compared to less disturbed environments (forested, non-urbanized). These results are in accordance with current literature indicating disturbed environments provide conditions suitable for *M. ulcerans* growth by affecting the physiochemical properties of water [73, 37, 53, 47]. For example, deforestation depletes riparian cover which may increase the temperature in aquatic systems to a degree necessary for *M. ulcerans* growth. Furthermore, urbanization can result in increased sedimentation in aquatic systems, attenuating UV penetration, and facilitating favorable conditions for *M. ulcerans* growth [37, 47].

3.5.3 Terrain Variables

Elevation, wetness index at the site, and variability of the wetness index in the vicinity of the site were found to be associated with *M. ulcerans* presence. Wetness index indicates the capacity for potential water pooling based on the slope and flow direction of the DEM, with higher values indicating higher potential for pooling. Our study found a negative association between *M. ulcerans* presence and wetness index at the site and a positive association with elevation. Areas of high wetness index or low elevation areas may be more prone to flooding or fast moving water that could wash out the natural habitat for *M. ulcerans*.

Wetness index variability had differing effects in the two regions. The positive association between wetness index variability and *M. ulcerans* presence in the Ashanti region could be attributed to variable wetness patterns enhancing conditions suitable for *M. ulcerans*. However, the negative association between *M. ulcerans* and wetness index variability in the Accra sites could be due to the sites' proximity to the coastline: conductivity measurements indicate that Greater Accra sites were saltier. If salt inhibits *M. ulcerans* growth, this could happen throughout the year in low lying sites during the dry season when the salinity of surface waters increase, but could change during the wet seasons when the landscape floods and water bodies are 'diluted', perhaps providing more suitable conditions for *M. ulcerans*.

3.5.4 Overall Model

The best fitting overall model contained elements from each category of covariates, which included ground-based measurements up to remotely sensed data. The residuals showed no evidence of spatial autocorrelation, indicating that a more sophisticated model taking into account the spatial locations of the sites was not necessary for our analysis. It is noteworthy that the semivariance of the residuals in Accra was greater than the semivariance in Ashanti, which indicated larger variability in the residuals of sites located in Greater Accra compared to Ashanti. The map of the predicted probability of *M. ulcerans* presence suggests more discordance between model prediction and observed outcomes in Greater Accra compared to Ashanti. Such discordant sites provide an opportunity for further investigation at specific spatial locations. *Mycobacterium ulcerans* negative sites with a high predicted probability of being positive could be re-sampled to verify the negative result, and *M. ulcerans* positive sites with a low predicted probability of being positive could be re-examined for unobserved covariates that may explain positive results.

3.5.5 Comparing Environmental Associations with *M. ulcerans* and Buruli ulcer

While *M. ulcerans* is the causative agent of Buruli ulcer, it is unclear whether we should expect similarities between environmental correlates of *M. ulcerans* presence and those of Buruli ulcer incidence and/or prevalence at a broad scale of observation. Our best-fitting overall model measuring associations with *M. ulcerans* presence contains similarities to and differences from published associations between comparable landscape covariates and reported cases of Buruli ulcer. In addition to differences in data quality and availability between disease surveillance and pathogen testing, simple presence of the pathogen in the environment may inhibit our ability to detect measurable local increases in reported cases. For example, certain environmental factors may provide suitable habitats for *M. ulcerans* in addition to being collocated with high human activity areas, thus possibly increasing exposure. Conversely, other environmental conditions associated with *M. ulcerans* presence may not promote human interaction with the environment, thus limiting exposure to the pathogen. Moreover, some of the variables associated with BU prevalence in other settings are defined on a broader geographic scale than our variables associated with *M. ulcerans*. Whereas coarse spatial BU disease patterns may be identifiable on a large geographic scale due to human behavior and broad environmental characteristics, fine scale geographic characteristics are likely more relevant to understanding the local ecology of *M. ulcerans.*

As specific examples, forest land cover and urbanization were shown to be positively and negatively associated with BU prevalence in Benin, respectively [55, 54]. Our results suggest associations with *M. ulcerans* in the opposite direction, showing negative associations with forest land cover and positive associations with urban land cover. A study in Côte d'Ivoire also demonstrated proximity to forest to be associated with higher BU incidence [60]. With respect to *M. ulcerans*, an argument can be made for forested areas both inhibiting and promoting *M. ulcerans* growth: forested areas with marshy ecosystems could act as a reservoir for the pathogen, however, lower temperatures due to riparian cover may not be conducive to *M. ulcerans* growth. Lastly, urbanization resulting in environmental disturbance may provide conditions suitable for *M. ulcerans*; nevertheless, activities occurring in such areas like the availability of pumped water may limit exposure to the pathogen.

As another example, our observed positive association between *M. ulcerans* presence and elevation differs from two different studies conducted in Benin showing negative associations between BU prevalence and altitude [54, 75]. In contrast, our measured associations between *M. ulcerans* and wetness index variability generally agree in direction with those observed with reported cases of BU. Many studies have implicated flooding as a risk factor for BU [49, 73, 76, 77, 37, 54], whereas low elevation areas with variable wetness patterns that could be prone to flooding could wash out the natural habitat for *M. uclerans.* Local variations and direction of effects demonstrate the need for additional focal studies.

3.5.6 Associations with Reported Buruli ulcer Cases

We investigated whether adding the presence of reported BU cases at the district or community level improved fit in our model. We did not find a significant unadjusted or adjusted association between *M. ulcerans* presence and either district level or community level case reporting history. If the case reporting history of villages is accurate, then the insignificance of these variables may suggest that *M. ulcerans* existence in nature is independent of human interaction [56]. The lack of association between *M. ulcerans* presence and case reporting could also signify that locations of reported BU cases are not limited to locations of *M. ulcerans* presence, implying a more complicated connection than simple collocation and suggesting that human behavior (particularly interaction with the environment) plays a role in transmission that has yet to be defined.

While continuing to improve in detail and accuracy, it is important to recognize that centralized case reports of Buruli ulcer represent a different type of data with accompanying challenges than the systematic testing of water bodies for *M. ulcerans* presence. Buruli ulcer surveillance involves a coordinated effort across institutions and treatment centers, relying on a variety of personnel from community health workers to district health officers. In Ghana, the surveillance system consists of both active and passive surveillance. Nevertheless, cases can be underreported for a variety of reasons including lack of awareness, stigma, costs associated with treatment, and proximity to health centers [78, 79]. We utilized two levels of case reporting in order to safeguard against potential misclassification of the villages' reporting status. The broader level of district case reporting may capture non-reporting communities that actually have cases; however, it may also incorrectly classify communities without BU cases. The narrower level of community reporting may not capture all communities with cases due to non- or delayed reporting. It should also be noted that locations of reported disease may not be the same as where disease acquisition occurred -

this could happen because an individual traveled for leisure or seasonal work, became infected while away, and then presented symptoms in their hometown. Therefore, it is possible that no associations were observed with case reporting due to the difficulty in correctly ascertaining a village's case reporting history or to the difficulty in correctly identifying the location of disease acquisition.

3.5.7 Limitations

Due to cost and time, sampling of each water body was performed only on a single day. The study systems are highly synergistic and hypereutrophic, meaning that they are nutrient rich and often subject to periods of excessive plant and other biomass growth and decay. This results in variability in the physiochemical properties of water throughout seasons or years which we were unable to capture in order to assess how it can affect *M. ulcerans*. Further, many of the sites were riverine wetlands that experience dynamic flooding and drying periods throughout the year. The ability to detect *M. ulcerans* could be sensitive to such weather events as natural habitats could be washed out. The fluctuations in water flow resulting from heavy and sporadic rainfalls render difficult categorization of a water body as lentic or lotic at a single point in time. Moreover, temperature of the aquatic site was assessed through point measurements whereas continuous temperature measurements are preferable to accurately quantify temperature. It is likely there were fine resolution temporal changes which occurred prior to sampling at some locations that we were unable to identify. For example, lack of precipitation data at the local scale inhibited our ability to address factors (e.g., rainfall and flooding) influencing temporal changes. Moreover, the complexities of the interactions between various components of water and their effect on aquatic ecosystems were difficult to disassemble and analyze separately. Temporal studies of both BU and *M. ulcerans* environmental distribution are needed.

We examined remotely sensed environmental covariates in buffers at relatively short distances (<5 km) under the assumption that the presence of *M. ulcerans* was highly dependent on the immediate surrounding environment. Note that groundtruthing of the LULC data was not performed as part of this study due to limited resources. To minimize the potential effect of misclassification, we used easily distinguished LULC categories. Small bodies of water could not be identified by the satellite imagery due to the coarse resolution, and therefore this LULC class could not be statistically analyzed in relation to *M. ulcerans* presence. It should be noted that the coarse resolution of the DEM data could underestimate the wetness variability in buffers surrounding sites. This study could be improved by the use of ground-truthed, high-resolution satellite imagery. Lastly, the dates of the satellite imagery (2000) did not coincide with the dates of our study time period (2005-2007). Therefore, there is a possibility that natural landscapes were urbanized, converted to agricultural practices, or stripped for mining during this time period which could affect our study results.

Our results are based on the presence or absence of *M. ulcerans* DNA as detected by PCR from suspended material in water and plant biofilm of environmental samples. This analysis focused on *M. ulcerans* positive water bodies, and did not consider other potential vectors or reservoirs of *M. ulcerans* such as aquatic insects or mammals. In addition, the number of positive samples and the DNA abundance were not quantified. The number of *M. ulcerans* positive samples could possibly be underestimated due to PCR inhibitors, though previous results suggest that detection methods employed were effective in eliminating PCR inhibitors [56].

3.6 Conclusion

The majority of findings of this study support previously posed hypotheses on the relationship between *M. ulcerans*, specific water conditions, and land use. Furthermore, we identified new associations between *M. ulcerans*, water hardness, and elevation.

Our research also demonstrated complex regional interactions limiting the ability to identify a specific set of universal factors which may be indicative of high risk environments for *M. ulcerans*. Covariates without regional interactions could potentially be used to create maps to identify areas suitable for *M. ulcerans*, whereas those with regional interactions merit further investigation into the underlying cause of the interaction. Continuous remotely sensed data (widely available) may be augmented by a well-planned water sampling strategy (much more time and resource intensive) to collect data for the creation of such maps. Furthermore, environmental sampling should be conducted over extended time periods (e.g., monthly for multiple years) as temporal changes in *M. ulcerans* and associated environmental conditions are needed to elucidate *M. ulcerans* ecology and BU transmission. As it appears that *M. ulcerans* is present in isolated pockets in the environment, we recommend utilizing high resolution remotely sensed data in targeted areas to better quantify these associations.

In contrast to other published research in which suitable habitats corresponded directly to disease risk [80, 68], areas suitable for *M. ulcerans* do not necessarily correlate to areas at high risk for acquiring Buruli ulcer as human interaction with the environment likely plays an important yet undefined role in disease acquisition. Locations of reported BU cases may differ from *M. ulcerans* positive locations. Identifying such discordant sites where *M. ulcerans* is present but no BU cases are reported or areas reporting BU cases with no local presence of the pathogen could help to elucidate human behaviors associated with disease acquisition. Moreover, future studies should include temporal aspects of pathogen detection and abundance along with identified or hypothesized environmental covariates. This could help identify environmental lag times necessary to detect *M. ulcerans* in specific habitats, much like modeling the effect of short term ambient air pollution on hospitalization due to cardiac or pulmonary disease or long term climate patterns that precede cholera outbreaks [81, 82, 83].

While few epidemiological studies have focused on the locations and environmental associations of BU disease, there have been no studies assessing the same for *M. ulcerans.* Knowledge of the ecology of *M. ulcerans* is crucial to understanding where the pathogen resides in the environment and factors which affect its growth. These details can highlight specific geographical areas in need of active disease surveillance, as well as provide insight into possible local modes of transmission. We found highly localized factors up to large-scale characterizations of environmental features were associated with the presence of *M. ulcerans*, and found no evidence of geographic clustering of *M. ulcerans* presence in neighboring aquatic systems. This research provides insights into conditions suitable for *M. ulcerans* growth and a basis for future research into the underlying ecology of the pathogen that causes Buruli ulcer disease.

Chapter 4

Assessments of and Modifications to Techniques Utilized for Data with False Zero Inflation

4.1 Overview

We are interested in formulating a statistical model to appropriately model cases of Buruli ulcer (BU) in Ghana. Buruli ulcer is a neglected tropical skin disease caused by the environmental pathogen *Mycobacterium ulcerans* (MU). Although many modes of transmission have been hypothesized, none have been conclusively identified (see discussion in Section 3.1). Our data are from the National Buruli ulcer Control Program (NBUCP) of Ghana, and represent Buruli ulcer case summaries from the 2008 district-level reports. Administratively, Ghana is divided into regions, and regions are subdivided into districts. In 2008, there were ten regions in Ghana comprised of 138 districts. Six of these ten regions reported BU cases. These six reporting regions were comprised of 89 districts: Ashanti (21), Brong Ahafo (19), Central (13), Eastern (17), Greater Accra (6), and Western (13) (Figure 4.1). A notable feature of our data is the substantial number of non-reports - 58/89 or 65% of the districts do not report BU cases. Programmatically, the NBUCP considers non-reports as reports of zero cases. While many of these non-reports represent the true absence of cases, there is reason to believe that some of these non-reporting districts are not truly disease-free. In the 31 reporting districts, the case counts range from 1 to 173, and the rate of cases per 10,000 individuals ranges from 0.01 to 16 (Figure 4.2).

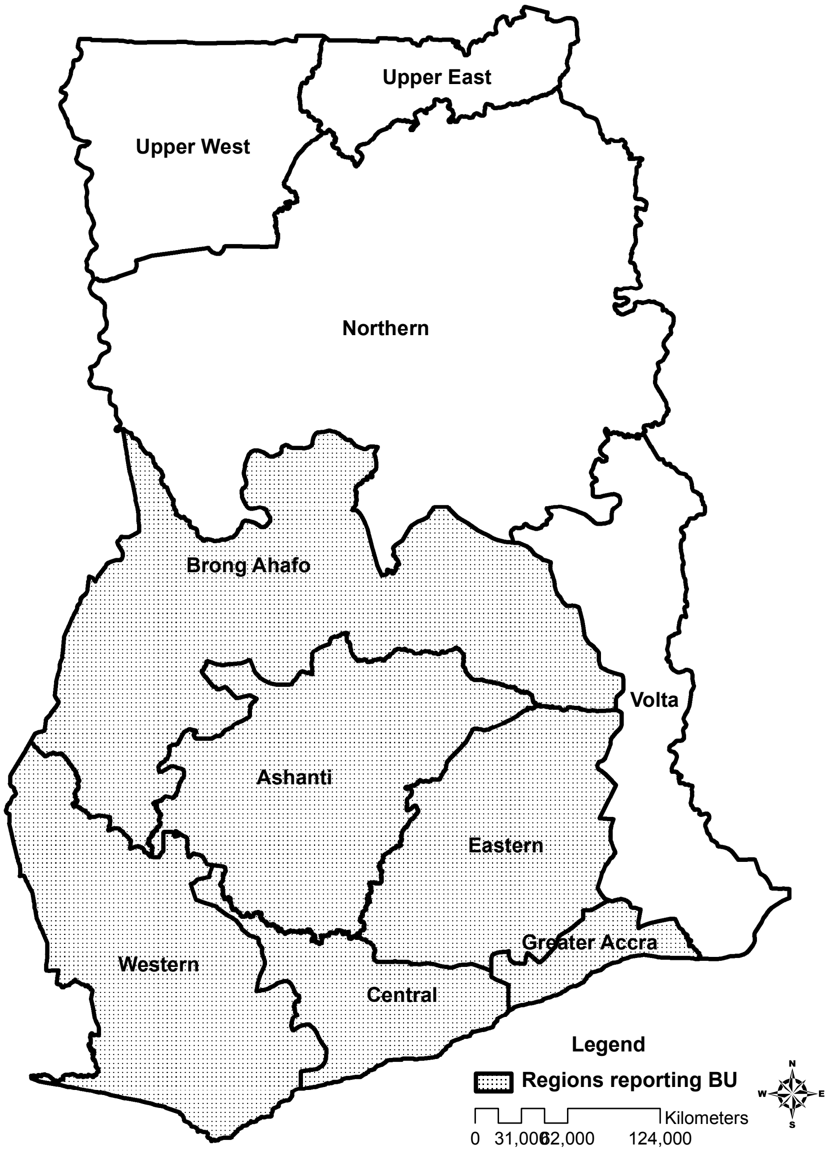


Figure 4.1: Six of ten regions in Ghana report cases of BU (2008)

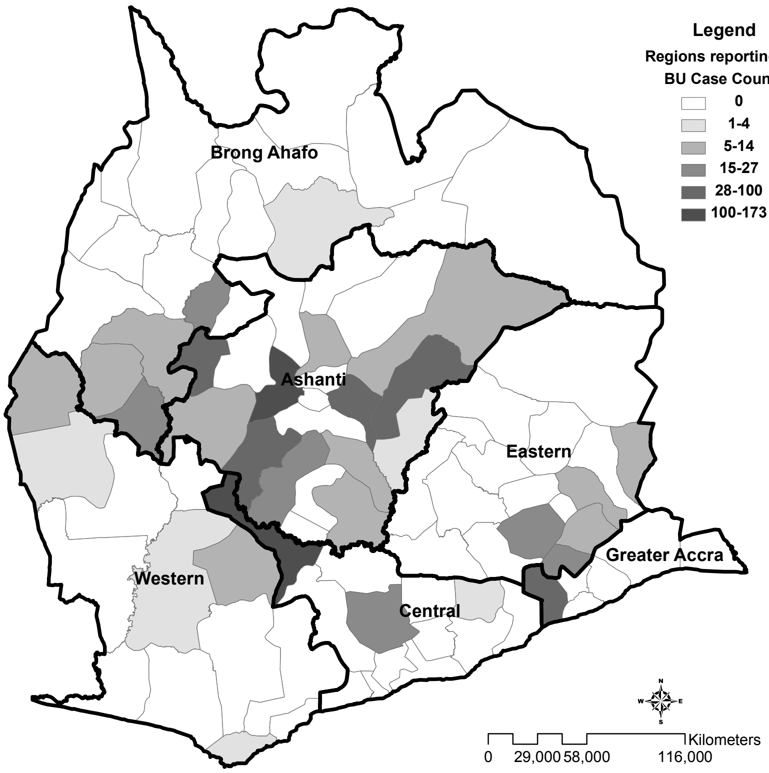


Figure 4.2: District counts of BU cases (2008). There are six regions affected by BU, comprised of 89 districts: Ashanti (21), Brong Ahafo (19), Central (13), Eastern (17), Greater Accra (6), and Western (13).

BU cases can be underreported at the individual level for a variety of reasons, including lack of awareness, stigma, costs associated with treatment, and proximity to health centers [78, 79]. Cases can also be subject to non-reporting at the district level. The NBUCP is a small office with four employees located on the southern coast of Ghana in the capital, Accra. Disease Control Officers (DCOs) at the district level are responsible for summarizing case reports within their district and forwarding this information to the regional office and then on to the NBUCP. The DCOs are often temporary 1-2 year assignments mandated by the national service requirements of Ghana. While many DCOs process and report cases in timely and accurate manner, some DCOs may not acclimate to the their position within the time frame of their temporary assignment, nor may they feel invested in their position. This could result in incomplete reporting. Moreover, as a neglected tropical disease competing with other high profile public health interests such as HIV/AIDs or tuberculosis, BU reporting may be low on the public health totem pole. The NBUCP lacks both the manpower and the resources to ensure accurate reporting from each DCO. It is also possible that the carbon copies of the BU case reports get misplaced or lost on the way to the capital, Accra. The most important consequence from all of this is that programmatically the NBUCP considers no reports of BU cases as reports of zero BU cases, and districts with no reports do not receive training in early case detection or case management.

The data we have received from the NBUCP only includes case counts from reporting districts, and it is assumed that non-reporting districts have zero cases. However, it is highly likely that some non-reporting districts are actually false zeros - districts which have BU cases that were not reported. It is our goal to develop statistical methodology to differentiate non-reporting districts are truly disease-free and non-reporting districts which are likely to have cases in order to make recommendations for allocation of resources.

As we observe a substantial number of non-reports that are considered to be reports of zero cases, we can consider our data to be zero-inflated. A plethora of statistical models have been proposed to model zero-inflated count data, including zero-inflated Poisson (ZIP) models and hurdle models [84]. Typically, zero-inflated models are used for scenarios in which the excess zeros arise by some mechanism generating true zeros. For example, in a manufacturing process monitoring the number of defects on a device, a near-perfect process would result in many instances of zero defects being present. However, zero inflation may also be generated by false zeros - observations in which counts are truly present but not observed. This could occur in ecological monitoring or disease surveillance. Although ZIP models are currently being utilized to model data with false zero inflation, to our knowledge the perfor-

Outcome	**Structural Zeros**	**Random Zeros**
Number of disease lesions on a plant [84]	A plant may have no lesions because it is resistant to disease	A plant may have no lesions because no spores have landed on it
Number of times a subject used medical services in the past year [94]	A patient may avoid doctors	A patient chooses not to visit doctors by chance
Number of animals in an area [98]	A species may be absent in a habitat because the habitat is unsuitable	A species may be absent in a habitat by chance due to the ecological dispersal process
		A species may be recorded as absent in a habitat by chance due to sampling or observer error
Number cases of dengue fever in Rio de Janeiro [99]	Cases may be absent due to the mosquito vector being absent	Cases may be absent due to an individuals immunologic resistance by chance
		Cases may be recorded absent due to underreporting

Table 4.1: Examples of traditional approach to zero inflated data with structural and random zeros. In the circumstance of imperfect detection, zeros can also be thought of as true and false zeros. Examples of false zeros are in red, whereas all other zeros may be considered as true zeros.

mance of ZIP models has not been evaluated for data which contain false zeros. In this last chapter we review the traditional ZIP model, discuss the application of ZIP models to data subject to false zeros, and propose a hierarchical zero-inflated model that accommodates false zeros with the ability to differentiate between true and false zeros.

4.2 Traditional ZIP Models

Excess zeros are said to be present in data when the observed frequency of zeros greatly exceeds the number expected given the distributional assumptions on the data.

Data with excess zeros can arise in many applications including industrial [85, 86], ecological [87, 88, 89, 90], horticultural, [91, 92, 93] and medical [94, 95, 96]. Excess zeros are classically described as structural versus random zeros. A structural zero occurs in situations where it is impossible to observe a response, whereas random zeros occur by chance according to the probability distribution describing the observation process [97]. Table 4.1 provides examples of the classical framework for zero inflated models, along with true and false zeros.

The most commonly utilized models for data with excess zeros are zero-inflated models and hurdle models [84]. Zero-inflated models are a mixture of a point mass at zero and a standard distribution for count data such as Poisson or negative binomial. A zero-inflated model utilizing a Poisson distribution is known as the zero-inflated Poisson (ZIP) model. Similarly, a hurdle model is a mixture of a point mass at a certain observation(s) and a truncated distribution for the remaining observations. A hurdle model often compared to ZIP model considers a point mass at zero, and a truncated Poisson distribution for the remaining observations. Hurdle models consider all zeros together, whereas zero inflated models should be able to partition the zeros into structural and random zeros. Other zero-inflated models account for overdispersion by modeling the count data with the negative binomial or generalized Poisson distributions [100, 98, 96, 84, 101].

A zero-inflated distribution is defined by

$$\begin{aligned} \Pr(Y = 0|p, \theta) &= p + (1 - p)f(0|\theta) \\ \Pr(Y = y|p, \theta) &= (1 - p)f(y|\theta), y > 0 \end{aligned} \tag{4.1}$$

for some parametric distribution $f(y|\theta)$, such as the Poisson distribution, where $f(y|\theta) = \dfrac{\theta^y \exp(-\theta)}{y!}$.

The zero-inflated distribution can be thought of as a joint distribution involving a latent random variable Z that indicates if the zero is a structural zero. In this situation, let $p = \Pr(Z = 1)$, which indicates the overall probability that an observation is an structural zero. The joint distribution of Y and Z is

$$\begin{aligned} &\Pr(Y = 0, Z = 1|p, \theta) = p \\ &\Pr(Y = y, Z = 0|p, \theta) = (1 - p)f(y|\theta). \end{aligned} \tag{4.2}$$

Conditionally,

$$\begin{aligned} &\Pr(Y = 0|Z = 1) = 1 \\ &\Pr(Y = y|Z = 0, p, \theta) = f(y|\theta) \\ &\Pr(Z = 1|Y = y > 0) = 0 \end{aligned} \tag{4.3}$$

More interestingly, we can calculate the probability that a zero is a structural zero given that we observe a zero:

$$\begin{aligned} \Pr(Z = 1|Y = 0, p, \theta) &= \frac{\Pr(Z = 1 \bigcap Y = 0|p, \theta)}{\Pr(Y = 0)} \\ &= \frac{p}{p + (1 - p)f(0|\theta)}. \end{aligned} \tag{4.4}$$

Using the latent variable approach, the full data likelihood for the model takes the form

$$\begin{aligned} L(p, \theta; Y, Z) &= \prod_{i=1}^{n} \Pr(Y_i = y_i|Z_i = Z_i)P(Z_i = z_i) \\ &= \prod_{i=1}^{n} p_i^{z_i}((1 - p_i)f(y_i|\theta_i))^{1-z_i} \\ &= \prod_{y_i>0} (1 - p_i)f(y_i|\theta_i) \prod_{y_i=0} p_i^{z_i}((1 - p_i)f(0|\theta_i))^{1-z_i}. \end{aligned} \tag{4.5}$$

The observed data likelihood is

$$L(p, \theta; Y, Z) = \prod_{i=1}^{n} p_i I(y_i = 0) + (1 - p_i) f(y_i|\theta_i). \tag{4.6}$$

Furthermore, the parameters p and λ can be modeled as linear functions of covariates. It is natural to model these two parameters in terms of their canonical links [85].

$$\begin{aligned} \log(\lambda) &= X\beta \\ \text{logit}(p) &= W\alpha \end{aligned} \tag{4.7}$$

Here, β and α represent parameter vectors associated with the rate of case counts and with the probability of being a structural zero, respectively. The sets of covariates W and X could be the same or they could be different. As Lambert noted, in some circumstances it may make sense that p and λ are functionally related [85]. That is, the same covariates may affect both p and λ, and rather than estimating two separate sets of coefficients for both the p model and the λ model, we can estimate one set of coefficients of which the other is a linear function. This makes sense as covariates that are associated with a higher Poisson mean would likely also be associated with a lower probability of zeros. In this circumstance,

$$\begin{aligned} \log(\lambda) &= X\beta \\ \text{logit}(p) &= \tau X\beta \end{aligned} \tag{4.8}$$

This is designated as the ZIP(τ) model.

In the first presentation of the ZIP and ZIP(τ) models, maximum likelihood estimates of parameters were obtained by the EM algorithm [85]. The performance of the ZIP and ZIP(τ) models was evaluated through simulation to explore conformance with asymptotic theory based on finite samples. Simulations showed some convergence issues for the ZIP(τ) model, but in general found MLE's to be approximately normally distributed for large sample sizes. Extensions to zero-inflated models within the maximum likelihood estimation framework include bivariate normally distributed random effects in the λ and p model compartments for longitudinal data [94], nested random effects for clustered data [102], and multivariate ZIP models [103].

Alternatively, many researchers are turning to Bayesian estimation for zero-inflated models. Ghosh et al. (2006) present a fully Bayesian approach to ZIP models with Markov Chain Monte Carlo (MCMC) simulation-based methods implemented through Gibbs sampling in WinBUGS [86]. Simulation results showed Bayesian estimation to be competitive with maximum likelihood estimation estimation, with improved small-sample performance. Moreover, Bayesian estimation performed better when the probability of observing a zero in the outcome Y was close to one. In general, a Bayesian approach to ZIP models allows for incorporation of prior information, facilitates estimation of functions of parameters, and reduces small-sample bias compared to maximum likelihood estimation.

Bayesian estimation of zero-inflated models has been developed for many applications. Dagne (2004) utilized independent normally distributed random effects in λ and p to analyze longitudinally correlated count data [104]. Neelon et al. (2010) discussed Bayesian longitudinal data analysis utilizing bivariate normal random effects in λ and p models for three types of zero-inflated models. Both Xue-Dong (2009) and Dagne (2010) presented a semi-parametric framework for longitudinal data analysis with random effects and a non-parametric component to model the effect of time or time-varying covariates [105, 91].

Extensions have also been made for the analysis of spatially correlated zero-inflated count data. Agarwal et al. (2002) employed a spatial Bayesian hierarchical model with a conditional autoregressive (CAR) prior distributions on random effects [87]. The CAR random effect was initially proposed by Clayton and Kaldor (1987)

[106]; Besag et al. utilized a fully Bayes implementation of the CAR prior [107]. This prior is constructed such that the effects ψ_i are Gaussian with the conditional prior mean of any spatial random effect defined as a weighted average of its neighboring effects ψ_j, $j \neq i$.

$$\psi_i|\psi_{j\neq i} \sim N\left(\frac{\sum_{j\neq i} c_{ij}\psi_j}{\sum_{j\neq i} c_{ij}}, \frac{1}{\nu_{CAR}}\sum_{j\neq i} c_{ij}\right), \quad i = 1, \ldots, N \tag{4.9}$$

Here, c_{ij} is a variable defining the neighborhood structure of the spatial data. Typically, $c_{ij} = 1$ if site j neighbors site i, and $c_{ij} = 0$ otherwise. However, other weighting options are available [108]. In specifying the variance of the prior distribution of the spatial random effect, ν_{CAR} is a hyperparameter for the conditional variance of ψ_i given other $\psi_j, j \neq i$ [26]. Agarwal et al. consider a CAR spatial random effect in the λ model, but forgo the CAR spatial random effect in the p model due to unstable model fitting. Argarwal et al. also discuss informative prior specification and issues of posterior propriety. The authors utilized an adjusted Gibbs sampler to perform posterior sampling for parameters. Gschlößl and Czado (2008) discuss the use of CAR random effects to account for overdispersion in various types of zero-inflated models [100].

The traditional framework of the zero-inflated model is well-equipped to handle the situation in which data are comprised of structural and random zeros which arise from natural processes and both the structural and random zeros can be considered as true zeros. However, zero-inflated models are currently being applied to processes which contain false zeros. We do not believe the existing framework of ZIP models is appropriate for such circumstances, as detailed in Section 4.3.

4.3 ZIP Models Applied to Processes with Imperfect Detection

Recently, data with excess zeros have been described in processes with imperfect detection such as ecological monitoring or disease surveillance. Imperfect detection occurs when data accuracy cannot be corroborated, resulting in false negative reports. This means that the item of interest was present, but unobserved, which happens often in animal surveillance. Martin et al. (2005) provide a nice overview of simple applications of zero-inflated models in ecology [98]. There is also there is a large body of literature on more complex models incorporating imperfect detection processes. This literature includes different types of zero-inflated models, and focuses on modeling heterogeneity in detection probability to explain the zeros in the data [109, 110, 111, 112, 113]. The foundation of such models is repeated measures in which specific sites are monitored multiple times for a species' presence.

Although multiple site observations are common in ecological studies, a few studies in ecology have utilized zero-inflated models in the absence of repeated measures. Flores et al. modeled tropical saplings density using the spatial ZIP model with CAR random effects in λ, and Kuhnert et al. modeled bird density with ZIP models incorporating expert opinion in prior distributions [88, 92]. These authors note the distinction between true and false zeros, and correspondingly describe the different sources of error in the observation process by which false zeros can arise [88, 98, 92]. They also claim that in general, true zeros can be structural or random, whereas false zeros arise from sampling mechanisms and can be considered as random zeros (Table 4.1). While we agree that false zeros can arise from observer error, we do not agree that false zeros should be considered as random zeros in the zero-inflated modeling framework.

In disease surveillance it is more common to obtain cross-sectional data that represent a snap-shot of disease occurrence at a specific time rather than repeated observations. While some disease surveillance scenarios many not have reason to suspect false zeros, in monitoring of neglected disease there should be cause for concern. For Buruli ulcer district-level reports, we would reasonably expect to observe case counts where cases are present, true zeros where districts truly are BU-free, and false zeros where districts truly have BU but do not report any cases. This is akin to ecology's vision of true and false zeros.

However, these types of zeros do not align easily with the concepts of structural and random zeros which traditionally consider a zero arising from reporting error as a random zero. We argue that a false zero should be considered a type of structural zero. However, since in general structural zeros refer to situations in which it is impossible to observe an outcome, we propose a more precise terminology. We can consider *distributional zeros* as zeros that can arise from the distribution under consideration with reasonable probability (formerly *random* zeros). *Excess zeros* are zeros that cannot reasonably arise from the distribution under consideration (formerly *structural* zeros). Distributional zeros are a type of true zero, whereas excess zeros can be generated by true or false zeros. We will use this framework to model cases of Buruli ulcer. Table 4.2 provides examples of our new conceptual framework.

Definitions

A correctly observed zero where the disease is truly absent can be a **distributional true zero**. This could represent a location which is disease susceptible, i.e., the pathogen is present in the environment, but cases do not occur.

A correctly observed zero where the disease is truly absent can also be an **excess true zero**. This could represent a location which is not susceptible to the disease, i.e., the pathogen is absent in the environment, and therefore it is impossible for cases to occur.

An incorrectly observed zero where cases are truly present but are not reported is an **excess false zero**. This is an observation that should follow the distribution at hand, e.g. Poisson, but is incorrectly observed as a zero.

In literature review, we have found only one application of ZIP models to disease surveillance with suspicion of false zeros. Fenandes et al. modeled cases of dengue fever in Brazil using a spatio-temporal ZIP model for 156 districts over 77 time points in which epidemiologists suspected a large amount of underreporting [99]. They utilized Equation 4.4 to estimate the probability than a non-reporting observation actually had cases present. They assert that if this probability estimate has a high standard error for a specific observation, then this might indicate districts “which are *suspicious* of having under-reporting.” This is the extent to which they assess false zeros.

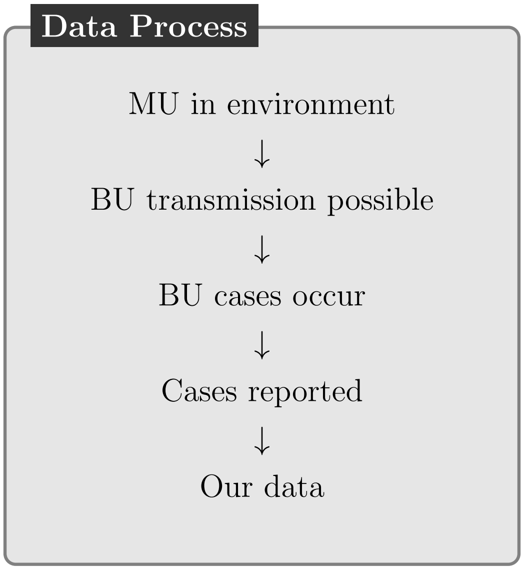
In the subsequent sections, we develop a hierarchical model for our new conceptual framework for zero-inflated models, describe estimation of the probability that an observation is a false zero, and provide simulations to evaluate the performance of both traditional ZIP models and our new hierarchical ZIP model in the presence of false zeros.

	Distributional Zeros	**Excess Zeros**	
	True	True	False
The idea	An event could happen, but does not.	It is impossible for the event to occur, so it does not.	The event does happen, but it is not monitored.
Animal	A habitat is suitable for an animal, but no species are present due to random dispersal process. That is, the animal could be there but is not.	A species is absent in a habitat because the habitat is unsuitable. That is, the animal is not observed because the animal is not and could not be there.	A species is present in a habitat, but the animals are not identified. This results from observer or sampling error.
Disease	Population is susceptible as pathogen and/or vector is present in the environment, but no cases occur. This could be due to immunologic resistance or because transmission does not occur. That is, the disease could be there but is not.	Population is not susceptible to disease because the vector and/or pathogen is not present in the environment. That is, the disease is not and could not be there.	Cases of disease are truly present, but no cases are reported. This results from reporting error.

Table 4.2: Examples of new approach to zero inflated data subject to imperfect detection with excess and distributional zeros. Here, excess zeros can be thought of as true or false zeros. Examples of false excess zeros are in red and true excess zeros are in blue, whereas all distributional zeros may be considered as true zeros.

4.4 Hierarchical Zero-Inflated Models

Accurate reporting of Buruli ulcer cases can be thought of as a multi-layer hierarchical process. First, the environmental pathogen that causes Burul ulcer, *M. ulcerans*, must be present in the environment. Second, given that MU is present in the environment, there must be circumstances related to transmission conditions under which it is possible to acquire Buruli ulcer. For example, these circumstances could be related to human interaction with the environment such as agricultural practices. Third, given that transmission is possible, Buruli ulcer cases may or may not occur for reasons such as an individual's disease susceptibility. Fourth, given that cases of the disease do occur, these cases must be accurately reported to the NBUCP. After all of the above conditions have been satisfied, we see our final data.



Data Process

MU in environment

↓

BU transmission possible

↓

BU cases occur

↓

Cases reported

↓

Our data

We can model this process as a hierarchical model involving four latent random variables. The first three indicate the underlying, unobserved truth about the state of three processes.

$$Z_{MU} = \begin{cases} 1 & \text{MU present in environment} \\ 0 & \text{MU absent in environment} \end{cases}$$

$$Z_{BU} = \begin{cases} 1 & \text{BU transmission is possible} \\ 0 & \text{BU transmission is not possible} \end{cases}$$

$$Z_{REP} = \begin{cases} 1 & \text{Reporting occurs at the district level} \\ 0 & \text{Reporting does not occur at the district level} \end{cases}$$

The variables Z_{MU}, Z_{BU}, and Z_{REP} are unobserved latent random variables that indicate MU presence, BU transmission, and reporting occurrence. These latent random variables describe the unknown, unobserved, but true state of the system. $Z_{MU} = 1$ means that *M. ulcerans* is present in the environment, and $Z_{BU} = 1$ implies that conditions for Buruli ulcer transmission were satisfied. $Z_{REP} = 1$ implies that all cases were reported, and $Z_{REP} = 0$ means that no cases were reported.

In addition to the these three latent random variables, we can utilize one last latent random variable, Y_{TRUE}, to model the true but unobserved distribution of case counts. Y_{TRUE} can be modeled by a Poisson distribution, where our observed outcome Y_{OBS} equals Y_{TRUE} if and only if $Z_{REP} = 1$. Note that in using a Poisson distribution to model case counts once conditions for transmission have been satisfied, we could still observe a distributional true zero with a certain probability. For example, if we were modeling a random variable Y with a Poisson distribution where $\lambda = 3$, then $\Pr(Y = 0|\lambda = 3) = \exp(-3) = 0.05$.

The probabilities associated with an event in each of the three latent random variables Z_{MU}, Z_{BU}, and Z_{REP} could be modeled as a generalized linear function of covariates using the logit link function. Also, the intensity of BU cases (λ) may be modeled as a generalized linear function of covariates using the log link function.

$$
\begin{aligned}
\text{logit}(p_{MU}) &= X_{MU}\beta_{MU} \\
\text{logit}(p_{BU}) &= X_{BU}\beta_{BU} \\
\log(\lambda) &= X_{CASES}\beta_{CASES} \\
\text{logit}(p_{REP}) &= X_{REP}\beta_{REP}
\end{aligned} \tag{4.10}
$$

The variables X_{MU}, X_{BU}, and X_{REP} represent sets of covariate matrices that may be associated with the latent random variables indicating MU presence, BU transmission, and reporting occurrence. X_{CASES} represents a set of covariates that may be related to mean case intensity once conditions for disease transmission have been satisfied.

In modeling the probability of reporting, we could also chose other functions. For instance, if there was reason to believe that the probability of reporting never reached 100% and plateaued at a certain percent less than 100, it could be appropriate to model this probability using the nonlinear logistic growth model discussed in Section 2.3. Moreover, it is also a reasonable scenario that the probability of reporting could also depend on the unobserved, true case counts, Y_{TRUE}, as districts with fewer cases may be less likely to report than districts with a substantial number of cases.

Equations 4.11 and 4.12 and below display two versions of our hierarchical modeling framework. In Model 4.11, Y_{TRUE} is explicitly modeled, and Y_{OBS} and Z_{REP}

are conditioned on Y_{TRUE}. In Model 4.12 Y_{TRUE} is not explicitly modeled.

$$\begin{aligned}
&Y_{OBS}|Z_{MU}, Z_{BU}, Y_{TRUE}, Z_{REP}, X_{MU}, X_{BU}, X_{REP}, X_{CASES} \sim \text{Poisson}(\lambda \times Z_{REP})\\
&Z_{REP}|Z_{MU}, Z_{BU}, Y_{TRUE}, X_{MU}, X_{BU}, X_{REP}, X_{CASES} \sim \text{Bernoulli}(p_{REP} \times \text{I}(Y_{TRUE} > 0))\\
&Y_{TRUE}|Z_{MU}, Z_{BU}, X_{MU}, X_{BU}, X_{CASES} \sim \text{Poisson}(\lambda \times Z_{MU})\\
&Z_{BU}|Z_{MU}, X_{MU}, X_{BU} \sim \text{Bernoulli}(p_{BU} \times Z_{MU})\\
&Z_{MU}|X_{MU} \sim \text{Bernoulli}(p_{MU})
\end{aligned} \tag{4.11}$$

$$\begin{aligned}
&Y_{OBS}|Z_{MU}, Z_{BU}Z_{REP}, X_{MU}, X_{BU}, X_{REP}, X_{CASES} \sim \text{Poisson}(\lambda \times Z_{REP})\\
&Z_{REP}|Z_{MU}, Z_{BU}, X_{MU}, X_{BU}, X_{REP} \sim \text{Bernoulli}(p_{REP} \times Z_{BU})\\
&Z_{BU}|Z_{MU}, X_{MU}, X_{BU} \sim \text{Bernoulli}(p_{BU} \times Z_{MU})\\
&Z_{MU}|X_{MU} \sim \text{Bernoulli}(p_{MU})
\end{aligned} \tag{4.12}$$

We present both Model 4.11 and Model 4.12 for multiple reasons. Model 4.11 is the model that best represents the actual data reporting process. In 4.11, $\text{I}(Y_{TRUE} > 0)$ indicates that cases are truly present, and therefore the occurrence of reporting is conditioned on the event that cases are truly present. This mirrors reality as we actually only observe case reports for districts that have cases, and we do not receive reports of zero cases from disease-free districts. This model also allows for Z_{REP} to be conditioned on Y_{TRUE} in another manner as well. It could be considered that reporting occurrence may also depend on the actual value of Y_{TRUE} (in addition to conditioning on the indicator that $Y_{TRUE} > 0$). For example, we could assume that $\text{logit}(p_{REP}) = X_{REP}\beta_{REP} + \alpha Y_{TRUE}$. This implies that the probability of reporting occurrence is positively associated with Y_{TRUE}, meaning that a larger number of cases would be more likely to be reported compared to a smaller number of cases. Lastly, the observed number of cases, Y_{OBS}, depends on the actual value of Y_{TRUE}

and whether or not reporting occurred, Z_{REP}.

Model 4.11 has four latent random variables with multiple conditioning arguments, and therefore it is likely to be challenging to fit and observe convergence. For these reasons, we present Model 4.12 as a relatively simpler alternative in that it ignores the underlying distribution of Y_{TRUE}. Here, Z_{REP} is conditioned on the possibility that cases are present ($Z_{MU} = 1$, $Z_{BU} = 1$), as opposed to explicitly conditioning on cases being truly present ($Y_{TRUE} > 0$). Y_{OBS} is then conditioned on reporting occurrence. Model 4.12 is justifiable as a sufficient alternative to Model 4.11 in that at this point in time, it is not an objective to estimate the true, but unobserved, case counts in non-reporting districts. While this is a logical extension of this research, for now we seek to determine the impact of false zeros on λ and the ability of the models to differentiate non-reporting districts as true zeros and false zeros.

In both Models 4.11 and 4.12 we have side-stepped the issue of whether or not true zeros should be subject to reporting occurrence, Z_{REP}. That is, if a district does not report cases because it truly has no cases, should we consider whether or not these true zeros are accurately reported? It is a sticky point, but ultimately has no effect on either Y_{TRUE} or Y_{OBS}. If $Y_{TRUE} = 0$ because $Z_{MU} = 0$, $Z_{BU} = 0$, or due to chance by the Poisson distribution, then regardless of whether or not $Z_{REP} = 0$ or $Z_{REP} = 1$, $Y_{OBS} = 0$ still. Therefore, the effect of Z_{REP} is only relevant when $Y_{true} > 0$, and does not have an effect on our observed outcome when $Y_{TRUE} = 0$. Again, this also mirrors the real reporting process as only case counts are actually reported, and not reports of zero cases.

Also, note that in this framework we are modeling the occurrence of reporting at the district level and not underreporting, or the possibility that individual cases may not be reported to the district to begin with. While this is an important and relevant consideration, here we are addressing the fact that no reports of BU at the district level are considered to be reports of zero cases.

This model assumes the underlying distribution of persons at risk to be homogenous. Note, however, that we could consider the scenario of heterogenous population distribution and model $Y_{TRUE} \sim Poisson(\lambda \times n \times Z_{BU})$.

The observed data likelihood for Model 4.11 is

$$\begin{aligned}
L(p,\theta;Y,Z) = \prod_{i=1}^{n} \{ &\mathrm{I}(y_i = 0)\,[(1 - p_{iMU}) \\
&+ p_{iMU}(1 - p_{iBU}) \\
&+ p_{iMU}p_{iBU}f_i(0) \\
&+ p_{iMU}p_{iBU}(1 - f_i(0))(1 - p_{iREP})] \\
&+ \mathrm{I}(y_i > 0)\,[p_{iMU}p_{iBU}f(y_i)p_{iREP}]\} \\
= \prod_{i=1}^{n} \{ &\mathrm{I}(y_i = 0)\,[1 - p_{iMU}p_{iBU}p_{iREP}(1 - f_i(0))] \\
&+ \mathrm{I}(y_i > 0)p_{iMU}p_{iBU}f(y_i)p_{iREP}\}. \qquad (4.13)
\end{aligned}$$

Because of the conditioning argument on Y_{TRUE} in Model 4.11, zeros and case counts contribute separately to this likelihood. The contribution given by observed zeros is the complement of the probability of observing a reported case count, and the contribution from a case count is the probability that we observed a reported case count.

The observed data likelihood for Model 4.12 is

$$L(p,\theta;Y,Z) = \prod_{i=1}^{n}\{I(y_i = 0)\,[1 - p_{iMU}p_{iBU}p_{iREP}] + p_{iMU}p_{iBU}p_{iREP}f(y_i)\}. \qquad (4.14)$$

This likelihood looks more similar to the likelihood for by the traditional ZIP model in Equation 4.6 because we are not conditioning on Y_{TRUE}. The presentation of this likelihood is based on the complements of elements in the traditional ZIP model since in the ZIP model $Z = 1$ implied a zero event, whereas in Model 4.11 and 4.12

$Z_{MU} = 1$, $Z_{BU} = 1$, and $Z_{REP} = 1$ implies a non-zero event.

For the sake of simplicity and generalizability there are some shorthand notations listed above that should be clarified. The expression $f_i(0)$ indicates the probability that we observe zero cases under some distributional function f for the i^{th} observation, which depends on the parameters β_{CASES} related to case intensity. For the model at hand, f is the Poisson distribution with mean λ, which is a function of β_{CASES} such that $\log(\lambda) = \beta_{CASES} X_{CASES}$. Therefore, $f_i(0) = \Pr(Y_{iTRUE} = 0|\beta_{CASES}) = \dfrac{\lambda_i^0 \exp(-\lambda_i)}{0!} = \exp(-\lambda_i) = \exp(-\exp(X_{iCASES}\beta_{CASES}))$.

Moreover, the probabilities associated with Z_{MU}, Z_{BU}, and Z_{REP} are also modeled as generalized linear functions of covariates using the logit link function. They could be expressed as, for example, $p_{iMU} = \Pr(Z_{iMU} = 1|\beta_{MU}) = \dfrac{\exp(X_{iMU}\beta_{MU})}{1 + \exp(X_{iMU}\beta_{MU})}$. For the sake of shorthand notation, these probabilities are simply referred to as p_{iMU}, p_{iBU}, and p_{iREP}, but it should be expressly noted that each of these probabilities depend on the parameters β_{MU}, β_{BU}, and β_{REP} (as well as their associated covariate values).

These models can be related back to the framework discussed in Section 4.3 regarding ZIP models for data subject to false zeros. Each type of data we could observe including case counts, distributional true zeros, excess true zeros, and excess false zeros can be represented by some combination of the latent random variables in the different model compartments (Table 4.3).

If Table 4.3 is re-organized as in Table 4.4 below, it more naturally follows our hierarchical framework and clearly illuminates the five types of data that we may observe in this framework. Note that we do not consider the effect of reporting on true zeros, i.e. we do not consider reported true zeros and non-reported true zeros, but rather just the event of a true zero.

We can express the five types of data provided by the hierarchical framework in terms of the events that must occur in the true underlying process that generates

Model compartment	Data type	Z_{MU}	Z_{BU}	Y_{TRUE}	Z_{REP}
Poisson distn	case count	$+$	$+$	$+$	$+$
	distributional zero	$+$	$+$	$-$	
Excess zeros	true zero	$-$			
	true zero	$+$	$-$		
	false zero	$+$	$+$	$+$	$-$

Table 4.3: Hierarchical zero-inflated model related back to ZIP framework incorporating false zeros. For the latent random variables, a plus indicates that the event occurred and a minus indicates that it did not. For Y_{TRUE}, the unobserved true number of cases, a plus indicates that $Y_{TRUE} > 0$ whereas a minus indicates $Y_{TRUE} = 0$.

Data type		Z_{MU}	Z_{BU}	Y_{TRUE}	Z_{REP}
1.	excess true zero	$-$			
2.	excess true zero	$+$	$-$		
3.	distributional true zero	$+$	$+$	$-$	
4.	excess false zero	$+$	$+$	$+$	$-$
5.	case count	$+$	$+$	$+$	$+$

Table 4.4: Five types of data observed in the hierarchical zero-inflated Poisson framework. For the latent random variables, a plus indicates that the event occurred and a minus indicates that it did not. For Y_{TRUE}, the unobserved true number of cases, a plus indicates that $Y_{TRUE} > 0$ whereas a minus indicates $Y_{TRUE} = 0$.

cases of BU. First, we observe an excess true zero when MU is absent in the environment, and therefore BU transmission cannot occur. Second, we can observe another type of excess true zero when MU is present in the environment, but conditions are not suitable for BU transmission. Third, we can observe a distributional true zero when MU is present in the environment, conditions are suitable for BU transmission, but due to random variation cases do not occur. Fourth, we can observe an excess false zero when MU is present in the environment, conditions are suitable for BU transmission, BU cases do occur, but reporting does not occur. Fifth and lastly, we can observe cases of BU when MU is present in the environment, conditions are suitable for BU transmission, BU cases do occur, and reporting occurs. We can write these event in terms of the latent random variables Z_{MU}, Z_{BU}, Y_{TRUE}, and Z_{REP}, and we can also identify the probability associated with each type of data.

1. **Excess true zero** (associated with MU)

   $Z_{MU} = 0$

   with probability $(1 - p_{MU})$

2. **Excess true zero** (associated with BU)

   $Z_{MU} = 1 \bigcap Z_{BU} = 0$

   with probability $p_{MU} \times (1 - p_{BU})$

3. **Distributional true zero**

   $Z_{MU} = 1 \bigcap Z_{BU} = 1 \bigcap Y_{true} = 0$

   with probability $p_{MU} \times p_{BU} \times f(0)$

4. **Excess false zero**

   $Z_{MU} = 1 \bigcap Z_{BU} = 1 \bigcap Y_{true} > 0 \bigcap Z_{REP} = 0$

   with probability $p_{MU} \times p_{BU} \times (1 - f(0)) \times (1 - p_{REP})$

5. **Non-zero count outcome**

   $Z_{MU} = 1 \bigcap Z_{BU} = 1 \bigcap Y_{true} > 0 \bigcap Z_{REP} = 1$

   with probability $p_{MU} \times p_{BU} \times (1 - f(0)) \times p_{REP}$

The same shorthand notation that applied to the observed data likelihoods in 4.13 and 4.14 also applies here.

Recall that with the standard ZIP model we can calculate the probability that an observed zero is an excess zero (Equation 4.4), which does not distinguish between true and false zeros. With the modified hierarchical ZIP framework, we can now calculate the conditional probability that an observed zero is an excess false zero. As noted above, we can observe zeros in four different cases. Only case 4 represents a false zero. Therefore, the conditional probability of a false zero given that a zero is observed is the probability of a false zero divided by the sum of the probabilities of all of the ways of observing a zero.

$$
\begin{aligned}
&\Pr_i(\text{false zero}|\text{obs zero}) \qquad (4.15)\\
&= \frac{\Pr_i(\text{false zero} \bigcap \text{obs zero})}{\Pr_i(\text{obs zero})}\\
&= \frac{\Pr_i(4)}{\Pr_i(1) + \Pr_i(2) + \Pr_i(3) + \Pr_i(4)}\\
&= \frac{p_{iMU}p_{iBU}(1 - f_i(0))(1 - p_{iREP})}{(1 - p_{iMU}) + p_{iMU}(1 - p_{iBU}) + p_{iMU}p_{iBU}f_i(0) + p_{iMU}p_{iBU}(1 - f_i(0))(1 - p_{iREP})}\\
&= \frac{p_{iMU}p_{iBU}(1 - f_i(0))(1 - p_{iREP})}{1 - p_{iMU}p_{iBU}p_{iREP}(1 - f_i(0))}
\end{aligned}
$$

The sum of of the probabilities of all of the ways to observe a zero in the denominator simplifies to the complement of the probability that a reported case count was observed, as in the observed data likelihood in 4.13.

Implementation of these models and estimation of their parameters is relatively straightforward using Bayesian analysis in WinBUGS. We can obtain estimates of the posterior distribution of Z_{MU}, Z_{BU}, Y_{TRUE}, Z_{REP}, β's, and Pr(false zero|obs zero) based on careful stipulations of the prior distributions of the parameters combined with the model likelihood and the observed data.

4.5 Assessment of Model Fit

There are a variety of methods to assess model fit in Bayesian analysis. These methods can fall into the category of model fit, model comparison, and model checking [96]. In order to compare the fit of models between naive and more informative models, we focus on the deviance information criterion (DIC). As models with increased complexity generally provide a better fit, this Bayesian information criterion adds a penalty for increased model complexity. The DIC estimates the effective number of parameters in a Bayesian hierarchical model in order to appropriately penalize for additional model complexity.

$$\begin{aligned} p_D &= \overline{D}(\theta) - \hat{D}(\theta) \\ &= \mathrm{E}[D(\theta)|y] - D[\mathrm{E}(\theta)|y] \end{aligned} \tag{4.16}$$

$$\begin{aligned} \mathrm{DIC} &= \overline{D}(\theta) + p_D \\ &= \overline{D}(\theta) + \overline{D}(\theta) - \hat{D}(\theta) \\ &= 2\overline{D}(\theta) - \hat{D}(\theta) \\ &= 2\mathrm{E}[D(\theta)|y] - D[\mathrm{E}(\theta)|y] \end{aligned} \tag{4.17}$$

The deviance, $\overline{D}(\theta)$, is an overall measure of model fit and is calculated by twice the negative log likelihood of the model. p_D is the estimate of the effective number of parameters and represents the model's complexity. DIC is the difference in twice the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters. Generally, if two models differ in DIC by more than 3, the model with the smaller DIC provides a better fit [114].

WinBUGS does not provide an estimate of DIC for models such as those presented in this research. While an estimate of the deviance, $\overline{D}(\theta)$, is provided by WinBUGS, $\hat{D}(\theta)$ is not. This can be calculated in R by evaluating the deviance at the posterior mean of the random variables. For this work, we calculate the *observed* deviance (and hence the observed DIC) based on the observed data likelihood presented in equations 4.13 and 4.14.

4.6 Simulations

The objectives of the following simulations are multi-fold. First, we evaluate the performance of the traditional ZIP model, both in the presence and absence of false zeros. Second, we evaluate the performance of the newly proposed hierarchical ZIP models in the presence of false zeros. We evaluate these two objectives considering both a mild and extreme relationship between reporting occurrence and Y_{TRUE}. In the mild (or weak) relationship between Y_{TRUE} and Pr(reporting), fewer cases are less likely to be reported and more likely to be false zeros in Y_{OBS}, the observed outcome. In the extreme (or strong) relationship between Y_{TRUE} and Pr(reporting), fewer cases are **much** less likely to be reported and **much** more likely to be false zeros in Y_{OBS}. For all simulations, we focus on the ability of the models to accurately estimate β_{0CASES} and β_{1CASES}, which are parameters defining the mean case intensity. For the traditional ZIP model, we qualitatively examine the ability of the model to distinguish between true distributional zeros and true excess zeros. For the new hierarchical ZIP models, we quantitatively and qualitatively examine the ability of the model to distinguish between excess true zeros and excess false zeros.

4.6.1 Generating the Data

For the following simulations we consider the scenario in which there is one dichotomous covariate associated with each of Z_{MU}, Z_{BU}, Y_{TRUE}, and Z_{REP}. See Appendix C for details on how parameter values were chosen for the scenario representing a mild relationship between reporting occurrence and Y_{TRUE}. Once those parameters were established, alternative for values β_{0REP} and β_{2REP} were chosen to represent the more extreme relationship. In addition to representing a more extreme association, the parameter values were also selected such that the number of observed false zeros would be the similar between the mild association and the extreme association with Y_{TRUE} and reporting. The parameter values used for simulating the data are in Table 4.5. Figure 4.3 shows the association between Y_{TRUE} and the probability of reporting for both mild and extreme scenarios, with covariate adjustment.

Parameter	β_{0CASES}	β_{1CASES}	β_{0MU}	β_{1MU}	β_{0BU}	β_{1BU}	β_{0REP}	β_{1REP}	β_{2REP}
Mild	1.1	1.1	0.8	0.2	0.4	1.4	0.6	-1.2	0.15
Extreme	1.1	1.1	0.8	0.2	0.4	1.4	-3.0	-1.2	1.0

Table 4.5: Parameter values chosen for simulation. The scenario representing a mild relationship between Y_{TRUE} and reporting occurrence has the exact same parameter values as the scenario representing a more extreme relationship between Y_{TRUE} and reporting occurrence, with the exception of β_{0REP} and β_{2REP}.

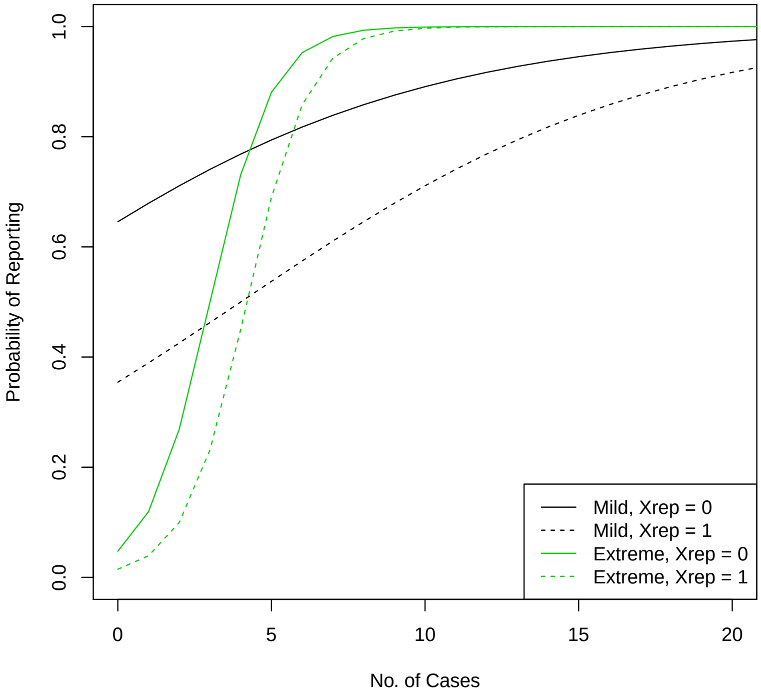


Figure 4.3: Depiction of how the probability of reporting occurrence varies with the covariate X_{REP} and with the number of true cases for both mild and extreme scenarios. This probability is displayed for the range of the observed simulated data.

The data generation steps are as follows:

1. Set all parameter values (β_{0MU}, β_{1MU}, β_{0BU}, β_{1BU}, β_{0CASES}, β_{1CASES}, β_{0REP}, β_{1REP}, and β_{2REP}).

2. Generate covariates X_{MU}, X_{BU}, X_{CASES}, and X_{REP} following Bern(0.5). These

are covariates that are associated with MU presence, BU presence, case intensity, and whether or not reporting occurs. These covariates remain fixed for each iteration.

3. Calculate $Pr(Z_{iMU} = 1)$ and $Pr(Z_{iBU} = 1)$ which remain fixed for each iteration, but depend on the i^{th} observation's covariates. This is the probability of the i^{th} observation having MU present, and the probability of the i^{th} observation having BU present.
$$p_{iMU} = \frac{\exp(\beta_{0MU} + \beta_{1MU}X_{iMU})}{1 + \exp(\beta_{0MU} + \beta_{1MU}X_{iMU})}$$
$$p_{iBU} = \frac{\exp(\beta_{0BU} + \beta_{1BU}X_{iBU})}{1 + \exp(\beta_{0BU} + \beta_{1BU}X_{iBU})}$$

4. Calculate λ_i, the mean rate of cases for the i^{th} observation.
$\lambda_i = \exp(\beta_{0CASES} + \beta_{1CASES}X_{iCASES})$

5. Begin simulation loop. The following quantities will vary with each iteration of the simulation.

   (a) Generate the latent random variables Z_{iMU} and Z_{iBU}. These latent random variables are unobserved, but represent if the i^{th} observation has MU present and if the i^{th} observation has BU present.

   $Z_{iMU} \sim \text{Bern}(p_{iMU})$

   $Z_{iBU} \sim \text{Bern}(p_{iBU} \times Z_{iMU})$

   (b) Generate the true number of observed cases Y_{iTRUE}. This quantity is unobserved.

   $Y_{iTRUE} \sim \text{Poisson}(\lambda_i \times Z_{iBU})$

   (c) Calculate $Pr(Z_{iREP} = 1)$ **for both mild and extreme scenarios**. This is the probability of the i^{th} observation reporting BU cases.
$$p_{iREP} = \frac{\exp(\beta_{0REP} + \beta_{1REP}X_{iREP} + \beta_{2REP}Y_{iTRUE})}{1 + \exp(\beta_{0REP} + \beta_{1REP}X_{iREP} + \beta_{2REP}Y_{iTRUE})}$$

(d) Generate the latent random variable Z_{iREP} **for both mild and extreme scenarios**. This is the unobserved latent random variable indicating if reporting occurred.

$Z_{iREP} \sim \text{Bern}(p_{iREP} \times \text{I}(Y_{iTRUE} > 0))$

6. Calculate the observed number of cases Y_{iOBS} **for both mild and extreme scenarios**.

   $Y_{iOBS} = Y_{iTRUE} \times Z_{iREP}$

We created and saved one master data set with 100 randomly simulated data realizations containing 100 observations. This same master data set is used to evaluate the models enumerated below. The master data set retains Y_{TRUE}, Y_{OBS} (mild), and Y_{OBS} (extreme).

4.6.2 Models Evaluated

All models are run in WinBUGS. Prior distributions only need to be specified for the β parameters. These prior distributions were set to be noninformative by a normal distribution with mean zero and a variance of 10 or 100, depending on the model. Each iteration in the simulation is run with 3 chains, with a burnin of 1000, and a thinning rate of 5. A total of 3000 MCMC samples are retained for which to base posterior inference. See Appendix D for the associated WinBUGS implementation code for each model.

I **The naive traditional ZIP model.**

This model is representative of the current yet naive implementations of the traditional ZIP model. For this model, we include the covariate X_{CASES} in the Poisson model for the case counts. We also include X_{MU} and X_{BU} in the model for the probability of an excess zero. This mimics current practice in that investigators may hypothesize that variables relating to either MU presence or

BU transmission settings may be associated with excess zeros. The latent random variable Z indicates if an observation is an excess zero or not. This model takes as inputs n, Y_{OBS}, X_{CASES}, X_{MU}, and X_{BU}. The prior distributions specified for all β's are N(0, 100).

$$
\begin{aligned}
&Y_i|Z_i \sim \text{Poisson}(\lambda_i(1 - Z_i)) \\
&Z_i \sim \text{Bernoulli}(p_{iEXCESS}) \\
&\log(\lambda_i) = \beta_{0CASES} + \beta_{1CASES}X_{iCASES} \\
&\text{logit}(p_{iEXCESS}) = \beta_0 + \beta_1 X_{iMU} + \beta_2 X_{iBU}
\end{aligned}
\tag{4.18}
$$

We run this model on both Y_{TRUE} and Y_{OBS} in order to assess the effect of false zero inflation on parameter estimation and distinguishing distributional and excess zeros.

II **The almost fully specified hierarchical ZIP model.** This model utilizes the hierarchical zero-inflated model presented in Section 4.4 defined by Model 4.12. Here, we model the distribution of Y_{OBS}, but ignore the underlying distribution of Y_{TRUE}. This model also does not condition on the value of Y_{TRUE}. This implementation includes covariates at each hierarchical level to model λ, p_{MU}, p_{BU}, p_{REP}. This model takes as inputs n, Y_{OBS}, X_{MU}, X_{BU}, X_{CASES}, and X_{REP}.

The prior distributions specified for all β's are N(0, 10).

$$\begin{aligned}
&Y_{iOBS}|Z_{iMU}, Z_{iBU}, Z_{iREP}, X_{iMU}, X_{iBU}, X_{iREP}, X_{iCASES} \sim \text{Poisson}(\lambda_i \times Z_{iREP})\\
&Z_{iREP}|Z_{iMU}, Z_{iBU}, X_{iMU}, X_{iBU}, X_{iREP} \sim \text{Bernoulli}(p_{iREP} \times Z_{iBU})\\
&Z_{iBU}|Z_{iMU}, X_{iMU}, X_{iBU} \sim \text{Bernoulli}(p_{iBU} \times Z_{iMU})\\
&Z_{iMU}|X_{iMU} \sim \text{Bernoulli}(p_{iMU})
\end{aligned}$$

$$\begin{aligned}
&logit(p_{iMU}) = \beta_{0MU} + \beta_{1MU}X_{iMU}\\
&logit(p_{iBU}) = \beta_{0BU} + \beta_{1BU}X_{iBU}\\
&\log(\lambda_i) = \beta_{0CASES} + \beta_{1CASES}X_{iCASES}\\
&logit(p_{iREP}) = \beta_{0REP} + \beta_{1REP}X_{iREP} \qquad (4.19)
\end{aligned}$$

III **The fully specified hierarchical ZIP model.** This model utilizes the hierarchical zero-inflated model defined in Section 4.4 defined by Model 4.11. Here, we explicitly model the distribution of Y_{TRUE}, and condition both Z_{REP} and Y_{OBS} on the value of Y_{TRUE}. This model has covariates at each level to model λ, p_{MU}, p_{BU}, p_{REP}, and expresses the probability of reporting as a function of Y_{TRUE}. This model takes as inputs n, Y_{OBS}, X_{MU}, X_{BU}, X_{CASES}, and X_{REP}. The prior distributions specified for all β's are N(0, 100).

$$
\begin{aligned}
&Y_{iOBS}|Z_{iMU}, Z_{iBU}, Z_{iREP}, X_{iMU}, X_{iBU}, X_{iREP}, X_{iCASES}, Y_{iTRUE}\\
&\quad \sim \text{Poisson}(\lambda_i \times Z_{iREP})\\
&Z_{iREP}|Z_{iMU}, Z_{iBU}, X_{iMU}, X_{iBU}, X_{iREP}, X_{iCASES}, Y_{iTRUE}\\
&\quad \sim \text{Bernoulli}(p_{iREP} \times \text{I}(Y_{iTRUE} > 0))\\
&Y_{iTRUE}|Z_{iMU}, Z_{iBU}, X_{iMU}, X_{iBU}, X_{iCASES} \sim \text{Poisson}(\lambda_i \times Z_{iBU})\\
&Z_{iBU}|Z_{iMU}, X_{iMU}, X_{iBU} \sim \text{Bernoulli}(p_{iBU} \times Z_{iMU})\\
&Z_{iMU}|X_{iMU} \sim \text{Bernoulli}(p_{iMU})
\end{aligned}
$$

$$
\begin{aligned}
&logit(p_{iMU}) = \beta_{0MU} + \beta_{1MU}X_{iMU}\\
&logit(p_{iBU}) = \beta_{0BU} + \beta_{1BU}X_{iBU}\\
&\log(\lambda_i) = \beta_{0CASES} + \beta_{1CASES}X_{iCASES}\\
&logit(p_{iREP}) = \beta_{0REP} + \beta_{1REP}X_{iREP} + \beta_{2REP}Y_{iTRUE}
\end{aligned}
\tag{4.20}
$$

4.6.3 Simulation Results

Table 4.6 provides a summary of the simulated data. Data realizations are summarized in terms of the median value observed and the range across the 100 simulated data sets. The three columns correspond to the true outcome, the observed outcome with a mild false zero relationship with reporting, and the observed outcome with an extreme false zero relationship with reporting. The first five rows correspond to the five types of observed data described in Section 4.4. The summaries of first three observation types (correspondingly, the first three rows in the table) are the same between Y_{TRUE}, Y_{OBS} (mild), and Y_{OBS} (extreme) by design. The remaining rows differ between the outcomes. In the unobserved true outcome, we do not observe

Summary	Y_{TRUE}		Y_{OBS} (mild)		Y_{OBS} (extreme)	
	Med.	*(Range)*	*Med.*	*(Range)*	*Med.*	*(Range)*
No. excess true zeros (MU)	29	(19, 40)	29	(19, 40)	29	(19, 40)
No. excess true zeros (BU)	20	(8, 31)	20	(8, 31)	20	(8, 31)
No. distributional true zeros	1	(0, 4)	1	(0, 4)	1	(0, 4)
No. excess false zeros	0	(0, 0)	15	(6, 21)	15	(10, 23)
No. non-zero observed count outcomes	50	(38, 64)	34	(26, 46)	33	(25, 46)
Total no. observed zeros	50	(36, 52)	66	(54, 74)	67	(54, 75)
Conditional proportion of false zeros	0	(0, 0)	0.23	(0.10, 0.37)	0.24	(0.14, 0.34)
Non-zero count data	5	(1, 20)	6	(1, 20)	8	(1, 20)

Table 4.6: Summary of simulated data with two different false zero generation processes (mild and extreme).

any false zeros. In the observed outcomes, there is a similar number of false zeros and conditional proportion of false zeros. The median of the non-zero observed count data is higher in the data realizations created by more extreme false zero generation because in this scenario lower case counts (Y_{TRUE}) are more likely to be false zeros in Y_{OBS}.

The objectives of the simulations are (1) determine each model's ability to estimate true parameter values, (2) assess the traditional ZIP model's ability to correctly distinguish distributional zeros, and (3) assess the hierarchical ZIP model's ability to correctly distinguish false zeros.

Model I results (assessing the effect of false zero inflation on parameter estimates)

Table 4.7 provides results from the traditional ZIP model using the unobserved true outcome and two versions of an observed outcome subject to false zero inflation. For the Poisson part of the model, we know the true values of β_{0CASES} and β_{1CASES} based on values we used to generate the simulated data. Using the true outcome, β_{0CASES} is slightly underestimated and β_{1CASES} is slightly overestimated. When there is a mild relationship between the true outcome and reporting occurrence in the observed outcome, the estimates of β_{0CASES} and β_{1CASES} are similar to those

	Model Parameters											
	Y_{TRUE}				Y_{OBS} (mild)				Y_{OBS} (extreme)			
	A	B	C	D	A	B	C	D	A	B	C	D
β_{0CASES} (1.1)	1.05	0.13	0.94	0	1.08	0.17	0.96	0	1.35	0.17	0.70	0
β_{1CASES} (1.1)	1.14	0.15	0.93	0	1.14	0.19	0.94	0	0.87	0.19	0.73	0
β_{0P}	0.36	0.36	N/A	0.80	0.93	0.36	N/A	0.23	1.03	0.39	N/A	0.20
β_{1p} (MU)	-0.11	0.50	N/A	0.94	-0.09	0.47	N/A	0.95	-0.21	0.48	N/A	0.94
β_{2p} (BU)	-0.80	0.46	N/A	0.59	-0.58	0.45	N/A	0.76	-0.52	0.43	N/A	0.79
	Model Performance											
	Y_{TRUE}				Y_{OBS} (mild)				Y_{OBS} (extreme)			
	$\overline{x}$	sd	L	U	$\overline{x}$	sd	L	U	$\overline{x}$	sd	L	U
DIC	353.99	23.24	297.3	429.58	286.9	26.12	234.88	349.87	284.14	26.41	221.55	367.52
pD	5.05	0.05	4.92	5.16	5.03	0.05	4.91	5.14	5.06	0.08	4.9	5.66
No. $\widehat{R} > 1.1$	15.1	3.97	7	27	13.82	5.75	3	27	12.73	18	1	184

Table 4.7: Simulation results on the traditional ZIP model (4.18) comparing the effect of knowing Y_{TRUE} vs Y_{OBS}. A=mean over all simulations of parameter point estimate, B=standard deviation over all simulations of parameter point estimate, C=coverage of parameter over all simulations, D=percent of simulations in which the credible set contained 0 (no effect), $\overline{x}$=the mean of the statistics over all simulations, sd=the standard deviation of the statistic over all simulation, L=the lowest observed statistic over all simulations (lower), U=the highest observed statistic over all simulations (upper).

for the true outcome. However, when there is an extreme relationship between the true outcome and reporting occurrence in the observed outcome, β_{0CASES} is more severely overestimated and β_{1CASES} is more severely underestimated. This means that the model is estimating that the baseline mean rate of cases is higher than the true baseline rate of cases (baseline meaning when X_{CASES}=0). Moreover, the effect of X_{CASES} is underestimated. The coverage for these parameters approximately reaches the 95% nominal level for the true outcome and the observed outcome with a mild relationship between Y_{TRUE} and reporting occurrence. However, the coverage for these parameters is poor (70% and 73%, respectively) for the observed outcome with an extreme relationship between Y_{TRUE} and reporting occurrence. Lastly, all credible sets for these two parameters were quite specific for all three outcomes, and none contained the null value of zero demonstrating no effect.

The true values of β_{0P}, β_{1p}, β_{2p} are not known because we did not use this model to simulate our data. Consequently, we do not calculate coverage for these parameters.

Nevertheless, since the covariates X_{MU} and X_{BU} are actually associated with the probability of being a zero in the data generation, it is reasonable to expect that estimates for these covariates would be non-zero. However, model results show that many credible sets were wide and contained the null value of zero, indicating no significant effect of X_{MU} or X_{BU} on the probability of an excess zero.

On average over the simulations, it actually appears that the traditional ZIP model fits the data best when the outcome is subject to an extreme false zero generation mechanism, as evaluated by the observed DIC. This may be counterintuitive. However, this outcome has more zeros and less low values (1, 2, etc.) in the observed count outcomes than the other two models. Therefore, this data may have a clearer separation in the Poisson model and the zero model, lending it to a better fit. The estimated number of effective parameters is similar between the the three outcomes, as it should be.

We also monitored each statistical node for convergence by Gelman and Rubin's $\widehat{R}$ statistic. The number of $\widehat{R} > 1.1$ indicates the number of nodes monitored in which $\widehat{R} > 1.1$ out of 419 nodes monitored for Model 4.18. Each outcome has some chains in which $\widehat{R} > 1.1$, indicating that the chains might not have converged yet. Although on average this occurs less frequently when using the outcome subject to false zero inflation with the extreme relationship between Y_{TRUE} and reporting, there is more variability in the number of nodes with $\widehat{R} > 1.1$ and wider range compared to the other outcomes. This is an indication that chains may need to be run longer to achieve more stable convergence.

Model I results (distinguishing types of zeros)

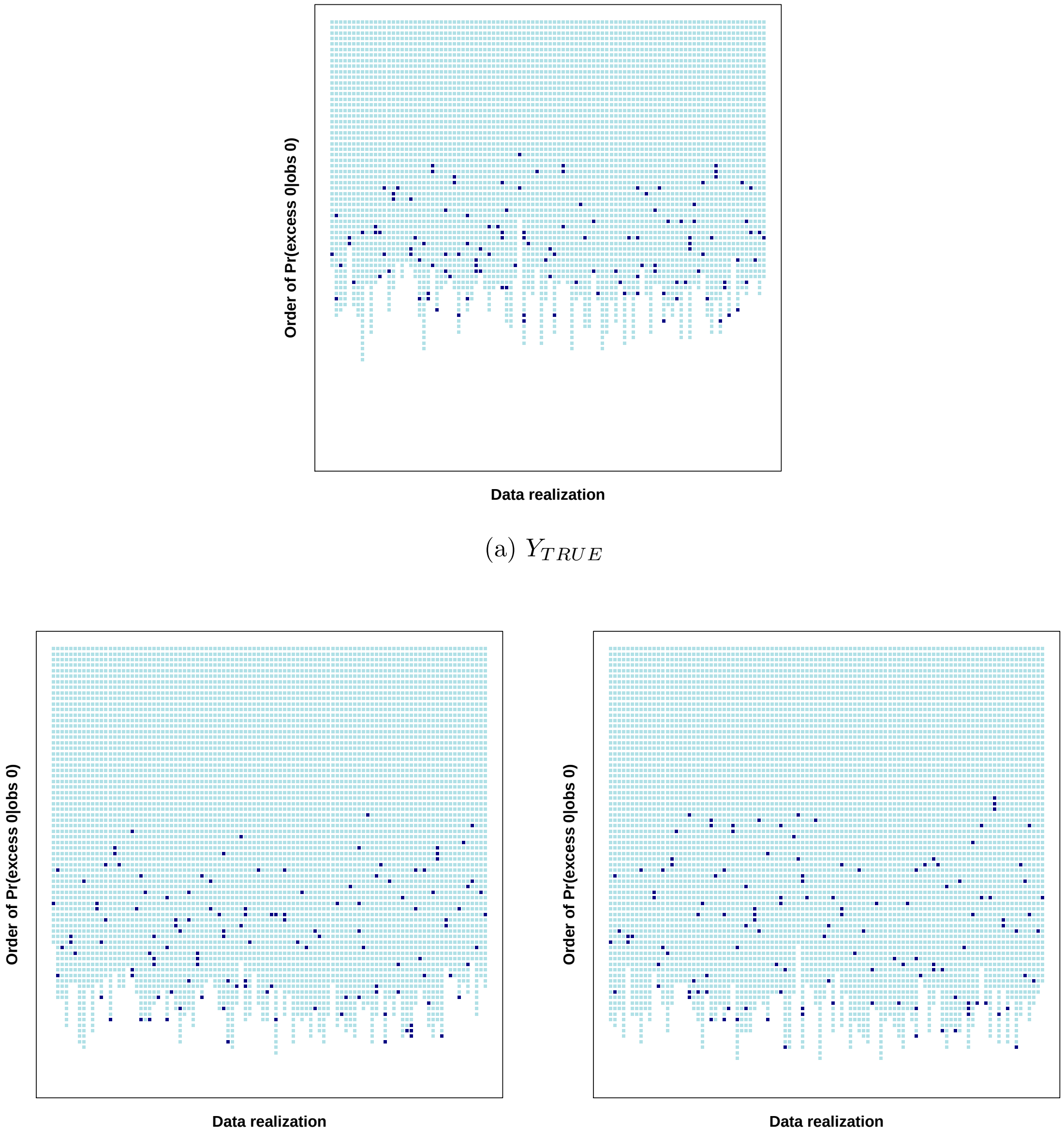
We qualitatively assessed the ability traditional ZIP model to distinguish between excess and distributional true zeros. Only 75% of the data realizations contain distributional zeros. For each 'zero observation' we estimate the Pr(excess zero|obs zero) as given in equation 4.4. For **distributional** zeros, we would hope for the conditional probability of an excess zero given that a zero was observed to be **low**. Each data realization takes on 7-8 unique values of Pr(excess zero|obs zero). This is because this probability is calculated based on three dichotomous covariates, and $2^3 = 8$ possible covariate combinations.

Figure 4.4 displays a visualization of this data. The length of the vertical lines is determined by the number of zeros in the data realization. Plots (b) and (c) show results from data realizations subject to false zero inflation, so these data realizations have more zeros present in the observed data. Within one data realization, the Pr(excess zero|obs zero) is ordered from highest to lowest. This order of Pr(excess zero|obs zero) is plotted on the y-axis such that observations at the top represent the largest conditional probability. The colors of the plot represent the truth behind the simulated data, which is not a model input. Light blue points represent excess zeros, and dark blue points represent distributional zeros. In the presence of ties with multiple observations having the same order of Pr(excess zero|obs zero), the dark blue distributional zeros are plotted on the bottom of that sequence. The x-axis is ordered by data realization and therefore has no meaningful ordering.

In Figure 4.4, it does appear that distributional zeros have a lower ordered conditional probability of being an excess zero. Nevertheless, it is noteworthy that the distributional zeros do not consistently have the lowest order of Pr(excess zero|obs zero). Plot (c) perhaps demonstrates a slightly wider spread in the order of the probability of excess zeros among the distributional zeros compared to plot (b).

Model II and III results (parameter estimates)

Tables 4.8 and 4.9 show results from the almost fully specified and the fully specified hierarchical ZIP models. Table 4.8 shows results when the outcome is subject to false zero inflation with a mild relationship between Y_{TRUE} and reporting. Table 4.9



(a) Y_{TRUE}

(b) Y_{OBS} (mild)

(c) Y_{OBS} (extreme)

Figure 4.4: This figure shows the distribution of Pr(excess zero|obs zero) in the excess zeros (light blue) and the distributional zeros (navy blue) with respect to their ranked order.

	Model Parameters (mild)							
	Model 4.19				Model 4.20			
	A	B	C	D	A	B	C	D
β_{0CASES} (1.1)	1.09	0.16	0.96	0.00	1.21	0.12	0.92	0.00
β_{1CASES} (1.1)	1.13	0.18	0.95	0.00	1.00	0.15	0.93	0.00
β_{0MU} (0.8)	2.02	0.69	0.99	0.94	5.46	0.12	0.00	0.00
β_{1MU} (0.2)	1.02	0.79	0.99	0.98	1.52	0.08	1.00	1.00
β_{0BU} (0.4)	1.26	0.93	0.99	0.99	5.16	0.20	0.01	0.01
β_{1BU} (1.4)	1.65	0.62	1.00	0.84	1.67	0.20	1.00	1.00
β_{0REP} (0.6)	1.97	0.52	0.98	0.92	-0.89	0.69	0.36	0.73
β_{1REP} (-1.2)	-0.48	0.97	0.99	0.90	-0.59	0.47	0.69	0.77
β_{2REP} (0.15)	N/A	N/A	N/A	N/A	0.07	0.09	0.84	0.87
	Model Performance (mild)							
	Model 4.19				Model 4.20			
	$\overline{x}$	sd	L	U	$\overline{x}$	sd	L	U
DIC	284.23	26.32	233.65	351.41	284.1	26.35	236.95	355.27
pD	4.37	0.74	2.23	5.65	3.89	3.42	-2.54	16.72
No. $\widehat{R} > 1.1$	31.12	21.84	4	141	14.14	26.21	2	233

Table 4.8: Simulation results comparing Model 4.19 (the almost fully specified hierarchical ZIP model) to Model 4.20 (the fully specified hierarchical ZIP model conditioning on $Y_{TRUE} > 0$). The outcome here is the observed Y subject to false zero inflation with a mild relationship between Y_{TRUE} and reporting. A=mean over all simulations of parameter point estimate, B=standard deviation over all simulations of parameter point estimate, C=coverage of parameter over all simulations, D=percent of simulations in which the credible set contained 0 (no effect), $\overline{x}$=the mean of the statistics over all simulations, sd=the standard deviation of the statistic over all simulation, L=the lowest observed statistic over all simulations (lower), U=the highest observed statistic over all simulations (upper).

shows results when the outcome is subject to false zero inflation with an extreme relationship between Y_{TRUE} and reporting.

In comparing the almost fully specified and the fully specified hierarchical ZIP models where the outcome is subject to false zero inflation with a mild relationship between Y_{TRUE} and reporting (Table 4.8), it appears that the almost fully specified hierarchical ZIP model performs better. The estimates of β_{0CASES} and β_{1CASES} exhibit greater bias and slightly lower coverage in the fully specified hierarchical ZIP model.

Similar to model evaluation with the traditional ZIP model, we cannot compare parameter estimates in MU, BU, and reporting from the almost fully specified hierarchical ZIP model to the true values because the data were generated by the fully specified hierarchical ZIP model. However, when using Model 4.19 the majority of the credible sets do contain both the true parameter value as well as the null value of zero. We can fairly compare the true parameter values to the models' point estimates with the fully specified hierarchical ZIP model. This model exhibits extreme bias in the parameter estimates for MU and BU effects. The intercepts have virtually no coverage and all credible sets contain the null value of zero. The covariate effects have 100% coverage and no credible sets contain the null value of zero.

With regards to the reporting covariates, the fully specified hierarchical ZIP model incorrectly specifies the direction of the intercept whereas the almost fully specified correctly specifies the direction of the intercept. The fully specified hierarchical ZIP model exhibits slight bias but reasonable estimates of β_{2REP}, the parameter defining the relationship between Y_{TRUE} and reporting. The credible sets for this parameter do have good coverage; however, 87% also contain zero indicating no significant relationship between Y_{TRUE} and reporting.

The DIC estimates show similar model fit between the almost fully specified and the fully specified hierarchical ZIP models where the outcome is subject to false zero

	Model Parameters (extreme)							
	Model 4.19				Model 4.20			
	A	B	C	D	A	B	C	D
β_{0CASES} (1.1)	1.35	0.17	0.66	0.01	1.45	0.10	0.35	0.00
β_{1CASES} (1.1)	0.87	0.19	0.71	0.00	0.77	0.12	0.43	0.00
β_{0MU} (0.8)	1.72	0.69	1.00	0.98	5.46	0.08	0.00	0.00
β_{1MU} (0.2)	1.18	0.82	0.98	0.98	1.52	0.08	1.00	1.00
β_{0BU} (0.4)	1.33	0.89	0.99	1.00	5.14	0.15	0.00	0.00
β_{1BU} (1.4)	1.57	0.61	1.00	0.88	1.73	0.19	1.00	1.00
β_{0REP} (-3.0)	1.96	0.49	0.99	0.93	-2.93	0.96	1.00	0.04
β_{1REP} (-1.2)	-0.28	0.91	1.00	0.96	-0.46	0.54	0.75	0.88
β_{2REP} (1.0)	N/A	N/A	N/A	N/A	0.33	0.12	0.06	0.11
	Model Performance (extreme)							
	Model 4.19				Model 4.20			
	$\overline{x}$	sd	L	U	$\overline{x}$	sd	L	U
DIC	281.72	26.62	215.59	365.51	273.08	26.54	215.37	351.46
pD	4.08	0.96	1.51	5.76	13.26	3.86	5.37	23.65
No. $\widehat{R} > 1.1$	29.61	47.09	4.00	405.00	12.40	9.00	1.00	66.00

Table 4.9: Simulation results comparing Model 4.19 (the almost fully specified hierarchical ZIP model) to Model 4.20 (the fully specified hierarchical ZIP model conditioning on $Y_{TRUE} > 0$). The outcome here is the observed Y subject to false zero inflation with an extreme relationship between Y_{TRUE} and reporting. A=mean over all simulations of parameter point estimate, B=standard deviation over all simulations of parameter point estimate, C=coverage of parameter over all simulations, D=percent of simulations in which the credible set contained 0 (no effect), $\overline{x}$=the mean of the statistics over all simulations, sd=the standard deviation of the statistic over all simulation, L=the lowest observed statistic over all simulations (lower), U=the highest observed statistic over all simulations (upper).

inflation with a mild relationship between Y_{TRUE} and reporting. However, in the fully specified hierarchical ZIP model the estimate of the number of effective parameters is low and attains negative values. Though counterintuitive, it is possible for this to occur in practice when the posterior distribution for a parameter is asymmetric [114]. The number of $\widehat{R} > 1.1$ indicates the number of nodes monitored in which $\widehat{R} > 1.1$ out of 910 nodes monitored for Model 4.19 or 1111 nodes for Model 4.20. Both models have statistical nodes in which convergence was not achieved as assessed by Gelman and Rubin's $\widehat{R}$ statistic.

In comparing the almost fully specified and the fully specified hierarchical ZIP models where the outcome is subject to false zero inflation with a extreme relationship between Y_{TRUE} and reporting (Table 4.9), it appears that the almost fully specified hierarchical ZIP model performs better again. Both models exhibit bias in the estimates of β_{0CASES} and β_{1CASES}, but this bias is greater in the fully specified hierarchical ZIP model. Moreover, though coverage of these true parameter values is poor for both models, coverage is worse in the fully specified hierarchical ZIP model.

In examining parameter estimates related to MU and BU in the almost fully specified hierarchical ZIP model, again the majority of the credible sets do contain both the true parameter value as well as the null value of zero. For the fully specified hierarchical ZIP model, we can fairly compare the true parameter values to the models' point estimates. Again, this model exhibits extreme bias in the parameter estimates for MU and BU effects. The intercepts have poor coverage and all credible sets contain the null value of zero, and the covariate effects have 100% coverage and again no credible sets contain the null value of zero.

With regards to the reporting covariates, this time the almost fully specified hierarchical ZIP model incorrectly specifies the direction of the intercept whereas the fully specified hierarchical ZIP model correctly specifies the direction of the intercept with minimal bias and 100% coverage. In this more extreme scenario, the fully specified hierarchical ZIP model exhibits greater bias in β_{2REP} with poor coverage (6%).

For this more extreme scenario, the DIC estimates show slightly better model fit for the fully specified hierarchical ZIP models. Moreover, the estimate of the number of effective parameters no longer attains negative values. Even though model 4.20 estimates only one additional parameter compared to model 4.19, the estimate of the number of effective parameters differs by quite a bit (4 compared to 13). The number of $\widehat{R} > 1.1$ indicates the number of nodes monitored in which $\widehat{R} > 1.1$ out of 910 nodes monitored for Model 4.19 or 1111 nodes for Model 4.20. Both models have statistical nodes in which convergence was not achieved as assessed by Gelman and Rubin's $\widehat{R}$ statistic, but this time it occurs more frequently in the almost fully specified hierarchical ZIP model.

Model II and III results (identifying false zeros)

In addition to evaluating the ability of the models to accurately estimate parameter values, we are also very interested in the models' capability to distinguish between true and false zeros. Towards this end, we summarized the distribution of the estimate of Pr(false zero|obs zero) for all zero observations within each data realization. Figure 4.5 displays the distribution of Pr(false zero|obs zero) within six data realizations for the 'mild' scenario under the almost fully specified model. From data realization to data realization, the range of Pr(false zero|obs 0) varies quite a bit. One range observed was 0.05 to 0.3, and another range observed was 0.1 to 0.7. When considering the distribution of Pr(false zero|obs zero) for the 'extreme' scenario, again we observe similar variability from data realization to data realization within approximately the same range. For either the mild or extreme scenario in the fully specified hierarchical ZIP model, the Pr(false zero|obs zero) falls within a much tighter range that is also much closer to one. Across all data realizations, the number of unique values for the Pr(false zero|obs 0) ranges from 14-16. This is because this probability is calculated based on 4 dichotomous covariates, and $2^4 = 16$ possible covariate combinations.

We utilized the unknown truth from the data realizations to assess model performance. Within each data realization, we compared the distribution of Pr(false zero|obs zero) among false zero observations to true zero observations by Wilcoxon rank-sum tests because of the sample size, the sometimes skewed distribution of the probability, and the fact on that this quantity takes on 16 discrete values. Figure 4.6 displays the distribution of the difference in medians (median among false zero observations minus median among true zero observations) and the distribution of the

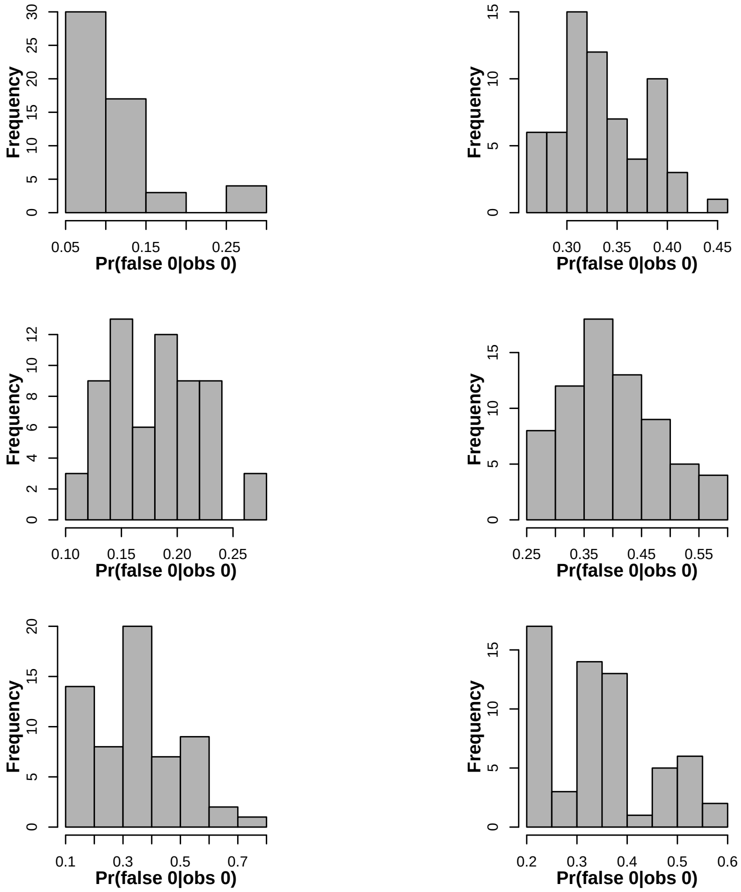


Figure 4.5: Example of the distribution of Pr(false zero|obs zero) estimated for each zero observation within six different data realizations for the almost fully specified hierarchical ZIP model 4.19 in the ‘mild’ scenario.

Model	Scenario	Mean Diff	Range	% Diff > 0	$\%p < 0.05$	$\%p < 0.05 \vert \text{Diff} > 0$
4.19	Mild	0.03	(-0.04, 0.16)	76	23	23
	Extreme	-0.01	(-0.09, 0.10)	39	8	5
4.20	Mild	0.02	(-0.04, 0.14)	65	12	10
	Extreme	-5.1×10^{-5}	(-0.06, 0.03)	55	23	14

Table 4.10: Results comparing the distribution of Pr(false zero|obs zero) among true zeros and false zeros for both mild and extreme false zero generation scenarios under the almost fully specified (4.19) and fully specified models (4.20).

p-value from the Wilcoxon rank-sum test for the almost fully specified hierarchical model for both mild and extreme false zero scenarios, Figure 4.7 does the same for the fully specified hierarchical model, and Table 4.10 presents the numeric results.

These results suggest that Model 4.19 (the almost fully specified hierarchical ZIP model) has promise to detect false zeros. In 76% of data realizations the difference in median Pr(false zero|obs zero) among the false zero observations and the true zero observations was greater than zero. In 23% of data realizations the p-value from the Wilcoxon rank-sum test is less than 0.05, indicating a significant difference in the distribution of Pr(false zero|obs zero) among the false zero observations and the true zero observations. For the same model, model performance is worse in the scenario representing the more extreme relationship between Y_{TRUE} and reporting compared to the scenario representing the mild relationship between Y_{TRUE} and reporting. In the more extreme scenario, fewer differences in medians were greater than zero, and fewer Wilcoxon p-values were less than 0.05. Moreover, three of the significant Wilcoxon p-values corresponded to situations where the difference in medians was less than zero, suggesting the distribution of Pr(false zero|obs zero) was lower among true zero observations compared to false zero observations.

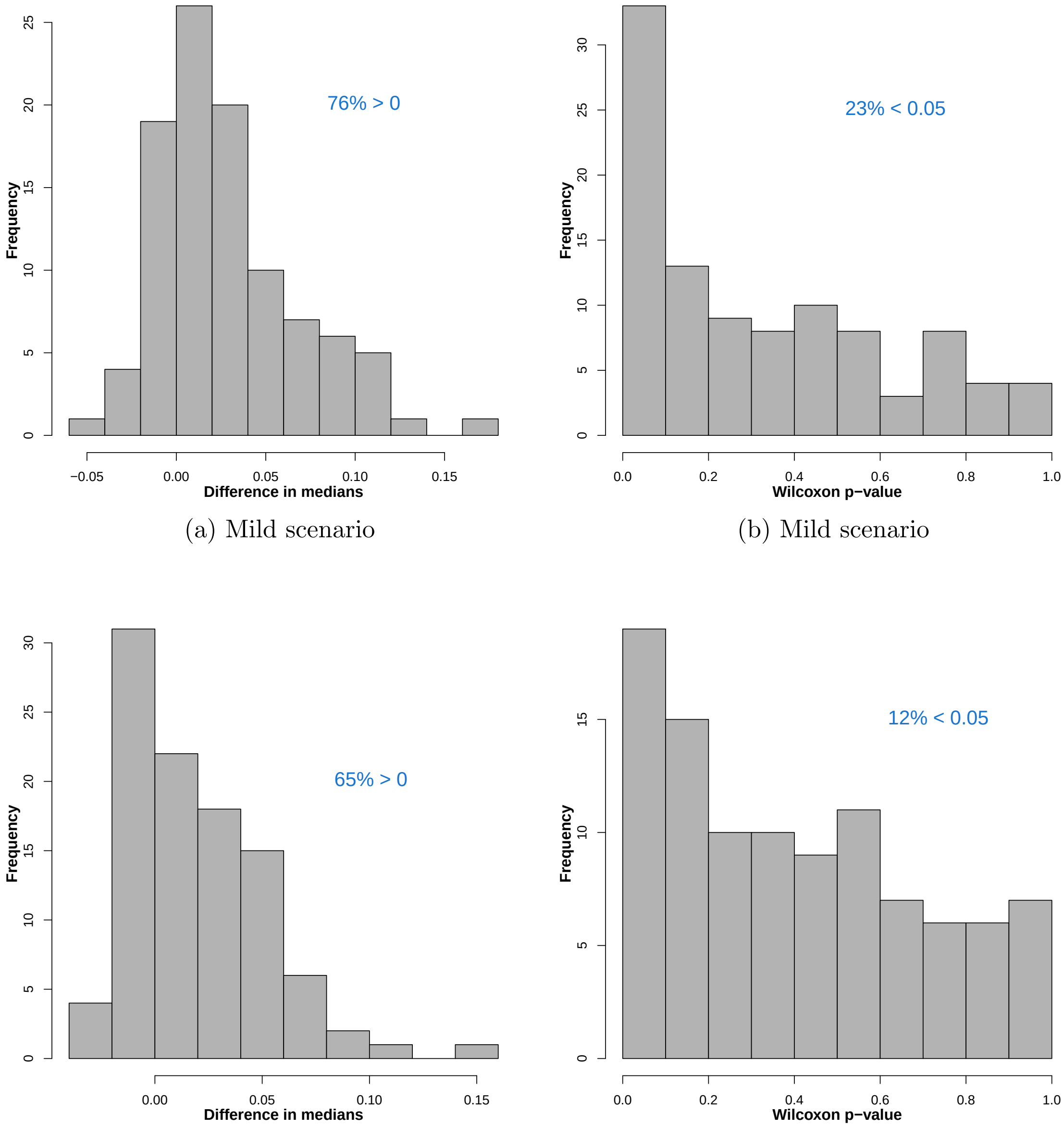
Figure 4.7 displays the same results for the fully specified hierarchical ZIP model (4.20). For this model, the distribution of the difference in medians is more symmetric for the mild scenario, and more skewed left for the extreme scenario. In the mild scenario we observe less median differences greater than zero and fewer significant

Wilxocon p-values compared to Model 4.19. In the extreme scenario, only 14 out of 23 significant p-values correspond to the desired directionality of the test. These results indicate that the almost fully specified hierarchical ZIP model is most promising to distinguish false zeros from true zeros in the mild false zero generation scenario.

The difference in medians and the results of the Wilcoxon rank-sum test demonstrate promise in Model 4.19 to detect false zeros with the mild false zero scenario; however, accurately identifying the false zeros is a different story. Figure 4.8 displays a visualization of this data. Within one data realization, the estimate of Pr(false zero|obs zero) is ordered from highest to lowest. This order is plotted on the y-axis such that observations at the top represent the largest value observed. Blue points represent true zeros, and red points represent false zeros. In the presence of ties with multiple observations having the same order of Pr(false zero|obs zero), the red false zeros were plotted on top of that sequence.

The x-axis represents a different ordering based on a summary value from that data realization. In the Data Realization plot, the x-axis is randomly ordered as the ordering is given by the sequence of the data realizations. In the subsequent plots the x-axis is ordered by meaningful values. For example, in the Wilcoxon p-value plot, the x-axis is ordered by the p-value from the Wilcoxon-rank sum test assessing if the distribution of Pr(false zero|obs zero) among true zeros is the same as the distribution of Pr(false zero|obs zero) among false zeros.

The Data Realization plot and the Number of Zeros plot show no pattern in the distribution of the order Pr(false zero|obs zero) among the false zeros. However, in the remaining four plots, slight clumping of the false zeros is visible. In the Wilcoxon p-value plot, there appears to be red clumping of false zeros in the upper left hand corner, such that data realizations with more significant p-values tend to have a lot of highly ordered Pr(false zero|obs zero) among the false zeros. In the other plots, there appears to be slight clumping of red false zeros in the upper right hand corner



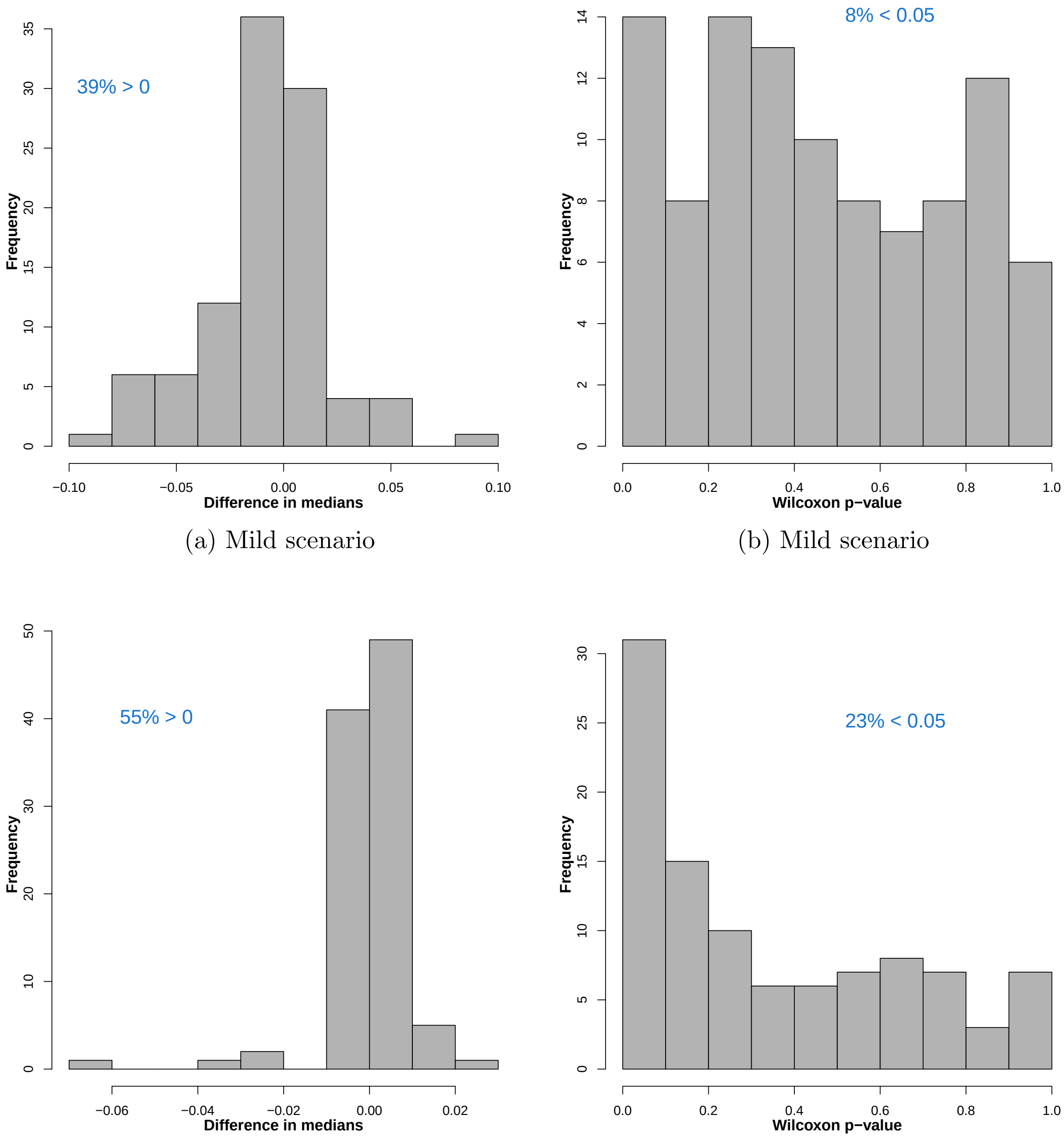
(a) Mild scenario

(b) Mild scenario

(c) Extreme scenario

(d) Extreme scenario

Figure 4.6: Wilcoxon rank-sum test results comparing the distribution of Pr(false zero|obs zero) in the false zeros compared to the distribution of Pr(false zero|obs zero) in the true zeros from 100 data realizations with the almost fully specified hierarchical ZIP model (4.19).



(a) Mild scenario

(b) Mild scenario

(c) Extreme scenario

(d) Extreme scenario

Figure 4.7: Wilcoxon rank-sum test results comparing the distribution of Pr(false zero|obs zero) in the false zeros compared to the distribution of Pr(false zero|obs zero) in the true zeros from 100 data realizations with the fully specified hierarchical ZIP model (4.20).

of the plots such that larger values for the x-axis quantity tend to have more false zeros with Pr(false zero|obs zero) near the top of the order.

The data visualization in Figure 4.8 shows that false zeros do not consistently take on the highest Pr(false zero|obs zero). Rather, a false zero(s) can be observed in each set of covariate combinations, and so within each discrete bin of Pr(false zero|obs zero) there may be both false zeros and true zeros. More work needs to be explored on accurately identifying false zero observations.

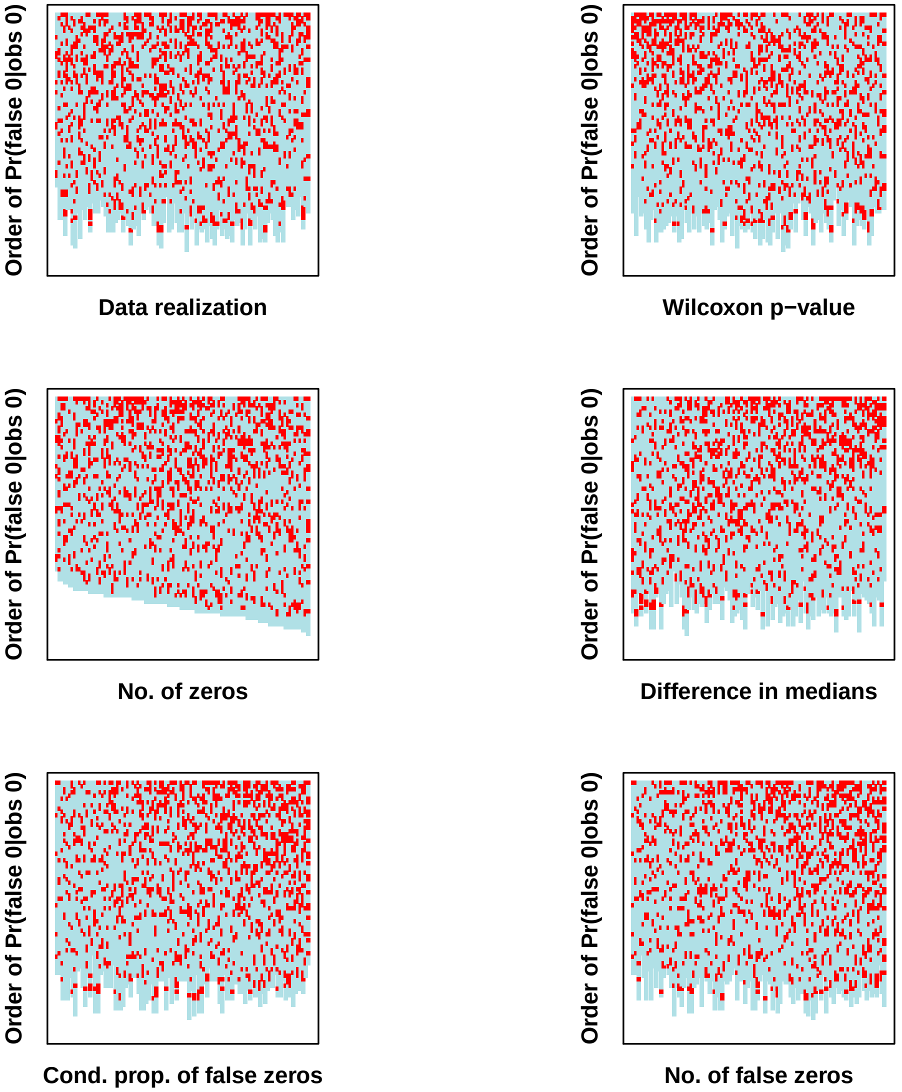


Figure 4.8: The distribution of Pr(false zero|obs zero) in the true zeros (blue) and the false zeros (red) with respect to their ranked order for the mild false zero generation scenario under the almost fully specified hierarchical ZIP model (4.19).

4.7 Back to the Motivating Data

4.7.1 Details on the Available Data

Annual case report summaries of Buruli ulcer for 2008 were obtained from the National Buruli Ulcer Control Programme of Ghana.

We obtained population data from the USAID Population Explorer[1]. The USAID Population Explorer was developed for the International Agency for International Development's Famine Early Warning System Network by Kimetrica. Its primary data source is Landscan, a high-resolution population dataset produced by Oak Ridge National Laboratory, under a US Department of Defense contract. The website consists of a world map to which one can add pre-loaded administrative boundaries. District administrative boundaries from the year 2000 are available for Ghana. Population estimates for the additional split districts that existed in Ghana in 2007 can be obtained by using the free-hand draw shape tool to represent new the newly formed districts. Then the website can estimate the population for that free-hand shape representing a district that did not exist in 2000, the year corresponding to the mapped districts on the website. So for districts that were split after 2000, we can still obtain rough population estimates using this tool. The USAID Population Explorer also provides three other variables for each district: total area (km^2), population density (population/total area), and the most populated km^2.

There are two districts in particular in which there is reason to doubt the population estimates. These districts are Adansi North and Obuasi Municipality within the Ashanti region. In 2000, only Adansi North district existed; Obuasi Municipality was carved out of Adansi North in 2003. Of Obuasi Municipality, www.ghanadistricts.com states, "The population of the Municipality is estimated at 205,000 using the 2000 Housing and Population Census as a base and applying a 4% annual growth rate."

[1] http://www.populationexplorer.com/

	Adansi North	Obuasi Municipal
Population Explorer 2007 estimate	235,103	25,052
www.ghanadistricts.com 2000 estimate	92,834	205,000

Table 4.11: Discrepancy in population estimates by two sources.

Of Adansi North, the same website states, "The district population stands at about 92,834 people as at the year 2000 when the last census was conducted, with a growth rate of 2.6% per annum." Moreover, "Migration is a major challenge in the district. This is because the main occupation in the district is agriculture and therefore those who are not interested in agriculture, especially the youth, migrate to nearby Obuasi Municipality where gold is being mined to seek for employment." The population estimates of these districts provided by the website is not consistent with those provided by the Population Explorer (Table 4.11).

The consensus reached by the Emory BU group is that district lines have not been drawn accurately. The district of Obuasi Municipality should contain the township of Obuasi and have the higher population. However, it does not. To move forward we must assign the geographic area called Obuasi Municipality the lower population estimate because that geographic region does appear to be sparsely populated. Note that this does bring into question reporting issues: If district lines are not drawn accurately, where would district cases actually get reported to?

4.7.2 Data We Would Like to Obtain

We would like to obtain covariates related to the presence of MU, BU, or reporting in order to implement the real data analysis. Table 4.12 shows a partial listing of such covariates. Currently, we have none of the environmental covariates. Although we used similar covariates in Chapter 3, the remotely sensed surfaces that we obtained did not cover the entire six regions of Ghana that reported BU cases. Rather, the surfaces only covered the extent of the sites tested for *M. ulcerans*. For human activity, we have the population density provided by the Population Explorer. For reporting covariates, the distance of each district's centroid to major cities would be easy to calculate in a GIS such as ArcMap.

4.8 Conclusion/Discussion

We assessed the performance of the the traditional ZIP model in the presence of false zeros. We showed that in some scenarios parameters corresponding to the rate of cases can be biased. We qualitatively showed that the traditional ZIP model adequately distinguishes between distributional and excess zeros.

We proposed a hierarchical ZIP model with the capacity to estimate the conditional probability that an observed zero was a false zero. Due to computational challenges with latent random variables, we evaluated two versions of this model that we called 'almost fully specified' and 'fully specified'. Although the 'fully specified' model is a more accurate representation of our data, the 'almost fully specified' model tends to perform better. In most scenarios, the parameters corresponding to the rate of cases can be biased, with more extreme bias when there is a more extreme association between the underlying unknown true outcome and the probability of reporting. Even though the model can estimate the conditional probability that an observed zero is a false zero, this only showed promise in actually distinguishing between false and true zeros for a mild association between the underlying unknown true outcome and the probability of reporting. We qualitatively showed that the model needs further development to better make prediction-specific results.

4.9 Future Directions

There are many future directions for this model. One, we would like to perform simulations with data subject to false zeros incorporating a spatial random effect as

	X_{MU}	X_{BU}	X_{REP}
Environmental			
Land use/land cover (urban, forested, cropland, indicators of deforestation - this can be an indicator of habitat, economy, or accessibility)	√	√	√
NDVI vegetation index. This could indicate habitat or accessibility.	√		√
Elevation (might be related to MU habitat, and can affect accessibility.)	√		√
Wetness index	√		√
Hydrology (can affect habitat, transmission, and water barriers could affect accessibility)	√	√	√
Rainfall (flooding might influence MU, transmission, and reporting)	√	√	√
Temperature (climate might affect BU)	√		
Human Activity			
Primary economy of district (agriculture, gold mining)	√	√	
Types of crops in a district (rice, maize, casava)	√	√	
Population density (less populated areas may not get BU as it is a rare event, but also less populated areas may not have good reporting)		√	√
Reporting			
Indicators of urban/rural (ease of access)			√
Amount of ‘paved’ roads (ease of access)			√
Number of health clinics (may facilitate reporting)			√
Distance to either Accra or Kumasi			√

Table 4.12: Covariates that may be associated with the presence of MU, BU transmission, or reporting. It is feasible for some covariates to be associated with more than one of these categories, as indicated by the check marks.

presented by Agarwal et al. [87]. This could spatially smooth parameter estimates and maybe even better highlight observations with a high conditional probability of being a false zero. Furthermore, we would like to explore incorporating multiple years of surveillance data, rather than just one year. This could involve incorporating elements from repeated measures ecological analysis presented in much research by Royle [113], or a latent temporal process as utilized by Fernandes [99]. Moreover, more work needs to be done to make prediction-specific results in order to better identify false zero observations. Lastly, we would like to obtain more data from Ghana in order to perform a more thorough data analysis.

Chapter 5

Conclusion

We presented model-based statistical methods to analyze data from three arenas of public health: modeling vaccination coverage, utilizing remotely sensed data to augment disease surveillance in remote locations, and addressing non-reporting in surveillance of neglected tropical diseases. All three topics described have the common theme that in public health surveillance, the data we want is not the data that we get. In monitoring vaccination coverage, we do not always obtain the age of vaccination among vaccinated children. In making inferences on the presence of neglected tropical disease and the corresponding disease causing pathogens, we can utilize remotely sensed satellite imagery to augment analysis on surveillance in remote locations. In surveillance of neglected tropical diseases, resources are often not available to confirm that non-reporting areas are actually disease-free. We proposed new statistical models to overcome these limitations in our data. We found that although such models present some challenges, they are ready to be used in practice. Moreover, these models show promise for even more future development.

Bibliography

[1] A. Jenny, “Public health surveillance.” http://www.who.int/immunization_monitoring/burden/routine_surveillance/en/index.html, 2010.

[2] L. Rodewald, E. Maes, J. Stevenson, B. Lyons, S. Stokley, and P. Szilagyi, “Immunization performance measurement in a changing immunization environment,” *Pediatrics*, vol. 103, pp. 889–97, 1999.

[3] A. E. Sommerfelt and A. L. Piani, “Childhood immunization: 1990-1994,” tech. rep., Macro International Inc., 1997.

[4] World Health Organization, “Control of neglected tropical diseases.” http://www.who.int/neglected_diseases/en/, 2010.

[5] M. Vaessen, “The potential of the demographic and health surveys (DHS) for the evaluation and monitoring of maternal and child health indicators,” tech. rep., Macro International Inc, 1997.

[6] T. W. Pullum, “An assessment of the quality of data on health and nutrition in the DHS surveys, 1999-2003,” tech. rep., Macro International Inc., 2008.

[7] “Kenya Demographic and Health Survey 2003,” tech. rep., Central Bureau of Statistics (Kenya), Ministry of Health (Kenya), ORC Macro, 2003.

[8] B. Laubereau, M. Hermann, H. J. Schmitt, J. Weil, and R. Von Kries, "Detection of delayed vaccinations: a new approach to visualize vaccine uptake," *Epidemiology and Infection*, vol. 128, no. 2, pp. 185–192, 2002.

[9] E. Dannetun, A. Tegnell, G. Hermansson, A. Torner, and J. Giesecke, "Timeliness of MMR vaccination–influence on vaccination coverage," *Vaccine*, vol. 22, no. 31-32, pp. 4228–32, 2004.

[10] G. H. Dayan, K. M. Shaw, A. L. Baughman, L. C. Orellana, R. Forlenza, A. Ellis, J. Chaui, S. Kaplan, and P. Strebel, "Assessment of delay in age-appropriate vaccination using survival analysis," *American Journal of Epidemiology*, vol. 163, no. 6, pp. 561–70, 2006.

[11] M. Ndiritu, K. D. Cowgill, A. Ismail, S. Chiphatsi, T. Kamau, G. Fegan, D. R. Feikin, C. R. Newton, and J. A. Scott, "Immunization coverage and risk factors for failure to immunize within the Expanded Programme on Immunization in Kenya after introduction of new *Haemophilus influenzae* type b and hepatitis b virus antigens," *BMC Public Health*, vol. 6, p. 132, 2006.

[12] A. Clark and C. Sanderson, "Timing of children's vaccinations in 45 low-income and middle-income countries: an analysis of survey data," *Lancet*, vol. 373, no. 9674, pp. 1543–9, 2009.

[13] J. C. Moisi, J. Kabuka, D. Mitingi, O. Levine, and J. A. G. Scott, "Spatial and socio-demographic predictors of time-to-immunization in a rural area in Kenya: Is equity attainable?," *Vaccine*, vol. 28, pp. 5725–30, 2010.

[14] D. Collett, *Modeling Binary Data, 2nd Ed.* Chapman and Hall/CRC, 2003.

[15] I. Seidl and C. A. Tisdell, "Carrying capacity reconsidered: from Malthus' population theory to cultural carrying capacity," *Ecological Economics*, vol. 31, no. 3, pp. 395–408, 1999.

[16] R. Pearl and L. J. Reed, "On the rate of growth of the population of the United States since 1790 and its mathematical representation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 6, pp. 275–288, 1920.

[17] R. Pearl and L. J. Reed, "A further note on the mathematical theory of population growth," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 8, pp. 365–368, 1922.

[18] J. Berkson, "Application of the logistic function to bio-assay," *Journal of the American Statistical Association*, vol. 39, no. 227, pp. 357–365, 1944.

[19] J. Berkson, "A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function," *Journal of the American Statistical Association*, vol. 48, no. 263, pp. 565–599, 1953.

[20] F. R. Oliver, "Methods of estimating the logistic growth-function," *The Royal Statistical Society Series C-Applied Statistics*, vol. 13, no. 2, pp. 57–66, 1964.

[21] D. Rodbard and G. Frazier, "Statistical analysis of radiogland assay data," *Methods in Enzymology*, vol. 37, pp. 3–22, 1975.

[22] D. A. Ratkowsky and T. J. Reedy, "Choosing near-linear parameters in the 4-parameter logistic model for radioligand and related assays," *Biometrics*, vol. 42, no. 3, pp. 575–582, 1986.

[23] J. C. Pinheiro and D. M. Bates, *Mixed-effects models in S and S-PLUS.* New York: Springer, 2000.

[24] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The Computer Journal*, vol. 7, no. 4, pp. 308–313, 1965.

[25] R. H. Byrd, P. H. Lu, J. Nocedal, and C. Y. Zhu, "A limited memory algorithm for bound constrained optimization," *Siam Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[26] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Boca Raton: Chapman and Hall/CRC, second ed., 2004.

[27] T. Louis, "Rhetoric matters," *Biometic Bulletin*, vol. 24, no. 3, 2007.

[28] R. Langsten and K. Hill, "The accuracy of mothers' reports of child vaccination: Evidence from rural Egypt," *Social Science and Medicine*, vol. 46, no. 9, pp. 1205–1212, 1998.

[29] K. George, S. Victor, and R. Abel, "Reliability of mother as an informant with regard to immunisation," *Indian Journal of Pediatrics*, vol. 57, pp. 588–590, 1990.

[30] E. Gareaballah and B. Loevinsohn, "The accuracy of mother's reports about their children's vaccination status," *Bulletin of the World Health Organization*, vol. 67, pp. 669–674, 1989.

[31] H. H. AbdelSalam and M. M. Sokal, "Accuracy of parental reporting of immunization," *Clinical Pediatrics*, vol. 43, no. 1, pp. 83–5, 2004.

[32] L. Suarez, D. M. Simpson, and D. R. Smith, "Errors and correlates in parental recall of child immunizations: effects on vaccination coverage estimates," *Pediatrics*, vol. 99, no. 5, 1997.

[33] R. Ramakrishnan, R. Venkata, L. Sundaramoorthy, and J. V, "Magnitude of recall bias in the estimation of immunization coverage and its determinants," *Indian Pediatrics*, vol. 36, pp. 881–885, 1999.

[34] J. J. Valadez and L. H. Weld, "Maternal recall error of child vaccination status in a developing nation," *American Journal of Public Health*, vol. 82, no. 1, pp. 120–2, 1992.

[35] "Buruli ulcer: *Mycobacterium ulcerans* infection," tech. rep., World Health Organization, 2000.

[36] P. Johnson, T. Stinear, P. Small, G. Plushke, R. Merritt, F. Portaels, K. Huygen, J. Hayman, and K. Asiedu, "Buruli ulcer (*M. Ulcerans* infection): New insights, new hope for disease control," *PLoS Medicine*, vol. 2, no. 4, p. e108, 2005.

[37] R. Merritt, M. E. Benbow, and P. Small, "Unraveling an emerging disease associated with disturbed aquatic environments: the case of Buruli ulcer," *Frontiers of Ecology and the Environment*, vol. 3, no. 6, pp. 323–331, 2005.

[38] T. S. van der Werf, Y. Stienstra, R. C. Johnson, R. Phillips, O. Adjei, B. Fleischer, M. H. Wansbrough-Jones, P. Johnson, F. Portaels, W. van der Graaf, and K. Asiedu, "*Mycobacterium ulcerans* disease," *Bulletin of the World Health Organization*, vol. 83, no. 10, pp. 785–791, 2005.

[39] A. J. Radford, "*Mycobacterium ulcerans* in Australia," *Australian and New Zealand Journal of Medicine*, vol. 5, no. 2, pp. 162–169, 1975.

[40] P. J. Mitchell, I. V. Jerrett, and K. J. Slee, "Skin ulcers caused by *Mycobacterium ulcerans* in koalas near Bairnsdale, Australia," *Pathology*, vol. 16, no. 3, pp. 256–260, 1984.

[41] F. Portaels, P. Elsen, A. Guimaraes-Peres, P. A. Fonteyne, and W. M. Meyers, "Insects in the transmission of *Mycobacterium ulcerans* infection," *Lancet*, vol. 353, no. 9157, p. 986., 1999.

[42] F. Portaels, K. Chemlal, P. Elsen, P. Johnson, J. Hayman, J. Hibble, R. Kirkwood, and W. Meyers, "*Mycobacterium ulcerans* in wild animals," *Reviews in environmental science and bio-technology*, vol. 20, pp. 252–264, 2001.

[43] L. Marsollier, R. Robert, J. Aubry, J. P. Saint Andre, H. Kouakou, P. Legras, A. L. Manceau, C. Mahaza, and B. Carbonnelle, "Aquatic insects as a vector for *Mycobacterium ulcerans*," *Applied and Environmental Microbiology*, vol. 68, no. 9, pp. 4623–4628, 2002.

[44] M. Eddyani, D. Ofori-Adjei, G. Teugels, D. De Weirdt, D. Boakye, W. Meyers, and F. Portaels, "Potential role for fish in transmission of *Mycobacterium ulcerans* disease (Buruli ulcer): an environmental study," *Applied and Environmental Microbiology*, vol. 70, no. 9, pp. 5679–5682, 2004.

[45] L. Marsollier, T. Stinear, J. Aubry, J. P. Saint Andre, R. Robert, P. Legras, A. L. Manceau, C. Audrain, S. Bourdon, H. Kouakou, and B. Carbonnelle, "Aquatic plants stimulate the growth of and biofilm formation by *Mycobacterium ulcerans* in axenic culture and harbor these bacteria in the environment," *Applied and Environmental Microbiology*, vol. 70, no. 2, pp. 1097–103, 2004.

[46] M. E. Benbow, H. Williamson, R. Kimbirauskas, M. McIntosh, R. Kolar, C. Quaye, F. Akpabey, D. Boakye, P. Small, and R. Merritt, "Aquatic invertebrates as unlikely vectors of Buruli ulcer disease," *Emerging Infectious Diseases*, vol. 14, no. 8, pp. 1247–1254, 2008.

[47] R. W. Merritt, E. D. Walker, P. L. Small, J. R. Wallace, P. D. Johnson, M. E. Benbow, and D. A. Boakye, "Ecology and transmission of Buruli ulcer disease: a systematic review," *PLoS Neglected Tropical Diseases*, vol. 4, no. 12, p. e911, 2010.

[48] W. Revill and D. Barker, "Seasonal distribution of mycobacterial skin ulcers," *British Journal of Preventative and Social Medicine*, vol. 26, pp. 23–27, 1972.

[49] D. J. P. Barker, "Epidemiology of *Mycobacterium Ulcerans* infection," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 67, no. 1, pp. 43–50, 1973.

[50] H. S. Thangaraj, M. R. W. Evans, and M. H. Wansbrough-Jones, "*Mycobacterium ulcerans* disease; Buruli ulcer," *Transactions of the Royal Society of Tropical Medicine and Hygiene*, vol. 93, no. 4, pp. 337–340, 1999.

[51] H. Aiga, T. Amano, S. Cairncross, J. A. Domako, O. K. Nanas, and S. Coleman, "Assessing water-related risk factors for Buruli ulcer: a case-control study in Ghana," *American Journal of Tropical Medicine and Hygiene*, vol. 71, no. 4, pp. 387–92, 2004.

[52] P. L. Raghunathan, E. A. S. Whitney, K. Asamoa, Y. Stienstra, T. H. Taylor, G. K. Amofah, D. Ofori-Adjei, K. Dobos, J. Guarner, S. Martin, S. Pathak, E. Klutse, S. Etuaful, W. I. A. van der Graaf, T. S. van der Werf, C. H. King, J. W. Tappero, and D. A. Ashford, "Risk factors for Buruli ulcer disease (*Mycobacterium ulcerans* infection): Results from a case-control study in Ghana," *Clinical Infectious Diseases*, vol. 40, no. 10, pp. 1445–1453, 2005.

[53] A. Duker, A. Stein, and M. Hale, "A statistical model for spatial patterns of Buruli ulcer in Amansie West district, Ghana," *International Journal of Applied Earth Observation and Geoinformation*, vol. 8, pp. 126–136, 2006.

[54] T. Wagner, M. E. Benbow, T. O. Brenden, J. Qi, and R. C. Johnson, "Buruli ulcer disease prevalence in Benin, West Africa: associations with land use/cover and the identification of disease clusters," *International Journal of Health Geographics*, vol. 7, 2008.

[55] T. Wagner, M. E. Benbow, M. Burns, R. C. Johnson, R. Merritt, J. Qi, and P. Small, “A landscape-based model for predicting *Mycobacterium ulcerans* infection (Buruli ulcer disease) presence in Benin, West Africa,” *EcoHealth*, vol. 5, pp. 69–79, 2008.

[56] H. Williamson, M. E. Benbow, K. Nguyen, D. Beachboard, R. Kimbirauskas, M. McIntosh, C. Quaye, E. Ampadu, D. Boakye, R. Merritt, and P. Small, “Distribution of *Mycobacterium ulcerans* in Buruli ulcer endemic and non-endemic aquatic sites in Ghana,” *PLoS Neglected Tropical Diseases*, vol. 2, no. 3, p. e205, 2008.

[57] B. C. Ross, P. D. R. Johnson, F. Oppedisano, L. Marino, A. Sievers, T. Stinear, J. A. Hayman, M. G. K. Veitch, and R. M. RobinsBrowne, “Detection of *Mycobacterium ulcerans* in environmental samples during an outbreak of ulcerative disease,” *Applied and Environmental Microbiology*, vol. 63, no. 10, pp. 4135–4138, 1997.

[58] T. Stinear, J. K. Davies, G. A. Jenkin, J. A. Hayman, F. Oppedisano, and P. D. R. Johnson, “Identification of *Mycobacterium ulcerans* in the environment from regions in southeast Australia in which it is endemic with sequence capture-PCR,” *Applied and Environmental Microbiology*, vol. 66, no. 8, pp. 3206–3213, 2000.

[59] L. Mosi, H. Williamson, J. R. Wallace, R. W. Merritt, and P. L. C. Small, “Persistent association of *Mycobacterium ulcerans* with West African predaceous insects of the family Belostomatidae,” *Applied and Environmental Microbiology*, vol. 74, no. 22, pp. 7036–7042, 2008.

[60] T. Brou, H. Broutin, E. Elguero, H. Asse, and J. F. Guegan, “Landscape diversity related to Buruli ulcer disease in Côte d’Ivoire,” *PLoS Neglected Tropical*

Diseases, vol. 2, no. 7, p. e271, 2008.

[61] NASA Landsat Program, "Landsat ETM+." `http://glcf.umiacs.umd.edu/`, 2000.

[62] K. Beven and M. Kirkby, "A physically based, variable contributing area model of basin hydrology," *Hydrological Sciences*, vol. 24, no. 1, pp. 43–69, 1979.

[63] L. Waller and C. Gotway, *Applied Spatial Statistics for Public Health Data*. Hoboken, NJ: John Wiley and Sons, 2004.

[64] R Development Core Team, "R: A language and environment for statistical computing." `http://www.sciviews.org/_rgui/`, 2011.

[65] M. Kulldorff, "A spatial scan statistic," *Communications in Statistics: Theory and Methods*, vol. 26, pp. 1481–1496, 1997.

[66] M. Kulldorff and I. M. Services Inc., "Satscan 8.0: Software for the spatial and space-time scan statistics." `www.satscan.org`, 2009.

[67] K. P. Burnham and D. R. Anderson, *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, 2nd ed.* Springer, 2002.

[68] P. J. Diggle, M. C. Thomson, O. F. Christensen, B. Rowlingson, V. Obsomer, J. Gardon, S. Wanji, I. Takougang, P. Enyong, J. Kamgno, J. H. Remme, M. Boussinesq, and D. H. Molyneux, "Spatial modelling and the prediction of *Loa loa* risk: Decision making under uncertainty," *Annals of Tropical Medicine and Parasitology*, vol. 101, no. 6, pp. 499–509, 2007.

[69] E. Marion, S. Eyangoh, E. Yeramian, J. Doannio, J. Landier, J. Aubry, A. Fontanet, C. Rogier, V. Cassisa, J. Cottin, A. Marot, M. Eveillard, Y. Kamdem, P. Legras, C. Deshayes, J. P. Saint-Andre, and L. Marsollier, "Seasonal and regional dynamics of *M. ulcerans* transmission in environmental context:

deciphering the role of water bugs as hosts and vectors," *PLoS Neglected Tropical Diseases*, vol. 4, no. 7, p. e731, 2010.

[70] J. C. Palomino, A. M. Obiang, L. Realini, W. M. Meyers, and F. Portaels, "Effect of oxygen on growth of *Mycobacterium ulcerans* in the BACTEC system," *Journal of Clinical Microbiology*, vol. 36, no. 11, pp. 3420–2, 1998.

[71] G. HIlson, "Harvesting mineral riches: 1000 years of gold mining in Ghana," *Resources Policy*, vol. 28, pp. 13–26, 2002.

[72] H. Boisvert, "Skin ulcer caused by *Mycobacterium ulcerans* in Cameroon," *Bulletin de la Societe de Pathologie Exotique et de ses Filiales*, vol. 70, no. 2, pp. 125–31, 1977.

[73] J. A. Hayman, H. B. Fleming, D. A. Monash, and I. M. Miller, "*Mycobacterium ulcerans* infection in paradise [letter]," *Medical Journal of Australia*, vol. 155, no. 2, p. 130, 1991.

[74] C. R. Horsburgh and W. M. Meyers, *Buruli Ulcer*, pp. 119–126. Washington, D.C.: American Society for Microbiology, 1997.

[75] G. E. Sopoh, R. C. Johnson, S. Y. Anagonou, Y. T. Barogui, A. D. Dossou, J. G. Houezo, D. M. Phanzu, B. H. Tente, W. M. Meyers, and F. Portaels, "Buruli ulcer prevalence and altitude, Benin," *Emerging Infectious Diseases*, vol. 17, no. 1, pp. 153–4, 2011.

[76] W. M. Meyers, N. Tignokpa, G. B. Priuli, and F. Portaels, "*Mycobacterium ulcerans* infection (Buruli ulcer): first reported patients in Togo," *British Journal of Dermatology*, vol. 134, no. 6, pp. 1116–21, 1996.

[77] P. Johnson, T. Stinear, and J. Hayman, "*Mycobacterium ulcerans* - a minireview," *Journal of Medical Microbiology*, vol. 48, pp. 511–513, 1999.

[78] K. Asiedu and S. Etuaful, "Socioeconomic implications of Buruli ulcer in Ghana: a three-year review," *American Journal of Tropical Medicine and Hygiene*, vol. 59, no. 6, pp. 1015–22, 1998.

[79] Y. Stienstra, W. T. van der Graaf, K. Asamoa, and T. S. van der Werf, "Beliefs and attitudes toward Buruli ulcer in Ghana," *American Journal of Tropical Medicine and Hygiene*, vol. 67, no. 2, pp. 207–13, 2002.

[80] M. Guerra, E. Walker, C. Jones, S. Paskewitz, M. R. Cortinas, A. Stancil, L. Beck, M. Bobo, and U. Kitron, "Predicting the risk of lyme disease: Habitat suitability for *Ixodes scapularis* in the north central United States," *Emerging Infectious Diseases*, vol. 8, no. 3, pp. 289–297, 2002.

[81] M. Pascual, X. Rodo, S. P. Ellner, R. Colwell, and M. J. Bouma, "Cholera dynamics and El Niño-southern oscillation," *Science*, vol. 289, no. 5485, pp. 1766–9, 2000.

[82] S. Altizer, A. Dobson, P. Hosseini, P. Hudson, M. Pascual, and P. Rohani, "Seasonality and the dynamics of infectious diseases," *Ecology Letters*, vol. 9, no. 4, pp. 467–84, 2006.

[83] F. Dominici, R. D. Peng, M. L. Bell, L. Pham, A. McDermott, S. L. Zeger, and J. M. Samet, "Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases," *Journal of the American Medical Association*, vol. 295, no. 10, pp. 1127–34, 2006.

[84] M. Ridout, C. Demetrio, and J. Hinde, "Models for count data with many zeros," 1998.

[85] D. Lambert, "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, vol. 34, no. 1, pp. 1–14, 1992.

[86] S. K. Ghosh, P. Mukhopadhyay, and J. C. Lu, "Bayesian analysis of zero-inflated regression models," *Journal of Statistical Planning and Inference*, vol. 136, no. 4, pp. 1360–1375, 2006.

[87] D. K. Agarwal, A. E. Gelfand, and S. Citron-Pousty, "Zero-inflated models with application to spatial count data," *Environmental and Ecological Statistics*, vol. 9, no. 4, pp. 341–355, 2002.

[88] P. M. Kuhnert, T. G. Martin, K. Mengersen, and H. P. Possingham, "Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion," *Environmetrics*, vol. 16, no. 7, pp. 717–747, 2005.

[89] T. G. Martin, B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham, "Zero tolerance ecology: improving ecological inference by modelling the source of zero observations," *Ecology Letters*, vol. 8, no. 11, pp. 1235–1246, 2005.

[90] D. I. Mackenzie, J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm, "Estimating site occupancy rates when detection probabilities are less than one," *Ecology*, vol. 83, no. 8, pp. 2248–2255, 2002.

[91] G. A. Dagne, "Bayesian semiparametric zero-inflated poisson model for longitudinal count data," *Mathematical Biosciences*, vol. 224, no. 2, pp. 126–30, 2010.

[92] O. Flores, V. Rossi, and F. Mortier, "Autocorrelation offsets zero-inflation in models of tropical saplings density," *Ecological Modelling*, vol. 220, pp. 1797–1809, 2009.

[93] J. Rodrigues, "Full Bayesian significance test for zero-inflated distributions," *Communications in Statistics-Theory and Methods*, vol. 35, pp. 299–307, 2006.

[94] Y. Min and A. Agresti, “Random effect models for repeated measures of zero-inflated count data,” *Statistical Modelling*, vol. 5, pp. 1–19, 2005.

[95] J. A. Tooze, G. K. Grunwald, and R. H. Jones, “Analysis of repeated measures data with clumping at zero,” *Statistical Methods in Medical Research*, vol. 11, no. 4, pp. 341–55, 2002.

[96] B. H. Neelon, A. J. O’Malley, and S. L. T. Normand, “A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use,” *Statistical Modelling*, vol. 10, no. 4, pp. 421–439, 2010.

[97] A. Agresti, *Categorical Data Analysis*. Hoboken: John Wiley and Sons, Inc., 2nd ed., 2002.

[98] T. G. Martin, P. M. Kuhnert, K. Mengersen, and H. P. Possingham, “The power of expert opinion in ecological models using Bayesian methods: Impact of grazing on birds,” *Ecological Applications*, vol. 15, no. 1, pp. 266–280, 2005.

[99] M. V. Fernandes, A. M. Schmidt, and H. S. Migon, “Modelling zero-inflated spatio-temporal processes,” *Statistical Modelling*, vol. 9, no. 1, pp. 3–25, 2009.

[100] S. Gschlossl and C. Czado, “Modelling count data with overdispersion and spatial effects,” *Statistical Papers*, vol. 49, no. 3, pp. 531–552, 2008.

[101] P. Vounatsou, G. Raso, M. Tanner, E. N’Goran, and J. Utzinger, “Bayesian geostatistical modelling for mapping schistosomiasis transmission,” *Parasitology*, vol. 136, pp. 1695–1705, 2009.

[102] R. J. Ma, M. T. Hasan, and G. Sneddon, “Modelling heterogeneity in clustered count data with extra zeros using compound Poisson random effect,” *Statistics in Medicine*, vol. 28, no. 18, pp. 2356–2369, 2009.

[103] C. S. Li, J. C. Lu, and J. H. Park, “Multivariate zero-inflated Poisson models and their applications,” *Technometrics*, vol. 41, no. 1, pp. 29–38, 1999.

[104] G. A. Dagne, “Hierarchical Bayesian analysis of correlated zero-inflated count data,” *Biometrical Journal*, vol. 46, no. 6, pp. 653–663, 2004.

[105] C. Xue-Dong, “Bayesian analysis of semiparametric mixed-effects models for zero-inflated count data,” *Communications in Statistics-Theory and Methods*, vol. 38, no. 11, pp. 1815–1833, 2009.

[106] D. Clayton and J. Kaldor, “Empirical bayes estimates of age-standardized relative risks for use in disease mapping,” *Biometrics*, vol. 43, no. 3, pp. 671–681, 1987.

[107] J. Besag, J. York, and A. Mollie, “Bayesian image-restoration, with 2 applications in spatial statistics,” *Annals of the Institute of Statistical Mathematics*, vol. 43, no. 1, pp. 1–20, 1991.

[108] N. G. Best, R. A. Arnold, A. Thomas, L. A. Waller, and E. M. Conlon, “Bayesian models for spatially correlated disease and exposure data,” *Bayesian Statistics 6*, pp. 131–156 867, 1999.

[109] J. A. Royle and J. D. Nichols, “Estimating abundance from repeated presence-absence data or point counts,” *Ecology*, vol. 84, no. 3, pp. 777–790, 2003.

[110] J. A. Royle, “N-mixture models for estimating population size from spatially replicated counts,” *Biometrics*, vol. 60, pp. 108–115, 2004.

[111] R. M. Dorazio, B. Mukherjee, L. Zhang, M. Ghosh, H. L. Jelks, and F. Jordan, “Modeling unobserved sources of heterogeneity in animal abundance using a Dirichlet process prior,” *Biometrics*, vol. 64, pp. 635–644, 2008.

[112] S. J. Wenger and M. C. Freeman, "Estimating species occurrence, abundance, and detection probability using zero-inflated distributions," *Ecology*, vol. 2008, pp. 2953–2959, 2008.

[113] J. A. Royle and R. M. Dorazio, *Hierarchical Modeling and Inference in Ecology.* Elsevier Ltd., 2008.

[114] D. J. Spiegelhalter, N. G. Best, B. R. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, vol. 64, pp. 583–616, 2002.

Appendices

Appendix A

Likelihood-based approach to non-linear logistic growth model

Model (2.8): The outcome Y_i is binary (0 or 1) where

$$P(Y_i = 1) = \frac{\phi_1}{1 + \exp\{-(x_i - \phi_2)/\phi_3\}}$$

We can write the likelihood as:

$$\begin{aligned}
\mathrm{L}(\phi) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\
\log \mathrm{L}(\theta) &= \sum_{i=1}^{n} \log\left[p_i^{y_i}(1-p_i)^{1-y_i}\right] \\
&= \sum_{i=1}^{n} \log\left[\left(\frac{\phi_1}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)}\right)^{y_i}\left(1-\frac{\phi_1}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)}\right)^{1-y_i}\right] \\
&= \sum_{i=1}^{n} y_i \log\phi_1 - \log\left\{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)\right\} \\
&\quad + (1-y_i)\log\left\{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)-\phi_1\right\}
\end{aligned}$$

Let $\theta = (\phi_1, \phi_2, \phi_3)$

Using the following substitutions:

$$\begin{aligned} a_i &= \frac{(x_i - \phi_2)}{\phi_3} \\ b_i &= 1 + \mathrm{e}^{-a_i} - \phi_1 \\ c_i &= 1 + \mathrm{e}^{-a_i} \end{aligned}$$

The asymptotic distribution of $\hat{\theta} = (\hat{\phi_1}, \hat{\phi_2}, \hat{\phi_3})$ is $\sqrt{n}(\hat{\theta} - \theta) \rightarrow N_3\left(0, \{nI_n(\theta)\}^{-1}\right)$ where

$$I_n(\theta) = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^{n} \frac{1}{\phi_1 b_i} & \sum_{i=1}^{n} -\frac{\mathrm{e}^{-a_i}}{\phi_3 b_i c_i} & \sum_{i=1}^{n} & -\frac{a_i \mathrm{e}^{-a_i}}{\phi_3 b_i c_i} \\ & \sum_{i=1}^{n} \frac{\phi_1 \mathrm{e}^{-2a_i}}{\phi_3^2 b_i c_i^2} & \sum_{i=1}^{n} & \frac{-\phi_1 a_i \mathrm{e}^{-2a_i}}{\phi_3^2 b_i c_i^2} \\ & & \sum_{i=1}^{n} & \frac{\phi_1 a_i^2 \mathrm{e}^{-2a_i}}{\phi_3^2 b_i c_i^2} \end{pmatrix}$$

Model (2.9): The outcome Y_i is binary (0 or 1) where

$$P(Y_i = 1) = \frac{\frac{1}{1+\exp(-\lambda)}}{1 + \exp\left(-\frac{(x_i - \phi_2)}{\phi_3}\right)}$$

We can write the likelihood as:

$$
\begin{aligned}
\mathrm{L}(\phi) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\
\log \mathrm{L}(\phi) &= \sum_{i=1}^{n} \log p_i^{y_i}(1-p_i)^{1-y_i} \\
&= \sum_{i=1}^{n} \log \left(\frac{\frac{1}{1+\exp(-\lambda)}}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^{y_i} \left(1 - \frac{\frac{1}{1+\exp(-\lambda)}}{1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)} \right)^{1-y_i} \\
&= \sum_{i=1}^{n} -y_i \log\left(1+\exp(-\lambda)\right) + (1-y_i)\log\left(1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right) - \frac{1}{1+\exp(-\lambda)}\right) \\
&\quad - \log\left(1+\exp\left(-\frac{(x_i-\phi_2)}{\phi_3}\right)\right)
\end{aligned}
$$

Let $\theta = (\lambda, \phi_2, \phi_3)$

Using the following substitutions:

$$
\begin{aligned}
g &= \mathrm{e}^{-\lambda} \\
h &= \frac{1}{1+g} \\
a_i &= \frac{-(x_i-\phi_2)}{\phi_3} \\
b_i &= 1 + \mathrm{e}^{a_i} - h \\
c_i &= 1 + \mathrm{e}^{a_i}
\end{aligned}
$$

The asymptotic distribution of $\hat{\theta} = (\hat{\lambda}, \hat{\phi}_2, \hat{\phi}_3)$ is $\sqrt{n}(\hat{\theta} - \theta) \rightarrow N_3\left(0, \{nI_n(\theta)\}^{-1}\right)$ where

$$
I_n(\theta) = \frac{1}{n}\begin{pmatrix}
\sum_{i=1}^{n} \frac{g^2}{c_i(1+g)^3}\left[1+\frac{1}{b_i(1+g)}\right] & \sum_{i=1}^{n} -\frac{g\mathrm{e}^{a_i}}{\phi_3 b_i c_i (1+g)^2} & \sum_{i=1}^{n} \frac{g a_i \mathrm{e}^{a_i}}{\phi_3 b_i c_i (1+g)^2} \\
 & \sum_{i=1}^{n} \frac{h\mathrm{e}^{2a_i}}{\phi_3^2 b_i c_i^2} & \sum_{i=1}^{n} -\frac{h a_i \mathrm{e}^{2a_i}}{\phi_3^2 b_i c_i^2} \\
 & & \sum_{i=1}^{n} \frac{h a_i^2 \mathrm{e}^{2a_i}}{\phi_3^2 b_i c_i^2}
\end{pmatrix}.
$$

Appendix B

Figures for logistic growth model results

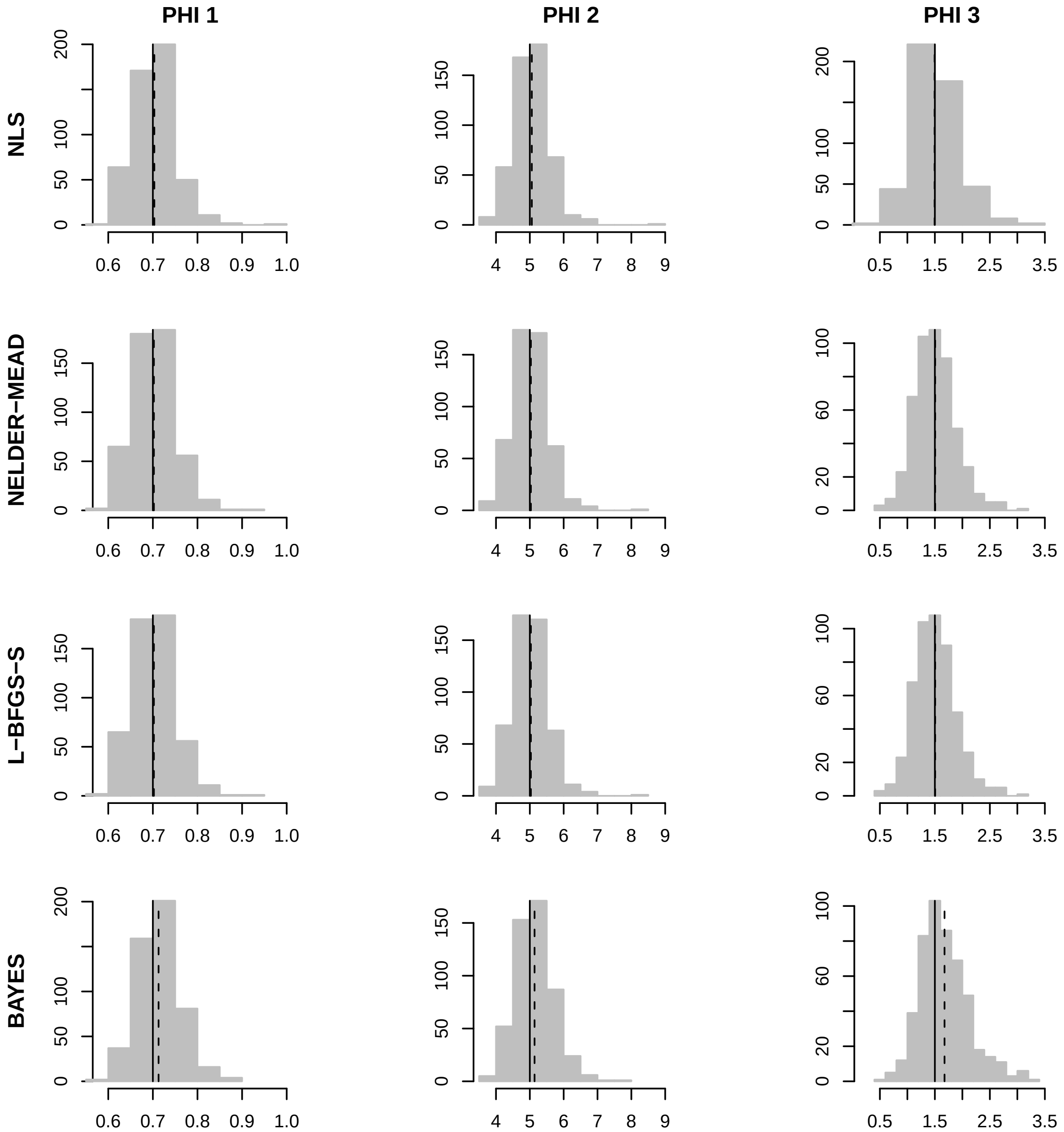


Figure B.1: Histograms of $\hat{\phi}$ for Model (2.8). The solid black line is the true value of the parameter, and the dashed black line is the mean value of the parameter estimates over the 500 simulations.

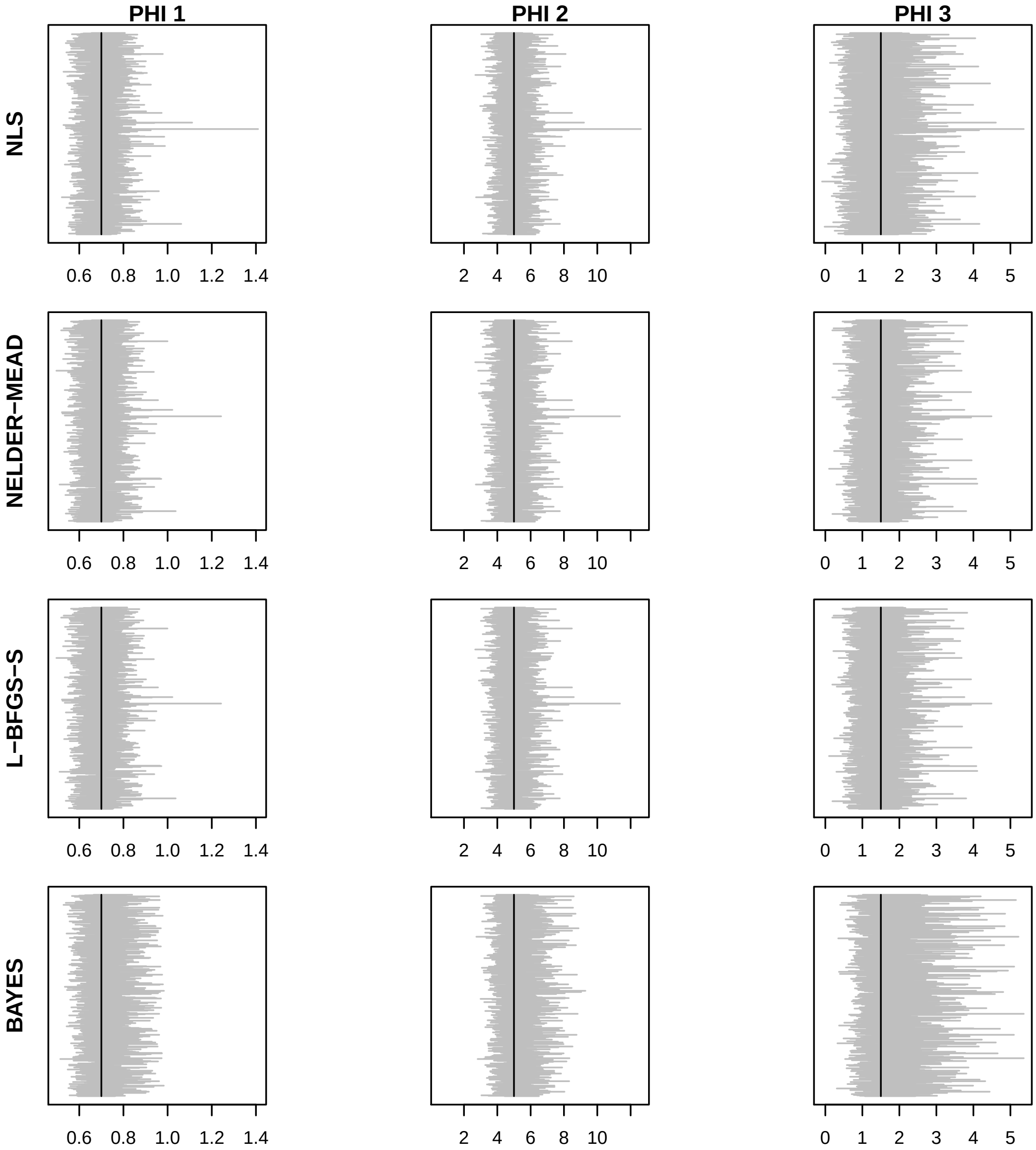


Figure B.2: Confidence intervals/credible sets of $\hat{\phi}$ for Model (2.8). The solid black line is the true value of the parameter.

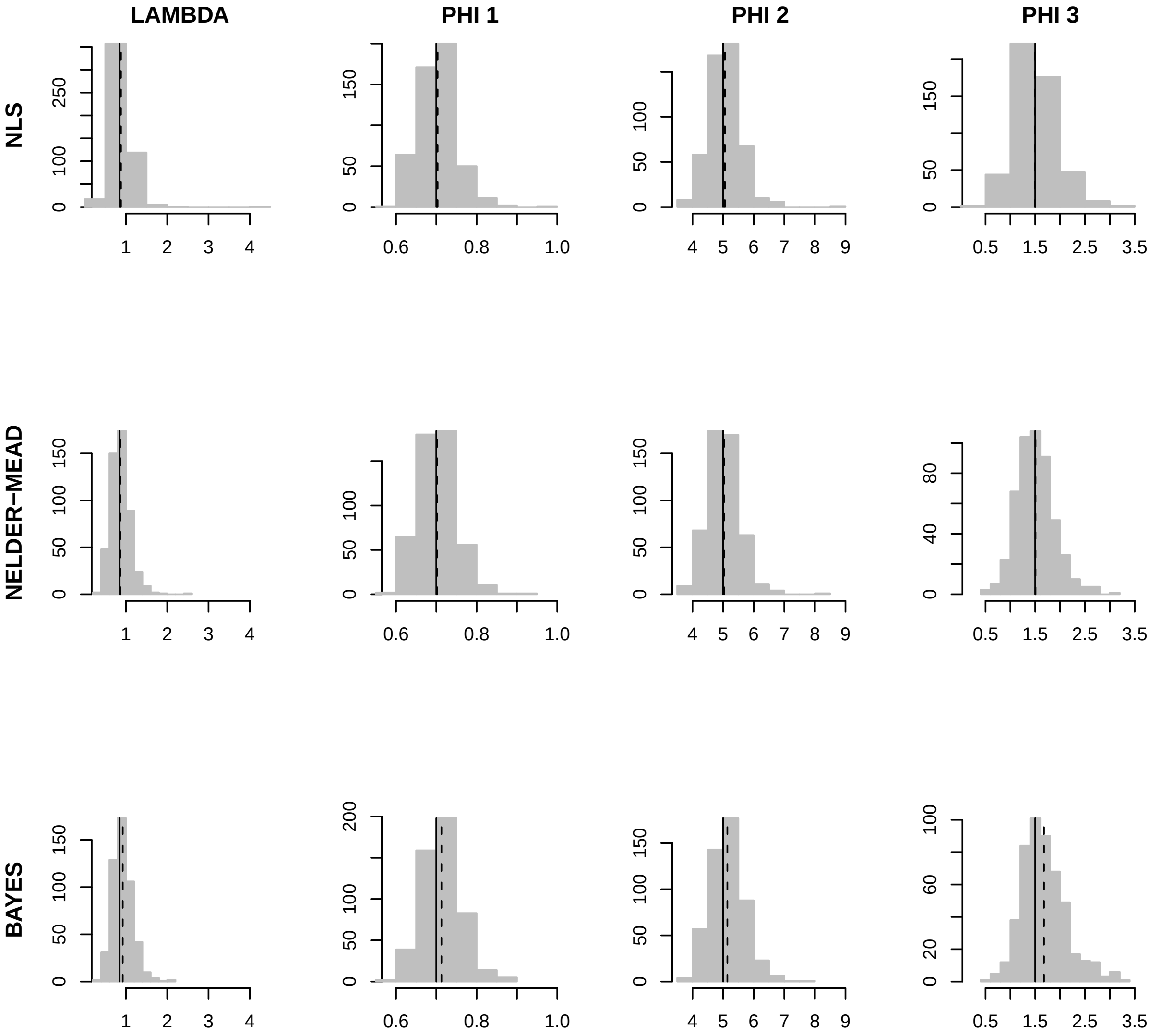


Figure B.3: Histograms of $\hat{\lambda}$ and $\hat{\phi}$ for Model (2.9). The solid black line is the true value of the parameter, and the dashed black line is the mean value of the parameter estimates over the 500 simulations.

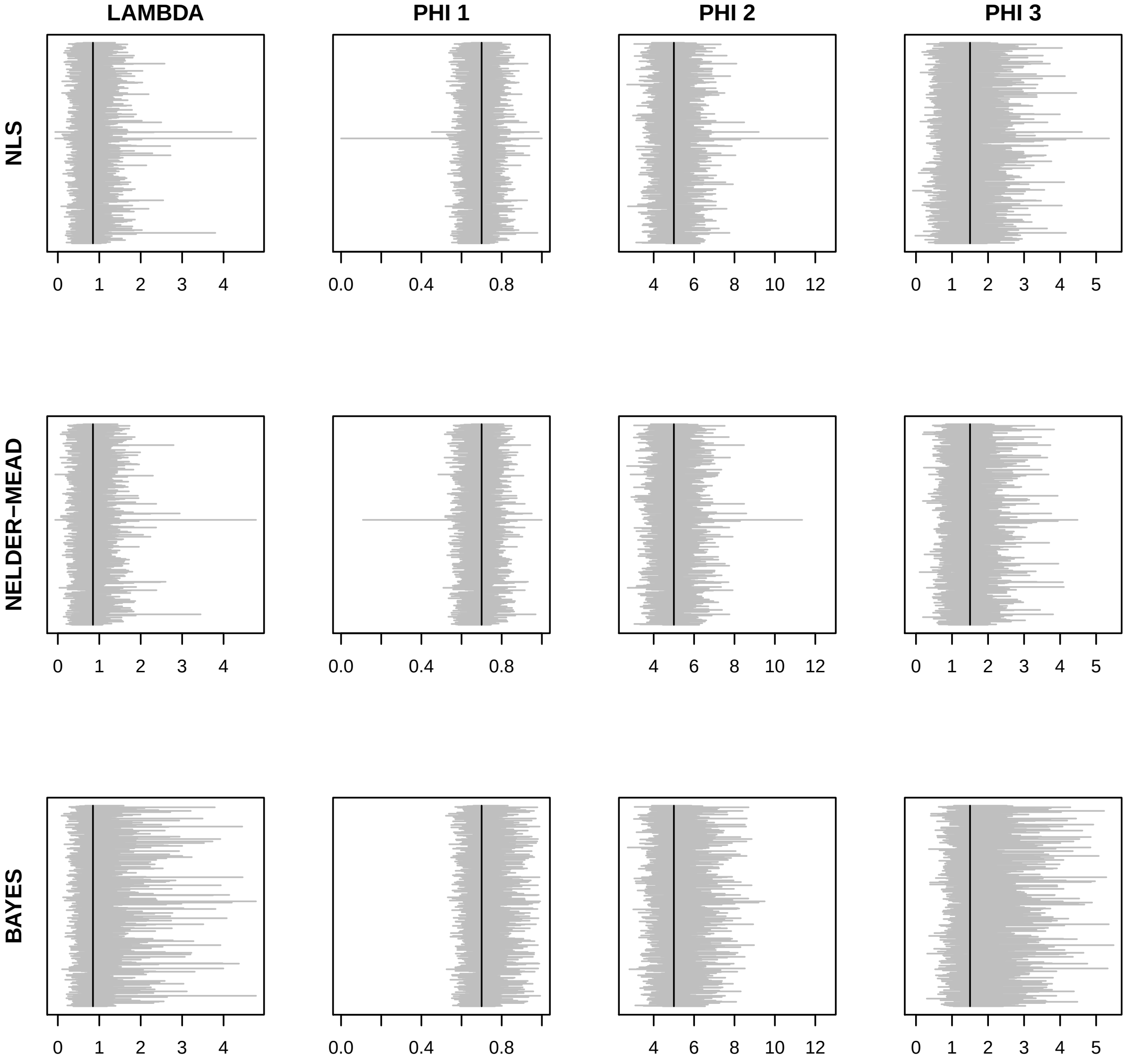


Figure B.4: Confidence intervals/credible sets of $\hat{\lambda}$ and $\hat{\phi}$ for Model (2.9). The solid black line is the true value of the parameter. The confidence interval for one simulation extends beyond the displayed range for λ in NLS and Nelder-Mead.

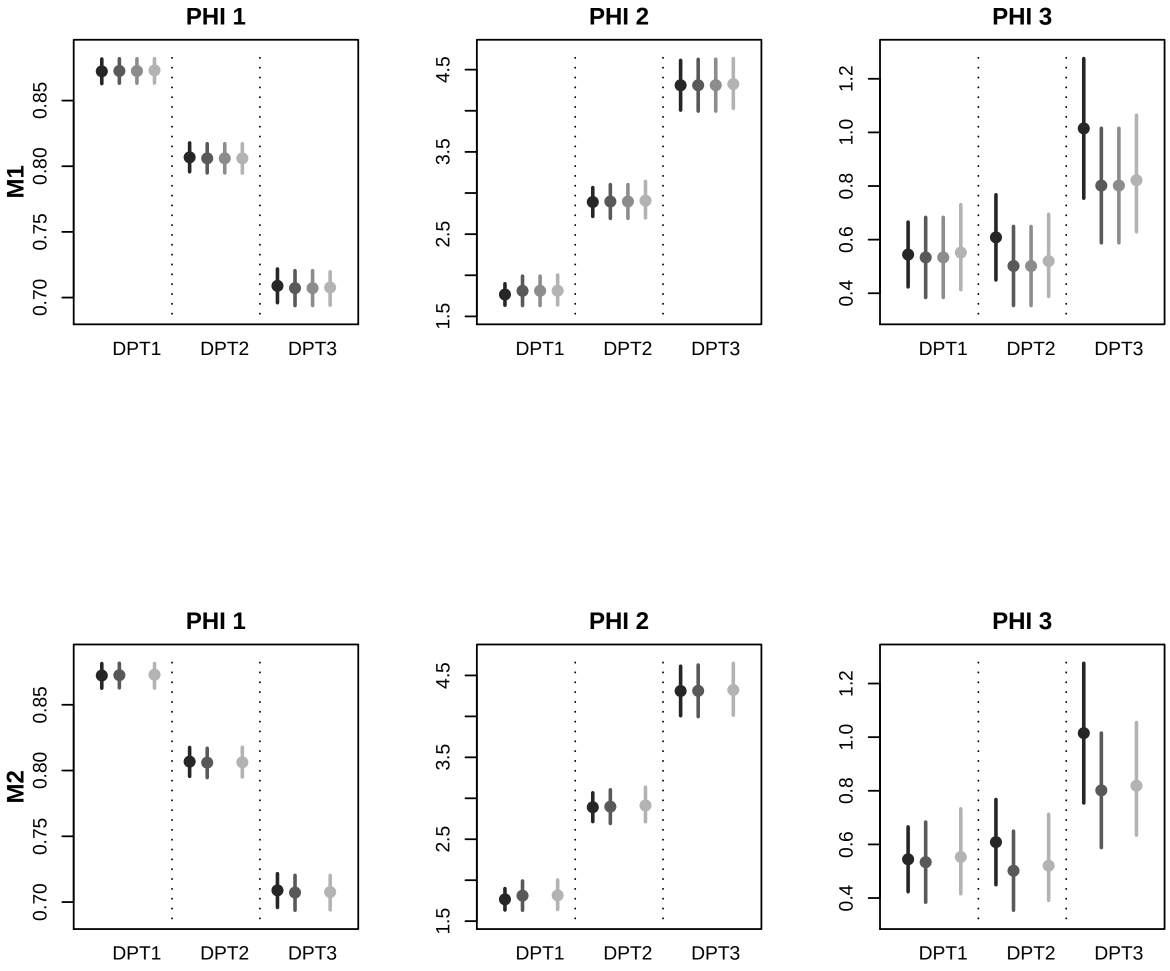


Figure B.5: Point estimates and 95% confidence intervals/credible sets for DPT1, DPT2, and DPT3 coverage from the 2003 Kenya DHS. The four lines in decreasing gray scale indicate: (1) nonlinear least squares, (2) Nelder-Mead, (3) L-BFGS-S, (4) Bayesian estimates. L-BFGS-S was not used for Model (2.9) as it would produce the same results as the Nelder-Mead algorithm.

Appendix C

Selection of parameter values for hierarchical ZIP data generation

Here we describe in detail how parameter values were chosen by which to generate the data. Parameters were chosen first for the data realization process where false zero generation was only mildly related to Y_{TRUE}. Afterwards, alternate values of β_{0REP} and β_{2REP} were selected to mimic the original data realization process, but to also have a more extreme relationship between false zero generation and Y_{TRUE}.

Great care was taken to ensure appropriate parameter values for all β's. Our goal was to closely mimic our actual data set, which had 31 out of 89 districts report cases of Buruli ulcer, representing 35% of the data which had non-zero observations. Therefore, we decided to simulate a sample size of $n = 100$ and ensure that after all of the steps of the hierarchial process had been completed that we would have on average 35 non-zero observations.

We achieved this by performing a brief simulation over different values of β's to identify optimal combinations of β's that would provide our desired end result. We considered a sample size of 100 where the covariates X_{MU}, X_{BU}, X_{CASES}, and X_{REP}, were fixed at the same values throughout the entire simulation with $Pr(X_{MU} = 1) =$

Variable	Rationale	Values Considered
β_{0MU}	> 0 in order to have baseline $Pr(Z_{MU} = 1) > 0.5$	0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0
β_{1MU}	> 0 so that the covariate X_{MU} is associated with higher $Pr(Z_{MU} = 1)$	0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0
β_{0BU}	> 0 in order to have baseline $Pr(Z_{BU} = 1) > 0.5$	0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0
β_{1BU}	> 0 so that the covariate X_{BU} is associated with higher $Pr(Z_{BU} = 1)$	0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0
β_{0CASES}	Baseline rate of cases is 3	log(3)
β_{1CASES}	Rate of cases when $X_{REP} = 1$ is 9 (RR=3)	log(3)
β_{0REP}	> 0 in order to have baseline $Pr(Z_{REP} = 1) > 0.5$	0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0
β_{1REP}	< 0 so that the covariate X_{REP} is associated with lower $Pr(Z_{REP} = 1)$	-2, -1.6, -1.4, -1.2, -0.8, -0.4
β_{2REP}	> 0 so that more cases are associated with higher $Pr(Z_{REP} = 1)$	0.05, 0.10, 0.15, 0.20, 0.25, 0.30

Table C.1: Candidate values considered for each covariate in the modified hierarchical ZIP model.

0.5, $Pr(X_{BU} = 1) = 0.5$, $Pr(X_{CASES} = 1) = 0.5$, and $Pr(X_{REP} = 1) = 0.5$. We considered a set of candidate values for each of the parameters (Table C.1). For each combination of parameter values, 100 simulations of the data were performed in which we recorded the total number of $Z_{MU} = 1$, $Z_{BU} = 1$,and $Z_{REP} = 1$, in order to represent the number of MU+ sites, the number of BU+ sites, and the number of reporting+ sites for those parameter values. A six number summary of the 100 repetitions was recorded to represent the minimum, first quartile, median, mean, third quartile, and maximum of the number of MU+ sites, the number of BU+ sites, and the number of reporting+ sites for those parameter value combinations. We then made restrictions on the summary numbers in order to narrow down the combinations of parameter values (Table C.2). After the restrictions in Table C.2 were satisfied, this narrowed it down to less than 100 suitable parameter value combinations. After that, arbitrary restrictions were made such that $\beta_{0MU} \neq \beta_{1MU}$, $\beta_{0BU} \neq \beta_{1BU}$, and $\beta_{1MU} \neq \beta_{1BU}$ in order to have varied effect sizes, which resulted in 17 possible parameter value

Minimum number of MU+ districts	> 60
Maximum number of MU+ districts	< 80
1st quartile of BU+ districts	> 40
3rd quartile of BU+ districts	< 60
Median number of Rep+ districts	$= 35$

Table C.2: Restrictions made on model summary in order to identify appropriate candidate parameter values.

								No. MU+			No. BU+			No. Rep+		
	β_{0MU}	β_{1MU}	β_{0BU}	β_{1BU}	β_{0REP}	β_{1REP}	β_{2REP}	*Min*	*Med*	*Max*	*Min*	*Med*	*Max*	*Min*	*Med*	*Max*
1	0.2	1.2	0.6	2.0	0.4	-1.2	0.15	61	68.0	78	44	54	65	25	35	46
2	0.2	1.2	0.8	1.8	0.2	-1.2	0.15	62	68.0	78	44	57	68	25	35	47
3	0.2	1.4	0.2	0.6	1.2	-1.2	0.15	61	69.5	79	29	44	59	20	35	48
4	0.2	1.4	0.2	1.6	0.6	-1.2	0.15	61	71.0	79	42	50	60	27	35	45
5	0.2	1.4	0.4	1.2	0.6	-1.2	0.15	61	70.0	79	37	50	65	24	35	45
6	0.2	1.4	0.4	2.0	0.4	-1.2	0.15	62	71.0	79	40	54	63	24	35	45
7	0.2	1.4	0.6	0.8	0.6	-1.2	0.15	61	70.0	79	40	49	63	23	35	45
8	0.2	1.4	0.8	0.4	0.6	-1.2	0.15	61	70.0	78	38	51	60	21	35	45
9	0.2	1.6	0.8	0.2	0.6	-1.2	0.15	61	71.0	79	38	51	61	25	35	48
10	0.2	1.6	0.8	1.0	0.4	-1.2	0.15	62	71.5	79	44	55	65	26	35	47
11	0.2	1.6	1.0	0.4	0.4	-1.2	0.15	62	71.0	79	35	54	66	23	35	48
12	0.2	1.8	0.4	1.0	0.6	-1.2	0.15	61	71.0	79	37	50	61	25	35	44
13	0.2	1.8	0.6	1.6	0.2	-1.2	0.15	63	74.0	79	47	57	67	25	35	45
14	0.4	1.0	0.2	1.4	0.8	-1.2	0.15	62	71.0	79	36	49	60	23	35	46
15	0.4	1.0	0.8	2.0	0.2	-1.2	0.15	61	70.0	79	46	57	67	22	35	45
16	0.4	1.2	0.8	0.4	0.6	-1.2	0.15	61	71.0	79	41	52	64	24	35	47
17	0.8	0.2	0.4	1.4	0.6	-1.2	0.15	61	70.0	79	41	52	63	23	35	49

Table C.3: Combinations of parameter values that provide appropriate summary results.

combinations (Table C.3). From these 17 combinations, we chose line 17 for our true parameter values because both the effects β_{1MU} and β_{1BU} are relatively small when compared to the values in the other candidate sets. Smaller effect sizes better mirror processes that may occur in reality.

For these candidate values on line 17, we repeated the simulation 1000 times to verify the ranges observed, holding the covariates X fixed. We also generated a new set of covariates for each simulation to assess the sensitivity of the number reporting to the covariate values. Moreover, we calculated the marginal proportion of false zeros in the data set as well as the conditional proportion of false zeros in the data set. A false zero is defined as when $Y_{iTRUE} > 0$ and $Z_{iREP} = 0$. The *marginal* proportion of false zeros is calculated as the number of false zeros in the data set divided by the

		Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Fixed	No. MU+	59	68	70	70.8	74	82
Covariates	No. BU+	37	48	51	51.3	55	64
	No. Rep+	22	32	34	34.1	37	46
	Marginal prop. false zeros	0.09	0.13	0.15	0.16	0.17	0.23
	Conditional prop. false zeros	0.13	0.21	0.23	0.24	0.27	0.35
Random	No. MU+	57	68	71	71.2	74	83
Covariates	No. BU+	35	48	52	51.7	55	69
	No. Rep+	20	32	35	34.7	38	51
	Marginal prop. false zeros	0.05	0.13	0.15	0.16	0.18	0.27
	Conditional prop. false zeros	0.07	0.20	0.24	0.24	0.28	0.45

Table C.4: Final assessment of parameter values to ensure appropriate data.

total number of observations, $\frac{\sum_{i=1}^{n} \text{I(false zero)}}{n}$. The *conditional* proportion of false zeros is calculated as $\frac{\sum_{i=1}^{n} \text{I(false zero)}}{\sum_{i=1}^{n} \text{I}(Y_{iOBS} = 0)}$, the total number of false zeros divided by the total number of observed zeros. The simulation results do not appear to be too sensitive to the covariate values as they show similar values for when the covariates X are fixed for the entire simulation and for when the covariates X are random for each simulation (Table C.4).

We can summarize each data set in terms of the five types of observations defined in Section 4.4. For example, in the first 20 data sets generated, Table C.5 shows the distribution of the types of observations, as well as the conditional proportion of false zeros.

	Excess true zero (MU)	Excess true zero (BU)	Distn'l true zero	Excess false zero	Observed case count	Cond. prop. false zero
1	28	13	3	15	41	0.25
2	27	26	1	14	32	0.21
3	25	18	3	19	35	0.29
4	28	22	3	11	36	0.17
5	33	20	4	20	23	0.26
6	29	15	1	12	43	0.21
7	28	21	0	19	32	0.28
8	25	12	4	17	42	0.29
9	26	25	0	15	34	0.23
10	29	20	1	17	33	0.25
11	35	16	2	15	32	0.22
12	28	18	3	17	34	0.26
13	33	26	1	8	32	0.12
14	30	21	1	14	34	0.21
15	26	19	3	10	42	0.17
16	18	21	0	18	43	0.32
17	35	20	0	10	35	0.15
18	27	18	2	14	39	0.23
19	25	23	3	9	40	0.15
20	32	17	2	16	33	0.24

Table C.5: Summary of 20 randomly generated data sets that shows the distribution of each of the five types of observations defined in 4.4, as well as the conditional proportion of false zeros out of all observed zeros. Each row is a different data set, and the numbers in each column represent the number of observations out of 100 that fell in that category.

Appendix D

WinBUGS code for ZIP Models

WinBUGS code for the naive traditional ZIP model, Model 4.18.

```
model{
  for (i in 1:n){
    Y.obs[i]~dpois(mu.obs[i])
    Z.excess[i]~dbern(p.excess[i])
    mu.obs[i] <- (1-Z.excess[i])*lambda[i]
    log(lambda[i]) <- b0.cases + b1.cases*x.cases[i]
    logit(p.excess[i]) <- b0.p + b1.p.mu*x.mu[i] + b2.p.bu*x.bu[i]
    #the probability that an observation takes that specific
    #value given Poisson(lambda)
    f[i]<- exp( -lambda[i] + Y.obs[i]*log(lambda[i]) - loggam(Y.obs[i]+1) )
    #log likelihood
    ll[i]<-log(  p.excess[i]*equals(Y.obs[i],0) + (1-p.excess[i])*f[i]  )
    #the probability of a distributional zero under a Poisson distribution
    #for count data
    f.0[i]<-exp(-lambda[i])
    #the conditional probability of observing an excess zero given
    #that a zero was observed
    p.0.cond[i]<- p.excess[i]/(p.excess[i]+(1-p.excess[i])*f.0[i])
    }
  b0.cases ~ dnorm(0.0,1.0E-2)
  b1.cases ~ dnorm(0.0,1.0E-2)
  b0.p ~ dnorm(0.0,1.0E-2)
  b1.p.mu ~ dnorm(0.0,1.0E-2)
  b2.p.bu ~ dnorm(0.0,1.0E-2)
  lambda0<-exp(b0.cases)
  lambda1<-exp(b0.cases + b1.cases)
  my.dev<- -2*sum(ll[1:n])
}
```

WinBUGS code for almost fully specified hierarchical ZIP model, Model 4.19.

```
model{
  for (i in 1:n){
   #the distribution of Z.MU is not conditioned on anything
   Z.MU[i]~dbern(p.MU[i])

   #the latent random variable Z.BU is conditioned on Z.MU
   p.BU.Z[i]<-p.BU[i]*Z.MU[i]
   Z.BU[i]~dbern(p.BU.Z[i])

   #the latent random variable Z.REP is conditioned on Z.BU
   p.REP.Z[i]<-p.REP[i]*Z.BU[i]
   Z.REP[i]~dbern(p.REP.Z[i])

   #the observed Y follows a Poisson distribution conditioned on Z.REP
   mu.obs[i] <- lambda[i]*Z.REP[i]
   Y.obs[i]~dpois(mu.obs[i])

   log(lambda[i]) <- b0.cases + b1.cases*x.cases[i]
   logit(p.MU[i]) <- b0.mu    + b1.mu*x.mu[i]
   logit(p.BU[i]) <- b0.bu    + b1.bu*x.bu[i]
   logit(p.REP[i]) <- b0.rep  + b1.rep*x.rep[i]

   #probability of excess true zero (MU)
   p.f1[i] <- 1 - p.MU[i]
   #probability of excess true zero (BU)
   p.f2[i] <- p.MU[i] * (1 - p.BU[i])
   #probability of zero under Poisson distribution
   f.0[i]  <- exp(-lambda[i])
   #probability of distributional true zero
   p.f3[i] <- p.MU[i] * p.BU[i] * f.0[i]
   #probability of excess false zero
   p.f4[i] <- p.MU[i] * p.BU[i] * (1 - f.0[i]) * (1-p.REP[i])
   #conditional probability of false zero given that a zero was observed
   p.false.zero[i] <- p.f4[i]/(p.f1[i] + p.f2[i] + p.f3[i] + p.f4[i])

   #the probability that an observation takes that specific value
   #given Poisson(lambda)
   f[i]<- exp( -lambda[i] + Y.obs[i]*log(lambda[i]) - loggam(Y.obs[i]+1) )
   p.all[i] <- p.MU[i] * p.BU[i] * p.REP[i]
   #log likelihood
   ll[i]<-log( equals(Y.obs[i],0)*(1 - p.all[i]) + p.all[i]*f[i]  )
   }
  b0.cases ~ dnorm(0.0,1.0E-2)
  b0.mu ~ dnorm(0.0,1.0E-2)
  b0.bu ~ dnorm(0.0,1.0E-2)
  b0.rep ~ dnorm(0.0,1.0E-2)
  b1.cases ~ dnorm(0.0,1.0E-2)
  b1.mu ~ dnorm(0.0,1.0E-2)
  b1.bu ~ dnorm(0.0,1.0E-2)
  b1.rep ~ dnorm(0.0,1.0E-2)

  my.dev<- -2*sum(ll[1:n])
}
```

WinBUGS code for the fully hierarchical ZIP model, Model 4.20.

```
model{
  for (i in 1:n){
    #the distribution of Z.MU is not conditioned on anything
    Z.MU[i]~dbern(p.MU[i])

    #the latent random variable Z.BU is conditioned on Z.MU
    p.BU.Z[i]<-p.BU[i]*Z.MU[i]
    Z.BU[i]~dbern(p.BU.Z[i])

    #the unobserved true Y follows a Poisson distribution conditioned on Z.BU
    mu.true[i] <- lambda[i]*Z.BU[i]
    Y.true[i]~dpois(mu.true[i])

    #the latent random variable Z.REP is conditioned on Y.true>0
    #This creates an indictor such that I.Ytrue=1 when Y.true>0
    #equals(arg1,arg2)=1 when arg1=arg2
    I.Ytrue[i] <- 1 - equals(Y.true[i],0)
    p.REP.Z[i]<-p.REP[i]*I.Ytrue[i]
    Z.REP[i]~dbern(p.REP.Z[i])

    #the observed Y follows a Poisson distribution conditioned on Z.REP
    mu.obs[i] <- lambda[i]*Z.REP[i]
    Y.obs[i]~dpois(mu.obs[i])

    log(lambda[i])  <- b0.cases + b1.cases*x.cases[i]
    logit(p.MU[i]) <- b0.mu    + b1.mu*x.mu[i]
    logit(p.BU[i]) <- b0.bu    + b1.bu*x.bu[i]
    logit(p.REP[i])<- b0.rep   + b1.rep*x.rep[i] + b2.rep*Y.true[i]

    #probability of excess true zero (MU)
    p.f1[i] <- 1 - p.MU[i]
    #probability of excess true zero (BU)
    p.f2[i] <- p.MU[i] * (1 - p.BU[i])
    #probability of zero under Poisson distribution
    f.0[i]  <- exp(-lambda[i])
    #probability of distributional true zero
    p.f3[i] <- p.MU[i] * p.BU[i] * f.0[i]
    #probability of excess false zero
    p.f4[i] <- p.MU[i] * p.BU[i] * (1 - f.0[i]) * (1-p.REP[i])
    #conditional probability of false zero given that a zero was observed
    p.false.zero[i] <- p.f4[i]/(p.f1[i] + p.f2[i] + p.f3[i] + p.f4[i])

    #the probability that an observation takes that specific value
    #given Poisson(lambda)
    f[i]<- exp( -lambda[i] + Y.obs[i]*log(lambda[i]) - loggam(Y.obs[i]+1) )
    p.all[i] <- p.MU[i] * p.BU[i] * p.REP[i]
    #log likelihood
    ll[i]<-log( equals(Y.obs[i],0)*(1 - p.all[i]*(1-f.0[i]))
                 + (1-equals(Y.obs[i],0))*p.all[i]*f[i]  )
     }
  b0.cases ~ dnorm(0.0,1.0E-2)
  b1.cases ~ dnorm(0.0,1.0E-2)
  b0.mu ~ dnorm(0.0,1.0E-2)
  b1.mu ~ dnorm(0.0,1.0E-2)
  b0.bu ~ dnorm(0.0,1.0E-2)
  b1.bu ~ dnorm(0.0,1.0E-2)
  b0.rep ~ dnorm(0.0,1.0E-2)
  b1.rep ~ dnorm(0.0,1.0E-2)
  b2.rep ~ dnorm(0.0,1.0E-2)
  my.dev<- -2*sum(ll[1:n])
}
```