# DETECTING AND CONTROLLING INSECT VECTORS IN URBAN ENVIRONMENTS: NOVEL 

 BAYESIAN METHODS FOR COMPLEX SPATIAL DATAErica Marie Weinmann Billig
A DISSERTATION
in
Epidemiology and Biostatistics
Presented to the Faculties of the University of Pennsylvania
in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
2017

Supervisor of Dissertation

Jason A. Roy
Associate Professor of Biostatistics

Co-Supervisor of Dissertation

Michael Z. Levy
Associate Professor of Epidemiology

Graduate Group Chairperson

Nandita Mitra, Professor of Biostatistics

Dissertation Committee
Michelle E. Ross, Assistant Professor of Biostatistics
Ricardo Castillo Neyra, Scientific Director, Universidad Peruana Cayetano Heredia
F. Ellis McKenzie, Senior Scientist, Fogarty International Center (in memoriam)

DETECTING AND CONTROLLING INSECT VECTORS IN URBAN ENVIRONMENTS: NOVEL BAYESIAN METHODS FOR COMPLEX SPATIAL DATA
(C) COPYRIGHT

2017

Erica Marie Weinmann Billig

This work is licensed under the
Creative Commons Attribution
NonCommercial-ShareAlike 3.0
License

To view a copy of this license, visit
http://creativecommons.org/licenses/by-nc-sa/3.0/

## ACKNOWLEDGEMENT

The writing of this dissertation would not have been possible without the support and insight of many others. I would first like to thank my advisors, Dr. Jason Roy and Dr. Michael Levy, for their dedication, advice, patience, and guidance over the years. I am grateful that even though they had not worked together before, they agreed to co-advise me. They spent much time and energy working with me on my dissertation, and were patient and kind as I made mistakes. Every time I reached an obstacle, they helped me find a way to a solution. I truly believe that because of them, I have continued to enjoy working on this project.

I would also like to thank my committee, Dr. Michelle Ross, and Dr. Ricardo Castillo Neyra, for their ideas and help along the way. Dr. Ross met with me frequently to teach me spatial statistics. She had insights that yielded results included here. Dr. Castillo Neyra knows these data better than anyone, and he readily answered my questions that helped me understand the content of the project. I especially want to thank the late Dr. F. Ellis McKenzie, with whom I worked for two years at the Fogarty International Center at NIH. He was a wise and thoughtful mentor, a brilliant scientist, a role model, and initial member of my committee. Although he was not able to see the final result, this dissertation bears his mark. His influence is evident not only here, but through the work of his many students.

I would like to thank Dr. Christopher Hunter and the support of the Parasitology Training Grant, 5T32AI007532. The grant has allowed me to focus my time on my dissertation, and supported me to present the results at several conferences. I would like to thank the team in Peru that has spent so much time collecting, managing, and studying these data : Ministerio de Salud del Perú (MINSA), the Dirección General de Salud de las Personas (DGSP), the Estrategia Sanitaria Nacional de Prevención y Control de Enfermedades Metaxénicas y Otras Transmitidas por Vectores (ESNPCEMOTVS), the Dirección General de Salud Ambiental (DIGESA), the Gerencia Regional de Salud de Arequipa (GRSA).

I would like to thank the faculty, staff, and students at the University of Pennsylvania in the Department of Epidemiology and Biostatistics. I am especially grateful to Dr. Justine Shults, Dr. Mary Sammel, Dr. Mary Putt, and Dr. Scarlett Bellamy, who provided me with invaluable guidance and advice at critical times. I would also like to thank Kwonsang Lee for his help with the simulations
presented in Chapter 3. I feel very fortunate that my graduate education has enabled me to connect with many wonderful mentors, colleagues, and friends.

Lastly, I would also like to thank my family and friends, who were so supportive as I have pursued my PhD. My parents, Dr. Gail Weinmann and Dr. Nathan Billig, have instilled in me a love of science, and continued to encourage me to find a way to reach my potential. My fiancé, Jason Rose, has been by my side since I began at Penn, and has always encouraged and believed in me, even when I was unsure of myself.


#### Abstract

DETECTING AND CONTROLLING INSECT VECTORS IN URBAN ENVIRONMENTS: NOVEL BAYESIAN METHODS FOR COMPLEX SPATIAL DATA

Erica Marie Weinmann Billig

Jason A. Roy

Michael Z. Levy

Efforts to control the spread of vector-borne diseases often focus on the vector itself. Here, we develop novel methods to strategically guide the search for vectors over urban landscapes. The methodology is motivated by Triatoma infestans, the vector of Chagas disease, a re-emerging vector in Arequipa, Peru. We first propose a novel stochastic epidemic model that incorporates both the counts of disease vectors at each observed house and the complex spatial dispersal dynamics. The goal of our analysis is to predict and identify houses that are infested with $T$. infestans for entomological inspection and insecticide treatment. A Bayesian method is used to augment the observed data, estimate the insect population growth and dispersal parameters, and determine posterior infestation probabilities of households. We investigate the properties of the model through simulation studies and implement the strategy in a region of Arequipa by inspecting houses with the highest posterior probabilities of infestation and report the results from the field study. After piloting this model in the field and assessing the strengths and weaknesses, we propose a much faster method that extends a Gaussian Field (GF) model to incorporate the urban landscape. GF models can be used to create risk maps of vector presence across large urban environments. However, these models do not typically account for the possibility that city streets function as permeable barriers for insect vectors. We extend GF models to account for this urban landscape. We demonstrate our method on simulated datasets and then apply it to data on T. infestans. We estimate that streets increase the effect of distance on the probability of vector presence at least 1.5 fold compared to the undivided environment. Lastly, we propose a Bayesian generalized multivariate conditional autoregressive approach to jointly model the distribution of vectors, T. infestans, with the proportion of vectors that carry the parasite of Chagas disease, Trypanosoma cruzi. We demonstrate the properties of the model using simulation studies, and apply the method to data from Arequipa.


## TABLE OF CONTENTS

ACKNOWLEDGEMENT ..... iii
ABSTRACT ..... v
LIST OF TABLES ..... viii
LIST OF ILLUSTRATIONS ..... x
CHAPTER 1: Introduction ..... 1
1.1 Vector-borne diseases ..... 1
1.2 Motivation: Chagas disease ..... 1
1.3 Statistical background and developments ..... 2
Chapter 2: Real-Time Epidemiological Search Strategy ..... 6
2.1 Introduction ..... 6
2.2 Methods ..... 7
2.3 Simulations ..... 17
2.4 Data Results ..... 18
2.5 Discussion ..... 20
CHAPTER 3: RISk maps for cities: Incorporating streets into geostatistical models ..... 23
3.1 Introduction ..... 23
3.2 Methods ..... 24
3.3 Simulations ..... 30
3.4 Data Results ..... 34
3.5 Discussion ..... 37
CHAPTER 4: Generalized Multivariate Conditional Autoregressive Model for Vector-borne Disease Data ..... 39
4.1 Introduction ..... 39
4.2 Methods ..... 41
4.3 Simulations ..... 47
4.4 Data Results ..... 48
4.5 Discussion ..... 50
CHAPTER 5: Discussion ..... 55
5.1 Conclusions ..... 55
5.2 Limitations ..... 57
5.3 Future directions ..... 58
APPENDICES ..... 62
BIBLIOGRAPHY ..... 68

## LIST OF TABLES

TABLE 2.1 : Simulation results. Each estimate represents the mean of given quartile for the set of 200 replicates. We use the median estimate to verify our approach obtains accurate parameter estimates and report the median AUC of each parameter set.
TABLE 3.1 : Description of map distortion in Figure 3.1 scaled so the spatial unit is the average distance between nearest neighbors on the same block. Interpretation of $S$ varies by specific map due to variability in sizes and shapes of city blocks. $S$ describes ratio of distance between geographic median of each city block relative to the true map (which is equivalent to $S=1$ ). Table summarizes how this distortion corresponds to additional distance between houses on different blocks using mean and standard deviation ( $s d$ ). The distortion varies block by block due to irregular grid. .
TABLE 3.3 : Difference in number of positive houses in the top $30 \%$ of probabilities of infestation using our approach of map distortion compared to using the true map with no distortion ( $S=1$ ). Model fit with true values of $S=1, S=2.5$, and $S=4$ when a randomly selected one-third of points were observed. Intercept fixed at $\beta_{0}=-5$. 100 simulated datasets were run at each value. We report number of positive houses in the top $30 \%$ of probabilities which were treated as unobserved (ie. no gain for houses that were observed as positive).33

TABLE 4.1: Median estimates with median of the $95 \%$ credible interval of 40 simulated
datasets. ..... 48
TABLE 4.2 : Data results using three models (Eq. 4.1, 4.2, 4.3). The full model allows all parameters to vary. The partial model allows $\eta_{20}$ and $\eta_{30}$ to vary, but fixes $\eta_{21}=0$ and $\eta_{31}=0$. The independent model fixes all linking parameters, $\boldsymbol{\eta}=0$. In all models, covariables are the same: $\beta_{i 0}$ corresponds to the model intercept, $\beta_{i 1}$ is the estimated effect of guinea pigs (cuy), $\beta_{i 2}$ is the estimated effect of dogs, $\beta_{i 3}$ is the estimated effect of poultry, $\beta_{i 4}$ estimates the effect of the presence of housing materials that are good habitats for vectors, and $\beta_{i 5}$ estimates the effect of the presence of housing materials that are poor habitats for vectors for $i=1,2,3$. Coefficient credible intervals of covariables that did not contain 0 are highlighted in red. ..... 50
TABLE B.1: Simulation results with the variation in intercept $\beta_{0}$. ..... 65
TABLE B. 2 : Simulation results with the variation in intercept $\kappa$. ..... 66
TABLE B. 3 : Simulation results with the variation in intercept $\beta_{0}$. ..... 66
TABLE B. 4 : Results of grid sampling $S$ on dataset, including estimates and standard deviations of $\theta_{1}, \theta_{2}$, and $\beta_{0}$. The log-likelihood is maximized when $S=1.5$, and the bolded results are reported in the main text. ..... 69

## LIST OF ILLUSTRATIONS

FIGURE 2.1: ROC curves describing the ability of the RJMCMC to uncover unreported infestations in a simulated vector-borne epidemic. In each simulation, $1 / 3$ of infestations were randomly removed to recover. 50 randomly selected ROC curves of each parameter regime are shown. (a) $\{\beta, r\}=\{0.02,1.6\}$ (b) $\{\beta, r\}=\{0.7,1.8\}$ (c) $\{\beta, r\}=\{0.3,2.1\}$

FIGURE 2.2 : Posterior estimate of $\beta$, the probability of successful invasion of a migrating bug, at the beginning of the inspection campaign in each community, indicating median and $95 \%$ credible interval of RJMCMC chains: (a) Community $1: 0.67(0.33,0.91)$ (b) Community $2: 0.59(0.22,0.90)$ (c) Community 3: $0.51(0.15,0.85)$
FIGURE 2.3 : Path of inspectors in study region from October 2015 through December 2015. Sites of attempted inspections (blue circles) and successfully inspected houses (black circle) shown, as well as daily routes (blue lines). Previous confirmed infestations (red plus signs) also shown. Points are jittered to de-identify data.

FIGURE 3.1 : A) $S=1$, which corresponds to the true map of section in Mariano Melgar, Arequipa, Peru used for simulations. There is no map distortion at this scale. B) Distorted map to a scale of $S=1.5$, where the distance between geographic medians of blocks are 1.5 fold the true distance. C) Map distorted to scale $S=2.5$, where distance between geographic medians of blocks are 2.5 fold the true distance. Note distance between houses within block is maintained but distance between houses on different blocks is stretched.
FIGURE 3.2 : In this figure, we demonstrate how streets between larger blocks result in a larger barrier effect. A. Four hypothetical blocks super-imposed. The smallest block is simply one house. The largest block has an X distance of 4 from the block median to the edge of the block. B. Same hypothetical blocks after distortion with $S=2$. C. The ratio of street width for this distortion at $S=2$ compared to no distortion at $S=1$. The street width doubles if the block is a single house and the distance from block median to edge is zero. Otherwise, the street width more than doubles, and the effect increases as the block size increases.
FIGURE 3.3 : Mesh over the map of the simulation region (houses in blue). Maximum edge length is constrained to $100 S$ to keep meshes consistent between scales. A. $S=1$ B. $S=1.5$
FIGURE 3.4 : Comparing the estimated Matérn covariance function (gray) with the true Matérn covariance function (black) under four parameter sets.32

FIGURE 3.5 : Three unidentifiable and identifiable log-likelihoods and the corresponding simulated datasets. Unidentifiable landscapes were uncommon, (rates varied based on the true parameter values of $\kappa$ and $\sigma_{u}^{2}$ ) and in most cases had scarce infestations (panels a and c). Occasionally, an unidentifiable landscape was oversaturated and also unidentifiable (panel b). For comparison, most simulated datasets were identifiable with clear maximums of the log-likelihoods (panels d, e, and f)

FIGURE 3.6 : Map of the study region, the district of Mariano Melgar, Arequipa, Peru, which consists of 12,069 houses and 724 blocks. Color corresponds to number of known infested houses on the block.

35
FIGURE 3.7 : Posterior distributions of estimated parameters when $S=1.5$. A. Posterior distribution of $\sigma_{u}^{2} \mathrm{~B}$. Posterior distribution of $\kappa$. C. Estimated posterior distribution of Matérn covariance, as a function of distance. For reference, when the map is scaled to $S=1.5$, the average distance between nearest neighbors on the same block is $10.2(s d=5.5)$ and the average distance between nearest neighbors on different blocks is 62.4. $(s d=18.0)$. . . .
FIGURE 3.8 : Risk map of predicted probabilities of infestation using A) $S=1$ (true map) B) $S=1.5$ and C) $S=3$. The last panel shows differences in risk between scales of the area enclosed in the black rectangle in more detail. The color scale shows $P$ (infestation) and ranges from 0.74 (red) to 0.00 (purple) .

FIGURE 4.1: Data used in analysis. Dataset contains 577 sites (contained within 67 houses; house level data not shown). Sites with vectors (red) and number of vectors (size of point on log scale) shown with number of vectors that tested positive for T. cruzi (size of yellow point on log scale).
FIGURE 4.2 : Unfilled gray points indicate all sites. (a) Expected probability of vector presence at each site (size of red point relative to gray point) (b) Expected count of vectors, conditional on the probability of vector presence (size proportional to log count) (c) Expected proportion of positive vectors, conditional on expected vector distribution (size of gold point relative to red point). Red point is relative to the size of the log of the number of observed T. infestans at that site.

FIGURE A.1 : Bounded growth using the Beverton-Holt model compared to unbounded $\begin{aligned} & \text { growth. Difference in slopes at time of new infestation was incorporated } \\ & \text { into hazard function to quantify infestation severity into the probability of } \\ & \text { infesting a neighboring house. . . . . . . . . . . . . . . . . . . . . } 62\end{aligned}$
FIGURE A.2 : Ranking of each house across 3 RJMCMC chains. There were a few houses that changed significantly, but most houses remained within a few rankings between chains. We use the median ranking between chains to give inspectors.

FIGURE A. 3 : Ranking of each house across 5 potential carrying capacities. There were
a few houses that changed significantly, but most houses remained within
a few rankings between carrying capacity values. ..... 64

FIGURE B.1: Log-likelihood analysis across different scales, $S$. In most cases, $S$ is
clearly identifiable, but in some cases (red rectangles) it is not. ..... 67
FIGURE B. 2 : Log-likelihood analysis across different scales, $S$. Likelihood is maximized at $S=1.5$ indicating the model of best fit. ..... 68

## CHAPTER 1

## INTRODUCTION

### 1.1. Vector-borne diseases

Vector-borne diseases are increasingly common in urban areas, and efforts to control these diseases are often targeted at the vector itself. However, detecting populations of disease vectors in large urban environments is especially complex (Weaver, 2013, Knudsen and Slooff, 1992). Poor and unplanned urban environments can create ideal breeding grounds for many vectors, facilitating increased transmission of vector-borne diseases in population dense areas (Knudsen and Slooff, 1992, Bowman et al., 2008, Levy et al., 2006). Several arboviruses, including Dengue, Chikungunya, West Nile, and Zika, have emerged repeatedly in urban areas (Haley, 2012, Sikka et al., 2016). Parasitic diseases, such as malaria and Chagas disease, once considered rural problems, have become common in cities (LaDeau et al., 2015, Delgado et al., 2013). In this dissertation, we develop methods for real-time surveillance of the re-emerging vector of Chagas disease, Triatoma infestans, in Arequipa, the second largest city in Peru.

### 1.2. Motivation: Chagas disease

Chagas disease has long been endemic in Central and South America and is caused by the parasite Trypansoma cruzi (Mathers, Fat, and Boerma, 2008, Bern, 2015, Rassi and Marin-Neto, 2010). Once infected with the parasite, the host will first suffer from the acute phase of the disease, which presents symptoms like that of the flu. Infected individuals may suffer symptoms such as fever, swollen lymph nodes, and, rarely, the classic sign of Chagas, Romana's sign, where the periorbita becomes inflamed (Bern, 2015). The acute phase typically lasts $4-8$ weeks, at which point the disease enters the chronic phase. During this phase, the patient is unlikely to have symptoms. Approximately 20-30\% of individuals in this phase will suffer a cardiac event over the course of a lifetime, while others will enter a symptomatic chronic phase, which affects the nervous, digestive, and cardiac systems (Bern, 2015). One of the challenges of assessing the burden of Chagas disease lies in the complex and wide range of symptoms. Patients often go undiagnosed and deaths are mis-attributed.

An estimated 5.7 million people are currently living with Chagas disease (Bern, 2015). The prevalence has dropped substantially over the last few decades, largely due to a successful public health campaign to control vector populations (Dias, Silveira, and Schofield, 2002). Arequipa has only one vector of $T$. cruzi transmission, $T$. infestans, a species of triatomine that thrives in urban settings. The species prefers environments such as guinea pig pens, common in Peru, and housing materials with dark cracks and crevices (Levy et al., 2006). Since the insect rarely flies, there is a highly spatial aspect to the observed vector distribution patterns. Previous studies have shown that the vectors are much more likely to move within city blocks than cross a street (Barbu, Dumonteil, and Gourbière, 2010, Barbu et al., 2013).

To control the spread of Chagas disease, Arequipa began an inspection and spray campaign to target T. infestans in 2003 (Barbu et al., 2014). During the first phase of the campaign, preliminary inspections were conducted on a locality level, followed by a treatment phase during which insecticide was applied to the targeted areas. These phases were followed by surveillance, during which residents report infestations, which are followed-up by an inspection to that location, and surrounding locations if a true vector is found. From the spray campaign and subsequent surveillance, detailed data was collected on vector prevalence, including the date, location, and number of vectors observed at houses throughout the city. We are motivated to use this data to guide a new campaign to target the vectors that have re-emerged since the initial campaign - we want to inspect high risk houses for vectors, and update the risk as inspections are completed.

Here, we develop methods to search urban environments for disease vectors, and apply them to the ongoing efforts to eliminate the spread of $T$. infestans, in Arequipa, Peru. Different from other types of search strategies, where the coveted item is not moving or spreading, controlling the spread of disease vectors in real time requires identifying and treating infested households before they move or infest others. We also present a method to estimate the proportion of vectors infected with $T$. cruzi, conditional on the distribution of $T$. infestans. Our methods expand upon existing Bayesian spatial, temporal, and epidemic modeling tools.

### 1.3. Statistical background and developments

In the second chapter, we propose a novel stochastic compartmental model to estimate the probability of infestation, in the context of the re-emerging Chagas disease vector, Triatoma infestans,
in Arequipa, to guide our inspection strategy. Stochastic epidemic models are a popular tool to describe the course of an infectious disease epidemic. Using this approach, we treat the spread of $T$. infestans through houses in the city like an infectious disease spreading through a population of individuals. Fitting stochastic epidemic models to data is challenging due to the detailed data necessary to retain a tractable likelihood (Andersson and Britton, 2012). We use a susceptible-infected-removed (SIR) model: every house in the system is in one of these three states (susceptible, infected, or treated) at any given time point, and parameters describe the rates of transition between states. However, the likelihood of this model includes the time of infestation and treatment of every house (Becker, 1989). In practice, we do not observe the true infestation time of each house, and many houses in the system have never been inspected. O'Neill et al. developed an approach to augment the data and retain a tractable likelihood when the infestation times and total number of infested houses are unobserved (O'Neill and Roberts, 1999). His key development is the use of a reversible-jump Markov chain Monte Carlo, an estimation algorithm used when the number of parameters is unknown. In this case, the total number of infested houses is unknown, and thus there is an unknown number of unobserved infestation times, which are treated as parameters in the model.

We expand off more recent methods, Jewell et al., 2009a and Jewell et al., 2009b, that incorporate the 'notification time' of the house, or the time that the house is observed as infested into O'Neill's approach. Jewell's method uses a susceptible-infected-notified-removed (SINR) model and therefore has additional parameters to capture the transition from the infected to notified, and notified to removed, states of the system. The approach was developed in the context of notifiable diseases, and the critical assumption necessary for a tractable likelihood of the SINR model, is a known distribution that captures the true infestation time as a function of the later, observed notification time. Due to the nature of how our data was collected, as part of both the spray campaign and subsequent surveillance, we cannot make that key assumption.

We incorporate a house-level vector population growth model and use the counts of disease vectors at each observed house to estimate the unobserved infestation time, by assuming houses with more observed insects were infested less recently than houses with fewer insects. In addition, we attempt to capture the complex spatial dispersal dynamics by using a kernel that incorporates city streets (Barbu et al., 2013). The goal of our analysis is to predict and identify houses that
are infested with T. infestans for entomological inspection and insecticide treatment. Our Bayesian method is used to augment the observed data, estimate the insect population growth and dispersal parameters, and determine posterior infestation probabilities of households. We investigate the properties of the model through simulation studies. We implement the strategy in a region of Arequipa by inspecting houses with the highest posterior probabilities of infestation and report the results from the field study. After implementing this model in the field, we identify strengths and weaknesses of this approach. While the model captures the spatial heterogeneity and dynamic movement of the vectors, the reversible-jump Markov chain Monte Carlo used for estimation is computationally intensive and does not consistently converge quickly enough for real-time updates.

In the third chapter, we are motivated to develop a model to improve upon the weaknesses we encountered in our field implementation, including speed of convergence, while estimating the effect of the urban landscape on the vector distribution. We use Gaussian field (GF) model, a type of spatial model used for point-process, or geostatistical, data. Recent advances, both in theory and computation, have enabled fast parameter estimation using this approach. Lindgren et al. found a link between the GF and Gaussian Markov Random Field (GMRF) with the use of the Matérn covariance function, popular for geostatistical data (Lindgren, Rue, and Lindström, 2011). Using this link, we can estimate the parameters of the Matérn covariance function using nested integrated Laplace approximations. Rue et al. developed a sophisticated R package, 'INLA', to implement this estimation algorithm, an alternative to Markov chain Monte Carlo, greatly easing the speed and computational burden (Rue, Martino, and Chopin, 2009). Using a GF, we assume the observed vector distributions are a realization of a stochastic process that is occurring over the landscape. However, the GF assumes the process is occurring over a continuous, smooth, space, but the city streets of the urban landscape create barriers for vector dispersal. Unlike true barriers, vectors can cross city streets, but previous studies have shown that $T$. infestans are more likely to move a given distance within a city block than between city blocks (Barbu et al., 2013). We propose using an additional parameter to capture the heterogeneity of the city landscape, by effectively distorting the city map and widening the streets by an estimated distance, creating permeable barriers. Our approach fits into the existing estimation software, making it accessible for researchers across disciplines, and easy to apply to different districts of Arequipa while maintaining the ability to regularly update the model for 'real-time' estimation. Including the spatial structure of city streets may more accurately describe spatial associations characteristic of insect infestations and therefore help in
developing effective public health interventions to reduce transmission in population dense areas (Weaver, 2013). We demonstrate our method on simulated datasets and then apply it to data on Triatoma infestans, the principal vector of Chagas disease in Arequipa Peru.

In the fourth chapter, we develop a multivariate spatial model to study the proportion of vectors infested with the parasite, T. cruzi, conditional on the count and spatial distribution of the population of vectors themselves. In this approach, we build off of the traditional conditional autoregressive (CAR) spatial model (Besag, 1975). Jin et al. developed a bivariate generalized conditional autoregressive model that uses linking parameters to create dependencies between spatial models (Jin, Carlin, and Banerjee, 2005). We extend this model to link three CAR models: a logistic model of the proportion of $T$. cruzi infected vectors, and the two pieces of the zero-inflated Poisson mixture model of the vector distributions. This approach has an inherent order in the structure and interpretation. We first model the probability of vector presence, including a spatial random effect. We then model the count of vectors, with a different spatial random effect, conditional on the probability of vector presence. These two models together are the zero-inflated Poisson model and are linked through the multivariate distribution of the random effects. We then model the proportion of vectors infested with $T$. cruzi conditional on the zero-inflated Poisson model, and estimate the parameters linking these three pieces. Again, this logistic model is linked to the zero-inflated Poisson model through the multivariate distribution of the random effects. We fit the model using OPENBugs, verify the model on simulated data, and apply the data to an area in Arequipa, Peru.

Lastly, in the fifth chapter, we discuss our conclusions, on-going field work, and further directions for the study.

## CHAPTER 2

## Real-Time Epidemiological Search Strategy

### 2.1. Introduction

Over the past twenty years, the prevalence of Chagas disease, a deadly disease caused by the parasite Trypanosoma cruzi, has dropped substantially, largely due to successful campaigns to control insect vectors (Dias, Silveira, and Schofield, 2002, Tarleton et al., 2007). During this time, however, some of the species have invaded urban environments creating new control challenges (Longo and Bern, 2015, Levy et al., 2006). The city of Arequipa, Peru, is at the tail end of a campaign to eliminate T. infestans. Over 140,000 households have been treated with insecticide during the 'attack' phase of this effort. Following the 'attack' phase of this campaign, households then enter a community-based surveillance phase, during which residents are encouraged to report suspected infestations. Reported households are verified by a trained entomological inspector, and, if found to be positive, retreated with insecticide along with their immediate neighbors. To complement the community based efforts, trained vector control personnel proactively inspect houses where the infestation status is unknown. Thus, we are motivated to develop a search strategy that adapts to real time information on the distribution of vectors as we identify new infestations and treat older ones. It is in this context that we developed and fielded a method to proactively search a landscape for infested households.

Our models build on a rich literature of susceptible-infectious-removed (SIR) models. Likelihoodbased methods have been developed to fit SIR models to data when the infection times and total epidemic size are known (Becker, 1989, Andersson and Britton, 2012, Ross, 1996). These methods were extended, using a reversible-jump Markov Chain Monte Carlo algorithm (RJMCMC), to cases when these parameters are unobserved (Gibson and Renshaw, 1998, O'Neill and Roberts, 1999, O'Neill, 2002). Jewell et al. extended the methodology further to notifiable diseases by describing the length of time from the unobserved infection time to some later, observed 'notification' time (Jewell et al., 2009b, Jewell et al., 2009a). We extend these methods to both cross-sectionally and longitudinally collected data on insect populations. By adding structured growth of the insect populations, we are able to estimate the infectious period and thereby fit an SIR model to the
data. Our model incorporates both spatial heterogeneity and the severity of infestation into the transmission process.

We first present our model and methods, both theoretically and in the field, and verify the approach using simulations. We then discuss the successes and failures in its implementation on the spread of $T$. infestans in Arequipa, Peru.

### 2.2. Methods

We use a stochastic epidemic modeling approach to estimate the posterior probabilities of infestation of each house within a given region. Our SIR model is set on the household level. We begin by describing the observed data and some model assumptions. We then describe the model, its components, and the RJMCMC algorithm used for estimation.

At any given time point, a house may be in one of the following states:

1. Currently infested and not yet treated: We assume that these houses continue to be infested until treatment.
2. Previously infested and treated within the last $\mathbf{1 0 0}$ days: We assume these houses are removed from the system and not susceptible to re-infestation.
3. Previously infested and treated more than $\mathbf{1 0 0}$ days ago: We assume these houses are susceptible and may be infested any time after the effective interval of insecticide, thought to be 100 days post-treatment (Palomino et al., 2008).
4. Previously inspected and known to have been uninfested at some point in the past: We assume these houses are susceptible at any point after the previous inspection.
5. Never inspected and without known information: We assume these houses are susceptible at any time point since the initial infestation in the area.

We incorporate data from multiple sources into our approach. In some houses, a resident has reported an infestation, while in other houses inspectors have pro-actively searched during a surveillance campaign. In the vast majority of inspected houses no bugs are found; we assume the house is uninfested at the time of inspection, but may have been infested since.

Our model uses the number of bugs found in each inspected house at the most recent inspection, in combination with a household-level insect population growth model, to estimate the unobserved true infestation time. Insecticide treatment times were all observed and treatment was assumed to be 100\% effective.

### 2.2.1. Notation

Before introducing the model and likelihood, we define some notation:

- Define $I_{i}, D_{i}$, and $R_{i}$ as the infestation time, detection time, and treatment time of the $i$ th house, respectively
- Define $h_{i j}(t)$ as the probability that house $i$ infests house $j$ at time $t$
- Define $b_{i, t}$ as the number of bugs in the $i$ th house at time $t$, and $b_{i}$ is the set of all bug counts of house $i$ over all time points.
- Define $r$ as the growth rate of the insect population. We describe the growth rate as the bug population increase per time unit $t$ of 90 days. We may estimate $r$ from the data if there are enough observed infestations. Otherwise, we fix $r$ to values identified from previous studies (Rabinovich, 1972).
- Define $K$ as the carrying capacity, or the number of bugs a single household can ecologically support. We assume each house to have the same carrying capacity, and pick $K$ to be large enough that it is a reasonable estimate of a carrying capacity seen in Arequipa.
- Define $\lambda_{t}$ as the expected number of bugs at time $t$ given the time of infestation $I_{i}$ according to the assumed population growth model (we assume one insect at the time of infestation).
- Define $N, N_{I}$, and $N_{D}$ as the total number of houses, number of infested houses, and number of detected infested houses respectively at the current time (today), $T_{\max } . N$ and $N_{D}$ are observed data. However, $N_{I}$ is unobserved and is presumably larger than $N_{D}$.
- Define $\beta$ as the probability of successful invasion given a migrating bug. We estimate $\beta$ from the data.
- Define $t_{\text {insp, }, i}$ as the inspection time of the $i$ th house and $T_{\max }$ as the current time


### 2.2.2. Likelihood

Next, we introduce the complete-data likelihood (the likelihood if we observe infection status and bug counts at all houses at all times) and then go on to describe models for each part of the likelihood. We will then describe prior distributions and inference algorithms.

The complete-data likelihood is:

$$
\begin{aligned}
L(I, R, b \mid \Theta) \propto \prod_{j \neq k}^{N_{I}}( & \left.\sum_{I_{i}<I_{j}} h_{i j}\left(I_{j}\right)\right) \times \exp \left(-\int_{I_{k}}^{T_{\max }} \sum_{i=1}^{N_{I}} \sum_{j=1}^{N} h_{i j}(t) d t\right) \\
& \times \prod_{j=1}^{N_{I}} \prod_{t \geq I_{j}}^{\tau_{i}} \frac{\lambda_{t}^{b_{j, t}}}{b_{j, t}} \exp \left(-\lambda_{t}\right)
\end{aligned}
$$

where $\Theta=\left\{\beta, r, K, \lambda_{t}, I_{\kappa}\right\}, \tau_{i}=\min \left(R_{i}, T_{\max }\right)$ and $I_{\kappa}$ is the initial infestation time. The parameters $\beta$ and $r$ are unknown, and therefore will need to be assigned prior distributions. We fix $K$ and $I_{\kappa}$ to plausible field estimates. If we do not have many infestations in a region of the city, we may also fix $r$, but we demonstrate that our model is capable of estimating both $\beta$ and $r$ in our simulations. In our model, $R_{i}, T_{\text {max }}, N_{D}, N$ are observed and $N_{I}$ and $I_{j}$ is unobserved. We observe $b_{i, t_{\text {insp }}}$ but all other $b_{i, t}$ are unobserved and will be imputed using a household-level insect population growth model. $\lambda_{t}$ is estimated using the population growth model, described later.

The first piece of the likelihood describes the probability that the $i$ th house infests the $j$ th house at the time the $j$ th house was infested, $I_{j}$. The second piece of this likelihood describes the cumulative infectious pressure. The infectious pressure captures the effect over time of each infested house on every other uninfested house. In other words, at a given time $t$, a house that is surrounded by infested houses will be much less likely to escape infestation than a house that is surrounded by bug-free houses. This kind of pressure, over time, is captured by the second term in the likelihood. The third piece of the likelihood describes the probability of observing the number of bugs at each time point of each infested house, which we assume follows a Poisson distribution. Specific models for the transmission process $h_{i j}(t)$, which incorporates spatial heterogeneity, and the bug count rate $\lambda_{t}$, are described below.

It has previously been shown that the integral in the likelihood can be written in a simpler and
intuitive way (Jewell et al., 2009b):

$$
\int_{I_{\kappa}}^{T_{\max }} \sum_{j \in \mathcal{I}} \sum_{i \in \mathcal{S}} h_{i j}(t) d t=\sum_{i=1}^{N_{I}} \sum_{j=1}^{N} H_{i j}(t)=\sum_{i=1}^{N_{I}} \sum_{j=1}^{N} \sum_{s=1}^{t_{i j}} h_{i j}(s)
$$

where $t_{i j}=\min \left(R_{i}, I_{j}, T_{\max }\right)-\min \left(I_{i}, I_{j}\right)$. The likelihood then becomes:

$$
\begin{align*}
L(I, R, b \mid \Theta) & \propto \prod_{j \neq k}^{N_{I}}\left(\sum_{I_{i}<I_{j}} h_{i j}\left(I_{j}\right)\right) \times \exp \left(-\sum_{i=1}^{N_{I}} \sum_{j=1}^{N} \sum_{s=1}^{t_{i j}} h_{i j}(s)\right)  \tag{2.1}\\
& \times \prod_{j=1}^{N_{I}} \prod_{t \geq I_{j}}^{\tau_{i}} \frac{\lambda_{t}^{b_{j t}}}{b_{j t}} \exp \left(-\lambda_{t}\right) .
\end{align*}
$$

To account for the possibility that a given house $j$ is infested by two houses at the same time point, we calculate the probability that house $j$ is infested at time $t$ by calculating $1-P$ (house $j$ avoids infestation at time $t$ ):

$$
\begin{aligned}
p_{j, t} & =P(\text { house } j \text { becomes infested at time } t \mid \Theta) \\
& =1-P(\text { house } j \text { does not become infested at time } t \mid \Theta) \\
& =1-\prod_{i \neq j}\left(1-h_{i j}(t)\right) .
\end{aligned}
$$

In addition, since we are working in the Bayesian framework, we put prior distributions on the parameters that we are estimating. We chose a beta prior distribution on $\beta$ and a gamma prior distribution on $r$.

### 2.2.3. Bug infectiousness

We allow for heterogeneous transmission relative to the number of insects in each infested house. We assume that houses with more vectors are more likely to infest their neighbors than houses with only a few vectors. We characterize this heterogeneity in infectivity using a house-level population growth model. We use the Beverton-Holt model, which has two parameters: a house-level carrying capacity and bug growth rate (Beverton and Holt, 2012, Varley, Gradwell, and Hassell, 1974). The Beverton-Holt model is a key element of our extension that captures the distribution of the infectious
period.

The model is:

$$
\lambda_{t}=\frac{K \lambda_{0}}{\lambda_{0}+\left(K-\lambda_{0}\right) r^{-t}},
$$

where $\lambda_{0}$ is the number of bugs at the initial time, $\lambda_{t}$ is the expected number of bugs at time $t, K$ is the carrying capacity, and $r$ is the growth rate per generation. In our Chagas disease example, we assume that $\lambda_{0}=1$ and $K=1000$, implying that that each household infestation begins with one bug and has a carrying capacity of 1000 bugs. In truth, the carrying capacity is unknown, and we have conducted sensitivity analyses on this assumption (Appendix A.2). However, we believe this is a realistic estimate of carrying capacity because although we do find approximately 1000 bugs in a house, we rarely find more than that. We estimate $r$ through the RJMCMC algorithm.

Using this model, when there are an expected $\lambda_{t}$ insects in a given house, the time of the first bug can be solved for explicitly:

$$
t=\frac{-\log \frac{K-\lambda_{t}}{\lambda_{t}(K-1)}}{\log (r)} .
$$

We assume that the rate at which bugs migrate to find new houses is the difference between the unbounded and bounded growth rates. As described above, the expected bug population size at time $t$ under the bounded (Beverton-Holt) model is $\lambda_{t}$. The bounded growth rate is therefore $\frac{d \lambda}{d t}$. If there were unlimited resources, then bug populations would grow exponentially. Define $\lambda_{t}^{\prime}$ as the number of bugs expected at time $t$ assuming exponential growth, $\lambda_{t}^{\prime}=\lambda_{0} r^{t}$. The unbounded growth rate is $\frac{d \lambda^{\prime}}{d t}$. The bug infectiousness of each house is the difference in slope between these two models at a given time point $t$.

In practice, we are unable to observe the bug count at every time point. At each inspected house, we observe the bug count at one time point and the rest can be imputed using data augmentation. We assume bug counts follow a Poisson distribution, centered around the Beverton-Holt function, $\lambda_{t-I_{i}}$ where $t-I_{i}$ is the time since infestation:

$$
b_{i, t} \sim \operatorname{Poisson}\left(\lambda_{t-I_{i}}\right)
$$

where $b_{i, t}$ is the number of bugs in house $i$ at time $t$. This Poisson distribution is the third component
in likelihood (2.1).

We can now incorporate information about bug counts, spatial heterogeneity, and bug infectivity into the vector transmission function. We assume the following vector transmission function $h_{i j}(t)$ :

$$
h_{i j}(t)=1-\left(1-\beta_{i j}\right)^{\frac{d \lambda^{\prime}}{d t}-\frac{d \lambda}{d t}}
$$

where $\beta_{i j}=\beta \times \delta_{i j}$. We estimate $\beta$ through the RJMCMC and $\delta_{i j}$ is the estimated spatial kernel, described below.

Although this transmission function is mathematically complex, it has a nice interpretation. Given the infestation status of house $i$ and the distance between houses $i$ and $j, \beta_{i j}$ describes the probability that house $i$ infests house $j$. Thus, $1-\beta_{i j}$ describes the probability that house $j$ escapes infestation from house $i$. The probability that house $j$ escapes infection exponentially decreases as the number of bugs in house $i$ increases. We characterize the dynamic of migrating bugs as the difference in population growth between unbounded population growth and bounded population growth. Thus, subtracting this quantity from one gives the probability that house $i$ does in fact infest house $j$.

The transmission function appears twice in likelihood (2.1), both as the infectious pressure of each infested $i$ house on every $j$ house, $\int_{I_{k}}^{T_{\text {max }}} h_{i j}(t) d t$ and cumulative hazard function of each infested house $i$ on each $j$ house, $H_{i j}(t)=\sum_{s=1}^{t} h_{i j}(s)$.

### 2.2.4. Spatial Dynamics

There is a highly spatial component to the movement of $T$. infestans, and we also incorporate this spatial heterogeneity into the hazard function, $h_{i j}(t)$. It is important to note that any number of spatial kernels can be implemented into this approach. A modified exponential transmission kernel was identified that incorporates both distance between houses and a city block indicator (Barbu et al., 2013). We use this kernel to describe the probability of a bug migrating from house $i$ to house $j$ given the distance $d_{i j}$ between the two houses. Using this kernel, we also incorporate city streets as barriers, since previous studies have shown $T$. infestans are less likely to move a given distance between city blocks compared to within city blocks (Barbu et al., 2013). We use normal
prior distributions on $\zeta$ and $\rho$, centered around values identified from previous work (Barbu et al., 2013):

$$
\delta_{i j}=\rho \exp \left(\frac{d_{i j}}{\zeta}\right)
$$

where $d_{i j}$ defines the Euclidean distance between houses $i$ and $j, \zeta \sim \mathrm{~N}\left(9.00, \sigma_{\zeta}\right)$, and $\rho=1$ if houses $i$ and $j$ are on the same city block. If houses $i$ and $j$ are not on the same city block, $\rho \sim \mathrm{N}\left(0.30, \sigma_{\rho}\right)$.

### 2.2.5. Implementation

To fit our model, we use a reversible-jump Markov Chain Monte Carlo (RJMCMC) algorithm using R v3.2 and the 'Rcpp' package (Eddelbuettel et al., 2011). The RJMCMC algorithm allows for changes in the dimensionality of the parameter space (Green, 1995, Gibson and Renshaw, 1998). We allow for the possibility of 'adding' a potential unobserved infestation, removing an already added unobserved infestation, or shifting an infestation in time. Although the parameter estimates converge quickly, we run the RJMCMC for much longer for adequate convergence of the posterior probabilities of infestation. We run the algorithm for a minimum of 1 million iterations, which takes about 12 hours. The algorithm is run using the Sun Grid Engine, which is comprised of a variety of hosts running Red Hat Linux, and RAM is dynamically allocated. After running the algorithm for $M$ iterations, each house with an unknown infestation status is added as infested for $m_{i} \leq M$ iterations. Thus, the calculated posterior probability of infestation of each house is $P$ (infestation) $=\frac{m_{i}}{M}$. The simulation code is available at https://github.com/ebillig/Search-Strategy.

We update the likelihood using the following RJMCMC algorithm:

1. Initialize $\beta^{1}$ and $r^{1}$.
2. Initialize infection times $I^{1}$. Initial infestation $I_{\kappa}=1$ and all other observed infestations initialized at $I=2$. All other infestations set to $I=\infty$.
3. Initialize $t_{\text {insp, } i}$ and $R_{i}$. If house $i$ has been inspected andor treated, these are set to the respective times. All dates are converted to time since initial infestation, using 90 day intervals. For example, if house $i$ was infested 200 days after the initial infestation, this infestation time
is considered $I_{i}=3$. If a house has not yet been inspected and treated, $t_{\mathrm{insp}, i}=R_{i}=\infty$. $T_{\text {max }}$ is set to the present day.
4. Initialize bug counts of each infested house given the infestation times $I^{1}$. This is done using the Beverton-Holt Model:

$$
\begin{gathered}
\lambda_{t}=\frac{K}{1+(K-1) r^{-t}} \\
b_{i, I_{i}}=1 \\
b_{i, t-I_{i}} \sim \operatorname{Poisson}\left(\lambda_{t-I_{i}}\right)
\end{gathered}
$$

Replace $b_{i, t_{\text {insp }, i}}=B_{i, t_{\text {insp }, i}}$, the observed bug counts.
5. Update $r^{m+1}$. Propose $r^{\star} \sim \operatorname{Normal}\left(r^{m}, 0.05\right)$. Update all bug counts $b^{\star}$ given $r^{\star}$ using the Beverton-Holt model.

$$
R=\min \left(1, \frac{L\left(r^{\star} \mid I, \beta^{m}, \lambda_{t}\right)}{L\left(r^{m} \mid I^{m}, \beta^{m}, \lambda_{t}\right)} \frac{p\left(r^{\star}\right)}{p\left(r^{m}\right)}\right)
$$

where $p$ is the prior distribution of $r$. We define $p \sim \operatorname{Gamma}\left(a_{r}, b_{r}\right)$.

$$
U \sim \operatorname{Uniform}(0,1)
$$

If $U<R$ then $r^{m+1}=r^{\star}$. Else, $r^{m+1}=r^{m}$.
6. Update $\beta^{m+1}$. Propose $\beta^{\star} \sim \operatorname{Normal}\left(\beta^{m}, 0.2\right)$. The proposal is constrained to $(0,1)$.

$$
R=\min \left(1, \frac{L\left(\beta^{\star} \mid I^{m}, r^{m+1}, \lambda_{t}\right)}{L\left(\beta^{m} \mid I^{m}, r^{m+1}, \lambda_{t}\right)} \frac{p\left(\beta^{\star}\right)}{p\left(\beta^{m}\right)}\right)
$$

where $p$ is the prior distribution of $\beta$. We define $p \sim \operatorname{Beta}\left(a_{\beta}, b_{\beta}\right)$.

$$
U \sim \operatorname{Uniform}(0,1)
$$

If $U<R$ then $\beta^{m+1}=\beta^{\star}$. Else, $\beta^{m+1}=\beta^{m}$.
7. Propose moving, adding or removing an infestation, each with equal probability. To move an infestation:
(a) Update I. Select uniformly from the set of infested houses, $N_{I}$. If $i$ has not yet been
inspected, propose new infestation time $I_{i}^{\star} \sim \operatorname{Uniform}\left(I_{k}+1, T_{\max }\right)$. If $i$ has been inspected but no bugs were found at that time, propose new infestation time $I_{i}^{\star} \sim$ Uniform $\left(t_{\text {insp }, i}, T_{\text {max }}\right)$. If $i$ has been inspected and at least one bugs was found at that time, propose new infestation time $I_{i}^{\star} \sim \operatorname{Uniform}\left(I_{k}+1, t_{\text {insp }, i}\right)$.

$$
\begin{gathered}
R=\min \left(1, \frac{L\left(I^{\star} \mid \beta^{m+1}, r^{m+1}, \lambda_{t}\right)}{L\left(I^{m} \mid \beta^{m+1}, r^{m+1}, \lambda_{t}\right)}\right) \\
U \sim \operatorname{Uniform}(0,1)
\end{gathered}
$$

If $U<R$ then $I^{m+1}=I^{\star}$. Else, $I^{m+1}=I^{m}$. If $I^{m+1}=I^{\star}$, update $b_{i}$ (all bug counts corresponding to this infestation) so that at $I_{i}, b_{i 1}=1$.

To add an infestation:
(a) Propose $i$ uniformly from $\mathcal{S}$, the set of susceptible houses.
(b) If $i$ has not yet been inspected, propose $I_{i}^{\star} \sim \operatorname{Uniform}\left(I_{k}+1, T_{\max }\right)$. If $i$ has been inspected, propose $I_{i}^{\star} \sim \operatorname{Uniform}\left(t_{\text {insp }, i}, T_{\max }\right)$. Propose $b_{i}^{\star} \sim \operatorname{Poisson}\left(\lambda_{t-I_{i}^{\star}}\right)$.

$$
R=\min \left(1, \frac{L\left(I^{\star} \mid r^{m}, \beta^{m}, \lambda_{t}\right)}{L\left(I^{m} \mid r^{m+1}, \beta^{m+1}, \lambda_{t}\right)} * p_{i}\left(I_{i}^{\star}\right) * \frac{1}{q\left(b_{i}^{\star} \mid \lambda_{t-I_{i}}\right)}\right)
$$

where $p_{i}$ is the prior probability that house $i$ is infested and $q$ is the proposal distribution of the bug counts $b_{i}$ (Poisson).

$$
U \sim \operatorname{Uniform}(0,1)
$$

If $U<R$ then $I^{m+1}=I^{\star}$. Else, $I^{m+1}=I^{m}$. If a house is added, the corresponding bug counts but also be added.

To remove an infestation:
(a) Propose $i$ uniformly from the set of occult infestations (previously added, but unobserved, houses).

$$
R=\min \left(1, \frac{L\left(I^{\star} \mid r^{m}, \beta^{m}, \lambda_{t}\right)}{L\left(I^{m} \mid r^{m+1}, \beta^{m+1}, \lambda_{t}\right)} * \frac{1}{p_{i}\left(I_{i}^{\star}\right)} * q\left(b_{i}^{\star} \mid \lambda_{t-I_{i}}\right)\right)
$$

where $p_{i}$ is the prior probability that house $i$ is uninfested and $q$ is the distribution of the
bug counts $b_{i}$ (Poisson).

$$
U \sim \operatorname{Uniform}(0,1)
$$

If $U<R$ then $I^{m+1}=I^{\star}$. Bug counts corresponding with the removed infestation must also be deleted. Else, $I^{m+1}=I^{m}$.
8. Repeat steps (e)-(g) for a total of $M$ iterations.

In the field implementation, $r$ was not estimated, and thus step 5 was omitted. In general, the acceptance rates of parameters ranged from $15 \%$ to $60 \%$, including those in the reversible-jump component. Posterior probabilities of infestation ranged from $0 \%$ to $20 \%$, but mostly stayed below $10 \%$.

Parameter estimates converge quickly, but it is difficult to assess convergence of posterior probabilities of infestation. We visually assess convergence by plotting the ranking of each house between chains of the RJMCMC (Appendix A.2). Rankings are consistent in that top ranked houses were ranked highly across all chains. However, the specific ranking of a given house varies between chains. A typical convergence plot is shown in Figure A.2. A few houses seemed to get stuck at high rankings each chain that did get picked up in other chains. By using the median ranking for each house across chains, we hope to minimize this effect on inspections.

We pilot our model in near real-time in the city of Arequipa. Inspectors search houses with the highest posterior probability of infestation in three adjacent localities of the city. Each night, we run five chains and obtain a ranking for each house by using the median ranking across the chains for each locality. After a day of inspections, we re-run the model to include observed data from the most recent day, creating a new ranking list for the following day. Houses that refuse or do not answer are kept in the algorithm as unknown infestations; those that did not answer may be revisited at a later date, while those that refuse inspection are removed from the pool. Abandoned houses are assumed to be uninfested and are not included in the algorithm. We believe this is a reasonable assumption because these sites contain no food sources for vectors. Each night, we run the model for as many iterations as possible for a daily update, over a million iterations. Houses outside of the study site, but within 50 meters of the border of the study site, are included in the model, however they are not included in the potential inspection pool. Including these nearby houses allows for the possibility that vectors enter the study site from a neighboring locality.

### 2.3. Simulations

We test our algorithm on simulated data generated on a real landscape. We choose a subset of the study region (173 houses) on which to simulate an epidemic. We randomly selected one house to be the initial infestation, $\kappa$, and set this house to have 1 bug at time $t=1, b_{\kappa, 1}=1$. We then forward simulate, using the Beverton-Holt model $b_{\kappa, t} \sim \operatorname{Poisson}\left(\lambda_{t}\right)$ until $t_{\text {max }}$. At time $t=2$, house $\kappa$ infests each other house $j, j \neq \kappa$ with probability $h_{\kappa j}(t)=h_{\kappa j}(2)$. As more houses become infested, these houses can then infest other houses. If house $i$ is not yet infected at time $t$, then $h_{i j}(t)=0$ (ie. the probability that house $i$ infests house $j$ at time $t$ is zero). The inspection time of each house is randomly selected $t_{\text {insp }, i} \sim$ Uniform $\left(I_{i}, t_{\max }\right)$. The initial infestation time $I_{\kappa}$ is the only infestation time that is considered observed. To simulate unreported infestations, we randomly choose $1 / 3$ of the infestations to treat as unobserved.

We simulate the data under three sets of parameter values $\{\beta, r\}=\{0.3,2.0\},\{0.7,2.5\},\{0.05,3.0\}$. To create realistic data, we are limited to simulated datasets under a subset of parameter regimes in which vector spread occurs. For each simulation, a wide beta prior distribution is used for $\beta$ and a wide gamma distribution is used for $r$, both centered around the true value.

Under each set of parameter regimes, we simulate 200 datasets. For each dataset, we run 3 chains at different starting values. We run each chain for 500,000 iterations, and discard the first 5000 iterations as burn-in. We assess convergence by plotting the rankings of all chain pairs (more details in Appendix A.2). From each simulation, we calculate the sensitivity and specificity by recovering occult infestations under various thresholds of inspection criteria. We create receiver operating characteristic (ROC) curves to examine the simulation results (Figure 2.1). We also estimate the parameters of interest, $\beta$ and $r$. For each simulated dataset, we record the median, $95 \%$ credible interval and AUC and report the mean of all datasets (Table 2.1). The credible intervals for our parameter estimates are wide, however we are trying to estimate a lot of information with little observed data. From the simulation results, we verify that our approach accurately estimates the parameters of interest, as our median estimates were close to the true values. In all cases, the AUC shows that our model performs better than random inspections. In addition, the AUC suggests that the model performs best under low values of $\beta$. With high values of $\beta$, the success of the model varies. This may result from a larger number of vectors traveling farther distances, compared to rare
long-distance movements when $\beta$ is small. If these maps become very saturated with infestations, the model may miss large foci of unobserved infestations. The acceptance rate of the RJMCMC ranged from 20-50\%.


Figure 2.1: ROC curves describing the ability of the RJMCMC to uncover unreported infestations in a simulated vector-borne epidemic. In each simulation, $1 / 3$ of infestations were randomly removed to recover. 50 randomly selected ROC curves of each parameter regime are shown. (a) $\{\beta, r\}=$ $\{0.02,1.6\}$ (b) $\{\beta, r\}=\{0.7,1.8\}$ (c) $\{\beta, r\}=\{0.3,2.1\}$

| Parameter Estimation |  |  |  |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  | $2.5 \%$ | $50 \%$ | $97.5 \%$ | $2.5 \%$ | $50 \%$ | $97.5 \%$ |  |
| $\beta=0.02, r=1.60$ | 0.0004 | 0.03 | 0.60 | 1.27 | 1.62 | 1.94 | 0.83 |
| $\beta=0.70, r=1.80$ | 0.31 | 0.76 | 0.98 | 1.21 | 1.73 | 1.95 | 0.57 |
| $\beta=0.30, r=2.10$ | 0.05 | 0.26 | 0.62 | 1.94 | 2.04 | 2.13 | 0.61 |

Table 2.1: Simulation results. Each estimate represents the mean of given quartile for the set of 200 replicates. We use the median estimate to verify our approach obtains accurate parameter estimates and report the median AUC of each parameter set.

### 2.4. Data Results

We implement our approach across three communities of the city of Arequipa, Peru. Each community had been treated with insecticide in 2004; we use data collected during this treatment to calculate prior probabilities of infestation for each household following Barbu (Barbu et al., 2014). Since 2004, residents have reported suspected infestations, which were then verified by our field staff and the Minstry of Health entomological inspectors. If a true infestation was reported, adjacent houses were inspected. Thus, houses were inspected at unique time points, either due to a resident report, or due to a positively infested neighboring house. The communities contain between

551 and 669 houses. There had been 11, 10 and 9 infestations observed at communities 1, 2, and 3 respectively at time points between 2004 and 2014. The observed bug counts range from 0 to 212 , and the median bug count of an infested house across all communities is 3 . Due to small number of infestations and the low bug counts in these areas, we do not attempt to estimate the bug growth rate, $r$. Instead, we draw $r$ from a normal distribution centered around 3.66 as found in Rabinovich, 1972 and only estimate $\beta$. The posterior distribution of $\beta$ at the beginning of the study is summarized in Figure 2.2. The estimates are similar between communities. We observe a wide $95 \%$ credible interval on the posterior distribution of $\beta$ across the three communities. The estimates of $\beta$ remain consistent throughout the implementation.


Figure 2.2: Posterior estimate of $\beta$, the probability of successful invasion of a migrating bug, at the beginning of the inspection campaign in each community, indicating median and $95 \%$ credible interval of RJMCMC chains: (a) Community 1: 0.67 ( $0.33,0.91$ ) (b) Community 2: $0.59(0.22,0.90)$ (c) Community 3: $0.51(0.15,0.85)$

Between October and December of 2015 we made a total of 835 household visits to 409 distinct households across the three study communities. Within each day, inspectors traveled within the same community. Only 135 households agreed to participate and permitted us to inspect their premises. We did not find a single instance of infestation. The low participation rate was due to many factors. Some people refused inspection or did not answer the door. Additionally some houses were abandoned.

Despite the low participation rate (33\%), the absolute absence of vectors in the 135 high-risk houses that we inspected indicates that the prevalence of infestation in the area is extremely low, much lower than we had anticipated at the beginning of the study. From the map, we can see that
inspectors were directed to areas near previous reports of infestation (Figure 2.3). Based on the results (of not finding any bugs), we conclude that this region may have lower levels of infestation than we anticipated, a very promising result more than 10 years after the original control campaign.


Figure 2.3: Path of inspectors in study region from October 2015 through December 2015. Sites of attempted inspections (blue circles) and successfully inspected houses (black circle) shown, as well as daily routes (blue lines). Previous confirmed infestations (red plus signs) also shown. Points are jittered to de-identify data.

### 2.5. Discussion

Vector-borne diseases continue to emerge and spread, especially throughout developing nations where observing detailed data, such as the time of individual infections, can be difficult (Hong et al., 2015, Weaver, 2013, Tarleton et al., 2007). We present a novel dynamic method that can guide the control of re-emerging insects by updating posterior probabilities of infestation as inspection data are collected. We demonstrate our method through simulations and then implement it in a campaign to curb the re-emergence of triatomine vectors of Chagas disease.

Methods have been developed to handle the incomplete data often encountered when observing infectious disease outbreaks, including unobserved infection times and incomplete epidemics (O'Neill and Roberts, 1999, Jewell et al., 2009b). We commonly observe the status of a subset of individuals at some time point after the true infection has occurred. Several alternatives to likelihood-based
methods have been developed, including approximate Bayesian computation (ABC) (Csilléry et al., 2010, Beaumont, 2010) and the synthetic likelihood (Wood, 2010). In addition, alternative methods of inference have been used, such as iterated particle filtering (lonides et al., 2011). We extend previous methods using a Bayesian model developed by Jewell et al., 2009a for notifiable diseases to cross-sectional observations of insect infestations. We provide a framework that can be easily extended to other insect infestations, including the recent bed bug epidemic. We then implement this method in the field in near real-time, updating the model daily to reflect the latest known observations. Real-time implementation creates additional challenges that are not present in simulated epidemic-control studies, such as resident participation (Buttenheim et al., 2014), geographic constraints, and computational power.

In addition to limitations of the model, the merging of the theoretical approach with practical constraints raises limitations that need to be addressed in future studies. The model itself is dependent on several assumptions made to retain a tractable likelihood. We assume a specific spatial kernel and insect population growth model. The insect population growth model is key to our ability to perform a likelihood-based analysis by enabling the estimation of the unobserved infestation times. Using these models, we must fix some parameters. We test the sensitivity of our results to these parameters (see Appendix A.2). In addition, we assume perfect inspections and spray effectiveness. Although the inspectors are highly skilled and search thoroughly, they may miss the presence of insects, especially the early stage nymphs which are small and difficulty to see. In addition, there is heterogeneity in their detection accuracy (Hong et al., 2015). Similarly, while the insecticide treatments are known to be highly effective, we cannot verify that the infestations were eliminated, especially in severe cases. Lastly, determining convergence of the posterior probabilities of infestation is difficult. The chains are qualitatively consistent in their rankings of houses (in terms of posterior probability of infestation) but the rankings themselves are not identical between chains, indicating limited convergence. We use the median ranking across chains to minimize this limitation.

Logistically, we encountered several limitations during the field implementation. The algorithm often sent our fieldworkers back and forth on semi-arduous treks across hillsides (Figure 2.3). In the future, we could gain efficiency by balancing visits to highest risk houses with decreasing daily travel costs. In addition, the inspectors often gather additional information on possible infestations based
on local knowledge-such information, while crucial, is not easily incorporated into our models. We are working to implement an inspection campaign that combines model-based risk maps with inspector knowledge.

Our modeling approach is purely exploitative, as opposed to exploratory. Inspectors go to houses with the highest posterior probability of infestation given the current knowledge of the system. Due to our exploitative approach, our conclusion of low prevalence may be incorrect if people from different areas of the city report infestations at different rates. Our model cannot uncover infestations in areas mistakenly considered to be low risk. In this case, our infestation estimates would be artificially low. An inspection algorithm that appropriately balances both an exploitative and explorative model may be more appropriate in a search strategy.

We have developed a model to search for vectors of disease by understanding their movement through an urban landscape. Our method could also be applied to other settings where little information is known about the pathogen, but more is known about the vector. Our framework enables a dynamic approach to make informed vector surveillance decisions and efficiently inspect houses for disease agents.

## CHAPTER 3

## RISK MAPS FOR CITIES: Incorporating streets into geostatistical MODELS

### 3.1. Introduction

This chapter is motivated by the need to understand vector distribution patterns in the urban environment of Arequipa, Peru. In particular, we are interested in models that can help guide search strategies for Triatoma infestans, the principal vector of Chagas disease in the region. Chagas disease, caused by the parasite Trypansoma cruzi, causes significant mortality in the Americas (Dias, Silveira, and Schofield, 2002, Bern, 2015). Due to successful vector control campaigns and surveillance, both the parasite and vector are currently relatively rare in metropolitan Arequipa (Barbu et al., 2014).

Risk maps can be used to guide search strategies. Creating risk maps using Gaussian fields (GFs) is an area of active research and development (Oluwole et al., 2015, Adigun et al., 2015, Jaya et al., 2016). Until recently, fitting GF models was computationally difficult, due to large matrix calculations (the big $n$ problem) in covariance estimation. However, recent advances in theory and computation, discussed below, have alleviated this problem. Lindgren et al. described the relationship between GFs and Gaussian Markov Random fields (GMRFs); an R package was developed to implement these analyses using nested integrated Laplace approximations, easing the computational burden of these models (Blangiardo et al., 2013, Lindgren, Rue, and Lindström, 2011, Rue, Martino, and Chopin, 2009).

However, geostatistical models using Gaussian fields do not typically incorporate the structure of urban landscapes (Diggle, Ribeiro Jr, and Christensen, 2003). Vector-borne diseases are increasingly common in urban areas (Weaver, 2013). Poor and unplanned urban environments can create ideal breeding grounds for many vectors, facilitating increased transmission of vector-borne diseases in population dense areas (Knudsen and Slooff, 1992, Bowman et al., 2008, Levy et al., 2006). Several arboviruses, including Dengue, Chikungunya, West Nile, and Zika, have emerged repeatedly in urban areas (Haley, 2012, Sikka et al., 2016). Parasitic diseases, such as malaria and

Chagas disease, once considered rural problems, have become common in cities (LaDeau et al., 2015, Delgado et al., 2013). Including the spatial structure of city streets may more accurately describe spatial associations characteristic of insect infestations and therefore help in developing effective public health interventions to reduce transmission in population dense areas (Weaver, 2013).

In this paper we develop an approach to predict the probability of urban insect vector infestation using a geostatistical model that incorporates city streets as permeable barriers. Our approach estimates the reduced movement of vectors between blocks, compared to within blocks, and predicts the heterogeneous urban vector distributions using a Gaussian field. Our model is computationally efficient and easily adaptable to other cities and vectors. Here, we present our methodology, demonstrate our approach on simulated data, and apply it to data on Chagas disease vectors in a district of the city of Arequipa, Peru.

### 3.2. Methods

### 3.2.1. Gaussian field approach

In this subsection we review the Gaussian field approach that can be used to create risk maps, ignoring the issue of streets as barriers. Gaussian fields are often used to model various types of point-level data, also known as geostatistical data. These models are popular for their flexibility and ability to capture complex processes across a wide range of applications such as epidemiology, ecology, and imaging (Rossi et al., 1992, Brooker, 2007, Diggle, Ribeiro Jr, and Christensen, 2003). Using this modeling approach, we assume the data, the presence of T. infestans, is a continuous stochastic process, with observations over a two-dimensional landscape, at locations $c$.

Denote by $y_{i}$ the indicator variable for vector presence at house $i$. We can model the probability of vector presence at house $i, \pi_{i}$, using a logistic model with intercept and Gaussian field. Although we do not use any household-level covariates in our model, they can easily be incorporated into the formula (3.1) with additional regression parameters. We use the model:

$$
\begin{align*}
\operatorname{logit}\left(\pi_{i}\right) & =\beta_{0}+u_{i}  \tag{3.1}\\
\boldsymbol{u} & \sim N(0, \Sigma)
\end{align*}
$$

where $u_{i}$ is a realization of the GF, $x\left(c_{i}\right)$, and $c_{i}$ is the location of house $i$. We use a Matern covariance structure, which is commonly used in spatial statistics. Specifically, the covariance between $u_{i}$ and $u_{j}$ is

$$
\Sigma_{i j}=\sigma_{u}^{2} \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\kappa\left\|c_{i}-c_{j}\right\|\right)^{\nu} K_{\nu}\left(\kappa\left\|c_{i}-c_{j}\right\|\right)
$$

where $\sigma_{u}^{2}$ is the marginal variance of the random effect, $\|\cdot\|$ is the Euclidean distance, $K_{\nu}$ is the modified Bessel function of the second order, and $\kappa$ and $\nu$ are parameters. The parameter $\nu$ describes the smoothness of the stochastic process and therefore controls the shape of the covariance function (the function is $\lfloor\nu-1\rfloor$ times differentiable). The $\kappa$ parameter characterizes how quickly the correlation between two points decreases as the distance between them increases, capturing the scale of the relationship.

The Matérn covariance structure has become widely accepted for GF models because of its link to GMRFs (Besag, 1975, Rue and Held, 2005, Rue, Martino, and Chopin, 2009, Lindgren, Rue, and Lindström, 2011). Using the Matérn covariance, $x(c)$ is a solution to the linear fractional stochastic partial differential equation (SPDE):

$$
W(\boldsymbol{c})=x(\boldsymbol{c})\left(\kappa^{2}-\Delta\right)^{\alpha / 2}
$$

where $\alpha=\nu+d / 2$ and $\Delta$ is the Laplacian

$$
\Delta=\sum_{i=1}^{d} \frac{\partial^{2}}{\partial x(\boldsymbol{c})_{i}^{2}}
$$

and $d=2$ since $c \in \Re^{2}$ (Whittle, 1954, Whittle, 1963).

Using this solution, Lindgren, Rue, and Lindström, 2011 found a direct link between the GF and GMRF, which greatly eases the computational burden of GF estimation. Using this link, we can estimate the precision matrix that represents the GF accurately, over a wide range of marginal variances, with sparse matrix calculations. For more details on this relationship, see Lindgren, Rue, and Lindström, 2011.

We follow the parameterization of Lindgren, Rue, and Lindström, 2011 and Krainski and Lindgren,

2013, defining the covariance function in terms of $\theta=\left\{\theta_{1}, \theta_{2}\right\}$, functions of $\kappa$, the scale parameter, and $\sigma_{u}^{2}$, the marginal variance:

$$
\begin{aligned}
& \theta_{1}=-\log \left(4 \pi \kappa^{2} \sigma_{u}^{2}\right) / 2 \\
& \theta_{2}=\log (\kappa)
\end{aligned}
$$

We also fix $\alpha$, as defined earlier, to $\alpha=2$. This setting is a natural choice for two dimensional problems, as argued in Whittle, 1954, but researchers may vary the value of $\alpha$ as needed. Together, $\theta$ and $\alpha$ define the Matérn covariance, $\Sigma$, of the GMRF.

### 3.2.2. Incorporating streets as barriers

The Gaussian field assumes a continuous, non-linear, yet smooth relationship between houses as a function of the Euclidean distance between them. However, urban streets create an uneven landscape on which the stochastic process occurs. Previous studies have shown the vector of Chagas disease, T. infestans, is less likely to move between city blocks compared to within blocks (Barbu et al., 2013).

One option to capture the heterogeneity of the urban grid would be to add a parameter into the covariance function to indicate whether two points are on the same city block as each other. This approach would allow the relationship of the outcome and distance to depend on whether or not streets separate the points. However, the Matérn covariance is the key element that links the GF to GMRF and changing the function itself may impact the relationship, which is key to efficient parameter estimation. We propose an alternative approach that fits into existing GF estimation software and incorporates an additive effect, so if points are separated by multiple streets, the barrier increases.

Our approach uses a single additional parameter, $S$, that influences the covariance function directly through the distances between houses by distorting the city map. Using $S$, we create additional Euclidean distances between houses on different blocks, but maintain the Euclidean distances between houses on the same block, creating permeable barriers. We define $S$ as the ratio of the distorted distance between geographic block medians, calculated as the median $c$ of each block,


Figure 3.1: A) $S=1$, which corresponds to the true map of section in Mariano Melgar, Arequipa, Peru used for simulations. There is no map distortion at this scale. B) Distorted map to a scale of $S=1.5$, where the distance between geographic medians of blocks are 1.5 fold the true distance. C) Map distorted to scale $S=2.5$, where distance between geographic medians of blocks are 2.5 fold the true distance. Note distance between houses within block is maintained but distance between houses on different blocks is stretched.
compared to the true distance (Figure 3.1). This additional distance between blocks influences the model directly through the spatial covariance structure, $\left\|c_{i}-c_{j}\right\|$ by widening the streets between blocks. With this approach, if $S$ is known, the Gaussian field model can be directly used. In addition, this approach has an additive barrier effect, rather than simply an indicator of whether or not two houses are on the same block. If points are more than one block apart, the width of each street between the points is modified. We assume streets do not facilitate improved vector movement and restrict $S \geq 1$, with $S=1$ representing the true map.

In other words, $S$ reduces the correlation between houses that are a distance apart but on different city blocks. In practice, this additional parameter is a flexible, usable tool to characterize a heterogeneous landscape using a continuous latent field. On different types of landscapes, $S$ could be used to define other potential permeable barriers such as rivers, valleys, or mountains.

### 3.2.3. Estimation and Interpretation

We now describe our approach to incorporate $S$ into the model, interpret the map distortion, and estimate the parameters. Each house $i$ is located on a known city block $j$. We first mark the spatial center of each block by finding the median coordinates of each block,

$$
\left\{X_{j}^{M}, Y_{j}^{M}\right\}=\left\{\operatorname{median}\left(X_{j}\right), \operatorname{median}\left(Y_{j}\right)\right\} .
$$

We then define the location of each house in relation to the center of the house's block,

$$
\left\{\bar{X}_{i}, \bar{Y}_{i}\right\}=\left\{X_{i}-X_{j}^{M}, Y_{i}-Y_{j}^{M}\right\} .
$$

We stretch the map by moving the block centers by scale $S$ and recording house coordinates relative to the block center, $\left\{\bar{X}_{i}, \bar{Y}_{i}\right\}$, at the true distance. The distorted map coordinates, $\left\{\hat{X}_{j}, \hat{Y}_{j}\right\}$, become

$$
\left\{\hat{X}_{i}, \hat{Y}_{i}\right\}=\left\{X_{j}^{M} S+\bar{X}_{i}, Y_{j}^{M} S+\bar{Y}_{i}\right\} .
$$

This approach retains the block structure but enables the manipulation of blocks relative to each other. The interpretation of the distorted map is somewhat dependent on the map itself due to the irregular size and shape of each individual city block. $S$ describes the additional distance between the geographic median of each city block relative to the true distance. For example, $S=1.5$ corresponds to adding $50 \%$ of the true distance between the geographic medians of each block. Due to the irregular grid, this distortion corresponds to varying degrees of street distortion depending on the size of the blocks on either side of the street. The effect of the map distortion of the width of streets in the map in Figure 3.1 is summarized in Table 3.1. Larger city blocks result in greater distortion because the houses are located farther from the geographic median of the block. If blocks are all the same shape and size, the increase in street width will be proportional for all streets, however this is rare in practice. For more details on the interpretation of how $S$ corresponds to the width of the street, see Figure 3.2.

To use the GMRF representation on irregular points, we divide our landscape into non-intersecting triangles, as described in Lindgren, Rue, and Lindström, 2011. A Delaunay triangulation is created over the landscape, forming a mesh. Each house is located at a vertex in the mesh. To ensure a regular triangulation, the maximum edge length is specified to be $100 S$ to appropriately adjust for map distortion. Although meshes are not identical between grid samples of $S$, controlling the maximum edge length as a function of $S$ ensures that the triangulation is similar (Figure 3.3).

We implement this methodology using the R package 'INLA’ (Rue, Martino, and Chopin, 2009, Rue et al., 2014, Krainski and Lindgren, 2013). The ability of this approach to fit into the existing R package makes the method easy to implement for researchers in different fields. INLA is a relatively

|  | $S=1$ | $S=1.5$ | $S=2.5$ |
| :--- | :--- | :--- | :--- |
| Average distance (and sd) between <br> nearest neighbors on the same block <br> (ie. no barrier) | $1.0(0.3)$ | $1.0(0.3)$ | $1.0(0.3)$ |
| Average distance (and sd) between <br> nearest neighbors on different blocks <br> (ie. one barrier) | $3.6(1.1)$ | $8.0(1.5)$ | $16.0(2.3)$ |
| Ratio of average distance between <br> nearest neighbors on different blocks <br> compared to same distance when $S=$ <br> 1 | - | $2.2(0.4)$ | $4.4(0.9)$ |

Table 3.1: Description of map distortion in Figure 3.1 scaled so the spatial unit is the average distance between nearest neighbors on the same block. Interpretation of $S$ varies by specific map due to variability in sizes and shapes of city blocks. $S$ describes ratio of distance between geographic median of each city block relative to the true map (which is equivalent to $S=1$ ). Table summarizes how this distortion corresponds to additional distance between houses on different blocks using mean and standard deviation $(s d)$. The distortion varies block by block due to irregular grid.


Figure 3.2: In this figure, we demonstrate how streets between larger blocks result in a larger barrier effect. A. Four hypothetical blocks super-imposed. The smallest block is simply one house. The largest block has an X distance of 4 from the block median to the edge of the block. B. Same hypothetical blocks after distortion with $S=2$. C. The ratio of street width for this distortion at $S=2$ compared to no distortion at $S=1$. The street width doubles if the block is a single house and the distance from block median to edge is zero. Otherwise, the street width more than doubles, and the effect increases as the block size increases.
new, yet powerful tool, and the package is designed for flexible and complex model development. The package is continually updated to incorporate new developments in spatial statistics. We chose to implement our approach using INLA due to the speed, package flexibility, and ease of implementation for future researchers. We plan to use our method in real-time in the field to guide vector surveillance, and therefore these strengths are of particular significance.

In the last few years, INLA has become a popular method to fit GFs. The algorithm is an alternative to Markov chain Monte Carlo algorithms, which are common for geostatistical models but can be


Figure 3.3: Mesh over the map of the simulation region (houses in blue). Maximum edge length is constrained to $100 S$ to keep meshes consistent between scales. A. $S=1$ B. $S=1.5$
problematic in GF estimation due to non-convergence and long computation times. The latent field $\boldsymbol{u}$ tend to be highly correlated, and $\boldsymbol{u}$ also tend to be dependent on the model hyperparameters, a common issue that arises in MCMC algorithms. Rue et al. found estimation of GMRF models using integrated nested Laplace approximations was more precise and significantly faster (Rue, Martino, and Chopin, 2009). This approach uses Gaussian approximations, nested Laplace approximations, and numerical integration to estimate the marginal distributions of the latent field and hyperparameters. The approximations are especially precise for GF estimation. For details on the algorithm, see Rue, Martino, and Chopin, 2009.

To estimate $S$, we fit the model at several values of $S$ and compare the log-likelihoods. We estimate $S$ as the value that maximizes the log-likelihood of the model. Using the INLA estimation algorithm, it is not possible to update the estimate of $S$ in the same way as an MCMC algorithm. In addition, changing $S$ changes the mesh used in the model. However, for the purposes of creating risk maps, estimating a full posterior distribution for $S$ does not provide a benefit over using the estimate of $S$ directly.

### 3.3. Simulations

To evaluate the performance of our proposed method, we simulate data on a subset of the study region consisting of 2265 houses over 93 city blocks (Figure 3.1A). Our barrier effect parameter, $S$, and three parameters ( $\kappa, \sigma_{u}, \beta_{0}$ ) in the model (3.1) with $2 \times 2 \times 2=8$ simulation scenarios with a

| True values |  |  |  |  |  |  | Estimates |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\kappa$ | $\sigma_{u}^{2}$ | $S$ | $\beta_{0}$ | $\hat{S}$ | $\hat{\beta}_{0}$ | Coverage $\left(\beta_{0}\right)$ | Identification $(S)$ |  |
| 0.005 | 5 | 1.5 | -3 | $1.57(0.35)$ | $-3.11(0.89)$ | 1.00 | 0.95 |  |
|  |  | 2.5 | -3 | $2.48(0.53)$ | $-3.13(0.61)$ | 1.00 | 0.73 |  |
|  | 10 | 1.5 | -3 | $1.61(0.33)$ | $-3.13(1.26)$ | 0.97 | 0.98 |  |
|  |  | 2.5 | -3 | $2.77(0.55)$ | $-3.05(0.94)$ | 0.97 | 0.89 |  |
| 0.01 | 5 | 1.5 | -3 | $1.64(0.27)$ | $-2.97(0.53)$ | 0.97 | 1.00 |  |
|  |  | 2.5 | -3 | $2.76(0.48)$ | $-2.91(0.35)$ | 0.96 | 0.76 |  |
|  | 10 | 1.5 | -3 | $1.58(0.19)$ | $-3.16(0.77)$ | 0.96 | 1.00 |  |
|  |  | 2.5 | -3 | $2.76(0.55)$ | $-3.00(0.49)$ | 0.95 | 0.92 |  |

Table 3.2: Results from 100 Monte Carlo simulations for each parameter set. The parameter estimates are shown with the corresponding estimated standard deviations with the true values set for the simulations. The coverage of $\left(\hat{\beta}_{0}\right)$ is the average rate that the credible interval captures the true value of $\beta_{0}$ for identified simulation cases. The last column is the proportion of identifiable simulated datasets.
fixed intercept $\beta_{0}=-3$ are considered: (1) $\kappa$ is either 0.005 or 0.01 , (2) $\sigma_{u}^{2}$ is either 5 or 10 , and (3) $S$ is either 1.5 or 2.5 . We choose these parameter values because they reflect those similar to those that we observe in our data. Additional simulation scenarios are considered in Appendix B.1. Each simulated data is generated with respect to the true values of the four parameters in each simulation scenario.

Table 3.2 shows the simulation results from 100 Monte Carlo simulations for each scenario. The table summarizes the true parameter values, the parameter estimates using our approach, and the proportion of simulations with successful identification of $S$. It is important to tune the parameters so that they produce realistic infestation patterns; under some parameter regimes, extremely sparse or over-saturated landscapes will be produced. In some cases, the log-likelihood continually increases, discussed in more detail later, resulting in cases we call unidentifiable. These cases are not exactly unidentifiable, as these likelihoods suggest that in these cases streets are impermeable barriers, and $S$ is large. These cases occurred most frequently under large true values of $S$ and small values of $\sigma_{u}^{2}$. We believe this occurs because this parameter combination produces sparse landscapes more often than other parameter combinations. We remove these cases in our summarized simulation results. The simulation results demonstrate the ability of this approach to successfully identify $S$ in most cases. The estimates of $\beta_{0}$ are very close to the true value in all conditions. The coverage of $\beta_{0}$ was high for very small $\kappa$ and $\sigma_{u}^{2}$, but otherwise coverage was close to $95 \%$.

In addition, the model captures the covariance function remarkably well, using only observed binary data (Figure 3.4). The specific parameter estimates of $\kappa$ and $\sigma_{u}^{2}$ have a slight bias, but when combined, the overall covariance function captures the true covariance function well. We are not interested in the underlying parameter values but in the covariance function as a whole, and thus report our results in Figure 3.4.


Figure 3.4: Comparing the estimated Matérn covariance function (gray) with the true Matérn covariance function (black) under four parameter sets.

We conduct additional simulations to quantify the gain in infestation prediction using permeable barriers. To do so, we simulate datasets and fit the model assuming one-third of the infestations are observed. We simulate the data under $S=1$ (true map - no additional barrier effect), $S=2.5$, and $S=4$, and fit the model using both the true map and distorted map under the estimated scale, $\hat{S}$ (Table 3.3). To describe the results, we report the difference in number of positive houses discovered if inspectors searched the unobserved houses with the top $30 \%$ of probabilities. These simulation results indicate that the model using barriers better guides risk-based searches, especially when streets are strong barriers (Table 3.3). When streets are not barriers ( $S=1$ ), using the approach does not hinder the number of positive houses discovered, but also does not improve it. In addition, even when the majority of houses are unobserved, our approach identifies and estimates $S$ reasonably well (Table 3.3).

| $\kappa$ | $\sigma_{u}^{2}$ | $S$ | $\hat{S}$ | + houses under $\hat{S}$ | + houses under $S=1$ | Difference |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.005 | 20 | 4 | 4.4 | 316 | 302 | 14 |
|  |  | 2.5 | 3.1 | 324 | 319 | 5 |
|  |  | 1 | 1.1 | 315.5 | 316.5 | -1 |
|  | 10 | 4 | 4.3 | 230 | 216 | 14 |
|  | 2.5 | 3.4 | 245 | 241 | 4 |  |
|  |  | 1 | 1.1 | 252 | 252 | 0 |

Table 3.3: Difference in number of positive houses in the top $30 \%$ of probabilities of infestation using our approach of map distortion compared to using the true map with no distortion $(S=1)$. Model fit with true values of $S=1, S=2.5$, and $S=4$ when a randomly selected one-third of points were observed. Intercept fixed at $\beta_{0}=-5.100$ simulated datasets were run at each value. We report number of positive houses in the top $30 \%$ of probabilities which were treated as unobserved (ie. no gain for houses that were observed as positive).

From the simulations, we identified two issues. First, for a small subset of simulated infestation patterns, it is difficult to identify the scale parameter $S$ (Figure 3.5). The identification issue can be understood from the following example. Consider two infected houses on separate blocks and assume that the spatial components of the two locations are weakly correlated. This weak correlation can appear either due to a large value of the street barrier effect parameter $S$, or to a rapidly decreasing Matérn covariance function. The identification issue is easily overcome when a larger number samples is provided-but in the case of very scarce infestations, it is difficult to capture $S$ (Figure 3.5ac). Rarely, we observe an oversaturated landscape, with few uninfested blocks, which also creates an unidentifiable pattern (Figure 3.5b). For more details on identification of $S$, see Appendix B.1. Secondly, the estimation of Matérn covariance parameters $\kappa$ and $\sigma_{u}^{2}$ is not consistent. In Figure 3.4, each estimated covariance function is plotted and compared with the true covariance function. We can see that though $\hat{\kappa}$ and $\hat{\sigma}_{u}^{2}$ are biased, the estimated covariance function itself is remarkably close to the true function (Figure 3.4). Our main interest lies in capturing the function, rather than the individual covariance parameters, to create useful risk maps.


Figure 3.5: Three unidentifiable and identifiable log-likelihoods and the corresponding simulated datasets. Unidentifiable landscapes were uncommon, (rates varied based on the true parameter values of $\kappa$ and $\sigma_{u}^{2}$ ) and in most cases had scarce infestations (panels a and c). Occasionally, an unidentifiable landscape was oversaturated and also unidentifiable (panel b). For comparison, most simulated datasets were identifiable with clear maximums of the log-likelihoods (panels d, e, and f)

### 3.4. Data Results

We apply the method to data collected during a vector control campaign in the district of Mariano Melgar in Arequipa in 2009. The district contains 12,069 houses, of which 586 were found to be infested with insect vectors (Figure 3.6). To fit the model to the dataset, we sample $S$ over (1, 4) by increments of 0.1 and also estimate $\kappa, \nu$, and $\beta_{0}$. We use the value of $S$ that maximizes the log-likelihood of the model as our estimate, $\hat{S}$, and the corresponding model parameters, $\hat{\theta}$, and $\hat{\beta_{0}}$.


Figure 3.6: Map of the study region, the district of Mariano Melgar, Arequipa, Peru, which consists of 12,069 houses and 724 blocks. Color corresponds to number of known infested houses on the block.

Using this approach, we find the the log-likelihood is maximized at $S=1.5$. Our result indicates that streets are permeable barriers in the distribution of $T$. infestans, in agreement with previous studies (Barbu et al., 2013). We conclude that streets create barriers at least 1.5 fold the additional width. Our map contains blocks of varying sizes and shapes, and therefore our estimate, $\hat{S}$, means the average minimum distance between houses on different blocks is increased 2.1 fold with a standard deviation of 0.5 . A scaled subsection of the map, incorporating this additional distance, can be seen in Figure 3.1B. Using this $S$, the covariance function is described using the estimates of $\kappa=0.009$ with a $95 \%$ posterior credible interval of $(0.007,0.013)$ and $\sigma_{u}^{2}=7.716$ with a $95 \%$ posterior credible interval of $(5.492,10.047)$ (Figure 3.7). We estimate the model intercept, $\beta_{0}=-6.10$ ( 0.41 ). Estimates of all parameters across values of $S$ are summarized in Appendix B.2.


Figure 3.7: Posterior distributions of estimated parameters when $S=1.5$. A. Posterior distribution of $\sigma_{u}^{2}$ B. Posterior distribution of $\kappa$. C. Estimated posterior distribution of Matérn covariance, as a function of distance. For reference, when the map is scaled to $S=1.5$, the average distance between nearest neighbors on the same block is $10.2(s d=5.5)$ and the average distance between nearest neighbors on different blocks is 62.4. $(s d=18.0)$

Using the scale of $S=1.5$, we develop a risk map, representing the probability of infestation for each household (Figure 3.8). For comparison, we also present the risk map at $S=1$, the true city map, and $S=3$, additional widening of the barriers. Using this map, we can visualize the areas with elevated probability of infestation and compare the risk to the analysis without incorporating streets as barriers. Our estimates of the additional parameters, the covariance parameters $\theta$, and the model intercept $\beta_{0}$, suggest there is significant spatial correlation between houses both within and between city blocks.


Figure 3.8: Risk map of predicted probabilities of infestation using A) $S=1$ (true map) B) $S=1.5$ and C) $S=3$. The last panel shows differences in risk between scales of the area enclosed in the black rectangle in more detail. The color scale shows $P$ (infestation) and ranges from 0.74 (red) to 0.00 (purple)

### 3.5. Discussion

We presented an approach to assess the significance of the urban landscape on the spatial distribution of disease vectors and quantify the effect of city streets in the distribution of the Chagas disease vector T. infestans in Arequipa, Peru. We estimated that streets add a distance of $50 \%$ to the true street width in the spatial distribution of vectors in the study region. Our estimate is qualitatively similar to Barbu et al. who estimated a fixed additional distance for each street regardless of the original width, however a direct comparison is difficult due to the difference in approaches (Barbu et al., 2013). The flexibility, generalizability, and computational efficiency make our approach a promising tool for real-time risk map creation.

Risk maps are often used to develop epidemic predictions and intervention strategies. The obser-
vation of identifiable permeable barriers raises new potential targets for public health interventions in urban landscapes. For example, given our results, it may be more effective to inspect houses on the same city block as known infested houses rather than inspect houses within a set radius of them. We intend to use this modeling approach to guide inspections in Arequipa, Peru. As inspections are completed, we will update the model to reflect the latest observed infestations.

Our model has limitations, both theoretically and in its interpretation. From our simulation studies, we observed specific unidentifiable datasets. We suspect that these parameter values occasionally generate simulated datasets where the pattern of infested houses is clustered such that the $S$ is difficult to identify, including heavy clustering within a city block and lower total numbers of infested houses. From the likelihood plots (Figure 3.5), it is clear which datasets are unidentifiable, as they peak and then do not decrease as $S$ increases. The unidentifiability of certain datasets is also a practical limitation, as scarce infestations are expected after control actions, and thus the barrier may not be identifiable.

In addition, our model assumes a constant scale factor for all barriers. In fact some streets, such as paved streets, may present more of a hindrance to insects than others. It would be quite difficult to incorporate covariates in the effects of barriers within the framework presented here. Hong developed a method that considers barriers such as streets as 'sunken' in relation to the remainder of the Gaussian field, rather than stretching the streets as we do here (Hong, 2013). Using grid methods similar to those we employ he estimated the degree of 'sinkage' of each barrier. His method would allow for greater flexibility to assess barriers of different types. In addition, his method is not affected by the irregularities of the specific urban grid. However 'sinking' the streets requires enormous and tedious manipulation of the triangulation used to approximate the Gaussian field, while stretching the streets is simple and easily incorporated into the existing R package, INLA. The ability to quickly incorporate our approach into existing software makes our method accessible to researchers across many disciplines for real-time risk map creation.

## CHAPTER 4

## Generalized Multivariate Conditional Autoregressive Model for Vector-borne Disease Data

### 4.1. Introduction

The complex transmission dynamics of vector-borne diseases make analyzing spatial patterns especially difficult (Jones et al., 2008, Ostfeld, Glass, and Keesing, 2005, Reisen, 2010, Focks et al., 1999). For successful disease transmission between hosts, the pathogen must be successfully transmitted from host to vector, and then from the infected vector to a new host. Thus, the spatial dynamics and external factors associated with the pathogen, infected hosts, and both infected and uninfected vectors can be especially difficult to tease apart. Statistical spatial analyses of vectorborne diseases tend to focus on one element of the system, or model the vector and pathogen independently (Winters et al., 2010, Kitron and Kazmierczak, 1997). However, the distribution of a pathogen is dependent on the distribution of the vector due to the inherent dependency of the biological transmission process. We develop a generalized multivariate conditional autoregressive model to examine the proportion of pathogen infected vectors conditional on the spatial distribution of all observed vectors, both infected and uninfected.

We are motivated to understand the spatial heterogeneity of the primary Chagas disease vector, both infected and uninfected, Triatoma infestans, in our study region, Arequipa, Peru. Chagas disease, caused by the parasite Trypansoma cruzi causes significant mortality in the Americas (Dias, Silveira, and Schofield, 2002, Bern, 2015). After more than a decade of vector control campaigns and surveillance, both the parasite and vector are currently relatively rare in the city of Arequipa (Barbu et al., 2014). We observe spatially correlated clusters of vectors (Figure 4.1), although the number of vectors varies from 1 to more than 1,000 in a given house (Levy et al., 2006). The proportion of vectors that contain parasite also varies - most houses contain low proportions of positive vectors, but some houses contain very high proportions (Figure 4.1). It is unknown what causes the high observed proportion of infected vectors. We are interested in understanding whether the number of vectors in a given house is associated with the proportion of infected vectors, and whether there is additional spatial correlation among positive vectors once adjusted for the


Figure 4.1: Data used in analysis. Dataset contains 577 sites (contained within 67 houses; house level data not shown). Sites with vectors (red) and number of vectors (size of point on log scale) shown with number of vectors that tested positive for $T$. cruzi (size of yellow point on log scale).
distribution of all vectors.

To analyze parasite and vector data, we use a multivariate conditional autoregressive spatial model. Conditional autoregressive (CAR) models are popular spatial models due to their flexible parameterization and robust estimation (Wall, 2004, Besag, 1974). Until recently, spatial data analyses were entirely focused on univariate outcomes. Then, within a short period of time, several authors introduced multivariate CAR models, but with improper posterior distributions (ie. the normalizing constant was unidentifiable) (Kim, Sun, and Tsutakawa, 2001, Sain and Cressie, 2002). In 2003, Gelfand and Vounatsou developed an approach to extend these multivariate models to obtain a proper posterior distribution (Gelfand and Vounatsou, 2003). Then, Jin, Carlin, and Banerjee, 2005. developed an approach to obtain an explicit joint distribution of the multivariate spatial field, through simple conditional distributions. We extend the work of Jin, Carlin, and Banerjee, 2005, which was originally applied to two Poisson models on related cancers using the generalized multivariate conditional autoregressive (GMCAR) model. We develop a three-dimensional spatial model using a zero-inflated Poisson distribution for vector counts, and a binomial model for the proportion of positive vectors. We fit the spatial random effects conditionally, demonstrate the applicability
of the method on zero-inflated data and outcomes from different distributions on simulations, and apply the method to our Chagas vector and parasite data in Arequipa, Peru.

### 4.2. Methods

### 4.2.1. Univariate CAR

First, we briefly introduce the univariate CAR model on which the multivariate model is based. We define a univariate spatially random variable $\phi_{i}$ at $n$ locations. Then, the full conditional distribution is

$$
p\left(\phi_{i} \mid \phi_{j, j \neq i}, \tau_{i}^{-1}\right)=\mathrm{N}\left(\alpha \sum_{i \sim j} b_{i j} \phi_{j}, \tau_{i}^{-1}\right)
$$

where $i \sim j$ means $i$ and $j$ are neighbors.

The joint conditional distribution is

$$
\phi \sim \mathrm{N}\left(\mathbf{0},\left[D_{\tau}(I-\alpha B)\right]^{-1}\right)
$$

where $D_{\tau}=\operatorname{Diag}\left(\tau_{i}\right)$, and $\tau$ is a precision parameter. $B$ is $n \times n$ with $b_{i i}=0$ and $b_{i j}>0$ if houses are neighbors; the definition of $b_{i j}$ varies depending on the specific CAR implementation. $\alpha$ is a smoothing parameter varying from 0 to 1 that describes the spatial association (Besag, 1974, Wall, 2004, Waller and Gotway, 2004).

### 4.2.2. Generalized Multivariate $C A R$

Before describing the multivariate structure of the model, we first introduce the notation and structure for our zero-inflated Poisson and binomial models. We use a zero-inflated Poisson model to describe the distribution of all vectors, and a binomial model for the distribution of positive vectors, conditional on the observed number of vectors.

Define $Z_{i}$ as the total number of vectors found at house $i$ and $Y_{i}$ as the total number of positive vectors found at house $i . Y_{i}$ is always less than or equal to $Z_{i}$, as it is a subset of the sample. We use a zero-inflated Poisson model to describe the distribution of all vectors, $Z_{i}$, and a binomial model for the distribution of positive vectors, $Y_{i}$, conditional on the observed number of vectors.

Consider $Z_{i}$ to be zero-inflated Poisson, with probability $\pi$ of being in the Poisson distribution where $\pi$ is Bernoulli 1 or 0 .

$$
Z_{i} \sim \pi_{i} \text { Poisson }\left(\lambda_{i}\right)+\left(1-\pi_{i}\right) \delta_{0}
$$

Then, we define the models on $Z_{i}$ as:

$$
\begin{gathered}
\operatorname{logit}\left(\pi_{i}\right)=X_{i 3} \beta_{k 3}+\phi_{i 3} \\
\log \left(\lambda_{i}\right)=X_{i 2} \beta_{k 2}+\phi_{i 2}
\end{gathered}
$$

where $X$ is a $n \times k$ matrix of $k-1$ fixed covariates. In our analysis, we include covariables that have previously been shown to be associated with Chagas disease vectors, including the presence of guinea pigs, dogs, and birds (chickens and ducks). We also include information on housing materials that are known to be good habitats for vectors, including stacked bricks, cement blocks, and adobe, as well as housing materials that are known to be poor vector habitats, including chicken wire and smooth stucco. In our analysis, we use the same covariables between models. However, the method allows for the covariables to differ between models.

We similarly define a zero-inflated binomial model, however the distribution probability is observed, conditional on $Z_{i}$ :

$$
\begin{gathered}
Y_{i} \mid Z_{i} \geq 1 \sim \operatorname{Binomial}\left(Z_{i}, p_{i}\right) \\
\operatorname{logit}\left(p_{i}\right)=X_{i 1} \beta_{k 1}+\phi_{i 1}
\end{gathered}
$$

We restrict this model so sites may only contain positive vectors if at least one vector was observed. Otherwise, $p_{i}=0$. In this model, we use the same covariables as the other models (ie. $X_{i 1}=X_{i 2}=$ $X_{i 3}$ ), however they may vary if different covariables are associated with different outcomes.

In each model, we have unique spatial random effects, $\phi_{1}, \phi_{2}, \phi_{3}$. If these three models were estimated using three univariate CAR models, the random effects would be spatially correlated within each model, but treated independently between models. Using the multivariate CAR approach, these random effects are dependent both within and between models.

We now introduce the GMCAR structure, which links the previously described models (Jin, Carlin, and Banerjee, 2005). Define

$$
\boldsymbol{\phi}=\left[\begin{array}{l}
\phi_{1} \\
\phi_{2} \\
\phi_{3}
\end{array}\right] \sim N\left(\left[\begin{array}{l}
0 \\
0 \\
0
\end{array}\right],\left[\begin{array}{ccc}
\Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\
\Sigma_{12}^{\prime} & \Sigma_{22} & \Sigma_{23} \\
\Sigma_{13}^{\prime} & \Sigma_{23}^{\prime} & \Sigma_{33}
\end{array}\right]\right)
$$

where each $\Sigma_{i j}$ is an $n \times n$ covariance matrix. We can extend the bivariate case in Jin et al. to the trivariate case, by specifying the conditional distributions,

$$
\begin{gathered}
p\left(\phi_{1} \mid \phi_{2}, \phi_{3}\right) \\
p\left(\phi_{2} \mid \phi_{3}\right)
\end{gathered}
$$

and the marginal distribution

$$
p\left(\phi_{3}\right)
$$

By describing our models as three conditional models, we can easily then write out the joint distribution. The joint distribution becomes:

$$
p(\phi)=p\left(\phi_{1} \mid \phi_{2}, \phi_{3}\right) p\left(\phi_{2} \mid \phi_{3}\right) p\left(\phi_{3}\right)
$$

To describe the conditional distributions, it is easiest to use partitioning to write out the components of the variance matrix. To define $p\left(\phi_{1} \mid \phi_{2}, \phi_{3}\right)$, we can partition the distribution into

$$
N\left(\left[\begin{array}{l}
0 \\
0 \\
0
\end{array}\right],\left[\begin{array}{ccc}
\Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\
\Sigma_{12}^{\prime} & \Sigma_{22} & \Sigma_{23} \\
\Sigma_{13}^{\prime} & \Sigma_{23}^{\prime} & \Sigma_{33}
\end{array}\right]\right)=N\left(\left[\begin{array}{l}
0 \\
0
\end{array}\right],\left[\begin{array}{cc}
\Sigma_{11} & \Sigma_{12: 13} \\
\Sigma_{12: 13}^{\prime} & \Sigma_{22: 33}
\end{array}\right]\right)
$$

where

$$
\Sigma_{12: 13}=\left[\begin{array}{ll}
\Sigma_{12} & \Sigma_{13}
\end{array}\right]
$$

and

$$
\Sigma_{22: 33}=\left[\begin{array}{cc}
\Sigma_{22} & \Sigma_{23} \\
\Sigma_{32} & \Sigma_{33}
\end{array}\right]
$$

Define $A_{2}=\Sigma_{12: 13} \Sigma_{22: 33}^{-1}$ (which will be $n \times 2 n$ ), and we can define

$$
E\left(\phi_{1} \mid \phi_{2}, \phi_{3}\right)=A_{2}\left(\phi_{2}, \phi_{3}\right)
$$

$$
\operatorname{Var}\left(\phi_{1} \mid \phi_{2}, \phi_{3}\right)=\Sigma_{11}-A_{2} \Sigma_{12: 13}^{\prime}
$$

from the bivariate normal conditional distribution.

Define $A_{3}$ as the lower half $(n \times n)$ of $A_{2},\left(\Sigma_{23} \Sigma_{33}^{-1}\right)$. We can also define

$$
E\left(\phi_{2} \mid \phi_{3}\right)=A_{3} \phi_{3}
$$

$$
\operatorname{Var}\left(\phi_{2} \mid \phi_{3}\right)=\Sigma_{22}-A_{3} \Sigma_{23}^{\prime}
$$

and $\phi_{3} \sim N\left(0, \Sigma_{33}\right)$.

We then rewrite the conditional and marginal distributions into the form presented as the univariate case. By writing the distribution in this form, we can interpret the parameters similarly to that of the
univariate case, with a within-model spatial association parameter, $\alpha$, and a precision parameter, $\tau$, which eases the interpretation.

$$
\begin{aligned}
\phi_{1} \mid \phi_{2}, \phi_{3} & =N\left(A_{2}\left(\phi_{2}, \phi_{3}\right),\left[\left(D-\alpha_{1} W\right) \tau_{1}\right]^{-1}\right) \\
\phi_{2} \mid \phi_{3} & =N\left(A_{3} \phi_{3},\left[\left(D-\alpha_{2} W\right) \tau_{2}\right]^{-1}\right) \\
\phi_{3} & \sim N\left(0,\left[\left(D-\alpha_{3} W\right) \tau_{3}\right]^{-1}\right)
\end{aligned}
$$

where $D=\operatorname{Diag}\left(m_{i}\right)$ where $m_{i}$ is the number of neighbors of house $i$ and $W$ is the adjacency matrix (ie. 1 if houses are neighbors and 0 otherwise) and $0<\alpha_{i}<1$.

These models are linked through the matrix $A_{2}$ and $A_{3}$, which describes the between-model spatial association. We define the components of $A_{2}$ and $A_{3}$ as:

$$
a_{i j}= \begin{cases}\eta_{k 0} & \text { if } j=i \\ \eta_{k 1} & \text { if } j \sim i \\ 0 & \text { otherwise }\end{cases}
$$

where $i$ and $j$ index sites 1 to $N$ and $k=2,3$ correspond to the appropriate model link.

We interpret the $\eta$ parameters as the between-model spatial association parameters. The parameter $\eta_{k 0}$ describes the spatial association of a given site to that site in other models. The parameter $\eta_{k 1}$ describes the spatial association of a given to site to neighboring sites between models. And, together with $\alpha$, which describes the within-model association, these parameters capture the system spatial heterogeneity.

We then re-write $A_{2}$ and $A_{3}$, and the expectation of $\phi_{1}$ and $\phi_{2}$ accordingly:

$$
A_{2}=\left[\begin{array}{l}
\eta_{20} I \\
\eta_{30} I
\end{array}\right]+\left[\begin{array}{l}
\eta_{21} W \\
\eta_{31} W
\end{array}\right]
$$

and

$$
A_{3}=\eta_{30} I+\eta_{31} W
$$

Putting all this together, we can write out the joint distribution

$$
\left[\begin{array}{l}
\phi_{1} \\
\phi_{2} \\
\phi_{3}
\end{array}\right] \sim N\left(\left[\begin{array}{l}
0 \\
0 \\
0
\end{array}\right],\left[\begin{array}{ccc}
\Sigma_{11}+A \Sigma_{22: 33} A^{\prime} & A_{2} \Sigma_{22}+A_{3} \Sigma_{33} A_{3}^{\prime} & A_{2} A_{3} \Sigma_{33}+A_{3} \Sigma_{33} \\
\left(A_{2} \Sigma_{22}+A_{3} \Sigma_{33} A_{3}^{\prime}\right)^{\prime} & \Sigma_{22}+A_{3} \Sigma_{33} A_{3}^{\prime} & A_{3} \Sigma_{33} \\
\left(A_{2} A_{3} \Sigma_{33}+A_{3} \Sigma_{33}\right)^{\prime} & \left(A_{3} \Sigma_{33}\right)^{\prime} & \Sigma_{33}
\end{array}\right]\right)
$$

When interpreting the results, it is important to note the conditional interpretation of the parameters. For example, $\alpha_{2}$ describes the spatial association of the number of vectors conditional on the spatial random effects, $\phi_{3}$, estimated for vector presence. Similarly, $\alpha_{1}$ describes the spatial association of the proportion of positive vectors conditional on the spatial effects of the vector distribution, captured by $\phi_{2}$ and $\phi_{3} . \eta_{30}$ describes the spatial association between the same house in the Poisson model with the probability of the house having zero vectors. Similarly, $\eta_{31}$ describes the spatial association between a house in the Poisson model with the neighboring houses of the probability of having zero vectors.

In addition, the order of the models is important. There is a one-directional interpretation of the conditional relationship. While in some contexts, this may be a barrier, the order of the distributions aligns well with the application of parasite infection and vector distributions - it makes sense to interpret parasite presence conditional on the vector distributions, but not vice versa.

We put prior distributions on parameters. We use normal priors with precision of 0.1 for the regression coefficients and a precision of 1 for the linking $\eta$ parameters, all with mean 0 . We tested simulated data with smaller precisions, but the results did not vary much for precisions smaller than those we chose. We use gamma prior distributions with shape of 1 and scale of 50 on the precision parameters $\tau$, and we use uniform priors $[0,1]$ on the within-model spatial association parameters $\alpha$. The model was sensitive to priors on the precision and $\alpha$ parameters. We chose a gamma prior to keep the precision small, which allows the probabilities to vary from 0 to 1 , but with values distributed throughout that range and not only at the extremes. We chose an uninformative prior
for $\alpha$ that restricted the possible values to those used by other authors (Jin, Carlin, and Banerjee, 2005).

The simulations and data analysis are conducted using MCMC algorithms in OPENBugs. In OPENBugs, the built-in function car.proper is used to define the spatial associations within models, which decreases the computation time to run the analysis. The computational time is heavily dependent on the size of the dataset and the number of neighbors per each site.

### 4.3. Simulations

To conduct simulations, we randomly simulate $n$ sites on an $n \times n$ grid. We define a neighbor as sites within a radius of $n / 8$ and require each site to have at least one neighbor. We generate simulated outcome variables of interest, $Y$ and $Z$, the count of vectors and positive vectors, using the conditional distributions. Each simulated dataset was run with $n=160$ for 10,000 iterations, with a burn-in period of 5,000. All parameter estimates converged with a Gelman-Rubin statistic $<1.10$.

We first run our model on simulated data to verify the model and examine the computational properties. We run 40 simulated datasets under 2 parameter regimes to ensure the model is working correctly. In Table 4.1 we summarize our simulation results. Overall, the model estimates the parameters well, although there is some variation in the estimates compared to the true value. The estimates of the $\beta$ converged quickly to the true values. The estimates of $\alpha$, in particular for small values, were sensitive to changes in the true value of $\tau$. In general, the estimates of $\tau$ and $\alpha$ were correlated. The credible intervals for $\alpha$ were very large, however they did converge well. For all parameters, the median of the $95 \%$ credible interval contained the true parameter value.

|  | True Value | Estimate (Median 95\% CI) | True Value | Estimate (Median 95\% CI) |
| ---: | ---: | ---: | ---: | ---: |
| $\alpha_{1}$ | 0.20 | $0.11(0.02,0.71)$ | 0.20 | $0.33(0.01,0.72)$ |
| $\alpha_{2}$ | 0.50 | $0.43(0.03,0.87)$ | 0.80 | $0.73(0.15,0.93)$ |
| $\alpha_{3}$ | 0.80 | $0.77(0.23,0.96)$ | 0.20 | $0.23(0.01,0.63)$ |
| $\beta_{10}$ | 1.00 | $1.05(0.40,1.59)$ | -1.00 | $-0.80(-1.46,-0.18)$ |
| $\beta_{20}$ | 3.00 | $3.04(2.64,3.39)$ | 3.00 | $3.05(2.70,3.40)$ |
| $\beta_{30}$ | 1.00 | $0.70(0.20,1.42)$ | 0.00 | $0.07(-1.17,1.01)$ |
| $\eta_{20}$ | 0.20 | $0.27(-0.24,0.70)$ | 0.30 | $0.16(-0.75,1.12)$ |
| $\eta_{21}$ | 0.05 | $0.03(-0.09,0.18)$ | 0.10 | $0.12(-0.24,0.52)$ |
| $\eta_{30}$ | 0.20 | $0.16(-0.17,0.57)$ | 0.01 | $-0.01(-0.15,0.11)$ |
| $\eta_{31}$ | 0.05 | $0.08(0.01,0.28)$ | 0.01 | $0.003(-0.02,0.03)$ |
| $\tau_{1}$ | 0.10 | $0.11(0.07,0.15)$ | 0.08 | $0.06(0.04,0.09)$ |
| $\tau_{2}$ | 0.40 | $0.44(0.29,0.91)$ | 1.00 | $1.15(0.69,2.32)$ |
| $\tau_{3}$ | 0.06 | $0.16(0.02,0.78)$ | 0.08 | $0.02(0.002,0.09)$ |

Table 4.1: Median estimates with median of the $95 \%$ credible interval of 40 simulated datasets.

### 4.4. Data Results

The dataset analyzed contains 577 sites within 67 houses (Figure 4.1). Each site has a unique set of coordinates. Vectors were found at 98 of these sites; the median number of $T$. infestans at sites where at least one triatomine was found was 4 , and the maximum number of triatomines found in one site was 1106. Of all triatomines analyzed, there were 29 sites containing triatomines with $T$. cruzi. Of sites with at least one triatomine infected with $T$. cruzi, the median proportion of infected triatomines was 0.50 , but in some sites $100 \%$ of triatomines collected contained parasite.

The dataset is analyzed on the site-level; multiple sites are contained within a house, such as rooms of a house and outdoor spaces. We ran 10,000 iterations and used a burn-in period of 3,000 iterations, and all parameters converged with a Gelman-Rubin statistic of $<1.10$. The following models are used for the analysis of the data:

$$
\begin{equation*}
\operatorname{logit}(\pi)=\beta_{30}+\beta_{31} \text { Cuy }+\beta_{32} \text { Dog }+\beta_{33} \text { Poultry }+\beta_{34} \text { Materials } 1+\beta_{35} \text { Materials2 }+\phi_{3} \tag{4.1}
\end{equation*}
$$

$$
\begin{equation*}
\log (\lambda)=\beta_{20}+\beta_{21} \text { Cuy }+\beta_{22} \operatorname{Dog}+\beta_{23} \text { Poultry }+\beta_{24} \text { Materials } 1+\beta_{25} \text { Materials2 }+\phi_{2} \tag{4.2}
\end{equation*}
$$

$$
\begin{equation*}
\operatorname{logit}(p)=\beta_{10}+\beta_{11} \text { Cuy }+\beta_{12} \text { Dog }+\beta_{13} \text { Poultry }+\beta_{14} \text { Materials1 }+\beta_{15} \text { Materials2 }+\phi_{1} \tag{4.3}
\end{equation*}
$$

These models contain five covariates that have previously been associated with our outcomes (Levy et al., 2006). We incorporate the presence of guinea pigs, or cuy, at each site, as well as the presence of dogs, and chickens or ducks (poultry). The variable 'Materials1' defines the presence of walls or animal corrals made of adobe, cement blocks with or without mortar, and bricks with or without mortar. These are materials that are good habitats for T. infestans. 'Materials2' defines the presence of walls or animal corrals out of materials that are poor habitats for $T$. infestans, including wire cage and smooth stucco.

To fit the data, we fit three variations of the multivariate model. First, we fit the full model, allowing all parameters to vary. We compare this model to a reduced multivariate model, where we fix $\eta_{i 1}=0$ for $i=1,2$, but allow $\eta_{i 0}$ to vary. In the reduced model, by fixing $\eta_{i 1}=0$ we do not directly link sites to neighboring sites between models. However, we still allow sites to directly link to themselves between models. In this reduced models, sites are still indirectly linked to neighboring sites through the within spatial association parameter, $\alpha$. We also compare these models to independent CAR models, where $\eta_{i j}=0$ for all $i$ and $j$. The results from the data analysis are summarized in Table B.4. We report the DIC for each model, and use the model with the smallest DIC as the model as best fit. We define DIC as

$$
D I C=D(\theta)+\operatorname{Var}(\hat{D}(\theta))
$$

where $D(\theta)=-2 l l$ (Gelman et al., 2014).

We run the model on our dataset and report the results in Table 4.2. From the DIC, it is clear the partially multivariate model, where sites are only linked between models to themselves, fits the data better than the fully multivariate model, where sites are linked between models both to themselves and neighboring sites (Table 4.2). Both the fully and partially multivariate models fit the data better than the independent model, where there is only within-model spatial association and no link between models. From these results, we visualize the predicted probability of vector presence, expected number of vectors conditional on the expected presence of vectors, and expected proportion of vectors that contain parasite conditional on the expected distribution of vectors (Figure 4.2).


Table 4.2: Data results using three models (Eq. 4.1, 4.2, 4.3). The full model allows all parameters to vary. The partial model allows $\eta_{20}$ and $\eta_{30}$ to vary, but fixes $\eta_{21}=0$ and $\eta_{31}=0$. The independent model fixes all linking parameters, $\boldsymbol{\eta}=\mathbf{0}$. In all models, covariables are the same: $\beta_{i 0}$ corresponds to the model intercept, $\beta_{i 1}$ is the estimated effect of guinea pigs (cuy), $\beta_{i 2}$ is the estimated effect of dogs, $\beta_{i 3}$ is the estimated effect of poultry, $\beta_{i 4}$ estimates the effect of the presence of housing materials that are good habitats for vectors, and $\beta_{i 5}$ estimates the effect of the presence of housing materials that are poor habitats for vectors for $i=1,2,3$. Coefficient credible intervals of covariables that did not contain 0 are highlighted in red.

### 4.5. Discussion

The GMCAR approach enables a desirable conditional interpretation of spatial parameters, which is not possible using independent models, due to inherent dependencies and associations between the biological components of the parasite-vector system. Vector-borne pathogens have complex spatial dynamics, and it is difficult to tease apart factors associated with different components of the transmission process. However, identifying variables associated with specific elements of the vector-borne cycle may help to target these elements for effective control measures. Using this conditional modeling approach, we can describe our data on Chagas disease in parts: presence of vectors, number of vectors, and number of infected vectors, and condition our models accordingly.


Expected probability of vector presence



Figure 4.2: Unfilled gray points indicate all sites. (a) Expected probability of vector presence at each site (size of red point relative to gray point) (b) Expected count of vectors, conditional on the probability of vector presence (size proportional to log count) (c) Expected proportion of positive vectors, conditional on expected vector distribution (size of gold point relative to red point). Red point is relative to the size of the log of the number of observed T. infestans at that site.

We determined the reduced multivariate GMCAR model fit the data better than the full multivariate GMCAR model and independent model because it had the smallest DIC. In the reduced multivariate model, we observed that housing materials with non-smooth surfaces, including stacked bricks and cement blocks, were positively associated with the presence of vectors, but not our other outcomes. This finding supports those found by previous studies, and supports the hypothesis that $T$. infestans prefer dwelling in the dark cracks and crevices of these materials (Levy et al., 2006, Barbu et al., 2014). In this model, we also found that guinea pigs, or cuy, were positively associated with both the number of vectors observed and the proportion of vectors that are infected with T. cruzi. This finding supports previous studies and the hypothesis that $T$. infestans use guinea pigs as a food source (Levy et al., 2006, Barbu et al., 2014). Often, guinea pig corrals are made of stacked bricks or cement blocks, which was found to be associated with the presence of vectors. Thus, the combination of guinea pigs and corrals made from these materials may fuel vector growth. We found that housing materials that are poor habitats for vectors, including chicken wire and smooth stucco, were negatively associated with the number of $T$. infestans expected at a given site. This finding supports previous studies that have found that vectors were less likely to be found in these environments, as they did not contain dark cracks or crevices the insects prefer (Levy et al., 2006).

Interestingly, we found that the same covariates that were associated with the number of bugs in a given site, were also associated with the proportion of vectors that were infected with parasite. We found that guinea pigs were positively associated with the proportion of vectors that were infected with T. cruzi, and poor habitat materials were negatively associated with the proportion of vectors that were infected.

In the reduced multivariate GMCAR model, both linking parameters, $\eta_{20}$ and $\eta_{30}$ were positive, suggesting positive spatial association between both the number of vectors and vector presence, as well as the proportion of vectors infected with $T$. cruzi and the distribution of vectors. In addition, the smaller DIC of the reduced multivariate model compared to the full model indicates that outcomes are linked between models for a given site, but are not directly linked between outcomes between neighboring sites. However, there is still spatial association between outcomes between neighboring sites indirectly in the reduced model, through the $\eta$ and $\alpha$ parameters.

These results highlight the dependencies of the outcomes - there is significant spatial association between them. In Figure 4.1, we can see the spatial clustering of the vectors and positive vectors. However, it is difficult to visually assess whether there is additional spatial association within the proportion of vectors that are infected with $T$. cruzi, once accounting for the spatial association of vectors in general. In our analysis, we estimate $\alpha_{1}=0.98$, suggesting high spatial association of positive vectors, conditional on the distribution of vectors. This analysis suggests it is important to find infected vectors to reduce Chagas disease transmission. In Arequipa, vectors are present in low levels, and positive vectors are rare. Infected vectors tend to be highly spatially clustered, and it is essential to find infected clusters to eliminate the presence of infected vectors.

Compared to analyzing the data independently, we were able to identify more significant covariables using the multivariate model. In the independent model, poor habitat materials was not significantly associated with lower numbers of vectors. However, this covariable was significantly associated with lower number of vectors in the multivariate analysis.

This analysis had several limitations. At each site, we only considered T. infestans that were then tested for $T$. cruzi. 14\% of captured $T$. infestans were not tested for the presence of $T$. cruzi and were not used in this analysis. There may be some sampling error, as young T. infestans are difficult to observe due to their small size. In addition, we limited our analysis to a small subset of data due to
the computational time of the analysis. In the future, we hope to use this method on the full dataset and analyze the results.

Our results suggest that covariables associated with large numbers of $T$. infestans are also associated with high proportions of infected vectors. Thus, control efforts should be targeted at controlling large populations of $T$. infestans to limit the transmission of $T$. cruzi. In addition, due to the high spatial associations both within and between outcomes, searches for vectors should be focused around sites that are known to be infested with vectors. We hope to expand our models by continuing to identify covariables that are associated with our outcomes to develop the model of best fit. In addition, we could extend to the model to include more components of the biological system, such as $T$. cruzi infected host data.

## CHAPTER 5

## Discussion

### 5.1. Conclusions

In this dissertation, we developed Bayesian methods to detect and control the insect vector, Triatoma infestans in the city of Arequipa, Peru. In addition, we presented a method to model the proportion of vectors infected with Trypansoma cruzi, the parasite that causes Chagas disease, conditional on the distribution of insect vectors. In Chapter 2, we developed a novel stochastic epidemic model that we used to guide searches for T. infestans in Arequipa, Peru. Stochastic epidemic models are particularly difficult to fit to real-time data due to the detailed-level of data needed for a tractable likelihood. Using a reversible-jump Markov chain Monte Carlo algorithm, we were able to augment the data and retain a tractable likelihood even though we did not observe the true infestation time of each house. To augment the data, we presented a house-level insect population growth model that accounted for the number of insects each house was able to support (ie. house-level carrying capacity), assuming additional insects dispersed to neighboring houses. In addition, we incorporated a spatial kernel that assumes insects were more likely to disperse to neighboring houses, and further were more likely to disperse to houses on the same city block than those across city streets. We demonstrated our method on simulated data, and piloted the algorithm over three months in the field. During the pilot, inspectors searched houses with the highest posterior probabilities of infestation, and we updated the model to reflect the latest observed data. No positive infestations were found during the pilot, and it may be useful to use this method again in an area of the city with higher prevalence of infestations. Alternatively, it may be interesting to use other measurements to determine high risk houses. For example, houses with large infectious pressure, or houses that would are most infectious, may be most important to treat.

During this pilot, we encountered implementation barriers, such as long computation times, convergence issues, and inefficient search paths throughout the city. To address these constraints, we worked to develop an algorithm, presented in Chapter 3, that was more computationally efficient and could be used with efficient search paths.

In Chapter 3, we presented an alternative algorithm that fit the data using iterative nested Laplace
approximations (INLA), an alternative to the traditional MCMC algorithms. Using this estimation approach, we were able to fit much larger datasets in minutes. In addition, we presented our results as a risk map, rather than an ordered list of houses, and we are currently piloting this approach in the field. Using the risk map, inspectors can use the visualization of house-level risk to guide their path, rather than searching houses in a specific order, such as a list. Using a map, inspectors can search a given locality in a path that is most convenient for them, while still using the latest observed data.

Using INLA, we developed a novel method to incorporate the urban landscape into a geostatistical model. Gaussian fields (GFs) are a popular tool used for creating risk maps of point-level data. However, GFs assume a smooth landscape. Previous studies showed the urban grid of Arequipa significantly affected the distribution of $T$. infestans; insects were more likely to move within a city block than between city blocks (cite). The objective of our method was to incorporate city streets into a logistic Gaussian field, treating the streets as permeable barriers in the distribution of vectors. To create permeable barriers, we presented an approach, with one additional parameter, $S$, that distorted the city map by widening the distance between city blocks but retained the distance between houses within the city block. We demonstrated the identifiability of the distortion parameter, $S$, through simulations, and applied the method to a district of Arequipa. We found that the effect of the width of streets was equivalent to at least 1.5 times the true width. In other words, for a given distance, houses located on the same block had higher spatial correlation than houses located on different city blocks. Using the results from this approach, we are currently implementing this model in the field; inspectors can view the current risk-map on tablets or smart phones in the field, enter the results, and the model can be updated each night. Based on the outcome of the pilot study, we can continue to improve the model as needed.

Using this approach, there were some patterns of data that resulted in an unidentifiable distortion parameter, $S$. In these cases, the log-likelihood continued to increase over $S$, suggesting streets were impermeable barriers. By examining these patterns individually, we suspect both very sparse and heavily saturated landscapes may have unidentifiable $S$. In cases where a map is very sparsely infested, it may be more beneficial to conduct ring-searches around houses that were observed as infested. In cases where the map is heavily saturated, it is most likely necessary to inspect as many houses as possible. Thus, this identifiability restraint may not have implications for field inspections.

In Chapter 4, we presented an approach to jointly model the distribution of vectors, and vectors infected with T. cruzi. Analyzing the patterns of vector-borne disease pathogens is especially difficult due to the complex transmission dynamics. For example, T. cruzi can infect many hosts, including humans, dogs, and guinea pigs, all prevalent in Arequipa, Peru, but must be transmitted between hosts through the vector. Thus, it is difficult to statistically differentiate factors associated with vectors, and infected vectors.

In our approach, we had three unique spatial random effects in three distinct models: the probability of vector presence, the expected count of vectors (together a zero-inflated Poisson), and the proportion of infected vectors expected given the distribution of observed vectors. We then linked these models through additional spatial parameters, capturing spatial dependencies between models. Thus, in addition to covariables, the models accounted for both within-model spatial correlation and between-model spatial correlation. Using this structure, we were able to interpret the models conditionally on each other. Thus, the covariate effects associated with the number of vectors was interpreted conditional on the probability of vector presence. Similarly, the covariates associated with observed $T$. cruzi infected vectors was conditional on the distribution of all vectors. This conditional interpretation of parameters is desirable for modeling infected vectors, as it is difficult to decipher variables associated with infected vectors compared to those associated with vectors in general, especially given the highly spatial nature of the data. In the future, more consideration into model covariables may provide insight into stronger relationships with vectors and infected vectors.

### 5.2. Limitations

Throughout this work, we encountered many statistical and epidemiological challenges. In Chapter 2, we attempted to implement a sophisticated, yet computational intensive model. Using this approach, we were able to capture the spatio-temporal dynamics using compartmental stochastic methods. To implement this type of model using limited data, we made many strong assumptions. If these assumptions were incorrect, the model may not perform well. Since our data was limited, it was difficult to verify some of assumptions, such as the rate of vector growth in houses. Without strong prior information and limited data, it may not be possible to apply this type of model. In addition, we encountered difficulties with convergence. This model was developed for use with a real-time field implementation, and the computational barriers were significant. The model had to
be run for at least twelve hours on a cluster for adequate convergence, although after this amount of time convergence was still not ideal.

In the third chapter, we developed an alternative approach where we attempted to improve upon these restraints. The logistic spatial model we developed used less assumptions and was significantly faster computationally. However, this model does not incorporate the dynamic aspect of the infectious process that was captured using the stochastic model. It is difficult to assess how much detail we lose by not incorporating the dynamic aspect of the infectious process. The model presented in Chapter 3 does not incorporate vector counts, however this model could easily be extended to a Poisson regression to do so. In the future, a comparison of these approaches, in particular comparing the stochastic epidemic model to the Poisson, urban-grid, Gaussian field model, may be valuable to quantify the differences between these approaches.

These models also contain many additional constraints. Often, much more data was collected than we were able to incorporate into the models. Inspectors take detailed notes of each site, including housing and animal corrals, inhabitants, and more. We were only to incorporate some of the collected data, and we may have missed key observations in our analysis. Inspectors may observe characteristics of a site that are not able to be quantified or recorded in a dataset easily. In addition, participation of residents in vector surveillance was limited. Many residents do not wish to have their homes inspected with such detail. It may be beneficial to incorporate participation analysis into future models to add additional information. Previous studies have shown that houses that do not participate are less likely to be infested (Hong et al., 2015). Therefore, this information could be used to add information to models where many houses have not been inspected.

Additional details were not included in our model, such as quality of inspections, inspector sensitivity, and quality of treatments. Often, inspectors are able to search part of a residence, but not the entire premises; including this information in the model is difficult. Previous studies have shown inspections may vary by inspector (Hong et al., 2015).

### 5.3. Future directions

In this dissertation, we attempted to capture spatial heterogeneity to guide vector search inspections. The goal of our work was to find the highest risk houses. In Chapter 2, we presented a
method that incorporated the number of vectors found at each house, which seems to be important in the probability that a given house infects a neighbor. In Chapter 3, we presented a more computationally efficient approach. While the approach in Chapter 3 incorporates a more detailed spatial model, it simply models the probability of vector presence and does not incorporate other covariates, such as the number of vectors observed at infested houses, housing material, or animals present. Including these additional elements into this spatial model may result in a better risk map, which may provide more information to inspectors. It may be possible to incorporate multiple types of barriers into this Gaussian field model. For example, in Arequipa, there are both streets and water channels. It may be possible to first distort the map by city block and estimate $S_{1}$. Then, by subsequently distorting the map by regions divided by water channels, it may be possible to estimate a second barrier, $S_{2}$.

In Chapter 4, we presented a method to jointly, spatially model the presence of vectors, the count of vectors, and the proportion of vectors infected with T. cruzi. Using this approach, variable selection is complex, and more thought should be put into the covariables used. In addition, the analysis was done on the site level. However, it may be important to include additional effects on the houselevel, since sites within a house are more likely related than sites between houses. Lastly, it may be possible to incorporate a more sophisticated spatial correlation structure into the joint model. We used a conditional autoregressive model, but it maybe possible to jointly model these outcomes using a Gaussian field approach.

Using our presented methods, inspectors searched the highest-risk houses for infestations. However, we suspect there is value in inspecting both high-risk houses, and houses where the risk level is unknown. For example, an area in which little inspection has been done may become high risk if inspected. Using our approaches, these areas may never be inspected. Thus, in the future, work should be done to examine the optimal amount of time that should be spent searching around high-risk houses (ie. 'exploitation' of data) compared to the amount of time that should be spent searching areas with little information (ie. 'exploring' to collect data). This balance may be challenging to quantify - how do we measure the gain in information of each inspected houses? A positive house provides important data, but negative houses also provide data, especially in areas where no inspections have been done.

We have not addressed or answered the question of whether a unified search strategy is beneficial.

For ease of implementation, a unified, continuing search strategy is clear to inspectors, who know what to expect and gain experience. However, switching search strategies at different stages of an epidemic may result in finding more infested houses. In the beginning of an epidemic, it may be better to 'explore' and gather as much information as possible. Later in an epidemic, it may be better to 'exploit' and use what has been collected.

Currently, we are piloting a tablet and smartphone application that creates a real-time visualization of a risk-map for inspectors to use in the field. Using this application, inspectors searching the landscape for vectors, can see house-level risk in the current location using GPS tracking. The inspectors can also input the results of their work directly into the application for real-time model updates. Each time a model is re-run incorporating new results, the inspectors instantly see the new risk-map on the tablet or smart phone. In this pilot, we are using a variation of the model presented in Chapter 3 of this dissertation. This space-time variation of the model in INLA incorporates a map distortion of $S=1.5$ (see Chapter 3 for more details on this method and parameter), and inspection data that has been collected since 2004 to predict the presence of vectors on a locality level in Arequipa. In this model, we also use covariate information from the spray campaign in 2004, including insecticide participation history. From the results of this pilot, we hope to identify further future directions for our research. We have not yet studied how to motivate inspectors to use the risk-map provided. Inspectors may not be interested in traveling longer distances to inspect more high-risk houses, compared to inspecting more low-risk houses in a smaller area. We are interested in studying how inspectors use the information presented in the risk-map, and how different motivation schemes may affect their search path.

Throughout this dissertation, we applied our methods to control the re-emerging vector of Chagas disease, T. infestans in Arequipa, Peru. However, our methods are appropriate to many types of vector-borne diseases, including Chikungunya, West Nile, and Dengue, as well as the recent outbreak of the Zika virus. Vector-borne diseases are increasingly occurring in cities, facilitated by rapid transmission and creating new challenges. Understanding how to efficiently control these outbreaks is becoming increasingly crucial as diseases travel quickly between cities and countries. As data is collected faster and better, updating risk maps and models in real time is essential to understand the current state of any epidemic. In the future, we hope to extend our methods to other diseases and vectors, with unique spatial kernels and population dynamics. Our methods have the
potential to assess real-time risk across large sets of data across many types of urban outbreaks.

## APPENDIX A

## Chapter 2 Details

## A.1. Growth Dynamics



Figure A.1: Bounded growth using the Beverton-Holt model compared to unbounded growth. Difference in slopes at time of new infestation was incorporated into hazard function to quantify infestation severity into the probability of infesting a neighboring house.

## A.2. Sensitivity Analysis

For all simulations and data applications, we fixed the carrying capacity, $K=1000$. We did some sensitivity analyses to assess the importance of this assumption. We ran 3 RJMCMC chains on one locality with 5 different carrying capacities, $K=\{100,500,800,1000,1500\}$. We obtained the median ranking of each house across the chains for each $K$, and then plotted the rankings against each other (Figure A.3). We can see the heterogeneity in ranking is similar to that between chains within each carrying capacity. Interestingly, the rankings stayed the same between $K=1000$ and $K=1500$.


Figure A.2: Ranking of each house across 3 RJMCMC chains. There were a few houses that changed significantly, but most houses remained within a few rankings between chains. We use the median ranking between chains to give inspectors.


Figure A.3: Ranking of each house across 5 potential carrying capacities. There were a few houses that changed significantly, but most houses remained within a few rankings between carrying capacity values.

## APPENDIX B

## Chapter 3 Details

## B.1. Detailed simulation results

Our simulation studies suggest we need a minimum amount of observed information to identify the scale parameter. More research must be done to identify the requirements for identifiability of $S$. Our initial investigations suggest that the scale is identifiable when at least $2 \%$ of houses are infested, however this approximation also seems sensitive to the specific infestation pattern. The requirements may vary by the specific map and model used. In our testing, when the parameter is unidentifiable, the log-likelihood sharply increases near the true value but then does not decrease as the scale increases. It may be possible to identify specific patterns when these unidentifiable cases occur. As a general rule, we noticed unidentifiable likelihoods in cases with very low levels of infestation. Occasionally, we observed an unidentifiable likelihood when there was a high number of infested houses.
$\kappa=0.01, \sigma_{u}^{2}=5, S=2.5$ are fixed. The intercept $\beta_{0}$ varies from -6 to -3 . The intercept plays a

Table B.1: Simulation results with the variation in intercept $\beta_{0}$.

| True values |  |  |  |  |  |  | Estimates |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\kappa$ | $\sigma_{u}^{2}$ | $S$ | $\beta_{0}$ | $\hat{S}$ | $\hat{\beta}_{0}$ | Coverage $\left(\beta_{0}\right)$ | Identification |  |
| 0.01 | 5 | 2.5 | -3 | $2.68(0.48)$ | $-2.92(0.35)$ | 0.96 | 0.76 |  |
|  | 10 |  |  | $2.69(0.55)$ | $-2.99(0.49)$ | 0.95 | 0.92 |  |
|  | 25 |  |  | $2.65(0.45)$ | $-3.11(0.79)$ | 0.97 | 0.93 |  |
|  | 50 |  | $2.65(0.40)$ | $-3.42(1.21)$ | 0.95 | 0.94 |  |  |
|  | 100 |  | $2.67(0.44)$ | $-3.31(1.86)$ | 0.95 | 0.98 |  |  |
|  | 200 |  |  | $2.64(0.38)$ | $-4.02(3.24)$ | 0.97 | 0.98 |  |

Table B.2: Simulation results with the variation in intercept $\kappa$.

| True values |  |  |  |  |  | Estimates |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\kappa$ | $\sigma_{u}^{2}$ | $S$ | $\beta_{0}$ | $\hat{S}$ | $\hat{\beta}_{0}$ | Coverage $\left(\beta_{0}\right)$ | Identification |
| 0.001 | 5 | 1.5 | -3 | $1.42(0.52)$ | $-2.35(4.19)$ | 0.93 | 0.69 |
| 0.002 |  |  |  | $1.52(0.51)$ | $-2.56(1.34)$ | 0.98 | 0.89 |
| 0.005 |  |  |  | $1.57(0.35)$ | $-3.11(0.89)$ | 1.00 | 0.95 |
| 0.01 |  |  |  | $1.64(0.27)$ | $-2.97(0.53)$ | 0.97 | 1.00 |
| 0.001 | 10 | 1.5 | -3 | $1.70(0.64)$ | $-4.24(6.78)$ | 0.97 | 0.73 |
| 0.002 |  |  |  | $1.55(0.45)$ | $-2.63(1.99)$ | 1.00 | 0.93 |
| 0.005 |  |  |  | $1.61(0.33)$ | $-3.13(1.26)$ | 0.97 | 0.98 |
| 0.01 |  |  |  | $1.58(0.19)$ | $-3.16(0.77)$ | 0.96 | 1.00 |

Table B.3: Simulation results with the variation in intercept $\beta_{0}$.

| True values |  |  |  |  |  |  | Estimates |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\kappa$ | $\sigma_{u}^{2}$ | $S$ | $\beta_{0}$ | $\hat{S}$ | $\hat{\beta}_{0}$ | Coverage $\left(\beta_{0}\right)$ | Identification |  |
| 0.005 | 5 | 2.5 | -3 | $2.76(0.48)$ | $-2.91(0.35)$ | 0.96 | 0.76 |  |
|  |  |  | -4 | $2.64(0.53)$ | $-4.07(0.46)$ | 0.98 | 0.69 |  |
|  |  |  | -5 | $2.52(0.58)$ | $-5.06(0.67)$ | 0.92 | 0.66 |  |
|  |  |  | -6 | $2.12(0.74)$ | $-6.25(1.17)$ | 0.94 | 0.52 |  |

important role that decides the total number of infestations. A larger value of the intercept $\beta_{0}$ implies more infestations. Table B. 3 shows the simulation results with the variation in $\beta_{0}$. As $\beta_{0}$ decreases, the variation in the estimation of $S$ increases and the identification rate decreases.


Figure B.1: Log-likelihood analysis across different scales, $S$. In most cases, $S$ is clearly identifiable, but in some cases (red rectangles) it is not.

## B.2. Additional Data Results



Figure B.2: Log-likelihood analysis across different scales, $S$. Likelihood is maximized at $S=1.5$ indicating the model of best fit.

Table B.4: Results of grid sampling $S$ on dataset, including estimates and standard deviations of $\theta_{1}, \theta_{2}$, and $\beta_{0}$. The log-likelihood is maximized when $S=1.5$, and the bolded results are reported in the main text.

| $S$ | $\hat{\theta}_{1}$ | $\hat{\theta}_{2}$ | $\hat{\beta}_{0}$ | $l l$ |
| :---: | :---: | :---: | :---: | :---: |
| 1.0 | $2.06(0.14)$ | $-4.30(0.14)$ | $-6.14(0.45)$ | -1809.18 |
| 1.1 | $2.08(0.15)$ | $-4.36(0.17)$ | $-6.16(0.44)$ | -1799.65 |
| 1.2 | $2.13(0.16)$ | $-4.41(0.20)$ | $-6.15(0.43)$ | -1795.19 |
| 1.3 | $2.19(0.18)$ | $-4.47(0.21)$ | $-6.13(0.42)$ | -1793.33 |
| 1.4 | $2.36(0.23)$ | $-4.63(0.21)$ | $-6.10(0.42)$ | -1793.35 |
| 1.5 | $2.35(0.18)$ | $-4.63(0.17)$ | $-6.10(0.41)$ | -1791.74 |
| 1.6 | $2.45(0.19)$ | $-4.71(0.19)$ | $-6.07(0.42)$ | -1793.44 |
| 1.7 | $2.49(0.18)$ | $-4.75(0.19)$ | $-6.06(0.42)$ | -1793.56 |
| 1.8 | $2.54(0.19)$ | $-4.79(0.19)$ | $-6.06(0.42)$ | -1793.02 |
| 1.9 | $2.61(0.19)$ | $-4.86(0.18)$ | $-6.04(0.42)$ | -1794.38 |
| 2.0 | $2.66(0.19)$ | $-4.91(0.18)$ | $-6.03(0.42)$ | -1794.74 |
| 2.1 | $2.71(0.19)$ | $-4.96(0.18)$ | $-6.03(0.42)$ | -1794.87 |
| 2.2 | $2.77(0.19)$ | $-5.00(0.17)$ | $-6.02(0.43)$ | -1795.97 |
| 2.3 | $2.82(0.15)$ | $-5.07(0.15)$ | $-6.00(0.42)$ | -1795.65 |
| 2.4 | $2.85(0.20)$ | $-5.09(0.18)$ | $-6.01(0.42)$ | -1796.46 |
| 2.5 | $2.88(0.21)$ | $-5.12(0.19)$ | $-6.01(0.42)$ | -1796.70 |
| 2.6 | $2.95(0.20)$ | $-5.15(0.18)$ | $-6.01(0.43)$ | -1797.74 |
| 2.7 | $3.00(0.19)$ | $-5.19(0.17)$ | $-6.00(0.43)$ | -1798.51 |
| 2.8 | $3.02(0.15)$ | $-5.26(0.15)$ | $-5.99(0.42)$ | -1797.84 |
| 2.9 | $3.06(0.20)$ | $-5.25(0.18)$ | $-6.00(0.43)$ | -1799.25 |
| 3.0 | $3.11(0.16)$ | $-5.31(0.15)$ | $-5.99(0.43)$ | -1799.15 |
| 3.1 | $3.11(0.16)$ | $-5.36(0.16)$ | $-5.97(0.43)$ | -1799.84 |
| 3.2 | $3.16(0.16)$ | $-5.34(0.15)$ | $-5.98(0.43)$ | -1800.82 |
| 3.3 | $3.21(0.16)$ | $-5.38(0.15)$ | $-5.98(0.44)$ | -1801.66 |
| 3.4 | $3.24(0.16)$ | $-5.40(0.15)$ | $-5.97(0.44)$ | -1802.66 |
| 3.5 | $3.30(0.15)$ | $-5.48(0.14)$ | $-5.97(0.44)$ | -1802.78 |
| 3.6 | $3.31(0.17)$ | $-5.48(0.16)$ | $-5.97(0.45)$ | -1803.63 |
| 3.7 | $3.37(0.15)$ | $-5.54(0.14)$ | $-5.96(0.45)$ | -1804.08 |
| 3.8 | $3.35(0.16)$ | $-5.68(0.19)$ | $-5.93(0.44)$ | -1803.83 |
| 3.9 | $3.39(0.16)$ | $-5.72(0.19)$ | $-5.93(0.44)$ | -1804.50 |
| 4.0 | $3.40(0.17)$ | $-5.76(0.19)$ | $-5.92(0.45)$ | -1805.19 |
|  |  |  |  |  |

## BIBLIOGRAPHY

Adigun, AB, Gajere, EN, Oresanya, O, and Vounatsou, P (2015). Malaria risk in Nigeria: Bayesian geostatistical modelling of 2010 malaria indicator survey data. Malaria journal 14.1, 1.

Andersson, H and Britton, T (2012). Stochastic epidemic models and their statistical analysis. Vol. 151. Springer Science \& Business Media.

Barbu, C, Dumonteil, E, and Gourbière, S (2010). Characterization of the dispersal of non-domiciliated Triatoma dimidiata through the selection of spatially explicit models. PLoS Negl Trop Dis 4.8, e777.

Barbu, CM, Hong, A, Manne, JM, Small, DS, Calderón, JEQ, Sethuraman, K, Quispe-Machaca, V, Ancca-Juárez, J, Carpio, JGC del, Chavez, FSM, et al. (2013). The effects of city streets on an urban disease vector. PLoS Comput Biol.

Barbu, CM, Buttenheim, AM, Pumahuanca, M-LH, Calderón, JEQ, Salazar, R, Carrión, M, Rospigliossi, AC, Chavez, FSM, Alvarez, KO, Carpio, JC del, et al. (2014). Residual Infestation and Recolonization during Urban Triatoma infestans Bug Control Campaign, Peru. Emerging infectious diseases 20.12, 2055.

Beaumont, MA (2010). Approximate Bayesian computation in evolution and ecology. Annual review of ecology, evolution, and systematics 41, 379-406.

Becker, NG (1989). Analysis of infectious disease data. Vol. 33. CRC Press.
Bern, C (2015). Chagas' disease. New England Journal of Medicine 373.5, 456-466.
Besag, J (1974). Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society. Series B (Methodological), 192-236.

Besag, J (1975). Statistical analysis of non-lattice data. The statistician, 179-195.
Beverton, RJ and Holt, SJ (2012). On the dynamics of exploited fish populations. Vol. 11. Springer Science \& Business Media.

Blangiardo, M, Cameletti, M, Baio, G, and Rue, H (2013). Spatial and spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology 7, 39-55.

Bowman, NM, Kawai, V, Levy, MZ, Del Carpio, JGC, Cabrera, L, Delgado, F, Malaga, F, Benzaquen, EC, Pinedo, VV, Steurer, F, et al. (2008). Chagas disease transmission in periurban communities of Arequipa, Peru. Clinical Infectious Diseases 46.12, 1822-1828.

Brooker, S (2007). Spatial epidemiology of human schistosomiasis in Africa: risk models, transmission dynamics and control. Transactions of the Royal Society of Tropical Medicine and Hygiene 101.1, 1-8.

Buttenheim, AM, Paz-Soldan, V, Barbu, C, Skovira, C, Calderón, JQ, Riveros, LMM, Cornejo, JO, Small, DS, Bicchieri, C, Naquira, C, et al. (2014). Is participation contagious? Evidence from a household vector control campaign in urban Peru. Journal of epidemiology and community health 68.2, 103-109.

Csilléry, K, Blum, MG, Gaggiotti, OE, and François, O (2010). Approximate Bayesian computation (ABC) in practice. Trends in ecology \& evolution 25.7, 410-418.

Delgado, S, Ernst, KC, Pumahuanca, MLH, Yool, SR, Comrie, AC, Sterling, CR, Gilman, RH, Náquira, C, and Levy, MZ (2013). A country bug in the city: urban infestation by the Chagas disease vector Triatoma infestans in Arequipa, Peru. International journal of health geographics 12.1, 1.

Dias, JCP, Silveira, AC, and Schofield, CJ (2002). The impact of Chagas disease control in Latin America: a review. Memórias do Instituto Oswaldo Cruz 97.5, 603-612.

Diggle, PJ, Ribeiro Jr, PJ, and Christensen, OF (2003). An introduction to model-based geostatistics. In: Spatial statistics and computational methods. Springer, 43-86.

Eddelbuettel, D, François, R, Allaire, J, Chambers, J, Bates, D, and Ushey, K (2011). Rcpp: Seamless $R$ and $C_{++}$integration. Journal of Statistical Software 40.8, 1-18.

Focks, DA, Brenner, RJ, Chadee, DD, and Trosper, JH (1999). The use of spatial analysis in the control and risk assessment of vector-borne diseases. American Entomologist 45.3, 173-183.

Gelfand, AE and Vounatsou, P (2003). Proper multivariate conditional autoregressive models for spatial data analysis. Biostatistics 4.1, 11-15.

Gelman, A, Carlin, JB, Stern, HS, and Rubin, DB (2014). Bayesian data analysis. Vol. 2. Taylor \& Francis.

Gibson, GJ and Renshaw, E (1998). Estimating parameters in stochastic compartmental models using Markov chain methods. Mathematical Medicine and Biology 15.1, 19-40.

Green, PJ (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82.4, 711-732.

Haley, RW (2012). Controlling urban epidemics of West Nile virus infection. JAMA 308.13, 13251326.

Hong, AE (2013). Gaussian Markov random field models for surveillance error and geographic boundaries. PhD thesis. University of Pennsylvania.

Hong, AE, Barbu, CM, Small, DS, Levy, MZ, et al. (2015). Mapping the spatial distribution of a disease-transmitting insect in the presence of surveillance error and missing data. Journal of the Royal Statistical Society: Series A (Statistics in Society) 178.3, 641-658.

Ionides, EL, Bhadra, A, Atchadé, Y, King, A, et al. (2011). Iterated filtering. The Annals of Statistics 39.3, 1776-1802.

Jaya, IGNM, Abdullah, AS, Hermawan, E, and Ruchjana, BN (2016). Bayesian Spatial Modeling and Mapping of Dengue Fever: A Case Study of Dengue Feverin the City of Bandung, Indonesia. International Journal of Applied Mathematics and Statistics ${ }^{T M}$ 54.3, 94-103.

Jewell, CP, Kypraios, T, Christley, R, and Roberts, GO (2009a). A novel approach to real-time risk prediction for emerging infectious diseases: a case study in Avian Influenza H5N1. Preventive veterinary medicine 91.1, 19-28.

Jewell, CP, Kypraios, T, Neal, P, Roberts, GO, et al. (2009b). Bayesian analysis for emerging infectious diseases. Bayesian Analysis 4.3, 465-496.

Jin, X, Carlin, BP, and Banerjee, S (2005). Generalized hierarchical multivariate CAR models for areal data. Biometrics 61.4, 950-961.

Jones, KE, Patel, NG, Levy, MA, Storeygard, A, Balk, D, Gittleman, JL, and Daszak, P (2008). Global trends in emerging infectious diseases. Nature 451.7181, 990-993.

Kim, H, Sun, D, and Tsutakawa, RK (2001). A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. Journal of the American Statistical association 96.456, 1506-1521.

Kitron, U and Kazmierczak, JJ (1997). Spatial analysis of the distribution of Lyme disease in Wisconsin. American journal of Epidemiology 145.6, 558-566.

Knudsen, AB and Slooff, R (1992). Vector-borne disease problems in rapid urbanization: new approaches to vector control. Bulletin of the World Health Organization 70.1, 1.

Krainski, E and Lindgren, F (2013). The R-INLA tutorial: SPDE models.
LaDeau, SL, Allan, BF, Leisnham, PT, and Levy, MZ (2015). The ecological foundations of transmission potential and vector-borne disease in urban landscapes. Functional ecology 29.7, 889901.

Levy, MZ, Bowman, NM, Kawai, V, Waller, LA, Carpio, JGC del, Benzaquen, EC, Gilman, RH, and Bern, C (2006). Periurban Trypanosoma cruzi-infected Triatoma infestans, Arequipa, Peru. Emerging infectious diseases 12.9, 1345.

Lindgren, F, Rue, H, and Lindström, J (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73.4, 423-498.

Longo, DL and Bern, C (2015). Chagas' disease. New England Journal of Medicine 373.5, 456466.

Mathers, C, Fat, DM, and Boerma, JT (2008). The global burden of disease: 2004 update. World Health Organization.

Oluwole, AS, Ekpo, UF, Karagiannis-Voules, D-A, Abe, EM, Olamiju, FO, Isiyaku, S, Okoronkwo, C, Saka, Y, Nebe, OJ, Braide, EI, et al. (2015). Bayesian geostatistical model-based estimates of soil-transmitted helminth infection in Nigeria, including annual deworming requirements. PLoS Negl Trop Dis 9.4, e0003740.

O'Neill, PD and Roberts, GO (1999). Bayesian inference for partially observed stochastic epidemics. Journal of the Royal Statistical Society: Series A (Statistics in Society) 162.1, 121-129.

ISSN: 1467-985X. DOI: 10.1111/1467-985X.00125. URL: http://dx.doi.org/10.1111/1467985X. 00125.

O'Neill, PD (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. Mathematical biosciences 180.1, 103-114.

Ostfeld, RS, Glass, GE, and Keesing, F (2005). Spatial epidemiology: an emerging (or re-emerging) discipline. Trends in ecology \& evolution 20.6, 328-336.

Palomino, M, Villaseca, P, Cárdenas, F, Ancca, J, and Pinto, M (2008). Eficacia y residualidad de dos insecticidas piretroides contra triatoma infestans en tres tipos de viviendas: Evaluación de campo en Arequipa, Perú. Revista Peruana de Medicina Experimental y Salud Pública 25.1, 9-16.

Rabinovich, JE (1972). Vital statistics of Triatominae (Hemiptera: Reduviidae) under laboratory conditions. I. Triatoma infestans Klug. J Med Entomol 9.4, 351-70.

Rassi, A and Marin-Neto, JA (2010). Chagas disease. The Lancet 375.9723, 1388-1402.
Reisen, WK (2010). Landscape epidemiology of vector-borne diseases. Annual review of entomology 55, 461-483.

Ross, SM et al. (1996). Stochastic processes. Vol. 2. John Wiley \& Sons New York.
Rossi, RE, Mulla, DJ, Journel, AG, and Franz, EH (1992). Geostatistical tools for modeling and interpreting ecological spatial dependence. Ecological monographs 62.2, 277-314.

Rue, H and Held, L (2005). Gaussian Markov random fields: theory and applications. CRC Press.
Rue, H, Martino, S, and Chopin, N (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71.2, 319-392.

Rue, H, Martino, S, Lindgren, F, Simpson, D, Riebler, A, and Krainski, ET (2014). INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximaxion. R package version 0.0-1389624686.

Sain, SR and Cressie, N (2002). Multivariate lattice models for spatial environmental data. Proc. Statist. Environ. Sect. Am. Statist. Ass, 2820-2825.

Sikka, V, Chattu, VK, Popli, RK, Galwankar, SC, Kelkar, D, Sawicki, SG, Stawicki, SP, and Papadimos, TJ (2016). The emergence of zika virus as a global health security threat: A review and a consensus statement of the INDUSEM Joint working Group (JWG). Journal of global infectious diseases 8.1, 3.

Tarleton, RL, Reithinger, R, Urbina, JA, Kitron, U, and Gürtler, RE (2007). The challenges of Chagas disease-Grim outlook or glimmer of hope? PLoS Med 4.12, e332.

Varley, GC, Gradwell, GR, and Hassell, MP (1974). Insect population ecology: an analytical approach. Univ of California Press.

Wall, MM (2004). A close look at the spatial structure implied by the CAR and SAR models. Journal of statistical planning and inference 121.2, 311-324.

Waller, LA and Gotway, CA (2004). Applied spatial statistics for public health data. Vol. 368. John Wiley \& Sons Sons.

Weaver, SC (2013). Urbanization and geographic expansion of zoonotic arboviral diseases: mechanisms and potential strategies for prevention. Trends in microbiology 21.8, 360-363.

Whittle, P (1954). On stationary processes in the plane. Biometrika, 434-449.
Whittle, P (1963). Stochastic-processes in several dimensions. Bulletin of the International Statistical Institute 40.2, 974-994.

Winters, AM, Eisen, RJ, Delorey, MJ, Fischer, M, Nasci, RS, Zielinski-Gutierrez, E, Moore, CG, Pape, WJ, and Eisen, L (2010). Spatial risk assessments based on vector-borne disease epidemiologic data: importance of scale for West Nile virus disease in Colorado. The American journal of tropical medicine and hygiene 82.5, 945-953.

Wood, SN (2010). Statistical inference for noisy nonlinear ecological dynamic systems. Nature 466.7310, 1102-1104.

