1-1-2013

Gaussian Markov Random Field Models for Surveillance Error and Geographic Boundaries

Andrew Ernest Hong
University of Pennsylvania, ahon@wharton.upenn.edu

Gaussian Markov Random Field Models for Surveillance Error and Geographic Boundaries

Abstract
This dissertation addresses two basic problems in epidemiological surveys of insect distributions: the uncertainty in the surveillance process conducted by human inspectors and the modeling of geographic barriers in spatial analysis.

In the first work, we propose a Bayesian hierarchical model which models the accuracy of human inspectors. We apply this model to analyze an entomological survey conducted by the Peruvian Ministry of Health in Mariano Melgar, Peru to locate areas of underreporting of insect infestation. We consider how the household assignment of inspectors influences this identifiability problem. We introduce a simulation paradigm where the strength of confounding may be controlled. Through these simulations, we demonstrate how practically implementable assignment recommendations can mitigate the error in infestation estimates created by this confounding.

In the second work, we study a method for modeling geographic boundaries. We parameterize the shape of these barriers to vary according to intensity of these effects. We demonstrate the model's properties on simulated data and show the efficiency of Bayesian procedures. We then apply the model to the above data set by modeling streets in Mariano Melgar. We quantify this barrier effect and after performing sensitivity analysis, conclude that streets are a major barrier. Lastly, we discuss some extensions and open possibilities with our approach.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Dylan S. Small

Subject Categories
Ecology and Evolutionary Biology | Statistics and Probability

GAUSSIAN MARKOV RANDOM FIELD MODELS FOR SURVEILLANCE ERROR AND GEOGRAPHIC BOUNDARIES

Andrew E. Hong

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

Dylan S. Small, Professor of Statistics

Graduate Group Chairperson

Eric T. Bradlow, Professor of Marketing, Statistics, and Education

Dissertation Committee

Dylan S. Small, Professor of Statistics

Michael Z. Levy, Professor of Biostatistics and Epidemiology

Lawrence D. Brown, Professor of Statistics

GAUSSIAN MARKOV RANDOM FIELD MODELS FOR SURVEILLANCE ERROR AND GEOGRAPHIC BOUNDARIES

ACKNOWLEDGEMENT

The completion of this dissertation is due largely to the patience and guidance of my advisors Dylan and Mike. I can not express how fortunate I have been to be able to work with you. I would also like to thank my other committee member, Professor Brown, for his insight and apt suggestions, which have saliently guided the direction of this work. I owe a great deal to Chris Paciorek and Tony Smith for their generosity in terms of ideas and time. I want to thank all of the faculty in the Statististics department, in particular our chair - Ed George, for their support and their dedication to seeing the best in their students. Lastly, I would like to thank Corentin Barbu for his friendly guidance and advice.

ABSTRACT

GAUSSIAN MARKOV RANDOM FIELD MODELS FOR SURVEILLANCE ERROR AND GEOGRAPHIC BOUNDARIES

Andrew E. Hong

Dylan S. Small

This dissertation addresses two basic problems in epidemiological surveys of insect distributions: the uncertainty in the surveillance process conducted by human inspectors and the modeling of geographic barriers in spatial analysis.

In the first work, we propose a Bayesian hierarchical model which models the accuracy of human inspectors. We apply this model to analyze an entomological survey conducted by the Peruvian Ministry of Health in Mariano Melgar, Peru to locate areas of underreporting of insect infestation. We consider how the household assignment of inspectors influences this identifiability problem. We introduce a simulation paradigm where the strength of confounding may be controlled. Through these simulations, we demonstrate how practically implementable assignment recommendations can mitigate the error in infestation estimates created by this confounding.

In the second work, we study a method for modeling geographic boundaries. We parameterize the shape of these barriers to vary according to intensity of these effects. We demonstrate the model's properties on simulated data and show the efficiency of Bayesian procedures. We then apply the model to the above data set by modeling streets in Mariano Melgar. We quantify this barrier effect and after performing sensitivity analysis, conclude that streets are a major barrier. Lastly, we discuss some extensions and open possibilities with our approach.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

The application of spatial statistics to epidemiology centers on two complementary problems: surveillance and control. The problem of surveillance is to visualize the geographic pattern of an infestation over space and infer its evolution over time. The problem of control is the strategic allocation of resources to mitigate or dampen the damage caused by an epidemic. The focus of this dissertation is on the modeling of natural phenomena that arise in the course of surveillance problems and demonstration how statistical methods based on these models overcome these problems. A primary motivation of these developments is the proliferation of data rich environments where efficient computation to handle the sheer volume and scale of the data is necessary.

The aim of this thesis is to address omnipresent problems that arise in epidemiology that are not addressed in the literature due to deficiency in existing statistical models. The two central contributions are investigations into the problem of surveillance error in epidemiological surveys and the development of the modeling of geographic boundaries in spatial analysis. In the former, we are interested in inferring the hidden or occluded sources of infestation in spatial data which is underreported due to insensitivity of inspection processes. In the latter, we introduce a method for modeling geographic boundaries in spatial data.

One primary application of this work is to analyze epidemiological data collected throughout the district of Mariano Melgar in the city of Arequipa, Peru by the Peruvian Ministry of Health. The purpose of these surveys is to identify the locations of infested households of *Triatoma infestans*, an insect which spreads a disease-causing protozoan *Trypanosoma cruzi*, for the application of insecticide. The resulting Chagas disease from this protozoan is a major epidemic in South America and has spurred interest in statistical methods to aid public health campaigns, which are limited by the availability of resources, Levy et al. (2010).

1.1. Background

Most models for spatial data, whether they be for continuous or discrete data, are based on the concept of a random field or a continuous stochastic process usually defined on $\mathbb{R}^2$. Because most spatiotemporal data is observed at a finite number of locations, $\{\omega_i\}_{i=1}^n$, this continuous model induces a marginalized, multivariable random variable $x = \sum_{i=1}^n x(\omega_i)\mathbf{e}_i$ with a joint distribution $f(x)$.

The vast majority of spatial modeling is done for cases in which the random field is assumed to be Gaussian, such as the Brownian sheet, so that the induced finite collections are jointly Gaussian. The advantage of Gaussian models is that their distributions are summarized by their first two moments. Then the primary point of focus in spatial statistics are covariance models where the covariance between two observations on the field is a function of the locations or, for centered fields, $c(x(\omega_i), x(\omega_j)) = \mathbb{E}\{x(\omega_i)x(\omega_j)\} = c(\omega_i, \omega_j)$. Of particular interest are *weakly stationary* random fields with shift invariant covariance functions $c(\omega_i, \omega_j) = c(0, \omega_i - \omega_j) = c(\omega_i - \omega_j)$. An even stronger assumption that is sometimes made is to assume that the covariance is isotropic or rotationally invariant $c(\omega_i - \omega_j) = c(\|\omega_i - \omega_j\|)$.

The usual practice for fitting spatial models to data is to assume that the covariance is isotropic and then to fit to a parametric covariance function. Standard selection of a covariance model involves the variogram or $\gamma(\|\omega_i - \omega_j\|) = \frac{1}{2}\mathbb{E}\{x(\omega_i) - x(\omega_j)\}^2$ where $h = \|\omega_i - \omega_j\|$ is the lag. The variogram γ is usually taken to be continuous except for the jump discontinuity at $h = 0$, which is known as the *nugget*. The *sill* is the limit $\lim_{h\to\infty}\gamma(h)$, which for ergodic random fields is just simply the variance $c(0)$. The *range* is the distance r needed such that $\gamma(r) \approx \lim_{h\to\infty}\gamma(h)$, which in practice is taken to be the distance where the correlation is equal to 0.005. One example of these parametric models is the spherical variogram model,

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 & \text{if } t \geq 1/\phi \\ \tau^2 + \sigma^2(3\phi h - (\phi h)^3)/2 & \text{if } 0 < t \leq 1/\phi \end{cases} \tag{1.1.1}$$

where τ, σ, ϕ are all positive, Banerjee et al. (2004). The interpretation of the parameters is as follows: τ^2 is the nugget, $\tau^2 + \sigma^2$ is the sill, and $h = 1/\phi$ is the range. Given data, the variogram may be fit to a class in a number of ways, either by calculating the method of moments estimator $\hat{\gamma}$, using various choices of lags, and fit to γ using weighted least squares or by (restricted if containing covariates) maximum likelihood methods using the multivariate Gaussian likelihood.

Another approach to spatial modeling is the use of graphical models defined on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The multivariate random variable $\mathbf{x}$ is defined on each node or vertice in $\mathcal{V}$ and generally the dependence structure in $\mathbf{x}$ is reflected in the connectivity among the vertices through the edges $\mathcal{E}$. The canonical example of these models is the multivariate Gaussian where the graphical structure is contained in the precision matrix $\mathbf{Q}$. A *clique* is a set of vertices where every possible pairwise connection is contained in $\mathcal{E}$, and a *maximal clique* is a clique not contained in another clique in the graph. A graph may be factored into a collection of maximal cliques. For a precision $\mathbf{Q}$ defined on a graph factorized into the collection of maximal cliques $\{C_i\}_{i=1}^m$ with n nodes, the density of $\mathbf{x}$ is,

$$f(\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^\intercal \mathbf{Q}\mathbf{x}\right\} = \exp\left\{-\frac{1}{2}\sum_{v\in\mathcal{V}} \mathbf{Q}_{v,v}{x_v}^2 - \sum_{u<v} \mathbf{Q}_{u,v}x_u x_v\right\} \tag{1.1.2}$$

$$= \exp\left\{-\sum_{i=1}^{m} \Phi(\mathbf{x}_{C_i})\right\} \tag{1.1.3}$$

The *potential* function Φ is then a function over the maximal cliques of $\mathcal{G}$. If $\mathbf{Q}_{i,j} = 0$ then there exists no edge between vertices i and j and one has the conditional independence property: $x_i \perp x_j | \mathbf{x}_{-\{i,j\}}$. This feature is a kind of Markov property, however Gaussian Markov random fields (GMRF) possess a stronger Markov property. Namely for three subsets of vertices of $\mathcal{V}$, we call C a *seperating set* from A to B if every path from a vertice in A to a vertice in B contains a vertice in C. GMRFs possess the global Markov property:

if C is a seperating set from A to B then $\mathbf{x}_A \perp \mathbf{x}_B|\mathbf{x}_C$, Rue and Held (2005). Interestingly for continuous random fields, a the notion of seperation can be extended: if removing C from the ambient space induces the closures of A and B to be disjoint. A continuous random field is Markovian if and only if its power spectra is an inverse polynomial, Rozanov (1977).

The problem with GMRF is that for general graphs is that there is no interpretation for their covariance structure. However, these models may be interpreted as discretized solutions to stochastic differential equations (SDE). The most popular GMRF model is the conditionally autoregressive (CAR) model introduced in Besag (1974), which we specify in general in equation 2.3.3. A simple example of this connection is the AR(1) model or the CAR model on a regular line has precision,

$$\mathbf{Q} = \begin{bmatrix} 1 & -1 & & & & & & \\ -1 & 2 & -1 & & & & & \\ & -1 & 2 & -1 & & & & \\ & & \ddots & \ddots & \ddots & & & \\ & & & -1 & 2 & -1 & & \\ & & & & -1 & 2 & -1 & \\ & & & & & -1 & 2 & -1 \\ & & & & & & -1 & 1 \end{bmatrix} \qquad \mathbf{D} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix} \tag{1.1.4}$$

Consider the SDE, $\frac{\partial x}{\partial t} = W$, where W is white noise in time by discretizing the derivative as the foward difference for the regularly spaced locations ($\Delta t = \Delta t_i = t_{i+1} - t_i$), $\frac{\partial x(t_i)}{\partial t} \approx \frac{x(t_{i+1}) - x(t_i)}{\Delta t}$. The discretization implies that $\Delta t^{-1}\mathbf{D}\mathbf{x} \sim \mathrm{N}(\mathbf{0}, \Delta t^{-1}\mathbf{I})$, which has density proportional to $\exp\{-1/(2\Delta t)(\mathbf{D}\mathbf{x})^\intercal(\mathbf{D}\mathbf{x})\} = \exp\{-1/(2\Delta t)\mathbf{x}^\intercal\mathbf{Q}\mathbf{x}\}$, or that $\mathbf{x} \sim \mathrm{N}(\mathbf{0}, \Delta t^{-1}\mathbf{Q})$. For simple models, the connection is a curiosity. Later, we will exploit this connection fully for more complex problems in chapter 3 to generate GMRF approximations to continuous models.

With the fit covariance (or precision) fixed, the problem of prediction or interpolating values of the random field at unobserved locations is calculated using the conditional mean, conditioning on the observed values. This procedure, also known as *kriging*, is the best linear unbiased estimator. Because this conditional distribution involves inversion of the covariance matrix, the computational complexity of kriging scales as $\mathcal{O}(n^3)$, where n references the number of locations observed. Although while in this work we take the Bayesian approach to estimation, one returns to similar problems of solving systems of linear equations involving large covariance matrices. Hence as efficient computation is one of our central interests, we focus on GMRF where these matrices are usually very sparse or where a majority of the entries are zero. For large sparse matrices, systems of equations involving covariances can be solved using sparse Cholesky solvers, which scale based on the width of the banding post-permutation.

1.2. Research Objectives

In chapter 2, we propose a Bayesian hierarchical model that accounts for the uncertainty in the observation process for presence and absence of infestation data collected by human inspectors. Modeling each inspector's sensitivity as well as the spatial process separately, we estimate the true infestation rates to determine regions of underreporting. As there is little prior data regarding individual inspector accuracy, we study the problem of parameter identifiability between inspector sensitivity and spatial intensity of the infestation. This problem of identifiability occurs when there is correlation in space between the intensity of the infestation and the accuracy of the inspector observing the data. We create a simulation paradigm where inspectors are assigned to households randomly based on their accuracy and the regional infestation intensity. By tuning the assignment distribution, we control the likelihood of confounding and demonstrate the associated increase in estimation error. Frequently in epidemiological surveys, there is no way to diagnose the degree of confounding. As a result, we propose assignment recommendations, which perform well even in the worst case scenario. We conclude that models which do not account for the heterogeneity in

inspection error perform poorly in the presence of confounding, regardless of assignment. However, we demonstrate that modeling the heterogeneity in addition to our assignment recommendations vastly improves the infestation estimation. This improvement follows even in the absence of informative priors for inspector accuracy. For the Mariano Melgar survey in Arequipa, Peru, we simulate data where inspectors are assigned to households uniformly at random and contrast the estimation error to simulations using the fixed assignment to conclude that there is some evidence that there is no danger of strong confounding. We identified four at risk localities that fell under the Peruvian Ministry of Health's threshold for infested households, which as a result were later sprayed with insecticide treatment.

In chapter 3, we introduce a mechanistic approach for modeling the effect on spatial dispersion of geographic boundaries. While the importance of the boundaries is intuitive to investigators, the ability to incorporate the geometry of these boundaries has been outside the scope of existing spatial methods. Our approach is to model geographic boundaries as deformations of the surface, on which a stochastic partial differential equation is defined. We take as the spatial effect, in our model, the approximate solution to this SPDE by using the finite element method. Because of the SPDE formulation, the approach avoids the problem of the positivity requirement for covariance models that usually arises when attempting to adapt models for curved surfaces. We first review some standard background regarding finite element analysis required for our implementation and then detail our contribution. In addition to modeling fixed deformation, we demonstrate the ability to parameterize the deformations. We then explore the efficiency of statistical methods on simulated data for inferring these deformation parameters. As an application of our method, we model the major roadways of Mariano Melgar and use Bayesian techniques to infer the effect of streets in the spatial dispersion of the insects. Our analysis shows that streets are a highly non-trivial barrier to the dispersion of insects. The ability, or lack thereof, of infestations of insects to traverse streets has been an open question to entomologists studying *Triatoma infestans*. Lastly, we conduct some sensitivity tests of our findings on the importance of streets to the model's other parameters.

The approaches and models discussed in this dissertation work were motivated by the problem of Chagas in South America but are applicable to other statistical problems in epidemiology. Most of the data we worked with was discrete, but we discuss simulations and implementations for the continuous case. Lastly, we worked with communities which sometimes contained over fifteen thousand households. We found our methods to scale very efficiently for large data sets. On a final note, in chapter 3, we summarize briefly our findings and discuss some immediate extensions such as spatiotemporal modeling and more complex models for capturing the shapes of spatial deformations.

CHAPTER 2 : Surveillance Error in Epidemiological Surveys

2.1. Introduction

The rise of urbanization around the developing world has been met with the increasing risk of epidemics of vector-borne disease. The use of spatial analysis in aiding public health officials in controlling these disease outbreaks is documented in: Dengue, Teixeira et al. (2002); Malaria, Trape et al. (1992); and Chagas disease, Corrasco et al. (2005).

Chagas disease is one of the most prevalent of the debilitating diseases in South America. The disease is due to the *Trypanosoma cruzi* parasite, transmitted by triatomine insect vectors. Most policy for Chagas disease control has centered around eliminating the insect vector. While modern practices of vector control have been applied, the triatomine insect has proven to be a continually re-emergent problem Levy et al. (2006). Given the disease carrier's persistence and abundant distribution, matched by a serious strain on public resources (manpower, insecticide, and so on), there has been an increasing need and interest for statistical methods to understand the spatial distribution of triatomine infestations in order to apply resources and direct public health campaigns Levy et al. (2010).

Currently, the Peruvian Ministry of Health (MoH) is working in Arequipa, Peru to control an epidemic of *Trypanosoma cruzi* spread by *Triatoma infestans*, Levy et al. (2006). Control efforts are guided by raw household level survey data such as that shown in figure 1. Prior to insecticide application, all localities in Arequipa are surveyed to assess the severity of infestation. Although the inspectors conducting these surveys are all trained by the MoH for this purpose, the process to determine if there is an infestation is not straightforward and inspectors naturally have varying degrees of skill and experience. Further, it is unfeasible in terms of cost to conduct repeated household level surveys. Hence, accurate prevalence estimates from a single survey are needed for prioritizing insecticide treatment. Although policy is fluid, at the time of this study localities with infestation prevalences exceeding 10 % were prioritized for insecticide application. Given this policy, the primary concern of our

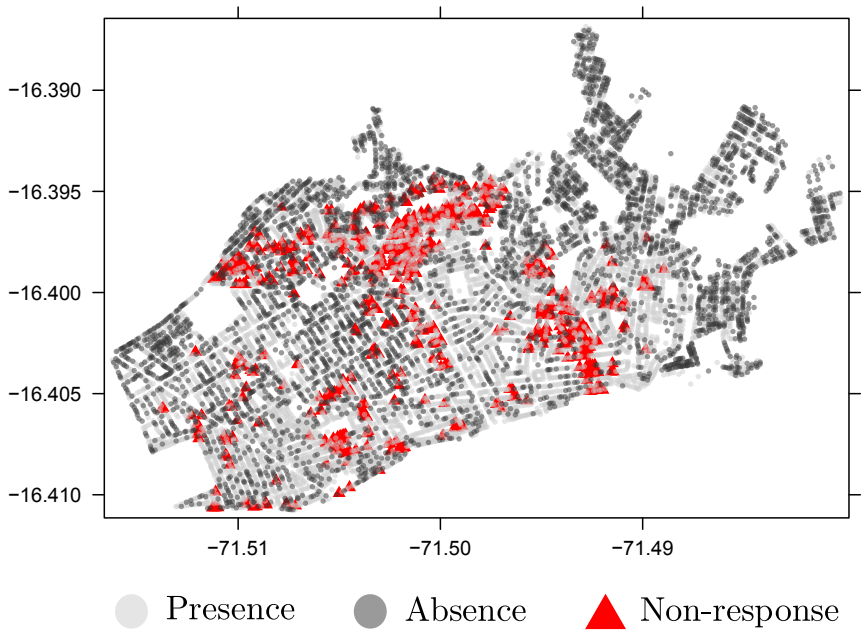

Figure 1: Raw results of a preliminary survey for *Triatoma infestans* in the district of Mariano Meglar, Arequipa, Peru, 2009

study is to address the under-reporting of infestation rates due to surveillance error and to apply appropriate statistical methods to correct for this problem. Although the magnitude of the surveillance error in this setting is only roughly known in advance, the identities of which inspectors were assigned to which households was carefully documented. In order to take advantage of this information, we construct a Bayesian hierarchical model to infer the heterogeneity in surveillance error across different inspectors. We will argue through simulation, that modeling the heterogeneity in inspector surveillance error improves overall identification of the infestation and, further, that ignoring the heterogeneity may occlude regions of serious infestation, hidden by insensitivity on the part of inspectors of that region.

2.2. Data

The dataset in consideration consists of 12,070 household-level entries from Mariano Melgar, a district of the city of Arequipa, Peru. Inspection was on a voluntary basis, which accounts

for the large amount of missing data (approximately 34 % of the total number of houses); the most common reason for missing data is simply absence - as the owner or any household members were unavailable at the time of inspection, permission to inspect the household could not be granted. We mapped the position of all households and the delimitation of city blocks in the district, comparing satellite imagery in Google Earth™ to field maps drawn by the personnel of the MoH Google (2009). Households were then aligned with their city block according to their respective coordinates. For each entry, the locations of the household are supplied along with its inspected status: infested, un-infested, and non-response. The entries are not evenly divided amongst the inspectors, nor are the inspectors themselves evenly distributed across space. Part of the interest of this study is to make future recommendations for inspector assignment.

2.3. Model

To model spatial correlation between the infestation status of the vector in each household, we used a Gaussian linear model, along with a probit link to the binary infestation status of each data point. A complete diagram of the hierarchical specification is shown in figure 2. Similar approaches can be found in a variety of public health applications in Banerjee et al. (2003), Banerjee et al. (2004). The parameter of interest $\mathbf{y}$ is binary, denoting the true infestation status of each household. In contrast, the actual data recorded by the inspector is $\mathbf{z}$, which is the observed infestation status. The infestation status, $\mathbf{y}$, is for the practical purposes of this study unobservable and must be estimated by the model.

2.3.1. Household Risk

The spatial correlation between the true infestation status of the household, $\mathbf{y}$, is modeled through the latent probability of infestation at each household denoted by $\mathbf{x}$. In order to perform spatial regression on the binary outcome, we use the probit link:

$$\mathbb{P}(y_i = 1|x_i) = \Phi(x_i) \tag{2.3.1}$$

Conditional on the spatial risk field $\mathbf{x}$, each variable y_i is conditionally independent from one another.

The latent risk, $\mathbf{x}$, is taken to be a linear model composed of a fixed effect t and a spatially-correlated household level effect u_i:

$$\mathbf{x} = t + \mathbf{u} \quad \text{where} \quad \mathbf{1}^\intercal \mathbf{u} = 0 \tag{2.3.2}$$

The spatial effect $\mathbf{u}$ is modeled using the sparse conditionally auto-regressive (CAR) model introduced in Besag et al. (1991) also known as a Gaussian Markov random field, Rue and Held (2005). Although $\mathbf{u}$ does not follow a proper Gaussian distribution, through abuse of notation its distribution is often denoted as $\mathrm{N}(\mathbf{0}, k_u \mathbf{\Lambda})$, where $k_u \mathbf{\Lambda}$ is the precision matrix. In this application, $\mathbf{\Lambda}$ is fixed.

$\mathbf{\Lambda}$ will be defined as $\Lambda_{ij} = -1/d(i,j)$, where $d(i,j)$ is the Euclidean distance between the locations indexed by i and j. In order to ease the computation, truncation or tapering will be applied in the following manner: if $d(i,j) > \kappa$, then $\Lambda_{ij} = 0$. From previous work in Barbu et al. (2013), spatial autocorrelation in data collected from an adjacent region was shown to drop sharply at around 50 meters, and in this work, we take 50 meters to be our threshold. The diagonal entries of $\mathbf{\Lambda}$ are defined as $\Lambda_{ii} = \sum_{j \neq i} -\Lambda_{ij}$. With this specification, the conditional distribution of each household is,

$$u_i | \mathbf{u}_{-i} \sim \mathrm{N}\left(\frac{\sum_{\{j \in N_i : j \neq i\}} d(i,j)^{-1} u_j}{\sum_{\{j \in N_i : j \neq i\}} d(i,j)^{-1}}, k_{\mathbf{u}} \sum_{\{j \in N_i : j \neq i\}} d(i,j)^{-1} \right) \tag{2.3.3}$$

where $\mathbf{u}_{-i}$ denotes all the households except for i. Because of the specification $\Lambda_{ii} = \sum_{j \neq i} -\Lambda_{ij}$; $\mathbf{\Lambda}$ is not invertible, which necessitates the sum to zero constraint in Equation 2.3.2. In deriving Monte Carlo methods, it is often helpful to consider the "joint-density,"

$$f(\mathbf{u}) \propto \exp\left(-\frac{k_u}{2}\mathbf{u}^t\mathbf{\Lambda}\mathbf{u}\right) = \exp\left(-\frac{k_u}{2}\sum_{i<j} d(i,j)^{-1}(u_i - u_j)^2 \mathbb{1}_{d(i,j)\leq\kappa}\right) \tag{2.3.4}$$

The conditional distribution of the spatial risks is centered on the weighted sums of the neighboring points, where the notion of a neighborhood for a specific point is every other household within a certain distance. From the conditional distribution, the more neighbors a household possesses, the smaller the conditional variation is for that household. The closer household j is to household i, the more influential the value of the field at j is on the conditional distribution of u_i.

Together k_u and $\mathbf{\Lambda}$ determine the strength of the spatial relationship between the household level risks. Larger values of k_u penalize the likelihood of rougher fields favoring flatter ones. From the conditional distribution, larger values of k_u allow the conditional distribution of a point to vary less from the weighted sum of its neighbors, increasing the relevance of neighboring values. In practice having fixed $\mathbf{\Lambda}$, a less informative prior will be placed on $k_{\mathbf{u}}$, as the degree of the spatial correlation in the vector distribution is not known in advance.

2.3.2. Inspection Process

The objective in modeling the inspection process is to be able to account for the imperfect surveillance and heterogeneity in inspector sensitivities. z_{ij} is the observed status of household i recorded by inspector j as: positive (insects found), negative (no insects found), or non-response. As a first approximation, we model this surveillance process as a Bernoulli random variable,

$$\mathbb{P}(z_{ij} = 1 | y_i, \beta_j, I_{\text{NA}i}) = \begin{cases} \beta_j y_i & \text{if } I_{\text{NA}i} = 1 \\ \text{NA} & \text{if } I_{\text{NA}i} = 0 \end{cases} \tag{2.3.5}$$

where $\mathbf{I}_{\text{NA}}$ is a binary vector containing each of the households, denoting a one if the household was inspected and a zero otherwise. β_j represents the inspector's sensitivity, or

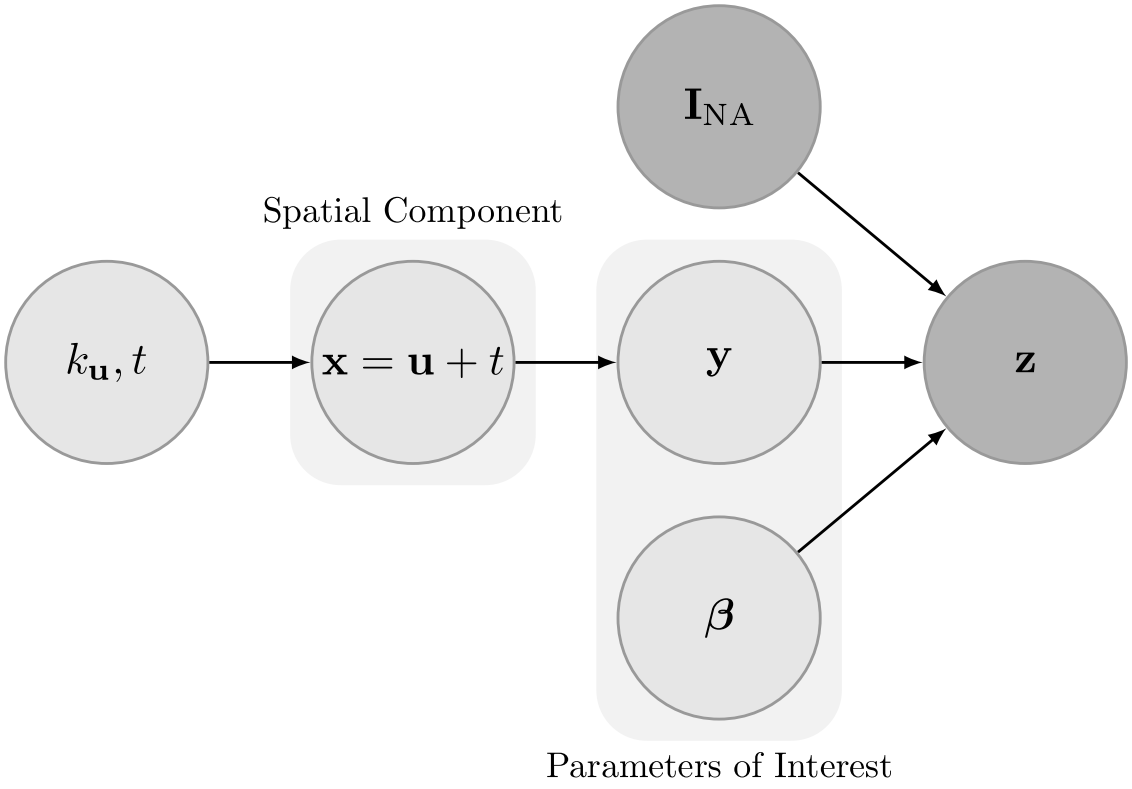

Figure 2: Hierarchical diagram of the model, where $\mathbf{I}_{\text{NA}}$ and $\mathbf{z}$ are the survey data denoting whether or there is non-response and the presence or absence of an infestation. The parameters of interest are $\mathbf{y}$ (infestation status) and $\boldsymbol{\beta}$ (inspector sensitivity). $\mathbf{u}$ is the latent process we used for spatial smoothing.

the probability the inspector reports the insect vector in a truly infested household. Larger values of β_j correspond to more accurate inspectors and are distributed according to a beta distribution. In general, the accuracy of the inspectors is unknown. Prior sensitivity is discussed in Section 2.5.4. Given the ease of identification of the triatomine insect when it is encountered, the possibility of false positives on the part of inspectors is negligible.

We allow for spatial correlation in the location of the missing data, but we assume that this missingness is independent of the risk of infestation, $\mathbb{P}(\mathbf{I}_{\text{NA}}|\mathbf{x}) = \mathbb{P}(\mathbf{I}_{\text{NA}})$.

We will contrast our heterogeneous surveillance error rate model with a simpler, homogeneous error rate model where β is flat for all inspectors and the inspector labels for the observations are ignored.

2.3.3. Prior specification

For the precision parameter $k_{\mathbf{u}}$, we chose to use the standard diffuse gamma $\Gamma(1, 100)$ prior, as recommended in Paciorek (2007). Diffuse folded-t priors were also implemented as recommended by Gelman (2006), but led to no discernible difference in terms of Monte Carlo performance or estimated results. For the fixed risk level t, we used a centered diffuse normal $\mathrm{N}(0, 1 \cdot 10^4)$ prior.

Inspector sensitivity plays a central role in our model. To set an informed prior on inspected sensitivity, we made use of data that the MoH had collected from spray campaigns completed in previous districts, where infestation data on the vector was collected and compared to the initial assessments from surveillance campaigns. Based on this prior data, we set as our informative prior that each β_j is distributed independently from the same $\mathrm{B}(6.5, 2)$ distribution.

2.3.4. Computation

The model was implemented through a Gibbs sampler detailed in Appendix A.3. The implementation of the model on a district of standard size is straightforward in R due to the release of a number of sparse matrix packages (see for example: **spam**, Furrer and Sain (2010)). We found that the slowest mixing and most autocorrelated parameter in the model was $k_{\mathbf{u}}$. A burn-in cut-off of 10,000 steps was determined through running chains from multiple starting points and using the Gelman and Rubin diagnostic on $k_{\mathbf{u}}$, Gelman and Rubin (1992). Similarly, the effective sample size of the estimates and later simulations were based on the autocorrelation of $k_{\mathbf{u}}$.

2.4. Results

Figure 3 is a map of the posterior probability of infestation, $\mathbf{y}$ in each household produced by the heterogeneous inspector model. Table 1 displays the locality-wide infestation estimates, comparing the results produced by the heterogeneous inspector model to those produced

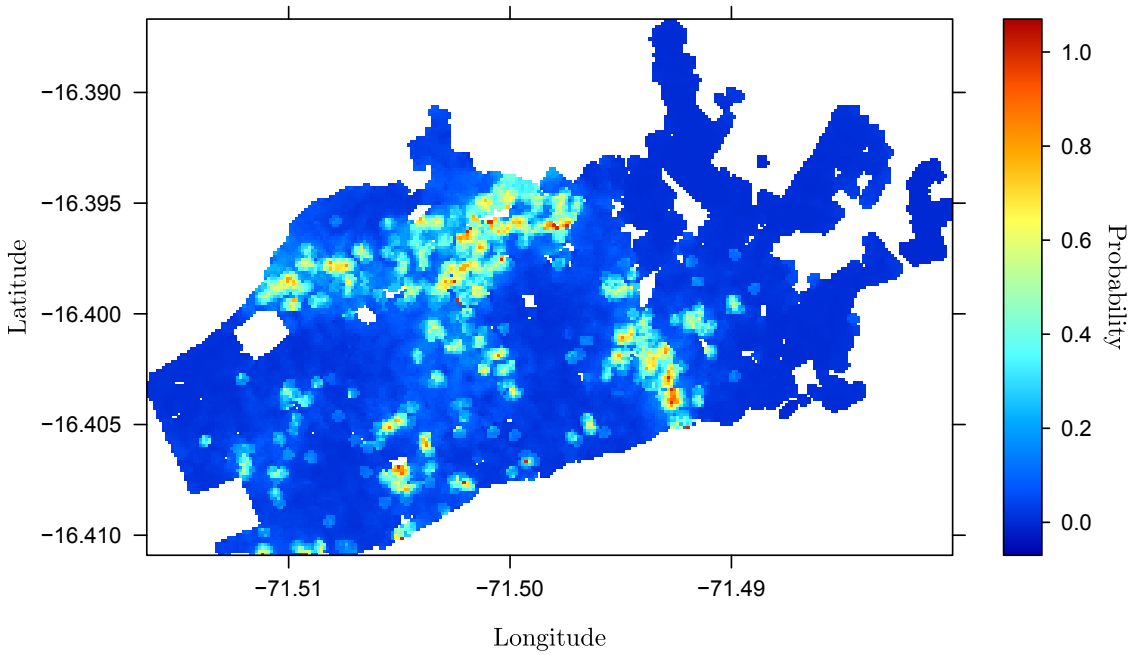

Figure 3: Posterior mean of infestation probabilities estimated from the Mariano Melgar data set, using the heterogeneous inspection model.

by the homogeneous inspector model and the no surveillance error model. At the time of the study, the MoH's use of these results was to apply insecticide to all of the households in localities for which the aggregated locality-wide estimated infestation prevalence exceeded ten percent. Based on our findings, we recommended that a number of localities (localities 2, 7, 11, 15 and 37) be treated that the MoH had not planned to treated, and the MoH followed our recommendations. For localities with presence levels below the 10 % threshold, insecticide application decisions were made on a case-by-case basis, and insecticide was often targeted to certain city blocks only. Thus, block level infestation estimates are also important for decision making in addition to the locality infestation estimates. A comparison of the block level infestation estimates between our model and smoothing which ignores surveillance error are shown in figure 5 showing the emergence of new blocks at risk for infestation.

Posterior Probability of Infestation Estimates for Models

Locality	Heterogeneous	Homogeneous	No Surveillance Error	% NA
1	0.0114	0.0113	0.0021	0.5884
2	0.1128	0.1214	0.0876	0.3831
3	0.0954	0.0906	0.0481	0.4613
4	0.0188	0.0228	0.0051	0.3529
5	0.0323	0.0261	0.0060	0.5976
6	0.0432	0.0421	0.0142	0.6164
7	0.1437	0.1460	0.0845	0.6220
8	0.0259	0.0296	0.0059	0.3243
9	0.1907	0.2041	0.1486	0.0093
10	0.2912	0.2913	0.2245	0.5238
11	0.1363	0.1416	0.0857	0.2879
12	0.3696	0.3706	0.2999	0.4956
13	0.3273	0.3391	0.2575	0.0828
14	0.0323	0.0308	0.0115	0.4082
15	0.1056	0.1095	0.0771	0.2857
16	0.0323	0.0297	0.0243	0.4253
17	0.0394	0.0410	0.0169	0.4690
18	0.0044	0.0015	0.0008	0.5611
19	0.0029	0.0012	0.0007	0.5917
21	0.0029	0.0011	0.0003	0.5428
22	0.0057	0.0035	0.0006	0.5926
23	0.0248	0.0235	0.0127	0.3352
24	0.0123	0.0108	0.0070	0.4489
25	0.0031	0.0014	0.0007	0.5000
26	0.0066	0.0019	0.0028	0.6504
28	0	0	0	0.5734
30	0.0044	0.0025	0.0016	0.5728
31	0.0057	0.0039	0.0041	0.5849
32	0.0029	0.0020	0.0013	0.3636
33	0.0073	0.0044	0.0015	0.6951
34	0.0013	0.0011	0.0004	0.4400
35	0.0079	0.0073	0.0053	0.5438
36	0.0564	0.0574	0.0445	0.1798
37	0.1075	0.1113	0.0804	0.1845
38	0.0780	0.0812	0.0575	0.2026

Table 1: Infestation estimates averaged by locality in Mariano Melgar using the heterogeneous and homogeneous inspection error models. Estimates which ignore inspection error, but are spatially interpolated for non-response are also provided.

From the Mariano Melgar data, the group average of the estimated inspector accuracies was 76.08 %, very nearly the prior mean of 76.47 %. Figure 4 demonstrate some of the heterogeneity of posterior distributions of the inspector sensitivities and how they deviate from the prior.

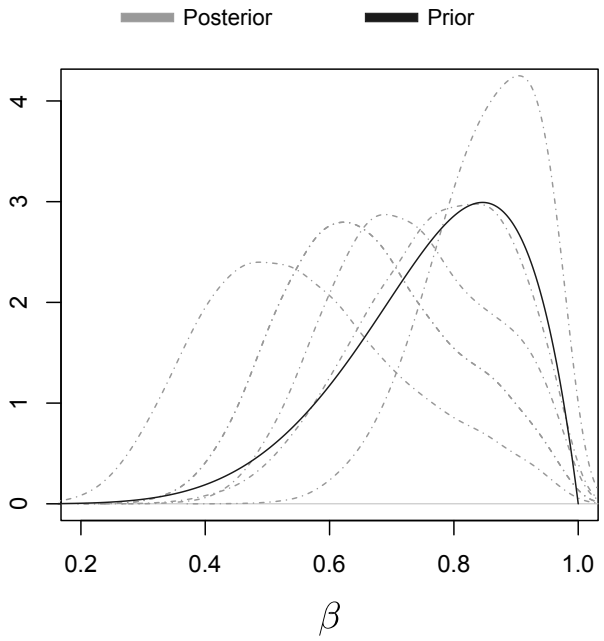

Figure 4: Posterior distributions of the five example inspectors' accuracies representing the variability in detection sensitivity across the inspectors in this study.

Although we have confidence in the informative prior used, an open question is how much information is present in the data to identify inspector sensitivity. We believe that there is a reasonable amount as although overall accuracy rates depend on the prior specification, inspector sensitivity rankings relative to one another are consistent when using weaker less informative priors such as the B $\left(\frac{1}{2}, \frac{1}{2}\right)$ (group posterior mean of 0.5297) and B$(1, 1)$ (group posterior mean of 0.5557). While the relative performance of the inspectors in Mariano Melgar were consistent with our prior knowledge, we must caution that these estimates are highly sensitive, especially in terms of overall performance, to the prior specification. Further, we will explore phenomena in section 2.5.2 that can severely bias these estimates, and the sensitivity estimates should be judged with caution. However, given the overall consistency in ranking, we believe there is sufficient information in the data to take advantage of the additional structure in the heterogeneous model.

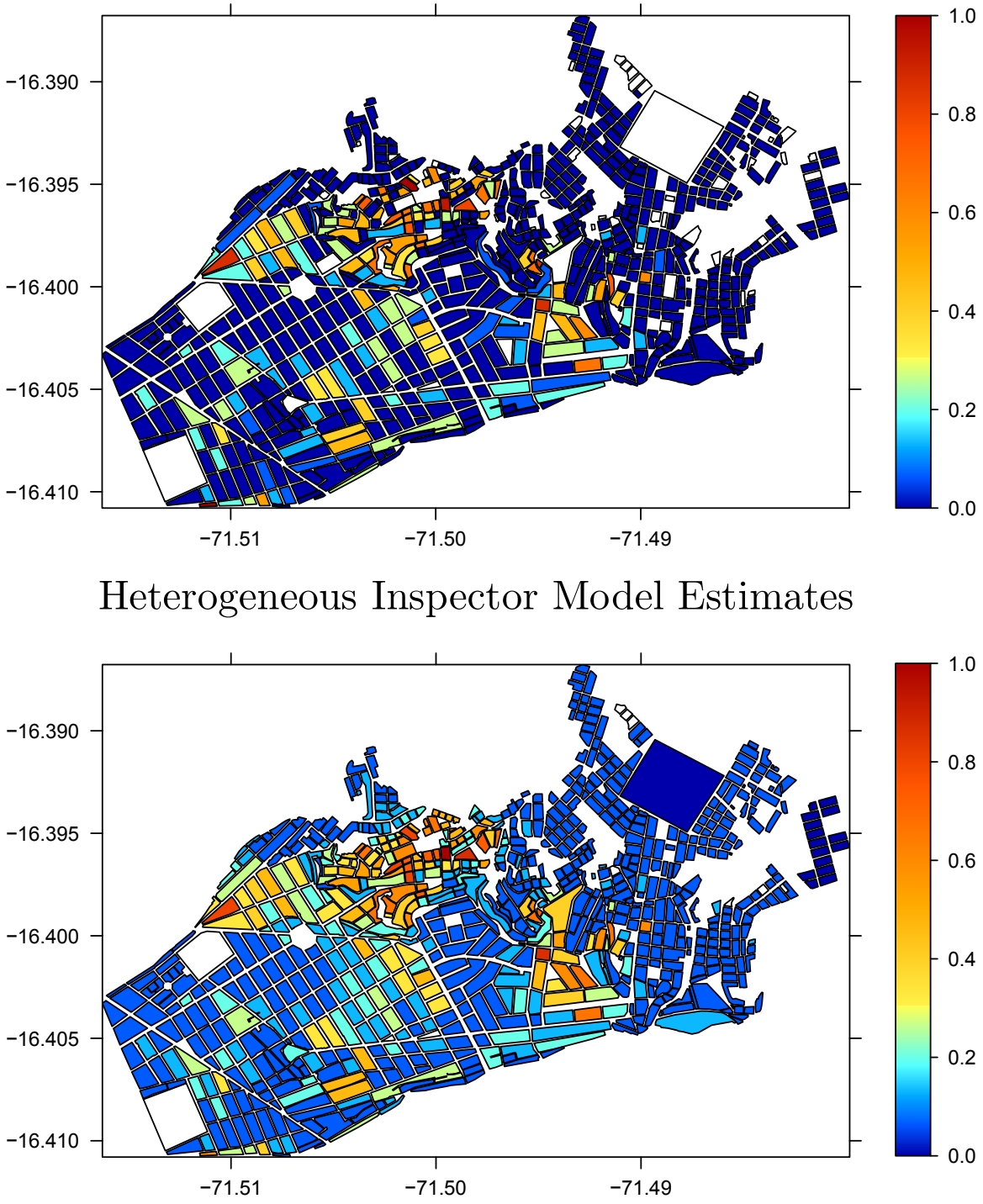

Figure 5: Top plot is the raw infestation prevalence, collected in the survey, at the block level. The bottom plot is the estimated infestation prevalence, produced by our model, aggregated at the block level.

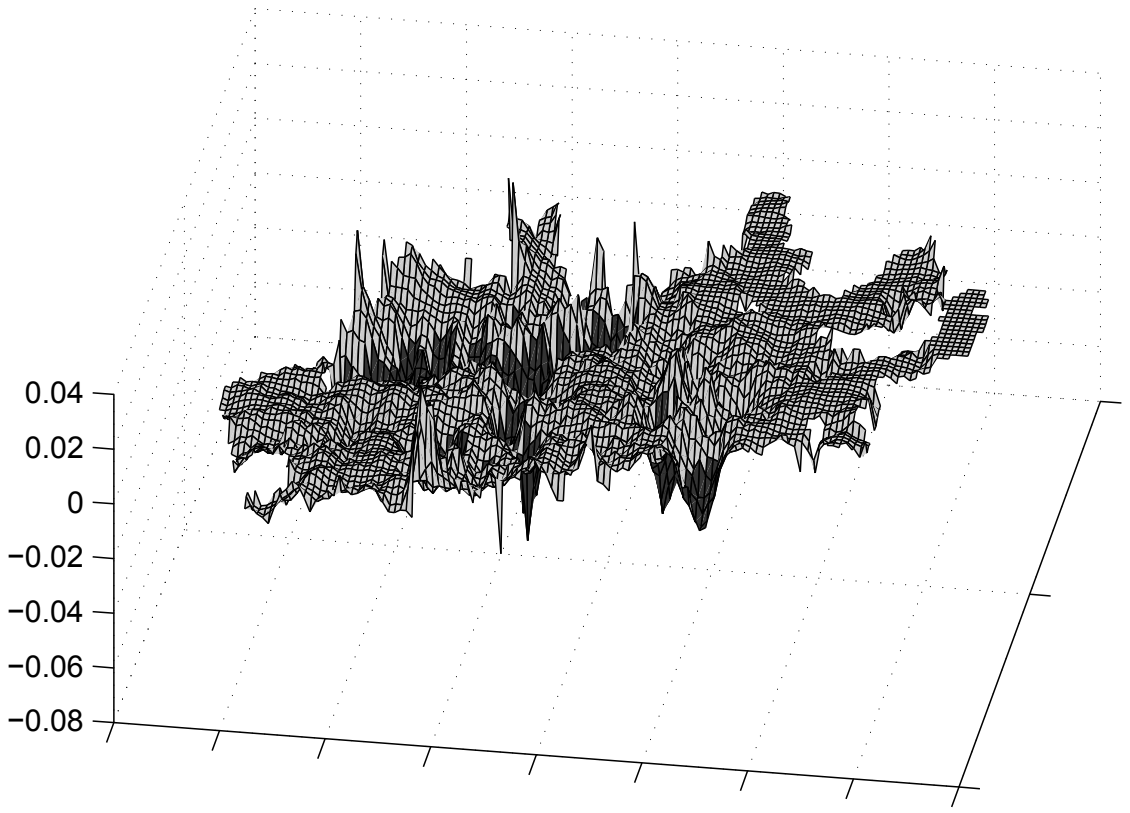

Figure 6: Difference field in infestation, $\bar{\mathbf{y}}|\mathbf{z}_{\text{hetero}} - \bar{\mathbf{y}}|\mathbf{z}_{\text{homo}}$, between the estimates produced by the heterogeneous and homogeneous inspection models.

2.4.1. Model Comparison

There are substantive differences between the estimated infestation probabilities produced by the homogeneous and heterogeneous inspector models at a fine scale. The difference between the two probability fields, shown in figure 6, indicates that the estimates produced by the heterogeneous inspector model are much more peaked and concentrated than those produced by the homogeneous inspector model. The posterior means for $k_{\mathbf{u}}$ in the heterogeneous and homogeneous inspector model are 1.7288 and 1.6667 respectively, indicating that the posterior distributions, induced by the two models, for the spatial component are quite similar.

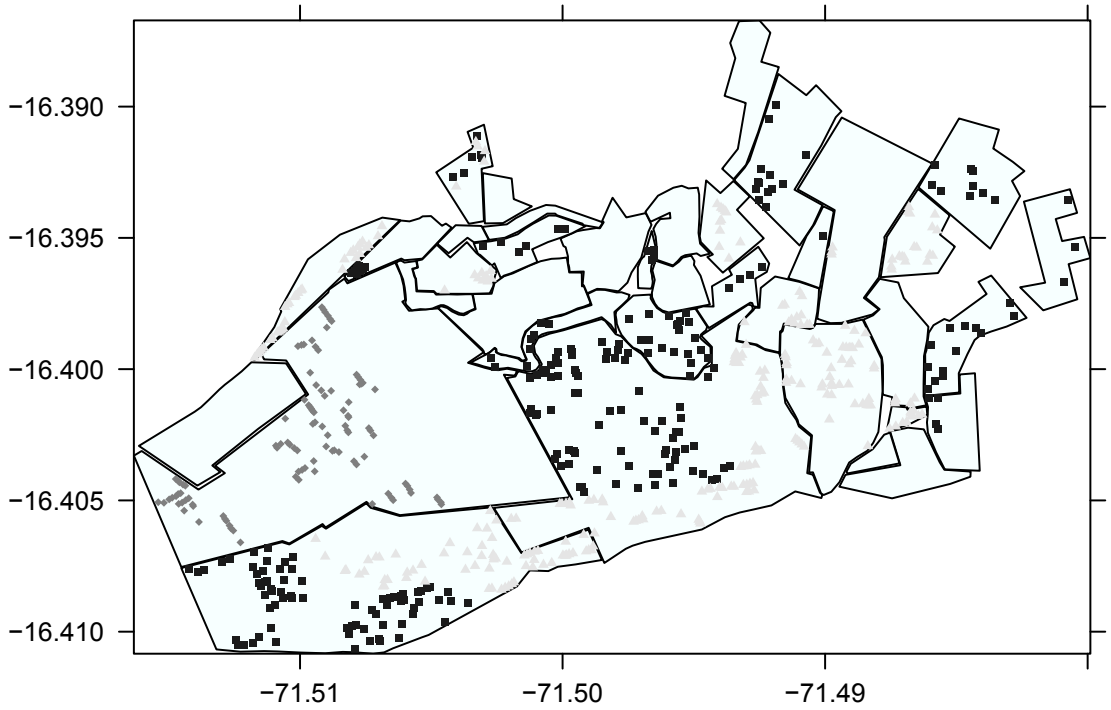

Figure 7: Distribution of households surveyed by a few example inspectors across Mariano Melgar, Arequipa, Peru demonstrating the aggregation in space of their assignment.

2.5. Impact of Inspector Distribution on Estimates

From figure 7, it is clear that inspectors are not distributed evenly or randomly across the region. One point of interest in this analysis is how different distributions of inspectors in space affect the accuracy of our model estimates. Particularly, we are interested in the case where the observation error becomes spatially correlated due to how the inspectors are distributed or assigned to households. In these scenarios, a motivating concern is to what degree one can obtain meaningful estimates of infestation probability and inspector sensitivity.

If there was no spatial correlation in the risk of infestation, then for any observed realization of the data, $\mathbf{z}$, there are many combinations of inspector sensitivity and true infestation that can lead to similar likelihoods. A decrease in the accuracy of the inspector combined with a corresponding increase in the amount of true infestation in a household leads to a

similar likelihood for the household. However, in our model, the similarity of the risk of infestation between households in close proximity can prevent this confusion between the many possible realizations of $\mathbf{y}$. Hence, in the absence of strong prior information regarding the accuracy of inspectors, the exact set of households inspected by each inspector becomes vitally important as it determines the extent to which the spatial component of the model can separate among the many possible combinations of hierarchical variables.

By this line of reasoning, distributing inspectors uniformly at random over the map would reduce this confounding between inspector sensitivity and infestation. In this section, we will investigate by simulation the effects of different strategies for assigning inspectors. Consensus on infestation status by a number of inspectors in an area increases the validity of the evidence, while repeated disagreement suggests insensitivity on the part of some inspectors.

2.5.1. Comparing Randomized to Actual Inspector Assignments

Distributing inspectors uniformly at random reduces confounding between inspector sensitivity and infestation. There are a number of questions we investigate through simulation. **Question 1.1**: how much better is a uniform assignment than the actual assignment? To approach this question, we will generate data from the binomial inspection model and our fitted Gaussian field, and compare the resulting estimates from our model under the two cases that 1) the inspector-to-household labels are identical to those found in the Mariano Melgar data; 2) inspectors are reassigned to households at random. We hypothesize that randomized assignment will have lower error in terms of estimating infestation prevalence, but that the estimation error would ideally not be too dissimilar - a dramatic drop in error through randomization would indicate the possibility of serious confounding when analyzing the Mariano Melgar data. **Question 1.2**: how does uniform random assignment of inspectors compare to the actual assignment for comparing the sensitivity of different inspectors based on the data? This question can be investigated through simulation as inspector sensitivities can be drawn from a specified β-distribution. We then compare the estimation of $\mathbf{y}$

when the prior on $\boldsymbol{\beta}$ differs from the generating distribution and when it matches perfectly, which would yield the lowest estimation error.

Example 1 We simulate the insect infestation, $\mathbf{y}$, from the posterior distribution from the Mariano Melgar data, and inspector sensitivity, $\boldsymbol{\beta}$, from the inspector sensitivity prior $\mathrm{B}(6.5, 2)$. Briefly, inspector-to-household labels are drawn uniformly at random for each inspector (while maintaining the total number of households inspected by each inspector as in the Mariano Melgar data). Simulated data is then drawn according to the binomial inspection model. To measure the performance in estimation accuracy, we consider the squared distance discrepancy between the simulated discrete infestation field $\mathbf{y}$, and the estimated posterior probability of infestation from the simulated data (also known as the or the Brier score for probability forecasts Brier (1950)).

To answer **question 1.1**, (when a $\mathrm{B}(1, 1)$ prior is used to analyze the data), the difference in mean Brier scores between data generated by the Mariano Melgar inspector assignment and random assignments of inspectors has a p-value of 0.0073 in favor of the randomized assignments. The significance of these results is similar when other priors are used. Thus, there is evidence that a randomized assignment of inspectors would have been better than the actual assignment in which most regions were inspected by only a subset of the inspectors but the difference is moderate. For **question 1.2**, the difference in mean Brier scores when using a misspecified $\mathrm{B}(1, 1)$ compared to using the specified generating prior of $\mathrm{B}(6.5, 2)$ has a p-value of 0.0004, using the Mariano Melgar inspector assignment. Under the randomized assignments, the p-value is less significant at 0.0185. These results suggest that error induced by possible misspecification is much more significant in the Mariano Melgar analysis than would be the case if the inspectors were randomly assigned to households.

Effect of randomized inspector-to-household assignment

Model								
	Homogeneous		Heterogeneous					
Prior	$\lvert\hat{\mathbf{y}}-\mathbf{y}\rvert_2$		$\lvert\hat{\mathbf{y}}-\mathbf{y}\rvert_2$		$\lvert\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\rvert_2$		$\mathrm{Cor}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})$	
$n=50$	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Actual assignments								
B(1, 1)	20.9611	0.3053	21.2609	0.2835	1.4122	0.0806	0.4456	0.0682
B(5, 5)	21.0010	0.2810	22.3966	0.2542	1.8430	0.0411	0.5333	0.0708
B(6.5, 2)	20.9363	0.2377	20.9187	0.2308	0.6613	0.0479	0.6824	0.0627
Randomized assignments								
B(1, 1)	20.7721	0.3049	21.0009	0.2850	1.2454	0.1289	0.5400	0.1125
B(5, 5)	20.8264	0.2660	22.1084	0.2473	1.7918	0.0486	0.6084	0.0961
B(6.5, 2)	20.7660	0.2892	20.7710	0.3225	0.6848	0.0561	0.6486	0.0784

Table 2: Estimation error in the infestation ($|\hat{\mathbf{y}}-\mathbf{y}|_2$) and inspector sensitivities ($|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}|_2, \mathrm{Cor}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})$) for simulated data when inspectors are assigned randomly to locations compared to the inspector assignment in the Mariano Melgar data. For comparison, the infestation estimation error for the homogeneous inspection error model is given in contrast to the error in the heterogeneous model.

We conclude that randomization reduces estimation error of the infestation and, randomization increases resiliency to misspecification of inspector sensitivity priors. We hypothesize the reason for this behavior is not simply the lack of spatial similarity in the distribution of inspectors, but rather the lack of correlation between how inspector sensitivities are distributed and how the insect infestation is distributed across space. It is difficult to attribute an observed mild infestation to a truly mild infestation observed by very accurate inspectors or to a very intense infestation observed by very unreliable inspectors. This scenario occurs when for a given household; the sensitivity of the assigned inspector is strongly negatively correlated with the point-wise risk of infestation.

2.5.2. Strength of Confounding

Question 2.1: Does increasing the correlation between distributions of inspectors and the distribution of insect infestation increase estimation error? Increasing the strength of the negative correlation in these cases should not only increase the estimation error of the

infestation, but reduce our ability to accurately estimate the inspector sensitivities. Further, stronger negative correlation should magnify the error induced by prior misspecification. Consequently, **question 2.2** does weakening this correlation reduces estimation error?

To address these questions we first partition the households of Mariano Melgar into regions (see Appendix A.1.1 for the elementary schema used). We then simulate infestations of varying intensity region-by-region and inspector sensitivities. Lastly, we assign inspectors to regions such that the correlation is negative between the inspectors accuracy and the intensity of the infestation within the region. Actual inspectors to household assignments are simulated last using a simple GMRF field model to ensure that they are spatially concentrated by inspector, within region (algorithm given in Appendix A.2). If $\mathbf{y}$ is the infestation, then we introduce $\bar{\mathbf{y}}$, which is the average rate of infestation by region. With $\boldsymbol{\beta}$ being the vector of inspector accuracies, the inspector-to-region assignments are drawn according to the following Gibbs distribution,

$$f(\bar{\mathbf{y}}, \boldsymbol{\beta}) \propto \exp\{-k(1 + \mathrm{Cor}(\mathbf{y}, \boldsymbol{\beta}))^2\} \tag{2.5.1}$$

where this distribution is normalized over all of the finite possible inspector-to-region assignments. The positive constant k controls the degree of negative correlation in the resulting sample assignment. Small values of k will result in assignments with low sample correlation, whereas large values of k will result in assignments with strong negative correlation between infestation intensity and inspector sensitivity.

Example 2 The Mariano Melgar household locations are divided into six regions. The vector presence in each region is simulated with an identical precision, $k_{\mathbf{u}}$ of 4.64 (estimated from the Mariano Melgar data) and intercepts of $[-1, 0, -4, -1.5, -1, -4]$ which correspond to low to high (from -4 to 0) infestation intensity. Thirty-two inspector accuracies are drawn: half from a $\mathrm{B}(3, 7)$ distribution, a quarter from a $\mathrm{B}(6, 4)$ distribution, and a quarter from a $\mathrm{B}(8, 2)$ distribution which correspond to low, moderate, and high inspector sensitivity.

Three correlation strengths are considered: k equal to $[0.2, 20, 2000]$, corresponding to low, medium, and high negative correlation. The primary interest is the effect of this correlation strength on estimation quality.

The impact of k on estimating $\mathbf{y}$ and $\boldsymbol{\beta}$ is shown in table . From figures 8 and 9, stronger negative correlation creates greater overall estimation error in infestation and inspector sensitivity. This effect is evident even when priors are exactly specified, or it is known in advance which inspectors belong to which low, medium, or high-accuracy distribution. However, the increased error is especially evident when the prior is misspecified.

Further as inspector accuracy is highly heterogeneous by design, the homogeneous inspection error model is very unsuitable. The difference in mean Brier scores when $k = 20$ and $k = 0.2$ has a Z-score of 15.95 (p-value < 0.0001) and between $k = 200$ and $k = 20$ is similarly large at 18.68 (p-value < 0.0001). Although these values are when a $\mathrm{B}(1, 1)$ is used and for the heterogeneous inspection model, the results are similar for other priors and for the homogeneous inspection model.

To contrast the examples, because the inspector error rates were drawn from a single unimodal distribution in example 1, it was more reasonable to use a single error rate to describe the group performance of the inspectors. However in example 2, inspectors were drawn from a combination of unimodal distributions; hence, the use of a single homogeneous error rate to describe their sensitivity resulted in poorer infestation estimates. We have shown that the increased correlation between inspector sensitivities and the infestation induces significant estimation error.

2.5.3. Frame of Reference to Improve Estimation

In most applications, the strength of the correlation between infestation intensity and inspector sensitivity is unknowable. With access only to reported infestation rates, it is impossible to diagnose how much confounding is present. However, there is a mollifier to this confounding that is effective even when the confounding is very strong. Our proposed

solution is to find, before the data is collected, a region of the area of interest where a strong infestation is likely to be present and then have a majority of the inspectors assigned to households in this region uniformly. This frame of reference region can be used to learn the relative inspector accuracies and accurately infer the infestation. The heuristic is that by contrasting the repeated failure of certain inspectors to detect insects in a region of intense infestation, the model is able to attribute this observed absence to insensitivity on the part of said inspectors. If these insensitive inspectors are distributed uniformly across the frame of reference region and said insensitive inspectors were mistakenly taken to be sensitive, then the implied roughness of the infestation field in the frame of reference region will be incompatible with the relative smoothness in the other regions.

Example 3 The parameterization of this simulation is identical to the one in example 2. The difference is the presence of a frame of reference region, which in these simulations is region 2, where the infestation prevalence is the highest among the regions. Each inspector is first assigned to a random, uniform subset of households in the frame of reference region. We then simulate the inspector assignments for the remaining five regions identically to example 2.

Effect of correlation between inspector assignments and infestation intensity

		Model							
		Homogeneous		Heterogeneous					
Prior $n = 50$	k	$\lvert\hat{\mathbf{y}} - \mathbf{y}\rvert_2$ Mean	SD	$\lvert\hat{\mathbf{y}} - \mathbf{y}\rvert_2$ Mean	SD	$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rvert_2$ Mean	SD	$\text{Cor}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ Mean	SD
		With no frame of reference							
	0.2	35.3138	0.6005	34.4806	0.9162	0.9729	0.2485	0.8327	0.0885
B(1, 1)	20	36.1165	0.6059	35.6554	1.0014	1.2491	0.2520	0.7296	0.1169
	2000	37.7844	0.7692	37.0555	0.9851	1.5087	0.2562	0.6052	0.1735
	0.2	35.1861	0.5709	34.1913	0.8635	0.9002	0.1316	0.8652	0.0530
B(5, 5)	20	36.0754	0.5300	35.3829	0.9296	0.9930	0.1368	0.8228	0.0633
	2000	37.7328	0.5589	36.8144	1.1503	1.1244	0.1867	0.7504	0.1115
	0.2			34.4256	0.8840	0.7101	0.1017	0.9475	0.0164
True	20			35.5826	0.9465	0.6887	0.1063	0.9487	0.0175
	2000			36.7599	0.9412	0.6950	0.1241	0.9476	0.0174

Continued on next page

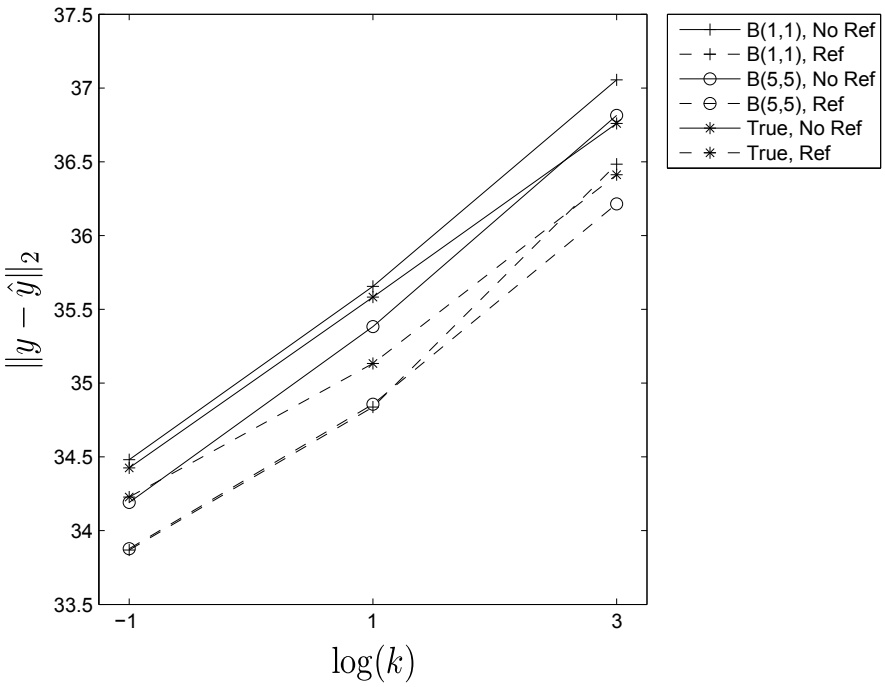

Figure 8: Brier scores, averaged across simulations, plotted against the correlation strength coefficient k.

Table 3 – continued from previous page

		Homogeneous		Heterogeneous					
Prior	k	$\lvert\hat{\mathbf{y}} - \mathbf{y}\rvert_2$		$\lvert\hat{\mathbf{y}} - \mathbf{y}\rvert_2$		$\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rvert_2$		$\text{Cor}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$	
$n = 50$		Mean	SD	Mean	SD	Mean	SD	Mean	SD
				With frame of reference					
	0.2	34.9523	0.4277	33.8690	0.4585	0.4513	0.0724	0.9676	0.0093
B(1, 1)	20	35.7650	0.4948	34.8374	0.4855	0.4360	0.0683	0.9686	0.0099
	2000	37.6855	0.5992	36.4842	0.3298	0.4185	0.0733	0.9713	0.0099
	0.2	34.9010	0.4013	33.8781	0.4563	0.5880	0.0592	0.9637	0.0096
B(5, 5)	20	35.8024	0.4652	34.8565	0.4557	0.5904	0.0525	0.9621	0.0096
	2000	37.7361	0.5760	36.2146	0.3275	0.5809	0.0556	0.9620	0.0109
	0.2			34.2282	0.5130	0.6693	0.0625	0.9638	0.0071
True	20			35.1329	0.4601	0.6470	0.0689	0.9659	0.0095
	2000			36.4124	0.3318	0.6166	0.0540	0.9663	0.0062

Table 3: Estimation error in the infestation ($\lvert\hat{\mathbf{y}} - \mathbf{y}\rvert_2$) and inspector sensitivities ($\lvert\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\rvert_2, \text{Cor}(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$) for simulated data when the good inspectors inspect low infestation regions and bad inspectors inspect high infestation regions (the extremity measured in k). For comparison, we consider two regimes when the frame of reference is present and absence, to demonstrate the reduction in estimation error across model parameters.

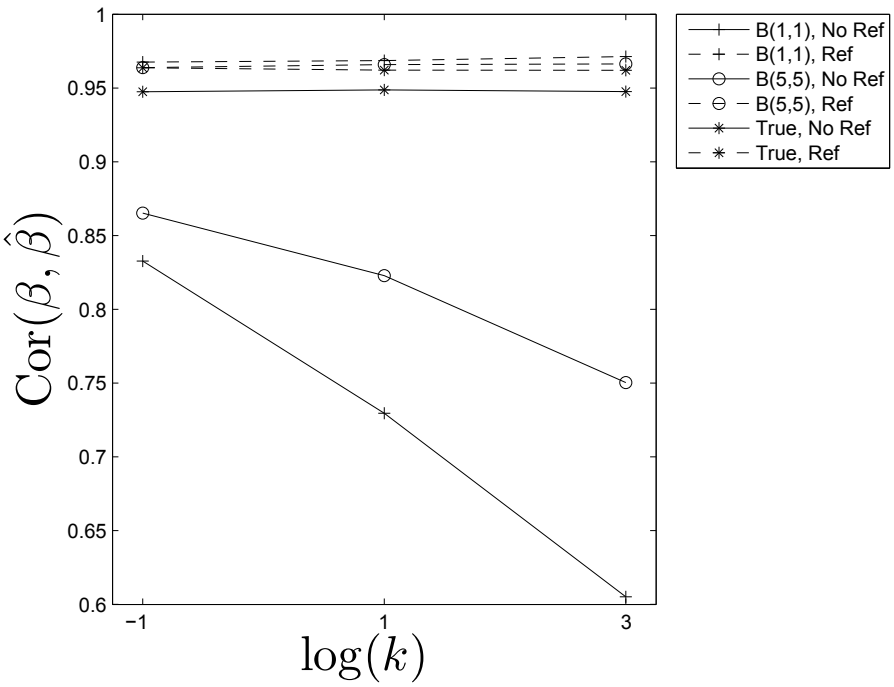

Figure 9: Correlations between estimated inspector sensitivities and simulated sensitivities, averaged across simulations, plotted against the correlation strength coefficient k.

From figures 8 and 9, the inclusion of the frame of reference region appears to remove the error introduced through prior misspecification on the part of inspector as well as the correlation between inspector sensitivities and infestation intensities. The improvement in estimation due to the frame of reference region is noticeable, especially in the estimation of inspector sensitivities. Even in the absence of accurate prior information, identifying the sensitive and insensitive inspectors, we obtain better estimates using the heterogeneous inspector model compared to the homogeneous model. As the heterogeneity in the inspector sensitivities is driving the confounding in this simulation, the homogeneous inspector model is unable to identify which regions have low true presence rates and which regions have low observed presence due to insensitive inspectors, as seen in figure 10. Without this additional model hierarchy and the inspector labeling information, the homogeneous inspector model appears to perform again poorly, similar to when the reference region is absent. The difference in mean Brier scores when the reference region is used and absent

has a Z-score of 3.89 (p-value < 0.0001) in favor of the reference region, when $k = 2000$ and a $B(1, 1)$ prior is used. The difference in mean Brier scores between the homogeneous and heterogeneous inspector model results has a Z-score of 12.42 (p-value < 0.0001) in favor of the heterogeneous model when $k = 2000$ and a $B(1, 1)$ prior is used. Again, there are similar significance results to the above for other values of k and other prior specifications.

Example 3 demonstrates that even if the prior on inspector sensitivities is misspecified, by modeling the heterogeneity in inspector errors and including a frame of reference region, one obtains very accurate estimates of presence absence and is able to identify insensitivity on the part of particular inspectors. The higher the correlation between inspector sensitivities and infestations, the more that stands to be lost by failing to model this heterogeneity. From these simulations, the ability to estimate the true rate of presence is intertwined with the ability to estimate the inspector accuracies both of which benefit from modeling the heterogeneity.

2.5.4. Prior Sensitivity

As we have seen in example 1, in section 2.5.1, the difference in estimation error due to prior misspecification is much smaller for the randomized inspector assignment compare to the actual assignment in Mariano Melgar. As a follow-up consider $k = 2000$, when the estimates are most adversely affected by the correlation between infestation intensities and inspector sensitivities, the p-value between the mean Brier scores when the generating prior is used compared to a default $B(1, 1)$ prior is 0.0625. The above statistic is when the reference region is absent, but when the reference region is present, the difference is much less significant at a p-value of 0.8611. We believe that the use of this frame of reference reduces this prior sensitivity.

2.6. Discussion

The quantification of the underreporting of spatially-correlated data when surveillance error itself is spatially-correlated is a common issue in applied spatial analysis. In our simulations,

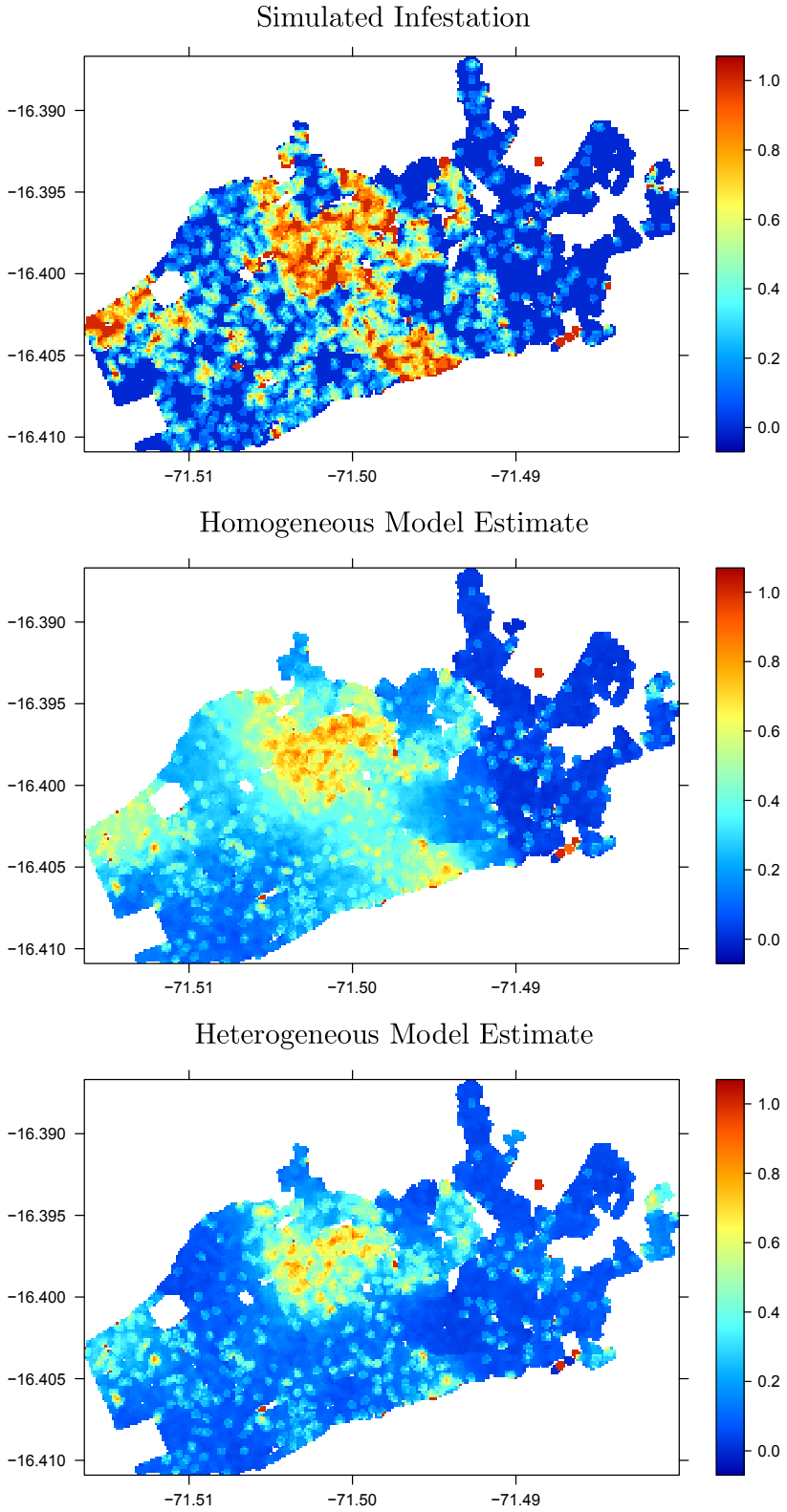

Figure 10: Simulated infestation (top) from example 3, when $k = 2000$ and resulting model estimates produced by the heterogeneous (middle) and homogeneous (bottom).

we have demonstrated that the correlation between the surveillance error and infestation intensity in space can result in severely increased estimation error. While existing work in epidemiology has employed GMRF models for spatial smoothing, these lines of analysis have largely assumed that there is no uncertainty in the surveillance process, Boyd et al. (2005). This raises some difficulties as although we have repeated measurements from the units of surveillance, these measurements are taken over differing locations. We have demonstrated that choosing to model heterogeneous surveillance error and using the inspector-to-household assignment information can be of great use in reducing the error in infestation estimates. Under suitable circumstances and design choices, this model can estimate varying levels of surveillance error across space from a single survey, which is useful for detecting insensitivity in terms of surveillance, as well as providing more accurate measures of intensity.

In spite of these improvements, there are a few limitations. Firstly, while the model is able to distinguish the relative accuracies of inspectors, the overall performance of inspectors and hence overall infestation rates are dependent on the prior. While implementations such as the frame of reference help in this respect, it does not completely remove the influence of the prior. Without any additional information, we do not believe it is possible to determine the exact amount of underreporting from just the observed data. The prior placed on inspector sensitivities dictates in this model the overall amount of underreporting, whereas the value of our model lies in identifying the location of the underreporting. An additional concern to the model specification is the conditionally-autoregressive specification, which is common to all Gaussian Markov random field analyses. We base the spatial interpolation of the infestation risk on the Euclidean distance between household locations and ignore households whose distance is above a threshold. These approaches have no interpretation in terms of a covariance function and have a highly fixed, discrete structure. Another concern is that the non-response in the data is taken to be independent of the risk of infestation. This assumption may be unrealistic, and we may be discarding vital information and biasing our infestation estimates.

In Mariano Melgar, we have identified localities at highest risk of infestation underreporting, due to surveillance error, and guided MoH officials to invest additional resources in controlling transmitters of Chagas disease in these regions. We believe that by identifying and addressing the location of underreporting we can reduce the continual problem of infestation. Furthermore, increased accuracy in mapping of insect infestation also benefits the modeling of the spread and evolution of insect populations.

CHAPTER 3 : Geographic Barrier Modeling

3.1. Introduction

The interest of this section is modeling spatial processes on non-flat surfaces and the accompanying statistical inference. While all the surfaces in consideration are subsets of Euclidean space, the application of traditional spatial models is inappropriate as it does not respect the geometry of the surface. For instance, using an isotropic covariance model for observations on the surface of the sphere would essentially be using chordal distance which in some applications would be unnatural. In particular, we are interested in deformations of surfaces which can be used to model the effects of barriers in spatially distributed data. For ecological data in particular, we are interested in developing a method that would allow one to infer the significance of geographic barriers such as streets on the distributions of insect populations in a city.

In general it is difficult to purpose a valid covariance model which not only respects the geometry of the surface, but also satisfies the positivity requirement. However recent work on the connection between spatial models and the solutions to stochastic partial differential equations circumvents some of these difficulties Lindgren et al. (2011). Further in keeping with the original intentions of the authors, these resulting models are very computationally efficient for large observation windows. Outside of the interest in modeling barriers, these models on flat surfaces are an improvement over the Gaussian markov random field models (GMRF) used in chapter 2. In contrast to other computational methods for large data sets such as: covariance tapering Kaufman et al. (2008), likelihood approximation (in the spatial domain Stein et al. (2004) and in the spectral domain Fuentes (2007)), and fixed rank kriging Cressie and Johannesson (2010), GMRFs avoid modeling the covariance of the spatial process and focus on the graphical interpretation of the precision matrix. The consequence is that it is unclear what the resulting covariance induced by the GMRF is. Further, the process exists only on the specified nodes of the graph - there is no notion of

the process's continual existence between two points. Therefore, it is more appropriate to refer to the GMRF as a graphical model induced by space rather than a spatial model.

To approach the problem of urban boundaries in this work, we emphasize the fact that Mátern equation (3.2.1) is well-defined on smooth surfaces or manifolds. The typical domain of study for spatial data are compact subsets of $\mathbb{R}^2$; the adaptation of covariance models to more complex domains such as spheres is of current interest Gneiting (2012), where the primary technical challenge is that the covariance functional of a process must be positive. The approach introduced by Lindgren et al. (2011) circumvents this difficulty as the stochastic partial differential equation solution is a valid Gaussian process defined on the manifold.

Hence, we will begin as usual and then deform the flat subset of the plane so that it becomes a surface located in $\mathbb{R}^3$. On a computer representation of this surface, the finite element approach has been specifically researched and developed to be suitable for finding approximate solutions on a variety of domains of relevance to engineering and applied problems.

The outline of this chapter is as follows: in section 3.3, we describe in detail the contributions of Lindgren et al. (2011). For practical implementations, we review some classical computational results in section 3.4. We explain our methodology in section 3.5 and report the results on the Mariano Melgar data in section 3.6.

3.2. Background

The Matérn covariance function is one the most widely studied models in spatial statistics Stein (1999). The central fact of this model to this work is that the Matérn covariance can be given as the solution to the following stochastic partial differential equation,

$$(\kappa^2 - \Delta)^{\alpha/2} x(\omega) = W(\omega) \tag{3.2.1}$$

where $\Delta = \nabla^\intercal \nabla$ and $W(\omega)$ is the Gaussian "white-noise" process on $\mathbb{R}^d$ Whittle (1963). The derivation is done using Fourier techniques which emphasizes that the Matérn function describes the covariance of stationary solutions.

If $\mathcal{L}$ is a linear operator such that $\mathcal{L}x = W$, where c is the covariance of the process x, then under the assumption of stationarity, $\mathcal{L}^2 c(\mathbf{d}) = \mathcal{L}^2 c(\mathbf{s}, \mathbf{t}) = \mathcal{L}^2 \mathbb{E}x(\mathbf{s})x(\mathbf{t}) = \mathbb{E}\mathcal{L}x(\mathbf{s})\mathcal{L}x(\mathbf{t}) = \mathbb{E}W(\mathbf{s})W(\mathbf{t}) = \boldsymbol{\delta_0}(\mathbf{s} - \mathbf{t})$. Hence, if x is the solution to the equation $\mathcal{L}x = W$, then c is the Green's function of $\mathcal{L}^2$. Following the above heuristic, c can be found by solving the deterministic equation,

$$(\kappa^2 - \Delta)^\alpha c(\omega) = \delta_0(\omega) \tag{3.2.2}$$

Appplying the Fourier transform to both sides of the equation,

$$(2\pi)^d(\kappa^2 + \|\xi\|^2)^\alpha \mathcal{F}c = \mathcal{F}\{(\kappa^2 - \Delta)^\alpha c\}(\xi) = \mathcal{F}\delta_0(\xi) = 1 \tag{3.2.3}$$

$$\implies \mathcal{F}c(\xi) = \frac{1}{(2\pi)^d} \frac{1}{(\kappa^2 + \|\xi\|^2)^\alpha} \tag{3.2.4}$$

the left-hand side is well-defined even for fractional values of α, see Samko et al. (1993) for a formal treatment. Then, c can be found by inverting the transformation, the well-posedness of this inversion and spectral representation for c is due to the stationarity of x. As the density (3.2.4) depends only on $\|\xi\|$, c is given by the Hankel transformation,

$$c(\omega) = \mathcal{F}^{-1}\left\{ \frac{1}{(2\pi)^d} \frac{1}{(\kappa^2 + \|\xi\|^2)^\alpha} \right\} = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{\exp\{2\pi\xi \cdot \omega\}}{(\kappa^2 + \|\xi\|^2)^\alpha} d\xi \tag{3.2.5}$$

$$= \frac{\|\omega\|^{1-d/2}}{(2\pi)^{d/2}} \int_0^\infty \frac{J_{d/2-1}(\|\omega\|\|\xi\|)\|\xi\|^{d/2}}{(\kappa^2 + \|\xi\|^2)^\alpha} d\|\xi\| \tag{3.2.6}$$

$$= \frac{(\|\omega\|/\kappa)^{\alpha-d/2} K_{\alpha-d/2}(\kappa\|\omega\|)}{(2\pi)^{d/2} 2^{\alpha-1}\Gamma(\alpha)} \tag{3.2.7}$$

where J_n is a Bessel function of the first kind and K_n is a modified Bessel function of the second kind, Abramowitz and Stegun (1972). Using the fact that the leading behavior of $K_n(z) \sim \frac{1}{2}\Gamma(n)(\frac{1}{2}z)^{-n}$ for $n > 0$ as $z \to 0$, the marginal variance is given by,

$$\lim_{\|\omega\| \to 0} c(\omega) = \frac{1}{2^d \pi^{d/2} \kappa^{2\alpha-d}} \frac{\Gamma(\alpha - d/2)}{\Gamma(\alpha)} \tag{3.2.8}$$

3.3. Finite Elements

Let $\mathcal{H}$ be a Hilbert space with the inner product $\langle f, g \rangle = \int_\Omega f(\omega)^\intercal g(\omega) d\omega$. $\mathcal{H}$ is a *seperable* Hilbert space if there exists a countable basis $\{\phi_i\}_{i=1}^\infty$ such that under the induced norm, one can approximate any element $h \in \mathcal{H}$ with a suitable large number of basis elements.

For a sequence $\{f_n\}$ in the Hilbert space $L^2(\Omega)$, we say $\{f_n\}$ converges to f *weakly* if for all $g \in L^2(\Omega)$, $\langle g, f_n \rangle \to \langle g, f \rangle$. This notion is relevant to the discussion of partial differential equations because often times a 'solution' to a differential equation cannot be strictly verified pointwise. However, the problem still has a well-posed weak solution which behaves appropriately when integrated against *test functions* g. Therefore if we are willing to relax the criterion for accepting a solution, we need only to verify its behavior with respect to the inner product against these test functions. Secondly, it may be necessary to restrict the space in which a solution lies - often there are restrictions to the space and

to the qualities of the solution which make a problem well-posed. Computationally, for an infinite dimensional space, it is impossible to evaluate this infinite criterion against all of $L^2(\Omega)$.

In finite elements there are two truncations, the first truncation being the set of test functions $\{\psi_n\}$ necessitated by the limitation that we can only evaluate a finite number of test constraints. The second truncation being $\{\phi_n\}$ due to the necessity of representing the approximate solution, $\tilde{x} = \sum_{i=1}^{n} u_i \phi_i$, with a finite number of terms. When the two sets coincide with one another, we call this approach the *Galerkin method*. In this setting an important distinction is made. For an infinite dimensional space $\mathcal{H}$, a finite dimensional subspace $\mathcal{H}_n$ is a subset spanned by a finite number n of bases. The intuition is that for a nested, increasing set of bases, the weak approximation found by the Galerkin method is a projection of the weak solution onto these nested subspaces. This intuition is made precise in Brenner and Scott (2008). The passing of Neumann boundary constraints to its weak formulation and then its finite element solution is given by,

$$\begin{cases} \mathcal{L}x = f \quad \text{on } \Omega \\ x = g \quad \text{on } \partial\Omega \end{cases} \implies \begin{cases} \text{find } x \in \mathcal{H} \text{ such that} \\ \langle g, \mathcal{L}x \rangle = \langle g, f \rangle \quad \text{for all } g \in \mathcal{H}^{\intercal} \\ x = g \quad \text{on } \partial\Omega \end{cases} \tag{3.3.1}$$

$$\implies \begin{cases} \text{find } x \in \mathcal{H}_n \text{ such that} \\ \langle g, \mathcal{L}x \rangle = \langle g, f \rangle \quad \text{for all } g \in \mathcal{H}_n^{\intercal} \\ x = g \quad \text{on } \partial\Omega \end{cases} \tag{3.3.2}$$

For the equation $\mathcal{L}x = W(\omega)$, we call x_n the weak solution if for all $h \in \mathcal{H}$, $\langle h, x_n \rangle$ converges in distribution to $\langle h, x \rangle$. As random variables of the form $\langle h, W \rangle$ are Gaussian, this convergence can be determined by $\mathbb{E}\langle h, x_n \rangle \to \mathbb{E}\langle h, x \rangle$ and $\mathbb{E}\langle f, x_n \rangle \langle g, x_n \rangle \to \mathbb{E}\langle f, x_n \rangle \langle g, x_n \rangle$

for all $f, g, h \in \mathcal{H}$.

3.3.1. Identities

The gradient of an element of $\mathcal{H}$ is given by $\nabla = \left[\frac{\partial}{\partial \omega_i}\right]$ or $\nabla f = \left[\frac{\partial f}{\partial \omega_i}\right]$. The Laplacian Δ is then defined as $\Delta = \nabla^\intercal \nabla = \sum_{i=1}^{d} \frac{\partial^2}{\partial {\omega_i}^2}$. For $\omega \in \partial\Omega$, the unit length, outward normal vector at ω is given by $\mathbf{n}(\omega)$ and the normal derivative of f at ω is the directional derivative given by $\partial_{\mathbf{n}} f(\omega) = \mathbf{n}(\omega)^\intercal \nabla f(\omega)$.

Theorem 3.3.1 (Stochastic Green's first identity)**.** *If $\nabla f \in L^2(\Omega)$ and Δg is $L^2(\Omega)$ bounded then the following holds with probability one,*

$$\langle f, -\Delta g \rangle_\Omega = \langle \nabla f, \nabla g \rangle_\Omega - \langle f, \partial_{\mathbf{n}} g \rangle_{\partial\Omega} \tag{3.3.3}$$

A proof of this identity can be found in the appendix of Lindgren et al. (2011). The computational application of this identity is to work with the gradients of these bases rather than the Laplacians. The complementary important fact is the following for f, g that have suitably regular gradients (more specifically belong to the Sobolev space $\mathcal{H}^1(\Omega)$), then the following holds for compact Ω,

$$\langle \Delta^{1/2} f, \Delta^{1/2} g \rangle_\Omega = \langle \nabla f, \nabla g \rangle_\Omega \tag{3.3.4}$$

or when $\langle f, \partial_{\mathbf{n}} g \rangle = \langle \partial_{\mathbf{n}} f, g \rangle = 0$.

3.3.2. Weak Solutions

Consider the Galerkin approximation to the weak solution to the spde give in (3.2.1) with the Neumann boundary conditions,

$$\partial_{\mathbf{n}} (\kappa^2 - \Delta)^i x(\omega) = 0, \ \omega \in \Omega, \ i = 0, \ldots, \lfloor (\alpha - 1)/2 \rfloor \tag{3.3.5}$$

which according to Lindgren et al. (2011) 'mollifies' the problem of uniqueness. For computational convenience, the number of boundary conditions will become clearer.

Under the bases and test functions $\{\phi_i\}_{i=1}^n$ define the following matrices,

$$\mathbf{C}_{ij} \triangleq [\langle \phi_i, \phi_j \rangle]_{i,j=1}^n \tag{3.3.6}$$

$$\mathbf{G}_{ij} \triangleq [\langle \nabla\phi_i, \nabla\phi_j \rangle]_{i,j=1}^n \tag{3.3.7}$$

then the approximate weak solution of (3.2.1) is given by the precision of the coefficients $\mathbf{u}$, which for a given α is defined as,

$$\mathbf{Q}_{\alpha=1} \triangleq \kappa^2\mathbf{C} + \mathbf{G} \tag{3.3.8}$$

$$\mathbf{Q}_{\alpha=2} \triangleq (\kappa^2\mathbf{C} + \mathbf{G})\mathbf{C}^{-1}(\kappa^2\mathbf{C} + \mathbf{G}) \tag{3.3.9}$$

$$\mathbf{Q}_{\alpha>2} \triangleq (\kappa^2\mathbf{C} + \mathbf{G})\mathbf{C}^{-1}\mathbf{Q}_{\alpha-2}\mathbf{C}^{-1}(\kappa^2\mathbf{C} + \mathbf{G}) \tag{3.3.10}$$

Note that only (3.3.9) is the Galerkin approximation. The exact nature of the other approximate solutions is clear through their derivations.

We begin with the derivation of the Galerkin approximate solution, as it is the most straight forward in terms of precedence. Recall (3.3.2) where $\mathcal{L} = \kappa^2 - \Delta$ then the coefficients $\mathbf{u}$ may be described by the following system,

$$\left[\langle \phi_j, (\kappa^2 - \Delta)\tilde{x} \rangle\right]_{j=1}^n \stackrel{d}{=} [\langle \phi_j, W \rangle]_{j=1}^n \tag{3.3.11}$$

Using the linearity of the operators and the fact that $\tilde{x} = \sum_{i=1}^n u_i\phi_i$, then the left hand

side can be simplied as,

$$\left[\langle\phi_j,(\kappa^2-\Delta)\tilde{x}\rangle\right]_{j=1}^{n} = \left[\sum_{i=1}^{n}\langle\phi_j,(\kappa^2-\Delta)\phi_i\rangle u_i\right]_{j=1}^{n} \tag{3.3.12}$$

$$= \left[\sum_{i=1}^{n}(\kappa^2\langle\phi_j,\phi_i\rangle - \langle\phi_j,\Delta\phi_i\rangle)u_i\right]_{j=1}^{n} \tag{3.3.13}$$

$$= \left[\sum_{i=1}^{n}(\kappa^2\langle\phi_j,\phi_i\rangle + \langle\nabla\phi_j,\nabla\phi_i\rangle)u_i\right]_{j=1}^{n} \tag{3.3.14}$$

$$= (\kappa^2\mathbf{C}+\mathbf{G})\mathbf{u} \tag{3.3.15}$$

The line (3.3.13) follows from using the first Neumann boundary condition and the Green's identity (3.3.3), where the boundary contribution is nullified.

Lemma 3.3.2. *If $f, g \in L^2(\Omega)$ and W is a cylindrical Wiener process then,*

$$\mathbb{E}(\langle f,W\rangle\langle g,W\rangle) = \langle f,g\rangle \tag{3.3.16}$$

This statement is the formalization of the calculation:

$$\mathbb{E}\prod_{i=1}^{2}\int_{\Omega_i} f_i(\omega_i)W(\omega_i)d\omega_i = \mathbb{E}\int_{\prod_i\Omega_i}\prod_{i=1}^{2} f_i(\omega_i)W(\omega_i)d\prod_i\omega_i \tag{3.3.17}$$

$$= \int_{\prod_i\Omega_i}\prod_{i=1}^{2} f_i(\omega_i)\boldsymbol{\delta_0}(\omega)d\prod_i\omega_i \tag{3.3.18}$$

$$= \int_{\Omega}\prod_{i=1}^{2} f_i(\omega)d\omega \tag{3.3.19}$$

because the 'white noise' process has infinite trace, the interchange of operations above is invalid and the rigorous derivation is given in Da Prato and Zabczyk (1992). Hence $(\kappa^2\mathbf{C}+\mathbf{G})\mathbf{u} \sim \mathrm{N}(\mathbf{0},\mathbf{C})$ or $\mathbf{u} \sim \mathrm{N}(\mathbf{0},(\kappa^2\mathbf{C}+\mathbf{G})^{\intercal}\mathbf{C}^{-1}(\kappa^2\mathbf{C}+\mathbf{G}))$.

To obtain the case $\alpha = 1$, the Galerkin approximation is not used. Although $\{\phi_i\}_{i=1}^n$ remains the basis, the test functions in this case are the functions $\{(\kappa^2 - \Delta)^{1/2}\phi_i\}_{i=1}^n$. Then the weak solutions are given by the system of stochastic equations,

$$\left[\langle(\kappa^2 - \Delta)^{1/2}\phi_j, (\kappa^2 - \Delta)^{1/2}\tilde{x}\rangle\right]_{j=1}^n \stackrel{d}{=} \left[\langle(\kappa^2 - \Delta)^{1/2}\phi_j, W\rangle\right]_{j=1}^n \tag{3.3.20}$$

which written in matrix form of (3.3.15) as, where the use of (3.3.4) substitutes the use of (3.3.3)

$$(\kappa^2\mathbf{C} + \mathbf{G})\mathbf{u} \stackrel{d}{=} \left[\langle(\kappa^2 - \Delta)^{1/2}\phi_j, W\rangle\right]_{j=1}^n \tag{3.3.21}$$

by using (3.3.4) again and the covariance of the Wiener process, (3.3.16), we have that $(\kappa^2\mathbf{C}+\mathbf{G})\mathbf{u} \sim \mathrm{N}(\mathbf{0}, (\kappa^2\mathbf{C}+\mathbf{G}))$. Then the precision of $\mathbf{u}$ is $(\kappa^2\mathbf{C}+\mathbf{G})^\intercal(\kappa^2\mathbf{C}+\mathbf{G})^\dagger(\kappa^2\mathbf{C}+\mathbf{G}) = (\kappa^2\mathbf{C}+\mathbf{G})$, where $\dagger$ is the generalized inverse. Hence it is not quite obvious the space spanned by the test functions in this case, the solution is not a valid Galerkin approximation.

For higher order solutions consider the following heuristic, let $\tilde{x}_{(\alpha)}$ be the approximate solution for α. Let $\tilde{x}$ be the weak solution to the following: $(\kappa^2 - \Delta)\tilde{x} = \tilde{x}_\alpha$ then applying the k-th order operator to both sides yields $(\kappa^2 - \Delta)^{(\alpha+2)/2}\tilde{x} = (\kappa^2 - \Delta)^{\alpha/2}(\kappa^2 - \Delta)\tilde{x} = (\kappa^2 - \Delta)^{\alpha/2}\tilde{x}_\alpha = W$. Hence, one find the distribution for the coefficients of the case $\alpha + 2$ from the distribution of the coefficients for α. In the language of finite elements, we require for all the test functions using Green's identity again (3.3.3),

$$(\kappa^2\mathbf{C} + \mathbf{G})\mathbf{u}_{(\alpha+2)} = \left[\langle\phi_j, (\kappa^2 - \Delta)\tilde{x}_{(\alpha+2)}\rangle\right]_{j=1}^n = \left[\langle\phi_j, \tilde{x}_{(\alpha)}\rangle\right]_{j=1}^n = \mathbf{C}\mathbf{u}_{(\alpha)} \tag{3.3.22}$$

If the precision of $\mathbf{u}_{(\alpha)}$ is $\mathbf{Q}_\alpha$, then given that $\mathbf{C}$ is symmetric, the precision of $\mathbf{C}\mathbf{u}_{(\alpha)}$ is $\mathbf{C}^{-1}\mathbf{Q}_\alpha\mathbf{C}^{-1}$. Hence, the precision of $\mathbf{C}\mathbf{u}_{(\alpha+2)}$ is then $(\kappa^2\mathbf{C}+\mathbf{G})^\intercal\mathbf{C}^{-1}\mathbf{Q}_\alpha\mathbf{C}^{-1}(\kappa^2\mathbf{C}+\mathbf{G})$. For a given integer α, all the precision matrices for the coefficients describing the weak solution

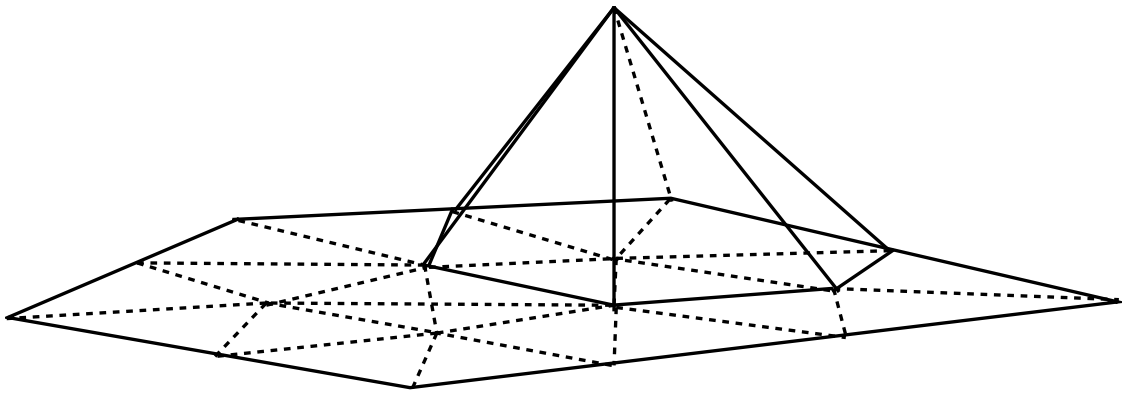

Figure 11: Example of a B-spline basis function on a triangulated subset of the plane.

can be calculated started from one of the two base cases.

3.4. Meshes & Splines

We consider solution domains approximated as triangulated meshes. The surface is represented by a series of nodes or points in space and edges or links between pairs of nodes. As the surface is triangulated, another description of the surface may be given in triples of nodes representing points in a triangle. On a triangulated area, the outer boundary is piece-wise linear. If the sampling location of the data is the collection $\{s_i\}_{i=1}^m$ then we will refer to the collection of nodes in the meshas $\{\omega_i\}_{i=1}^n$. For the B-spline basis, the piece-wise linear basis ϕ_i is indexed by the node i,

$$\phi_i(\omega) = \begin{cases} 1 & \omega_i \\ 0 & \omega_j \in N_i \end{cases} \tag{3.4.1}$$

where if N_i is the set of points connected to i in the mesh through the edges of a triangle. On $\mathbb{R}^2$, the B-spline basis can be easily visualized in figure 11.

The exact numerical calculation from working with these B-spline functions is outlined in the following.

3.4.1. Calculation of Covariance Matrices

Let $\mathcal{T}$ be the set of triangles forming the triangulation and for each node i, let $\mathcal{T}_i$ be the set of triangles supporting the function ϕ_i - all triangles who share the node i in common. To compute the precision matrix $\mathbf{Q}$, one must evaluate the following integrals:

$$\tilde{\mathbf{C}}_{i,i} = \langle \phi_i, 1 \rangle_\Omega = \int_{\mathcal{T}_i} \phi_i(\omega) dS(\omega) \tag{3.4.2}$$

$$\mathbf{C}_{i,j} = \langle \phi_i, \phi_j \rangle_\Omega = \int_{\mathcal{T}_i \cap \mathcal{T}_j} \phi_i(\omega)\phi_j(\omega) dS(\omega) \tag{3.4.3}$$

$$\mathbf{G}_{i,j} = \langle \nabla\phi_i, \nabla\phi_j \rangle_\Omega = \int_{\mathcal{T}_i \cap \mathcal{T}_j} \nabla\phi_i(\omega)^\intercal \nabla\phi_j(\omega) dS(\omega) \tag{3.4.4}$$

$$\mathbf{B}_{i,j} = \langle \phi_i, (\nabla\phi_j)^\intercal \mathbf{n} \rangle_{\partial\Omega} = \int_{\partial(\mathcal{T}_i \cap \mathcal{T}_j) \cap \partial\Omega} \phi_i(\omega)\{\nabla\phi_j(\omega)^\intercal \mathbf{n}(\omega)\} dS(\omega) \tag{3.4.5}$$

Each integral over the set of triangles can be decomposed into a sum over each triangle, ie. $\mathbf{C}_{i,j} = \int_{T \in \mathcal{T}_i \cap \mathcal{T}_j} \phi_i(\omega)\phi_j(\omega) dS(\omega)$. Thus, the integrals are evaluated over each triangle seperately then aggregated together.

3.4.2. Quadrature

Each point in triangle T can be parameterized in two dimensions (θ_1, θ_2) by noting that T is convex. If the three corner nodes of the triangle are (x_i, y_i, z_i), then each point on the interior can be represented as,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = (1 - \theta_1 - \theta_2) \begin{bmatrix} x_1 \\ y_1 \\ z_1 \end{bmatrix} + \theta_1 \begin{bmatrix} x_2 \\ y_2 \\ z_2 \end{bmatrix} + \theta_2 \begin{bmatrix} x_3 \\ y_3 \\ z_3 \end{bmatrix} \tag{3.4.6}$$

Similarly the mapping $\mathbf{F} : (\theta_1, \theta_2) \to (x, y, z)$ is,

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \\ z_2 - z_1 & z_3 - z_1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{3.4.7}$$

The Jacobian of this transformation is then,

$$\det \begin{bmatrix} 1 & x_2 - x_1 & x_3 - x_1 \\ 1 & y_2 - y_1 & y_3 - y_1 \\ 1 & z_2 - z_1 & z_3 - z_1 \end{bmatrix} = \det \begin{bmatrix} 1 & 1 & 1 \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \end{bmatrix} = 2\,|T| \tag{3.4.8}$$

For a basis ϕ_i supported on the triangle T at the node i. It is immaterial the labeling of the other nodes. To evaluate the following integrals, re-parameterize the domain of integration to the cannonical triangle T_0 (given by the vertices $(0,0),(1,0),(0,1)$). For the integrals in 3.4.2,

$$\int_T \phi_i(\mathbf{s}) d\mathbf{s} = \int_{T_0} \phi_i \circ \mathbf{F}(\boldsymbol{\theta}) \det(\mathbf{F}) d\boldsymbol{\theta} \tag{3.4.9}$$

$$= \det(\mathbf{F}) \frac{|T_0|}{3} \sum_{i=1}^{3} \phi_i([x_i, y_i, z_i]^\intercal) = \frac{|T|}{3} \tag{3.4.10}$$

(3.4.10) is using the three point quadrature rule for approximating integrals over triangular regions. For the B-spline basis, ϕ is a first order polynomial and the formula is exact. Lastly, recall that $\det(\mathbf{F}) = 2|T|$.

$$\int_T \phi_i(\mathbf{s})d\mathbf{s} = \int_{T_0} \phi_i \circ \mathbf{F}(\boldsymbol{\theta}) \det(\mathbf{F}) d\boldsymbol{\theta} \tag{3.4.11}$$

$$= \det(\mathbf{F}) \frac{|T_0|}{3} \sum_{i=1}^{3} \phi_i([x_i, y_i, z_i]^\intercal) = \frac{|T|}{3} \tag{3.4.12}$$

Similarly for the integrals in 3.4.3,

$$\int_T {\phi_i}^2(\mathbf{s})d\mathbf{s} = \det(\mathbf{F}) \frac{|T_0|}{3} \sum_{i=1}^{3} {\phi_i}^2\Big(\Big[x_i^{(m)}, y_i^{(m)}, z_i^{(m)}\Big]^\intercal\Big) = \frac{|T|}{6} \tag{3.4.13}$$

where $\Big\{\Big[x_i^{(m)}, y_i^{(m)}, z_i^{(m)}\Big]^\intercal\Big\}_{i=1}^3$ is the set of midpoints on the triangle T. Here because for the B-spline basis ϕ^2 is a second order polynomial, we require a higher order quadrature scheme to obtain an exact value for the integral. Again for the cross-terms, the midpoint rule is needed,

$$\int_T \phi_i\phi_j(\mathbf{s})d\mathbf{s} = \det(\mathbf{F}) \frac{|T_0|}{3} \sum_{i=1}^{3} \phi_i\phi_j\Big(\Big[x_i^{(m)}, y_i^{(m)}, z_i^{(m)}\Big]^\intercal\Big) = \frac{|T|}{12} \tag{3.4.14}$$

Note, every parameterization and labeling of i, j yields the same value, and hence not much attention is paid to this choice.

3.4.3. Sparsification

For the precision matrices of the weak solutions for $\alpha = 2$, (3.3.9), and the higher order solutions,(3.3.10), the choice of $\mathbf{C}^{-1}$ is often replaced by $\tilde{\mathbf{C}}^{-1}$ (from (3.4.2)). The practical reason for this is that for the B-spline basis although $\mathbf{C}$ is sparse, there is no garauntee on the sparsity of its inverse and hence the sparsity of the precision. Recall that $C_{ij} = \langle \phi_i, \phi_j \rangle_\Omega$ and the row sum of $\mathbf{C}$ is then $\sum_{j=1}^n \langle \phi_i, \phi_j \rangle = \langle \phi_i, 1 \rangle$ because $\{\phi_i\}$ forms a *partition of unity.* Hence this approximation is called the *loaded mass* approximation as it concentrates the elements of $\mathbf{C}$ onto the diagonal of the matrix.

An interpretation of this is as follows, recall the entries of $\mathbf{C}$ and the integral, but instead of using the midpoint quadrature rule to compute the exact value, consider the first order scheme using the verticies, for $T \in \mathcal{T}_i \cap \mathcal{T}_j$

$$\int_T \phi_i \phi_j(\mathbf{s}) d\mathbf{s} \approx \frac{|T|}{3} \delta_0(i,j) \tag{3.4.15}$$

For sufficiently fine grids, this approximation is negligible, but being that $\tilde{\mathbf{C}}$ is diagonal and hence it's inverse will preserve the sparsity of $\mathbf{Q}$.

3.4.4. Gradient

Although the integrals in equation 3.4.4 can be calculated in a similar manner to the above using a change of variables, for this particular basis, it is simpler to derive the expressions directly based on its simplicity. For a given triangle T with nodes labeled $\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2$, define the vectors $\mathbf{v}_0 = \mathbf{s}_2 - \mathbf{s}_1, \mathbf{v}_1 = \mathbf{s}_0 - \mathbf{s}_2, \mathbf{v}_2 = \mathbf{s}_1 - \mathbf{s}_0$ and let ϕ_i be the linear basis function centered on node i. There are a few equivalent expressions for the values of the gradients, in maintaining convention with Lindgren and Rue (2007), $\nabla\phi_0$ and $\nabla\phi_1$ are derived here using $\mathbf{v}_0$ and $\mathbf{v}_1$.

As the B-spline function ϕ_0 is equal to one at $\mathbf{s}_0$ and must be equal to zero alone the line $p\mathbf{s}_1 + (1-p)\mathbf{s}_2$. As $\mathbf{v}_0$ and $\mathbf{v}_1$ share node $\mathbf{s}_2$ in common, $\frac{\mathbf{v}_0{}^\intercal \mathbf{v}_1}{\mathbf{v}_0{}^\intercal \mathbf{v}_0}\mathbf{v}_0$ is the projection of $\mathbf{v}_1$ in the direction of $-\mathbf{v}_0$. The vector $\mathbf{v}_1 - \frac{\mathbf{v}_0{}^\intercal \mathbf{v}_1}{\mathbf{v}_0{}^\intercal \mathbf{v}_0}\mathbf{v}_0$ is then the rejection or the shortest path from the vector $\mathbf{v}_0$ to the point $\mathbf{s}_0$. As the derivative in the direction $\frac{\mathbf{v}_0{}^\intercal \mathbf{v}_1}{\mathbf{v}_0{}^\intercal \mathbf{v}_0} - \mathbf{v}_1$ should be normalized to be -1, the normalized gradient is,

$$\nabla\phi_0 = \frac{\mathbf{v}_1 - \frac{\mathbf{v}_0{}^\intercal \mathbf{v}_1}{\mathbf{v}_0{}^\intercal \mathbf{v}_0}\mathbf{v}_0}{\|\mathbf{v}_1 - \frac{\mathbf{v}_0{}^\intercal \mathbf{v}_1}{\mathbf{v}_0{}^\intercal \mathbf{v}_0}\mathbf{v}_0\|^2} \tag{3.4.16}$$

The reason that the direction of $\nabla\phi_0$ must be along the rejection is that it must be orthogonal to $\mathbf{v}_0$ in the plane formed by $\{\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2\}$. If the form of ϕ_0 is $\phi_0(\mathbf{s}) = 1 + \langle \nabla\phi_0, \mathbf{s} - \mathbf{s}_0 \rangle$,

then for all $\mathbf{s}_p = p\mathbf{s}_1 + (1-p)\mathbf{s}_2$, $\phi_0(x)$ must equal zero.

$$1 - \langle \nabla\phi_0, \mathbf{s}_0 \rangle + \langle \nabla\phi_0, \mathbf{s}_p \rangle = 1 + p\langle \nabla\phi_0, \mathbf{v}_2 \rangle + (1-p)\langle \nabla\phi_0, -\mathbf{v}_1 \rangle \tag{3.4.17}$$

$$= 1 - \langle \nabla\phi_0, \mathbf{v}_1 \rangle - p\langle \nabla\phi_0, \mathbf{v}_0 \rangle \tag{3.4.18}$$

As this expression must equal zero for all $p \in [0,1]$, it must be that $\langle \nabla\phi_0, \mathbf{v}_0 \rangle$ is also equal to zero.

Now as the B-spline basis forms a partition of unity ($\phi_0 + \phi_1 + \phi_2 = 1$), it is only necessary to compute $\nabla\phi_1$ as well to obtain all three gradients. To express $\nabla\phi_1$ in terms of $\mathbf{v}_0$ and $\mathbf{v}_1$, compute the shortest path from $\mathbf{v}_1$ to $\mathbf{s}_1$ which can be expressed by projecting $-\mathbf{v}_0$ onto $\mathbf{v}_1$ then taking the rejection as $-\mathbf{v}_0 - \frac{\mathbf{v}_1{}^\intercal(-\mathbf{v}_0)}{\mathbf{v}_1{}^\intercal\mathbf{v}_1}\mathbf{v}_1$. Similarly to normalize the derivative in the direction of the shortest path from $\mathbf{s}_1$ to $\mathbf{v}_1$, the final expression is,

$$\nabla\phi_1 = \frac{-\mathbf{v}_0 - \frac{\mathbf{v}_1{}^\intercal(-\mathbf{v}_0)}{\mathbf{v}_1{}^\intercal\mathbf{v}_1}\mathbf{v}_1}{\|\mathbf{v}_0 - \frac{\mathbf{v}_1{}^\intercal\mathbf{v}_0}{\mathbf{v}_1{}^\intercal\mathbf{v}_1}\mathbf{v}_1\|^2} \tag{3.4.19}$$

These gradients are constant over each triangle T when taken with respect to the canonical basis as the B-spline basis is linear.

$$\int_T \nabla\phi_i(\omega)^\intercal \nabla\phi_j(\omega) dS(\omega) = |T| \nabla\phi_0{}^\intercal \nabla\phi_1 \tag{3.4.20}$$

Let θ_2 be the angle between $-\mathbf{v}_0$ and $\mathbf{v}_1$, then the inner product can be calculated using the fact that $-\mathbf{v}_0{}^\intercal\mathbf{v}_1 = \|\mathbf{v}_0\|\|\mathbf{v}_1\|\cos\theta_2$ and $|T| = \frac{1}{2}\|\mathbf{v}_0\|\|\mathbf{v}_1\|\sin\theta_2$,

$$\nabla\phi_0{}^\intercal\nabla\phi_1 = \frac{\mathbf{v}_1 - \frac{\mathbf{v}_0{}^\intercal\mathbf{v}_1}{\mathbf{v}_0{}^\intercal\mathbf{v}_0}\mathbf{v}_0}{\|\mathbf{v}_1 - \frac{\mathbf{v}_0{}^\intercal\mathbf{v}_1}{\mathbf{v}_0{}^\intercal\mathbf{v}_0}\mathbf{v}_0\|^2}\frac{-\mathbf{v}_0 - \frac{\mathbf{v}_1{}^\intercal(-\mathbf{v}_0)}{\mathbf{v}_1{}^\intercal\mathbf{v}_1}\mathbf{v}_1}{\|\mathbf{v}_0 - \frac{\mathbf{v}_1{}^\intercal\mathbf{v}_0}{\mathbf{v}_1{}^\intercal\mathbf{v}_1}\mathbf{v}_1\|^2} \tag{3.4.21}$$

$$= \frac{\mathbf{v}_1 + \mathbf{v}_0\|\mathbf{v}_1\|\cos\theta_2/\|\mathbf{v}_0\|}{\|\mathbf{v}_1\|^2\sin^2\theta_2}\frac{-\mathbf{v}_0 - \mathbf{v}_1\|\mathbf{v}_0\|\cos\theta_2/\|\mathbf{v}_1\|}{\|\mathbf{v}_0\|^2\sin^2\theta_2} \tag{3.4.22}$$

$$= \frac{\mathbf{v}_0{}^\intercal\mathbf{v}_1(1-\cos^2\theta_2)}{\|\mathbf{v}_0\|^2\|\mathbf{v}_1\|^2\sin^4\theta_2} \tag{3.4.23}$$

$$= \frac{\mathbf{v}_0{}^\intercal\mathbf{v}_1}{\|\mathbf{v}_0\|^2\|\mathbf{v}_1\|^2\sin^2\theta_2} \tag{3.4.24}$$

$$= \frac{\mathbf{v}_0{}^\intercal\mathbf{v}_1}{4|T|^2} \tag{3.4.25}$$

By symmetry, $\nabla\phi_0{}^\intercal\nabla\phi_2 = \frac{\mathbf{v}_0{}^\intercal\mathbf{v}_2}{4|T|^2}$ and $\nabla\phi_1{}^\intercal\nabla\phi_2 = \frac{\mathbf{v}_1{}^\intercal\mathbf{v}_2}{4|T|^2}$. As $\sum_{i=1}^3\nabla\phi_i$ is equal to zero, then so must $\nabla\phi_j{}^\intercal\sum_{i=1}^3\nabla\phi_i$ equal zero therefore $\mathbf{v}_i{}^\intercal\mathbf{v}_i = \frac{\|\mathbf{v}_i\|^2}{4|T|}$. Hence the contribution of the triangle T to the matrix given by 3.4.4 is,

$$[\mathbf{G}_{i,j}(T)]_{i,j=0,1,2} = \frac{1}{4|T|}\begin{bmatrix}\|\mathbf{v}_0\|^2 & \mathbf{v}_0{}^\intercal\mathbf{v}_1 & \mathbf{v}_0{}^\intercal\mathbf{v}_2\\ \mathbf{v}_1{}^\intercal\mathbf{v}_0 & \|\mathbf{v}_1\|^2 & \mathbf{v}_1{}^\intercal\mathbf{v}_2\\ \mathbf{v}_2{}^\intercal\mathbf{v}_0 & \mathbf{v}_2{}^\intercal\mathbf{v}_1 & \|\mathbf{v}_2\|^2\end{bmatrix} \tag{3.4.26}$$

3.4.5. Boundary

The last component to be evaluated are the contributions from the boundary integrals of the form,

$$\int_{\partial(\mathcal{T}_i\cap\mathcal{T}_j)\cap\partial\omega}\phi_i(\omega)\nabla\phi_j(\omega)^\intercal\mathbf{n}(\omega)dS(\omega) \tag{3.4.27}$$

$\mathbf{n}(\omega)$ is the outward pointing normal vector with respect to the boundary. For triangles contained in the common set $\mathcal{T}_i\cap\mathcal{T}_j$, these are boundary (line) integrals over the edges which are in common with the boundary of the domain of interest $\partial\omega$. For a given triangle,

it is possible for any number of its edges to belong to the boundary and depending on which edge of the triangle is the boundary and the (node) labeling of the function, the evaluation will vary. As a starting point consider the evaluation for ϕ_0 and ϕ_1 over the edge E_1. Recall that E_1 is the edge connecting points $\mathbf{s}_2$ to $\mathbf{s}_0$. As E_1 is a line, then the normal vector relative to this boundary is constant in ω, thus the integral can be simplified in this setting as follows,

$$\int_{E_1} \phi_0(\omega) \nabla \phi_1(\omega)^\intercal \mathbf{n}(\omega) dS(\omega) = \nabla \phi_1{}^\intercal \mathbf{n} \int_{E_1} \phi_0(\omega) dS(\omega) \tag{3.4.28}$$

As the boundary is E_1, the orthogonal direction to $\mathbf{v}_1$ is the same direction as $\nabla\phi_1$ in the plane formed by $\{\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2$. Because it is outward with respect to the boundary, the correct vector is $-\nabla\phi_1$ or pointing away from $\mathbf{s}_0$ and towards E_1. After normalizing $-\nabla\phi_1$ to be unitary, $\mathbf{n}$ is $-\nabla\phi_1 \|\mathbf{v}_0\| \sin(\theta_2)$ (where θ_2 is the angle between $-\mathbf{v}_0$ and $\mathbf{v}_1$). To evaluate the integral parameterize the line integral using $\mathbf{v}_1 t + \mathbf{s}_2$ for $t \in [0,1]$ and using the change of variables $d\omega = \|\mathbf{v}_1\| dt$. Using the fact that $\langle \nabla\phi_0, \mathbf{v}_1 \rangle$ is equal to one the integal is evaluated as,

$$\begin{aligned}
\int_{E_1} \phi_0(\omega) dS(\omega) &= \int_{\omega \in E_1} \phi_0(\omega) d\omega && (3.4.29)\\
&= \int_{\omega \in E_1} 1 - \langle \nabla\phi_0, \mathbf{s}_0 \rangle + \langle \nabla\phi_0, \omega \rangle d\omega && (3.4.30)\\
&= \int_{t \in [0,1]} \Big[1 - \langle \nabla\phi_0, \mathbf{s}_0 \rangle + \langle \nabla\phi_0, \mathbf{s}_2 + \mathbf{v}_1 t \rangle \Big] \|\mathbf{v}_1\| dt && (3.4.31)\\
&= \int_{t \in [0,1]} \Big[1 - \langle \nabla\phi_0, \mathbf{v}_1 \rangle (1-t) \Big] \|\mathbf{v}_1\| dt = \frac{\|\mathbf{v}_1\|}{2} && (3.4.32)
\end{aligned}$$

From equation 3.4.25,

$$\nabla\phi_1{}^\intercal \mathbf{n} \int_{E_1} \phi_0(\omega) dS(\omega) = -\langle \nabla\phi_0, \nabla\phi_1 \rangle \frac{1}{2} \|\mathbf{v}_0\| \|\mathbf{v}_1\| \sin(\theta_2) \tag{3.4.33}$$

$$= -\langle \nabla\phi_0, \nabla\phi_1 \rangle |T| \tag{3.4.34}$$

$$= \frac{-\mathbf{v}_0{}^\intercal \mathbf{v}_1}{4|T|} \tag{3.4.35}$$

By symmetry then the contribution from E_1 of a triangle for the three basis functions supported on the triangle is for the integral in equation 3.4.27 is,

$$[\mathbf{B}_{i,j}(E_1)]_{i,j=0,1,2} = -\frac{1}{4|T|} \begin{bmatrix} \mathbf{v}_0{}^\intercal \mathbf{v}_1 & \|\mathbf{v}_1\|^2 & \mathbf{v}_2{}^\intercal \mathbf{v}_1 \\ 0 & 0 & 0 \\ \mathbf{v}_0{}^\intercal \mathbf{v}_1 & \|\mathbf{v}_1\|^2 & \mathbf{v}_2{}^\intercal \mathbf{v}_1 \end{bmatrix} \tag{3.4.36}$$

The middle row follows from the fact that ϕ_1 is equal to zero on the line E_1 and the middle column follows $B_{i,1} = -B_{i,0} - B_{i,2}$ as $\sum_i \nabla\phi_i$ is equal to zero. Further, the contributions when the edges E_0 and E_2 belong to the boundary are as follows,

$$[\mathbf{B}_{i,j}(E_0)]_{i,j=0,1,2} = -\frac{1}{4|T|} \begin{bmatrix} 0 & 0 & 0 \\ \|\mathbf{v}_0\|^2 & \mathbf{v}_1{}^\intercal \mathbf{v}_0 & \mathbf{v}_2{}^\intercal \mathbf{v}_0 \\ \|\mathbf{v}_0\|^2 & \mathbf{v}_1{}^\intercal \mathbf{v}_0 & \mathbf{v}_2{}^\intercal \mathbf{v}_0 \end{bmatrix} \tag{3.4.37}$$

$$[\mathbf{B}_{i,j}(E_2)]_{i,j=0,1,2} = -\frac{1}{4|T|} \begin{bmatrix} \mathbf{v}_0{}^\intercal \mathbf{v}_2 & \mathbf{v}_1{}^\intercal \mathbf{v}_2 & \|\mathbf{v}_2\|^2 \\ \mathbf{v}_0{}^\intercal \mathbf{v}_2 & \mathbf{v}_1{}^\intercal \mathbf{v}_2 & \|\mathbf{v}_2\|^2 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.4.38}$$

To compute the entry $\mathbf{B}_{i,j}$ examine all the triangles in the common support $T_i \cap T_j$ who have edges in common with the boundary. For each edge type on a triangle labeling, the contribution is computed as above and summed over edges.

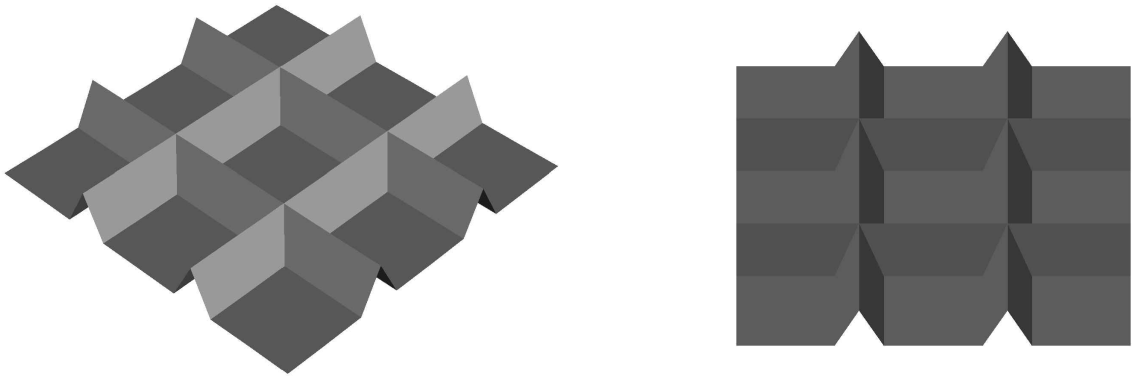

Figure 12: Example four intersecting tent deformations, representing four streets, on a regular grid of $[0, 1] \times [0, 1]$.

3.5. Methodology

The most basic approach for representing street barriers on a mesh is to model city blocks of households as polygons and to insert internal nodes in between these blocks along street paths. Initially, all of these inserted internal nodes are in two-dimensions. However, we can parameterize the altitude of these nodes which controls the height of these tent-like deformations. Larger values of h are more extreme deformations which we hope to represent larger barriers to the spatial distribution.

Proposition 3.5.1. *For triangles in a triangulated mesh in $\mathbb{R}^2$, reflecting the altitude coordinates of the nodes across the origin does not alter the entries of the precision matrix of* $\mathbf{Q}$*, constructed using the B-spline basis.*

In other words, the sign of h has no effect, nor can it be identified. The reason is obvious, from the construction of $\mathbf{Q}$, changing the sign of the altitude in the triangle preserves the area of the triangle and all the angles, which is all that is used to construct $\mathbf{Q}$.

3.5.1. Inference

The interest in this work is performing inference on the parameter h. For a single parameter h this problem can be implemented using standard Bayesian techniques so long as the likelihood of h, f_h, is not computationally expensive. After a certain magnitude, it is not

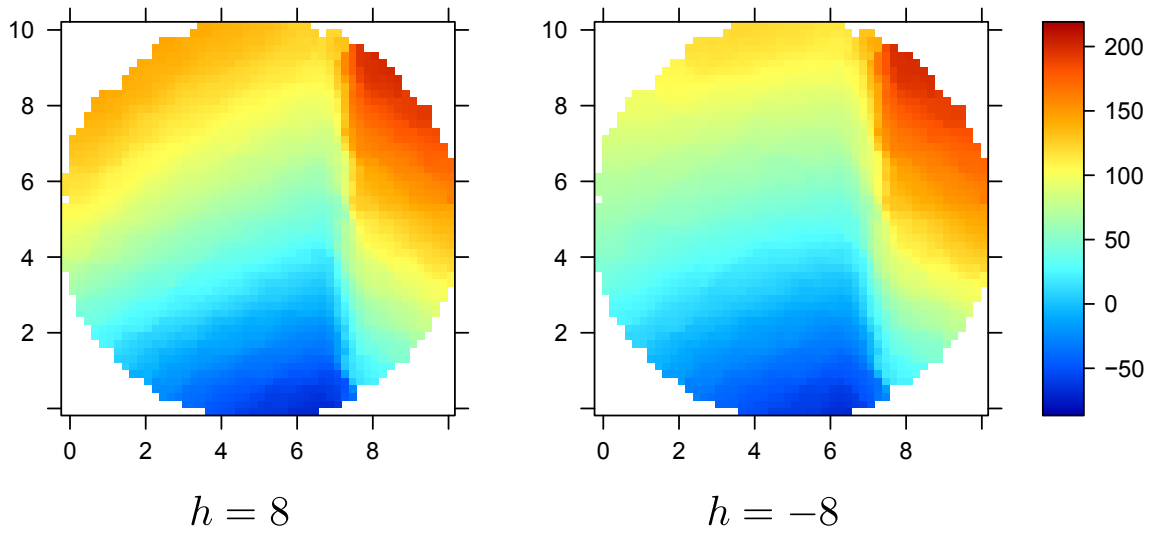

Figure 13: For $\alpha = 2$, continuous fields simulated using identical deformation magnitudes but opposite signs. Single vertical fold beginning at $x = 6.5$ and ending at $x = 7.5$, the raised nodes are the ones along $x = 7$.

clear how well h may be identified. For this reason a good starting point for priors on h, is the exponential distribution, $f(h) \propto \exp\{-\lambda h\}$ for a positive hyper-parameter λ, where the mass of the distribution declines rapidly.

For the continuous observation process of corruption by unit white noise, the process is laid out as,

$$h \sim f_h \tag{3.5.1}$$

$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}(h)) \tag{3.5.2}$$

$$\mathbf{x} = \mathbf{\Phi u} \tag{3.5.3}$$

$$\mathbf{y} \sim \mathrm{N}(\mathbf{x}, \mathbf{I}) \tag{3.5.4}$$

$\Phi_{i,j} \triangleq \phi_j(s_i)$ is the evaluation of the j-th basis function on the location of the i-th observation s_i. This matrix $\mathbf{\Phi}$ brings up a brief technical aside. In this application, $\{s_j\}_{j=1}^m \in \mathbb{R}^2$

and for our simple deformation model, none of the sampling locations lie in triangles on the deformations (ie, we have no observations that lie in between city blocks). However, in simulations similar to the one in figure 13, there are a number of sampling points located on the deformed triangles. In these cases, after deformation, the original sampling location is no longer a point in the mesh and it is necessary to alter the sampling location as well. There are two immediate principled approaches that use the plane defined by the three triangle nodes after deformation: first, projecting the sampling location into the altered plane or second using the equation of the altered plane and the first two coordinates of the sampling location to obtain the altitude of the location. For very large deformations (meshes with large values of h), relative to the area of the triangle, the projection approach produces undesirable results. To produce the simulations in figure 13, we took the second approach.

3.5.2. Metropolis Sampler

We now describe a standard Metropolis-Hastings algorithm for inference on h, whose input is a prior on h and a noisy observation $\mathbf{y}$. In this implementation, new values of h are proposed on the log-scale to enforce positivity and τ_h is used to tune the acceptance rate. The proposal for sampling a new set of coefficients $\mathbf{u}$ is similar to a Gibbs step, using the conditional distribution given the other parameters in the model. We found that given the strong auto-correlation between h and $\mathbf{u}$ that it is best to jointly accept and reject all the parameters at once. The performance of this sampler on data generated using the continuous observation process when $h = 8$ under the Mátern model with $\alpha = 2$ and $\kappa = 1e - 8$ is shown in figure 13. The posterior mean of h using an exponential prior for h with $\lambda = 1$ is reasonably close to the generating value of h.

Metropolis-Hastings Sampler

1. For $i = 1, \ldots, N$ for a fixed N

2. Draw $[h_{i+1}, \mathbf{u}_{i+1}]^\intercal$ from a proposal distribution

2.1. $\log h_{i+1} \sim \mathrm{N}(\log h_i, \tau_h)$

2.2. Assemble $\mathbf{Q}_{i+1} \triangleq \mathbf{Q}(h_{i+1})$

2.3. ‡ : $\mathbf{u}_{i+1} \sim f(\mathbf{u}_{i+1}|\mathbf{Q}_{i+1}, \mathbf{y}) = \mathbf{N}((\mathbf{Q}_{i+1} + \mathbf{\Phi}^\intercal\mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal\mathbf{y}, \mathbf{Q}_{i+1} + \mathbf{\Phi}^\intercal\mathbf{\Phi})$

3. Calculate the transition probabilities

3.1. $q_{i+1} \triangleq q(\mathbf{u}_{i+1}|\mathbf{Q}_{i+1})q(h_{i+1}|h_i)$

3.2. $q_i \triangleq q(\mathbf{u}_i|\mathbf{Q}_i)q(h_i|h_{i+1})$

4. Calculate the model likelihood $p_{i+1} \triangleq p(\mathbf{y}|\mathbf{u}_{i+1})p(\mathbf{u}_{i+1}|\mathbf{Q}_{i+1})p(h_{i+1})$

5. Accept $[h_{i+1}, \mathbf{u}_{i+1}]^\intercal$ with probability $\min\left(\frac{p_{i+1}q_i}{p_iq_{i+1}}, 1\right)$

‡: if the sampling locations lie on modified triangles it becomes necessary to introduce the deterministic dependency between $\mathbf{\Phi}$ and h, for example $\mathbf{\Phi}_{i+1} = \mathbf{\Phi}(h_{i+1})$.

3.5.3. Laplace Approximation

For the typical applications in this subject, where the data is usually discrete (number of insects found) or binary (infestation status), a link function such as the probit or logit links are introduced to produce the observed data. We focus on the probit link for the binary infestation status application. For the probit link on simulated data, we found that the estimates for h were severely biased compared to the generating values of h.

In contrast, the authors of Lindgren et al. (2011) implemented their work using the integrated nested Laplace approximation (INLA) as researched by one of the authors in Rue et al. (2009). The necessity of nested Laplace approximations is necessitated by the need to produce marginals of $\{u_i|\mathbf{y}\}_{i=1}^n$ where $\mathbf{y}$ is the observed data and coefficients for the weak solution. However, to approximate the marginal $h|\mathbf{y}$, it is sufficient to use a single Laplace approximation as in Tierney and Kadane (1986). The model here is similar to (3.5.1) except for the observation process $\mathrm{Probit}(y_i) = \sum_{j=1}^n u_j\phi_j(s_i)$. The Laplace approximation is

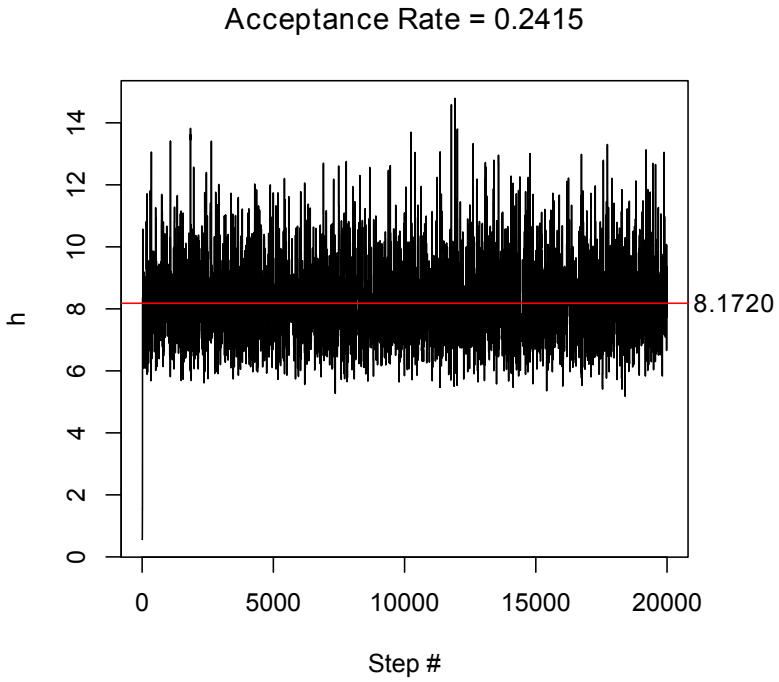

Figure 14: Trace plot for h for data generated using the Mátern-like model ($\alpha = 2$ and $\kappa = 1e-8$ and $h = 8$) and the continuous observation process. The posterior draws of h, using the prior: $\mathbb{P}(h) \propto \exp(-h)$, are well centered around the generating value of h.

given in (3.5.7),

$$
\begin{aligned}
p(h|\mathbf{y}) &\propto p(\mathbf{y}, h) && (3.5.5)\\
&= \int_{\mathbf{u}} p(\mathbf{y}|\mathbf{u}) p(\mathbf{u}|h) p(h) d\mathbf{u} && (3.5.6)\\
&\approx p(h) p(\mathbf{y}, \mathbf{u}^\star|h) \int_{\mathbf{u}} \exp\left\{ -\frac{1}{2} (\mathbf{u} - \mathbf{u}^\star)^\intercal \mathbf{H} (\mathbf{u} - \mathbf{u}^\star) \right\} d\mathbf{u} && (3.5.7)\\
&= \frac{p(h) p(\mathbf{y}, \mathbf{u}^\star|h)}{\sqrt{\det(\mathbf{H}/2\pi)}} && (3.5.8)
\end{aligned}
$$

where $\mathbf{u}^\star$ is the mode or $\mathbf{u}^\star = \arg\max p(\mathbf{y}, \mathbf{u}|h)$ and the dependence of $\mathbf{u}^\star$ on h is clear and $H_{ij} = -\left[\frac{\partial^2}{\partial u_i \partial u_j} \log p(\mathbf{y}, \mathbf{u}|h) \right]_{\mathbf{u}=\mathbf{u}^\star}$.

Laplace Approximation to $p(h|\mathbf{y})$

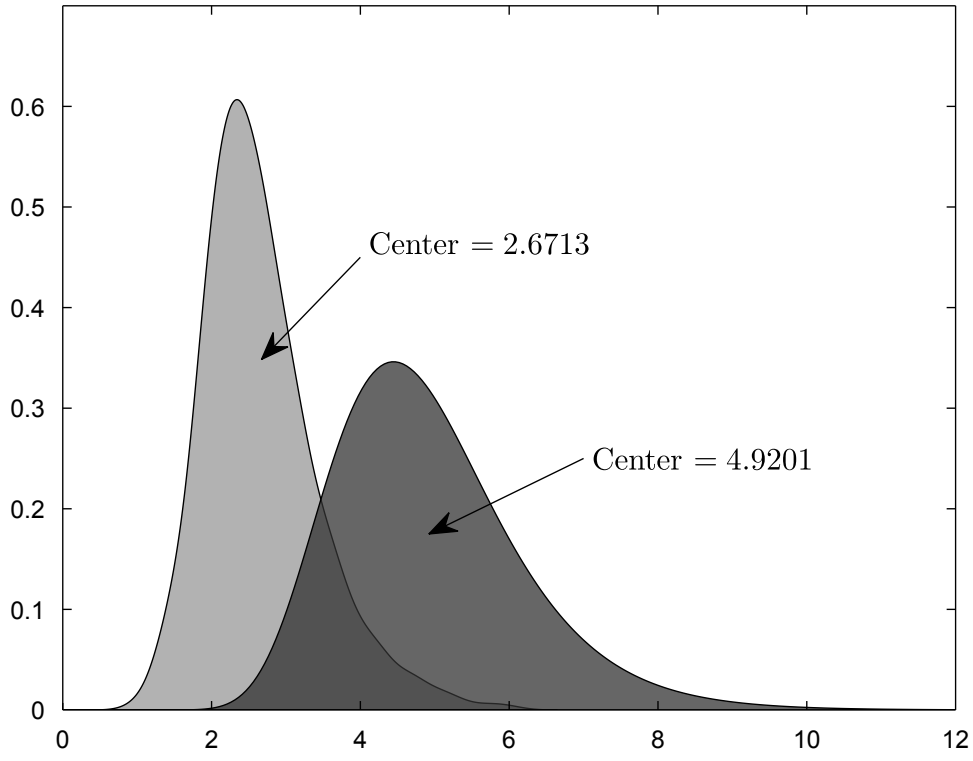

Figure 15: For data generated with a single fold (as in figure 13) and $h = 5$, a comparison of the Laplace approximation to the density $\mathbb{P}(h|\mathbf{y})$ compared to the empirical histogram of the posterior samples of h drawn by the MCMC, demonstrating the bias in Metropolis sampler, for binary response data.

1. Initialize a sequence of $\{h_i\}_{i=1}^N$, for each value h_i

2. Find the maxmimum of $\log p(\mathbf{y}, \mathbf{u}|h_i)$ at the mode $\mathbf{u}_i^\star \triangleq \mathbf{u}^\star(h_i)$ using a Newton or quasi-Newton method

3. Calculate for $\mathbf{H}_i \triangleq \mathbf{H}(\mathbf{u}_i^\star)$, $\det \mathbf{H}_i$

4. After enumeration, fit a spline through the approximate values of the density and integrate to find the normalizing constant

The size of the discretization for h dictates the run time, where for the last step the density must decay reasonably quickly with respect to the limits of the discretization to obtain a

reasonably correct normalization. For certain observation models, the Newton step may require modification for semi-definiteness of the Hessian matrix, see Nocedal and Wright (1999) for modified schemes. Given the long mixing times of the Metropolis sampler and high auto-correlation between samples, the Laplace approximation provides an attractive deterministic alternative. However, to obtain posteriors for $\mathbf{u}$ requires the implementation of the full integrated nested Laplace approximation. For the purposes of this work, the center calculated from the approximate density appeared to be much less biased than the Monte Carlo estimate.

3.6. Results

In order to analyze the Mariano Melgar data, we constructed a mesh using the household locations in the survey along with the corner nodes of the block polygons. We added additional nodes along the centers of the six widest streets, which run in between the city block polygons and constructed a Delaunay triangulation, taking the block polygons as constraints.

Our model to analyze the effect of streets in Mariano Melgar is the following,

$$\log\left(\frac{\mathbb{P}(y_i = 1|x_i, t)}{1 - \mathbb{P}(y_i = 1|x_i, t)}\right) = \sum_{j=1}^{n} u_j \phi_j(s_i) + t \qquad \mathbf{u} \sim \text{Mátern}(\alpha, \kappa, h) \tag{3.6.1}$$

where s_i is the sampling location of observation i, α is the smoothness parameter which is fixed at 1 (the roughest possible), and κ is the range parameter. The parameter of interest, h, is the height parameter of the tent deformations representing the streets. For the Mariano Melgar data, we used a uniform prior between 0 and 500 on h, reasoning that an effective value of h higher than 500 would be unreasonably high, given that distances in the UTM coordinate system are based in meters. We used a standard diffuse $\text{N}(0, 1e-8)$ prior on the intercept t and selected a $\kappa = 0.004$, based on the likelihood values at the mode calculated

Figure 16: Constraint Delaunay triangulation of Mariano Melgar, Arequipa, Peru containing 17,674 nodes and 35,275 triangles, taking polygons, representing the city blocks, as constraints.

using the Laplace approximation ($\mathbf{u}^\star$ from equation (3.5.8)). Visually, the maximizer for the spatial effect, calculated using this value of κ, is appropriately smooth as shown in figure 19.

One point of interest was the sensitivity of our results to the choice of the range parameter κ on the posterior of the deformation parameter given the data. Based on the maximizing likelihoods calculated in the process of computing the Laplace approximation (section 3.5.3) as well as visual checks of the maximizing value of the latent spatial parameter, $\mathbf{u}$, such as in figure 19, we narrowed the reasonable range for κ down to $[0.001, 0.010]$. To evaluate

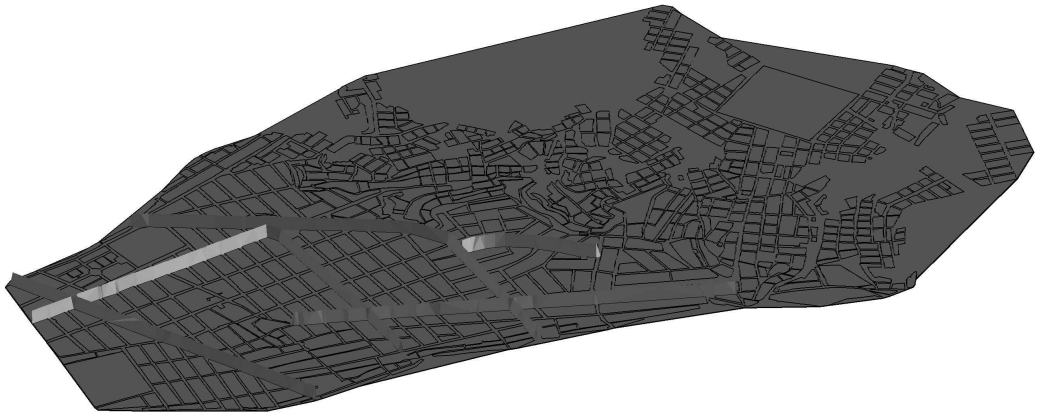

Figure 17: For a deformation parameter of $h = 53$, these are the six widest roadways of Mariano Melgar, Arequipa, Peru of interest modeled using the tent deformation.

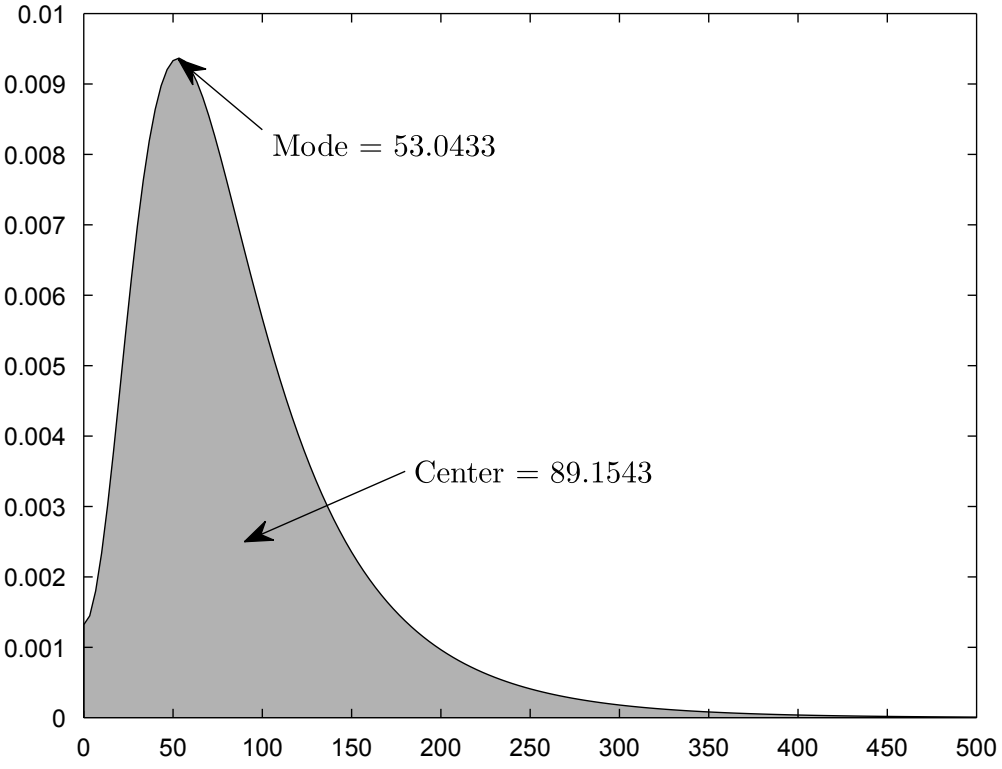

Figure 18: Approximate posterior distribution, $\mathbb{P}(h|\mathbf{y})$, using the Laplace approximation for the Mátern model with $\alpha = 1$ and $\kappa = 0.004$ on the survey data.

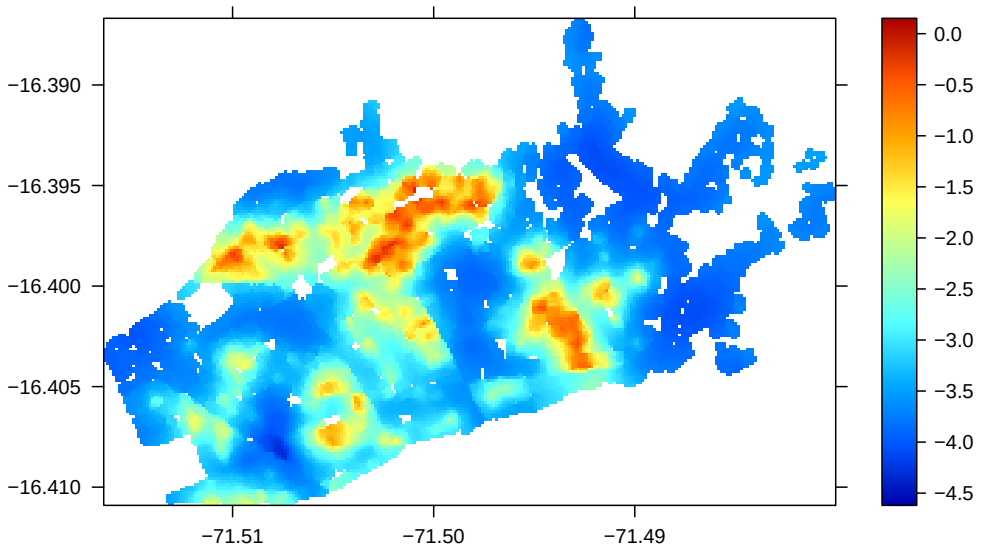

Figure 19: Plot of the maximizing field of the model likelihood induced by $h = 53$, which was the estimated posterior mode from the data.

the sensitivity of our findings to κ, we found the approximate posterior distribution for κ varying on a grid from 0.001 to 0.010 by increments of 0.001 and noted the posterior mode and means of these distributions. The resulting modes and centers are shown in table 4 Although larger range parameters decrease the posterior mode of h, the posterior mean of h is consistently higher than 70 meters, which indicates that these are major barriers to the distribution of the insect. As these are the six widest streets or more appropriately boulevards, it confirms the belief that insect infestations should have difficulty crossing these barriers.

Sensitivity Analysis on Approximate Posterior to Choice of Range Parameter κ

κ	Mode	Center
0.001	113.5499	163.9871

Continued on next page

Table 4 – continued from previous page

κ	Mode	Center
0.002	88.1494	130.1704
0.003	67.8249	105.5682
0.004	53.0433	89.1542
0.005	42.2523	78.7410
0.006	34.2092	72.7253
0.007	28.0824	70.0058
0.008	23.3328	69.8496
0.009	19.5946	71.7986
0.010	16.6102	75.5831

Table 4: Mode and means of the posterior distribution of the deformation parameter $\mathbb{P}(h|\mathbf{y})$ on the Mariano Melgar survey due to varying κ, the range parameter in the Mátern model.

3.7. Discussion

We found by modeling the six widest streets of Mariano Melgar that these streets are major barriers to the distribution of infestation across the region. We are confident that its posterior mean of this height is above seventy meters, regardless of the choice of κ. Further across all choices of parameters, the model indicates that the posterior distribution of the height is well away from zero. We feel that there is strong evidence to support the conclusion that streets do act as a physical barrier to how insects are distributed in an urban environment.

The methodology that we propose can be well-integrated into existing Bayesian approaches to data analysis. While we propose a simple model for these street barriers, this approach can be expanded to include other classes of shapes for these barriers. These results indi-

cate that streets are an important urban barrier, but it is also raises additional questions regarding the shapes of these barriers, which our analysis does not yet address. If streets are the primary urban boundaries, these results suggest that sampling and spraying should be based around areas cordoned off by major roadways. We hope that these results can be used to aid the design of sampling and insecticide application for public health campaigns.

CHAPTER 4 : Conclusion

In this thesis, we have investigated two issues in epidemiological studies, observation error and spatial boundaries, using Bayesian hierarchical models. In order to approach these issues, we have adapted and proposed new models based on using Gaussian Markov random fields to capture the spatial variation present in the data. The motivation behind using these models primarily is not only their flexibility but also their computational tractability for large data sets. We have shown that these models fit well on simulated data as well as actual data collected by the Peruvian Ministry of Health, and that the frameworks created by these models answer interesting and often proposed questions by epidemiologists.

4.1. Summary

In chapter 2, we proposed a Bayesian hierarchical model that models the heterogeneity in observation error from human inspectors. Because the the insect infestation is spatially-distributed, the fact that inspectors themselves are aggregated in space created an issue of identifiability. We concluded that post-observation the issue of identifiability can not be remedied. However, we demonstrated through simulation that the correlation between inspector sensitivity and infestation intensity through space drove the estimation error. We showed that the greater the degree of this correlation, the greater the amount of error using the Bayesian posterior estimates. We demonstrated how randomized household assignment of inspectors produces better estimates of infestation probability per household. Because of the impracticality of this assignment for large regions of space, we showed how randomization over a smaller positive region can be used to learn the relative accuracies of various inspectors and reduce error over the entire region.

In chapter 3, we introduced a novel approach for spatial models on curved surfaces, where physical boundaries in space are modeled using these curved deformations. Because most spatial modeling is done on a flat domain, this function approximation approach using stochastic partial differential equations circumvents a number of difficulties. We parame-

terized these deformations and took the Bayesian approach to parameter estimation. We showed on simulated data that existing techniques such as Monte Carlo and Laplace approximation that inference is well-behaved. As an application of our methodology, we demonstrated on the Mariano Melgar survey data that streets are a major geographic barrier to the distribution of the *Triatoma infestans* insect.

4.2. Extensions & Future Work

One open extension is to analyze repeated measurements over time using the following stochastic partial differential equation,

$$\frac{\partial x(\omega,t)}{\partial t} + (\kappa^2 - \Delta)x(\omega,t) = W(\omega,t) \tag{4.2.1}$$

where $W(\omega,t)$ is stochastically-white in space and time noise. Representing the above equation as, $dx + (\kappa^2 - \Delta)xdt = Wdt$ and using the approximation $x = \sum_{i=1}^{n} u_i(t)\phi$, then the quantities $\langle dx, \phi_j\rangle$ and $\langle(\kappa^2 - \Delta)xdt, \phi_j\rangle$ have the following representations:

$$\langle dx, \phi_j\rangle = \sum_{i=1}^{n} \langle \phi_i, \phi_j\rangle du_i(t) \tag{4.2.2}$$

$$\langle(\kappa^2 - \Delta)xdt, \phi_j\rangle = \sum_{i=1}^{n} \langle(\kappa^2 - \Delta)\phi_i, \phi_j\rangle u_i(t)dt \tag{4.2.3}$$

If $\mathbf{C}_{ij} \triangleq \langle \phi_i, \phi_j\rangle$ and $\mathbf{H}_{ij} \triangleq \langle(\kappa^2 - \Delta)\phi_i, \phi_j\rangle$, then the weak solution to the above equation may be represented as,

$$\mathbf{C}du_t + \mathbf{H}u_t dt = \epsilon_t \tag{4.2.4}$$

where $\epsilon_t \sim \mathrm{N}(0, ,\mathbf{C}dt)$, then using a foward Euler scheme with the increment $\Delta t_i = t_{i+1} - t_i$,

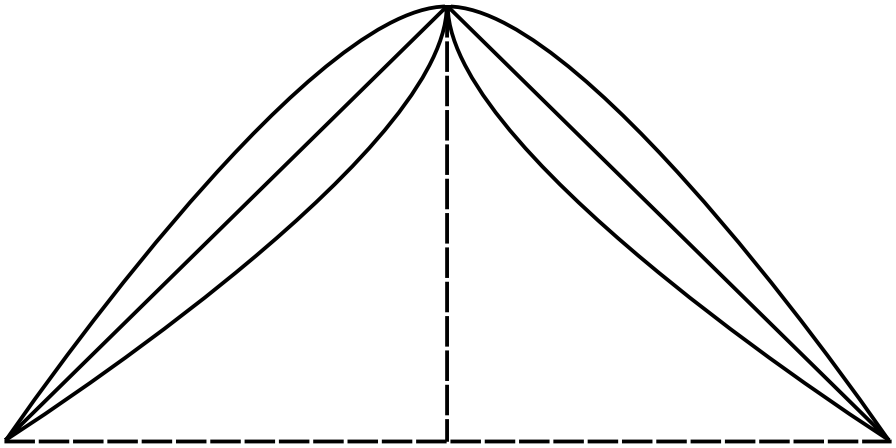

Figure 20: After adding additional nodes along boundaries, additional parameters can be added to select for even more complex boundary deformations.

the system may be written down as,

$$\mathbf{C}(\mathbf{u}_{t_{i+1}} - \mathbf{u}_{t_i}) + \Delta t_i \mathbf{H}\mathbf{u}_{t_i} = \boldsymbol{\epsilon}_{t_i} \tag{4.2.5}$$

where $\boldsymbol{\epsilon}_{t_i} \sim \mathrm{N}(0, , \Delta t_i \mathbf{C})$. The evolution of u_t during each time step is as follows,

$$\mathbf{u}_{t_{i+1}} = (\mathbf{I} - \Delta_t \mathbf{C}^{-1}\mathbf{H})\mathbf{u}_{t_i} + \mathbf{w}_{t_i} \tag{4.2.6}$$

where $\mathbf{w}_{t_i} \sim \mathrm{N}(\mathbf{0}, \Delta t_i \mathbf{C}^{-1})$ with similar boundary equations to simplify $\mathbf{H}$, this evolution equation can be integrated straightforwardly into existing methods for state space estimation.

Another extension of interest to increase the complexity of the deformation shapes. Currently we use a 3-point, tent to represent a street, it is certainly possible to add additional points along the streets so that the deformations are depicted as Bézier curves. Depending on the location and height of these knots, it's possible to represent the thickness of the streets. Adding additional parameters to represent these features, it is possible to determine in greater detail the shape of the deformation.

APPENDIX

A.1. Correlated Inspector Distribution Simulation Details

A.1.1. Region Division

In order to divide the Mariano Melgar data into regions, the simple approach taken here was divide the data along the vertical axis in half, and across the horizontal axis in thirds. To minimize the difference in the number of households between the most and least populated regions the coordinate system was first rotated.

Although the Mariano Melgar data set contains forty unique inspectors, after examining the distribution of the number of total households labeled by each inspector, it was found that a small handful inspected less than 50 households in total. Further, the average number of households inspected by each inspector was in the neighborhood of 200 houses.

To determine the number of inspectors held in each region, the region was divided into sub-blocks of around 200 houses which for Mariano Melgar meant that the regions respectively held: 10, 13, 7, 12, 12, and 7 inspectors. Excluding the largest region, the second, this amounts to 48 sub-blocks to be assigned. In following the data, 32 unique inspectors were chosen such that a subset of 16 would be assigned to two regions and the rest only one.

A.1.2. Presence-absence Simulation

To create, heterogeneity among the regions in terms of vector presence, six separate GMRF were used to simulate the presence-absence data $\mathbf{y}$. In practice, the precision parameter was kept constant and only the intercept was varied to induce different levels of intensity.

In general, the intensity levels were chosen to match the inspector sensitivities such that rate of infestation $\times$ inspector accuracy remained constant over regions 1, 4, and 5. Region 2 was selected as the reference region, as the level of infestation was high. Due to the large amounts of data missing in regions 3 and 6, very low infestation intensity were specified. The

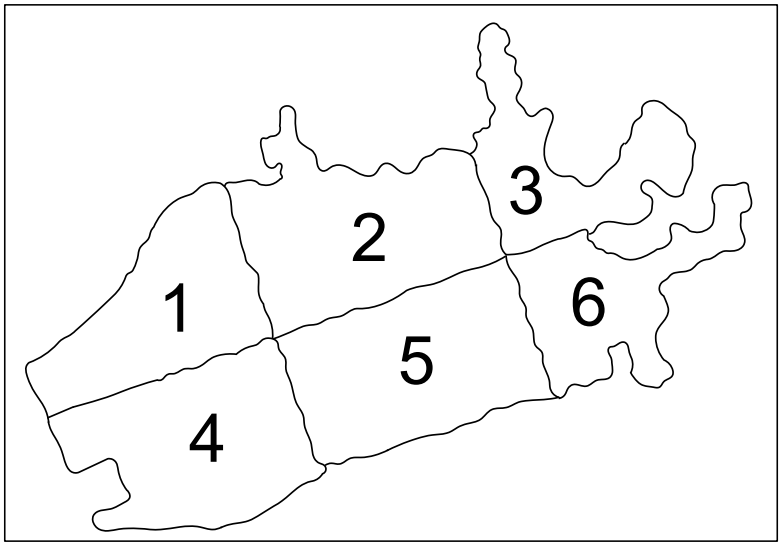

Figure 21: Division of Mariano Melgar into six regions for simulated examples.

reason being firstly, we require areas of low infestation to assign highly accurate inspectors. More importantly in practice, mixing for this region is very slow due to the amount of missing data - strong dependence on a good starting point for the sampling is required in these regions.

A.2. Inspector Assignments and Household Labeling

Once inspector accuracies and presence-absence is simulated, inspectors are drawn according to the discrete distribution given in equation 2.5.1. The partition function is taken over all configurations such that a unique subset of half the inspectors is assigned to different regions. This last requirement is made so that a randomly selected inspector chosen to appear twice in the assignment is not assigned to identical regions. The intention of these multiple assignments is so that some inspectors are distributed in aggregations in multiple regions of the map as in the data.

This process is done in the simulations at first to the exclusion of region 2. For simulations in which the frame of reference is excluded, a random subset of 13 of the total 32 inspectors is drawn and assigned to this region.

Once inspector to region assignments are complete, individual household level assignments are done according to the following,

For each region i, let $\{n_i, I_i, L_i\}_{i=1}^m$ be a tupple consisting of a constant n_i denoting the size of each block in region i, I_i an index set of the inspectors assigned to region i from above, and lastly L_i an index set denoting the individual household locations belonging to region i. Note $|_{L_i}$ denotes the restriction of the matrix to the locations in L_i

Algorithm 1 Sampling Household Labels

procedure SAMPLEH($\{n_i, I_i, L_i\}_{i=1}^m, t, k, H$)
  for all $i \in \{1, \ldots, m\}$ **do**
    for all $k \in \{1, \ldots, |I_i|\}$ **do**
      if $|L_i| > n_i$ **then**
        Sample $\{x_j\}_{j \in L_i} \sim \mathrm{N}(t, k\mathbf{Q}|_{L_i})$
        Sample $\{y_j\}_{j \in L_i} \sim \mathrm{Bern}(p = \Phi(\{x_j\}_{j \in L_i}))| \sum_{j \in L_i} y_j = n_i$
        for all $l \in \{j : y_{j \in L_i} = 1\}$ **do**
          $H_l \leftarrow k$
        end for
        $L_i \leftarrow L_i - \{j : y_{j \in L_i} = 1\}$
      else
        for all $l \in L_i$ **do**
          $H_l \leftarrow k$
        end for
      end if
    end for
  end for
end procedure

Procedures for simulating draws from a conditional Bernoulli distribution are given in Chen and Liu (1997). Simulated data is drawn according to the binomial model using the simulated $\mathbf{y}$ and $\boldsymbol{\beta}$ with the exact pairing of the indices as drawn from above.

A.3. Gibbs Sampler

The Gibbs sampling scheme for the model is provided below. In practice the transfer from the continuous risk to the binary presence-absence outcome is handled through the use of the parameter expansion as in Albert and Chib (1993). An analogous formulation using the logit link can be found in Holmes and Held (2006).

$$\mathbf{y}_0 = \mathbf{u} + t + \boldsymbol{\epsilon} \tag{A.3.1}$$

$$\mathbf{y}_1 = \mathbb{1}_{y_{0,i}>0} \tag{A.3.2}$$

where recalling $\mathbf{u}$ is the centered GMRF $\mathrm{N}(\mathbf{0}, k_{\mathbf{u}}\mathbf{Q})$ with the sum-to-zero constraint and t is the intercept. The if the prior on t is given by $\mathrm{N}(\mu, \tau)$, where τ is the precision, the prior on $k_{\mathbf{u}}$ is given by $\Gamma(k, \theta)$, where k and θ represent the scale and shape parameters, and the prior on $\boldsymbol{\beta}$ is $\mathrm{B}(a, b)$, then the Gibbs sampler is given by,

1. $(k_{\mathbf{u}}|\mathbf{u}) \sim \Gamma\big(\frac{n-1}{2} + k, \frac{1}{2}\mathbf{u}^t\mathbf{Q}\mathbf{u}\big)$

2. $\left(\begin{bmatrix}\mathbf{u}, t\end{bmatrix}^t | k_{\mathbf{u}}, \mathbf{y}_0\right) \sim \mathrm{N}\left(\begin{bmatrix} k_{\mathbf{u}}\mathbf{Q} + \mathbf{I} & \mathbf{1} \\ \mathbf{1}^t & n + \tau \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{1}^t\mathbf{y}_0 + \mu + \tau \end{bmatrix}, \begin{bmatrix} k_{\mathbf{u}}\mathbf{Q} + \mathbf{I} & \mathbf{1} \\ \mathbf{1}^t & n + \tau \end{bmatrix}\right)$

3. $(y_{0,i}|u_i, t, y_{1,i}) \sim \begin{cases} \mathbf{N}(u_i + t, 1|y_{0,1} > 0) & \text{if } y_{1,i} = 1 \\ \mathbf{N}(u_i + t, 1|y_{0,1} < 0) & \text{if } y_{1,i} = 0 \end{cases}$

4. $(y_{1,i}|u_i, \beta_j) \sim \mathrm{Bern}\left(p_i = \begin{cases} \frac{(1-\beta_j)\Phi(u_i+t)}{(1-\beta_j)\Phi(u_i+t)+(1-\Phi(u_i+t))} & \text{if } I_{\mathrm{NA}i} = 0 \\ \Phi(u_i + t) & \text{if } I_{\mathrm{NA}i} = 1 \end{cases}\right)$, where location i is labeled with inspector j

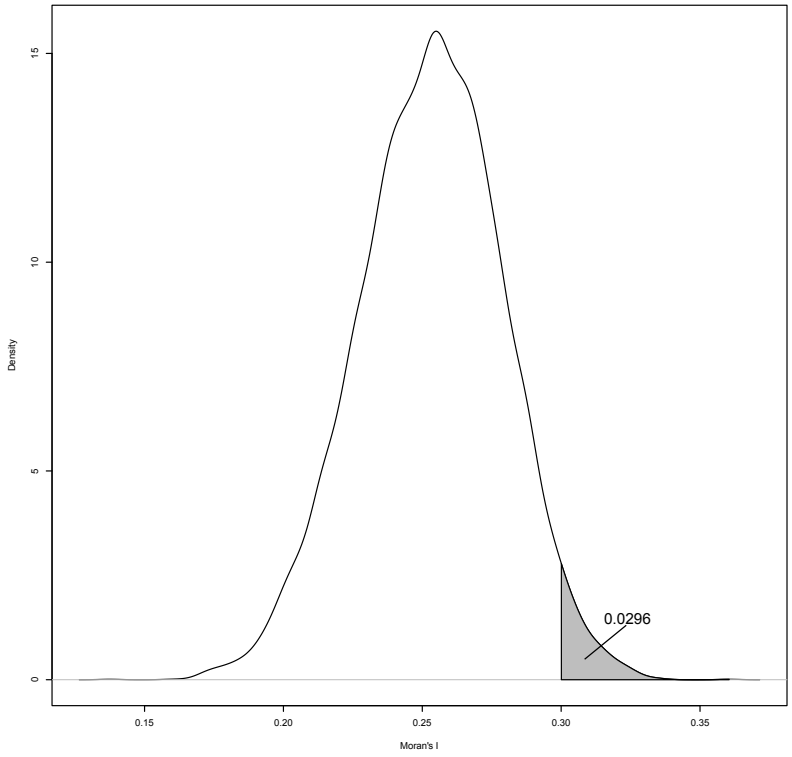

Figure 22: Posterior predictive distribution of the Moran's I statistics calculated from simulated data from posterior draws of $\mathbf{y}$ and $\boldsymbol{\beta}$.

5. $(\beta_i|\mathbf{y}|_{n_i}, \mathbf{z}|_{n_i}) \sim \mathrm{B}(\sum_{j\in n_i} y_j z_j + a, \sum_{j\in n_i} y_j(1-z_j) + b)$, where n_i denotes the households labeled inspector i

A.4. Posterior Predictive Check

A posterior predictive check using the Moran's I statistic was used to address the appropriateness of the binomial inspection error model. Using the sampled values of presence-absence, $\mathbf{y}$ and the inspection sensitivities, $\boldsymbol{\beta}$, 'newly-observed' data sets were sampled according to the binomial inspection model. From figure 22, the Moran's I found in the actual data tends to be on the higher side compared to simulated data from posterior draws. Although the model captures a reasonable large amount of the spatial correlation found in the data, these results suggest some room for improvement for future modeling.

BIBLIOGRAPHY

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, ninth edition, 1972.

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679, 1993.

S. Banerjee, M. M. Wall, and B. P. Carlin. Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, 4:123–142, 2003.

S. Banerjee, B. P. Carlin, , and A. E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall, first edition, 2004.

C. Barbu, A. Hong, J. M. Manne, D. Small, J. E. Calderón, K. Sethuraman, V. Quispe-Machaca, J. Ancca-Juárez, J. G. Cornejo del Carpio, F. S. Chavez, et al. The effects of city streets on an urban disease vector. *PLoS Computational Biology*, page e1002801, 2013.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B*, 36(2):192–236, 1974.

J. Besag, J. York, and A. Mollié. Bayesian image restoration with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, 43(1):1–59, 1991.

H. Boyd, W. Flanders, D. Addiss, and L. Waller. Residual spatial correlation between geographically referenced observations: a bayesian hierarchical modeling approach. *Epidemiology*, 16(4):532–541, 2005.

S. Brenner and L. Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, third edition, 2008.

G. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1–3, 1950.

S. X. Chen and J. Liu. Statistical applications of the poisson-binomial and conditional bernoulli distributions. *Statistica Sinica*, 7:875–892, 1997.

H. Corrasco, A. Torrellas, C. García, M. Segovia, and M. Feliciangeli. Risk of Trypanosoma cruzi I (Kinetoplastida: Trypanosomatidae) transmission by Panstrongylus geniculatus (Hemiptera: Reduviidae) in Caracas (Metropolitan District) and neighboring States, Venezuela. *International Journal for Parasitology*, 35(13):1379–1384, 2005.

N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, 70(1):209–226, 2010.

G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*. Cambridge University Press, first edition, 1992.

M. Fuentes. Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association*, 102(477):321–331, 2007.

R. Furrer and S. Sain. spam: A sparse matrix r package with emphasis on mcmc methods for gaussian markov random fields. *Journal of Statistical Software*, 36(10):1–25, 2010.

A. Gelman. Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–533, 2006.

A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

T. Gneiting. Strictly and non-strictly positive definite functions on spheres. Technical report, Institute of Applied Mathematics, University of Heidelberg, 2012.

Google. Google earth, 2009. URL `\url{http://earth.google.com/}`. Accessed on 17/08/2009.

C. Holmes and L. Held. Bayesian auxiliary variable models for binary and polychotomous regression. *Bayesian Analysis*, 1(1):145–168, 2006.

C. Kaufman, M. Schervish, and D. Nychka. Covariance tapering for likelihood-based estimation in large spatial datasets. *Journal of the American Statistical Association*, 103 (484):1556–1569, 2008.

M. Levy, N. Bowman, V. Kawai, L. Waller, J. Cornejo del Carpio, E. Cordova Benzaquen, R. Gilman, and C. Bern. Periurban Trypanosoma cruzi-infected Triatoma infestans, Arequipa, Peru. *Emerging Infectious Diseases*, 12(9):1345–1352, 2006.

M. Levy, F. Malaga Chavez, J. Cornejo Del Carpio, D. Vilhena, F. McKenzie, and J. Plotkin. Rational spatio-temporal strategies for controlling a Chagas disease vector in urban environments. *Journal of the Royal Society Interface*, 7:1061–1070, 2010.

F. Lindgren and H. Rue. Explicit construction of GMRF approximations to generalised Matérn fields on irregular grids. Technical report, Lund Institute of Technology, Lund University, 2007.

F. Lindgren, H. Rue, and J. Lindström. 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*, 73(4):423–498, 2011.

J. Nocedal and S. Wright. *Numerical Optimization*. Springer-Verlag, first edition, 1999.

C. Paciorek. Computational techniques for spatial logistic regression with large data sets. *Computational Statistics & Data Analysis*, 51(8):3631–3653, 2007.

J. A. Rozanov. Markov random fields and stochastic partial differential equations. *Math. USSR Sb.*, 32(4):515–534, 1977.

H. Rue and L. Held. *Gaussian Markov Random Fields Theory and Applications*. Chapman & Hall, first edition, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B*, 71(2):319–392, 2009.

S. Samko, A. Kilbas, and O. Marichev. *Fractional Integrals and Derivatives: Theory and Applications*. Gordon and Breach Science Publishers S.A., first edition, 1993.

M. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, first edition, 1999.

M. Stein, Z. Chi, and L. Welty. Approximating likelihoods for large spatial datasets. *Journal of the Royal Statistical Society: Series B*, 66(2):275–296, 2004.

M. Teixeira, M. Barreto, M. Costa, L. Ferreira, P. Vasconcelos, and S. Cairncross. Dynamics of dengue virus circulation: a silent epidemic in a complex urban area. *Tropical Medicine & International Health*, 7(9):757–762, 2002.

L. Tierney and J. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

J. Trape, E. Lefebvre-Zante, F. Legros, G. Ndiaye, H. Bouganali, P. Druilhe, and G. Salem. Vector density gradients and the epidemiology of urban malaria in Dakar, Senegal. *American Journal of Tropical Medicine and Hygiene*, 47(2):181–189, 1992.

P. Whittle. Stochastic processes in severl dimensions. *Bulletin of the International Statistical Institute*, 40:974–994, 1963.