Lancaster University
Medical School

CHICAS
combining
health information,
computation and statistics

Bayesian non-parametric models for zoonotic disease applications

by

Poppy Purefoy Miller

Supervisor(s):

Dr. Chris Jewell and Dist. Prof. Peter Diggle

Thesis submitted in partial fulfilment for the degree of

Doctor of Philosophy in Epidemiology and Statistics

in the Faculty of Health and Medicine

Lancaster Medical School

Lancaster - England

April 2020

Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussion with my supervisors, Dr. Chris Jewell and Professor Peter J. Diggle.

Chapter 2 of this thesis has been published in the following academic publication (with an associated `R` package available on CRAN):

Miller, P., Marshall, M., French, N., and Jewell, C., 2017. `sourceR`: Classification and Source Attribution of Infectious Agents among Heterogeneous Populations. *PLoS Computational Biology.*

Chapter 3 of this thesis is in preparation for publication in PLoS Neglected Tropical Diseases.

Poppy P. Miller, BSc(Hons)
Lancaster University, UK
April 2020

Abstract

Advanced statistical models are a key tool in developing interventions to reduce disease incidence, particularly in low resource settings. Infectious diseases often have complex infection processes and pathways, particularly zoonotic diseases which often have direct and indirect routes of infection. This makes epidemiological studies aimed at identifying and/or quantifying risk factors challenging, as they typically include complexities such as multi-level dependency structures, correlated covariates, missing data, and high noise. Often, data are only partially observed due to censoring and structurally missing information, and are often observational rather than the result of direct treatments. This thesis explores novel methods and models designed to tease out pathways and factors that contribute to risk of disease in humans for zoonotic pathogens. Chapter 2 develops a Bayesian non-parametric model to estimate the proportion of cases attributable to known sources of disease, and identify sub-types of pathogens which are unusually dangerous. This model was applied to a campylobacteriosis data set from New Zealand with results showing chicken from a single supplier was likely the source of approximately 70% of cases in the data set, and identified 9 particularly dangerous subtypes. Chapter 3 widens the scope to consider the relative contributions of many potential risk factors for disease (causal or not). Our model considers many environmental and social risk factors for leptospirosis in complex urban environments, including rat exposure. We estimate both rat exposure and leptospirosis risk using a Bayesian non-parametric cut model which correctly accounts for uncertainty in the rat exposure predictions. The results identify groups of high risk individuals, based on socio-economic data and environmental risk factors, that could be targeted using interventions. This chapter highlighted a significant limitation in many epidemiological studies which use inaccurate diagnostic techniques. Chapter 4 develops methodology to address this limitation by modelling within-host immune responses to pathogenic challenge. This was done by integrating a mechanistic ordinary differential equation model with a Bayesian censored noise model. Our model estimates expected changes in antibody levels after challenge with different pathogens, and indicates possible differences in immune response that may be responsible. The model is also able to estimate time of challenge at an individual level. The model is applied to a leptospirosis challenge data set in sheep, and shows significant differences in immune response to serovars Pomona and Hardjobovis. Integration of this methodology with epidemiological studies (like those in Chapters 2 and 3) will allow for more accurate estimation of relative risks and enable more effective intervention strategies to be developed.

Acknowledgements

I'd like to thank my supervisors, Dr. Chris Jewell and Professor Peter Diggle, for their continuous support and encouragement throughout my PhD. I deeply appreciate the time they spent discussing project ideas, directions, methodologies and impacts with me. I particularly appreciate Peter involving me in the EEID program (Ecoepidemiology of Leptospirosis in the Urban Slums of Brazil, Fogarty International Center at the U.S. National Institutes of Health). This project gave me invaluable experience collaborating with a wide range of scientists from across the world and deepened my understanding of global health challenges (and gave me the chance to enjoy amazing food, company and caipirinha's at some of the most gorgeous beaches in the world). I would like to thank Chris for opening my eyes to the joys of programming and tramping in the Lake District/ Yorkshire Dales and for welcoming my partner Cyrus and I into his home to celebrate Christmas. I particularly appreciate all of the late night/ early morning Skype calls after I returned to New Zealand and the constant encouragement and motivation that were crucial during the seemingly never-ending write up period.

I would like to thank Jonathan Marshall, Geoff Jones and Nigel French for their invaluable contributions to the source attribution project and to the Massey University Epi-Centre for providing the dataset. I would also like to thank Jackie Benschop and Julie Collins-Emerson (Massey University) for their insightful comments and suggestions for the leptospirosis titre modelling project, and for providing the data. I am thankful to the EEID team (including Albert Ko, Mike Begon, James Childs, and Mitermayer Reis) for providing the Brazil leptospirosis dataset, a truly supportive vibrant environment during the yearly meetings in Salvador, and for being great company. I am especially grateful to

Kate Hacker who collected the data and provided biological insights which were crucial for model development and interpreting the results. Kate and the other students/post-docs in this project made me feel incredibly welcome in Brazil and I thank them for all the social events they organised when I visited.

I am very grateful for the PhD funding provided by the Health eResearch Centre (HeRC) and to my PhD confirmation panel, Luigi Sedda and Jon Read, for their insightful comments and suggestions. I particularly appreciated Luigi's incredibly positive outlook and constructive suggestions which gave me an enormous motivational boost. I would also like to thank Vanessa Cave for her support during the write up period, and also Kim Miller and Harold Henderson for proof reading my thesis.

I would like to thank the CHICAS group for providing a supportive, positive environment, and plenty of coffee, cake and great conversation. I am especially grateful to Lisa Koeppel, Annie Edwards, Alex Waldron and Rhiannon Edge for being incredibly supportive friends through the peaks and troughs of PhD life and beyond, providing great company, and making my time in the UK so memorable. I'd like to thank Cyrus for always believing in me and supporting me through everything. Finally, I am very thankful to my friends and family for their support and patience.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction to zoonotic disease applications and associated statistical modelling challenges

Infectious diseases are a large source of human morbidity and mortality worldwide, causing 8 million deaths during 2016 (World Heath Organisation, 2018). Zoonotic diseases are infectious diseases of animals that can be naturally transmitted to humans. These diseases, particularly those that are foodborne, are a major source morbidity, mortality and productivity losses in both humans and animal populations. The World Health Organisation (WHO) estimated that there were over 600 million cases of foodborne illness globally in 2010; of these, only 217 thousand were caused by chemicals or toxins rather than infectious organisms (World Health Organization, 2015). Infectious agents that cause diarrhoeal diseases accounted for the vast majority (550 million), in particular *Campylobacter spp.* which caused 96 million cases, and 21 thousand deaths (World Health Organization, 2015).

There is a considerable difference in the burden of zoonotic diseases between low and high-income regions (World Health Organization, 2015; Goarant, 2016) which suggests that a major proportion of this disease burden is preventable. Therefore, identifying the risk factors surrounding infection by zoonotic diseases is imperative to implement effective targeted interventions. This will be especially important in resource-poor settings, where these diseases are most prevalent.

In this thesis, we focus on two globally important zoonotic diseases: campylobacteriosis and leptospirosis. We used advanced statistical techniques to advance understanding of the hidden biological processes and how they interact with the environment to inform development of effective interventions. The rest of this chapter is split into five components: introduction to zoonotic diseases of interest (section 1.1), challenges common to statistical modelling of epidemiological data (section 1.2), relevant statistical methods (section 1.3), Bayesian inference (section 1.4), and relevant methods of fitting Bayesian models (section 1.5).

1.1 Introduction to relevant zoonotic diseases

Zoonotic diseases are caused by micro-organisms (such as bacteria, viruses, parasites, and fungi) and can be transmitted to humans via direct contact with an infected animal, environmental exposure, or through ingestion of contaminated foodstuffs. Many of these micro-organisms are commonly found in the gut of healthy food-producing animals. This means the risk of contamination can be high if control measures throughout the food processing chain are not tight. There are four main modes of zoonotic disease transmission between animals and humans: direct contact transmission, indirect contact transmission, vector borne transmission, and foodborne transmission. Infections from direct contact result from exposure to the bodily fluids of an infected animal i.e. saliva, blood, urine, nasal secretions, faeces, and other fluids; or through contaminated air, such as influenza. Indirect infections occur via contact with areas, objects, or surfaces that have been contaminated by an infected animal. Vector borne transmission results when the disease causing micro-organism is transmitted from an infected animal to a human, via another organism; this usually occurs through transmission of infected blood, i.e. when a vector feeds on a human after feeding on an infected animal. Finally, foodborne transmission results from eating food that has been contaminated by an infected animal. The contaminated food is often unpasteurised milk or under cooked meat/egg products, but can also include unwashed fresh produce that has been contaminated with faeces from an infected animal.

1.1.1 Campylobacter and campylobacteriosis

Campylobacter is the most common cause of acute bacterial gastroenteritis worldwide; in the UK alone, *Campylobacter* causes an estimated 700 000 infections each year (Tam et al., 2012) and results in an economic burden of over £1 billion per annum (Humphrey, O'Brien, and Madsen, 2007).

Campylobacteriosis is characterised by sudden onset of fever, abdominal pain and cramping, and diarrhoea containing blood and leukocytes (Blaser, 1997), however, many individuals are thought to be asymptomatic. It is estimated that the number of reported cases in the United Kingdom and the Netherlands (World Health Organisation, 2012) represents only about 10% of the true number of cases. It is likely that this percentage is lower in many countries where healthcare is less accessible. Campylobacteriosis is a major predisposing cause of the peripheral nervous system disorder Guillain-Barré Syndrome (Nachamkin, Allos, and Ho, 1998) and occasionally leads to other health sequelae such as reactive arthritis, meningitis, carditis and skin and urinary problems (Zia et al., 2003).

Campylobacter jejuni has been isolated from diverse animal, human, and environmental sources and has a large genetic diversity (Brownowski, James, and Winstanley, 2014). Due to the large number of potential infection sources (food and environmental), and the relatively long incubation time (between 24 hours and 7 days), it can be difficult to attribute campylobacteriosis cases to sources of infection. Comparing the distribution of the pathogen's sub-species within food sources, to that observed in samples from infected individuals, allows inference on the likely source of infection.

1.1.2 Leptospira and leptospirosis

Leptospirosis is a leading zoonotic cause of morbidity and mortality worldwide (Costa et al., 2015) with similar early symptoms to a number of unrelated infections such as influenza (World Health Organisation, 2003). It is estimated that the global annual incidence of endemic and epidemic human leptospirosis is 5 and 14 cases per 100,000 people, respectively, (World Health Organisation, 2011). Cumulatively, this incident rate results in an estimated 1.03 million cases and 58,900 deaths annually (Costa et al., 2015). Severe cases (diagnosed in 5-10% of patients) result in kidney or liver failure, leading to death in 10%-50% of these patients (McBride et al., 2005). Currently, available vaccines for human leptospirosis only provide partial protection and are often serovar specific, produce only a short duration of protection, and have side effects (World Health Organisation, 2003; Wang, Jin, and Wegrzyn, 2007; Xu and Ye, 2018).

Humans are typically infected through exposure to contaminated water or soil, particularly through damaged skin or mucous membranes. Leptospires are typically spread into the environment through urine of infected animals (many of which are maintenance hosts) where they can survive for months in moist conditions (World Health Organisation, 2003; Izurieta, Galwankar, and Clem, 2008; Haake and Levett, 2015). In developing countries, a tropical climate, high rainfall, disasters, and urban slums are strongly associated with leptospirosis (World Health Organisation, 2003; Costa et al., 2015). Rodents are one of the most important sources of human infections within urban slums (Haake and Levett, 2015) and are causing an increasing number of cases as urban slums expand worldwide (Costa et al., 2015). In contrast, cases in developed countries are typically associated with occupational exposure and freshwater recreational pursuits. High risk professions include those associated with farming, meat processing, veterinary, military, sewerage, and forestry (Costa et al., 2015; Haake and Levett, 2015).

1.2 Introduction to statistical challenges of modelling epidemiological data

There are many statistical challenges involved in modelling epidemiological data. Commonly observed issues include non-independence of data, missing or censored data, and complex data generating processes which can be difficult to model. In particular, this thesis includes methods for Poisson overdispersion in complex non-identifiable models, spatial autocorrelation in hierarchical cross-species models, and temporal correlation and censoring in complex dynamical biological systems.

1.2.1 Independence

The probability of two events A and B both occurring is defined as $Pr\left(A, B\right) = Pr\left(A|B\right) \times Pr\left(B\right)$. When the two events are statistically independent, knowing whether event B has occurred does not change the probability of event A occurring $Pr\left(A|B\right) = Pr\left(A\right)$. Therefore, the two events are statistically independent if and only if their joint probability equals the product of their probabilities $Pr\left(A, B\right) = Pr\left(A\right) \times Pr\left(B\right)$.

Many statistical tests and models assume that the data x are conditionally independent $Pr\left(y_i, y_j|x\right) = Pr\left(y_i|x\right) \times Pr\left(y_j|x\right)$; that is, conditional on the covariates x, knowing the value of one data point does not increase knowledge about the values of other data points. This is equivalent to stating that the model errors are independent.

Traditional statistical models use a likelihood function $L\left(\theta|\cdot\right)$ which is maximised to give the value of θ most consistent with the data y. The likelihood function utilises the multiplication rule of probability to calculate the joint probability of the data given the parameter vector θ. Therefore, the likelihood function is only valid when the y_i are conditionally independent.

$$L\left(\theta|\boldsymbol{y}\right) = Pr\left(y_1, y_2, ..., y_n|\theta\right) = \prod_{i=1}^{n} Pr\left(y_i|\theta\right)$$

The probabilities $Pr\left(y_i|\theta\right)$ are given by the probability mass function $p_\theta\left(y\right)$ for discrete data y and the probability density function $f_\theta\left(y\right)$ for continuous y, or a combination of the two.

Most epidemiological data sets have multiple dependencies at many levels, and identification of these dependencies is often as challenging as correctly adjusting for them. It is often impossible to identify and control all important covariates in epidemiological studies, particularly those involving humans, due to ethical, financial, and practical constraints. Therefore, it is necessary to adapt the statistical models and tests to account for the dependencies between observations.

One of the most common methods of adjusting for residual dependencies is incorporation of random effects. Random effects may be independent group level effects, such as individual level random intercepts or slopes, or be correlated in some way (e.g. spatial, temporal, or genetic correlation structures). These dependencies are sometimes caused by missing information for important covariates such as genetic relatedness or common environmental exposures over space and/ or time. It is often possible to detect dependencies by observing correlation in the residuals (although lack of observed correlation in the residuals does not guarantee their independence). Typically, groups of dependent data points have more similar values within than between groups (positive correlation), although negative correlations can also occur.

For example, if an experimenter wished to investigate the effect of several treatments on weight in adult sheep they might randomly assign individual animals to each treatment and weigh the animals post treatment. It may be appropriate to model this data using a linear model such as $\mathsf{weight}_{ij} \sim \mathsf{Normal}\left(\mu_j, \sigma\right)$ where $\mu_j = \beta_0+\beta_j$ is the linear predictor giving the mean weight for animals given treatment j and i is an index for each individual animal. If the animals were grouped in some way e.g. located in pen h, then pen (group level) random effects can be used to account for dependencies correlated with pen, such as competition for food or shared exposure. A random intercept model with Gaussian distributed random Pen effects would modify the linear predictor to give $\mu_{jh} = \beta_0 + \beta_j + u_h$ where $u_h \sim \mathsf{Normal}\left(0, \sigma_{\mathsf{pen}}\right)$. If measurements were taken on the same animals over time, more complex random effect structures are required. For example, the model could be extended to incorporate animal level random intercepts to account for animal level differences from the group mean (perhaps caused by unmeasured covariates such as initial weight). The model could also be extended to incorporate any other appropriate random effect structure, such as animal level random time trends, temporally correlated random effects, spatially correlated pen effects and/ or genetically correlated animal level random effects, given enough informative data.

1.2.2 Missing and censored data

Datasets often contain missing values, particularly epidemiological and ecological datasets where the experimental conditions cannot be tightly controlled. Missingness can occur for many reasons, such as study participant drop out, equipment failure, equipment with a restricted range of measurement or an inability to directly measure the variable of interest. Ideally, data will be missing completely at random (MCAR), which occurs when the cause of missingness is unrelated to all covariates, confounders, and the variable with the missing values itself. An example of this would be a sample being lost or damaged before it is analysed. Values are missing at random (more accurately called missing conditionally at random, MAR), when we are able to explain the probability of missingness at least partly using covariates, but it is still independent of the variable containing the missingess. For example, a child may miss a spelling test at school because they are sick. This may be partially explained by measured health covariates, but would typically not be related to the score the child would have received had they attended. Missing not at random (MNAR) occurs when the missingess is related to the variable that is missing. For example, an individual may miss a drug test because they took drugs the night before. Typically data where observations are MNAR are difficult to analyse because MNAR has a large potential to cause bias.

Missing data are problematic because it can cause a loss of statistical power (when incomplete cases are removed from the analysis) and bias (particularly for MNAR). Missing data typically occurs for a portion of data in the study, but may also occur for all data points in the special case where the variable of interest is not able to be measured directly. The Lotka-Volterra prey-species dynamic (Lotka, 1910) is a good example of this phenomenon. If a researcher was interested in modelling the number of rabbits in an area, in addition to covariates such as weather and food sources, they might expect that the distribution of predators, such as foxes, would also be a causative explanatory variable. It would be difficult to measure the fox population density and activity levels precisely, resulting in missing or low quality fox covariate data. Instead, they could estimate the fox density covariate by jointly sampling fox and rabbit data, and estimating the missing fox density data. Had the study ignored the fox variable, it would have likely resulted in unexplained spatial and or temporal dependencies in the data, due to the unobserved fluctuating fox densities (see the above section for a discussion on the impact of this). Had the study decided to directly incorporate some observed fox information, such as observed counts in each area, the model would likely generate inconsistent estimates including bias and underestimate uncertainty.

Censoring is a special case of MNAR where the reason the data are missing is that the value is outside some known range (usually in a known direction). This is extremely common in survival analysis applications where an individual or item may experience the event of interest before or after the pre-defined study time period, or where status changes are recorded after a block of time rather than continuously. It is also common in chemical analyses where concentrations outside the range of a standard curve are recorded as below or above the detectable limits. This type of missing information does not typically cause bias as the likelihood function can be adjusted for censored data. The likelihood is adjusted for censored values by replacing the probability that random variable X is observed to have value x given by $P(X = x)$ with the probability that it is in the observed range $P\left(X > x_U\right)$ (right censoring), $P\left(X < x_L\right)$ (left censoring) or $P\left(x_L < X < x_U\right)$ (interval censoring) where x_L and x_U are the lower (left) and upper (right) limits of the range. There can be multiple sets of ranges within a single data set. For example, an individual attending yearly health check-ups has censored data with year long censoring durations, whilst another individual may have 6 month long durations which may have limits at different times of the year.

1.2.3 Fitting models for data generated from complex biological system

The above issues become all the more challenging when combined in a model for a complex and dynamical biological system. The desired model is typically one which accurately reflects the data generating process; when this biological process is very complex, the resulting desired model is also complex. Due to practical constraints, it is often not possible to collect an ideal data set to support inference on all aspects of this model. This can result in non-identifiability due to negative degrees of freedom (more parameters than data points), correlations between covariates and/or structurally missing data. There are several methods of reducing the effective number of parameters such as simplifying the model, clustering techniques (e.g. using a Dirichlet Process) and partial pooling using random effects (via shrinkage). Although the resulting models may be mathematically straightforward, they can be difficult to fit with real data, and care must be taken when interpreting results.

For example, the desired model for chapter 2 data contains more parameters than data resulting in non-identifiability unless the model is modified to reduce the number of parameters (using clustering via a Dirichlet Process). It is difficult to identify and interpret the covariate effects in the model used in Chapter 3 due to complicated correlations between covariates. The desired model in Chapter 4 is not identifiable (the initial values must be fixed to constants and time of infection cannot be estimated jointly with other model parameters) due to structurally missing data, and was simplified (using random multiplicative effects rather than random slopes for individuals and assuming a simplified set of ordinary differential equations to describe the mean curve).

Despite the challenges involved, these models form an integral part of addressing social, environmental, and health inequalities globally; the results of which are used by government, non-governmental organisations (e.g. the WHO), private organisations (e.g. the Gates Foundation), and others to create and modify policies and interventions aimed at improving well-being globally.

1.3 Relevant stochastic processes

Bayesian non-parametrics and dynamical models allow far more flexible models to be fitted than would be possible using traditional techniques such as standard generalised linear models. Bayesian non-parametric hierarchical models, including Dirichlet and Gaussian Processes (DP and GP, respectively), allow the number of latent variables to grow as necessary depending on the data; however, the individual variables and the process controlling the growth of latent variables is parametric. Non-parametric techniques allow complex shapes to be fitted with a relatively simple model with few assumptions. Parametric alternatives, for example using a polynomial function of x and y to fit a spatial surface, often need an extremely large number of parameters and have extremely poor predictive ability outside the range of the data.

Bayesian models allow prior knowledge to be added to the model through prior distributions. Often, the domain knowledge is difficult to relate to specific parameters in the model and informative priors can be difficult to specify. Another method of incorporating prior information into models is by specifying the shape of the mean function. Linear models can be extended to non-linear models which allow more flexible curves to be fitted; however, it can be difficult to choose the correct non-linear form. Non-parametric techniques, such as splines, are extremely flexible and can fit complex data patterns well. However, they are unreliable in areas with little data, work particularly poorly outside the range of the observed data, and provide little additional understanding of the data generating process. If knowledge of the underlying causing mechanism suggests a functional form for the data generating process, an appropriate dynamical system can be incorporated into the model; this restricts the possible shape or form of the mean function and often makes model parameters more interpretable. It is particularly useful when the data are not strongly informative, such as when data are observed with large error, when predictions need to be made outside the range of the observed data; or when there are significant amounts of missing information e.g. from censoring. This is because the restriction on the shape of the fitted curve (from dynamical model) adds information. This is similar to the use of strong priors, however the restrictions are typically applied to the fitted values (e.g. shape of the fitted curve) rather than to individual model parameters.

The following three subsections introduce Dirichlet and Gaussian Processes and dynamical modelling.

1.3.1 Dirichlet distributions and processes

Chapter 2 of this thesis uses a non-parametric Dirichlet Processes; therefore, a short introduction to DP's is provided here. The information in this section is primarily from Whye Teh (2007), Paisley (2009), and Frigyik, Kapila, and Gupta (2010). Please refer to these resources for a more detailed introduction to Dirichlet Processes.

1.3.1.1 Dirichlet distribution

The Dirichlet distribution is a family of continuous multivariate probability distributions parameterised by a vector $(\alpha'_1, ..., \alpha'_K)$, $\alpha'_i = \alpha_i/\sigma$ of positive real numbers (the base measure) and a scale parameter $\sigma = \sum_i \alpha_i$. The support of the Dirichlet distribution is the set of K-dimensional vectors $\boldsymbol{g}$ whose entries are real numbers in the interval $(0, 1)$ which sum to 1. Therefore, the domain of the Dirichlet distribution is itself a set of probability distributions, specifically the set of K-dimensional discrete distributions, where the vector $\boldsymbol{g}$ can be viewed as the probabilities of a K-way categorical event. The set of points in the support of a K-dimensional Dirichlet distribution is referred to as the open-standard $(K-1)$-simplex.

Formally, let $G = [G_1, ..., G_K]$ be a random probability mass function (therefore, $G_i \geq 0$

for $i = 1, ..., K$ and $\sum_{i=1}^{K} G_i = 1$) and $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_K]$ be a vector of positive real numbers. G has a Dirichlet distribution with parameter $\boldsymbol{\alpha}$, denoted $G \sim \mathsf{Dir}(\boldsymbol{\alpha})$ if it has $f(\boldsymbol{g}; \boldsymbol{\alpha}) = 0$ if $\boldsymbol{g}$ is not a pmf, and

$$f(\boldsymbol{g}; \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} g_i^{\alpha_i - 1} \tag{1.1}$$

otherwise.

The base measure determines the mean distribution whilst altering the scale affects the variance

$$\mathsf{E}(g_i) = \frac{\alpha_i'}{\sigma} = \alpha_i \tag{1.2}$$

$$\mathsf{Var}(g_i) = \frac{\alpha_i(\sigma - \alpha_i)}{\sigma^2(\sigma + 1)} = \frac{\alpha_i'(1 - \alpha_i')}{(\sigma + 1)} \tag{1.3}$$

$$\mathsf{Cov}(g_i, g_j) = \frac{-\alpha_i \alpha_j}{\sigma^2(\sigma + 1)} \quad (i \neq j) \tag{1.4}$$

The Dirichlet distribution is the multivariate generalisation of the Beta distribution, and is commonly used in many areas of statistics and machine learning such as in Bayesian modelling (as the conjugate prior of the multinomial distribution) and in natural language processing (in the context of the compound Dirichlet distribution) (Frigyik, Kapila, and Gupta, 2010).

The marginal distributions of a Dirichlet distribution are beta distributions:

$$g_i \sim \mathsf{Beta}(\alpha_i, \sigma - \alpha_i) \tag{1.5}$$

A Dirichlet prior with small scale parameter σ, favours sparse distributions but this prior belief is very weak and is easily overwritten by data. Larger values of σ have smaller covariances, and thus the mass is more evenly dispersed. This favours variates that are dense, evenly distributed distributions (i.e. all the values within a single sample are similar to each other and clustered around the mean $\boldsymbol{\alpha}'$). As the scale parameter increases towards infinity, the variance and covariance tend towards zero; hence, the samples tend towards the base measure. The symmetric Dirichlet distribution is commonly used as a non-informative prior. It is a special case where all elements of the $\boldsymbol{\alpha}$ vector have the same value (called the concentration parameter). A symmetric Dirichlet distribution with concentration parameter equal to 1, is equivalent to a uniform distribution over the open standard (K-1)-simplex (i.e. it is uniform over all points in its support). See Figure 1.1 for examples of draws from various Dirichlet distributed variables.

1.3.1.2 Dirichlet Process

Dirichlet processes (DP) are a family of stochastic processes whose realisations are probability distributions; that is, it is a distribution over distributions. In the same way that the Dirichlet distribution is the generalisation of the Beta distribution, the DP generalises the Dirichlet distribution. It is classed as a non-parametric model because distributions drawn from a Dirichlet process are discrete, but cannot be described using a finite number of parameters. It is called a Dirichlet process because it has Dirichlet distributed finite dimensional marginal distributions. That is, if G is a probability distribution over a measurable space Θ, then $G \sim DP\left(\alpha, G_0\right)$ means that

$$\left(G\left(T_1\right), ..., G\left(T_K\right)\right) \sim \mathsf{Dirichlet}\left(\alpha G_0\left(T_1\right), ..., \alpha G_0\left(T_K\right)\right) \tag{1.6}$$

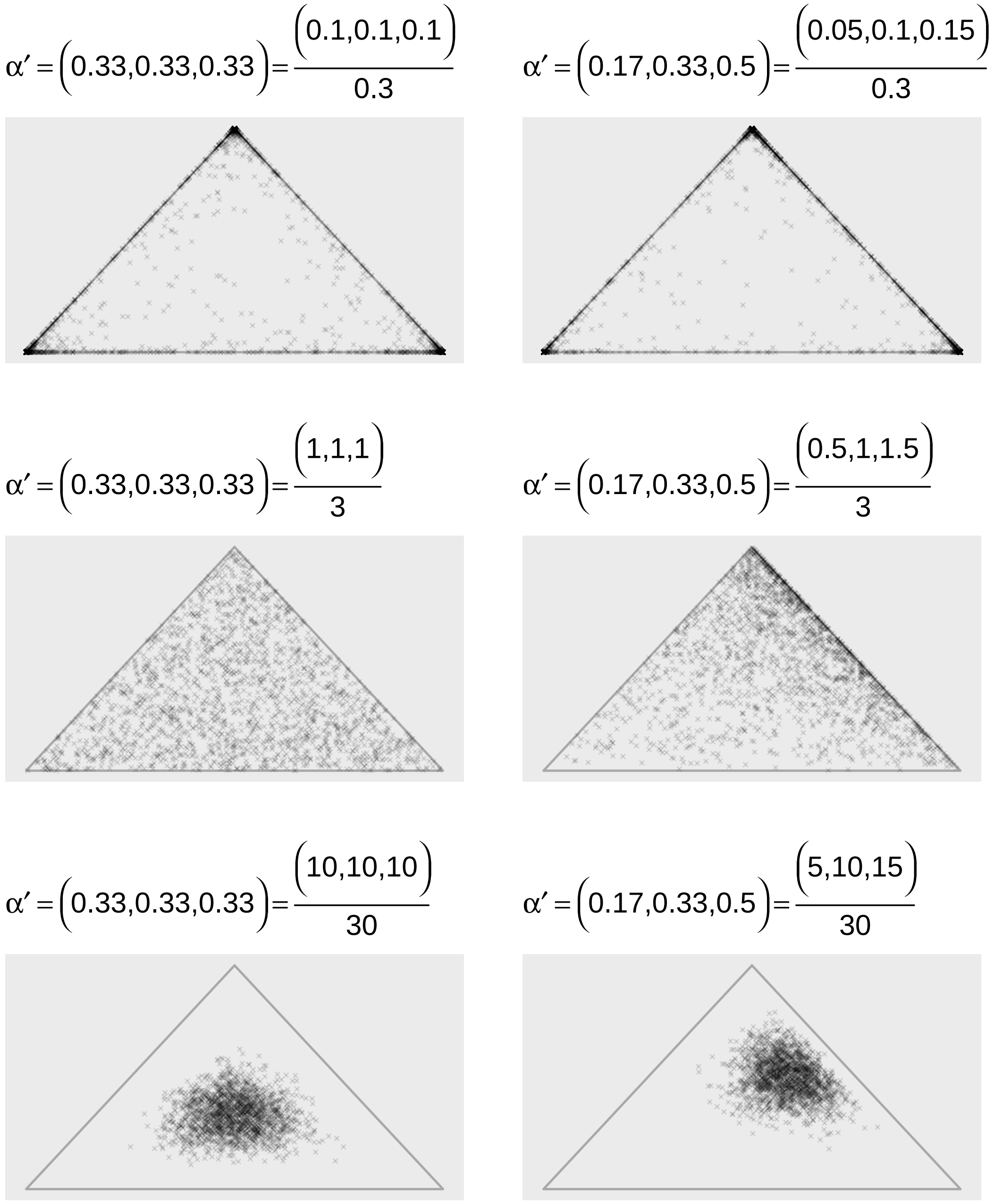


Figure 1.1: Plots of sample pmfs drawn from Dirichlet distributions over the probability simplex in $\mathbb{R}^3$ for various values of $\alpha'_i = \alpha_i/\sigma$ where $\sigma = \sum_i \alpha_i$. When $\alpha' = [c; c; c]$ for some $c > 0$, the density is symmetric about the uniform pmf (which occurs in the middle of the simplex), and the special case $\alpha' = [1; 1; 1]$ is the uniform distribution over the simplex. When $0 < c < 1$, there are sharp peaks of density almost at the vertices of the simplex and the density is tiny away from the vertices. When $c > 1$, the density becomes concentrated in the centre of the simplex. If α' is not a constant vector, the density is not symmetric. For more information see Frigyik, Kapila, and Gupta (2010). Each row has the same scale parameter σ and each column has the same α'.

for any finite partition $T_1, ..., T_K$ of Θ. Therefore, the probabilities that G assigns to any finite partition of Θ follow a Dirichlet distribution with parameters $\alpha G_0 (T_1), ..., \alpha G_0 (T_K)$.

This can be equivalently expressed as

$$\begin{aligned} G|\alpha, G_0 &\sim DP(\alpha, G_0) \\ \theta_i &\sim G \\ x_{ij}|\theta_i &\sim F(\theta_i) \end{aligned} \tag{1.7}$$

where x_{ij} is the jth observed data point in group i. The x_{ij} are $F(\theta_i)$ distributed where F is parametrised by θ_i.

The mean of a DP is its base distribution G_0, and the concentration parameter α can be thought of as an inverse variance. This means that on average, distributions drawn from a DP look like G_0. Because G is discrete, multiple θ_i's can take the same value resulting in draws that are always discrete, even if the base distribution is continuous. This discretisation means the DP can be seen as a mixture model where all x_i's with the same θ_i belong to the same cluster. The concentration parameter α specifies how strong the discretization is. In the limit of $\alpha \to 0$, the realisations are all concentrated at a single value, while in the limit of $\alpha \to \infty$ the realisations become continuous. Values of the concentration parameter between these two extremes results in realisations that are discrete distributions with decreasing concentration as α increases. When using the DP as a prior in a Bayesian non-parametric model, the concentration parameter controls the strength of that prior; hence, it is also referred to as the strength parameter.

Dirichlet Processes are used across a wide range of applications in Bayesian statistics and machine learning, such as Bayesian model validation, selection and averaging, density estimation, and clustering via mixture models (Frigyik, Kapila, and Gupta, 2010). Common

Dirichlet Process representations include the Chinese Restaurant Process (CRP) and the Stick Breaking Process (SBP) which I will briefly introduce in the following sections.

Chinese restaurant process: The intuition behind the CRP can be shown by visiting an imaginary chinese restaurant and choosing a table to sit at. When each new customer arrives, they must select from a countably infinite number of tables labelled $1, 2, ...\infty$. Their choice of table follows a random process where

1. The table chosen by the first customer is labelled 1.
2. The nth customer chooses a new table with probability $\frac{\alpha}{n-1+\alpha}$ and an existing table with probability $\frac{c}{n-1+\alpha}$ where c is the number of people already at the table

Equation 1.8 shows the probability of 2 sets of seating arrangements where $n = 6$.

$$\begin{aligned} Pr\left(z_1, ..., z_6\right) =& Pr\left(z_1\right) Pr\left(z_2 | z1\right) ... Pr\left(z_6 | z_1, ..., z_5\right) \\ Pr\left(1, 2, 3, 3, 1, 1\right) =& \frac{\alpha}{\alpha} \frac{\alpha}{1+\alpha} \frac{\alpha}{2+\alpha} \frac{1}{3+\alpha} \frac{1}{4+\alpha} \frac{2}{5+\alpha} \\ Pr\left(1, 2, 2, 3, 3, 3\right) =& \frac{\alpha}{\alpha} \frac{\alpha}{1+\alpha} \frac{1}{2+\alpha} \frac{\alpha}{3+\alpha} \frac{1}{4+\alpha} \frac{2}{5+\alpha} \\ =& 2\alpha^3 \sum_{i=1}^{n} \frac{1}{\alpha + (i-1)} \end{aligned} \tag{1.8}$$

The resulting table allocation sequences are different, but they have identical probabilities of occurring. Note that the denominators are identical for all sequences of length n. The only changes are in the numerator, and these consist only of permutations for all identically sized partitions which consist of the same number of components. This means the sequences can be exchanged with each other simply by switching the labels of the groups (tables).

A CRP mixture is constructed using the following algorithm

```
1 ) Draw a random partition using CRP: Customers $i = 1, 2, ..., n$ choose tables
 $z_1, ..., z_K \sim \mathsf{CRP}(\alpha)$;
2a) Draw a random parameter for each group $z_k; k = 1, ..., K$ (table) from $G_0$:
 $\theta_k^* \sim G_0$;
2b) Assign $\theta_i = \theta_k^*$ for each $i$ in cluster $z_k$;
3 ) Draw a sample data point from $G$: $x_i \sim G(\theta_i)$
```

Algorithm 1: Chinese Restaurant Process (CRP) algorithm used to create a mixture distribution.

Exchangeability allows us to draw a parameter θ_k^* from the prior G_0, then draw data x_i iid from that prior. This results in data that are independent conditioned on the parameter, but not independent in general.

More formally, de Finetti's exchangeability theorem (de Finetti, 1931) states that the exchangeability of a random sequence $x_1, x_2, ...$ is equivalent to having a parameter ψ drawn from a distribution $F(\cdot)$, then drawing x iid from the distribution implied by θ. In the CRP scenario, $F(\cdot) = \mathsf{Dirichlet\ Process}(\alpha G_0)$ and $\psi = G$.

$$\begin{aligned} G &\sim \mathsf{Dirichlet\ Process}(\alpha G_0) \\ \mu_{z_i} &\overset{iid}{\sim} G \end{aligned} \tag{1.9}$$

Stick-breaking process: An alternative constructive definition of a DP is called the stick-breaking process (Sethuraman, 1994). Here, we imagine a stick of length 1, which is recursively broken into portions of length $\pi_1, \pi_2, ...$ where the length of stick k is given by

$$\pi_k = \beta_k \prod_{i=1}^{k-1} (1 - \beta_i) \tag{1.10}$$

The locations of the breaks β_k are iid Beta distributed

$$\beta_k \sim \mathsf{Beta}\,(1, \alpha) \tag{1.11}$$

This creates a scenario in which the values of π are stochastically decreasing as the number of stick pieces increases. The vector of stick lengths $\pi_1, \pi_2, ...$ are used to define the probability of choosing each group $Pr\,(1) = \pi_1, Pr\,(2) = \pi_2, ...,$ and thus , the stick breaking process creates clusters.

We can then define

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*} \tag{1.12}$$

$$\theta_k^* \sim G_0 \tag{1.13}$$

$$G \sim \mathsf{DP}\,(\alpha, G_0) \tag{1.14}$$

where $\delta_{\theta_k^*}$ is the dirac delta measure and denotes a point mass at θ_k^*. The stick breaking distribution over π is sometimes written referred to as the Griffiths-Engen-McCloskey distribution $\pi \sim \mathsf{GEM}\,(\alpha)$. Although this creates an infinite number of possible groups, in practice, the maximum number K must be limited for computational efficiency. The value of K must be set to be much larger than the expected number of groups from the data, such that $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_i$ is very small.

We can construct a SBP as follows

```
1a) Draw a random vector of group probabilities using SBP:
 $\pi_1, \pi_2, ..., \pi_{K-1} \sim \mathsf{Truncated\ GEM}\,(\alpha)$;
1b) Set $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$;
2 ) Draw a random parameter for each group from $G_0$: $\theta_k^* \sim G_0$;
3 ) Assign each individual $i$ to a group with probabilities given by $\pi$;
4 ) Draw a sample data point from $G$: $x_i \sim G\,(\theta_i)$
```

Algorithm 2: Stick Breaking Process (STB) algorithm used to create a mixture distribution.

The CRP process has the benefit of allowing new groups to be Gibbs sampled which can be very efficient, whilst the SBP is more flexible allowing any combination of likelihood and base distribution.

1.3.2 Gaussian Process

Chapter 3 of this thesis uses Gaussian Processes (GP) to account for spatially correlated dependencies in the data; as such, a short introduction to GP's in the context of spatial modelling is given here. The information in this section is primarily from Diggle and Ribeiro (2007). See this book for a more detailed treatment of Gaussian Processes and model based geostatistics in general.

Gaussian Processes are often used in geostatistical models to account for spatial dependencies in point referenced data. The GP fits a continuous spatial surface with a defined correlation structure (Diggle, Tawn, and Moyeed, 1998). The GP distribution is a family of distributions over stochastic processes, where a stochastic process is a collection of random variables on some probability space indexed by a variable (here spatial location). A Gaussian process can be thought of as an infinite collection of random variables with the property that any finite subset has a multivariate normal distribution, with mean μ and covariance function Σ. A GP is isotropic and stationary when the covariance function only depends on the distance matrix and the mean is constant. This assumption is commonly made for simple geostatistical models, but may be relaxed if necessary.

We can define a GP spatial surface $S(\cdot)$ as follows

$$S(x) : x \in \mathbb{R}^2 \tag{1.15}$$

where any finite subset of $S(\cdot)$ is Multivariate Normal and

$$\mathsf{E}[S(x)] = \mu \tag{1.16}$$

$$\Sigma = \mathsf{Cov}[S(x), S(x')] = \sigma^2 \rho(||x - x'||) \tag{1.17}$$

The mean μ may be fixed to 0 (common when using a GP to model spatially correlated random effects) or specified by a set of parameters and covariates such as a linear predictor. The correlation function $\rho(\cdot)$ must be positive definite (as with the covariance matrix of a multivariate normal distribution). There are a wide range of suitable correlation functions, such as the exponential, Gaussian, power exponential and Matérn. These functions require estimation of parameters which control the shape, smoothness, and correlation decay rate. The roughness of a given GP is directly linked to the differentiability of its covariance function which can be easily calculated for many common correlation functions. Thus, the choice of correlation function is, in a sense, incorporating prior information about the expected smoothness of the spatial surface.

Two common covariance functions are the powered exponential and Matérn families, which also encompass the Gaussian and exponential functions as special cases.

1. Powered exponential family:

$$\rho\left(d;\phi,k\right)=\mathsf{exp}\left(-\left(\phi d\right)^{k}\right) \tag{1.18}$$

   where $k\in(0,2]$, $\phi>0$ is the scale parameter and d is the distance between any two points in the study region

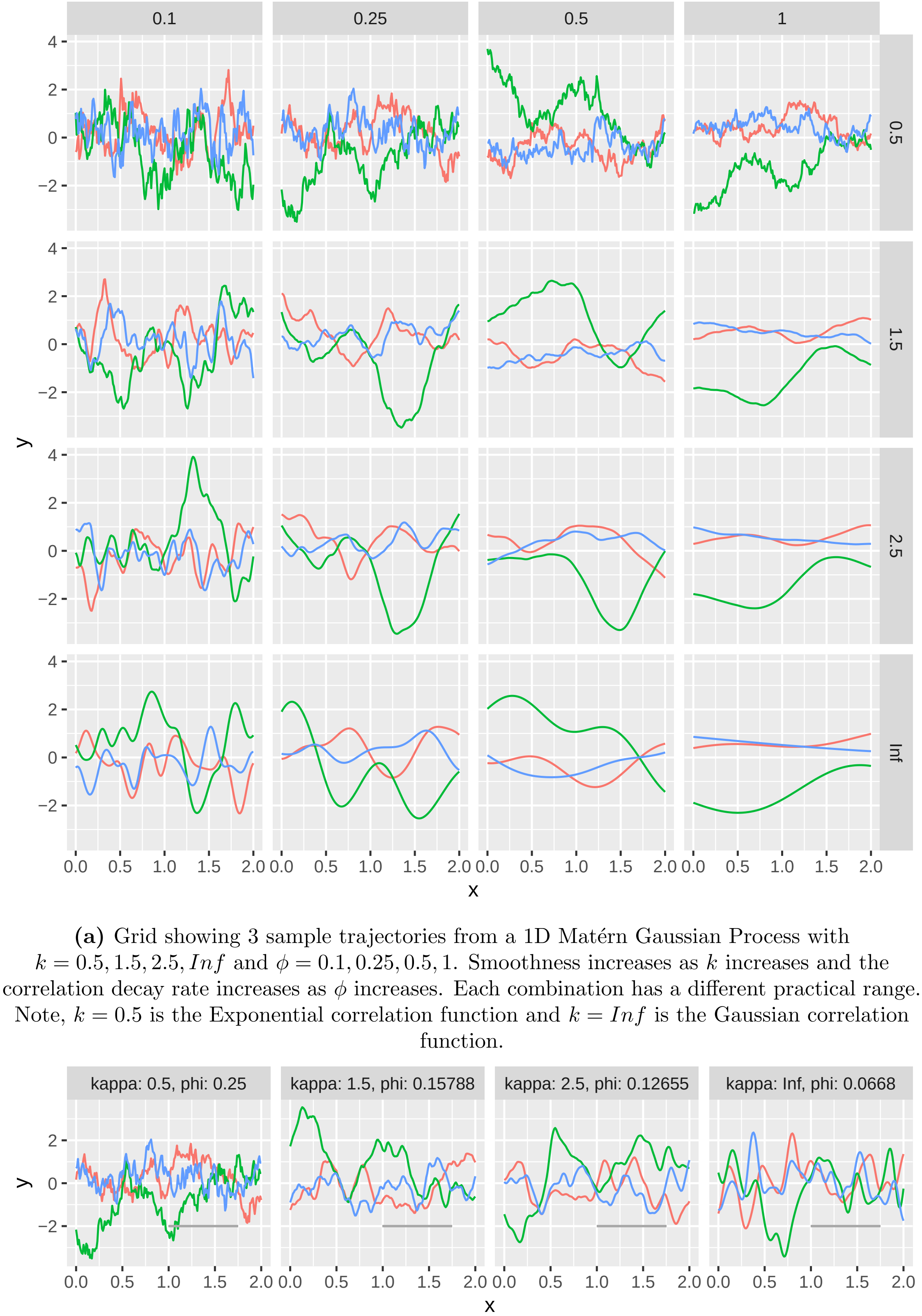
2. Matérn family:

$$\rho\left(d;\phi,k\right)=\frac{1}{2^{k-1}\Gamma\left(k\right)}\left(d\phi\right)^{k}K_{k}\left(d\phi\right) \tag{1.19}$$

   where $k>0$ is the number of times the function is mean-square differentiable, K_k is the modified Bessel function of the second kind, $\phi>0$ is the scale parameter and d is the distance between any two points in the study region. This equation can be simplified to the product of an exponential and polynomial of order p when $k=p+0.5, p\in\mathbb{N}^{+}$.

The scale parameter ϕ controls the decay rate of the correlation with distance, whilst k controls the degree of smoothing. The exponential correlation function (a special case of the powered exponential function with $k=1$ and the Matérn function with $k=0.5$) results in GP's which are not differentiable, and therefore are very rough. Conversely, Gaussian correlation functions result in processes which are infinitely differentiable ($k=2$ for the powered exponential and $k\rightarrow\infty$ for Matérn) and therefore extremely smooth. The practical range is defined as the distance at which the correlation is near 0 (usually the value chosen is 0.05). The practical range is affected by both parameters of the correlation function: k and ϕ. The effect of each parameter can be visualised by comparing sample trajectories from a single dimensional GP, with varying k and ϕ as in Figures 1.2a and 1.2b. The corresponding correlation functions are shown in figures 1.3a and 1.3b.

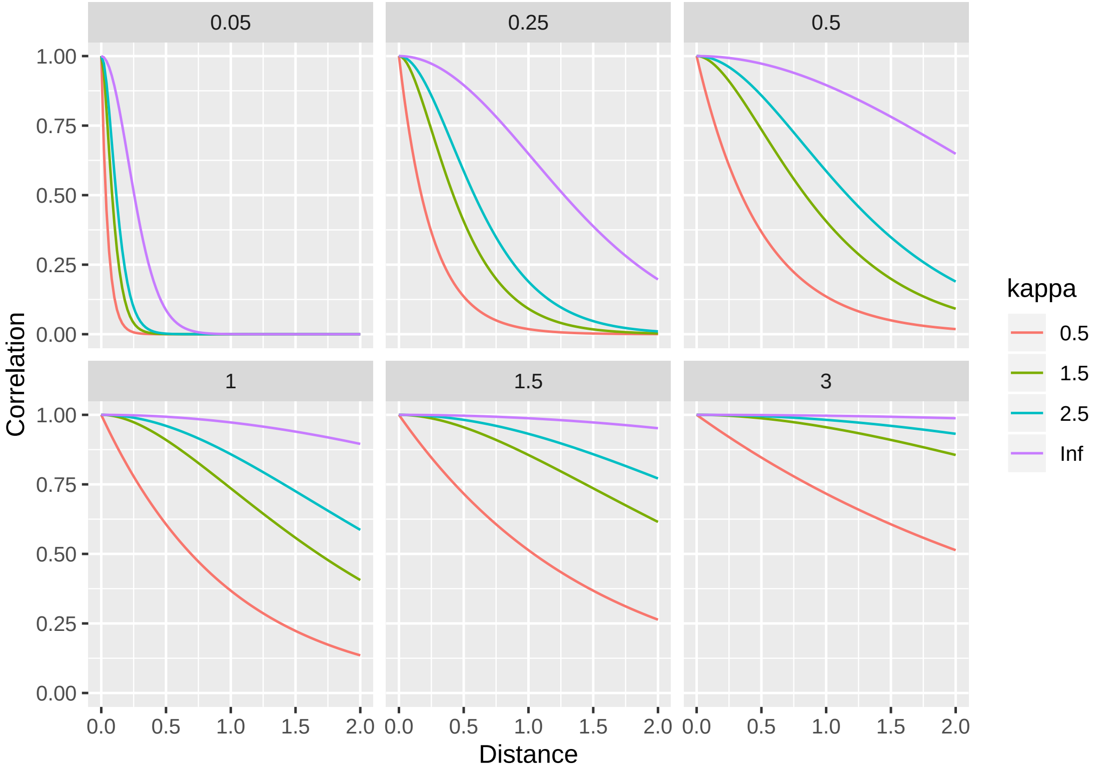
The information about k can be very flat for some values of ϕ (Schmidt, Conceição, and



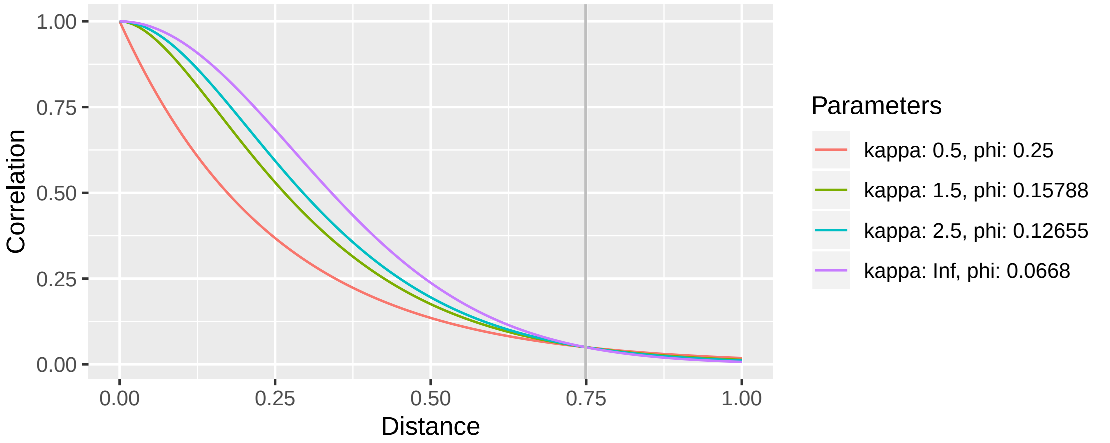
(a) Grid showing 3 sample trajectories from a 1D Matérn Gaussian Process with $k = 0.5, 1.5, 2.5, Inf$ and $\phi = 0.1, 0.25, 0.5, 1$. Smoothness increases as k increases and the correlation decay rate increases as ϕ increases. Each combination has a different practical range. Note, $k = 0.5$ is the Exponential correlation function and $k = Inf$ is the Gaussian correlation function.

(b) Grid showing 3 sample trajectories from a 1D Matérn Gaussian Process with $k = 0.5, 1.5, 2.5, Inf$ and ϕ chosen such that the practical range is the same for each. The practical range distance is shown as a horizontal grey line segment.

Figure 1.2: Matérn Gaussian Process sample trajectories (1D).



(a) Grid showing the Matérn correlation function with $k = 0.5, 1.5, 2.5, Inf$ and $\phi = 0.1, 0.25, 0.5, 1$. Smoothness increases as k increases and the correlation decay rate increases as ϕ increases. Each of the combinations has a different practical range. Note, $k = 0.5$ is the Exponential correlation function and $k = Inf$ is the Gaussian correlation function.



(b) Matérn correlation function curve with $k = 0.5, 1.5, 2.5, Inf$ and ϕ chosen such that the practical range is fixed at 0.75. Smoothness increases as k increases and the correlation decay rate increases as ϕ increases. Each of the combinations has a different practical range. The practical range is shown as a grey vertical line.

Figure 1.3: Matérn Gaussian Process functions (1D).

Alberti Moreira, 2008), and many combinations of k and ϕ can give the same practical range. This has lead many authors to fix k at a particular value, given a priori knowledge about the expected smoothness of the surface. This is easiest to do with the Matérn function as the relationship between k and the smoothness of the surface is easy to calculate, and the function can be significantly simplified when k is a half integer.

In addition to inferring the properties of the spatial process $S\left(.\right)$, geostatistical models often aim to predict the process at new locations of interest. The fitted model can be used for prediction by sampling from the predictive distribution $\left[\mathcal{F}\left(S\right)|Y\right]$ where $\mathcal{F}\left(S\right)$ is the target for prediction. This is achieved by drawing samples $S_i : i = 1, ..., N$ from $\left[S|Y\right]$ where $S = \left\{S\left(x\right) : x \in A\right\}$, and applying the selected transformation $\mathcal{F}$ to get $\mathcal{F}_i = \mathcal{F}\left(S_i\right)$.

1.3.3 Introduction to Bayesian dynamical modelling

A dynamic system is a stochastic process of the form $\left(X_n, Y_n\right)$, where X_n is the true state of the system at time n, and $Y_n = f\left(X_n\right)$ is an observation of some function of X_n. The system may run over either continuous time or discrete time, and may incorporate noise in either X_n, called dynamical noise; or Y_n, called observational noise; or both (McGoff et al., 2015).

It is often possible to describe a (usually simplified) stochastic process T_θ using a system of ordinary differential equations. The parameters of this system θ are usually unknown, and must be estimated from a time series of empirical observations which follow an observation model f_θ. A key feature of these processes is feedback, where the current values of each variable affect the future values of the other variables.

An example of a dynamic system is the Lotka-Volterra predator–prey model which describes how the populations of a pair of predator and prey animals change over time in response to each other. The simplest version of this model shows that the numbers of predator and prey animals in the environment fluctuates in a regular pattern (similar to simple harmonic motion), with the population of the predators trailing that of the prey in the cycle. This is because the prey population at time $t+1$ depends not only on the prey population at time t, but also the predator population at time t, and vice versa. Here, X_n gives the true population numbers of the predator and prey animals at time n, and Y_n is a (possibly noisy) observation of X_n e.g. the number of animals of each species trapped at time n.

It is possible to incorporate a dynamic system into a linear regression model by using the deterministic system to provide an expected population value; around which there will be variation due to both measurement error and simplifications in the scientific model.

The parameters θ and the initial state x_0 can be estimated in a Bayesian framework where the likelihood function (denoting the probability of observing y^{n-1} given the parameters θ and the true initial condition x_0) is combined with priors for each parameter to give

$$\pi\left(x_0, \theta | y^{n-1}\right) \propto P\left(y^{n-1} | x_0, \theta\right) \pi\left(x_0, \theta\right) \tag{1.20}$$

Simplified models are often preferred for modelling purposes as they capture the essential behaviours of the system without requiring huge numbers of parameters to be estimated from relatively small volumes of data. For a more detailed introduction to statistical inference for dynamical systems see McGoff et al. (2015).

1.4 Introduction to Bayesian inference

All models in this thesis were fitted in a Bayesian framework as this is a principled, coherent, unbiased, and flexible method of inference with very interpretable results. Bayesian inference uses probability to quantify uncertainty using probability distributions, and allows prior information to easily be incorporated into a model.

1.4.0.1 Bayes theorem

In a Bayesian framework, all model parameters are assigned a distribution, which is combined with the distribution of the data using Bayes theorem

$$E\left[P\left(\theta|D\right)\right] = \frac{P\left(\theta\right) \times P\left(D|\theta\right)}{P\left(D\right)} = \frac{P\left(\theta\right) \times P\left(D|\theta\right)}{\int P\left(\theta\right) P\left(D|\theta\right) d\theta} \tag{1.21}$$

where θ are the model parameters and D is the data. The likelihood $P\left(D|\theta\right)$ embodies our statistical model and gives the probability of the evidence (data) given the parameters θ. The prior $P\left(\theta\right)$ quantifies our beliefs about the value of each parameter prior to the current study. This information can come from expert opinion, previous studies, or be designed to have a weak influence on the results. The posterior $P\left(\theta|D\right)$ is a probability distribution describing the probability of the model parameters θ, given the evidence (data) D. The probability of the evidence $P\left(D\right)$ can be thought of as a normalising constant for the posterior distribution and is usually ignored as it is constant for a given data set. Therefore, we can rewrite Bayes theorem as

$$\textsf{Posterior probability} \propto \textsf{Prior probability} \times \textsf{Likelihood} \tag{1.22}$$

Every parameter in the likelihood requires a prior distribution. Often the parameters of the prior distributions are fixed values; however, priors can be put on these parameters forming a hierarchical model. Hierarchical models are especially important when data are available on several different levels of observational units, and when random effects are required.

1.4.0.2 Posterior summaries

As the posterior is a joint distribution over all model parameters (given the data), it is often very high dimensional and difficult to interpret directly. Marginal distributions for each parameter are usually summarised using point estimates (mean or median) and measures of uncertainty, such as percentile intervals or highest posterior density intervals, both of which are often referred to as credible intervals.

These summary statistics can be expressed in terms of posterior expectations of functions of θ. The posterior expectation of a function $f(\theta)$ is

$$E[f(\theta)] = \frac{\int f(\theta) P(\theta) P(D|\theta) d\theta}{\int P(\theta) P(D|\theta) d\theta} \tag{1.23}$$

It is usually impossible to algebraically solve the integrations in this expression, and numerical evaluation is difficult and inaccurate in high dimensions (W. Gilks, Richardson, and Spiegelhalter, 1996). Therefore, Monte Carlo integration (including Markov chain Monte Carlo (MCMC) methods, and Laplace approximations) is commonly used to evaluate these expectations. See section 1.5 for a high level introduction to some common MCMC algorithms, including those used to fit the models in this thesis.

Bayesian models can be easier to interpret than frequentist models because they allow direct probability statements to be made about the values of the parameter, rather than relying on binary hypothesis tests. For example, where a frequentist model might test whether a regression coefficient is significantly different to 0 at the 5% level, whilst a Bayesian model can directly calculate the probability that the parameter is larger or smaller than 0. If this probability is sufficiently large (or small), we can conclude that it is highly likely that the associated covariate has a non-trivial effect on the results. Note that the size of the coefficient still needs to be checked to make sure the covariate is practically significant in the specified context. Probabilistically interpreted credible intervals are also easier to interpret than confidence intervals, which must be interpreted in the context of hypothetical long run frequencies.

1.5 Relevant MCMC methods

Markov chain Monte Carlo (MCMC) methods are methods to perform Monte Carlo integration using Markov chains. There are a wide range of MCMC algorithms available with highly varying complexity and performance. Essentially these procedures all draw samples from a required distribution, usually called a target distribution $\pi(\cdot)$, then use sample averages to approximate the desired expectations.

First we introduce Monte Carlo integration and Markov chains, then introduce several common MCMC algorithms including those used to fit the models in this thesis.

The information in this section is primarily from W. Gilks, Richardson, and Spiegelhalter (1996) and C. Robert and Casella (2011). See W. Gilks, Richardson, and Spiegelhalter (1996) for a more detailed introduction to this topic, and C. Robert and Casella (2011) for an introduction and comprehensive history of MCMC methods.

1.5.1 Introduction to Monte Carlo integration and Markov chains

1.5.1.1 Monte Carlo integration

Let X be model parameters and missing data, and $\pi(\cdot)$ be the posterior distribution. We wish to evaluate the expectation for some function of interest $f(\cdot)$ which requires evaluation of

$$E\left[f\left(X\right)\right]=\frac{\int f\left(x\right)\pi\left(x\right)dx}{\int\pi\left(x\right)dx} \tag{1.24}$$

Often, we know $\pi\left(x\right)$ up to a constant of normalisation, which means we are unable to evaluate $\int\pi\left(x\right)dx$.

Monte Carlo integration evaluates $E\left[f\left(X\right)\right]$ by drawing samples $X_t, t=1,...,n$ from $\pi\left(\cdot\right)$ then approximates the population mean $f\left(X\right)$ with the sample mean

$$E\left[f\left(X\right)\right]=\frac{1}{n}\sum_{t=1}^{n}f\left(X_t\right) \tag{1.25}$$

The accuracy of the approximation can be increased simply by increasing n, given the samples are independent (laws of large numbers). MCMC methods draw these samples by running a cleverly constructed Markov chain for a long time. This results in samples which are not independent; however, this dependency is not an issue so long as the samples are drawn in the correct proportions throughout the support of $\pi\left(\cdot\right)$. This occurs when $\pi\left(\cdot\right)$ is chosen to be the stationary distribution of the Markov chain (W. Gilks, Richardson, and Spiegelhalter, 1996).

1.5.1.2 Markov chains

A first order Markov chain is a sequence of random variables $X_0, X_1, X_2, ...$ where the next state X_{t+1} is sampled from a distribution $P\left(X_{t+1}|X_t\right)$ which depends only on the current state of the chain, X_t. The chain is time-homogenous if the transition kernel $P\left(\cdot|\cdot\right)$ does not change over time i.e. it does not depend on t

$$T_m\left(X_m|X_{m+1}\right) \equiv P\left(X_{m+1}|X_m\right); T_1 = T_2 = ... = T_M = T \tag{1.26}$$

An ergodic Markov Process converges to its unique stationary distribution $\phi\left(\cdot\right)$, so long as certain conditions hold. Thus, after removing m burn-in iterations, the remaining points $X_t; t = m+1, ..., n$ are dependent samples approximately from $\phi\left(\cdot\right)$. These points can then be used to estimate the expectation $E\left[f\left(X\right)\right]$, where $X \sim \phi\left(\cdot\right)$. Removing burn-in iterations modifies the estimating equation for the expectation to

$$\bar{f} = \frac{1}{n-m}\sum_{t=m+1}^{n} f\left(X_t\right) \tag{1.27}$$

which is now called an ergodic average.

The ergodic theorem defines the set of conditions which must hold to guarantee that the resulting Markov chain converges to a limiting distribution $\pi\left(\cdot\right)$; therefore, converging to the required expectation $E\left[f\left(X\right)\right]$, regardless of the starting value X_0. The conditions are that the Markov chain is aperiodic, irreducible, and positive recurrent (W. Gilks, Richardson, and Spiegelhalter, 1996). Aperiodicity guarantees that the chain does not have a regular pattern determining when it returns to particular values. Irreducibility means that the chain can get between all possible states (possibly in several steps). Restricting the chain to be positive recurrent means that the expected time till returning to a given state is finite. These conditions are sufficient to guarantee that samples from the Markov chain are equivalent to a standard independent, identically distributed, simulation from the target distribution, with some loss of efficiency due to the dependent nature of the draws.

To guarantee that the samples from our chosen Markov chain meet the above criteria, we must determine when the chain has converged to $\pi(\cdot)$; that is, the number of burn-in states m to remove from the beginning of the chain. We also need to determine when we have sampled a sufficient number of states n to accurately approximate the required integral with the available expectation approximation. This can be very difficult to do in practice, especially when the chain is poorly mixing i.e. generating highly dependent samples. Two common methods include comparing parallel chains with highly dispersed starting values X_0, and running a single chain for a very long time and comparing multiple sub-sections of that chain to each other. Both of these methods can detect lack of convergence when differences are observed, but are not sufficient to conclude convergence. Some more formal convergence diagnostic methods are discussed in Roy (2019).

When using MCMC algorithms to fit a Bayesian model, the target distribution is typically the posterior $P(\theta|D)$.

Constructing a Markov chain, such that the stationary distribution is our distribution of interest, is reasonably straightforward. One of the simplest methods is the Metropolis-Hastings (MH) algorithm, named after (Metropolis et al., 1953; Hastings, 1970)

1.5.2 Metropolis-Hastings samplers

The MH algorithm chooses the next state X_{t+1} by sampling a candidate point X' from a proposal distribution $q(.|X)$ which may depend on the current state X_t. For example, $q(.|X)$ may be a Normal distribution with mean X_t and fixed variance. The proposed new state is accepted with probability $\alpha(X, X')$ where

$$\alpha(X, X') = \mathsf{min}\left(1, \frac{\pi(X')\, q(X|X')}{\pi(X)\, q(X'|X)}\right) \tag{1.28}$$

The first part of the acceptance equation $\pi\left(X'\right)/\pi\left(X\right)$ gives the ratio of probabilities of the proposed state X', and the current state X, under the target distribution $\pi\left(\cdot\right)$. The second part $q\left(X|X'\right)/q\left(X'|X\right)$ adjusts the acceptance rate for differing proposal distributions, and is often called a correction factor.

If the proposed step is accepted, the next state becomes $X_{t+1} = X'$, otherwise the current state is retained $X_{t+1} = X_t$. An easy way to choose whether to accept the move, given the acceptance probability $\alpha\left(X_t, X'\right)$, is to draw a random value $u \sim \mathsf{Uniform}\left(0, 1\right)$ and accept the proposed move only if $u < \alpha\left(X_t, X'\right)$.

The following shows the algorithm to draw M samples using the Metropolis-Hastings sampler:

```
Set $t = 0$;
Generate an initial state $X_0 \sim \pi_0$;
while $t < M$ do
    Set $t = t + 1$;
    Generate a proposal state $X'$ from $q\left(X|X_{t-1}\right)$;
    Calculate the proposal correction factor $c = \frac{q(X_{t-1}|X')}{q(X'|X_{t-1})}$;
    Calculate acceptance probability $\alpha = \mathsf{min}\left(1, \frac{\pi(X')}{\pi(X_{t-1})} \times c\right)$;
    Sample a Uniform(0, 1) random variable u;
    if $u \leq \alpha$ then
        set $X_t = X'$
    end
    else
        set $X_t = X_{t-1}$
    end
end
```

Algorithm 3: Metropolis-Hastings algorithm.

The proposal or candidate probability density function (pdf) q can have any form so long as it has wide enough support to eventually reach any region of the state space $\mathcal{X}$ with a positive mass under $\pi\left(\cdot\right)$.

To satisfy detailed balance, thus guaranteeing convergence to the target distribution $\pi(\cdot)$, the transition operator T must satisfy the following condition

$$\pi(X)\,T(X,X') = \pi(X')\,T(X',X) \tag{1.29}$$

where

$$T(X,X') = q(X'|X)\,\alpha(X,X') \tag{1.30}$$

for MH samplers.

Substituting equations 1.30 and 1.28 into equation 1.29 shows that detailed balance holds for MH samplers.

$$\begin{aligned}
\pi(X)\,T(X,X') &= \pi(X)\,q(X'|X)\,\alpha(X,X') \\
&= \mathsf{min}\left(\pi(X)\,q(X'|X),\pi(X')\,q(X|X')\right) \\
&= \mathsf{min}\left(\pi(X')\,q(X|X'),\pi(X)\,q(X'|X)\right) \\
&= \pi(X')\,q(X|X')\,\alpha(X',X) \\
&= \pi(X')\,T(X',X)
\end{aligned} \tag{1.31}$$

When using a vanilla Metropolis-Hastings sampler, the algorithm always accepts proposals which take the chain closer to a local mode. It also accepts moves which would take it further away with probability exactly equal to the relative “heights” of the target distribution at the proposed and current values. This results in the sampler spending most of its time exploring the main mass of the posterior.

Although any arbitrary proposal distribution can be used, the algorithms performance is highly dependent on the choice of q. A poorly chosen proposal distribution can result in slow convergence and/or slow exploration of the support of $\pi(\cdot)$ (i.e. poor mixing). Additionally, it is advantageous to choose a proposal distribution which can be easily sampled and evaluated for computational efficiency.

1.5.2.1 Gibbs sampler

The Gibbs sampler is a special case of the MH algorithm where a single component is updated by sampling from the full conditional distribution. This results in an acceptance probability of 1. The proposal distribution for updating the ith component of X is

$$q_i\left(X'|X\right) = \pi\left(X'|X\right) = \pi\left(X'\right) \tag{1.32}$$

where $\pi\left(X'|X\right)$ is the full conditional distribution. Substituting equation 1.32 into equation 1.28 gives an acceptance ratio of 1.

1.5.2.2 Metropolis sampler

The Metropolis sampler is another special case of the MH algorithm when symmetric proposal distributions are used $q\left(X'|X\right) = q\left(X|X'\right)$.

This simplifies the acceptance probability as the correction factor c is now 1

$$\alpha\left(X, X'\right) = \mathsf{min}\left(1, \frac{\pi\left(X'\right)}{\pi\left(X\right)}\right) \tag{1.33}$$

1.5.2.3 Random walk Metropolis and Random walk Metropolis-Hastings samplers

The single most widely used subclass of MCMC algorithms is based around the random walk Metropolis (RWM) (Sherlock, Fearnhead, and Roberts, 2010). It is a special case of the Metropolis sampler where the proposal is generated using a random walk $q(X'|X) = q(X|X')$. For example, a multivariate normal distribution with mean X_t and constant covariance matrix Σ, is a common proposal distribution for continuous X.

The Random walk Metropolis-Hastings (RWMH) algorithm extends the RWM algorithm by allowing non-symmetric proposal distributions. This can be useful when the values of X are restricted; for example, when X must be positive, as it can be constructed to never propose transitions to negative values of X. For example, a log-normal distribution with mean X_t and constant variance σ^2 for continuous X.

When using the RWM and RWMH algorithms, particular attention needs to be given to choosing an appropriate scale for the proposal distribution. To explore the target distribution efficiently, the proposal distribution needs to propose jumps of an appropriate size compared with some measure of scale of the target distribution. Jumps that are too small are almost always accepted, but result in slow exploration of the target distribution. Large jumps, which aim to move about the distribution more quickly, are often rejected stranding the chain in a single location for many iterations.

The most efficient scale parameter gives an acceptance rate of between 40 and 45% for a single parameter, and around 20 to 30% for a block update (Roberts, Gelman, and W.R. Gilks, 1997). Although this result specifically relates to a spherically symmetric proposal distribution targeting a unimodal elliptically symmetric distribution with components which are i.i.d up to a scale parameter, it works reasonably well in many situations (C.P. Robert, 2015). Tuning of the scale parameter can be done using pilot runs of the algorithm (which must be repeated for every new algorithm, and every new data set), or adapted on the fly.

1.5.2.4 Adaptive Metropolis-Hastings sampler

An adaptive MCMC approach can be used to automatically tune the proposal distribution for random walk Metropolis-Hastings algorithms. Adaptive MCMC is a technique which allows the algorithm to adaptively choose optimal parameters for a given proposal distribution during a single run. This is done by gradually adjusting the proposal distribution parameters such that the acceptance rate moves closer to the optimal values described above. Diminishing adaptation, where the size of the changes to the kernel tend to 0 as the number of iterations n tends to infinity, is a critical feature of this algorithm. Without this, the algorithm may produce a sample which does not converge to a stationary distribution.

There are several algorithms which adapt the kernel; however, here we will describe the version described by Roberts and Rosenthal (2009) and Sherlock, Fearnhead, and Roberts (2010). For a block random walk of dimension d with adaption, a new value at time t is sampled from a mixture of an adaptive Gaussian

$$x^* \sim \mathsf{Normal}\left(0, \frac{1}{d}2.38^2\hat{\Sigma}_t\right) \tag{1.34}$$

and a non-adaptive Gaussian distribution, where $\hat{\Sigma}_t$ is the variance matrix calculated from the previous $t-1$ iterations. This method satisfies the diminishing adaptation requirements as the variance matrix changes are $O\left(1/t\right)$ at the t^{th} iteration. Sampling from a mixture of the adaptive and non-adaptive kernels is important. This ensures that the chain does not get stuck in a region of the posterior; i.e. causing the chain to never visits some regions of high density due to the sampler adapting to suggest only small steps.

1.5.2.5 Block updates

The above algorithms (with the exception of the Gibbs sampler) can be implemented as single site or block updates. Single site updates iteratively update one parameter at a time, each with its own proposal distribution. Block samplers update a group of parameters jointly using a multivariate proposal density. Block updates can improve mixing when highly correlated components are jointly updated, but this depends on the choice of proposal (W. Gilks, Richardson, and Spiegelhalter, 1996).

1.5.3 Hamiltonian Monte Carlo and the No-U-turn Sampler

Random walk Metropolis-Hastings algorithms tend to scale poorly with increasing dimension, and complexity of the target distribution (particularly when it is non-orthogonal), and often require an unacceptably long time to converge to the target distribution (Betancourt, 2017). This is because it is particularly difficult to propose sensible values in high dimensional posterior spaces. As the number of dimensions increases, the volume exterior to the typical set (high probability region of the posterior distribution) overwhelms the volume interior to the typical set. This results in almost every Random Walk Metropolis proposal producing a point on the outside the typical set (towards the tails) (Betancourt, 2017). These proposals are usually rejected as the acceptance probability is negligible due to the extremely low density of these points. The low acceptance probability can be increased by choosing very small step sizes for proposals. However, both options result in Markov chains that barely move and take excessive time to adequately visit the entire posterior space (Betancourt, 2017).

Rather than proposing high dimensional jumps randomly (using the random walk), we can use information about the geometry of the typical set to guide the proposals (Betancourt, 2017). Hamiltonian Monte Carlo (HMC, (Duane et al., 1987)) is a type of

Metropolis-Hastings sampler that uses Hamiltonian dynamics to propose jumps using auxiliary variables. By utilising gradient information from the targeted posterior, it avoids the inefficiency of random walk behaviour and the associated sensitivity to correlated parameters (Hoffman and Gelman, 2014). This allows it to more efficiently converge to high-dimensional target distributions; however, it is extremely sensitive to the tuning parameters (step size and number of steps) and requires information about the gradient of the log-posterior. Gradient information can be estimated using automatic differentiation engines, such as Theano and Tensorflow. The step size can be easily adapted (using adaptive MCMC techniques described above or dual averaging (Nesterov, 2009)), and the number of steps can be tuned using the no-U-turn sampler (NUTS, (Hoffman and Gelman, 2014)) algorithm.

The NUTS algorithm transformed the usability of HMC by estimating appropriate tuning values on-the-fly, removing the need for user intervention and costly tuning runs. This allowed HMC to be incorporated into several general purpose inference engines such as Stan and PyMC3. These packages have greatly reduced the technical expertise and time required to fit complex Bayesian models. The geostatistical generalised linear model used in chapter 3 was fitted using PyMC3 using the NUTS for all parameters. This fitting algorithm is particularly useful as models incorporating Gaussian processes have large correlated sample spaces which can be difficult to traverse efficiently using many other sampling techniques, such as RWMH.

1.6 Thesis outline

Chapter 2 develops a joint Bayesian non-parametric source attribution model, which uses surveillance data from human cases of campylobacteriosis and food source samples, to estimate the proportion of campylobacteriosis cases associated with each food source. The model measures the force of infection from each source, allowing for varying survivability, pathogenicity, and virulence of pathogen strains using a Dirichlet Process; and varying abilities of the sources to act as vehicles of infection.

Chapter 3 expands the scope of zoonotic disease attribution to assess the effect of many potential risk factors on the probability of leptospirosis infection in urban slums. The Bayesian non-parametric model considers risk factors for leptospirosis in an urban slum in Salvador, Brazil, using a Gaussian Process to estimate rat exposure throughout the study area.

Chapter 4 develops a Bayesian dynamical model which captures the essential aspects of temporal antibody concentration changes in sheep after vaccination using antibody titre data. This model can be used to estimate time of infection for an individual, and may be extended to estimate disease status.

In chapter 5, the challenges and limitations common to these 3 projects are discussed in depth, and future work is proposed to address them. Note, code for the source attribution model in chapter 2 can be found at `https://CRAN.R-project.org/package=sourceR`, whilst code for the models in chapters 3 and 4 can be found at `https://gitlab.com/poppymiller/phd-thesis-related-code`.

Bibliography

Lotka, A.J., 1910. Contribution to the Theory of Periodic Reaction. *J. phys. chem.*

de Finetti, B., 1931. Funzione caratteristica di un fenomeno aleatorio. *Atti della r. academia nazionale dei lincei.*

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E., 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics* [Online], 21(6), pp.1087–1092. eprint: `https://doi.org/10.1063/1.1699114`. Available from: `https://doi.org/10.1063/1.1699114`.

Hastings, W.K., 1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika* [Online], 57(1), pp.97–109. Available from: `http://www.jstor.org/stable/2334940`.

Duane, S., Kennedy, A.D., Pendleton, B.J., and Roweth, D., 1987. Hybrid monte carlo. *Physics letters b* [Online], 195(2). Available from: `https://doi.org/https://doi.org/10.1016/0370-2693(87)91197-X`.

Sethuraman, J., 1994. A constructive definition of dirichlet priors. *Statistica sinica* [Online], 4(2), pp.639–650. Available from: `http://www.jstor.org/stable/24305538`.

Gilks, W., Richardson, S., and Spiegelhalter, D., eds, 1996. *Markov chain monte carlo in practice* [Online]. New York: Chapman and Hall/CRC. Available from: `https://doi.org/https://doi.org/10.1201/b14835`.

Blaser, M., 1997. Epidemiologic and clinical features of *Campylobacter jejuni* infections. *Journal of infectious diseases.*

Roberts, G.O., Gelman, A., and Gilks, W.R., 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. appl. probab.* [Online], 7(1) (), pp.110–120. Available from: `https://doi.org/10.1214/aoap/1034625254`.

Diggle, P.J., Tawn, J.A., and Moyeed, R.A., 1998. Model-based geostatistics. *Journal of the royal statistical society: series c (applied statistics)* [Online], 47(3), pp.299–350. eprint: `https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9876.00113`. Available from: `https://doi.org/10.1111/1467-9876.00113`.

Nachamkin, I., Allos, B., and Ho, T., 1998. *Campylobacter* species and Guillain-Barre syndrome. *Clinical microbiology reviews.*

World Health Organisation, 2003. *Human leptospirosis guidence for diagnosis, surveillance and control.* WHO Library Cataloguing-in-Publication Data.

Zia, S., Wareing, D., Sutton, C., Bolton, E., Mitchell, D., and Goodacre, J., 2003. Health problems following *Campylobacter jejuni* enteritis in a lancashire population. *Rheumatology.*

McBride, A., Athanazio, D., Reis, M., and Ko, A.I., 2005. Leptospirosis. *Current opinion on infectious diseases.*

Diggle, P.J. and Ribeiro, P.J., 2007. *Model-based geostatistics*, Springer Series in Statistics. Springer-Verlag New York.

Humphrey, T., O'Brien, S., and Madsen, M., 2007. Campylobacters as zoonotic pathogens: a food production perspective. *International journal of food microbiol.*

Wang, Z., Jin, L., and Wegrzyn, A., 2007. Leptospirosis vaccines. *Microbial cell factories.*

Whye Teh, Y., 2007. *Dirichlet process.* (Technical report). University College London.

Izurieta, R., Galwankar, S., and Clem, A., 2008. Leptospirosis: the mysterious mimic. *Journal of emerging trauma and shock.*

Schmidt, A., Conceição, M., and Alberti Moreira, G., 2008. Investigating the sensitivity of gaussian processes to the choice of their correlation function and prior specifications. *Journal of statistical computation and simulation* [Online], 78 (), pp.681–699. Available from: https://doi.org/10.1080/00949650701231983.

Nesterov, Y., 2009. Primal-dual subgradient methods for convex problems. *Math. program.* [Online], 120(1) (), pp.221–259. Available from: https://doi.org/10.1007/s10107-007-0149-x.

Paisley, J., 2009. *A tutorial on the dirichlet process for engineers.* (Technical report). Duke University.

Roberts, G.O. and Rosenthal, J.S., 2009. Examples of adaptive mcmc. *Journal of computational and graphical statistics* [Online], 18(2), pp.349–367. eprint: https://doi.org/10.1198/jcgs.2009.06134. Available from: https://doi.org/10.1198/jcgs.2009.06134.

Frigyik, B., Kapila, A., and Gupta, M., 2010. *Introduction to the dirichlet distribution and related processes.* (Technical report). University of Washington.

Sherlock, C., Fearnhead, P., and Roberts, G.O., 2010. The random walk metropolis: linking theory and practice through a case study. *Statist. sci.* [Online], 25(2) (), pp.172–190. Available from: https://doi.org/10.1214/10-STS327.

Robert, C. and Casella, G., 2011. A short history of markov chain monte carlo: subjective recollections from incomplete data. *Statist. sci.* [Online], 26(1) (), pp.102–115. Available from: https://doi.org/10.1214/10-STS351.

World Health Organisation, 2011. *Report of the second meeting of the leptospirosis burden epidemiology reference group*. WHO Library Cataloguing-in-Publication Data.

Tam, C., Rodrigues, L., Viviani, L., Dodds, J., Evans, M., Hunter, P., Gray, J., Hetley, L., Rait, G., Tompkins, D., and O'Brien, S., 2012. Longitudinal study of infectious intestinal disease in the uk (iid2 study): incidence in the community and presenting to general practice. *Gut*.

World Health Organisation, 2012. *The global view of campylobacteriosis: report of an expert consultation* [Online]. WHO. Available from: `http://apps.who.int/iris/bitstream/10665/80751/1/9789241564601-eng.pdf`.

Brownowski, C., James, C., and Winstanley, C., 2014. Role of environmental survival in transmission of *Campylobacter jejuni*. *Fems microbiology letters*.

Hoffman, M.D. and Gelman, A., 2014. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of machine learning research* [Online], 15, pp.1593–1623. Available from: `http://jmlr.org/papers/v15/hoffman14a.html`.

Costa, F., Hagan, J., Calcagno, J., Kane, M., Torgerson, P., Martinez-Silveira, M.S., Stein, C., Abela-Ridder, B., and Ko, A., 2015. Global morbidity and mortality of leptospirosis: a systematic review. *Plos neglected tropical diseases*.

Haake, D. and Levett, P., 2015. Leptospirosis in humans. *Current topics in microbial immunology*.

McGoff, K., Mukherjee, S., Nobel, A., and Pillai, N., 2015. Consistency of maximum likelihood estimation for some dynamical systems. *The annals of statistics* [Online], 43(1), pp.1–29. Available from: `http://www.jstor.org/stable/43556506`.

Robert, C.P., 2015. *The metropolis-hastings algorithm*. arXiv: `1504.01896 [stat.CO]`.

World Health Organization, 2015. *Who estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015* [Online]. available on the WHO web site (www.who.int) or can be purchased from WHO Press, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland. Available from: `http://apps.who.int/iris/bitstream/10665/199350/1/9789241565165_eng.pdf?ua=1`.

Goarant, C., 2016. Leptospirosis: risk factors and management challenges in developing countries. *Res rep trop med*.

Betancourt, M., 2017. *A conceptual introduction to hamiltonian monte carlo*. arXiv: `1701.02434 [stat.ME]`.

World Heath Organisation, 2018. *Global Health Estimates 2016 Summary Tables*. (Technical report). World Heath Organisation.

Xu, Y. and Ye, Q., 2018. Human leptospirosis vaccines in china. *Human vaccines & immunotherapeutics* [Online], 14(4). PMID: 29148958, pp.984–993. eprint: `https://doi.org/10.1080/21645515.2017.1405884`. Available from: `https://doi.org/10.1080/21645515.2017.1405884`.

Roy, V., 2019. *Convergence diagnostics for markov chain monte carlo*. arXiv: `1909.11827 [stat.CO]`.

Chapter 2

sourceR: Classification and Source Attribution of Infectious Agents among Heterogenous Populations

Poppy Miller[1,*], Jonathan Marshall[2,3], Nigel French[3,4,5], Chris Jewell[1]

1 CHICAS, Faculty of Health and Medicine, Lancaster University, Lancaster, England

2 Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

3 mEpiLab, Massey University, Palmerston North, New Zealand

4 New Zealand Food Safety Science and Research Centre

5 New Zealand Institute for Advanced Studies

Abstract

Zoonotic diseases are a major cause of morbidity, and productivity losses in both human and animal populations. Identifying the source of foodborne zoonoses (e.g. an animal reservoir or food product) is crucial for the identification and prioritisation of food safety interventions. For many zoonotic diseases it is difficult to attribute human cases to sources of infection because there is little epidemiological information on the cases. However, microbial strain typing allows zoonotic pathogens to be categorised, and the relative frequencies of the strain types among the sources and in human cases allows inference on the likely source of each infection. We introduce sourceR, an R package for quantitative source attribution, aimed at foodborne diseases. It implements a Bayesian model using strain-typed surveillance data from both human cases and source samples, capable of identifying important sources of infection. The model measures the force of infection from each source, allowing for varying survivability, pathogenicity and virulence of pathogen strains, and varying abilities of the sources to act as vehicles of infection. A Bayesian non-parametric (Dirichlet process) approach is used to cluster pathogen strain types by epidemiological behaviour, avoiding model overfitting and allowing detection of strain types associated with potentially high “virulence”.

sourceR is demonstrated using *Campylobacter jejuni* isolate data collected in New Zealand between 2005 and 2008. Chicken from a particular poultry supplier was identified as the major source of campylobacteriosis, which is qualitatively similar to results of previous studies using the same dataset. Additionally, the software identifies a cluster of 9 multilocus sequence types with abnormally high ’virulence’ in humans.

sourceR enables straightforward attribution of cases of zoonotic infection to putative sources of infection. As sourceR develops, we intend it to become an important and flexible resource for foodborne disease attribution studies.

2.1 Introduction

Zoonotic diseases are a major source of human morbidity world wide. In 2010, there were an estimated 600 million cases globally (Havelaar et al., 2015), of which 96 million were *Campylobacter spp.* resulting in 21,000 deaths (World Health Organization, 2015). Attributing cases of foodborne disease to putative sources of infection is crucial for identifying and prioritising food safety interventions, prompting routine national recording of human cases and surveillance of high-risk sources in many countries – for example FoodNet in the

US (Pires et al., 2009), the Danish Zoonosis Centre (`food.dtu.dk`), and the Ministry for Primary Industries in New Zealand (`foodsafety.govt.nz`).

Traditional approaches to source attribution include observational risk assessment, extrapolation of surveillance or outbreak data, and epidemiological field studies (Crump, Griffin, and Angulo, 2002). The results of such direct observational methods may be highly uncertain due to long and variable disease incubation times, and many exposures of an individual to multiple sources of infection. Nevertheless, statistical modelling of human case count data, incorporating molecular strain typing of pathogen isolates from national surveillance programmes, has shown promise for identifying important sources of foodborne illness (Hald et al., 2004; Müllner, Jones, et al., 2009).

The aim of this paper is to extend current approaches to statistical source attribution, and to provide a standard software package, `sourceR`, providing an intuitive interface to source attribution models for epidemiological domain specialists. Our principle innovation is a novel class of Bayesian non-parametric source attribution model which classifies strain types by differential epidemiological behaviour and accurately quantifies uncertainty. Furthermore, we allow for spatial and temporal heterogeneity in case and source data with the aim of detecting differential exposures to infection sources across space and time. `sourceR` represents the first standard software for source attribution, and is designed for use by epidemiologists and public health decision makers. It is written as an add-on package to `R`, the open-source lingua-franca for modern epidemiological analysis, and incorporates an object-orientated style to facilitate further model development and future maintainability.

The paper is structured as follows. We first introduce a motivating example and review existing source attribution models. The new model is described in the Design and Implementation section followed by a demonstration of model fitting using `sourceR` in the

Materials and Methods section. Results and Discussion sections follow, and it concludes with details of Availability and Future directions.

2.1.1 Example: *Campylobacter* food-poisoning in Manawatu, New Zealand

In 2006, New Zealand had one of the highest incidences of campylobacteriosis in the developed world, with an annual incidence in excess of 400 cases per 100,000 people (Baker et al., 2006). Our motivating data set was collected between 2005 and 2008 in the Manawatu region of New Zealand with the aim of identifying the most important sources of campylobacteriosis and implementing interventions. A campaign to change poultry processing procedures, supported in part by results from previous quantitative source attribution approaches, was successful in leading to a sharp decline in campylobacteriosis incidence after 2007 (Müllner, Jones, et al., 2009).

Campylobacter has many subtypes which are usually defined using Multilocus Sequence Typing (MLST), a commonly used genotyping method providing a relatively rapid method of characterising isolates. An MLST sequence type is a unique combination of alleles at specified gene loci, typically located in conserved regions of the genome (Dingle et al., 2001; Urwin and Maiden, 2003). The data set consists of the dominant MLST-genotype *Campylobacter* isolated from each source (potential food and environmental sources) and human sample. The data was first published in Müllner, Collins-Emerson, et al. (2010), and is described in detail (including data collection methods in French and Marshall (2009) and French and Marshall (2013)). These data are included in our `sourceR` package (named `campy`). We use this data set as a case study, and compare our results with previously published statistical approaches.

2.1.2 Existing methods of source attribution

The general structure of the source attribution model is that the observed case-counts y_i for strain i (occurring in a defined surveillance period) are mutually independent Poisson distributed with means

$$\lambda_i = \sum_{j=1}^{m} \alpha_j p_{ij}. \tag{2.1}$$

where p_{ij} is the prevalence of strain i in source j, and "source effects" $\boldsymbol{\alpha}$ measure each source's capacity to act as a vehicle of infection. The estimated number of cases attributed to a particular source j is

$$\hat{\xi}_j = \hat{\alpha}_j \sum_{i=1}^{n} p_{ij}. \tag{2.2}$$

Comparing the relative magnitudes of $\hat{\xi}_j$ provides a statistical method to prioritise intervention strategies to the most important sources of infection. The model is fitted in a Bayesian framework as posteriors for functions of parameters (such as ξ) are easily calculated, and to allow previous knowledge to be incorporated via informative priors.

A significant problem is that this model does not allow for some strain types have differential affinities for human infection resulting in over-dispersion of $\boldsymbol{y}$. Additionally, it does not allow for uncertainty in $\boldsymbol{P}$, inherent in sample based source data. In the rest of this section, we review current extensions to Equation 2.1 aimed at accounting for the Poisson over-dispersion in observed case numbers, and incorporating uncertainty in source surveillance data. In particular, the preliminary developments made by Hald et al. (2004) and Müllner, Jones, et al. (2009) form an ontology on which we base our innovations.

2.1.2.1 Over-dispersion

Hald et al. (2004) address the issue of Poisson over-dispersion in Equation 2.1 by introducing a "type effect" $\boldsymbol{q}$ accounting for some strain types being more adapted to human infection than others.

$$\lambda_i \quad = \quad q_i \sum_{j=1}^{m} \alpha_j c_j p_{ij}. \tag{2.3}$$

Additionally, they include an offset $\boldsymbol{c}$ representing known rates of consumption of each source foodstuff, allowing $\boldsymbol{\alpha}$ to be interpreted as a source-specific factor independent of exposure. However, the addition of $\boldsymbol{q}$ as a vector of uncorrelated unknowns over-specifies the model, with $m+n$ parameters but only n independent disease case count observations. Hald *et al.* therefore reduce the number of parameters by heuristic *a priori* grouping of the elements of $\boldsymbol{q}$, albeit with the generally undesirable property that quantification of uncertainty in the most appropriate choice of grouping is not readily permissible.

The "Modified Hald" model of Müllner, Jones, et al. (2009) treats $\boldsymbol{q}$ as log Normally distributed random effect, with unit mean and unknown variance τ^2

$$q_i \sim \mathrm{logNormal}(1, \tau^2) \tag{2.4}$$

with a Gamma-distributed prior distribution imposed on τ^2. However, this approach suffers from *a posteriori* non-identifiability of $\boldsymbol{q}$ and τ^2, hindering the performance of MCMC algorithms used to fit the model (Gelfand, Sahu, and Carlin, 1995). Though this may be ameliorated by choosing an informative prior for τ^2 with small mean, it results in severe shrinkage of $\boldsymbol{q}$ and inference which is sensitive to the choice of prior.

2.1.2.2 Uncertainty in source sampling

The Modified Hald model introduces uncertainty into the prevalences p_{ij} by modelling the source sampling process. Let s_j denote the total number of source samples collected from source $j = 1, \ldots, m$, of which x_{ij} are positive for pathogen type i. Normalisation of the number of positive samples x_{ij} gives the relative prevalence $r_{ij} = x_{ij}/\sum_{i=1}^{n} x_{ij}$ of type i in source j. The relative prevalence r_{ij} is then combined with the prevalence of positive samples $k_j = \sum_{i=1}^{n} x_{ij}/s_j$ to calculate the absolute prevalence $p_{ij} = r_{ij} \times k_j$ of strain i in source j. The Modified Hald model was fitted in WinBUGS using an approximate two stage process (Müllner, Jones, et al., 2009). First, a posterior distribution was estimated for the absolute prevalence of source types $\boldsymbol{p}$, using the model specified in Eqs 2.5 and 2.6 :

$$r_{\cdot j} \sim \mathsf{Dirichlet}(\mathbf{1}) \; \forall \; j \tag{2.5}$$

$$k_j \sim \mathsf{Beta}(1,1) \; \forall \; j \tag{2.6}$$

The marginal posterior for each element of $\boldsymbol{p}$ was then approximated by a Beta distribution

$$p_{ij} \sim \mathsf{Beta}(w_{ij}, v_{ij})$$

(using the method of moments to calculate w_{ij} and v_{ij}) which was then used as an independent prior in step 2. Since each isolate is assigned to only one type, we must observe $\sum_{i=1}^{n} r_{ij} = 1$, and therefore $\sum_{i=1}^{n} p_{ij} = k_j$. This is not enforced when using independent Beta priors for each p_{ij} which results in k_j (the probability of a sample being positive given the sample is from source j) no longer being constrained to be between 0 and 1.

2.2 Design and Implementation

Our approach addresses the deficiencies inherent in both the Hald and Modified Hald models by fitting a joint model for both source and human case sampling with non-parametric clustering of the type effects. This allows integration over uncertainty in the source sampling process without resorting to an approximate marginal probability distribution on $\boldsymbol{p}$.

The over-dispersion is solved by non-parametrically clustering the pathogen types using a Dirichlet process (DP) on the type effect vector $\boldsymbol{q}$. This is a data driven, automatic method which reduces the effective number of parameters in the model without requiring strong assumptions about τ^2 in Equation 2.4. Additionally, it quantifies the similarity between epidemiological characteristics (virulence, pathogenicity and survivability) of the subtypes forming the basis of future research on the genetic determinants of this behaviour. Often, human case data is associated with location such as urban/rural, or GPS coordinates whilst food samples are likely to be less spatially constrained (due to distances between production and sale locations). Both human and source data may exist for multiple time-periods. Therefore, we allow for spatial and temporal heterogeneity in the data.

2.2.1 HaldDP Model

As with the Hald model, we assume the number of human cases y_{itl} identified by isolation of subtype i in time-period t at location l is Poisson distributed

$$y_{itl} \sim \mathsf{Poisson}(\lambda_{itl} = q_i \sum_{j=1}^{m} \alpha_{jtl} p_{ijt}) \tag{2.7}$$

We allow for different exposures of humans to sources in different locations and times, by allowing the source effects to vary between times and locations, α_{jtl}.

For each source j, we model the number of positive source samples

$$\boldsymbol{x}_{jt} \sim \text{Multinomial}(s_{jt}^{+}, \boldsymbol{r}_{jt}) \tag{2.8}$$

where $\boldsymbol{x}_{jt} = (x_{ijt}, i = 1, ..., n)^T$ denotes the vector of type-counts in source j in time-period t, $s_{jt}^{+} = \sum_{i=1}^{n} x_{ijt}$ denotes the number of positive samples obtained, and $\boldsymbol{r}_{jt}$ denotes a vector of relative prevalences $Pr\left(\mathsf{type}_i | \mathsf{source}_j, \mathsf{time}_t\right)$. This automatically places the constraint

$\sum_{i=1}^{n} r_{ijt} = 1$. The source case model is then coupled to the human case model through the simple relationship

$$p_{ijt} = r_{ijt} k_{jt} \tag{2.9}$$

where k_{jt} is the prevalence of any isolate in source j in time-period t.

In principle, a Beta distribution could be used to model k_{jt}, arising as the conjugate posterior distribution of a Binomial sampling model for s_{jt}^{+} positive samples from s_{jt} tested, and a Beta prior on k_{jt}. We instead choose to fix the source prevalences at their empirical estimates ($k_{jt} = s_{jt}^{+}/s_{jt}$) because the number of source samples is typically high.

The type effects $\boldsymbol{q}$, which are assumed invariant across time or location, are drawn from a DP with base distribution Q_0 and a concentration parameter a_q

$$q_i \sim \text{DP}\left(a_q, Q_0\right). \tag{2.10}$$

The Dirichlet process is a probability distribution whose range is a set of probability distributions and is defined by a base distribution and concentration parameter (Ferguson, 1973). The concentration parameter of the DP a_q encodes prior information on the number of groups K to which the pathogen types are assigned. The Gamma base distribution of the DP Q_0 induces a prior for the cluster locations. The DP groups the elements of $\boldsymbol{q}$ into a finite set of clusters $1 : \kappa$ (unknown *a priori*) with values $\theta_1, ..., \theta_\kappa$ (drawn from the Gamma base distribution Q_0) which addresses the inevitable over-dispersion in the case counts $\boldsymbol{y}$ robustly and clusters subtypes into groups with similar epidemiological behaviour.

Heterogeneity in the source matrix $\boldsymbol{x}$ is required to identify clusters from sources, which may not be guaranteed *a priori* due to the observational nature of the data collection.

2.2.2 Inference

This section describes how the model is fitted in a Bayesian context by first describing the McMC algorithm used to fit this model, then developing the prior model.

2.2.2.1 MCMC algorithm

The joint model over all unobserved and observed quantities is fitted using Markov chain Monte Carlo (MCMC, full details in Full McMC Algorithm). The source effects and relative prevalence parameters are updated using independent adaptive Metropolis-Hastings updates (Roberts and Rosenthall, 2006). The type effects $\boldsymbol{q}$ are modelled using a DP (Eq 2.10) with a Gamma base distribution $Q_0 \sim \text{Gamma}(a_\theta, b_\theta)$. The choice of a Gamma base distribution with the Poisson likelihood (Eq 2.7) permits the use of a marginal Gibbs strategy for efficient sampling from the posterior ditribution of $\boldsymbol{q}$. Each observation i is assigned to a cluster k with value θ_k, such that $q_i \mapsto \theta_k$. The algorithm proceeds by alternately sampling from the posterior of the group assignments (adding new clusters or deleting empty clusters as necessary), and the posterior of θ_k for each cluster.

2.2.2.2 Priors

The parameters $\boldsymbol{\alpha}_{tl}$ and $\boldsymbol{q}$ account for a multitude of source and type specific factors which are difficult to quantify *a priori*. Therefore, with no single real-world interpretation, the distributional form of the priors were chosen for their flexibility. A Dirichlet prior is placed on each $\mathbf{r}_{jt}$ which suitably constrains the individuals r_{ij}s such that $\sum_{i=1}^{n} r_{ijt} = 1$. A Dirichlet prior is also placed on each $\boldsymbol{\alpha}_{tl}$, with the constraint $\sum_{j=1}^{m} \alpha_{jtl} = 1$ aiding identifiability between the mean of the source and type effect parameters. In `sourceR`, the concentration parameter of the DP α_q is specified by the analyst as a modelling decision.

We note that the choice of base distribution Q_0 may have a stronger effect than anticipated due to the small size of the relative prevalence and source effect parameters. This can been seen by considering the marginal posterior for θ_k

$$\theta_k|\cdot \sim \mathsf{Gamma}\left(a_\theta + \sum_{i:S_i=k} y_i, b_\theta + \sum_{i:S_i=k}\sum_{j=1}^{m} \alpha_j \cdot p_{ij}\right)$$

The term $\sum_{i:S_i=k}\sum_{j=1}^{m} \alpha_j \cdot p_{ij}$ is very small (due to the Dirichlet priors on α and $\mathbf{r}_j$), which can result in even a fairly small rate parameter (b_θ) dominating.

2.2.3 Code implementation

Standard McMC packages (e.g. WinBUGS, Stan, PyMC3) cannot implement marginal Gibbs sampling for Dirichlet processes, necessitating a custom McMC framework (see section 'Extensibility'). We chose R as a platform because of its ubiquity in epidemiology, and advanced support for post-processing of McMC samples. Dependencies on other R packages are required, but these are installed automatically by R's package manager.

`sourceR` uses an object-oriented design, which allows separation of the model from the McMC algorithm. Internally, the model is represented as a directed acyclic graph (DAG) in which nodes are represented by an R6 class hierarchy. Generic adaptive Metropolis Hastings algorithms are attached to each parameter node, with the conditional independence properties of the DAG allowing automatic computation of the required (log) conditional posterior densities.

A difficulty with the DAG setup is the representation of the DP model on the type effects $\boldsymbol{q}$, since each update of the marginal Gibbs sampler requires structural alterations. Therefore, we subsume the entire DP into a single node, with a bespoke marginal Gibbs sampling algorithm written for our Gamma base-distribution and Poisson likelihood model.

2.3 Materials and methods

The case study below illustrates how the `sourceR` package is used in practice. We compare the results of our approach with results from the Modified Hald, Asymmetric Island (see Island model overview and D. Wilson et al. (2008) and D. Wilson (2016)), and the "Dutch" model (see Dutch model overview and Pelt et al. (1999)). The priors for our model were selected to be minimally informative. The prevalence k_j is calculated by dividing the number of positive samples by the total number of samples for each source. In the data below, we note that for several samples the MLST typing failed, with the number of positive samples exceeding the apparent total number of MLST-typed isolates. Assuming MLST typing fails independently of pathogen type, this does not bias our results.

The model fitting process begins by formatting the data, constructing the HaldDP model and setting the McMC parameters before running the algorithm using the `update()` method.

```
## Format data
y <- Y(data = campy$cases,   # Cases
  y = "Human", type = "Type", time = "Time", location = "Location")

x <- X(data = campy$sources, # Sources
  x = "Count", type = "Type", time = "Time", source = "Source")

k <- Prev(data = campy$prev, # Prevalences
  prev = "Value", time = "Time", source = "Source")

## Set priors
priors = list(a_theta = 0.01, b_theta = 0.00001, a_alpha = 1, a_r = 0.1)

## Construct model
```

```
my_model <- HaldDP(y = y, x = x, k = k, priors = priors, a_q = 0.1)

## Set mcmc parameters
my_model$mcmc_params(n_iter = 1000, burn_in = 10000, thin = 500)

## Run model
my_model$update()
```

The `sourceR` package provides methods to extract and subset the complex posterior, calculate medians and credible intervals (with three possible methods percentile, SPIn (Liu, Gelman, and Zheng, 2015), or Chen-Shao (Chen and Shao, 1991)) and plot a heatmap with a dendrogram showing the clustering of the type effects.

```
my_model$extract()
my_model$summary(alpha = 0.05, CI_type = "percentiles")
my_model$plot_heatmap()
```

2.4 Results

Figure 2.1 shows the the proportion of cases attributed to each source. The HaldDP model identified the highest proportion of human campylobacteriosis cases as coming from chicken produced by supplier A (a median of 67 percent of cases attributed). A further 11 percent were attributed to Chicken from poultry supplier B and 17 percent to Ovine. The median values for the proportion of cases attributed to each source are qualitatively similar between all models except the Dutch method.

To visualise how the DP has clustered the type effects, Gower's distance (Gower, 1971) is used to compute a dissimilarity matrix between all pairs of types. Figure 2.2 shows that the DP identified four main type clusters (from 91 types). The violin plots of the marginal

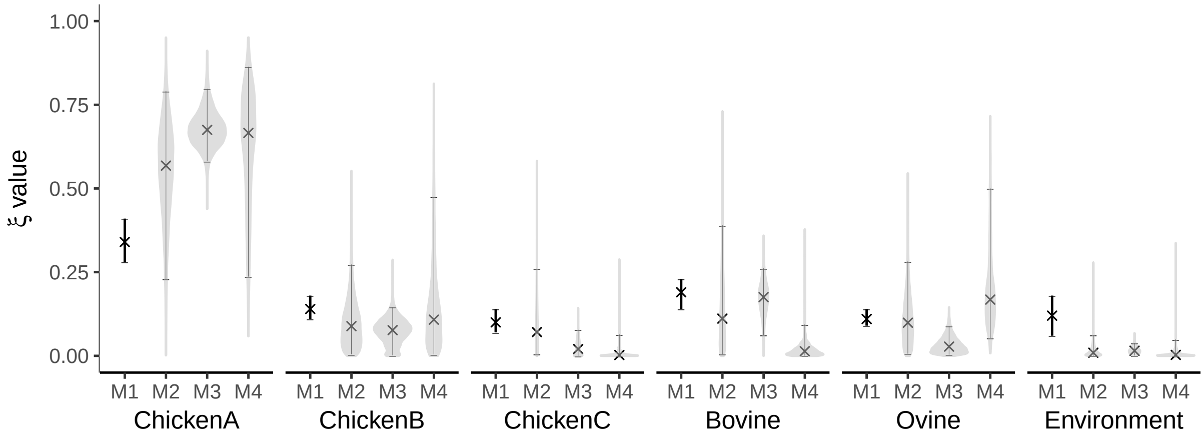


Figure 2.1: Comparison of the proportion of human campylobacteriosis cases attributable to each source. The models compared are: M1 (Dutch model), M2 (Modified Hald model), M3 (Island model) and M4 (HaldDP model). Error bars represent 95% percentile confidence or credible intervals with medians shown as a cross. Violin plots show the marginal posteriors of the ξ_j parameters.

posterior distributions for each type effect (Figure 2.3) show the largest group of types has very small type effects and wide credible intervals compared to the other groups.

Model fit and convergence was assessed visually using trace and autocorrelation plots (see Fig A and Fig B in Model fit and convergence diagnostic plots).

2.5 Discussion

`sourceR` represents a significant advance in source attribution modelling, and translation of advanced statistical methods into mainstream epidemiological use. In particular, the DP clustering results in a large decrease in the effective number of parameters in the model and allows detection of unusually virulent subtypes (group 2 in Figure 2.3) by epidemiological behaviour. The subtypes in each cluster have similar epidemiological traits (such as virulence, pathogenicity and survivability) which forms the basis for future

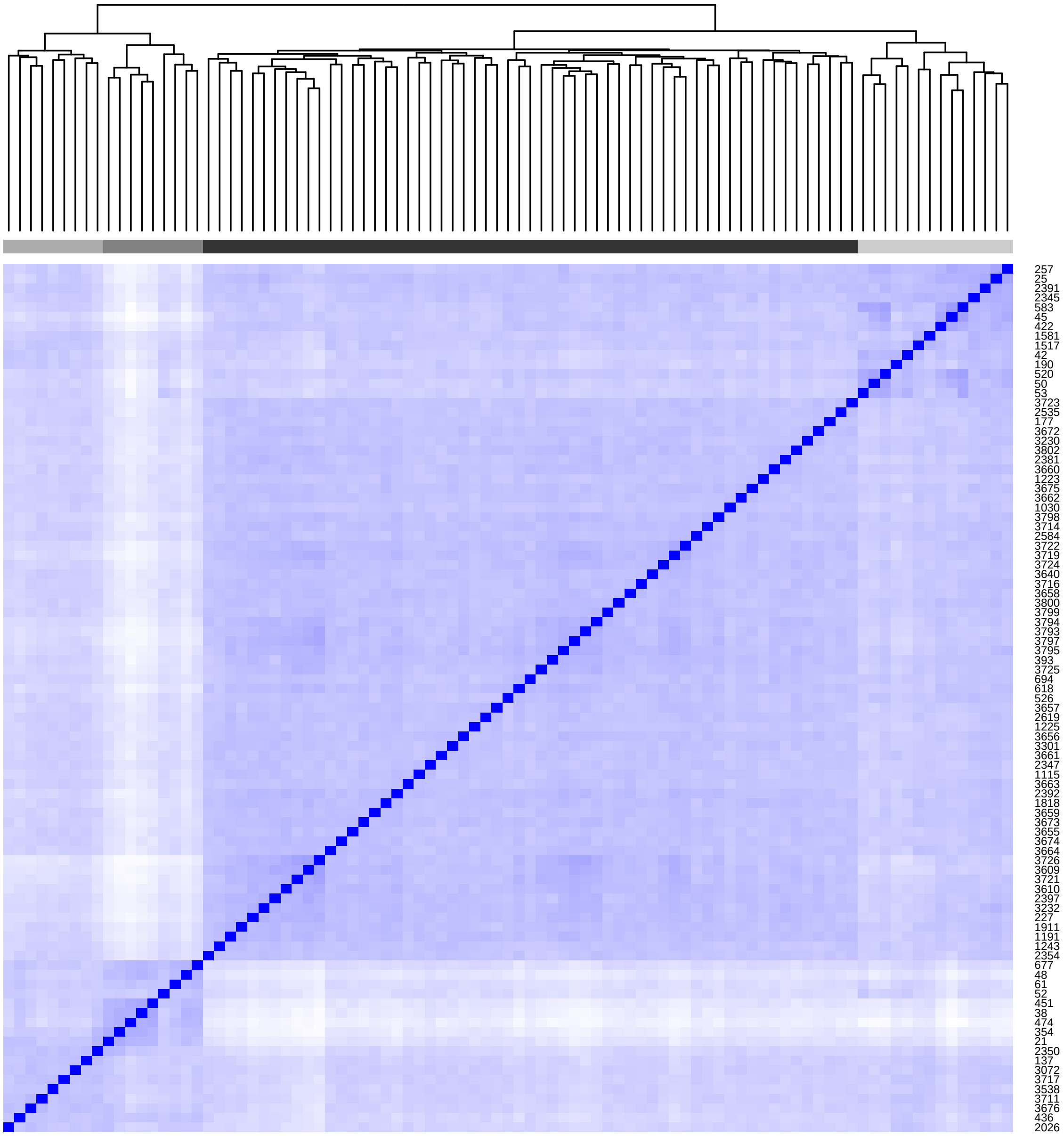


Figure 2.2: Heatmap showing the grouping of the type effects (q). A white pixel represents a dissimilarity value of 1 between a pair of sub types, whilst dark blue (see pixels on the diagonal) gives a value of zero. The grey coloured bar shows the groupings if the dendrogram is cut at 4 groups.

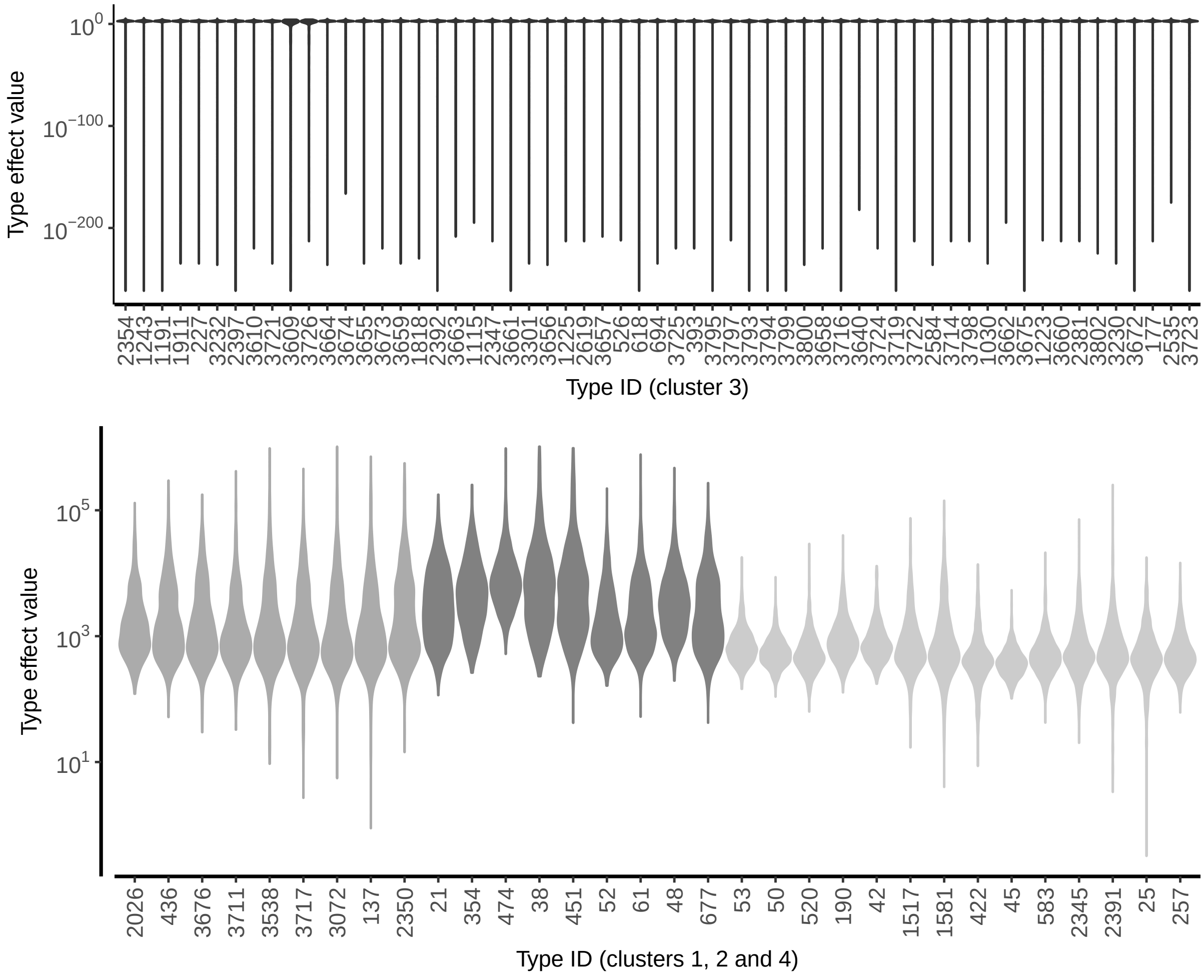


Figure 2.3: Violin plots of the marginal distributions of the type effects (q). Note that the y axis uses a log scale axis. The fill colour matches the coloured grouping bar on the heatmap.

research on genetic determinants of those traits. Additionally, if a particular type moved into the high virulence group when repeating the analysis with further data from a later time period, it would flag that type as possibly evolving to become more risky for humans. The type effects for group 3 subtypes have very wide credible intervals due to the sparsity of source samples and human cases for those types.

The relatively large uncertainty for the disease origin (the credible intervals of $\boldsymbol{\xi}$) is likely due to *C. jejuni's* complex epidemiology (Müllner, Jones, et al., 2009) giving rise to *a posteriori* correlations between components of $\boldsymbol{\alpha}$ and $\boldsymbol{q}$. This is expected due to bias/variance trade-off: the Dutch and Island models both lack type effects risking biased results due to not all types being equally likely to infect humans. The Island model also possesses inherently strong and difficult to verify *a priori* assumptions (see D. Wilson et al. (2008) and Island model overview) which are not subject to uncertainty quantification. Moreover, by removing the approximation inherent in the Modified Hald model, we expect the HaldDP model to more accurately reflect inferential uncertainty – this is particularly important for decision making in food hygiene policy, especially when commercial interests must be supported by rigorous scientific advice.

Mixing and *a posteriori* correlations of the HaldDP model are significantly decreased in comparison to the Modified Hald model, if not entirely resolved. Although heterogeneity in X is required to fit the models, a sparse or highly unbalanced source matrix increases posterior correlations between some source and type effects. In our experience, the algorithm works best when the source matrix has a moderate amount of heterogeneity.

Whilst the HaldDP results for $\boldsymbol{\xi}$ are qualitatively similar to those from the other models (Figure 2.1, we note an interesting disagreement between the Island and Hald model derivatives when comparing the the number of cases attributed to Ovine and Bovine. We conjecture that this may be due to some non-identifiability between Bovine and Ovine sources as both sources have high contamination from the same types increasing the sensitivity of $\boldsymbol{\xi}$ to sampling error. It may also be due to lack of explicit source and type effects in the Island model. Resolving this disparity is the subject of ongoing research.

2.6 Availability and Future Directions

The stable release version of `sourceR` is available from the Comprehensive R Archive Network, released under a GPL-3 licence. The development version is available at `http://fhm-chicas-code.lancs.ac.uk/millerp/sourceR`. As this package develops, we intend `sourceR` to become a platform for new source attribution model development, providing a central analytic resource for public health professionals.

The main focus of extending `sourceR` will be on modelling spatiotemporal correlation in the time and location dependent parameters. A spatiotemporal correlation model on $\boldsymbol{\alpha}_{tl}$ could be used to identify particular foci of source contamination, enabling targeted investigation of particular food supply regions. Implementation of time varying type effects may be appropriate as *Campylobacter* can evolve quickly and genetic variation conferring virulence may not be apparent from coarse-scale MLST typing (D.J. Wilson et al., 2009). Interaction terms between some sources and types would allow for the biologically plausible possibility that certain types are differentially likely to survive and cause disease, dependent on the food source they appear in. Additionally, water/ environmental samples could be attributed to the other sources of infection allowing estimation of the proportion of cases attributed to different paths of infection (direct infection from the source versus infection via the environment). However, including interaction terms and additional paths of infection would significantly increase the number of parameters and the number and strength of posterior correlations. With higher posterior correlations, the current Metropolis-Hastings based fitting algorithm would suffer from a loss of efficiency.

This could be addressed with gradient-based fitting algorithms such as Hamiltonian Monte Carlo (HMC) (Duane et al., 1987) which are designed to converge to high-dimensional, non-orthogonal target distributions much more quickly. In particular, the No U-Turn Sample (NUTS) presents an attractive method for tuning HMC adaptively, a quality which we consider necessary to minimise user intervention and maximise research productivity (Homan and Gelman, 2014).

With increased interest in source attribution models for both foodborne pathogens, `sourceR` has been written with extensibility in mind. In particular, the DAG representation allows for rapid construction of modified and new models. The package routines are written in R (as opposed to C or C++) to aid readability, with the node class hierarchy and three stage workflow designed to aid the addition of new model classes. All internal classes and methods are documented to enable prospective developers to familiarise themselves with the source code quickly, and an extensive test suite is provided. We note that the DAG framework is not limited solely to source attribution models and may used for other Bayesian applications, particularly those for which a Dirichlet process is required.

2.7 Conclusions

We have presented a novel source attribution model which builds upon, and unites, the Hald and Modified Hald approaches. It is widely applicable, fully joint, and does not require approximations or a large number of assumptions. Mixing and *a posteriori* correlations are significantly decreased in comparison to the Modified Hald model. Furthermore, it allows the data to inform type effect clustering using a Bayesian non-parametric model which identifies groups of sub types with similar putative virulence, pathogenicity and survivability. This is a significant improvement over the previous attempts to improve model identifiability (fixing some source and type effects *a priori*, or modelling the type effects as random using a 2 stage model). Like the Modified Hald model, the new model incorporates uncertainty in the prevalence matrix into the model, however, it does this by fitting a fully joint model rather than a 2 step model. This has the advantage of allowing the human cases to influence the uncertainty in the source data and preserves the restriction on the sum of the prevalences for each source. The `sourceR` package implements this model to enable straightforward attribution of cases of zoonotic infection to putative sources of infection by epidemiologists and public health decision makers.

2.8 Acknowledgments

We thank all members of the Hopkirk Molecular Epidemiology Team (Massey University), Environmental Science and Research, MidCentral Health, Public Health Services, MedLab Central, the New Zealand Food Safety Authority, Petra Müllner (for the Manawatu data set) and Geoff Jones (for his helpful input on automatic clustering methods).

Bibliography

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. *Biometrics.*

Ferguson, T., 1973. Bayesian analysis of some nonparametric problems. *Ann. stat.*, 1 (2), pp.209–230.

Duane, S., Kennedy, A., Pendleton, B.J., and Roweth, D., 1987. Hybrid monte carlo. *Physics letters b* [Online], 195(2), pp.216–222. Available from: `https://doi.org/http://dx.doi.org/10.1016/0370-2693(87)91197-X`.

Chen, M.-H. and Shao, Q.-M., 1991. Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of computational and graphical statistics.*

Gelfand, A.E., Sahu, S.K., and Carlin, B.P., 1995. Efficient parameterisations for normal linear mixed models. *Biometrika.*

Pelt, W. van, Giessen, A. van de, Leeuwen, W. van, Wannet, W., Henken, A., and Evers, E., 1999. Oorsprong, omvang en kosten van humane salmonellose. deel1. oorsprong van humane salmonellose met betrekking tot varken, rund, kip, ei en overige bronnen. *Infectieziekten bull.*

Dingle, K., Colles, F., Wareing, D., Ure, R., Fox, A., Bolton, F., Bootsma, H., Willems, R., Urwin, R., and Maiden, M., 2001. Multilocus sequence typing system for Campylobacter jejuni. *Journal of clinical microbiology.*

Crump, J.A., Griffin, P.M., and Angulo, F.J., 2002. Bacterial contamination of animal feed and its relationship to human foodborne illness. *Clinical infectious diseases* [Online], 35(7), pp.859–865. eprint: `http://cid.oxfordjournals.org/content/35/7/859.full.pdf+html`. Available from: `https://doi.org/10.1086/342885`.

Urwin, R. and Maiden, M., 2003. Multi-locus sequence typing: a tool for global epidemiology. *Trends in microbiology.*

Hald, T., Vose, D., Wegener, H., and Koupeev, T., 2004. A Bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. *Risk analysis*, 24(1), pp.255–269.

Baker, M., Wilson, R., Ikram, R., Chambers, S., Shoemack, S., and Cook, G., 2006. Regulation of chicken contamination urgently needed to control New Zealand's serious campylobacteriosis epidemic. *The new zealand medical journal.*

Roberts, G. and Rosenthall, J., 2006. *Examples of adaptive mcmc.* (Technical report). University of Toronto Department of Statistics.

Wilson, D., Gabriel, E., Leatherbarrow, A., Cheesebrough, J., Hart, C., and Diggle, P., 2008. Tracing the source of campylobacteriosis. *PLoS Genetics*.

French, N. and Marshall, J., 2009. *Dynamic modelling of Campylobacter sources in the Manawatu.* (Technical report). Prepared for Dr Donald Campbell, New Zealand Food Safety Authority. Hopkirk Institute, Massey University.

Müllner, P., Jones, G., Noble, A., Spencer, S., Hathaway, S., and French, N., 2009. Source attribution of food borne zoonoses in New Zealand: a modified hald model. *Risk analysis*, 29(7).

Pires, S.M., Evers, E.G., Pelt, W. van, Ayers, T., Scallan, E., and Angulao, F.J., 2009. Attributing the human disease burden of foodbourne infections to specific sources. *Foodborne pathogens and disease.*

Wilson, D.J., Gabriel, E., Leatherbarrow, A.J., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C.A., Diggle, P.J., and Fearnhead, P., 2009. Rapid evolution and the importance of recombination to the gastroenteric pathogen campylobacter jejuni. *Molecular biology and evolution*, 26(2), pp.385–397.

Müllner, P., Collins-Emerson, J., Midwinter, A., Carter, P., Spencer, S., van der Logt, P., Hathaway, S., and French, N., 2010. Molecular epidemiology of Campylobacter jejuni in a geographically isolated country with a uniquely structured poultry industry. *Applied and environmental microbiology*, 76(7), pp.2145–2154.

French, N. and Marshall, J., 2013. *Completion of sequence typing of human and poultry isolates and source attribution modelling.* (Technical report). Hopkirk Institute, Massey University.

Homan, M.D. and Gelman, A., 2014. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. mach. learn. res.*

Havelaar, A.H., Kirk, M.D., Torgerson, P.R., Gibb, H.J., Hald, T., Lake, R.J., Praet, N., Bellinger, D.C., Silva, N.R. de, Gargouri, N., Speybroeck, N., Cawthorne, A., Mathers, C., Stein, C., Angulo, F.J., Devleesschauwer, B., and World Health Organization Foodborne Disease Burden Epidemiology Reference Group, on behalf of, 2015. World health organization global estimates and regional comparisons of the burden of foodborne disease in 2010. *Plos med* [Online], 12(12) (), pp.1–23. Available from: `https://doi.org/10.1371/journal.pmed.1001923`.

Liu, Y., Gelman, A., and Zheng, T., 2015. Simulation-efficient shortest probability intervals. *Statistics and computing.*

World Health Organization, 2015. *Who estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015* [Online]. available on the WHO web site (www.who.int) or can be purchased from WHO Press, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland. Available from: `http://apps.who.int/iris/bitstream/10665/199350/1/9789241565165_eng.pdf?ua=1`.

Wilson, D., 2016. *`iSource`* [Online]. Available from: `%7Bhttp://www.danielwilson.me.uk/iSource.html%7D`.

2.9 Supporting Information

2.9.1 Full McMC Algorithm

This section gives the full details of the algorithm used to fit our fully joint non-parametric source attribution model. The outline McMC is shown in Algorithm 4.

```
Data: Human cases $\boldsymbol{y}$, source isolates $\boldsymbol{X}$, source prevalence $\boldsymbol{s}$
Initialize all parameters ;
for $z$ times do
    foreach $t$, $l$ do
1       Update $\boldsymbol{\alpha}_{tl}$ ;
    end
    foreach $j$,$t$ do
2       Update $\boldsymbol{r}_{jt}$ ;
    end
3   Update $\boldsymbol{q}$ ;
    Save chain state ;
end
```

Algorithm 4: Outline McMC algorithm for the `HaldDP` model.

The Dirichlet distributed source effects $\boldsymbol{\alpha}_{tl}$ across times t and locations l (Step 1), and the relative prevalences $\boldsymbol{r}_{jt}$ across sources j and times t (Step 2) are updated using a constrained adaptive multisite logarithmic Metropolis-Hastings update step for 95% of proposals, and a constrained adaptive multisite Metropolis-Hastings update step for the remainder to prevent the chain getting stuck at very low values (Roberts and Rosenthall, 2007). The adaptive algorithm updates the tuning value every 50 updates of the parameter. This is further explained in Algorithm 5.

For the Dirichlet process prior on $\boldsymbol{q}$, a marginal Gibbs sampler is constructed, as described in Algorithm 6. Let $\mathcal{H}$ denote a set of cluster identifiers, with the n-dimensional group assignment vector $\boldsymbol{c}$ associating elements of $\boldsymbol{q}$ with clusters, such that $c_i = h$ assigns q_i to

```
Input: $d$-dimensional Dirichlet($\boldsymbol{a}$) distributed random variable $\boldsymbol{W}$, tuning variance
       vector $\boldsymbol{\sigma}$, online acceptance rate vector $\boldsymbol{\rho}$, $z$ the current McMC iteration
       number.
Output: Updated $\boldsymbol{W}$ and $\boldsymbol{\sigma}$.
Let $\boldsymbol{W}' = \boldsymbol{W}$ ;
for $h$ times do
    Let $j \sim \text{UniformInteger}[1, d]$ ;
    Let $g \sim \text{Uniform}[0, 1]$;
    if $g > 0.05$ then
        Simulate $W_j' = W_j * \exp\left[\text{N}(0, \sigma_j)\right]$
        $\delta = \frac{\boldsymbol{W}'}{\boldsymbol{W}}$
    end
    else
        Simulate $W_j' = \text{N}(W_j, 0.1)$
        $\delta = 1$
    end
    Let $\boldsymbol{W}' = \boldsymbol{W}'/|\boldsymbol{W}'|$ ;
    Accept $\boldsymbol{W} = \boldsymbol{W'}$ with probability $1 \wedge \frac{f(\boldsymbol{W}'|\boldsymbol{a})}{f(\boldsymbol{W}|a)} \cdot \delta$ and update $\rho_j$ ;
    if $h \mod 50 = 0$ then
        if $\rho_j > 0.44$ then
            $\sigma_j = \exp\left[\log(\sigma_j) + \left(0.05 \wedge \frac{1}{\sqrt{(z)}}\right)\right]$
        end
        else
            $\sigma_j = \exp\left[\log(\sigma_j) - \left(0.05 \wedge \frac{1}{\sqrt{(z)}}\right)\right]$
        end
    end
end
```

Algorithm 5: Constrained adaptive multisite logarithmic random walk used for Dirichlet-distributed random variables.

cluster h. Furthermore, each cluster h assumes a value θ_h such that $q_i = \theta_{c_i}$.

In Step 1 of Algorithm 6, conjugacy between the Gamma-distributed base distribution P_0 and the Poisson data likelihood permits the calculation of Multinomial conditional posteriors for elements of $\boldsymbol{c}$ arising from the Chinese Restaurant Process construction. Here, the conditional posterior probability of type i being assigned to group h is as shown in Algorithm 6, with conjugacy permitting marginalisation with respect to the base distribution in order to calculate the probability of being assigned to a new group $h^\star$

$$p_{h^\star} = a_q \int_\Theta L(y_i^\star|\theta, \lambda_i^\star) dP_0(\theta) = \frac{b_\theta^{a_\theta}(a_\theta + y_i^*)}{\Gamma(a_\theta)(b_\theta + \lambda_i^\star)^{a_\theta + y_i^\star}}$$

with $y_i^\star = \sum_{t,l} y_{itl}$ and $\lambda_i^\star = \sum_{t,l} \boldsymbol{\alpha}_{tl}^T(\boldsymbol{r}_{it} \odot k_t)$

If a type is assigned to a new group, the set $\mathcal{H}$ is augmented and a corresponding cluster value is drawn from the posterior of $\theta_{h^\star}$. Conversely, $\mathcal{H}$ is shrunk if a particular group becomes empty.

In Step 2, the group values are drawn from the posterior, conditional on $\boldsymbol{c}$. The algorithm therefore alternates between updating group assignments $\boldsymbol{c}$ and group values $\boldsymbol{\theta}$. Hence, it explores the number of groups present, the type effects assigned to each group, and the values of each group.

```
Data: Human case counts $\boldsymbol{y}^\star = \sum_{t,l}\{y_{1tl}, \ldots, y_{ntl}\}$,
      source intensities $\boldsymbol{\lambda}^\star : \lambda_i^\star = \sum_{t,l} \boldsymbol{\alpha}_{tl}^T(\boldsymbol{r}_{it} \odot k_t)$
Input: $\mathcal{H}$ the set of cluster identifiers, $\boldsymbol{c}$ an $n$-dimensional vector of group allocators,
       $c_i \in \mathcal{H}$, $\boldsymbol{\theta}$ a $|\mathcal{H}|$-dimensional vector of cluster values
// Update group allocation $c$
for $i$ in $1:n$ do
    Sample $c_i$ from $k(c_i|\cdot) \sim \text{Multinomial}(\langle p_h : h \in \mathcal{H}, p_{h^\star}\rangle)$ where
        $p_h = |\mathcal{H}_h^{(-i)}| L(y_i^\star|\theta_h, \lambda_i^\star), \quad h \in \mathcal{H} \quad (2.11)$
        $p_{h^\star} = a_q \int_\Theta L(y_i^\star|\theta, \lambda_i^\star) dP_0(\theta), \quad h \notin \mathcal{H} \quad (2.12)$
    ;
    if $c_i = h^\star$ then
        Set $\mathcal{H} = \{\mathcal{H}, h^\star\}$ ;
        Sample $\theta_{h^\star} \sim \text{Gamma}(y_i^\star + a_\theta, \lambda_i^\star + b_\theta)$ ;
    end
    else if $|\mathcal{H}_h| = 0$ then
        Set $\mathcal{H} = \mathcal{H}^{(-h)}$ ;
    end
end
// Update cluster values $\boldsymbol{\theta}$
for $h$ in $\mathcal{H}$ do
    Update $\theta_h \sim \text{Gamma}(\sum_{i:c_i=h} y_i^\star + a_\theta, \sum_{i:c_i=h} \lambda_i^\star + b_\theta)$
end
```

Algorithm 6: Marginal Gibbs sampling algorithm using the Chinese Restaurant Process construction of a Dirichlet process

2.9.2 Island model overview

In the Asymmetric Island Model (Wilson et al., 2008; Wilson, 2016), the evolutionary processes (mutation, migration and recombination) of the sequence types are modelled to infer probabilistically the source of each human infection using genetic data from each subtype. The extra information in the genetic typing allows the model to attribute human cases from a type not observed in any sources to a likely source of infection by comparing the genetic similarity to other types that are observed in the sources (this is not possible using the Hald and Modified Hald models). However, this model makes strong assumptions about the evolutionary process (for example, constant molecular clocks) and the uncertainty in these assumptions is not easy to quantify. As the Island model requires sequence data, it cannot be used for source attribution where subtyping is based on phenotypic characteristics, such as serotype.

2.9.3 Dutch model overview

The Dutch method (Pelt et al., 1999) is one of the simplest models for source attribution. It compares the number of reported human cases caused by a particular subtype with the relative occurrence of that subtype in each source. The number of reported cases per subtype and reservoir is estimated by:

$$\lambda_{ij} = \frac{r_{ij}}{\sum_j r_{ij}} y_i \tag{2.13}$$

where r_{ij} is the relative occurrence of subtype i in source j, y_i is the estimated number of human cases of type i per year, λ_{ij} is the expected number of cases per year of type i from source j. A summation across types gives the total number of cases attributed to source j, denoted by ξ_j:

$$\xi_j = \sum_i \lambda_{ij} \tag{2.14}$$

As the Dutch model has no inherent statistical noise model, confidence intervals for the estimated total attributed cases $\hat{\xi}_j$ by bootstrap sampling over the data set. This model implicitly assumes that there are no source or type specific effects (such as differing virulence of types, or differing consumption of food sources) which is not plausible for most zoonoses.

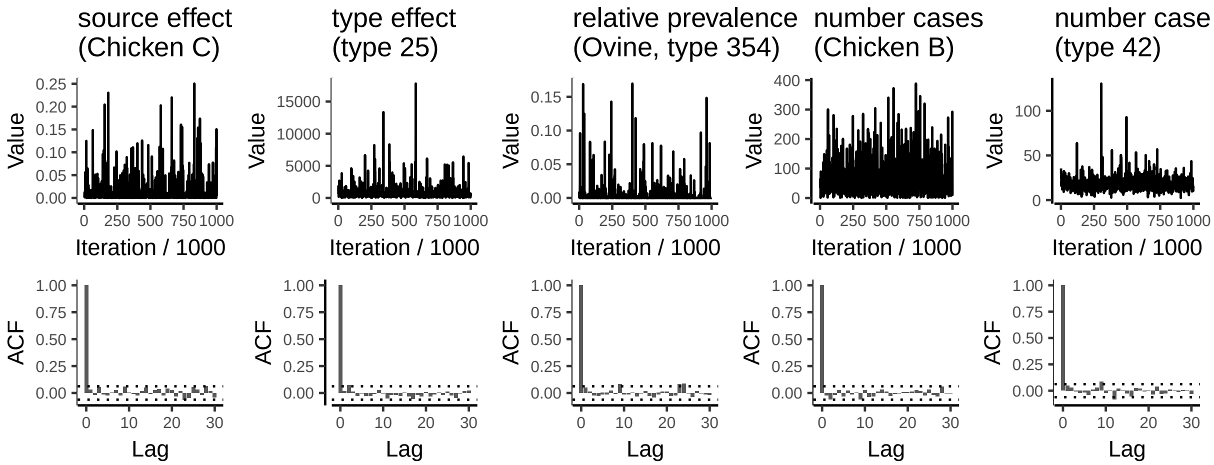


Figure 2.4: Trace and acf plots for a sample of the model parameters.

2.9.4 Model fit and convergence diagnostic plots

The trace and autocorrelation plots for a sample of the model parameters show that the model has converged (note, plots were assessed for other parameters to conclude that the model had converged). Comparing the marginal posteriors for each of the λ_i parameters with the associated observed number of human cases shows that the model fits well.

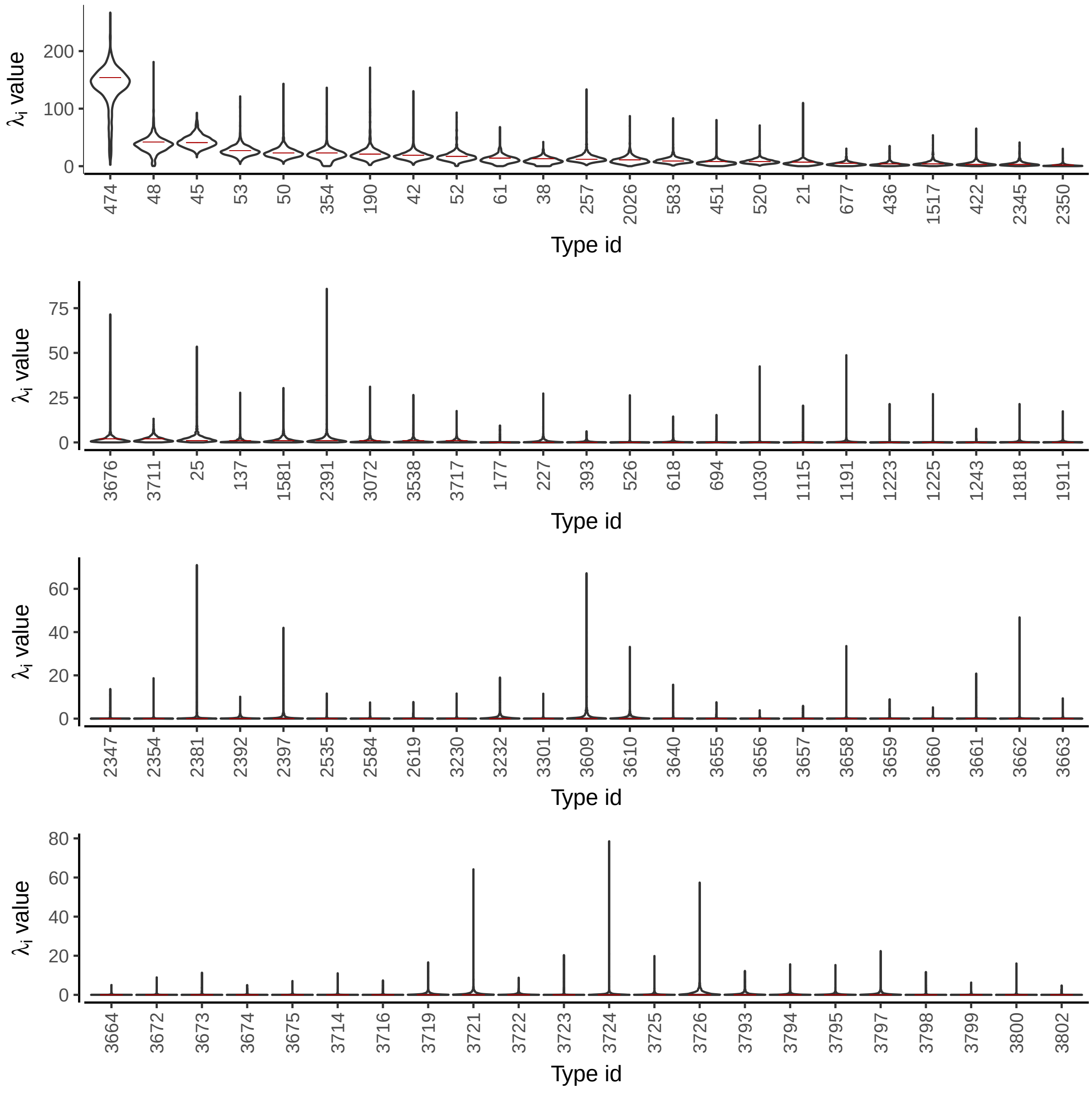


Figure 2.5: of each λ_i **(estimated number of cases attributed to each type)**. Observed number of cases for each type are shown as horizontal red lines.

Bibliography

Pelt, W. van, Giessen, A. van de, Leeuwen, W. van, Wannet, W., Henken, A., and Evers, E., 1999. Oorsprong, omvang en kosten van humane salmonellose. deel1. oorsprong van humane salmonellose met betrekking tot varken, rund, kip, ei en overige bronnen. *Infectieziekten bull.*

Roberts, G. and Rosenthall, J., 2007. Coupling and ergodicity of adaptive Markov Chain Monte Carlo algorithms. *Journal of applied probability.*

Wilson, D., Gabriel, E., Leatherbarrow, A., Cheesebrough, J., Hart, C., and Diggle, P., 2008. Tracing the source of campylobacteriosis. *PLoS Genetics.*

Wilson, D., 2016. *iSource* [Online]. Available from: %7Bhttp://www.danielwilson.me.uk/iSource.html%7D.

Chapter 3

Joint Spatial Modelling for Disease Risk in Wildlife Mediated Zoonotic Diseases

Poppy Miller[1], Kate Hacker[2], Peter Diggle[1, 2, 3], Albert Ko[2], Mike Begon[3], James Childs[2], Federico Costa[2], Mitermayer Reis[2], Chris Jewell[1]

1 CHICAS, Faculty of Health and Medicine, Lancaster University, Lancaster, England

2 Yale University, New Haven, Connecticut, United States of America

3 Liverpool University, Liverpool, England

Abstract

Leptospirosis is an emerging zoonotic disease in urban slums with few control strategies currently available. Developing effective, targeted interventions requires identification of the relative risks associated with correlated risk factors which vary at fine spatial scales. It is known that rats are carriers for leptospirosis and also shed leptospires into the environment where they can survive for long periods, especially in warm moist conditions. However, rodent control has so far been largely ineffective at reducing the burden of leptospirosis in urban slum environments where Norway rats (*Rattus norvegicus*) are the primary reservoir hosts. We estimate the spatio-temporal rat distribution and identify environmental features associated with high rat density. We then estimate comparative risk of rat exposure and other risk factors for leptospirosis for individuals living in an urban slum in Brazil.

We performed a prospective study of 1123 slum residents in the urban slum settlement of Pau da Lima, Salvador, Brazil during 2015 – 2016, whilst concurrently measuring rat presence throughout the study area. Household interviews, weather station data, serosurveys, Geographical Information System surveys, and rat tracking board studies were used to quantify individual exposure to these spatially heterogeneous risk factors. We used a spatio-temporal Bayesian cut model to estimate the environmental drivers of rat prevalence distributions and estimated the risk of leptospiral challenge in humans associated with each potential risk factor. Male gender, age (under 30), increased rainfall, decreased income, and increased rat prevalence were all associated with an increase in the risk of disease. However, many risk factors are highly correlated which complicates interpretation and prevents us from attempting to infer causal patterns. Spatially correlated random effects were required in the rat distribution model as the environmental fixed effects did not adequately explain the observed spatial variation. This allows our model to identify areas with higher than expected rat activity/density, which could in turn identify areas with an unexpectedly high risk of contracting leptospirosis. The human leptospirosis model did not show any evidence of requiring spatially correlated random effects.

The analysis clearly identifies high risk groups of individuals and suggests targets for interventions. Young black males, with low incomes, have a particularly high risk of contracting leptospirosis and also tend to live in high risk areas (near open sewers and rubbish dumps with high nearby rat activity). Our approach allowed us to incorporate highly heterogeneous spatio-temporally varying covariates (with uncertainty), and may be applicable for many other diseases with complex infection pathways.

3.1 Introduction

Effective public health interventions often require accurate estimation of spatially heterogeneous risk factors to be successful. In practice it can be challenging to collect and analyse such data. For this reason, spatially varying risk factors are often aggregated to block or census-tract levels which simplifies data collection and analyses, but which may average out the effect of factors which vary over short distances. This reduces the ability to accurately evaluate exposures and outcomes using spatially varying risk factors. It is essential to develop methods and analyses that preserve spatially heterogeneous data while still balancing field collection efforts.

In this paper, we address the spatial distribution of human leptospirosis cases in a Brazilian slum setting, relating case incidence to a presumptive reservoir in the local wild rat population. Urban rats (*Rattus* spp.) are responsible for the maintenance and transmission of a variety of pathogens important to public health (Glass et al., 1997; Himsworth et al., 2013; Costa, F.H. Porter, et al., 2014a; Walker et al., 2017) and provide an excellent system to address concerns arising from spatio-temporal heterogeneity. The habitat for the Norway rat (*Rattus norvegicus*) has expanded as the world urbanises, particularly in developing countries where expansion has been rapid and disorganised (A.I. Ko, Galvao Reis, et al., 1999; McBride, Athanazio, M. Reis, and A. Ko, 2005; Puckett et al., 2016). Nowhere is this more evident than in urban slums which are home to more than one billion people (PSUP Team Nairobi, 2016). Residents of urban slums are at a relatively high risk for zoonotic pathogens carried by Norway rats (Clinton, 1969; Mills and Childs, 1998; UNHS, 2003; Himsworth et al., 2013). One of the most striking examples is leptospirosis; a zoonotic disease with an estimated 1.03 million cases and 50,000 deaths occurring annually (Costa, Hagan, et al., 2015; Torgerson et al., 2015). Urban slums have experienced large outbreaks of leptospirosis in countries including Nicaragua, India, and Bangladesh (Varaiya et al. (2002), LaRocque et al. (2005), and Bacallao et al. (2014) respectively). These outbreaks are often associated with heavy rainfall events, where environmental damage compounds existing sanitational deficiencies, such as open sewers and rubbish dumps (A.I. Ko, Galvao Reis, et al., 1999; Barcellos and Sabroza, 2001; Karande et al., 2002; Kaur et al., 2003; A.I. Ko, Goarant, and Picardeau, 2009). Infections and severe cases of leptospirosis have also been associated with flood prone and rat infested areas, and proximity to open rubbish dumps and sewers (Sarkar et al., 2002; Maciel et al., 2008b; R.B. Reis et al., 2008a; Costa, F.H. Porter, et al., 2014a; Costa, Ribeiro, et al., 2014; Felzemburgh et al., 2014).

Within the urban slums in Brazil, and other temperate and tropical cities, the Norway rat is the primary reservoir and maintenance host for leptospirosis (Levett, 2001; McBride, Athanazio, M. Reis, and A.I. Ko, 2005; Maciel et al., 2008b; A.I. Ko, Goarant, and Picardeau, 2009). Infections are maintained within the rat population primarily by vertical and environmental transmission, via contact with contaminated urine in the environment, direct contact with an infected animal, and through sexual transmission World Health Organisation, 2003; Minter et al., 2017. Once infected, rats persistently shed leptospires into the environment (via urine) where they can survive for years in warm wet climates (Costa, Wunder, et al., 2015). Humans become infected with leptospires through contact with mud or water sources contaminated with rat urine, particularly through wounds or mucous membranes (Levett, 2001; McBride, Athanazio, M. Reis, and A.I. Ko, 2005; A.I. Ko, Goarant, and Picardeau, 2009). Forthwith, we use the term leptospirosis to refer to human disease only. Although rats are known maintainence hosts of leptospirosis, our epidemiological understanding of the contribution of relative rat abundance on leptospirosis risk, compared to other factors in urban slums, is limited. Control for rat-associated diseases has relied on rodent control campaigns, which despite major investments are largely ineffective in developing nations (Masi, P. Vilaca, and Razzolini, 2009; Masi, P.J. Vilaca, and Razzolini, 2009). Targeted control relies on accurate estimates of rat abundance and distribution and identification of environmental characteristics associated with rat density changes. Therefore, assessment of the relative importance of rat prevalence and rat risk factors, compared to addressing other social and environmental risk factors, is critical when resources for interventions are limited.

Evaluation of the effect rats have on leptospirosis incidence is extremely challenging, as it requires information about individuals' spatio-temporal exposure to rats (including direct contact and indirect exposures, such rat urine in the dirt). Instead of trying to directly measure exposure to rats, we can instead measure the relative abundance of rats near an individual's home as a proxy for their individual rat exposure. Estimating rat abundance comes with its own challenges, particularly in highly heterogeneous environments such as urban slums (Sandhu, 1987; UNHS, 2003; Kara Jose, 2008; Moreno, Oyeyinka, and Mboup, 2010; Hacker, Seto, et al., 2013). Assessing rat abundance using traditional capture-recapture or trapping studies is labour intensive and may produce unreliable estimates due to rat neophobia and trap avoidance (Ann Eileen Miller, 1985; Webster, Brunton, and Macdonald, 1994; Brunton, 1995). Indirect measures of presence and abundance, such as tracking plates where rat markings are recorded on ink-covered plates, are an attractive alternative in urban slums. They provide a cheap and effective method to not only assess the distribution of rat populations, but also estimate the relative abundance of rats across large and small spatial scales (Sheppe, 1965; 1967; Brown, 1969; Lord et al., 1971; Lord, 1983; Taylor and Raphael, 1988; Quy, Cowan, and Swinney, 1993; Drennan, Beier, and Dodd, 1998; Glennon, W.F. Porter, and Demers, 2002; Nams and Gillis, 2003; Connors et al., 2005; Promkerd et al., 2008). We previously validated the use of tracking plates to assess rat population abundance in an urban slum in Salvador, Brazil (Hacker, Minter, et al., 2016).

In this study, we followed a cohort of study participants in an urban slum in Salvador, Brazil where leptospirosis is endemic, and simultaneously sampled rodent abundance and distribution in the study area using tracking plates. A recent community-based cross-sectional survey of 3171 slum residents in Salvador found an overall prevalence of *Leptospira* antibodies of 15.4% (R.B. Reis et al., 2008b), with an estimated 60-80% of rats carrying leptospires (Costa, F.H. Porter, et al., 2014b; Costa, Wunder, et al., 2015). Dense sampling of rat activity allowed estimation of a high resolution spatio-temporal surface of rat activity, which can be used to estimate the relative rat exposure for each study participant. We additionally identified spatially relevant environmental features in the micro-environment associated with rat presence and activity, and the risk of contracting leptospirosis. We estimated the contribution of various environmental risk factors to relative rat abundance, and additionally we estimated the effect of social and environmental factors on the risk of contracting leptospirosis.

Whilst our main objective was to quantify the effects of risk factors of leptospirosis in urban slums, this study also serves as an example for examining spatial heterogeneous proxies, particularly by the study design and analysis. The results of this study aim to inform prevention and control strategies for leptospirosis, and in particular, may aid in the development of evidence-based rodent control campaigns in these complex urban areas.

The chapter is structured as follows. We first introduce the study area and experimental design. We then describe the exploratory analysis, model fit diagnostics, and the final statistical model. Results and Discussion sections follow, and the chapter ends with a conclusion.

3.2 Methods

3.2.1 Study Area

Our study was conducted from October 2015 - December 2016 across 3 valleys in the slum (*favela*) community of Pau da Lima. Pau da Lima is located on the periphery of Salvador and has been described in detail previously (Panti-May et al., 2016). The city of Salvador is the third largest city in Brazil with 2.7 million inhabitants, located on the north-east coast of Brazil (12°55' 34" southern latitude and 38°31' 12" western longitude, see Figure 3.2a) (R.B. Reis et al., 2008a). Salvador has a subtropical climate with temperatures remaining relatively constant across the year. Rainfall occurs year-round but is heaviest from April-July (mean 272.2 mm/mo) compared to the relatively dry season from September - December (mean 124.2 mm/mo).

The study site is composed of a series of valleys with a high population density and characterised by a lack of structural planning, basic sanitation, and trash collection (Figure 3.1a, 3.3a, 3.3b, 3.2b). Based on a 2013 census of community members in the study area, inhabitants are mainly squatters (88%) with low levels of education (66% did not finish primary school) and low income (mean per capita daily household income, US$ 2.60).

Chronic infection of Norway rat kidneys exceeds 50% in areas throughout Salvador (Faria et al., 2008; Costa, F.H. Porter, et al., 2014a; Costa, Wunder, et al., 2015). In this area, the mean annual incidence of hospitalised leptospirosis cases at the site between 1996 and 2002 was 57.8 cases per 100,000 population (Barcellos and Sabroza, 2001). In 2013, a single year seroincidence study of 2,003 residents at our study site estimated the leptospirosis infection rate to be 37.8 per 1,000 person-years.

The study area was constructed by creating a polygon within 3 connecting valley systems.

The polygon covered all areas that were less than 35m from the base of each valley as these areas had previously been identified as having an elevated risk of leptospirosis infection (A.I. Ko, Galvao Reis, et al., 1999; Maciel et al., 2008a; R.B. Reis et al., 2008a). The central portion of the northern most valley was not included in the study due to safety concerns, as there were high levels of gang related activities in the area. Rat presence and leptospirosis incidence were concurrently measured throughout the study area for two consecutive campaigns including a wet and dry season.

3.2.2 Sampling Design: Leptospirosis Incidence

All residents within the study area were invited to join the cohort which made up the leptospirosis data set. Blood samples from each participant were tested for the presence of pathogenic leptospires, using the microscopic agglutination test (MAT), before and after each rat tracking campaign. These paired measurements are used to classify participants as infected by defining infection as seroconversion (a titre increase from 0 to $\leq$1:50) or a 4-fold increase in titre during a campaign. See Chapter 4 for more details on the MAT method.

Of the 2076 participants invited to join the study, 1123 were retained for both campaigns and had three consecutive and successful blood draws. The locations of the individuals and their infection status is shown in Figure 3.4.

3.2.3 Sampling Design: Rat Abundance

Within the study area, an inhibitory plus close pairs design (Chipeta et al., 2017) was used to randomly select 440 sites. This design guarantees good spatial coverage by enforcing a minimum distance between a portion of the sampling locations, whilst still adhering to the principles of random sampling to avoid bias. The addition of close pairs enables

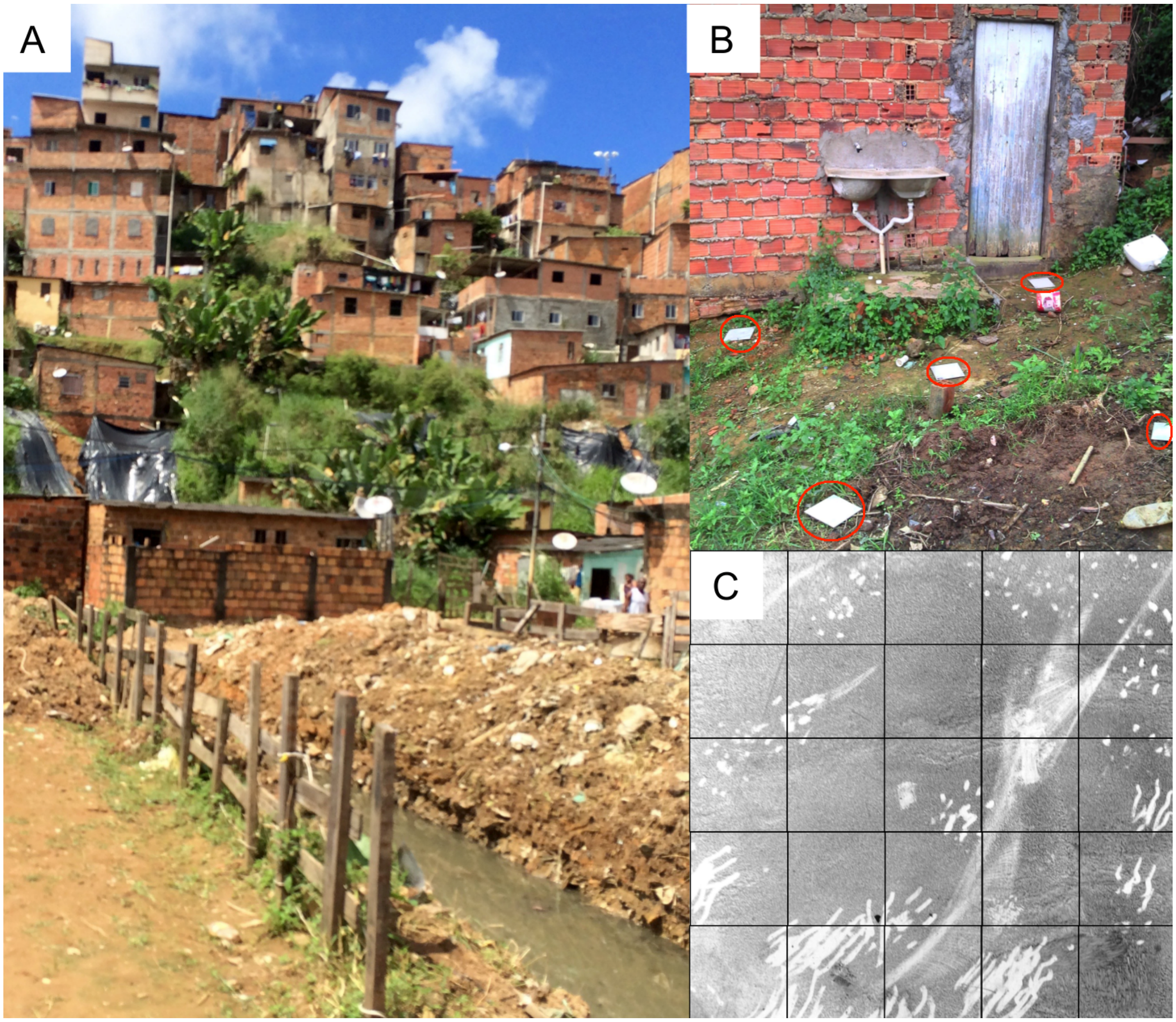
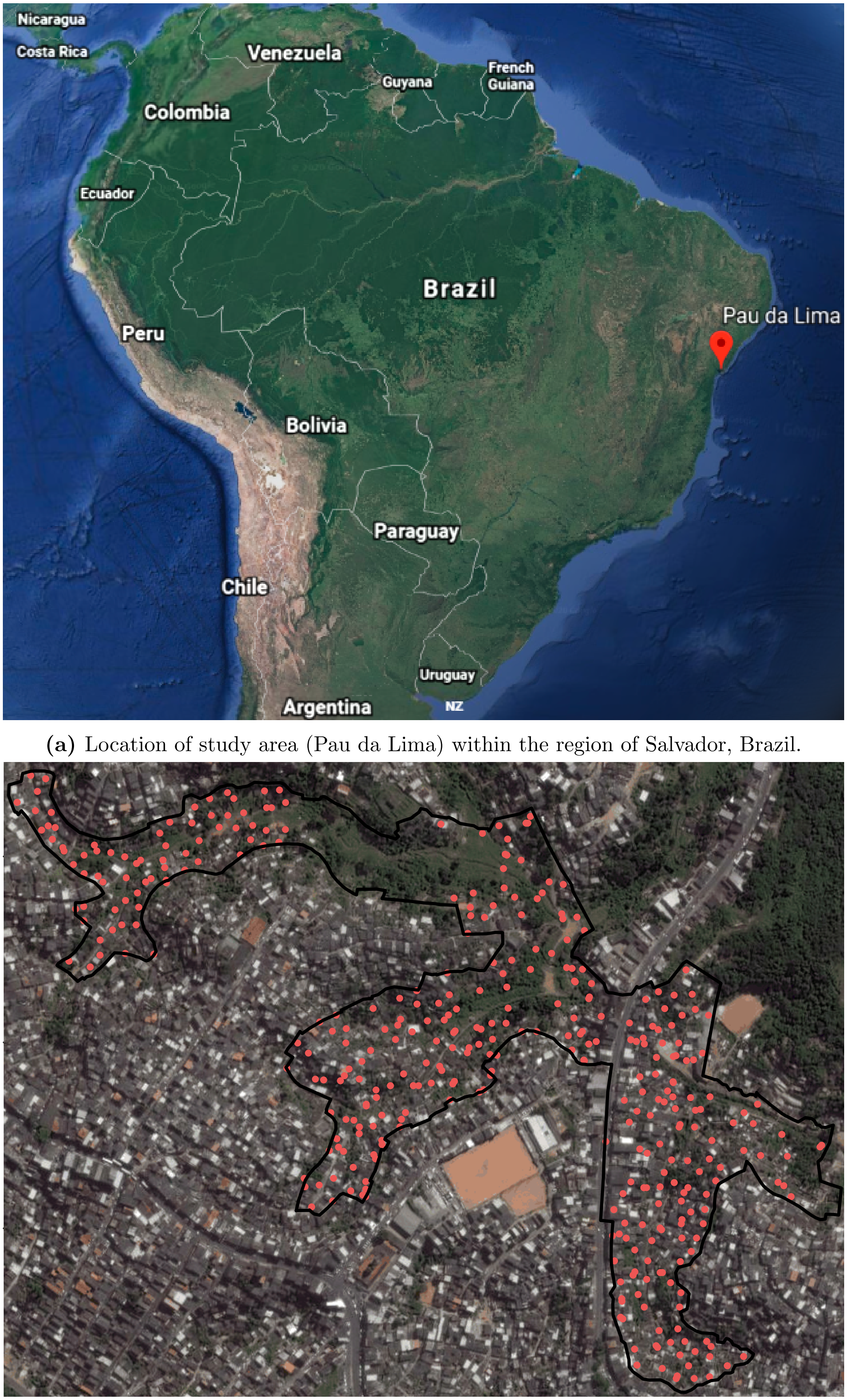


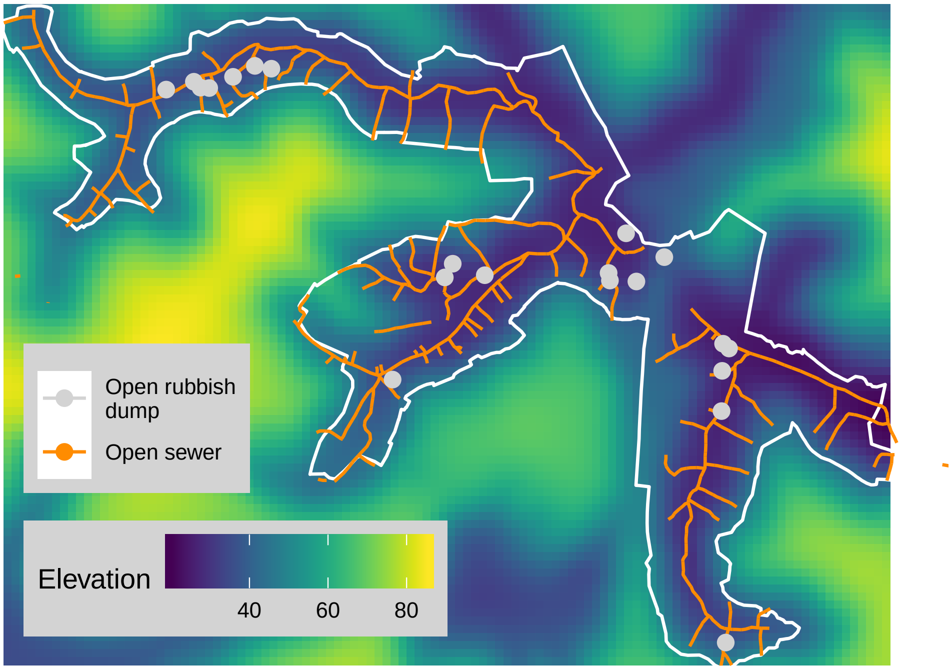
Figure 3.1: Field site in Pau da Lima. (A) Pau da Lima is a dense urban slum characterised by peaks and valleys, poor construction, and lack of sanitation, which provides ample habitat resources for Norway rats. (B) Five tracking plates were placed at each sampling point in the formation of a five on a die. Plates are circled in red in this photograph. (C) Tracking plates recorded rat paw prints, tail slides, and rat scratches. Tracking plates were scored by the presence/absence of rat marks by trained experts.



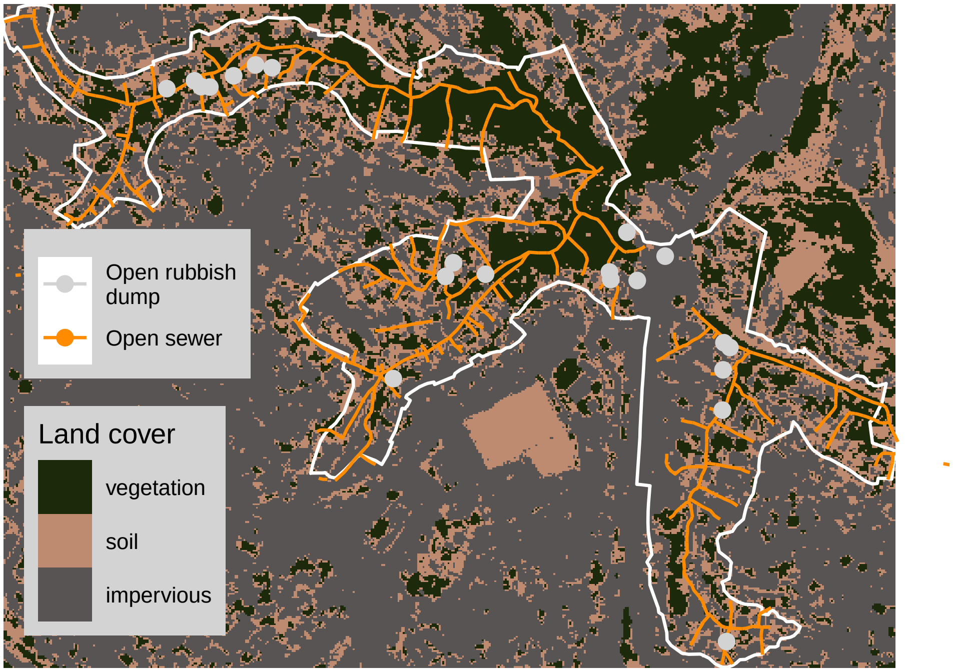
(a) Location of study area (Pau da Lima) within the region of Salvador, Brazil.

(b) Satellite map of Pau da Lima with location of tracking boards marked as red dots, and the study area enclosed in a black polygon.

Figure 3.2: Maps showing the location of the study area.



(a) Elevation map of study area with the study area enclosed in a white polygon, open rubbish dumps marked as grey dots, open sewers marked as orange lines, and elevation given by tile colour.



(b) Land cover map with the study area enclosed in a white polygon, public trash locations marked as light grey dots and open sewers marked with orange lines. The estimated land cover class is given by tile colour (green vegetation, brown soil and grey impervious surfaces).

Figure 3.3: Maps of the study area showing spatially varying covariates.

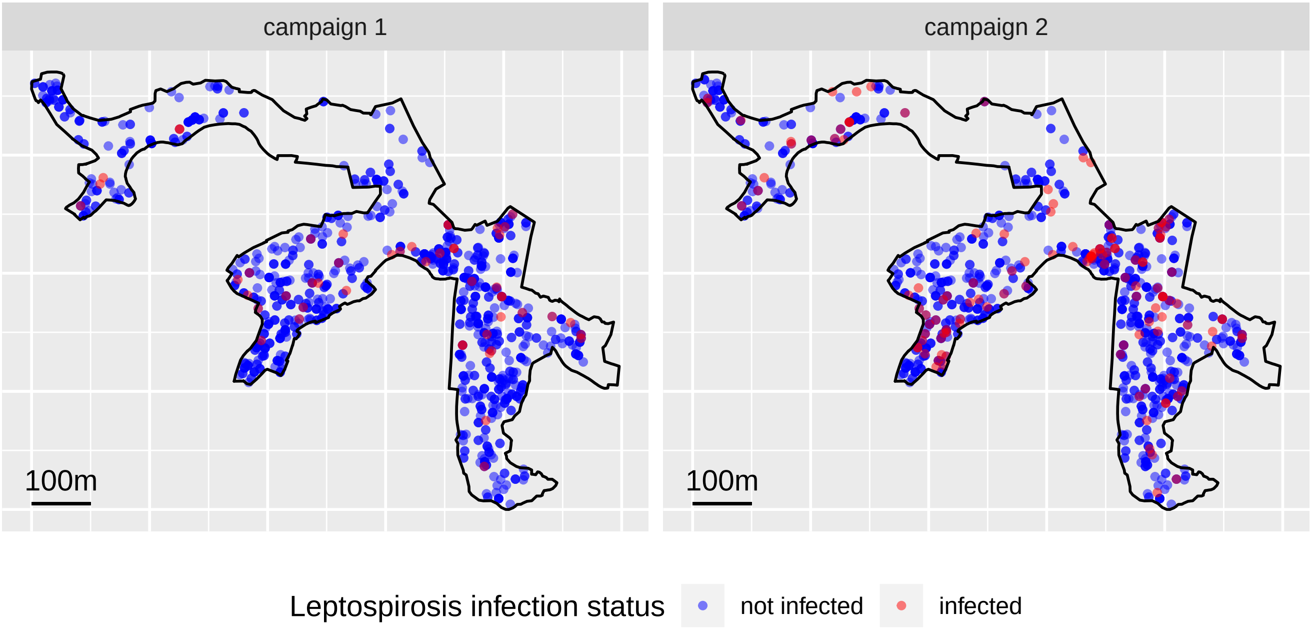


Figure 3.4: Study participant locations and infection status by campaign. Study participant locations are the addresses where they resided for the majority of the study period. Red dots indicate individuals who tested positive for leptospirosis, and blue dots indicate those who did not.

quantification of short range spatial variation, helping distinguish it from measurement error (Chipeta et al., 2017). In our study, 340 points were randomly selected to be at least 16.5m apart, with the remaining 100 sites were chosen to be at most 10m from one of the initial sampling locations (Figure 3.6). When the nugget variance (measurement error and short range correlation) is non-negligible, inhibitory plus close pairs designs demonstrate improved efficiency over designs without close pairs.

Of the original 440 randomised sites, 420 (95%) were located in areas that were physically accessible and were considered for further analysis. Sites were excluded when they were inaccessible for reasons such as: being located on a cliff, in a flooded bog, or being in a domestic environment where access consent was not able to be obtained from the homeowner. Of the 420 sites attempted, we successfully performed track plate surveys at 369 (89%) sites. To limit the effects of spatial-temporal confounding, the sites were grouped into 24 clusters, and 3 random clusters were sampled per week from October to December. Clusters were selected by assessing the number of points that were spatially feasible for the field team to collect in a single day (groups of approximately 20 closely spaced points).

Prior to placing track plates, each sampling point was identified using GPS and georeferenced maps. The sites were marked with a unique ID label and photographed for future reference. If the randomised point fell within an inaccessible area, and there were no suitable locations to place tracking plates within 5m of the original randomised point, the point was excluded from the study. If tracking plates could be placed within a 5m radius buffer of the inaccessible point, the new location was used instead.

Tracking plate surveys used standardised protocols refined previously (Hacker, Minter, et al., 2016). At each point, 5 tracking plates were placed in the geometric shape (as in the number 5 on a die) shown in Figure 3.1b. The central plate was placed as close as possible to the selected geo-referenced point, and the remaining 4 plates were evenly spread around it at a radius on 1m. When possible we avoided placing tracking plates in open areas but choose natural barriers or near visible rodent signs, within a the 1m radius buffer area.

Tracking plates were painted with lampblack on site, using methods described previously (Lord, 1983; Hacker, Minter, et al., 2016), and left in place for two consecutive tracking nights. Each morning, tracking plates were examined for rat activity (Figure 3.1c), photographed, and re-painted with lampblack solution. If tracking plates could not be placed for two nights within two weeks of the original sample date, they were treated as missing for the second night of tracking activity. All data documenting whether tracking plates were marked by a rat, missing, or moved was recorded using standardised questionnaires in the RedCap data system. All photos of tracking plates were saved on secure data servers and connected to the RedCap data system following quality control protocols. Plates were censored when >70% of the area on the plate was unreadable for rat-specific marks. Two independent scorers scored 20% of the tracking boards to assess for agreement using methods described previously and in Hacker, Minter, et al. (2016). The scores of the randomly selected sites (74 sites, 370 individual tracking plates) were highly correlated between the two reviewers (r-squared = 0.94, $p < 0.001$). Since the scores were highly correlated, a single scorer was designated to read the remaining tracking plates. This process was repeated twice at the same locations, once in the dry season and once in the wet season, producing two campaigns worth of data.

The proportion of boards that were lost at each location was low (overall 80% of boards could be read) with 43% of locations in campaign 1 and 46% of locations in campaign 2 having zero lost boards (Figure 3.5).

Of the 369 sampling sites, 328 recorded two days of track plate activity and 189 sites were positive for rats over the course of the tracking period (see Figure 3.5 and Figure 3.6).

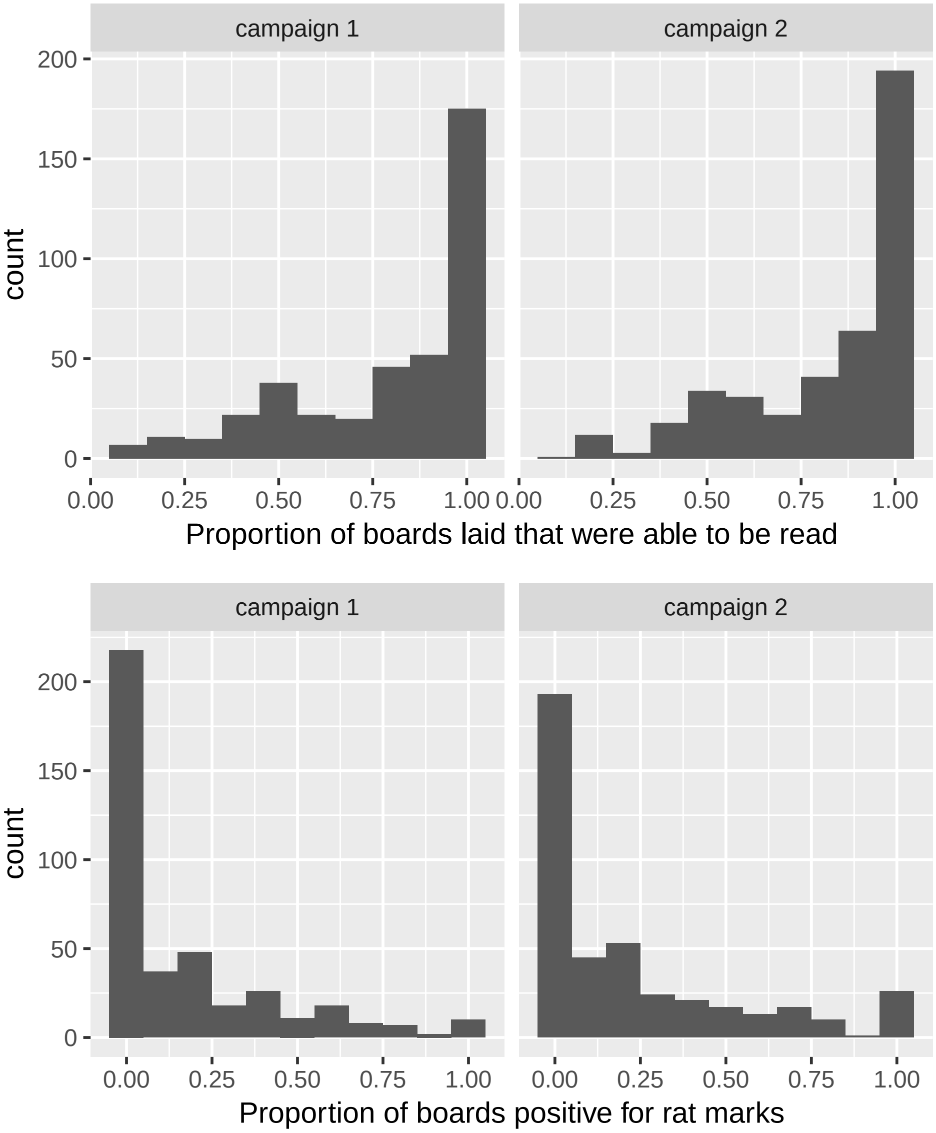


Figure 3.5: Rat tracking board summary distributions. The top histograms show the distribution of the proportion of rat tracking boards that were able to be read in each campaign. The bottom histograms show the distribution of proportion of boards positive at each site, within the two campaigns. The distributions show that most boards were successfully read in both campaigns, and that many locations had 0 boards with rat marks. Distributions were similar between campaigns, with some increase in proportion positive in campaign 2.

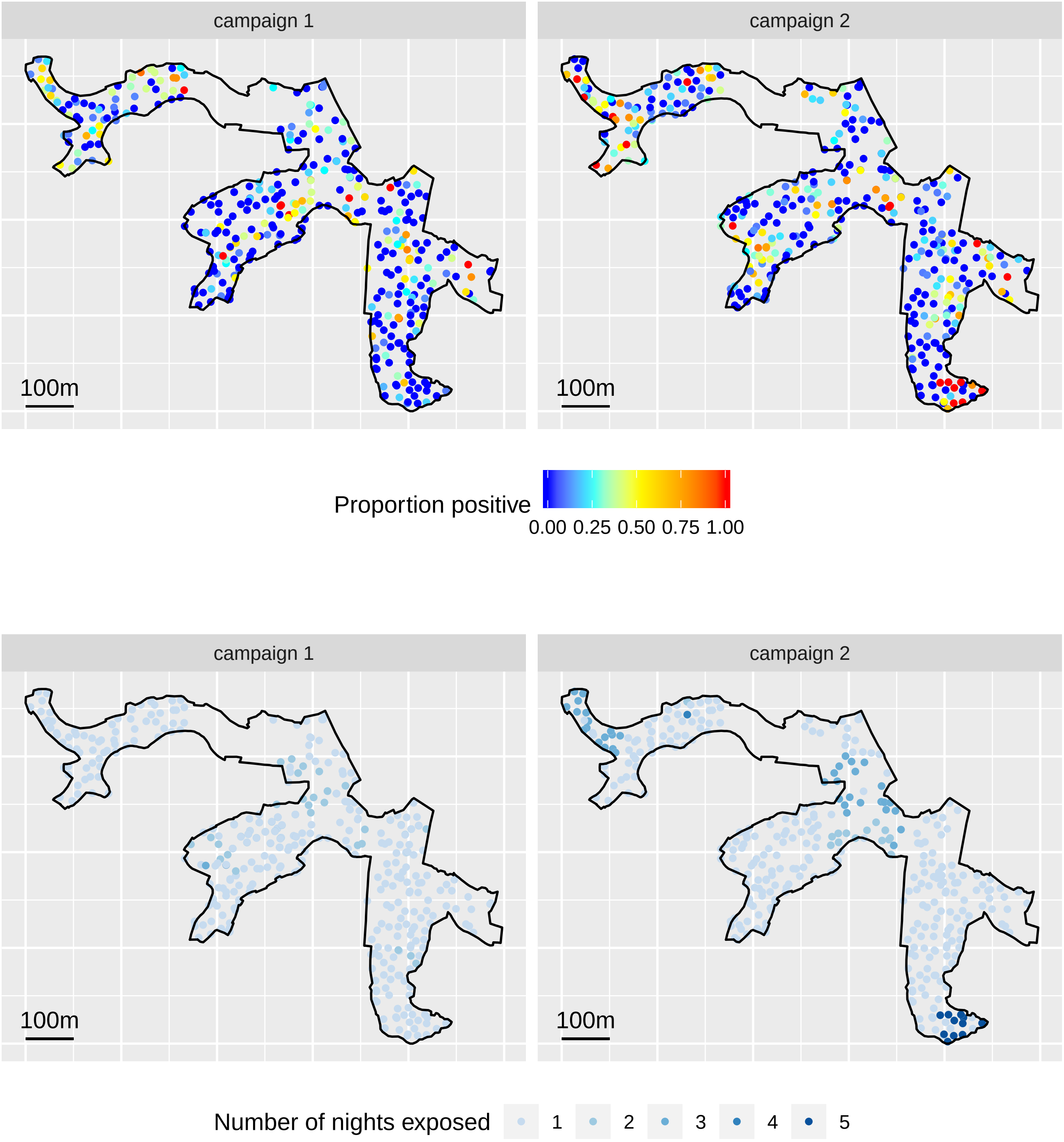


Figure 3.6: Map of rat tracking board locations, proportion observed positive and exposure time. The maps show distinct hotspots of rat prevalence (orange and red dots on top map), and highlights the effect of increased exposure time (dark blue dots on bottom maps) on the probability of detecting rat marks.

3.2.4 Covariate collection

3.2.4.1 Environmental covariates

Satellite imagery was used to generate spatial covariates that were included in both the rat abundance and leptospirosis models. All spatial data was recorded using either GPS or geo-referenced maps and were entered into a secure geo-database using ArcGIS (ArcGIS, 2012). In October 2014, a trained team of health care workers mapped the location of open sewers and public trash dumps throughout the study area and created geo-referenced shapefiles for use in ArcGIS.

For each sampling site, three-dimensional shortest path distances to the nearest open-sewer, valley bottom, and public trash dump were calculated in R using the polyline shapefiles and a geo-referenced Digital Elevation Model (DEM) raster (10m resolution). All shapefiles and raster datasets were projected to the Universal Transverse Mercator (UTM) South America Data zone 24S coordinate system.

Satellite imagery was acquired using Digital Globe's WorldView-2 satellite imagery (8 bands) on February 17, 2013. With this imagery we generated a supervised land cover classification model using a maximum likelihood classification algorithm in ENVI 2.0 (ENVI, 2013). Three classes of land cover were chosen: man-made structures (impervious surfaces including pavement, cement, and different types of roofs), vegetation, and exposed soil. Training data for the land cover types was collected from 20 sites throughout the study area. If land cover class was not present, or had changed, the training site was disregarded. At each sampling location the proportion of area covered by each cover type was calculated within a 5m radius.

Cumulative rainfall was calculated by the total amount of rainfall observed from the Canabrava weather station during the tracking plate period. The Canabrava weather station is the closest station to the study site and is located approximately 1.5km away.

3.2.4.2 Social covariates

A study team of various healthcare workers assessed various demographic, socio-economic, employment, and exposure features during household interviews and home inspections at recruitment. Interviewers and inspectors were trained on the study tools and interview techniques prior to initial data collection and all used a standardised questionnaire format. Information on ethnicity was self-reported. Income was defined as the household member who earned the highest monthly income. Literacy was assessed by the ability of the participant to read standardised sentences and interpret their meaning. Exposure status was evaluated by questioning the participant's about their contact with mud, floodwater, and sewer water. Household surveys were conducted to determine the presence of environmental features within a 5m radius of the study household.

3.2.5 Exploratory analysis

An exploratory analysis was performed to identify potential non-linearities in the relationships between the environmental covariate data and rat activity and the human leptospirosis response variables. Additionally, we looked for evidence of residual spatial variation to indicate whether spatial random effects were necessary.

3.2.5.1 Identifying non-linearities

For the first of these tasks, a Binomial generalised additive model (GAM, fitted using the R package mgcv version 1.8-24, (S. N. Wood, 2011)) was fitted to the rat presence and human incidence data sets, where all continuous variables were represented using the default 1D spline type in mgcv (penalised thin plate regression splines which are equivalent to natural cubic splines in 1D) and no random effects were incorporated. We chose to replace the natural cubic splines used in mgcv with simpler forms (quadratic or linear splines) to reduce the risk of over fitting and simplify model interpretation. Graphs of the partial residuals and fitted splines for each variable were used to choose an appropriate simple form for each covariate, and knots where appropriate. Partial residuals for a smooth term are the residuals that would be obtained by dropping the term concerned from the model, while leaving all other estimates fixed.

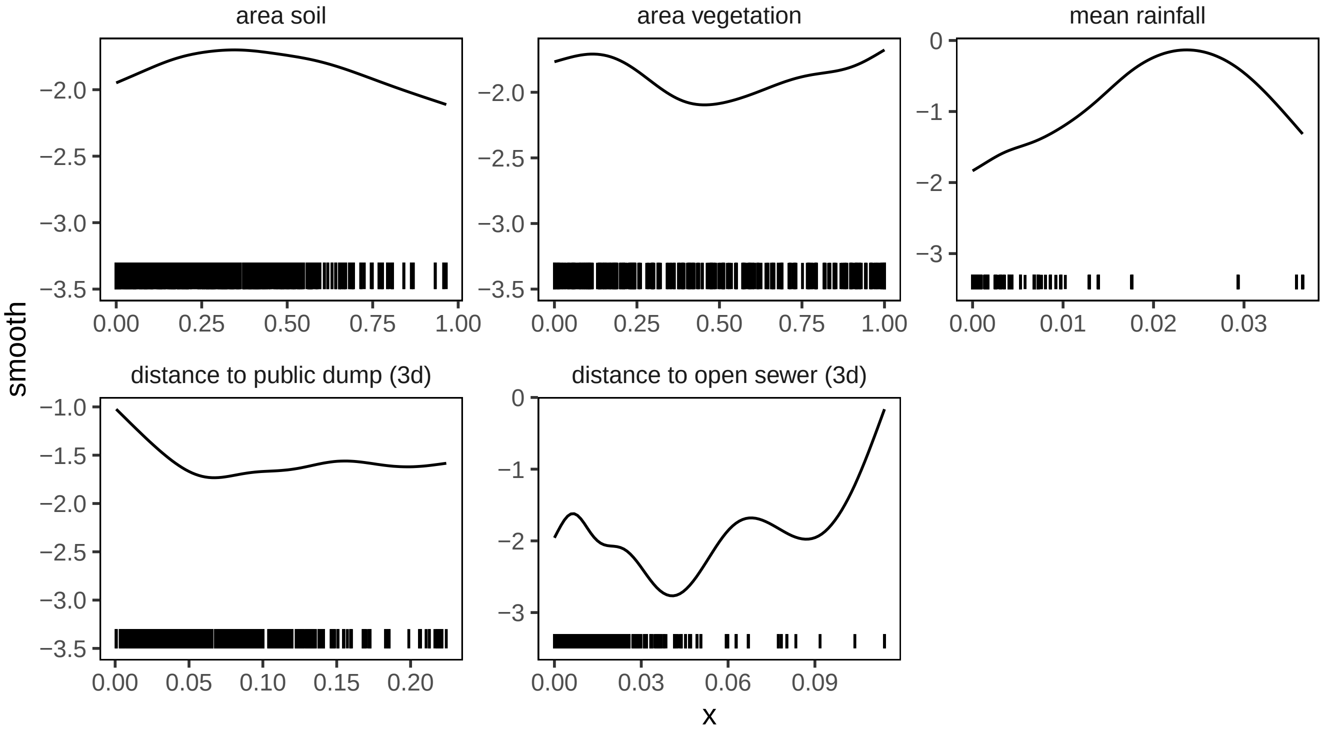


Figure 3.7: Preliminary GAM smooths for rat data. Estimated component smooth functions from the preliminary GAM model fitted to the rat tracking board data set. See 3.2.5.1 for more details on the model. Area soil and vegetation are represented as proportions of total area, distances to public dumps and open sewers are in km and mean rainfall is in m.

The exploratory GAM rat model indicated that several covariates had non-linear relationships (see Figure 3.7). Therefore, in the final model quadratic terms were introduced for the area covariates, and the distance covariates were fitted as piece-wise linear splines with knots at 70 meters (0.07km) for distance to public dump, and at 40 meters (0.04km) for

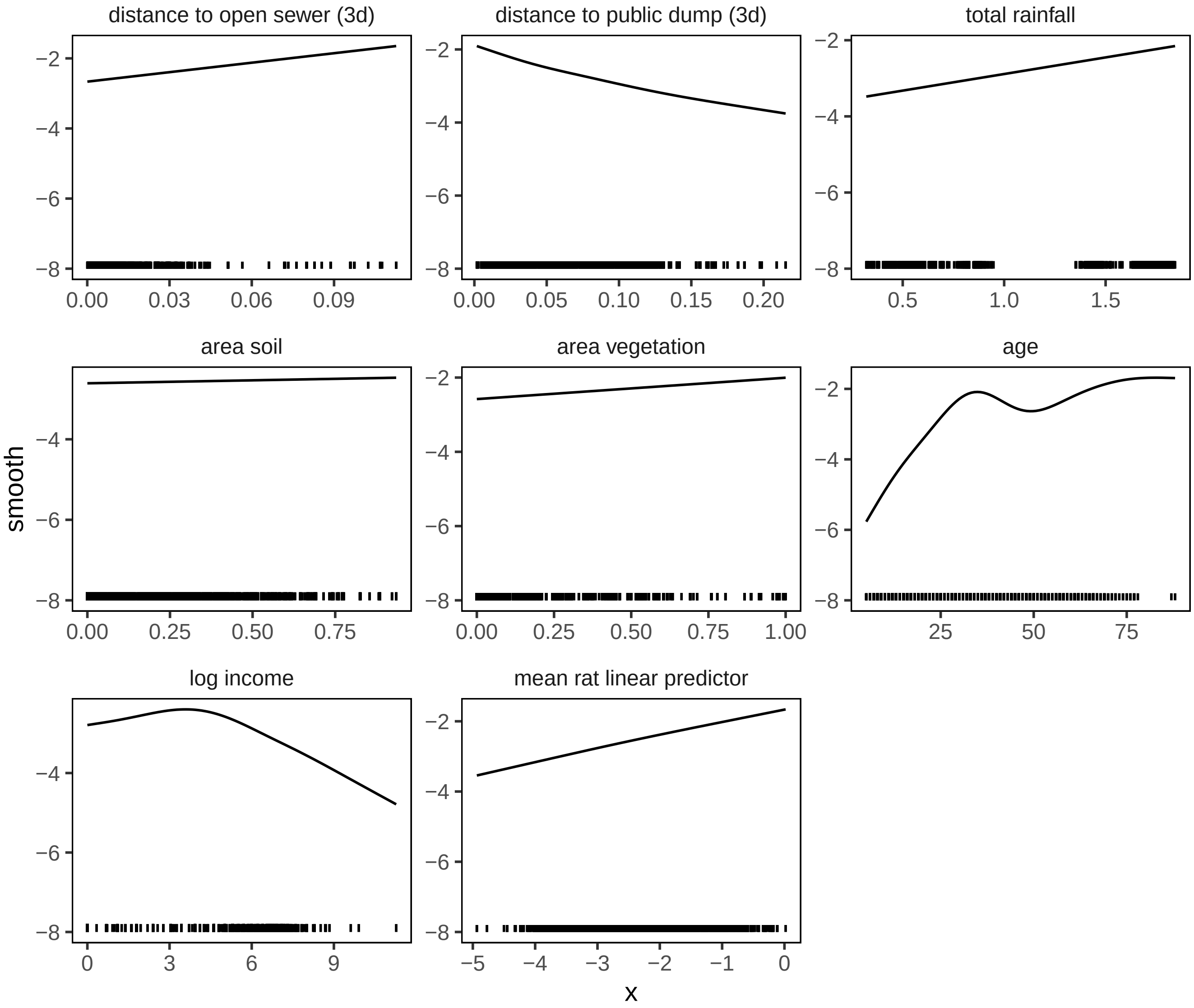


Figure 3.8: Preliminary GAM smooths for leptospirosis data. Estimated component smooth functions from the preliminary GAM model fitted to the leptospirosis case data. See 3.2.5.1 for more details on the model. Area soil and vegetation are represented as proportions of total area, distances to public dumps and open sewers are in km, mean rainfall is in m and age is in years. meanlp is the mean logit of the predicted probability of rats at the study participants home using the final Bayesian rat model (not the preliminary GAM rat model).

distance to open sewer. It was decided not to include a quadratic term for mean rainfall as the trend looked linear for the majority of points, with the smoothed curve shown at the end of the GAM mostly driven by very few boards with which experienced extremely high rainfall.

The linear splines were created as follows. For a model with a single covariate x with a knot at w, we have the following linear predictor

$$\alpha_0 + \alpha_L x + \alpha_U (x - w)_+ \tag{3.1}$$

where $(u)_+$ equals u, if u is positive, and 0 otherwise. This gives a slope of α_L for $x < w$ and $\alpha_L + \alpha_U$ for $x > w$, and forces the two sides of the spline to meet at w.

A similar GAM model was then fitted to the leptospirosis data set with the mean linear predictor from the final rat model (see section 3.2.7.1) as a covariate (see figure 3.8). Based on the results of this exploratory analysis, it was decided that the age and log income covariates should be represented by a piece-wise linear spline with knots at 30 and $\log(40) = 3.69$ respectively.

3.2.5.2 Identifying spatial correlation

A generalised linear mixed effects regression model, with independent random intercepts for each unique location-campaign combination, was then fitted to the rat data set. A similar model was fitted using the leptospirosis data set where the rat linear predictor covariate value was calculated using the mean posterior predicted value from the final rat model (see details of the final model in section 3.2.7.1). An empirical variogram of the random intercepts from the rat model r_i was used to check for the presence of residual spatial correlation in the rat data. The leptospirosis model was not able to converge with independent spatial random effects; instead, the model residuals were used to create the variogram.

The empirical variogram was calculated as follows. Let $v_{ij} = (r_i - r_j)^2/2$ and u_{ij} be the distance between sites i and j. Pick a grouping interval h, let n_r be the number of u_{ij} that lie between $(r-1)h$ and rh and $\bar{v}_r$ the sample mean of the corresponding v_{ij} . A plot of $\bar{v}_r$ against $(r-0.5)h$ is called the sample variogram. It estimates the function $V(u) = \sigma^2(1-\rho(u))$, called the theoretical variogram.

Figure 3.9 shows an example variogram with a Matérn covariance function. The sample variance of the residuals/iid random effects estimates the quantity $\tau^2 + \sigma^2$, although some shrinkage of the random intercepts is expected when compared to the fitted spatially correlated random effects. The nugget τ^2 is the variation attributed to measurement error and very short scale spatial variation (at distances smaller than the sampling distances). The sill is the value at which the semi-variogram levels out ($\tau^2 + \sigma^2$) and is the variance of the spatial random effects. The practical range is the distance at which the semi-variance reaches 95% of the sill value. As the distance between 2 sample locations increases, their correlation decreases. Locations further apart than the practical range have minimal correlation. When there is no evidence of spatial correlation, the plotted points will form an approximately flat line.

A fitted variogram can be used to visualise the effective range of the spatial correlation; that is, the distance beyond which the correlation between observations is less than or equal to 0.05. We can compare the empirical and fitted variograms to check that the spatial correlation model fits the data well.

It is also possible to detect spatial correlation using a variogram of the model residuals (with no random effect for location). This can be useful when there are very few data points at each location, making it difficult to fit a model with iid random effects at each

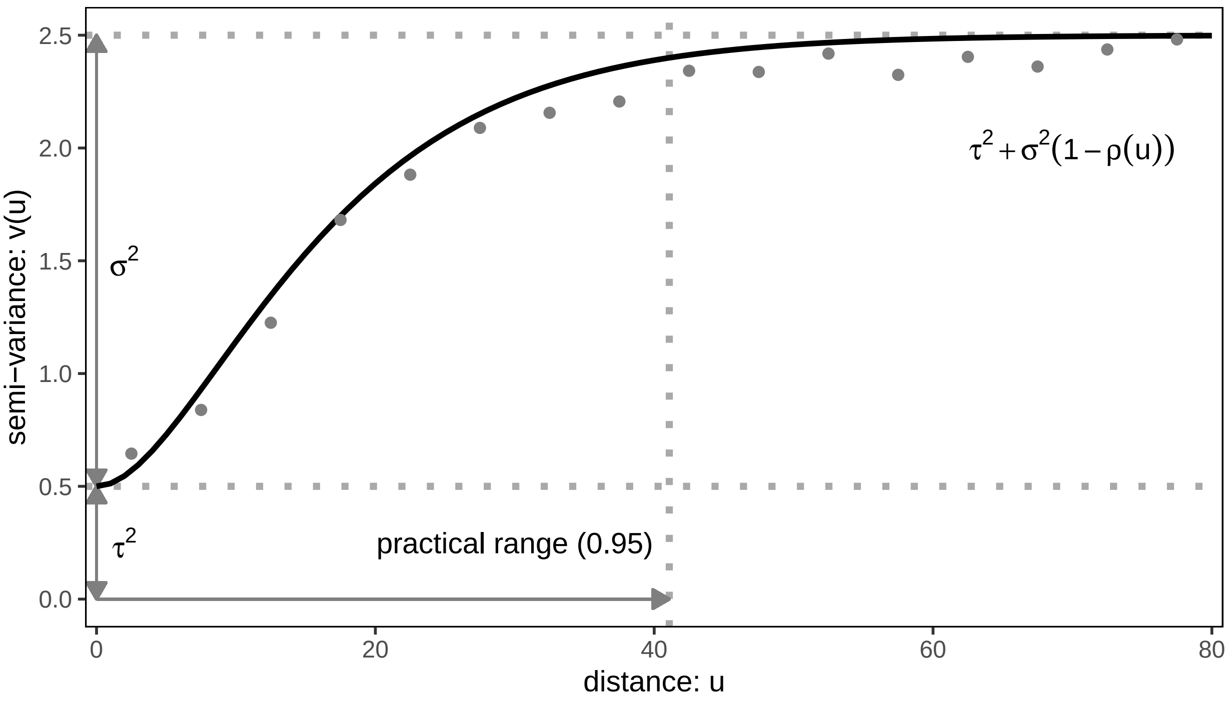


Figure 3.9: Example variogram showing relationship between model parameters and correlation patterns over space. The fitted theoretical variogram $V(u)$ gives a graphical representation of the estimated variance components τ^2 and σ^2 and of the correlation function $\rho(u)$. This plot shows an example of a semi-variogram with a Matérn 3/2 covariance structure with variance σ^2, nugget τ^2 and covariance $\tau^2 + \sigma^2(1 - \rho(u))$. Simulated data is shown as grey dots and the true semi-variance value is shown as a solid black line. Dotted grey lines show the location of the practical range beyond which the correlation is below 0.05.

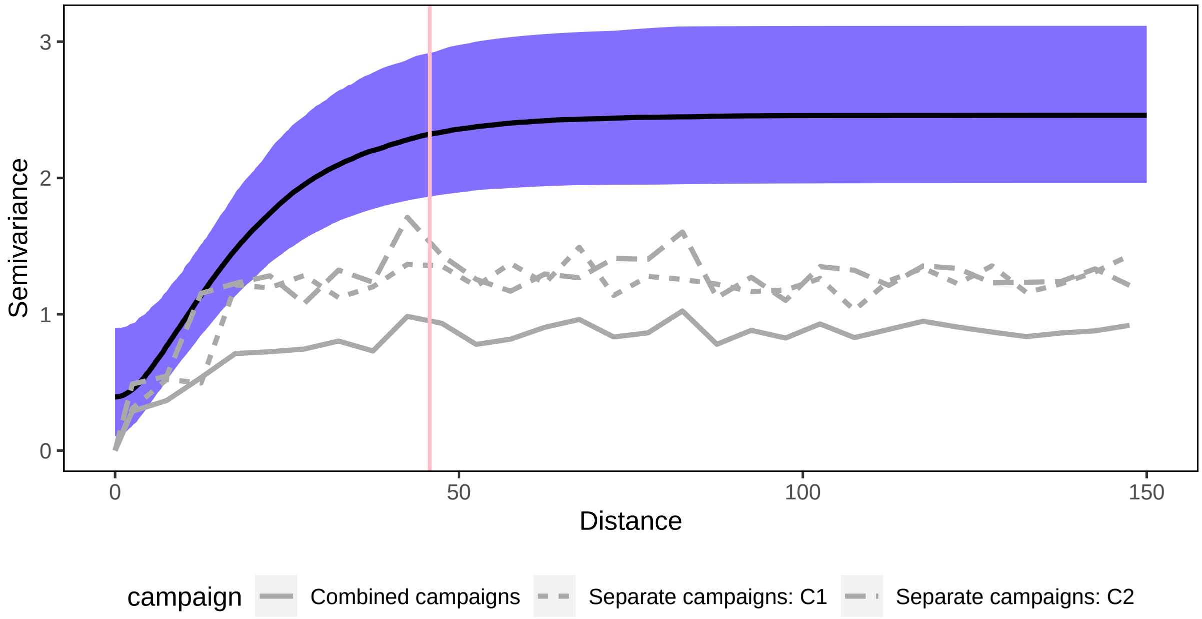
location.

The variogram produced using the random effects (per location and campaign) from the rat model indicated that there was significant spatial correlation at close ranges with a practical range of nearly 50 meters (see Figure 3.10a). It also showed that although the pattern of spatial correlation was very similar for the two campaigns, the correlation was smaller when both campaigns were combined, indicating that a separate spatial surface should be fitted for each campaign. The variogram produced using the leptospirosis residuals indicated no evidence of residual spatial correlation (see Figure 3.10b).

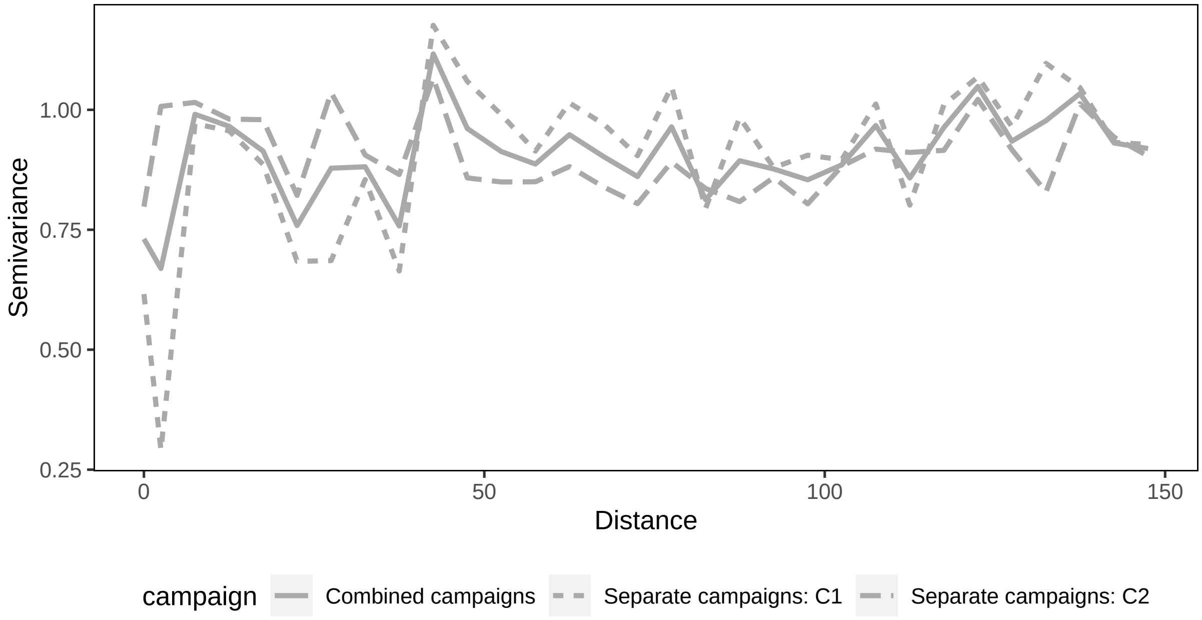
3.2.6 Residuals

Residuals for Binomial and Bernoulli generalised linear models often show patterns that do not indicate a lack of fit when n is small. There have been several methods proposed to detect whether the patterns are problematic. Comparing the residual plots visually with those produced using a parametric bootstrap indicates whether the observed patterns are unusual. However, this can be very computationally intensive which limits its use with models requiring a significant amount of time to fit. Hosmer and Lemeshow (1980) proposed grouping the residuals to reduce the patterns due to small n. The grouped observed and fitted values may also be used to test for lack of fit. However, this method can be very sensitive to the sample size and number of groups chosen (Xin and Liu, 2018).

Instead, we chose to use separation plots to visually assess model fit (Greenhill, Ward, and Sacks, 2011). Separation plots visually show the model's ability to attribute high predicted probabilities to actual occurrences of the event, and low probability predictions to non-events, whilst avoiding sensitivities to arbitrary probability thresholds. Separation plots show the observed data (0/1 for Bernoulli and empirical probability for Binomial)



(a) Variogram of iid random effects at each location from the rat glmm model. This plot shows clear evidence of residual spatial correlation in the iid random effects values, indicating spatially correlated random effects are needed in the rat tracking board model. The pink line shows the median fitted practical range. The blue ribbon shows the 95% credible interval from the fitted variogram from the final model, and the median fitted variogram is shown as a solid black line (see section 3.3.2.1 for more details).



(b) Variogram of residuals from the preliminary leptospirosis glm model. This plot shows no evidence of residual spatial correlation in the residuals, indicating spatially correlated random effects are not required for the leptospirosis case model.

Figure 3.10: Empirical variograms from preliminary models and fitted variogram from the rat model.

as coloured vertical lines ordered by fitted probability, with an additional curve showing the fitted values. If the model fits the data well, the colours associated with high observed probabilities will cluster on the right hand side of the plot. See section 3.3.2.2 for the separation plots for the models from this chapter.

3.2.7 Statistical modelling

A full probability model for the combined rat and leptospirosis datasets would allow information from both data sets to inform all model parameters. We chose to instead fit a cut model (Plummer, 2015), which controls the flow of information from data to parameters, because a fully joint model displayed significant computational issues. The cut model helps to make the model identifiable when the same environmental covariates are used to estimate both the rat prevalence and probability of leptospirosis. The cut model is implemented in practice by first fitting an appropriate model to the rat data (model 1), then fitting a model to the leptospirosis data (model 2), conditional on the predicted rat prevalences from model 1. Uncertainty in the predicted rat prevalence covariate in the leptospirosis model is incorporated by numerically integrating over the predictive posterior distribution of the rat model at the human locations. This means that the leptospirosis model is conditional on the fitted spatio-temporal distribution of rats from model 1. Both the rat and human data were modelled using Binomial generalised linear mixed models with appropriate random effects structures. The following section outlines the technical implementation of the model. The model parameters are defined in table 3.1 and a graphical representation of the model is given in Figure 3.11.

3.2.7.1 Rat prevalence model

Each board was classified as marked or unmarked by rats, and this data (along with a set of covariates) are used to produce a spatio-temporal binomial generalised linear model

(Diggle, Tawn, and Moyeed, 1998) of rat abundance/ activity allowing prediction of rat abundance at any location within the study area. Due to weather and safety concerns, some boards were left out for multiple nights before being photographed. Therefore, we added an offset to the model to account for the increased probability of rat marks as exposure time increases.

We then combined the information at each site i during campaign c to give a number of positive boards m out of the total number able to be read n (at most 10, made up of 5 per night). When some boards at a particular location were exposed for a longer period than others, we did not combine those boards in the model. Therefore, the number of boards positive for rat marks m_{ich} (observation h, at location i during campaign c and exposed for k nights) out of a total n_{ich} boards laid, is modelled using a Binomial likelihood with probability ν_{ich} of a rat mark being present.

$$m_{ich} \sim \mathsf{Binomial}\left(n_{ich}, \nu_{ich}\right) \tag{3.2}$$

A complementary log log (referred to as cloglog and defined as $\mathsf{cloglog}(x) = \log\left(-\log\left(1-x\right)\right)$) link function was chosen for the binomial model as it best reflects the data generating process and correctly offsets for differing exposure times. The following paragraph explains the relationship between the data generating process, the data measurement process, and the derivation of the cloglog link function linking the two together. It is plausible that rat marks occur on tracking boards according to a Poisson process. The count of rat marks on a board y_{ic} is recorded as a binary presence/ absence of marks m_{ic}. If we assume that the number of marks y_{ic} are Poisson distributed with some rate λ_{ic} per night (varying by location i and campaign c) then the probability that a board is marked

$P(m_{ic}=1) = P(y_{ic}>0)$ when exposed for k_{ic} nights is

$$P(y_{ic}>0|\lambda_{ic},k_{ic}) = 1 - P(y_{ic}=0|\lambda_{ic},k_{ic}) = 1 - \exp(-k_{ic}\lambda_{ic}) \tag{3.3}$$

Applying the cloglog transformation to this probability creates linear combination of the log offset $\log k_{ic}$ and log rate of rat marks per night $\log \lambda_{ic}$.

$$\mathsf{cloglog}\left(P(y_{ic}>0|\lambda_{ic},k_{ic})\right) = \log\left(-\log\left(1-\left(1-\exp\left(-k_{ic}\lambda_{ic}\right)\right)\right)\right) = \log \lambda_{ic} + \log k_{ic} \tag{3.4}$$

This allows us to estimate the log rate of rat marks per night using a linear combination of environmental risk factors Z_{ic} and random effects $S_c(i)$ using Binomial mixed effects model with a cloglog link function.

$$\mathsf{cloglog}(\nu_{ich}) = Z_{ic}\beta + S_c(i) + \log k_{ich} \tag{3.5}$$

$$S_c \sim \mathsf{Multivariate\ Normal}(0, \Sigma_c) \tag{3.6}$$

The spatial random effects S_{ic} are a set of values from a spatially continuous process $S_c(x)$ evaluated at locations $i = 1, ..., N$ which describes how the prevalence of rats in the c^{th} campaign varies over space, after all covariate effects have been adjusted for. The model assumes that the $S_c(x)$ are independent copies of a stationary Gaussian process with mean 0, and correlation function $\rho(d) = \mathsf{Corr}\left(S_{ic}, S_{(i-d)c}\right)$. We chose a Matérn covariance structure with ω, variance σ^2 and correlation $\rho\left(d\right) = \left(1 + \frac{d\sqrt{3}}{\phi}\right) e\left\{-\frac{d\sqrt{3}}{\phi}\right\}$, where ϕ describes the rate at which the correlation decays towards zero with increasing distance d. The value of ω determines the smoothness of the process; a process with $\omega = k + 1/2$ is k times mean square differentiable. In practice ω is often difficult to identify precisely from data (H. Zhang, 2004). We chose to fix it at 3/2 which assumes a reasonably smooth surface (i.e. it is more smooth than an Exponential correlation structure ($\omega = 1/2; k = 0$), but less than a Gaussian ($\omega = \text{inf}; k = \text{inf}$) correlation structure). We also include a nugget effect τ^2, such that $\Sigma_c\left(i, i\right) = \sigma^2 + \tau^2$, which represents both measurement error and the within campaign variation in rat prevalence on shorter distance scales than the shortest distances between measured locations.

The model only incorporates covariates that were not strongly correlated and were available for all possible points in the study area to enable predictions at all locations. These are 3-dimensional distance to nearest sewer and public trash, mean rainfall, area of vegetation and impervious surfaces within a 5 meter radius of each location, and a binary domestic status. Preliminary modelling (see section 3.2.5.1) suggested several covariates had non-linear effects. Hence, area of soil and vegetation were fitted with quadratic effects while distance to public dump and distance to open sewer were fitted using linear splines with one knot.

3.2.7.2 Leptospirosis model

The infection status b_{ljc} of individual l at household location j in campaign c is Bernoulli distributed with probability μ_{ljc} of being infected. We chose a logit link function as all individuals were exposed for the same time period (no offset required) and to enable interpretations of risk factors to be made on the odds (or log-odds) scale which is more commonly used in public health research.

$$b_{ljc} \sim \mathsf{Bernoulli}\left(\mu_{ljc}\right) \tag{3.7}$$

$$\mathsf{logit}\left(\mu_{ljc}\right) = \left[X_{jc}, Z^{*}_{jc}\right]\alpha + \theta w^{*}_{jc} + \kappa_l \tag{3.8}$$

$$\kappa_l \sim \mathsf{Normal}\left(0, \delta^2\right) \tag{3.9}$$

where x are leptospirosis model specific covariates, z are environmental covariates (that are also present in model 1), κ_l is a random intercept for individual l, and $w^{\star}$ is the predicted rat activity at location j in campaign c determined by the rat model (Equation 3.10). The marginal posterior distribution of α, θ, and δ is calculated as

$$\pi\left(\alpha, \theta, \delta | X, Z^{*}, Z, m\right) = E_w\left[\pi\left(\alpha, \theta, \delta, w^{*} | X, Z^{*}, Z, m\right)\right] = \int_W \pi\left(\alpha, \theta, \delta | X, Z^{*}, w^{*}\right)\pi\left(w^{*} | m, Z\right) dw^{*} \tag{3.10}$$

where $\pi\left(w^{*} | m, Z\right)$ is the predictive distribution of the predictor in equation 3.10, X_{jc} is the model matrix of leptospirosis only covariate effects, Z and Z^{*} are the model matrices of environmental covariates shared by both models, and $\hat{w}_{jc}$ is the expectation of the rat model posterior at location j in campaign c.

The parameter θ quantifies the effect of the rat prevalence near the subject's home, on the probability that they are infected with leptospirosis. Single point predictions using some measure of centre for each of the parameters does not account for the variation in our estimates of these parameters. Hence, we fit a separate leptospirosis model for each post burn-in iteration h in the rat model, where the predictions from the rat model are calculated using the estimates of the parameters at iteration h.

Parameter		Description
Rat model	h, i, c, k	Observation h from site i in campaign c exposed for k nights
	m_{ich}	Number of positive boards
	n_{ich}	Total number of boards able to be read
	Z_{ic}	Environmental covariates
	ν_{ich}	Fitted probability of rat mark
	β	Estimated covariate effects
	$S_c(i)$	Estimated spatial random effects
	Σ	Estimated covariance matrix (spatial random effects)
	ω, σ, ϕ	Matérn covariance parameters
	τ	Nugget effect
Leptospirosis model	l, c, j	Individual l in campaign c at household location j
	b_{ljc}	Leptospirosis status (0/ 1)
	Z^*_{jc}	Environmental covariates
	X_{ljc}	Other covariates
	μ_{ljc}	Fitted probability of leptospirosis
	α	Estimated covariate effects
	θ	Estimated rat effect
	w^*_{jc}	Estimated rat linear predictor
	κ_l	Individual level iid random effects
	δ^2	Random effect variance

Table 3.1: Model parameter descriptions. See sections 3.2.7.1 and 3.2.7.2 for more details about the rat and leptospirosis models.

The models are then combined to give estimates of the leptospirosis model parameters α, δ and θ which account for the variation in β, ϕ, σ^2, and τ^2 from the rat model. A joint leptospirosis-rat model would have accounted for this implicitly; however, it was not possible to fit this model with our data due to computational fitting issues.

3.2.8 Priors

Weakly informative Normal priors were selected for each of the fixed effect parameters in both the human leptospirosis and rat tracking board models.

$$\beta_w, \alpha_v \sim \mathsf{Normal}\left(0, 10^2\right)$$

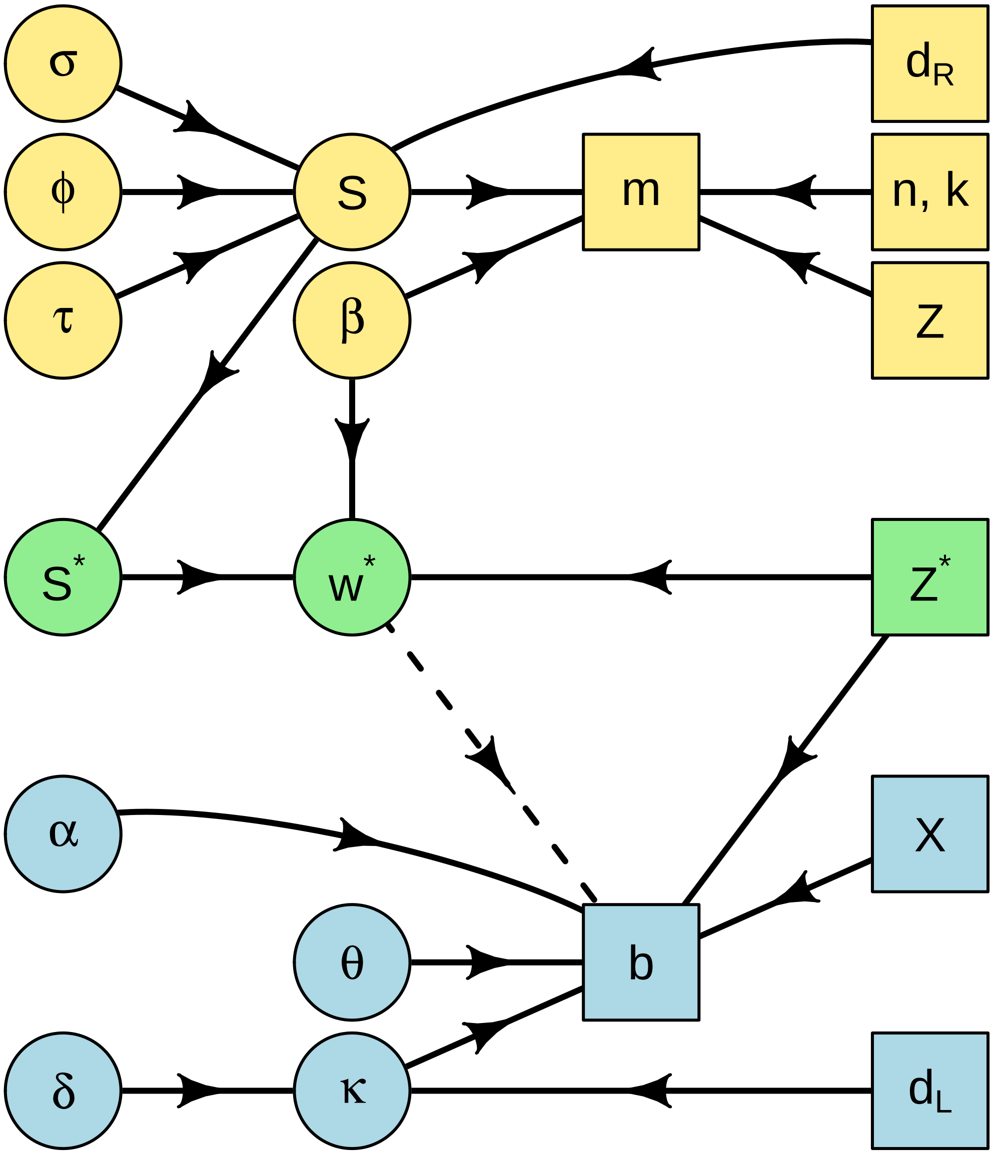


Figure 3.11: Diagram showing the full cut model for the rat and leptospirosis data set. The parameters and data relating to the rat model are shown in yellow (first 3 rows), the predictions (and prediction data) from the rat model are shown in green (4th row), and the parameters and data relating to the human model are shown in blue (rows 5-7). The dotted arrow shows the location of the cut in the model. The cut prevents information from flowing from the blue section of the model into the green or yellow sections. Note, d_R and d_L denote structural data such as locations of tracking boards and study participants which are accounted for using spatially correlated and iid random effects. Prior parameters are not shown in this diagram. See sections 3.2.7.1 and 3.2.7.2 for details of the models and table 3.1 for a table of parameter descriptions.

Priors for the covariance terms were selected to include a wide range of plausible values bounded away from zero for computational stability. Gamma priors, with shape 2 and a small rate, are suggested by Chung et al. (2013) as a good default for variance parameters which may have a mode near to their boundary. This prior has a positive constant derivative allowing the likelihood to dominate if it is strongly curved near 0. We decreased the shape of the Gamma prior for ϕ because it was expected to be reasonably large due to the scale of the spatial coordinates.

$$\tau^2, \sigma^2, \delta \sim \mathsf{Gamma}\,(2, 0.5)$$

$$\phi \sim \mathsf{Gamma}\,(1.5, 0.05)$$

3.2.9 Inference

All models were fitted in a Bayesian framework using the python package PyMC3 (Salvatier, Wiecki, and Fonnesbeck, 2016) using the No-U-Turn Sampler (NUTS). This algorithm is highly effective compared to traditional Metropolis samplers in high-dimensional, highly correlated sampling spaces, as it uses gradient information from the posterior to sample around correlated spaces.

The rat study model was run for 100000 iterations with a burn-in of 2000 iterations. The resulting 100000 iterations were thinned by 100, resulting in 1000 iterations. A leptospirosis model was run for 500 iterations (with a burn-in of 4000, and thinned by 5) using the rat logit probability from each of the 1000 rat model iterations to give a total of 100000 iterations.

3.2.10 Coefficient of determination

We calculate coefficients of determination (CoD), and partial coefficients of determination (PCoD), to show the relative proportions of variation in the dependent variable explained by the covariates included in each model. The CoD is well defined for linear models; however, it is not straightforward to apply to generalised linear models. Many different definitions have been proposed; we prefer to use the definition in D. Zhang (2017) as it reduces to the classical R^2 for linear models and does not overstate the proportion of variance explained by explanatory variables compared to other generalisations of the coefficient of determination to GLM's (Giorgi, 2018).

The variance function is used to define the total variation of the dependent variable after modelling the predictive effects of the independent variables. For a response variable with its mean changing from a to b, its variation moves accordingly along the variance function from $\phi V(a)$ to $\phi V(b)$ (where ϕ is a dispersion parameter). The variation change of the response variable can be measured by the squared length of the variance function $V(\cdot)$ between $V(a)$ and $V(b)$

$$d_V(a, b) = \left\{ \int_a^b \sqrt{1 + [V'(u)]^2} du \right\}^2$$

This measure of distance of variation is more appropriate than the distance used for linear models ($(a-b)^2$) when the underlying variance function is non-linear (as in many popularly considered exponential family distributions (Morris, 1982; 1983)).

This is used to define the CoD for a generalised linear model as follows

$$R_V^2 = 1 - \frac{\sum_{i=1}^n d_V(y_i, \hat{y}_i(X))}{\sum_{i-1}^n d_V(y_i, \hat{y}(1_n))}$$

where $\hat{y}_i\left(X\right)$ is the prediction for y_i based on the covariates X calculated by plugging-in the estimated regression coefficients via maximum likelihood and $\hat{y}\left(1_n\right)$ is the prediction in the null (intercept only) model. In a Bayesian context, we can estimate the distribution of R_V^2 values using the posterior distribution of fitted values.

This can be extended to a CoPD which measures the proportion of variation in the response variable, not explained by a set of predictors, that can be explained by an additional set of predictors. For example, if we consider two set of predictors X_1 and X_2, we can define

$$R_V^2\left(X_2|X_1\right) = \frac{R_V^2\left(X_1, X_2\right) - R_V^2\left(X_1\right)}{1 - R_V^2\left(X_1\right)} = \frac{\sum_{i=1}^{n} d_V\left(y_i, \hat{y}_i\left(X_1, X_2\right)\right)}{\sum_{i=1}^{n} d_V\left(y_i, \hat{y}_i\left(X_1\right)\right)}$$

This can be extended to generalised linear mixed models in the following way (see Giorgi (2018) for further details).

$$R_V^2 = 1 - \frac{E_{S|y,X}\left[\sum_{i=1}^{n} d_V\left(y_i, \hat{y}_i\left(X, S\right)\right)\right]}{\sum_{i-1}^{n} d_V\left(y_i, \hat{y}\left(1_n\right)\right)} \approx 1 - \frac{\frac{1}{B}\sum_{j=1}^{B}\left[\sum_{i=1}^{n} d_V\left(y_i, \hat{y}_i\left(X, S_j\right)\right)\right]}{\sum_{i-1}^{n} d_V\left(y_i, \hat{y}\left(1_n\right)\right)}$$

where S_j are predicted random effects (which may be spatially, temporally or otherwise correlated).

3.2.11 Model Interpretation

There are two measures of variable importance in the previous models. A covariate is defined as statistically significant if the probability that the associated coefficient is larger or smaller than 0 is above 0.95. Although this identifies variables which are highly likely to be associated with the response, we must also consider the effect size for practical significance. We assess the practical effect of each covariate by comparing predictions at lower and upper quartile values for each continuous covariate, and at each level of factors.

This gives an effect size that is roughly comparable between all covariates even when they have non-linear effects on the response or are measured on very different scales and enables an intuitive interpretation of the effect of the covariate on the outcome of interest, accessible to non-statisticians.

Although both the rat mark model and the leptospirosis incidence model use a Binomial likelihood, the differing link functions suggest different comparisons for practical importance. A cloglog link function was used for the rat model; therefore, we interpret the practical significance of covariates using rate ratios (RR) which compare the mean rate of the underlying rat mark deposition process (which we have converted to a binary presence/absence) at the upper and lower quartiles of each covariate. These can be interpreted in the following way: for an increase in covariate x from the lower quartile x_L to the upper quartile value x_U, the rate of marks is changes by a factor of $\mathsf{RR} = \exp\left(\beta_X\left(x_U - x_L\right)\right)$

As a logit link was used for the leptospirosis model, we interpret covariates using odds ratios (OR) where odds are compared at the upper and lower quartiles of each covariate. These can be interpreted in the following way: for an increase in covariate x from the lower quartile to the upper quartile value, the odds of infection are changed by a factor of $\mathsf{OR} = \exp\left(\beta_X\left(x_U - x_L\right)\right)$. Note that the odds of infection may change non-linearly for some covariates.

3.3 Results

3.3.1 Rat tracking board model

Maps of the median predicted probability, of rat marks for each campaign, showed that the distribution of rat activity was not homogeneous throughout the study site; rather, there were distinct hotspots of rat activity throughout the three valleys. Rat activity, particularly in valleys two and three, changed dramatically across short distances (see Figure 3.13), with a practical range of about 50 meters (see Figure 3.10a and Table 3.2).

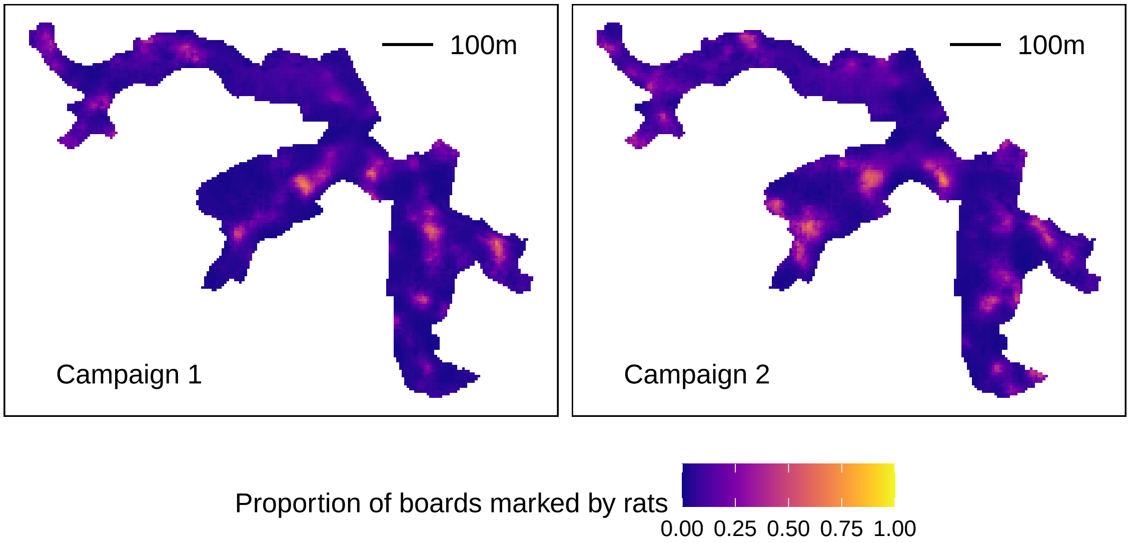


Figure 3.12: Rat model: map showing median predicted probability of rat marks. The predictions used a rainfall value of 1.62mm for campaign 1 and 7.25mm for Campaign 2. These were the observed mean rainfall values throughout the relevant campaign. The fitted probability maps show that rat prevalence/activity is highly heterogeneous within a campaign, and reasonably consistent between campaigns.

Several covariates are significant (see Table 3.2). However, the spatial random effects have a larger influence on the overall surface than the fixed effects (see Figures 3.13). Increasing mean rainfall (RR 1.46 CI: 1.26, 1.66) and increasing area of soil (RR 1.49 CI: 0.95, 2.26) were associated with increased probability of rat marks. Increasing area of vegetation (RR 0.71 CI: 0.43, 1.07), increasing distance to public dump (RR 0.60 CI: 0.38, 0.82) and increasing distance to open sewer (RR 0.76 CI: 0.60, 0.96) were all associated with a decrease in the probability of rat marks. The coefficients for distance to public dump indicate that the probability of rat marks decreases as distance increases, until about 70 meters, after which the probability of rat marks plateaus (increases by a small amount as distance increases). The same effect is observed in distance to open sewer (with the change at about 40 meters), although the increase in probability after 40m is driven by a small

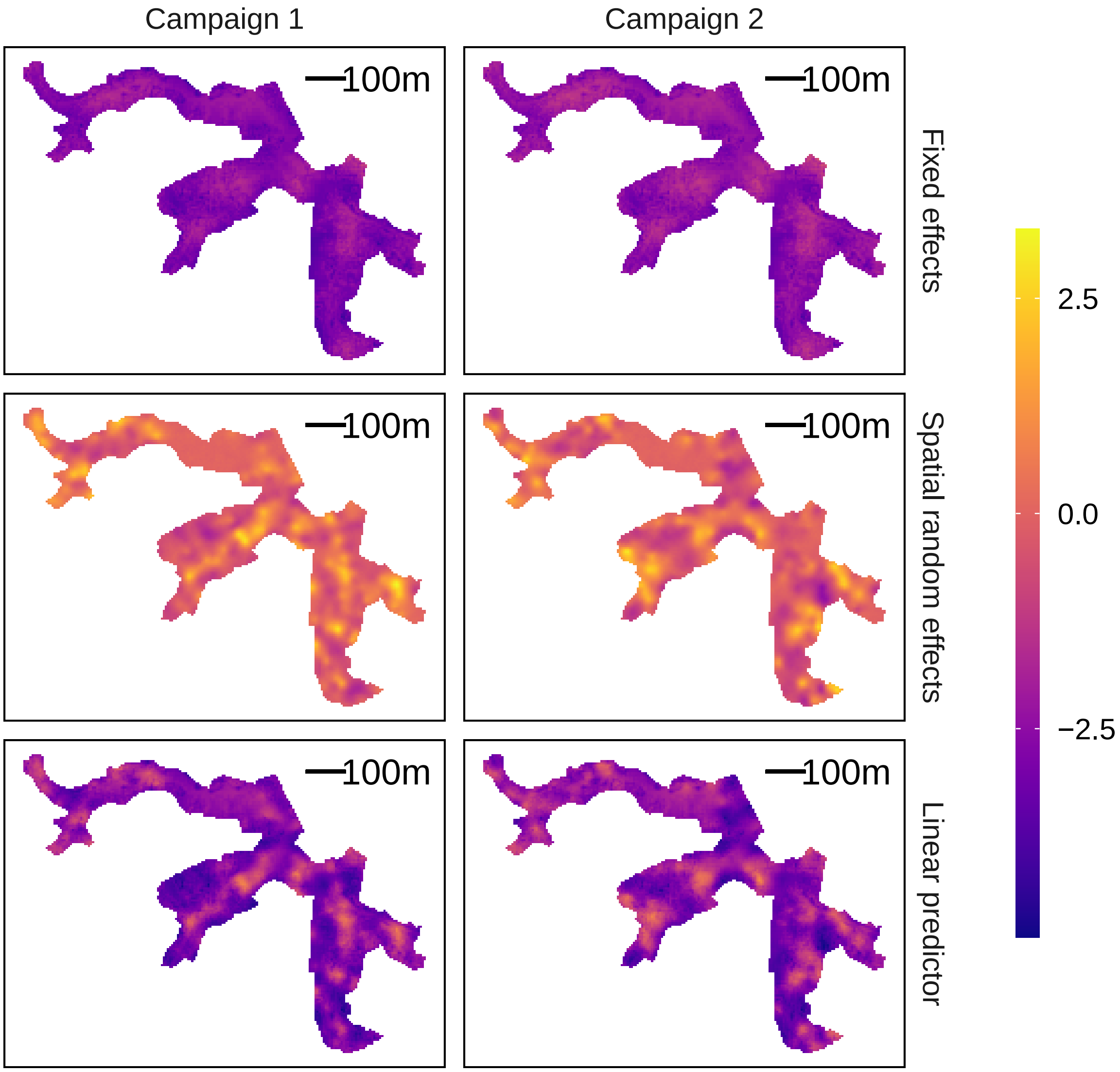


Figure 3.13: Rat model: fitted rat map comparisons. These maps show the relative contributions of the fixed effects (row 1) and the spatial random effects (row 2) to the overall probability of rat marks (shown on logit scale in row 3). Comparison of these maps show that there is significant spatial variation that is not explained by the environmental covariates (fixed effects). The spatial random effects allow identification of areas with unexpectedly high rat activity/ prevalence. Note, fixed effects maps are centred near -2.5 as this map includes the intercept term. The predictions used a rainfall value of 1.62mm for campaign 1 and 7.25mm for Campaign 2.

number of points, so may be spurious. We decided not to include a covariate for campaign as it was highly correlated with mean rainfall due to the two campaigns occurring in the dry and wet seasons respectively. Instead, we let the spatial random effects (which were spatially correlated within a campaign, and independent between campaigns) incorporate any residual temporal trend. If more time periods were available, a full spatio-temporal model could have been fitted; however, this is not suitable for only 2 time periods.

Parameter	**Median**	**Credible interval**	**Prob > 0**	**Prob < 0**	**Sig.**
Intercept	-3.06	(-3.52, -2.61)			
Area soil 5m	0.87	(-0.15, 2.00)	*0.943*		+
Area soil 5m squared	-3.76	(-6.89, -1.04)		**0.992**	**
Area veg 5m	-0.52	(-1.28, 0.17)		*0.934*	+
Area veg 5m squared	2.42	(0.04, 4.50)	**0.984**		*
Mean rainfall	58.32	(38.14, 80.39)	**1.000**		***
Distance 3d public dump	-16.11	(-26.50, -5.45)		**0.999**	***
Distance 3d public dump above 70m	20.52	(6.14, 36.22)	**0.998**		**
Domestic	0.13	(-0.17, 0.39)	0.795		
Distance 3d open sewer	-19.87	(-38.14, -3.29)		**0.983**	*
Distance 3d open sewer above 40m	51.69	(12.29, 84.97)	**0.997**		**
phi	17.31	(12.85, 22.86)			
sigmasq	2.03	(1.40, 2.71)			
tausq	0.39	(0.08, 0.85)			

Significance levels: *** [1 - 0.999) ** [0.999 - 0.99) * [0.99 - 0.95) + [0.95 - 0.90)

Table 3.2: Summary statistics for coefficients of the parameters in the rat tracking board model described in section 3.2.7.1. Credible intervals (CI) are 95% highest posterior density intervals. Statistically significant results (95% probability of being above or below 0) are shown in bold and borderline significant variables (90% probability of being above or below 0) are shown in italics. Distances measured in km; rainfall measured in m; areas measured in proportion.

3.3.2 Leptospirosis model

The interpretation of the leptospirosis model is complicated by the fact that several environmental covariates are present in both models. Due the the cut model implementation,

Covariate		**Rate Ratio**	**Data Quartiles**	
Continuous		$RR_{U/L}$	Data LQ	Data UQ
	Area soil 5m	1.49 (0.95, 2.26)	0.06	0.41
	Area veg 5m	0.71 (0.43, 1.07)	0.00	0.57
	Mean rainfall (m)	1.46 (1.26, 1.66)	0.0003	0.0068
	Distance 3d public dump (km)	0.60 (0.38, 0.82)	0.0305	0.0966
	Distance 3d open sewer (km)	0.76 (0.60, 0.96)	0.0096	0.0176
Binary		$RR_{1/0}$		
	Domestic / non-domestic	1.14 (0.82, 1.45)		

Table 3.3: Rate ratios for covariates in the rat tracking board model described in section 3.2.7.1 with 95% highest posterior density credible intervals. These compare the rate of rat mark deposition at the upper and lower quartile values of the covariates for continuous variables and at 0 and 1 for the binary variables. Distances measured in km; rainfall measured in m; areas measured in proportion.

this complication does not affect the rat model results. Although environmental covariates are unlikely to "cause" disease directly, they can be associated with an increase or decrease in risk for several reasons. Some environments are more or less desirable habitats for rats (which shed leptospires), some environments are more conducive to free leptospire survival (e.g. warm moist areas), whilst others may be associated with a change in risk due to a correlated unknown cause. When rat and environmental covariate levels cannot be controlled directly and are correlated, it is difficult to separately estimate the effect of environmental covariates independently of rat density. Hence, any effect of environmental covariates may be partially absorbed by the predicted rat linear predictor value w^*. Consequently, care must be taken when interpreting the estimated rat exposure and environmental covariate results on the risk of leptospirosis for this data set. For example, the results show that individuals living close to an open rubbish dump have a reduced risk (OR: 0.44, CI 0.28, 0.63); however, rat levels are high in these areas (RR: 0.60, CI 0.38, 0.82). Therefore, the individuals living near open rubbish dumps may have less risk than expected (given the high estimated rat prevalence), but may still have an overall high risk of challenge compared to individuals living in other areas.

The leptospirosis model shows that several covariates have a substantial impact on the probability of leptospirosis (see Tables 3.4 and 3.5) including θ (see Tables 3 and 4). Increasing total rainfall (OR 4.02, CI: 2.42, 6.14), being male (OR 3.75, CI: 2.01, 6.27), increasing age (OR 12.71, CI 5.15, 25.21), and increasing rat covariate (OR 1.03, CI 1.00, 1.07) were significantly associated with increased probability of leptospirosis challenge. Increasing distance to public dump (RR 0.44, CI: 0.28, 0.63) and increasing log income (OR 0.66, CI: 0.32, 1.11) were significantly associated with a decrease in the probability of leptospirosis challenge. The variables distance to open sewer, area soil, area vegetation, race, literacy status, sewer exposure, mud exposure, and flood exposure were not significant. Again, note that although some environmental covariates are not significant, they may still be associated with a change in risk of leptospirosis through their association with rat prevalence changes.

The largest increase in risk is that of age. This is a non-linear effect where increasing age is associated with increased odds of leptospirosis until approximately age 30, after which the odds of leptospirosis decrease slightly. The coefficients for log income indicate that as income increases, the odds of leptospirosis challenge increase a little, until about 40 reais a month, after which they decrease steeply.

Increasing rat prevalence at a person's home is borderline significantly associated with a small increase in their probability of having leptospirosis. This effect is very small compared to the increase in risk by some other covariates, such as being young, male, having a low income, living near a public dump, and being exposed to a large amount of rainfall.

The coefficient of determination for this model is 0.39, whilst partial coefficients of determination for different sets of parameters are as follows: 0.12 for social parameters only (age, income, gender, race, literacy), 0.33 for environmental parameters only (distances to open sewer and public dumps, total rainfall, area of soil and vegetation; the rat linear predictor; and exposure to sewers, mud and floods), and 0.02 for the rat linear predictor only. This shows that the majority of variance in the dependent variable is explained by the environmental variables, but overall, most of the variation is still unexplained. The variance explained by the rat linear predictor is very small, so although the covariate is significant in the model, it has a very small practical effect (also evidenced by the small odds ratio, see Tables 2 and 3). This weak evidence of an effect is not unexpected the estimated cloglog rat abundance covariate has been shrunk towards 0 due to it's errors-in-variables nature (regression to the mean). This makes it more difficult to quantify the effect of rats prevalence on risk of leptospirosis.

3.3.2.1 Variogram

A comparison of the empirical and fitted theoretical variogram from the final rat model shows that, modulo some shrinkage in the independent random intercepts, the fitted spatial random effects are accounting for the spatial correlation in the data well (see Figure 3.10a). Fitting a model allowing differing spatial parameters for each campaign showed that they had very similar posteriors, indicating the parameters were not significantly different between campaigns. Whilst the median fitted spatial surfaces for the two campaigns are similar, there have been some changes to the size and exact location of hotspots from campaign 1 to campaign 2 (see Figure 3.13). If more campaigns of data were available, it may be possible to explain the changes over time using a temporal correlation structure rather than fitting separate surface to each time. However, this is not possible with only 2 time points.

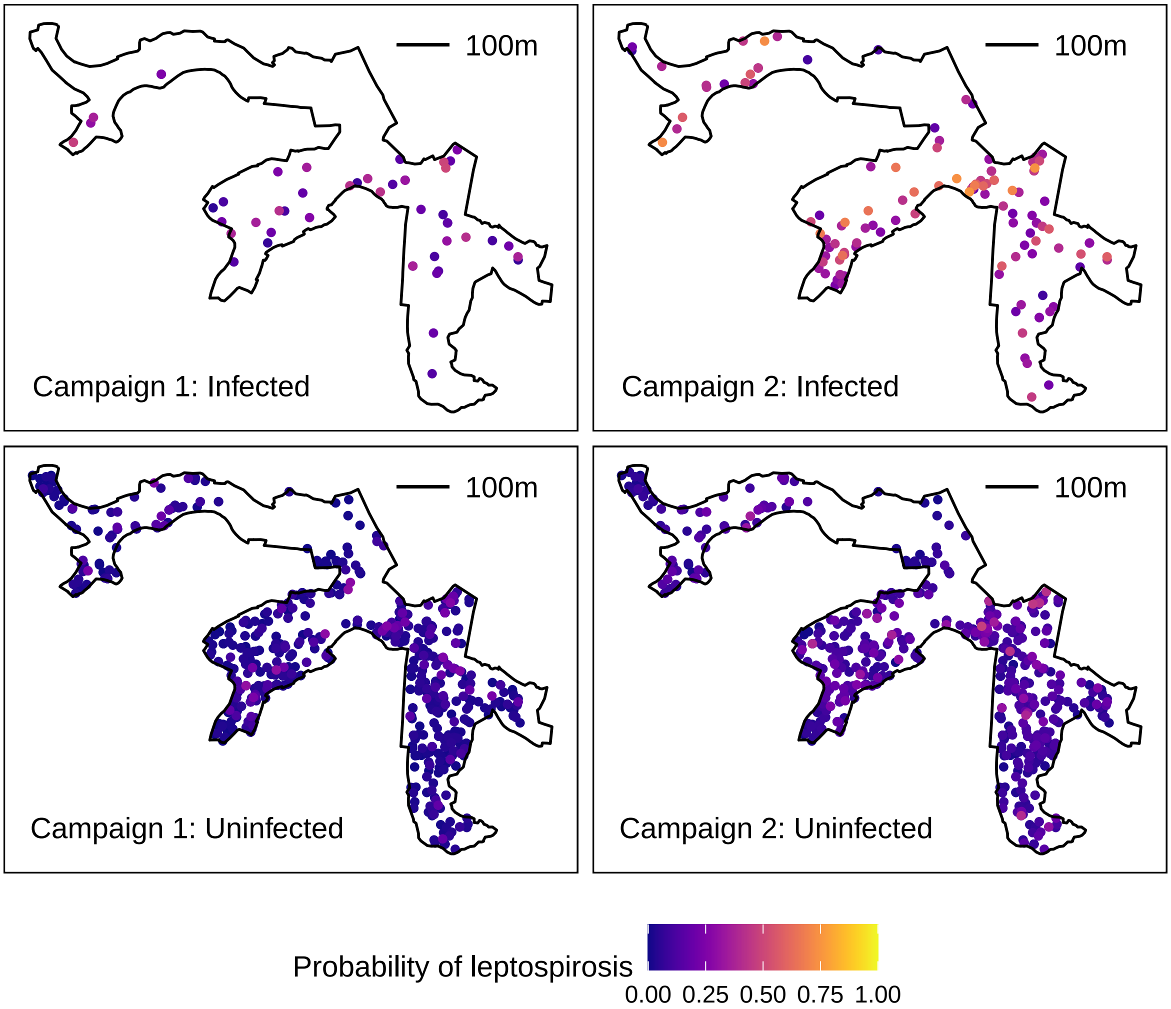


Figure 3.14: Map of fitted median probabilities of leptospirosis for each individual in each campaign. This map shows the fitted probability of infection (colour) by observed infection status in each campaign (facet). There is a clear increase in the risk of leptospirosis in season 2.

Parameter	**Median**	**Credible interval**	**Prob > 0**	**Prob < 0**	**Sig.**
Intercept	-2.12	(-3.36, -0.95)			
Distance 3d open sewer	10.47	(-3.12, 23.63)	*0.938*		+
Distance 3d public dump	-14.00	(-21.00, -7.10)		**1.000**	***
Total rainfall	1.22	(0.84, 1.64)	**1.000**		***
Area soil 5m	0.28	(-0.84, 1.39)	0.694		
Area veg 5m	0.83	(-0.18, 1.89)	*0.944*		+
Age	0.17	(0.12, 0.23)	**1.000**		***
Age (above 30 years)	-0.18	(-0.24, -0.11)		**1.000**	***
Sex (male = 1, female = 0)	1.32	(0.77, 1.88)	**1.000**		***
Race Black	-0.02	(-0.45, 0.46)		0.528	
Literate	-0.11	(-0.67, 0.47)		0.656	
Log income	0.25	(0.02, 0.49)	**0.984**		*
Log income above 40 reias per month	-0.72	(-1.24,-0.21)		**0.997**	**
Sewer contact	0.41	(-0.15, 0.97)	*0.927*		+
Mud contact	0.21	(-0.36, 0.75)	0.774		
Flood contact	0.01	(-0.54, 0.56)	0.513		
Rat linear predictor	0.17	(-0.02, 0.38)	**0.966**		*
σ (sd individual level random effect)	1.65	(1.09, 2.24)			

Significance levels: *** [1 - 0.999) ** [0.999 - 0.99) * [0.99 - 0.95) + [0.95 - 0.90)

Table 3.4: Summary statistics for coefficients of the parameters in the leptospirosis model described in section 3.2.7.2. Credible intervals are 95% highest posterior density intervals. Statistically significant results (95% probability of being above or below 0) are shown in bold and borderline significant variables (90% probability of being above or below 0) are shown in italics. Distances measured in km; rainfall measured in m; areas measured in proportion; age measured in years; income measured in reais per month. Contact variables (sewer, flood and mud) are evaluated near the study participant's home.

Covariate		**Odds Ratio**	**Data Quartiles**	
Continuous		$OR_{U/L}$	Data LQ	Data UQ
	Area soil 5m	1.10 (0.72, 1.57)	0.03	0.37
	Area veg 5m	1.16 (0.96, 1.38)	0.00	0.17
	Cumulative rainfall	4.02 (2.42, 6.14)	0.56	1.70
	Distance 3d public dump	0.44 (0.28, 0.63)	0.0327	0.0906
	Distance 3d open sewer	1.12 (0.97, 1.28)	0.0062	0.0169
	Age (years)	12.71 (5.15, 25.21)	15	42
	Log income	0.66 (0.32, 1.11)	0.00	6.59
	Rat linear predictor	1.03 (1.00, 1.07)	0.033	0.214
Binary		$OR_{1/0}$		
	Male / Female	3.75 (2.01, 6.27)		
	Race Black / Race Other	0.98 (0.58, 1.49)		
	Literate / Illiterate	0.90 (0.45, 1.49)		
	Sewer contact: Yes / No	1.51 (0.78, 2.49)		
	Mud contact: Yes / No	1.23 (0.65, 2.04)		
	Flood contact: Yes / No	1.01 (0.52, 1.64)		

Table 3.5: Odds ratios for covariates in the leptospirosis model described in section 3.2.7.2 with 95% highest posterior density credible intervals. These compare the odds of leptospirosis infection at the upper and lower quartile values of the covariates for continuous variables and at 0 and 1 for the binary variables. Distances measured in km; rainfall measured in m; areas measured in proportion; age measured in years; income measured in reais per month.

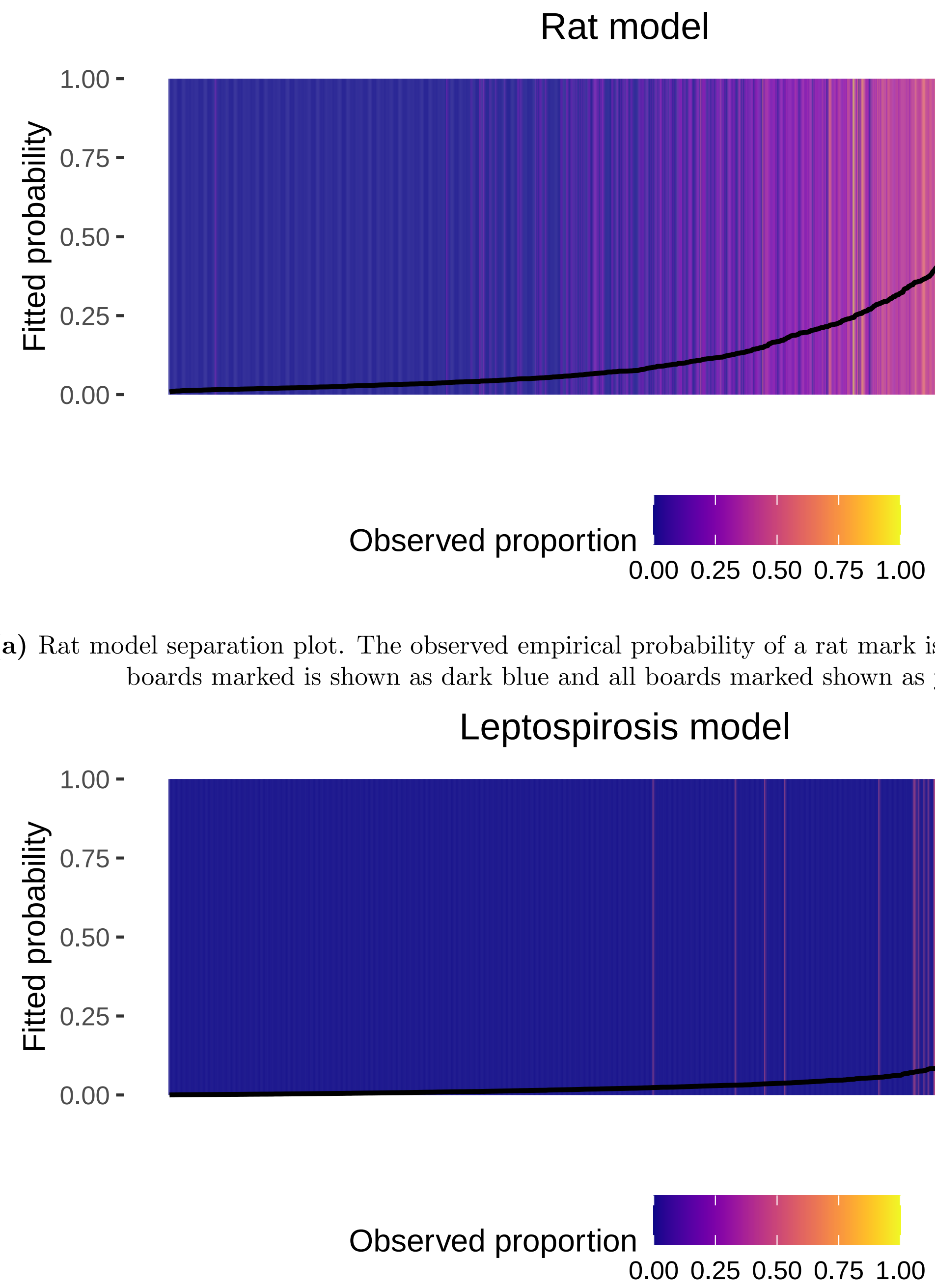
3.3.2.2 Residuals

The separation plots (see Figure 3.15) show that the rat and leptospirosis models fit reasonably well. Leptospirosis is a relatively rare disease even in our high prevalence study area. This means that individuals with covariates indicating they are at high risk for leptospirosis have a much lower than 1 probability of experiencing an event. This is shown on figure 3.15 as the median fitted probability line reaches about 0.75 for the highest risk individuals in our data.

3.4 Discussion

We present the results of an urban slum based study designed to estimate spatio-temporal rat prevalence and human leptospirosis incidence. We followed a cohort of residents for 1 year, and simultaneously tracked rat presence using tracking boards. We developed a temporal mixed model to estimate the contribution of various environmental and social covariates to the risk of leptospirosis, alongside the estimated risk attributable to rat prevalence near the residents' homes. We predicted rat prevalence near the residents homes using a spatio-temporal mixed effects model with environmental covariates. These analyses quantified the relative effects of social and environmental covariates on the risk of leptospirosis in urban slums. Our results agree with previous studies (Maciel et al., 2008a; R.B. Reis et al., 2008a; Costa, F.H. Porter, et al., 2014a; Costa, Ribeiro, et al., 2014; Felzemburgh et al., 2014), indicating that the risk of leptospirosis infection in urban slums is strongly affected by social and environmental features. Additionally, our study quantifies the effect of rat prevalence near the home on the risk of leptospirosis.

The model results clearly identify high risk groups of individuals for leptospirosis, and suggests targets for interventions. Young males with low incomes are at particularly high



(a) Rat model separation plot. The observed empirical probability of a rat mark is given by colour (0 boards marked is shown as dark blue and all boards marked shown as yellow).

(b) Leptospirosis model separation plot. The results are grouped by individual over the two campaigns. Individuals with 0 observed events are shown in blue, individuals with 1 observed event are shown in pink, and individuals with 2 observed events are shown in yellow.

Figure 3.15: Separation plots. Each observation in the model is shown as a vertical line, coloured by the observed empirical probability of success (i.e. leptospirosis positive or marked by a rat). The vertical bands are ordered by median fitted probability of success, and the estimated probability for each observation is shown as a solid black line. When the model has high predictive accuracy, the line should move from near 0 on the left to near 1 on the right, and the colours should reflect low observed probabilities on the left, and high observed probabilities on the right. A model with near 0 predictive accuracy would show a nearly flat black line at y = 0.5 and the coloured bars would be distributed randomly along the x axis.

risk. This agrees with previous studies in this population (Hagan et al., 2016), and studies in other populations (Adler and Pena Moctezuma, 2010; Lau et al., 2010; Goarant, 2016). This is likely due to young, poor males having greater exposure to leptospires in soil and standing water (for example, occupational exposure as a labourer). Although race was not identified as a strong predictor of leptospirosis risk, it is self identifying in our data and is a proxy for social class and poverty. In Pau da Lima, black individuals tend to be less educated, earn less, and live in less desirable areas (such as near open rubbish dumps and sewers). Overall, individuals in the high risk category for socio-economic factors are also more likely to live near open sewers and rubbish dumps, which tend to have high nearby rat activity. The high (sometimes non-linear) correlation between many risk factors makes it difficult to separate out the effect of rats, socio-economic, and environmental factors on the risk of leptospirosis, and highlights the large increase in risk shouldered by the poorest inhabitants of the community.

We show that although the spatial distribution of rat prevalences changes between campaigns, the hotspots remain in a similar location (qualitatively stationary). Although it is unclear whether this stability would be maintained over longer periods, it indicates that the study method is finding stable areas of high rat activity or prevalence. This is partly due to stationary risk factors for rats (open sewers, open rubbish dumps, presence of soil/ mud); however, the qualitative similarity of the two spatial random effect surfaces (Figure 3.13) suggests that there are unmeasured covariates which make some areas more attractive to rats. Many of these risk factors are likely to be stationary across small time scales (e.g. presence of a fruit tree), indicating that it is likely that at least some rat hotspots are reasonably stationary over time (with some stochastic temporal changes). More campaigns worth of rat tracking data is necessary to separate out the spatial and temporal correlations of rat prevalence in the study area. It was not possible to investigate many of these additional covariates as the main focus of the rat model was to build a predictive surface for use in the leptospirosis model. Therefore, only covariates which were able to be measured at all study locations were included in the model.

In our study area there is a distinct socio-economic gradient, with the most impoverished individuals living closer to the valley bottom near open sewers and rubbish dumps. These locations also tend to be associated with increased rat prevalence and increased damage during flooding events. Flooding events are not unusual during the wet season, often causing environmental damage such as mudslides and overflow from open-air sewers in urban slums with poor infrastructure and urban planning. Our study was conducted in an el niño year; significant environmental damage occurred in the study area during campaign 2, including slips, structural damage to housing, and overflowing sewers. These changes likely affected the rat prevalence and leptospirosis risk. This meant it was not possible to separate out the effects of rainfall and rat prevalence on leptospirosis risk from any time trend caused by other unmeasured factors. Due to this severe confounding, campaign was not put as a covariate into the leptospirosis model. Note, campaign was also excluded from the rat model; however, the GP was able to account for any campaign effects by having a non-zero posterior mean for the random effects. Further seasons of data are required to accurately quantify the effect of rainfall and rat prevalence on leptospirosis risk, and to more fully understand the effects of increased rainfall in the absence of environmental damage. However, there is strong evidence in the literature for an increase in leptospirosis following flooding events (A.I. Ko, Galvao Reis, et al., 1999; Barcellos and Sabroza, 2001; Flannery et al., 2001; Karande et al., 2002; Varaiya et al., 2002; Maskey et al., 2006; Lau et al., 2010; Agampodi et al., 2011; Hagan et al., 2016; Naing et al., 2019) which agrees with our model results.

This correlation of many risk factors makes it difficult to independently quantify the increase in risk attributable to high rat prevalence, rainfall and flooding, and living near open sewers and rubbish dumps. This may have contributed to the small estimated effect of rat prevalence on leptospirosis risk, given rats are known leptospirosis reservoirs and have high infection prevalences in the study area (Faria et al., 2008; Costa, F.H. Porter, et al., 2014a; Costa, Wunder, et al., 2015). A better estimate of the increase in risk, directly attributable to living in an area with high rat activity, could be estimated using a controlled factorial study where rat prevalences in different regions of the study area are altered independently of the correlated environmental risk factors (for example, using poisoning). An experiment of this nature has been performed in Pau da Lima, although results are not yet finalised and published.

Due to the correlation between distance to valley bottom, distance to open sewer, and absolute elevation it is not easy to uniquely and concurrently estimate the effects of each covariate. Therefore, I only included distance to open sewer in the model. However, we must consider that any, or all of these 3 variables, may contribute to the observed result of increased risk close to open sewers. It is likely that they contribute in a similar way. The study area was reasonably small, so weather and temperature differences attributable to elevation should be small. Additionally, it is likely that valleys without open sewers collect rubbish and storm water at the bottom, forming a makeshift sewer (particularly during flooding events).

Additional difficulties arise because rat activity/prevalence near an individuals home does not directly correlate with an individuals exposure to rats. Many individuals likely spend significant amounts of time away from home in areas with unknown rat activity, which changes their rat exposure in ways we did not observe, and thus cannot be accounted for in the model.

The MAT method used to detect leptospirosis infections has some significant drawbacks. It easily identifies the first time an individual is infected by detecting seroconversion (the presence of any antibodies). However, further challenges by leptospires are more difficult to detect due to the long decay period for antibodies in humans (Budihal and Perwez, 2014). As titre measurements are very noisy, the paired serology method typically requires a 4 fold increase in titre, between paired measurements, before designating an individual as infected to keep the test specificity high (Chirathaworn et al., 2014). However, it decreases the test sensitivity severely as it ignores titre decay over time. Sensitivity reduces further when measurements are so far apart that multiple challenge events may occur between them, complicating the decay pattern further. Additionally, a high titre measurement relies on an individual having a strong immune response to the pathogen. This can be difficult for immune-compromised individuals, such as those suffering from malnutrition (Bourke, Berkley, and Prendergast, 2016). This may explain why the model showed that individuals on extremely low incomes (under 40 reais a month) had a slightly increased probability of testing positive for leptospirosis as income increased.

3.5 Conclusion

This study combines ecological and epidemiological studies, and establishes a spatio-temporal link between rat prevalence and its effect on disease risk. It highlighted a common difficulty in observational epidemiological studies, trying to estimate the effect of correlated risk factors, and identified a more suitable experiment to directly estimate the effect of rat prevalence on leptospirosis incidence. Despite the limitations of the study, a robust analysis allows us to identify risk factors for leptospirosis which are consistent with results from other studies. In particular, individuals living near open sewers and rubbish dumps have increased risk of leptospirosis, which may be attributed directly to these attributes, or may be due to increased rat exposure and flooding risk during high rainfall events. Addressing these two variables may reduce leptospirosis more than expected by also reducing rat activity and flood risk during high rainfall events. Finally, our results suggest that socio-economic interventions should focus on individuals under 30 years old, particularly low income males.

Bibliography

Sheppe, W., 1965. Characteristics and Uses of Peromyscus Tracking Data. *Ecology*, 46(630-634).

Sheppe, W., 1967. Effect of Livetrapping on Movements of Peromyscus. *American midland naturalist*, 78(471-480).

Brown, L.E., 1969. Field Experiments on Movements of Apodemus Sylvaticus L Using Trapping and Tracking Techniques. *Oecologia*, 2.

Clinton, J.M., 1969. Rats in urban America. *Public health reports*, 84(1).

Lord, R.D., Vilches, A.M., Maiztegu, J., Hall, E.C., and Soldini, C.A., 1971. Frequency of rodents in habitats near Pergamino, Argentina, as related to junin virus. *American journal of tropical medicine and hygiene*, 20.

Hosmer, D.W. and Lemeshow, S., 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in statistics - theory and methods* [Online], 9(10), pp.1043–1069. eprint: `https://www.tandfonline.com/doi/pdf/10.1080/03610928008827941`. Available from: `https://doi.org/10.1080/03610928008827941`.

Morris, C.N., 1982. Natural exponential families with quadratic variance functions. *The Annals of Statistics*, 10.

Lord, R.D., 1983. Rodent control programs: use of the inked tracking board method in Mexico. *Bull pan am health organ*, 17.

Morris, C.N., 1983. Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, 11.

Ann Eileen Miller, B., 1985. NA. *The quarterly review of biology*, 60.

Sandhu, R.S., 1987. Not All Slums are Alike: A Comparison of Squatter Housing in Delhi and Amritsar. *Environment and behaviour*, 19(3).

Taylor, C. and Raphael, M., 1988. Identification of Mammal Tracks from Sooted Track Stations in the Pacific Northwest. *California fish and game*, 74.

Quy, R.J., Cowan, D.P., and Swinney, T., 1993. Tracking as an Activity Index to Measure Gross Changes in Norway Rat-Populations. *Wildlife society bulletin*, 21.

Webster, J.P., Brunton, C.F.A., and Macdonald, D.W., 1994. Effect of Toxoplasma-Gondii Upon Neophobic Behavior in Wild Brown-Rats, Rattus-Norvegicus. *Parasitology*, 109.

Brunton, C.F.A., 1995. Neophobia and Its Effect on the Macrostructure and Microstructure of Feeding in Wild Brown-Rats (Rattus-Norvegicus). *Journal of zoology*, 235(223-236).

Glass, G.E., Johnson, J.S., Hodenbach, G.A., Disalvo, C.L.J., Peters, C.J., Childs, J.E., and Mills, J.N., 1997. Experimental Evaluation of Rodent Exclusion Methods to Reduce Hantavirus Transmission to Humans in Rural Housing. *The american journal of tropical medicine and hygiene* [Online], 56(4), pp.359–364. Available from: `https://doi.org/https://doi.org/10.4269/ajtmh.1997.56.359`.

Diggle, P.J., Tawn, J.A., and Moyeed, R.A., 1998. Model-based geostatistics. *Journal of the royal statistical society: series c (applied statistics)* [Online], 47(3), pp.299–350. eprint: `https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9876.00113`. Available from: `https://doi.org/10.1111/1467-9876.00113`.

Drennan, J.E., Beier, P., and Dodd, N.L., 1998. Use of track stations to index abundance of sciurids. *Journal of mammalogy*, 79.

Mills, J.N. and Childs, J.E., 1998. Ecologic studies of rodent reservoirs: their relevance for human health. *Emerging infect. dis.*, 4(4), pp.529–537.

Ko, A.I., Galvao Reis, M., Ribeiro Dourado, C.M., Johnson, W.D., and Riley, L.W., 1999. Urban epidemic of severe leptospirosis in Brazil. Salvador Leptospirosis Study Group. *Lancet*, 354(9181), pp.820–825.

Barcellos, C. and Sabroza, P.C., 2001. The place behind the case: leptospirosis risks and associated environmental conditions in a flood-related outbreak in Rio de Janeiro. *Cad saude publica*, 17 Suppl, pp.59–67.

Flannery, B., Pereira, M.M., Velloso L de, F., Carvalho C de, C., De Codes, L.G., Orrico G de, S., Dourado, C.M., Riley, L.W., Reis, M.G., and Ko, A.I., 2001. Referral pattern of leptospirosis cases during a large urban epidemic of dengue. *Am. j. trop. med. hyg.*

Levett, P.N., 2001. Leptospirosis. *Clinical microbiology reviews* [Online], 14(2), pp.296–326. eprint: `https://cmr.asm.org/content/14/2/296.full.pdf`. Available from: `https://doi.org/10.1128/CMR.14.2.296-326.2001`.

Glennon, M.J., Porter, W.F., and Demers, C.L., 2002. An alternative field technique for estimating diversity of small-mammal populations. *Journal of mammalogy*, 83.

Karande, S., Kulkarni, H., Kulkarni, M., De, A., and Varaiya, A., 2002. Leptospirosis in children in Mumbai slums. *Indian j pediatr*, 69(10), pp.855–858.

Sarkar, U., Nascimento, S.F., Barbosa, R., Martins, R., Nuevo, H., Kalofonos, I., Kalafanos, I., Grunstein, I., Flannery, B., Dias, J., Riley, L.W., Reis, M.G., and Ko, A.I., 2002. Population-based case-control investigation of risk factors for leptospirosis during an urban epidemic. *Am. j. trop. med. hyg.*

Varaiya, D., Mathur, M., Bhat, M., Karande, S., and Yeolekar, M.E., 2002. An outbreak of leptospirosis in Mumbai. *Indian journal of medical microbiology.*

Kaur, I.R., Sachdeva, R., Arora, V., and Talwar, V., 2003. Preliminary survey of leptospirosis amongst febrile patients from urban slums of East Delhi. *J assoc physicians india*, 51, pp.249–251.

Nams, V.O. and Gillis, E.A., 2003. Changes in tracking tube use by small mammals over time. *Journal of mammalogy*, 84.

UNHS, P., 2003. *The Challenge of Slums: Global Report on Human Settlements Global Report on Human Settlements.* (Technical report). London and Sterling, VA: UN-HABITAT.

World Health Organisation, 2003. *Human leptospirosis guidence for diagnosis, surveillance and control.* WHO Library Cataloguing-in-Publication Data.

Zhang, H., 2004. Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics. *Journal of the american statistical association* [Online], 99(465), pp.250–261. eprint: `https://doi.org/10.1198/016214504000000241`. Available from: `https://doi.org/10.1198/016214504000000241`.

Connors, M.J., Schauber, E.M., Forbes, A., Jones, C.G., Goodwin, B.J., and Ostfeld, R.S., 2005. Use of track plates to quantify predation risk at small spatial scales. *Journal of mammalogy*, 86(991-996).

LaRocque, R.C., Breiman, R.F., Ari, M.D., Morey, R.E., Janan, F.A., Hayes, J.M., Hossain, M.A., Brooks, W.A., and Levett, P.N., 2005. Leptospirosis during dengue outbreak, Bangladesh. *Emerging infect. dis.*, 11(5), pp.766–769.

McBride, A., Athanazio, D., Reis, M., and Ko, A., 2005. Leptospirosis. *Current opinion in infectious diseases.*

McBride, A., Athanazio, D., Reis, M., and Ko, A.I., 2005. Leptospirosis. *Current opinion on infectious diseases.*

Maskey, M., Shastri, J., Saraswathi, K., Surpam, R., and Vaidya, N., 2006. Leptospirosis in Mumbai: Post-deluge outbreak 2005. *Indian Journal of Medical Microbiology.*

Faria, M.T. de, Calderwood, M.S., Athanazio, D.A., McBride, A.J.A., Hartskeerl, R.A., Pereira M. M.and Ko, A.I., and Reis, M.G., 2008. Carriage of Leptospira interrogans among domestic rats from an urban setting highly endemic for leptospirosis in Brazil. *Acta tropica*, 108.

Kara Jose, M., 2008. *Alagados : the story of integrated slum upgrading in Salvador (Bahia)* [Online]. (Technical report). Washington, DC: World Bank. Available from: `http://documents.worldbank.org/curated/en/982251468247212151/Alagados-the-story-of-integrated-slum-upgrading-in-Salvador-Bahia-Brazil`.

Maciel, E.A.P., Carvalho, A.L.F. de, Nascimento, S.F., Matos, R.B. de, Gouveia, E.L., Reis, M.G., and Ko, A.I., 2008a. Household transmission of Leptospira infection in urban slum communities. *PLoS Neglected Tropical Diseases*, 2.

Maciel, E.A.P., Carvalho, A.L.F. de, Nascimento, S.F., Matos, R.B. de, Gouveia, E.L., Reis, M.G., and Ko, A.I., 2008b. Household Transmission of Leptospira Infection in Urban Slum Communities. *PLoS Neglected Tropical Diseases* [Online], 2(1). Available from: `https://doi.org/10.1371/journal.pntd.0000154`.

Promkerd, P., Khoprasert, Y., Virathavone, P., Thoummabouth, M., Sirisak, O., and Jakel, T., 2008. Factors explaining the abundance of rodents in the city of Luang Prabang, Lao PDR, as revealed by field and household surveys. *Integrative zoology*, 3.

Reis, R.B., Ribeiro, G.S., Felzemburgh, R.D., Santana, F.S., Mohr, S., Melendez, A.X., Queiroz, A., Santos, A.C., Ravines, R.R., Tassinari, W.S., Carvalho, M.S., Reis, M.G., and Ko, A.I., 2008a. Impact of environment and social gradient on Leptospira infection in urban slums. *PLoS Neglected Tropical Diseases.*

Reis, R.B., Ribeiro, G.S., Felzemburgh, R.D., Santana, F.S., Mohr, S., Melendez, A.X., Queiroz, A., Santos, A.C., Ravines, R.R., Tassinari, W.S., Carvalho, M.S., Reis, M.G., and Ko, A.I., 2008b. Impact of environment and social gradient on Leptospira infection in urban slums. *PLoS neglected tropical diseases*, 2(4).

Ko, A.I., Goarant, C., and Picardeau, M., 2009. Leptospira: the dawn of the molecular genetics era for an emerging zoonotic pathogen. *Nat. rev. microbiol.*, 7(10), pp.736–747.

Masi, E. de, Vilaca, P., and Razzolini, M.T., 2009. Environmental conditions and rodent infestation in Campo Limpo district, Sao Paulo municipality, Brazil. *Int j environ health res.*

Masi, E. de, Vilaca, P.J., and Razzolini, M.T., 2009. Evaluation on the effectiveness of actions for controlling infestation by rodents in Campo Limpo region, Sao Paulo Municipality, Brazil. *Int j environ health res.*

Adler, B. and Pena Moctezuma, A. de la, 2010. Leptospira and leptospirosis. *Vet. microbiol.*

Lau, C.L., Smythe, L.D., Craig, S.B., and Weinstein, P., 2010. Climate change, flooding, urbanisation and leptospirosis: fuelling the fire? *Trans. r. soc. trop. med. hyg.*

Moreno, E.L., Oyeyinka, O., and Mboup, G., 2010. *State of the World's Cities 2010/2011: Bridging the Urban Divide.* (Technical report). UN Habitat.

Agampodi, S.B., Peacock, S.J., Thevanesam, V., Nugegoda, D.B., Smythe, L., Thaipadungpanit, J., Craig, S.B., Burns, M.A., Dohnt, M., Boonsilp, S., Senaratne, T., Kumara, A., Palihawadana, P., Perera, S., and Vinetz, J.M., 2011. Leptospirosis outbreak in Sri Lanka in 2008: lessons for assessing the global burden of disease. *Am. j. trop. med. hyg.*

Greenhill, B., Ward, M.D., and Sacks, A., 2011. The separation plot: a new visual method for evaluating the fit of binary models. *American journal of political science* [Online], 55(4). Available from: `https://doi.org/10.1111/j.1540-5907.2011.00525.x`.

S. N. Wood, 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1).

ArcGIS, 2012. *ArcGIS 10.1 Environmental Systems Resource Institute, Redlands, California.*

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., and Liu, J., 2013. A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika.*

ENVI, 2013. *Excelis Visual Information Solutions 2.0, Boulder, Colorado.*

Hacker, K.P., Seto, K.C., Costa, F., Corburn, J., Reis, M.G., Ko, A.I., and Diuk-Wasser, M.A., 2013. Urban slum structure: integrating socioeconomic and land cover data to model slum evolution in Salvador, Brazil. *Int j health geogr*, 12, p.45.

Himsworth, C.G., Parsons, K.L., Jardine, C., and Patrick, D.M., 2013. Rats, cities, people, and pathogens: a systematic review and narrative synthesis of literature regarding the ecology of rat-associated zoonoses in urban centers. *Vector borne zoonotic dis.*, 13(6), pp.349–359.

Bacallao, J., Schneider, M.C., Najera, P., Aldighieri, S., Soto, A., Marquino, W., Saenz, C., Jimenez, E., Moreno, G., Chavez, O., Galan, D.I., and Espinal, M.A., 2014. Socioeconomic factors and vulnerability to outbreaks of leptospirosis in Nicaragua. *International journal of environmental research and public health*, 11(8), pp.8301–8318.

Budihal, S.V. and Perwez, K., 2014. Leptospirosis diagnosis: competancy of various laboratory tests. *J clin diagn res.*

Chirathaworn, C., Inwattana, R., Poovorawan, Y., and Suwancharoen, D., 2014. Interpretation of microscopic agglutination test for leptospirosis diagnosis and seroprevalence. *Asian pac j trop biomed.*

Costa, F., Porter, F.H., Rodrigues, G., Farias, H., Faria, M.T. de, Wunder, E.A., Osikowicz, L.M., Kosoy, M.Y., Reis, M.G., Ko, A.I., and Childs, J.E., 2014a. Infections by Leptospira interrogans, Seoul Virus, and Bartonella spp. Among Norway Rats (Rattus norvegicus) from the Urban Slum Environment in Brazil. *Vector-borne and zoonotic diseases*, 14.

Costa, F., Porter, F.H., Rodrigues, G., Farias, H., Faria, M.T. de, Wunder, E.A., Osikowicz, L.M., Kosoy, M.Y., Reis, M.G., Ko, A.I., and Childs, J.E., 2014b. Infections by Leptospira interrogans, Seoul virus, and Bartonella spp. among Norway rats (Rattus norvegicus) from the urban slum environment in Brazil. *Vector borne zoonotic dis.*, 14(1), pp.33–40.

Costa, F., Ribeiro, G.S., Felzemburgh, R.D., Santos, N., Reis, R.B., Santos, A.C., Fraga, D.B., Araujo, W.N., Santana, C., Childs, J.E., Reis, M.G., and Ko, A.I., 2014. Influence of household rat infestation on leptospira transmission in the urban slum environment. *PLoS Neglected Tropical Diseases*.

Felzemburgh, R.D., Ribeiro, G.S., Costa, F., Reis, R.B., Hagan, J.E., Melendez, A.X., Fraga, D., Santana, F.S., Mohr, S., Santos, B.L. dos, Silva, A.Q., Santos, A.C., Ravines, R.R., Tassinari, W.S., Carvalho, M.S., Reis, M.G., and Ko, A.I., 2014. Prospective study of leptospirosis transmission in an urban slum community: role of poor environment in repeated exposures to the Leptospira agent. *PLoS Neglected Tropical Diseases*.

Costa, F., Hagan, J.E., Calcagno, J., Kane, M., Torgerson, P., Martinez-Silveira, M.S., Stein, C., Abela-Ridder, B., and Ko, A.I., 2015. Global Morbidity and Mortality of Leptospirosis: A Systematic Review. *PLoS Neglected Tropical Diseases*, 9(9).

Costa, F., Wunder, E.A., De Oliveira, D., Bisht, V., Rodrigues, G., Reis, M.G., Ko, A.I., Begon, M., and Childs, J.E., 2015. Patterns in Leptospira Shedding in Norway Rats (Rattus norvegicus) from Brazilian Slum Communities at High Risk of Disease Transmission. *PLoS Neglected Tropical Diseases*, 9.

Plummer, M., 2015. Cuts in Bayesian graphical models. *Statistical computing*, 25, pp.37–43.

Torgerson, P.R., Hagan, J.E., Costa, F., Calcagno, J., Kane, M., Martinez-Silveira, M.S., Goris, M.G., Stein, C., Ko, A.I., and Abela-Ridder, B., 2015. Global Burden of Leptospirosis: Estimated in Terms of Disability Adjusted Life Years. *PLoS Neglected Tropical Diseases*, 9(10).

Bourke, C.D., Berkley, J.A., and Prendergast, A.J., 2016. Immune Dysfunction as a Cause and Consequence of Malnutrition. *Trends immunol.*

Goarant, C., 2016. Leptospirosis: risk factors and management challenges in developing countries. *Res rep trop med.*

Hacker, K.P., Minter, A., Begon, M., Diggle, P.J., Serrano, S., Reis, M.G., Childs, J.E., Ko, A.I., and Costa, F., 2016. A comparative assessment of track plates to quantify fine scale variations in the relative abundance of Norway rats in urban slums. *Urban ecosystems.*

Hagan, J.E., Moraga, P., Costa, F., Capian, N., Ribeiro, G.S., Wunder, J., Elsio, A., Felzemburgh, R.D.M., Reis, R.B., Nery, N., Santana, F.S., Fraga, D., dos Santos, B.L., Santos, A.C., Queiroz, A., Tassinari, W., Carvalho, M.S., Reis, M.G., Diggle, P.J., and Ko, A.I., 2016. Spatiotemporal Determinants of Urban Leptospirosis Transmission: Four-Year Prospective Cohort Study of Slum Residents in Brazil. *PLoS Neglected Tropical Diseases* [Online], 10(1). Available from: https://doi.org/10.1371/journal.pntd.0004275.

Panti-May, J.A., Carvalho-Pereira, T.S.A., Serrano, S., Pedra, G.G., Taylor, J., Pertile, A.C., Minter, A., Airam, V., Carvalho, M., Junior, N.N., Rodrigues, G., Reis, M.G., Ko, A.I., Childs, J.E., Begon, M., and Costa, F., 2016. A two-year ecological study of norway rats (Rattus norvegicus) in a brazilian urban slum. *PLoS ONE* [Online], 11(3). Available from: https://doi.org/10.1371/journal.pone.0152511.

PSUP Team Nairobi, 2016. *Slum Almanac 2015/2016.* (Technical report). UN Habitat.

Puckett, E.E., Park, J., Combs, M., Blum, M.J., Bryant, J.E., Caccone, A., Costa, F., Deinum, E.E., Esther, A., Himsworth, C.G., Keightley, P.D., Ko, A., Lundkvist, A., McElhinney, L.M., Morand, S., Robins, J., Russell, J., Strand, T.M., Suarez, O., Yon, L., and Munshi-South, J., 2016. Global population divergence and admixture of the brown rat (Rattus norvegicus). *Proceedings of the royal society b: biological sciences* [Online], 283(1841). eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspb.2016.1762. Available from: https://doi.org/10.1098/rspb.2016.1762.

Salvatier, J., Wiecki, T.V., and Fonnesbeck, C., 2016. Probabilistic programming in Python using PyMC3. *PeerJ.*

Chipeta, M., Terlouw, D., Phiri, K., and Diggle, P., 2017. Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure. *Environmetrics* [Online], 28(1). Available from: https://doi.org/10.1002/env.2425.

Minter, A., Diggle, P.J., Costa, F., Childs, J., Ko, A.I., and Begon, M., 2017. Evidence of multiple intraspecific transmission routes for leptospira acquisition in norway rats (rattus norvegicus). *Epidemiology and infection* [Online], 145(16). Available from: https://doi.org/10.1017/S0950268817002539.

Walker, R., Carvalho-Pereira, T., Serrano, S., Pedra, G., Hacker, K., Taylor, J., Minter, A., Pertile, A., Panti-May, A., Carvalho, M., Souza, F.N., Nery, N., Rodrigues, G., Bahiense, T., Reis, M.G., Ko, A.I., Childs, J., Begon, M., and Costa, F., 2017. Factors affecting carriage and intensity of infection of Calodium hepaticum within Norway rats (Rattus norvegicus) from an urban slum environment in Salvador, Brazil. *Epidemiology and infection* [Online], 145(2). Available from: https://doi.org/10.1017/S0950268816002259.

Zhang, D., 2017. A coefficient of determination for generalized linear models. *The american statistician* [Online], 71(4), pp.310–316. eprint: `https://doi.org/10.1080/00031305.2016.1256839`. Available from: `https://doi.org/10.1080/00031305.2016.1256839`.

Giorgi, E., 2018. On the goodness-of-fit of generalized linear geostatistical models. *Spatial statistics* [Online], 28. One world, one health. Available from: `https://doi.org/https://doi.org/10.1016/j.spasta.2018.01.002`.

Xin, L. and Liu, L., 2018. A simple test procedure in standardizing the power of Hosmer–Lemeshow test in large data sets. *Journal of statistical computation and simulation* [Online], 88(13). Available from: `https://doi.org/10.1080/00949655.2018.1467912`.

Naing, C., Reid, S.A., Aye, S.N., Htet, N.H., and Ambu, S., 2019. Risk factors for human leptospirosis following flooding: A meta-analysis of observational studies. *PLoS ONE.*

Chapter 4

Inference for Partially Observed Seroservalence Data

Poppy Miller[1], Chris Jewell[1]

[1] CHICAS, Faculty of Health and Medicine, Lancaster University, Lancaster, England

Abstract

Infectious diseases, such as covid-19, are responsible for a large portion of the global burden of illness, disability and death (World Health Organization, 2015; World Heath Organisation, 2018). Identification of risk factors can significantly reduce disease burden through mitigating risk using targeted interventions. Identifying risk factors requires a comparison of potential factors among infected and uninfected individuals, which requires accurate disease status labels for all study participants. For many diseases, a significant portion of individuals have only mild, sub-clinical or non-specific symptoms, which makes diagnosis challenging unless testing methodologies have very high specificity and/or sensitivity. Serological diagnosis is used for many diseases including leptospirosis. This method measures antibody concentration or titre values at several time points. However, it results in noisy censored observations that make diagnosis challenging. The criteria for diagnosis often ignores the expected concentration time trends post challenge. These deficiencies make diagnosis challenging, and estimation of time of infection difficult.

We analysed time series data from an ovine leptospirosis challenge trial with 24 animals over 42 days. We developed a Bayesian model which uses a highly stylised mechanistic immune system model to generate a time series of fitted titre values for each individual (given parameters), with an interval censored noise model. This model was used to estimate the parameters governing the shape of the fitted titre curve over time. We then estimated the time of infection for hold-out individuals, given the estimated parameter distributions. Although the immune system is highly complex, our simplified mechanistic model fitted the data well and allowed estimation of reasonably accurate times of infection. The results indicate Pomona infected animals experience a quicker titre rise to a higher peak, followed by a more rapid decline of antibodies compared to Hardjobovis infected animals. This was attributed to differences between serovars in the parameters controlling the rates of pathogen growth and antibody (or equivalently B-cell) death, but not parameters responsible for the rates of antibody/B-cell growth or pathogen death. The results were reasonably robust to exclusion of an individual (using 24-fold cross validation by individual). The estimated times of infection were most accurate for Hardjobovis infected animals, with Pomona infected animals showing a bias of 2-3 days. Two animals had bimodal posterior distributions for time of infection, both of which had unusual titre patterns, and one of which may have in fact been infected prior to the challenge times.

This model shows that the highly complex immune system can be adequately modelled using a reasonably simple process model, and that accurate fitted curves and predictions can be generated even when the data is noisy, heavily censored and from a small sample of individuals. While the results from our particular data set cannot be generalised due to the experimental design, they are consistent with other observations in the literature. This model could be used to estimate titre and pathogen temporal patterns post challenge for many diseases in animals and humans. Additionally, it can be used to estimate the time of infection.

4.1 Introduction

Infectious diseases are responsible for a large portion of the global burden of illness, disability and death (World Health Organization, 2015; World Heath Organisation, 2018). The development of vaccines has significantly reduced the disease burden in the 20th century, however, they are not always available or may provide only partial protection. Where vaccines are not available, targeted interventions can be an effective disease reduction strategy. These interventions aim to reduce disease incidence and/or severity by reducing exposure to risk factors (e.g. environmental alterations, education and behaviour change or vector control) or by decreasing individual susceptibility (e.g. nutritional changes and preventative drugs). Identifying risk factors typically requires a comparison of potential risks among infected and uninfected individuals. An example of this can be seen in Chapter 3, where risk factors for leptospirosis in urban Brazilian slums were developed. A critical aspect of this methodology is identifying infected individuals, often at multiple time points. This can be challenging, particularly for diseases where many individuals have mild, subclinical or non-specific symptoms and/or where tests have low sensitivity/specificity. In chapter 3 we used a standard criteria for diagnosing leptospirosis infection, namely, a 4 fold increase in titre between two paired measurements on an individual. As discussed in Chapter 3, this criteria likely has low sensitivity as it ignores expected temporal titre changes, particularly, titre decay.

In this chapter, we use leptospirosis in sheep as a model system to develop a flexible method to model temporal titre patterns and predict time of infection at an individual level. We develop our model using data from an ovine leptospirosis vaccine trial conducted at Massey University, New Zealand (Fang, 2014). This data is used to estimate temporal titre trends for two serovars of leptospirosis and to estimate the time of infection.

4.1.1 Introduction to leptospirosis in New Zealand

Leptospirosis is a leading zoonotic cause of morbidity and mortality, with an estimated 1.03 million cases and 50,000 deaths occurring annually in the world (Costa, J.E. Hagan, et al., 2015; Torgerson et al., 2015). Less than 10% of challenges result in severe manifestations such as Weil's disease, with the rest being asymptomatic or producing mild generic symptoms such as myalgia and flu (Ko, Goarant, and Picardeau, 2009). Most of these cases occur in developing economies in warm wet climates (Costa, Wunder, et al., 2015) where humans become infected through contact with soil or water contaminated with the urine of an infected mammalian host, particularly through wounds or mucous membranes (P.N. Levett, 2001; McBride et al., 2005; Ko, Goarant, and Picardeau, 2009).

In recent years, the incidence of human leptospirosis in New Zealand has ranked high among developed countries with an average of 2.24 (95% CI 1.85-2.64) cases per 100,000 people a year during 2001-2014 (Health Intelligence Team. and Health Group., 2015). Approximately 80% of the notified cases in New Zealand are associated with farmers, abattoir workers and other occupations requiring frequent contact with animals (Hartskeerl, Collares-Pereira, and Ellis, 2011; Musso and La Scola, 2013; Costa, J. Hagan, et al., 2015; Haake and P. Levett, 2015; Health Intelligence Team. and Health Group., 2015; Frigolett, 2016), and 50% of notified cases require hospitalization (IESR, 2013). This is because unvaccinated livestock are a major reservoir for leptospires and infected animals routinely shed leptospires in their urine (Higgins et al., 1980; Cousins et al., 1989; Gerritsen et al., 1994; Magajevski et al., 2005). This indicates that ovine and bovine animal leptospirosis vaccines have the potential to significantly reduce the number of cases in New Zealand. Animal vaccines and targeted interventions also have the potential to relieve animal suffering due to clinical illness (Cordes et al., 1982; Ayanegui-Alcerreca et al., 2007) and reduce economic losses by reducing associated reproductive failure, decreased milk and meat production (Pearson, Mackie, and Ellis, 1980; Ellis, 1994; Langoni et al., 1999) and reduced growth (Subharat et al., 2012).

The gold standard for detection of exposure to leptospires is a serological method called the Microscopic Agglutination Test (MAT). The following section introduces serological methods in general, before briefly discussing the MAT method for leptospirosis.

4.1.2 General introduction to serology diagnosis methods

When an individual is challenged by an infectious organism, the antigens present on the pathogen trigger the adaptive immune system to begin a complex web of events designed to disable or destroy the invading organism. The adaptive immune system creates novel antigen-binding molecules (antibodies) with high specificity to an invading pathogen by somatically rearranging gene elements (Chaplin, 2010). There are a number of different methods to detect antigens and antibodies under the umbrella term of serological diagnostic methods. These methods exploit the fact that many antigens and their associated antibodies are highly specific to particular pathogens.

The temporal antibody concentration pattern depends on this highly complex immune response and may vary greatly between individuals (see Chaplin (2010) for an introduction to the immune system). Although the process is extremely complicated, a simple temporal antibody concentration pattern is commonly observed. After the body detects a pathogen and recognises specific antigens, B-cells are triggered to begin to proliferate and produce antibodies against them. Once an individual has produced detectable levels of antibodies, they are said to have seroconverted against the pathogen. The concentration of antibodies grows quickly until the infection is cleared, then begins to decline as antibodies die and are not replaced. The body often maintains low levels as an immunological memory long after the pathogen has been cleared to allow a quicker response to future challenges.

If antibody levels decrease to very low concentrations quickly after clearing the pathogen, then antibody presence is indicative of recent challenge. It is more complicated when the antibody concentration decay is slow, as the presence of antibodies in the body may be from historic exposure. This reduces the effectiveness of antibody presence/absence as a diagnostic tool for current infections. Observed antibody concentrations post challenge are typically highly variable between individuals (Fierz, 1998; Antia et al., 2018), and the antibody concentration peak is often short lived. This makes it difficult to diagnose disease based on a high concentration of antibodies as samples are unlikely to be taken whilst the concentration is near its peak, and it is difficult to tell what constitutes a high concentration at an individual level. Therefore, accurately detecting recent challenge requires more complex methods.

Paired serology is a useful diagnostic tool for individuals suspected to have the disease, as it can be repeated several times over a short time period until a diagnosis is confirmed, and results can be evaluated in combination with other diagnostic tests and clinical observations. As a public health research tool, it is more difficult to use, as accurate results require several closely spaced tests (relative to the immune response length). Many public health studies use a single paired serology test and class individuals as infected during the study period if they experienced a large increase in titre value. However, this does not account for the decay in antibody concentration, and likely has a high false negative rate for many diseases, particularly when the time points are far apart. The lack of precision when classifying infected individuals leads to difficulties when trying to correlate cases to potential risk factors.

4.1.3 Introduction to MAT method for leptospirosis

The MAT test detects antibodies to leptospires by incubating patient serum with various serovars of leptospire, and visualising antibody binding (agglutination) using fluorescent tags. If a sample contains the relevant antibodies, the plate will fluoresce indicating that the individual (from whom the sample was taken) has been challenged by the relevant serovar. Relative antibody concentrations can be estimated by recording the number of serial dilutions required to dilute the antibody concentration enough that it is no longer classified as a positive result. A positive result is usually chosen to be 50% agglutination (i.e. half the plate fluorescing). The MAT procedure records the highest 2 fold dilution at which agglutination levels on the plate drop below 50%, or some a priori maximum dilution is reached. As the maximum titre (or equivalently, dilution factor) at which the test changes from positive to negative is not observed directly, the results are heavily interval censored, which contributes to the difficulty in classifying individuals disease status correctly using this data.

Due to the expected large between individual variation, a common method of diagnosis paired serology compares titre values at two time points within an individual and defines a infection as seroconversion or a two or four-fold increase in titre between these measurements. Determination of a positive/negative result can be difficult and is somewhat subjective, resulting in large measurement errors. Seroconversion or a four fold increase in titre is often chosen as the cut off for infection diagnosis (Goris and Hartskeerl, 2014) as this reduces the number of false positives due to measurement error. However, this criteria likely results in many undiagnosed infections as it requires a large positive change in concentration and ignores the expected antibody decay between samples when taken after the peak (Cauchemez et al., 2012; Costa, J.E. Hagan, et al., 2015; Zhao et al., 2017). This is a particular problem when the disease is relatively asymptomatic in some individuals or if the disease causes generic symptoms as detection rates are already low. Cauchemez et al. (2012) and Zhao et al. (2017) suggest using a 2 fold change in titre which reduces the number of false negatives, at the expense of increasing the number of false positive results.

The MAT method requires a high level of technical expertise and a large panel of live *Leptospira* standard cultures (Niloofa et al., 2015) which creates a significant risk of laboratory acquired leptospirosis. Other methods of detection are not suitable for routine use due to technical limitations and low sensitivity (Niloofa et al., 2015). Although titre data does not measure antibody concentrations directly, they are strongly and positively correlated. This justifies using titre measurements as a proxy for antibody concentrations when modelling. Other methods of measuring antibody concentration, such as avidity, optical density or ELISA, are also suitable for modelling.

4.1.4 Current statistical models

There have been many approaches to modelling data of this type, particularly in the last 5 years (including 2 R packages). Our approach is most similar to those of de Graaf et al. (2014) and Borremans et al. (2016). de Graaf et al. (2014) developed a simple within-host model using a system of differential equations (ODE) describing the interaction between a pathogen and the immune system and the waning of immunity after clearing of the pathogen. Their initial model has been extended significantly, including the production of an R package *seroincidence* which estimates the frequency of seroconversions (infections) in a sampled population. This package does not allow for interval censored antibody concentrations, and uses cross sectional antibody data which makes it inappropriate for use with our data. Borremans et al. (2016) uses splines to estimate time-series antibody titre patterns (initial model), then predicts individual time of infection from cross sectional data using covariates such as age, season and presence of pathogen in a separate second model. Spline methods are extremely flexible and perform well when the time trend is complex and there are a large number of individuals with frequent titre measurements, as occurred in the example data for this paper. However, splines (and other semi-parametric methods) tend to perform poorly in the presence of sparse data. The authors note that this semi-parametric model can be replaced with a mechanistic model when a suitable one can be created (Borremans et al., 2016). The paper does not consider censored response data, but this could be incorporated simply by modifying the likelihood. We also use a 2 stage model which initially estimates titre dynamics, then use these results to predict time of infection at an individual level. However, we prefer to use a simple phenomenological within-host model to describe the temporal antibody patterns as this performs better with sparse data and gives additional insights into the differences in immune response between individuals and pathogen types. We additionally allow for censored biomarker data.

In contrast, A.J. Kucharski et al. (2015) and A. Kucharski et al. (2018) develop a model which combines individual specific infection history with a shared antibody response process. They use a dynamic model which multiplies a fixed underlying titre with changes expected due to processes such as antigenic superiority, boosting and cross-reactivity. In A.J. Kucharski et al. (2015) the log titre is modelled using a Poisson distribution which is only suitable when all log titres are measured using the same baseline dilution (to give integer log titres). Our data used two labs to analyse the titres, and each used a different baseline dilution, resulting in many non-integer log titre values. A. Kucharski et al. (2018) in contrast use an interval censored Normal likelihood for the log titre values which reflects the data generating process better than using a discrete observation distribution. This model can be used to infer complex interactions between related strains of a pathogen, but cannot be used to estimate the within-host immune response to different strains. Following on from this work Hay et al. (2019) created an R package called *serosolver* which jointly infers prior infections and cross-reactive antibody dynamics using antibody titre data. The process model used to generate temporal fitted antibody measurements in *serosolver* is essentially a linear combination of the contribution of antibody responses from each prior infection. Although this efficiently combines temporal antibody measurements from multiple strains of the pathogen, it does not provide additional information about the mechanistic antibody changes within host when challenged by the relevant pathogens.

In contrast, Owers and Diggle (n.d.) use a simple non-linear model to capture the temporal antibody titre pattern. The model allows for censored observations and incorporates covariate effects through a log-linear dependence of one of the model parameters on a vector of covariates x. This approach does not require large sample sizes (as do many of the non-parametric spline based models), but does not provide any additional insight into the possible mechanistic causes of observed differences in titre pattern between individuals or groups.

Our data contains time series interval censored titre measurements for individual animals with a single known challenge time. *seroincidence* is not suitable as it focuses on estimating infection frequency using cross sectional antibody data and does not allow for interval censored data. *serosolver* does not use a mechanistic within-host model and incorporates multiple infections with different strains over years. We do not have enough data to reliably estimate a time trend semi-parametrically as done in Borremans et al. (2016), and we do not have large numbers of additional covariates to include in the model. Hence, we develop a stylised mechanistic within-host model allowing for interval censored observations and prediction of infection time. Modelling the full immune system response mechanistically is not practical, as it is highly complex, and we only observe the relative concentration (or titre) of a particular antibody at a (usually small) number of times for a small number of individuals. Hence, we restrict ourselves to a highly stylised within-host model describing both the interaction between pathogen and the immune system and the waning of immunity after clearing of the pathogen. Although this model is highly simplified, it captures the essential time trend well and fits the data adequately.

4.2 Motivating dataset

We develop our method using data from an ovine leptospirosis vaccine trial which was previously analysed by Fang, 2014 as part of a PhD thesis at Massey University. This data is particularly useful for model development as it has multiple measurements per animal over time and a known (shared) infection time.

A commercial research organisation, Estendard Ltd., was contracted by a vaccine company to run challenge trials on sheep and cattle with serovars Hardjobovis (*L. borgpetersenii* sv Hardjobovis) and Pomona (*L. interrogans* sv Pomona). The Hopkirk Leptospirosis Research Laboratory at Massey University was given the data from these trials, which was subsequently analysed by Fang (2014) as part of a PhD thesis. Trial B consisted of 8 animals challenged with Hardjobovis, whilst trial C had 16 animals challenged with Pomona (AES approval number 019/09). Note, data from challenge A was not included as it was a pilot study; data from challenge D was not included as it was performed in cows rather than sheep. All animals were clinically healthy and screened to be seronegative (at a minimum dilution of 25). See Fang (2014) for more details around experimental design and data collection. Around 2mL of Hardjobovis or Pomona culture (containing between 10^7 - 10^9 leptospires) was administered to each animal (1mL administered into the nasal cavity, and 1mL administered through conjunctival instillation). The inoculations were administered on 3 successive days.

Titres were measured for each animal between 5 and 8 times over 42 days. This time range adequately captured both the rise and fall of titre values in response to leptospirosis challenge. Each serum sample was initially diluted either 1:24 or 1:25, followed by serial two fold dilutions until 50% agglutination no longer occurred or until 7 serial dilutions had been performed. This produces titre values y that are left ($0 < y <$ min_titre where min_titre was either 24 or 25), interval censored, or right ($y >$ max_titre where max_titre was either 3200 or 3072). Note, we could also define the left censored observations as interval censored, however, it is more efficient to calculate the likelihood when they are considered left censored (as 0 is the minimum supported value for the chosen likelihood distribution). Figures 4.1 and 4.2 show the interval censored titre values for each sheep in the study.

All animals experience some increase in titre values, with Pomona infected animals peaking earlier and higher than Hardjobovis infected animals. The titre patterns are reasonably consistent between animals within a serovar.

4.3 Methods

Our approach first develops a model which combines a highly stylised mechanistic process model with an interval censored observation model. We use this model to estimate the antibody titre dynamics, which are combined with new data to predict individual times of infection.

4.3.1 Model

The limiting dilution assay process creates interval-censored observations of the underlying antibody concentration. We first formulate a process model for the underlying antibody concentration at any time $t > \tau_0$, where τ_0 is the inoculation time, then develop an observation model which accounts for the interval censored nature of the measured data.

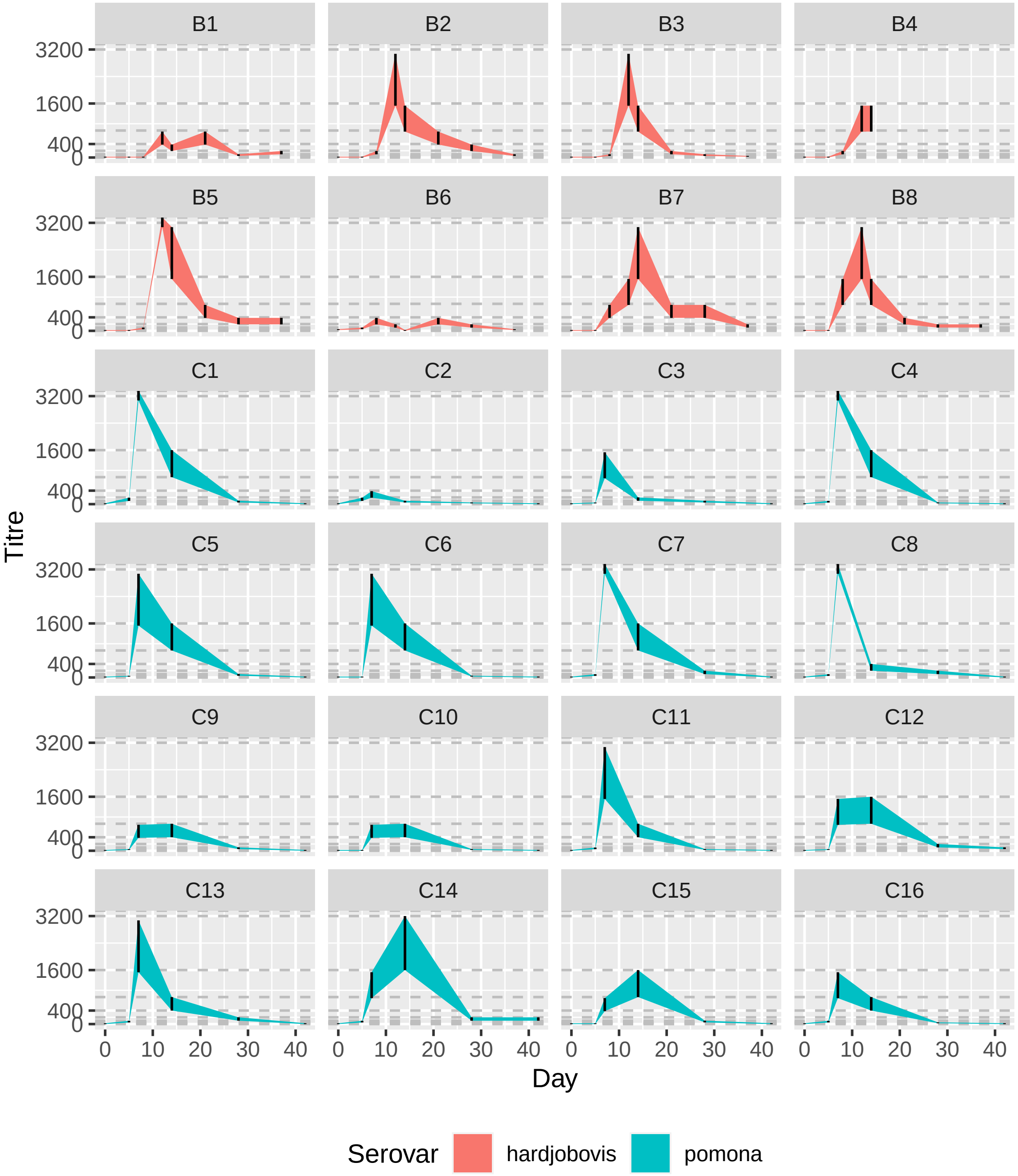


Figure 4.1: Observed censored titre values for each animal. The black vertical line segments show the observed titre intervals which are left ($0 < y <$ min_titre where min_titre was either 24 or 25), interval censored, or right ($y >$ max_titre where max_titre was either 3200 or 3072). The observations are interpolated using ribbons coloured by serovar (red for Hardjobovis and blue for Pomona). The facet labels give the individuals unique ID, and which trial they belonged to (B or C). The horizontal grey dotted lines show the censoring intervals.

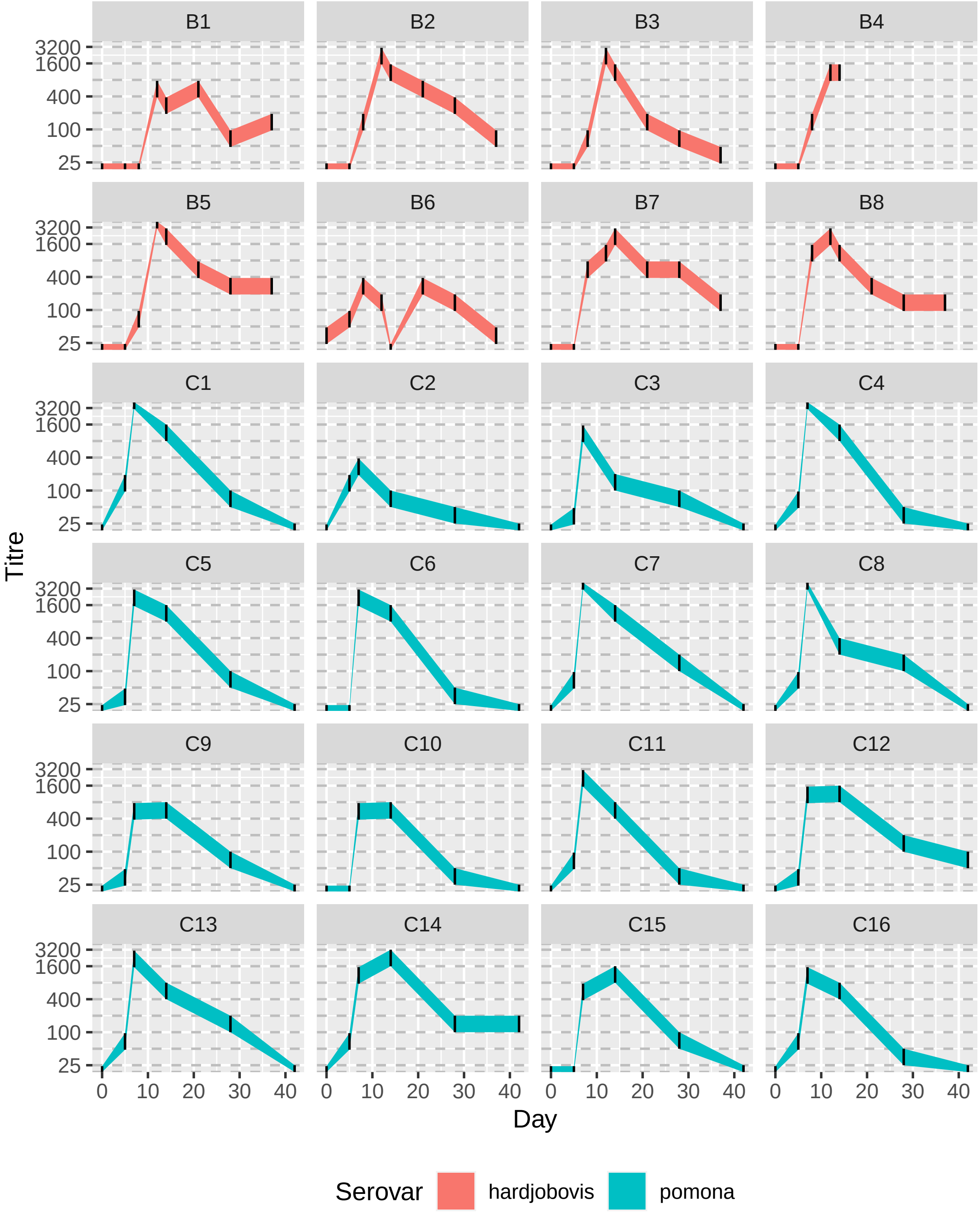


Figure 4.2: Observed censored titre values for each animal (log scale). The black vertical line segments show the observed titre intervals which are left ($0 < y <$ min_titre where min_titre was either 24 or 25), interval censored, or right ($y >$ max_titre where max_titre was either 3200 or 3072). The observations are interpolated using ribbons coloured by serovar (red for Hardjobovis and blue for Pomona). The facet labels give the individuals unique ID, and which trial they belonged to (B or C). The horizontal grey dotted lines show the censoring intervals. When plotted on the log scale, censoring intervals are approximately equal.

We extend the model to estimate of the time of infection for each individual using a cut model (as previously described in chapter 3).

4.3.1.1 Mechanistic antibody concentration pattern model

The trajectory of antibody concentration over time is modelled using a similar approach to a Lotka-Volterra predator prey model (Lotka, 1910; de Graaf et al., 2014) using a set of ordinary differential equations (ODE's) to describe the mean fitted curve over time. We assume that the presence of antigen (A) triggers clonal expansion of antibody-secreting B-cells (C) which produce antibodies in proportion to the number of cells currently present. The rate of B-cell expansion $\beta C_t A_t$ and the rate of antigen removal $\delta C_t A_t$ both depend on the current number of cells and antigens. We allow the antigen (or at least the pathogen to which the antigen is a part) to undergo a reproduction process, resulting in growth at rate γA_t which depends only on the current number of pathogens in the system. Finally, we assume that cells undergo a death process at rate ρC_t, which allows the antibody concentration to reduce to near 0 after clearing the antigens from the system.

We denote by A_t the antigen concentration (or any measure for the severity of the infection) and by C_t the B-cell concentration at time t. The dynamics of A and C are determined by

$$\frac{\mathsf{d}C}{\mathsf{d}t} = \beta C_t A_t - \rho C_t \tag{4.1}$$

$$\frac{\mathsf{d}A}{\mathsf{d}t} = \gamma A_t - \delta C_t A_t \tag{4.2}$$

where $\beta C_t A_t$ is the rate of B-cell replication, ρC_t is the rate of B-cell death, $\delta C_t A_t$ is the rate of antigen consumption as a result of antibody-antigen binding and γA_t is the rate of pathogen growth at time t. We note that the parameters β, δ, ρ and γ are quantities that describe the emergent properties of complex immunological interactions which we consider here to be linear with respect to C and A. This set of ODE's can result in cyclical patterns with respect to C and A (as see in typical Lotka-Volterra prey-predator curves), however, the data and priors are used to constrain the fitted curves to include only one peak (cycle) within the range of the data.

We observe a time series of C (the titre values) for each individual, but are missing information on the associated antigen concentrations A. This makes it impossible to estimate the starting values (A_0, C_0), therefore, these parameters either need strong informative priors, or can be fixed. Here, we choose to fix both initial values as follows: $\{A_0 = 1, C_0 = 1\}$.

4.3.1.2 Observation model

The censored titre observations for individual i at time t (lower limit y_{it}^L and upper limit y_{it}^U) are assumed to be Log Normally distributed with mean μ_t

$$\left(y_{it}^L, y_{it}^U\right) \sim \textsf{Censored Log Normal}\left(\mu_t \middle| \left\{A_0, C_0, \beta, \rho, \delta, \gamma\right\}, \sigma\right) \tag{4.3}$$

where $\mu_t = C_t$ is the solution to the set of differential equations in the previous section. The censored log normal density function is defined as follows for an observation y which may be uncensored (4.4), left censored (4.5), interval censored (4.6) or right censored (4.7) with censoring limits L and U

$$p\left(y|\mu,\sigma\right) = f\left(y|\mu,\sigma\right) \tag{4.4}$$

$$p\left(y > U|\mu,\sigma\right) = 1 - F\left(U|\mu,\sigma\right) \tag{4.5}$$

$$p\left(L < y < U|\mu,\sigma\right) = F\left(U|\mu,\sigma\right) - F\left(L|\mu,\sigma\right) \tag{4.6}$$

$$p\left(y < L|\mu,\sigma\right) = F\left(L|\mu,\sigma\right) \tag{4.7}$$

where

$$f\left(x|\mu,\sigma\right) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{\left(\log x - \mu\right)^2}{2\sigma^2}\right) \tag{4.8}$$

$$F\left(x|\mu,\sigma\right) = \Phi\left(\frac{\left(\log x\right) - \mu}{\sigma}\right) \tag{4.9}$$

and Φ is the cumulative distribution function of the standard normal distribution.

4.3.1.3 Covariates

The complex interaction between an individual's immune system and the invading pathogen results in highly variable antibody concentration curves between individuals, particularly if they are exposed to different pathogens. We can partially capture this variability by allowing different fitted curves for each pathogen or sub-type of a pathogen. However, there may still be unexplained individual specific differences that are not accounted for by this approach. We use a hierarchical model to pool information from many individuals, allowing estimation of unique individual specific titre curves which vary around a shared pathogen specific curve. As the fitted curves depend on a complex interaction between the ODE parameters (β, δ, ρ and γ), we allow all four parameters to vary with serovar, that is, we treat serovar as a fixed effect in the model. Although we could also allow each ODE parameter to vary by individual using random effects, this results in a complex model with many parameters which can easily over fit given that titre data is often sparse, noisy and heavily censored. Instead, we incorporate a multiplicative independent identically distributed (iid) individual level random effect ζ_i which applies the serovar level fitted curve. For example, a data set with serovars $j = 1, 2$ and individual ids $i = 1, 2, ..., I$ results in the following model

$$\frac{\mathsf{d}C_j}{\mathsf{d}t} = \beta_j CA - \rho_j C \tag{4.10}$$

$$\frac{\mathsf{d}A_j}{\mathsf{d}t} = \gamma_j A - \delta_j CA \tag{4.11}$$

$$\zeta_i \overset{\text{iid}}{\sim} \textsf{Log Normal}\left(0, \sigma_\zeta\right) \tag{4.12}$$

$$\mu_{ijt} = \mu_{jt}\zeta_i \tag{4.13}$$

$$\left(y^L_{ijt}, y^U_{ijt}\right) \sim \textsf{Censored Log Normal}\left(\mu_{ijt}, \sigma\right) \tag{4.14}$$

where serovar effects are defined multiplicatively. For example, the pathogen growth rate for each serovar γ_j can be defined as follows

$$\gamma_j = \begin{cases} \gamma_1, & \textsf{if serovar 1} \\ \gamma_1\gamma_2, & \textsf{if serovar 2} \end{cases} \tag{4.15}$$

This allows us to easily compare pathogen growth rates between serovars, as γ_2 defines the difference between Serovars 1 and 2 (note, $\gamma_2 = 1$ when there is no difference). Other functional forms are possible, but we leave this extension as future work.

4.3.1.4 Estimation of time of infection

An individual's time of infection is typically unknown, and of great interest to the researcher. Hence, we extend the model to enable estimation of the time of infection. The time of infection τ_0 is the estimated length of time before observations began (at $t = 0$) giving estimated observation times $t^* = t + \tau_0$. Therefore, the likelihood for observed titre range $\left\{y^L, y^U\right\}$ at time t^* is

$$\frac{\mathsf{d}C}{\mathsf{d}t^*} = \beta CA - \rho C \tag{4.16}$$

$$\frac{\mathsf{d}A}{\mathsf{d}t^*} = \gamma A - \delta CA \tag{4.17}$$

$$\left(y^L_{t^*}, y^U_{t^*}\right) \sim \textsf{Censored Log Normal}\left(\mu_{t^*}, \sigma\right) \tag{4.18}$$

Estimating the time of infection jointly with β, δ, ρ and γ over-parametrises the model. This is because there are 3 parameters (β, γ and τ_0) that can affect the shape of the initial increase post-challenge, causing strong posterior correlations particularly when data is sparse and heavily censored, and when pathogen concentrations are unobserved (see Figure 4.3). This correlation can be greatly reduced by simplifying the model to assume no pathogen growth (i.e. removing γ from the model), however, it is biologically known that the pathogen multiplies within host, and the fitted curve shows significant bias during the first 5 days when pathogen growth is ignored. Additionally, there is no data between times $t \in (-\tau_0, 0)$ to inform the fitted curve, which can result in poor fits (for example showing decreases in titre concentrations pre $t = 0$). It also results in wide posterior densities for all parameters, indicating that the model has serious non-identifiability issues. It is possible that this could be mitigated by constraining the fitted curve to be monotonically increasing until the peak concentration, however, this is left as future work. Instead, we

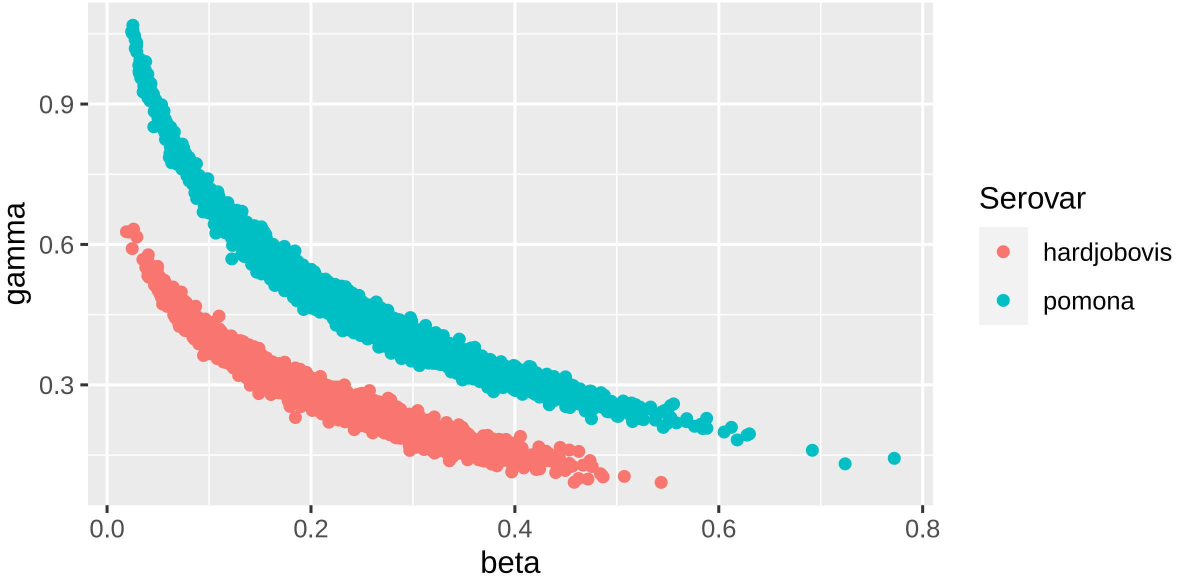


Figure 4.3: Correlation between model parameters. The scatter plot shows the strong observed correlation between model parameters β and γ. These parameters control the rate of antibody production and the rate of pathogen growth, respectively, and both affect the observed antibody concentration increase in the initial times post infection.

focus on a more practical approach, predicting the time of infection given "known" ODE distributions using a cut model.

The cut model is implemented in practice by first estimating the ODE parameters (β, δ, ρ and γ) using data from individuals with known infection times (henceforth called the "ODE model"). These parameters describe how the disease progresses over time and are assumed to be applicable to other similar individuals. The time of infection τ_0 is then estimated for a new individual, conditional on the previously estimated joint posterior distribution of the ODE parameters (henceforth referred to as the "τ_0 model"). Uncertainty in the ODE parameters is incorporated into the τ_0 model by numerically integrating over the posterior distribution of the ODE model. More formally, the marginal posterior distribution of τ_0 is calculated as

$$\pi\left(\tau_0|y, y^*\right) = E_\theta\left[\pi\left(\tau_0, \theta|y, y^*\right)\right] = \int_\Theta \pi\left(\tau_0|y^*, \theta\right)\pi\left(\theta|y\right)d\Theta \tag{4.19}$$

where $\theta = \{\beta, \rho, \gamma, \delta\}$ is the vector of ODE parameters, y is the data from the individuals used to fit the ODE model, and y^* is the data from the individuals used to fit the τ_0 model. This mimics a situation in which the dynamics of infection are "known" (for example, through previous studies), but the time of infection for an individual of interest is not. This is likely to be more practically useful than a fully joint model as most epidemiological studies are unlikely to contain enough informative data to estimate both the within-host antibody trend and the time of infection. Instead, studies could estimate the time of infection given a reference titre curve distribution estimated using alternative data sources (e.g. challenge or vaccination trials).

4.3.2 Massey Data

The Massey data set has few individuals and all have known infection times. Therefore, we use a leave-one-out cross validation procedure to estimate the time of infection for each individual. For each hold-out individual g, with associated data y_g, we first fit the ODE model using data y_{-g} (that is, data from all individuals except g). We then estimate τ_0 for individual g using y_g giving

$$\pi\left(\tau_0|y_{-g}, y_g\right) = E_\theta\left[\pi\left(\tau_0, \theta|y_{-g}, y_g\right)\right] = \int_\Theta \pi\left(\tau_0|y_g, \theta\right)\pi\left(\theta|y_{-g}\right)d\theta \tag{4.20}$$

In practice this is achieved by fitting the τ_0 model many times using draws from the posterior of the associated ODE model, and combining the resulting posterior distribu-

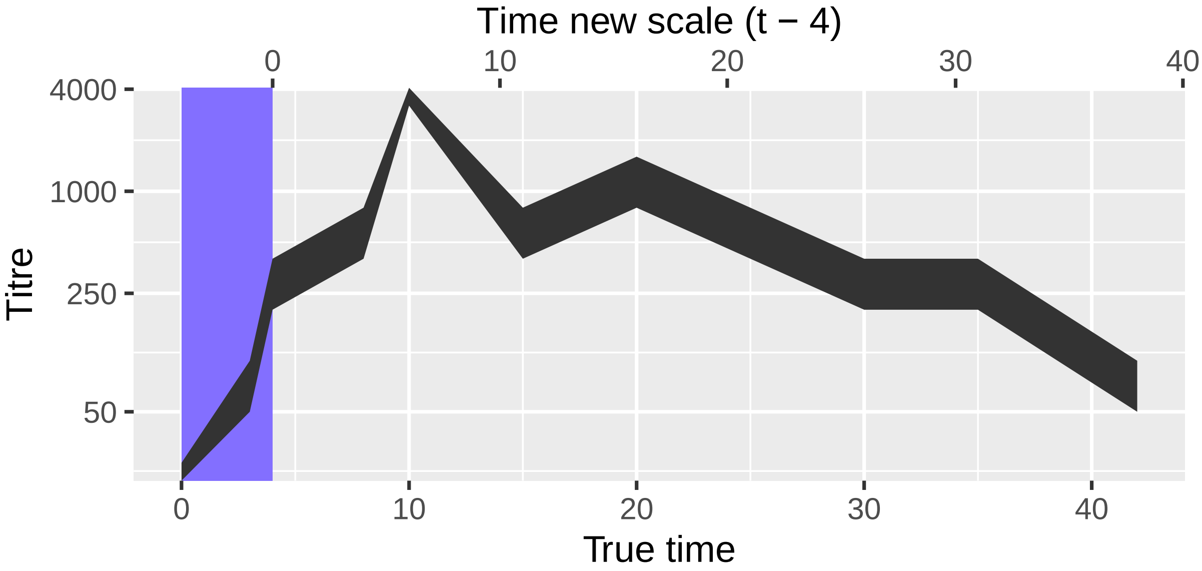


Figure 4.4: Massey data timeline shift. The grey ribbon shows an example observed titre trajectory over time. The true infection time occurred at $t = 0$. The new time scale is shifted by subtracting 4 days from the true observed times, and deleting any observations that occur before 0 in the new time scale (those falling in the blue zone).

tions.

The animals were challenged 3 times (once daily) from Day 0, therefore, the true time of infection is $t = \tau_0 = 0$, with additional doses at times $\tau_0 + 1$ and $\tau_0 + 2$. Our model assumes that the time of infection occurs before the first observation (prior to $t = 0$), which makes inference challenging due to the true parameter value occurring on a boundary. To mimic a situation in which the time of infection is unknown, we reset the time scale to begin at $t = -4$, and retain only data from day 4 onwards (see Figure 4.4). Day 4 was chosen so that only a small portion of the data was removed due to the small sample size. This means that the true time of infection is $\tau_0 = 4$, with additional doses given during the following 2 days (exact time of administration not recorded). Note that we plot all results on the original time scale. Given the small number of data points per individual (particularly after removing the measurements from before Day 4), we choose to fix the random effect ζ_i to 1 when predicting the time of infection.

Hardjobovis is considered to be in a maintenance host relationship with at least cattle and deer, and probably sheep, in New Zealand, and therefore, the disease caused by this serovar may behave differently to that of a Pomona infection (Heuer et al., 2012). Therefore, we allow unique but related mean antibody titre curves for each individual, and for each infecting serovar. We used fixed effects for each serovar and random effects for each individual. Note, in this dataset serovar and trial ID are totally confounded. Hence, whilst results may discuss differences between serovars, it is possible that the observed differences were caused by other trial specific effects. Whilst this is an unavoidable limitation with our data set, the data was sufficient for model development. The resulting model can easily be applied to other data sets.

4.3.3 Inference

Models were fitted in a Bayesian framework using the development version of `rstan` (Stan Development Team, 2020). This package implements No-U-Turn Samplers for each parameter (NUTS), and incorporates an ODE solver. This algorithm is highly effective compared to traditional Metropolis samplers in high-dimensional, highly correlated continuous sampling spaces as it uses gradient information from the posterior to sample around correlated spaces. This is particularly important when using ODE models as the parameters are nonlinearly related and often correlated. These models were first implemented using a custom program written in `R` (a heavily modified version of the code used to fit the `sourceR` models in chapter 2), however, the adaptive metropolis samplers were very inefficient, so the decision was made to re-implement the model in (Stan Development Team, 2020) as an ODE solver was recently added.

The ODE models (full, and excluding each ID) were run with 3 chains for 1500 iterations with the first 500 iterations discarded as burn-in. A cut model was implemented to estimate the time of infection τ_0 for each individual. Two hundred and fifty τ_0 models were fitted using randomly selected draws from the associated ODE model. The posteriors from the τ_0 models were combined together to form a single joint posterior. The τ_0 models were run with 2 chains for 300 iterations with the first 250 iterations discarded as burn-in. Model diagnostics (including divergence, tree depth, energy, effective sample size and Rhat values) were checked for each model.

Starting values cannot be totally randomly selected because many combinations of ODE parameters create invalid differential equations or fitted curves which are wildly inconsistent with the observed data. Hence, we set the starting values to be close to the "correct" values (as assessed using the parameters estimated with the prior generation algorithm as discussed in section 4.3.3.1). This makes it more challenging to find alternative modes if the posterior is multimodal, and to detect non-convergence, however, it is necessary to generate fitted models in a reasonable time frame.

4.3.3.1 Priors

The base model parameters β, δ, ρ and γ have independent half Normal (positive) priors. We chose to fit the random effects on the log scale for computational reasons. Hence, the random effects $\zeta_i^* = \log(\zeta_i)$ were Normally distributed with mean 0 and standard deviation σ_ζ. Both square root variance components σ and σ_ζ, and the initial time of infection τ_0 have independent Gamma distributed priors.

$$\beta, \rho, \delta, \gamma \sim \mathsf{Half\ Normal}(0, 1) \tag{4.21}$$

$$\zeta_i^* \sim \mathsf{Normal}(0, \sigma_\zeta) \tag{4.22}$$

$$\sigma_\zeta, \sigma, \tau_0 \sim \mathsf{Gamma}(2, 0.5) \tag{4.23}$$

In some situations, prior information may be available and the incorporation of informative priors may improve model identifiability and posterior results. However, it is challenging to create informative priors for the ODE parameters based on information about titre curves (which is typically what an expert would have). This can be accommodated by fitting a model where fit is judged using a weighted combination of squared deviations from a chosen set of values (for example, peak titre value, peak titre time, time to decay to below 50). This is easily accomplished using optimisation algorithms such as with the `optim` function in `R`. The optimisation can be repeated several times using a range of values (for example, minimum and maximum a priori expected peak titre time) to give a range of plausible ODE parameters. These can then be converted into reasonable prior distributions (using method of moments, or another suitable method).

4.4 Results

The stylised nature of the process model (alongside fixing A and C), means that interpretation of model parameters β, δ, ρ and γ requires caution, particularly when comparing between individuals who are suspected to have very different true initial values for C and A. The animals used in our example were all given a standard dose of the pathogen and had not been exposed to leptospirosis before, so it is reasonable to assume that their initial pathogen and antibody levels were similar to each other. Given known (or assumed constant) starting values for a set of individuals, comparison of the relative posterior distributions for the ODE parameters gives information about possible differences in antibody dynamics between individuals and serovars. In other situations, it may be better to only compare the fitted antibody titre curves and estimated times of infection. The results described in section 4.4.1 are from the ODE model fitted using all the data. The τ_0 model results in section 4.4.2 reflect the ODE and τ_0 models fitted using hold-out individuals.

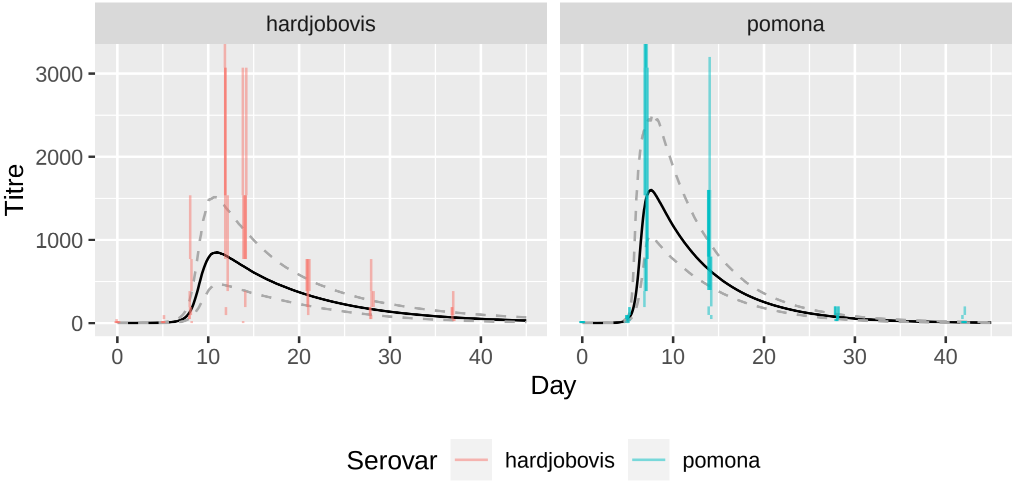
4.4.1 ODE model

Overall, the model fitted the data well, particularly when incorporating the individual level random effects (see Figures 4.5a, 4.5b and 4.6). The fitted curves show clear differences between serovars, and smaller differences between individuals within a serovar.

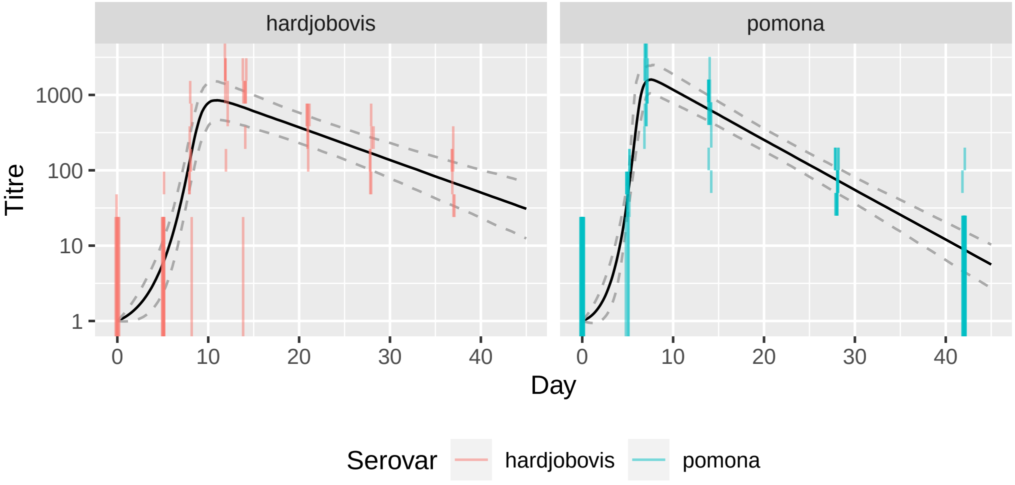
The fitted curves can be summarised in many ways. We choose to characterise the curves by estimating the duration and timing of high titre levels (where titres larger than 100 are considered high given the minimum detection level of 25 for a positive MAT result), when the peak titre occurs, and what the value of the peak titre is (see Figure 4.7 and Table 4.1 for results):

- `Day start`: first time at which the fitted titre exceeds 100
- `Day peak`: time when the fitted titre is highest
- `Day end`: first time, after Day peak, that the fitted titre decreases below 100
- `Titre peak`: titre value at Day peak

Alternative summary values may be more applicable depending on the shape of the curves, and questions of interest. The day start values are significantly earlier for Pomona infected animals with an average start day of around 5.4 for Pomona and 7.8 for Hardjobovis. The Pomona challenged animals experience their peak titre at approximately day 7.4, whilst Hardjobovis infected animals peaked on average 3.4 days later (on day 10.8). The Pomona infected animals also returned to low titre values more quickly, with an average end time of 26.3 (giving approximately 20.9 days where the mean titre was above 100). Hardjobovis animals had fitted titres above 100 for an average of 25.6 days (returning to sub 100 titres by day 33.4). The average maximum titre experienced by Hardjobovis infected animals was



(a) Titre curves.



(b) Log titre curves.

Figure 4.5: Fitted titre curves for each serovar. Median value given as a solid black line with 95% credible intervals shown as dashed grey lines. The observed data are shown as vertical segments coloured by Serovar (red for Hardjobovis and blue for Pomona). The observed data have been plotted with a small amount of jitter on the x axis and at 50% opacity to avoid over-plotting. The fitted curves were calculated without the random effects (equivalently, with random effects fixed at 1 for all animals), giving an overall mean fitted curve per serovar. The results are shown on the original (a) and log (b) scales.

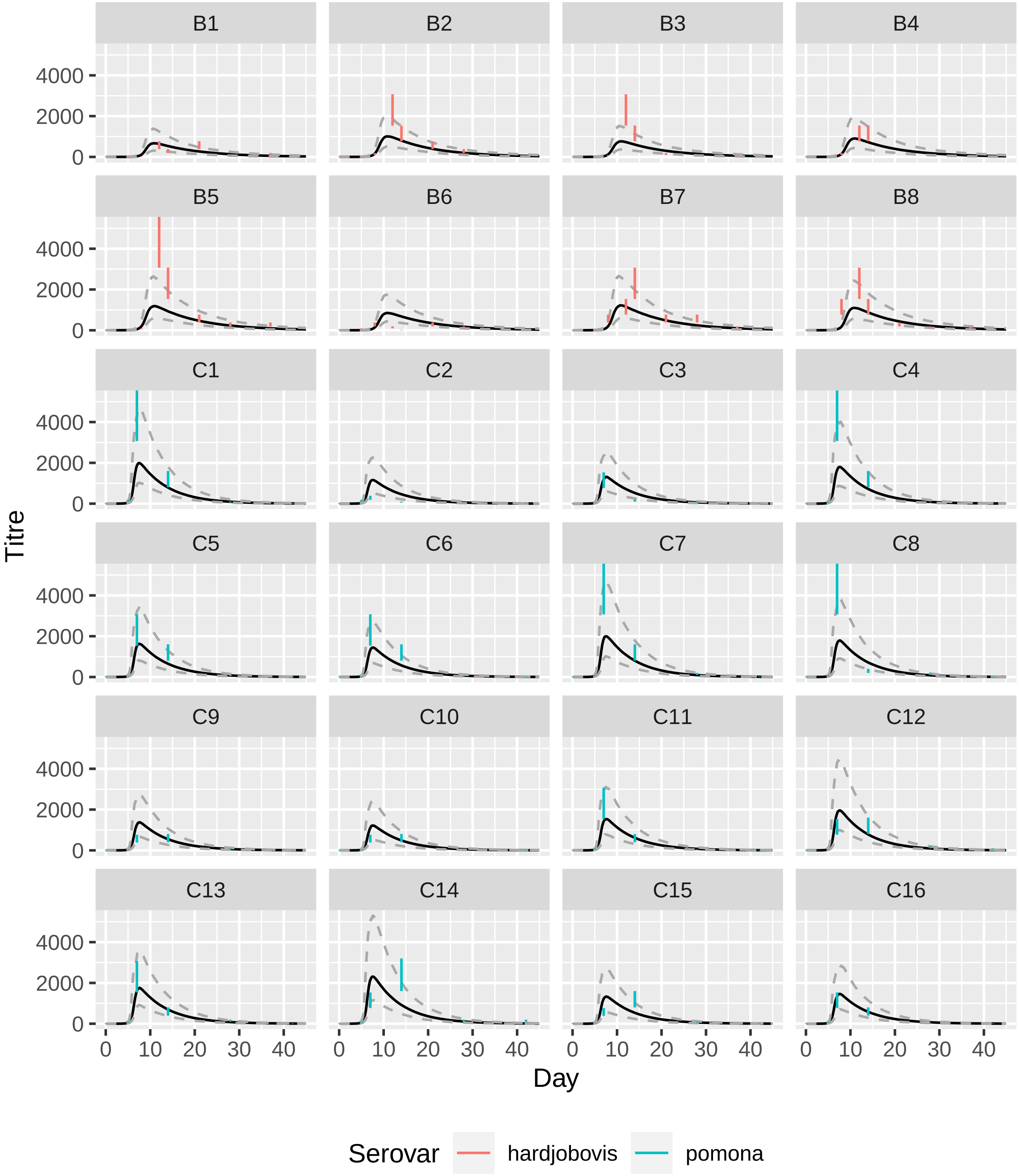


Figure 4.6: Fitted titre curves for each individual. Median value given as a solid black line with 95% credible intervals shown as dashed grey lines. The observed data are shown as vertical segments coloured by Serovar (red for Hardjobovis and blue for Pomona). The fitted curves were calculated with the random effects, giving an individual specific curve.

	Hardjobovis			**Pomona**		
	Mean	*Lower*	*Upper*	*Mean*	*Lower*	*Upper*
Day start	7.8	7.3	8.4	5.4	5.1	5.6
Day peak	10.8	9.8	12.1	7.4	6.5	8.2
Day end	33.4	28.5	39.5	26.3	24	28.5
Titre peak	900	510	1519	1725	1102	2736
Titre peak (log)	6.8	6.2	7.3	7.4	7.0	7.9

Table 4.1: Summary statistics for fitted curves by Serovar. The results give the mean and 95% credible interval values for each summary statistic (calculated without random effects). See definitions of each summary statistic in section 4.4.1.

900. Pomona infected animals experienced an average maximum titre of nearly double this value (1725). Note that the estimated time of peak titre is shared between all individuals within serovar by construction (as random effects per individual are multiplicative on the entire fitted curve).

The possible causes of the observed differences in fitted curves between serovars can be assessed by considering the relative values of the ODE parameters. The results show that there is no evidence that β or δ vary between serovars (the posterior distributions for β_{Pomona} and δ_{Pomona} are both centred on one, see Figure 4.8), whilst there is strong evidence that γ and ρ are larger for Pomona infected animals (the distributions of γ_{Pomona} and ρ_{Pomona} are located significantly above 1). This means that the estimated pathogen rate of growth and cell rate of death are both larger for Pomona challenged animals, whilst the rates of cell growth and pathogen death remain similar. This causes the fitted titre curves for Pomona challenged individuals to peak earlier, achieve a higher titre value and decay more rapidly (see Figures 4.7 and 4.5a).

The random effects variance is smaller than the error variance (see 4.8) indicating small estimated between host differences. This may be due to the large censoring hiding small differences between individuals, or may be due to the chosen model structure (where random effects are multiplicative on the entire fitted curve). It is possible that random

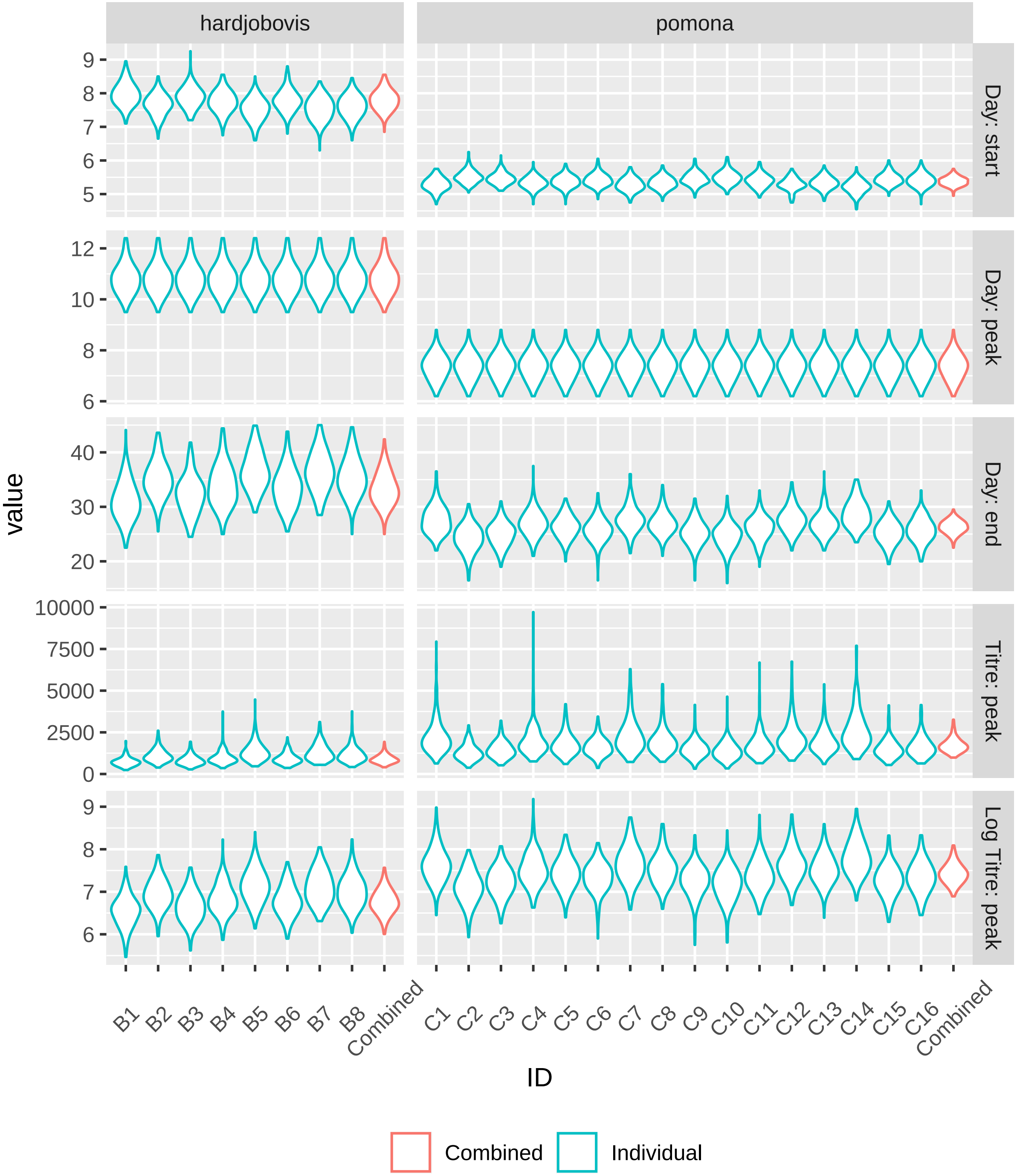


Figure 4.7: Distributions of summary statistics for fitted curves. The Individual (green) plots include the random effect, whilst the Combined (red) plots do not. Day start is the distribution of the day at which the fitted titre first exceeds 100. Day peak is the day at which the maximum titre occurs. Day end is the first time post peak that the fitted titre retreats to below 100. Max titre is the maximum fitted titre during the study period. The results show large between serovar differences, however, between individual differences are reasonably small.

effects on the ODE parameters would explain more of the variance, at the cost of a much more complex model. Although the individual differences are reasonably small on the log scale, they allow the upper bounds of the fitted curves (on the linear scale) to nearly double when comparing individuals at the extremes of a serovar (see Figures 4.6 and 4.7).

4.4.2 τ_0 models

The ODE model was fitted using data from all individuals and the results summarised in section 4.4.1. To estimate the times of infection for each individual, we refit the ODE model with each individual excluded from the data set in turn. The posterior densities of the ODE parameters from these models were used in the τ_0 models to estimate the time of infection for each holdout individual. This also enables us to assess the ODE model for robustness and sensitivity to the data from each individual by comparing the marginal posterior distributions for parameters of interest. The results show that overall the model was very robust, with all parameters showing similar marginal posterior distributions. Individual B6 was shown to have the largest influence, particularly on β, γ and σ. This is likely due to the observations for individual B6 following a different pattern compared to other Hardjobovis infected animals. Fang (2014) concluded that it was possible that individual B6 was infected before the trial began due to this unexpected pattern, particularly as B6 had a positive titre on Day 0. We chose to retain individual B6 in our data set as it is possible that the positive Day 0 titre was a false positive and we did not want to remove such a large portion of the data, particularly as it did not have an unduly large influence on the parameter estimates (see Figure 4.9). Additionally, it may be possible to estimate the true time of infection for B6, if it were indeed infected prior to Day 0. Individuals C2, C9 and C10 (from group Hardjobovis) had similar observed titre curves, which were much lower than the others in their group. Excluding only one of these individuals has little effect on the model results. This is likely due to two reasons, excluding only one

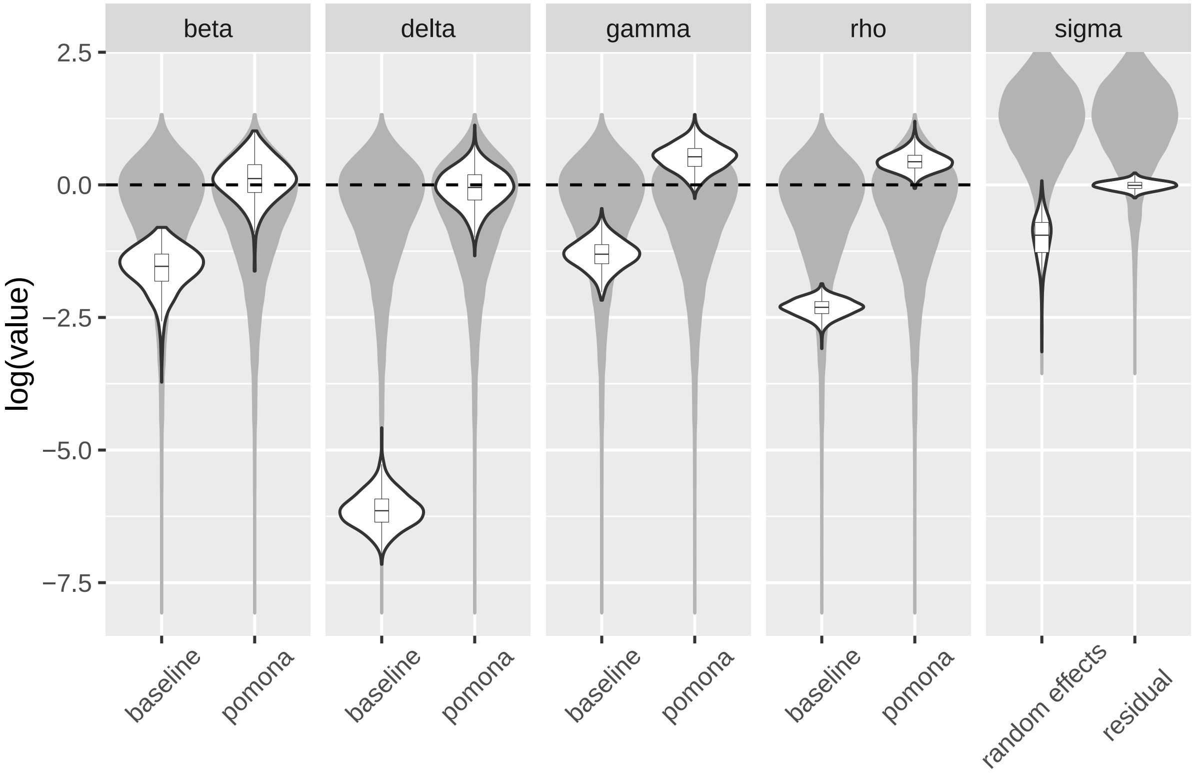


Figure 4.8: Log marginal posterior distributions for main ODE model parameters. The prior distributions are shown as solid grey violin plots and the posterior distributions are shown as white violin plots with a black outline. The baseline serovar in this model was hardjobovis, hence baseline ODE parameters (β, δ, ρ or γ) give the value for hardjobovis, and pomona parameters give the multiplier required with baseline for the pomona value. For example, $\beta^*_{\mathsf{hardjobovis}} = \beta_{\mathsf{baseline}}$ whilst $\beta^*_{\mathsf{pomona}} = \beta_{\mathsf{baseline}} \times \beta_{\mathsf{pomona}}$. A non-baseline ODE parameter is non-significant if it's value is 1, hence, a dotted black line has been added at $\log(1) = 0$. The results show that β and δ are not significantly different between Pomona and Hardjobovis challenged animals, whilst γ and ρ are significantly higher for Pomona infected animals.

individual means that their uncommon pattern is still represented in the data by the other two individuals, and because the sample size is larger for this group (16 versus 8 animals). This highlights the importance of collecting a large enough sample size (both within and between animals) to help ensure that the model is robust and generalisable.

The posterior distributions of τ_0 vary by ID, but are generally similar within serovar (see figure 4.10). Pomona infected individuals had earlier estimated times of infection compared to Hardjobovis infected animals. The posterior distributions for Hardjobovis individuals were centred around the true time of infection, whilst the Pomona infected individuals were centred 2-3 days prior to the true time of infection. Although there is a bias of several days, this is still a reasonably accurate estimate given very small quantities of individual level data. Additionally, if this bias were consistent enough, it could be incorporated into the model to correct the time of infection estimates. Individuals B6 and C2 had bimodal marginal distributions with long tails. These individuals are discussed further in the Discussion.

4.5 Discussion

We develop a stylised mechanistic immune system model for temporal censored antibody titre data. The model was used to estimate temporal titre trends for sheep challenged with *leptospira* serovars pomona and hardjobovis. The results suggest potential differences in immune response to the different leptospiral serovars, and enables estimation of the time of infection for each animal. The results were robust to exclusion of an individual (using 24-fold cross validation by individual). The estimated times of infection were most accurate for Hardjobovis infected animals, with Pomona infected animals showing a bias of 2-3 days.

There were significant negative posterior correlations between β and γ because an increase

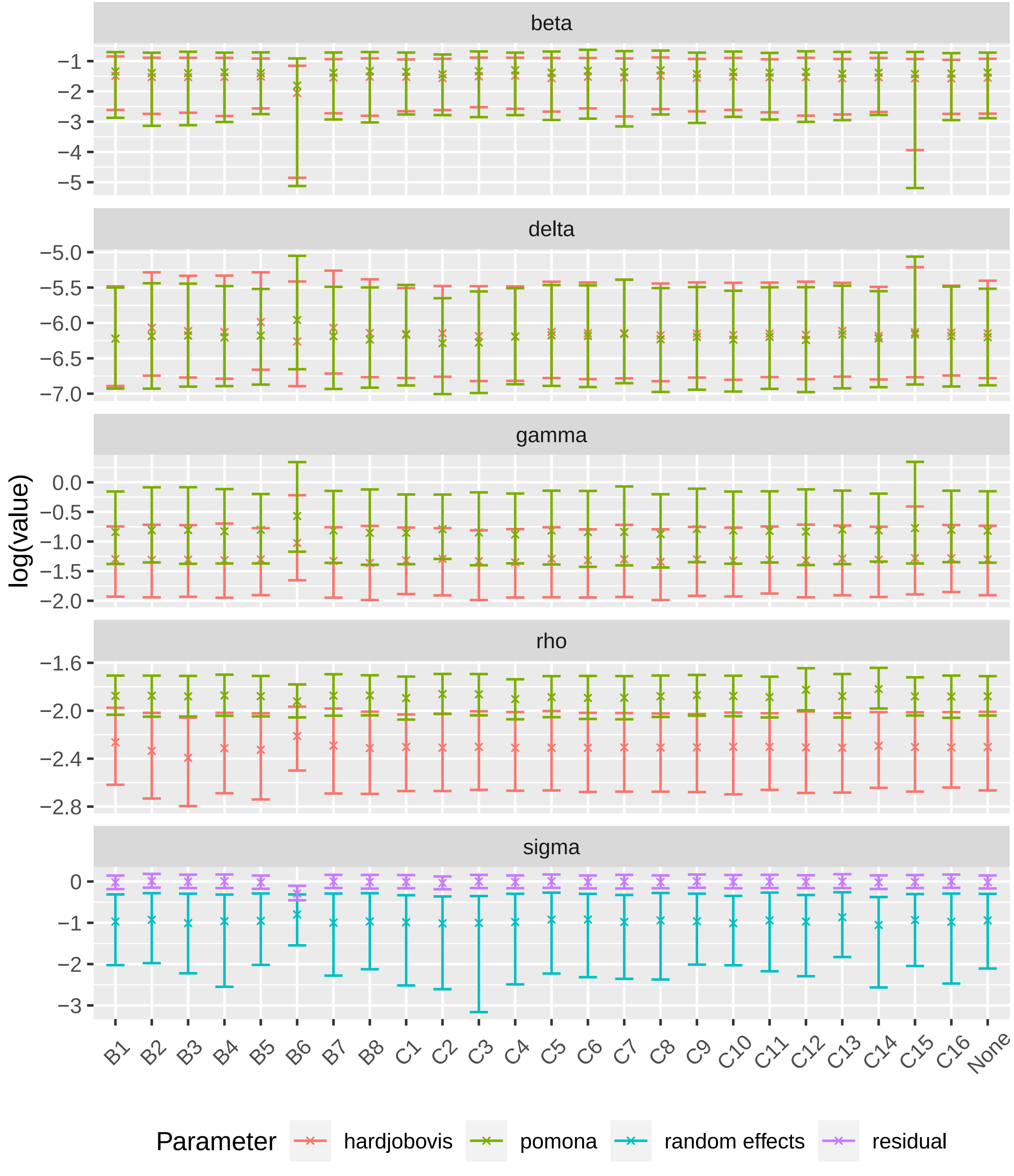


Figure 4.9: Summary statistics for ODE model parameters estimated from ODE models fitted with hold-out individuals. Median values shown with a cross, and 95% credible intervals shown using vertical line segments. The excluded individual is the label on the x axis. The results for the full model (excluding no individuals) is labelled "None". This plot shows influential individuals, and which parameters they influence most. No individuals have an unduly large effect, however, individual B6 is the most influential, with particular effects on β, γ and σ.

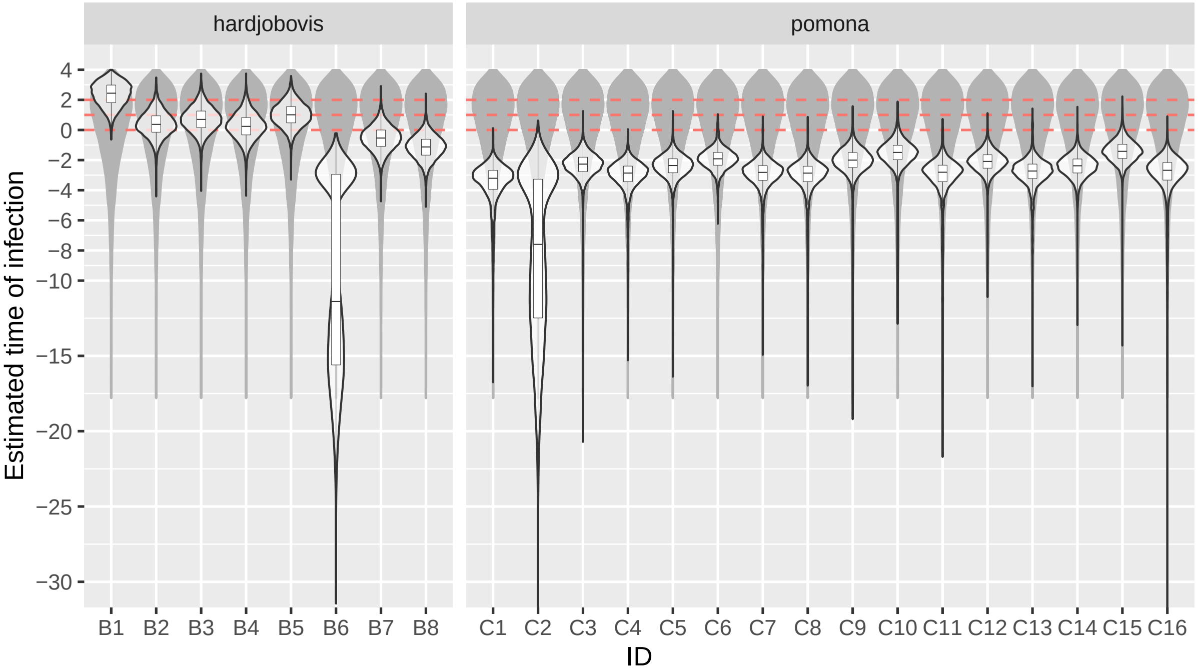


Figure 4.10: Marginal posterior distributions for time of infection $\tau_0 + 4$ for each individual. Note, the time scale was shifted by subtracting 4 Days before model fitting, hence, we add 4 days to the estimated time of infection so that the plot is on the original time scale. The posterior distributions are shown as black violin plots with the prior distribution shown as grey. The approximate true challenge times are shown as red dotted lines. Individuals B6 and C2 show evidence of multimodality. Pomona challenged individuals show evidence of a bias of 2-3 days.

in observed antibody concentrations can either be caused by pathogen concentrations quickly increasing and triggering B-cell production (large γ, small β), or by B-cells quickly reproducing even with relatively low pathogen concentrations (small γ, large β). The two processes are difficult to identify from each other as pathogen concentrations are not observed. This is a considerable problem when jointly estimating these parameters and τ_0, as there are now 3 parameters (β, γ and τ_0) which can affect the shape of the initial titre increase and, usually, very little informative data. The cut model we used circumvents this problem by estimating τ_0 after estimating the other two parameters. The cut model process was only possible as our data had known times of infection enabling accurate estimating of the ODE parameters independently of τ_0. This may not be possible for many other data sets, particularly for serious diseases affecting humans, as ethical concerns around inducing illness restrict researchers abilities to collect suitable data (Franklin and Grady, 2001). However, natural experiments such as point source exposures (Lupidi et al., 1991) or vaccine trial data could provide suitable data sets when controlled challenge studies are not available. Alternatively, strong informative priors could be combined with data where the time of challenge is unknown to jointly estimate the ODE parameters and τ_0. Although vaccine studies produce valuable data for this purpose, care must be taken when generalising results to the general population as individuals may respond differently when not in controlled experimental conditions. For example, vaccination titres can be lower than those induced by infection for leptospirosis in cattle (Kiesel and Dacres, 1959; Strother, 1974). Future work includes a simulation study to investigate the impact of sample size, observations times and various parameter values on the observed posterior correlations.

The timing of the observations is critical, particularly when estimating the antibody dynamics. The titre increase was very rapid, with the time taken to increase from undetectable levels (< 25) to peak titre values being in the range of 2-5 days. The model absolutely requires data during the upward phase to differentiate between pathogen γ and B-cell β rates of growth, and to estimate time of infection accurately. Alternative ODE equations and/or strong informative priors would need to be considered if data were only available during the downward trajectory. Although our model can estimate the time of infection using a single titre for an individual, the estimate will be highly uncertain, strongly affected by the chosen prior for τ_0, and likely bimodal (as it is unclear whether the titre is occurring on the increasing or decreasing phase of the temporal curve). The minimum number of titre observations per individual required to estimate the antibody titre dynamics and time of infection is highly dependent upon the timing of the measurements and the degree of censoring. For our data, the model could reliably estimate the time of infection with 5 temporally spaced observations, although some bias was seen for Pomona animals.

Although collecting more data at pre-peak times would greatly improve predictions of τ_0, this is typically very difficult due to added expense and diagnosis typically occurring some time after challenge. The diagnosis delay has many intertwined causes including many diseases displaying a lag phase where there is little to no increase in antibody concentration for a short period after exposure (Nicholson, 2016). Although our model allows for a phase where titres remain very low, the structure of our ODE model requires that fitted titre values continuously change over time. Comparison of figures 4.5a and 4.5b shows that, for our data, the titres continuously increase from the time of challenge even when the fitted curves appear near flat on the original scale. If titre values were collected beginning immediately post-challenge (with much smaller censoring intervals), it would be possible to estimate an additional lag time (where antibody levels do not increase). Alternatively, the model could be modified to add on a predetermined lag time using expert opinion.

MAT titres detect antibodies which are based on the infected individuals immune response to the pathogen (here leptospires). This means that it can be unclear whether an infection is active, particularly with only one or two samples for an individual, and cannot be used to differentiate between a vaccine induced response and a pathogenic challenge. This means that any data reliant on measuring antibody concentrations will not work reliably if individuals in the data set have previously been vaccinated against the disease of interest. However, if the time of vaccination was known, and enough correctly spaced measurements were observed, it would likely be possible to estimate additional challenges post vaccination.

Although the multiplicative random effects captured much of the between animal variation, there was still evidence of some systematic additional between animal variation. For example, animals B6 and C14 do not follow the typical trajectory given their serovar. It is unclear with so few animals whether this result is a normal (albeit uncommon) titre pattern, if there was some additional unmeasured covariate affecting the result, or if it was simply noise. The random effects would likely explain significantly more variance if they affected the ODE parameters directly (for example, $\beta_{ij} = \beta_j \zeta_i$). However, this would greatly increase the model complexity, and risk over-fitting and non-identifiability problems, particularly with temporally sparse, heavily censored data and small sample sizes. The animals in our data set displayed reasonably small between animal differences. This is likely due to the controlled experimental setting where all animals in the trial had a similar initial dose of pathogen at same time and experienced the same diet and other environmental factors. The large between individual variability observed in some other epidemiological studies may be due to larger environmental differences, or may vary by pathogen type and host species. We chose not to jointly estimate the individual level random effect alongside τ_0 because the data for an individual animal is very sparse (typically 6 - 8 temporally spaced measurements), and we felt it was excessive to estimate two parameters from so little data. Therefore, our predicted time of infection is given the posterior distribution of fitted curves for an average animal from the appropriate serovar. The random effect could easily be added into the τ_0 model if enough data was present to estimate it well (e.g. more measurements per individual, or smaller censoring ranges). The marginal posterior distribution for initial infection time τ_0 was bimodal for individuals B6 and C2. Individual B6 is a clear outlier, modes centred 3 and 14 days prior to the true time of infection. It was suspected that B6 was infected prior to the study as the observed titre before challenge ($t = 0$) was non-zero, whilst a screening measurement at $t = -16$ showed no antibodies present. B6 also showed an atypical titre pattern with relatively high titres before $t = 10$ (see Figures 4.1 and 4.11). If individual B6 was infected prior to the initial challenge at $t = 0$, then the expected infection time would have likely been between $t = 0$ and $t = -16$. This range of times is consistent with the marginal posterior distribution for τ_0 which has modes near $t = -3$ and $t = -14$. Individual C2 also showed a bimodal posterior distribution for τ_0 with most mass near $t = -3$ (similar to other animals in the same trial), but with significant mass near $t = -12$. This result is likely due to the unusually low titre values recorded for this animal which, combined with the small sample size, high noise and censoring, mean that the titre pattern is consistent with a very early time of infection (see Figure 4.11).

The Massey data was not designed and collected for the purpose of modelling antibody titre dynamics and estimating times of infection, hence, it has some limitations in this context. Ideally, titre measurements would have been more frequent, particularly during the rapid initial rise in titre. Had this data been available, the shape of the titre curve could have been ascertained with more certainty. Additionally, the trial ID is totally confounded with serovar (one trial per serovar). We have so far interpreted results as though observed differences are serovar effects, however, they could plausibly be due to any (unrecorded) differences between the two trials. This confounding makes it impossible

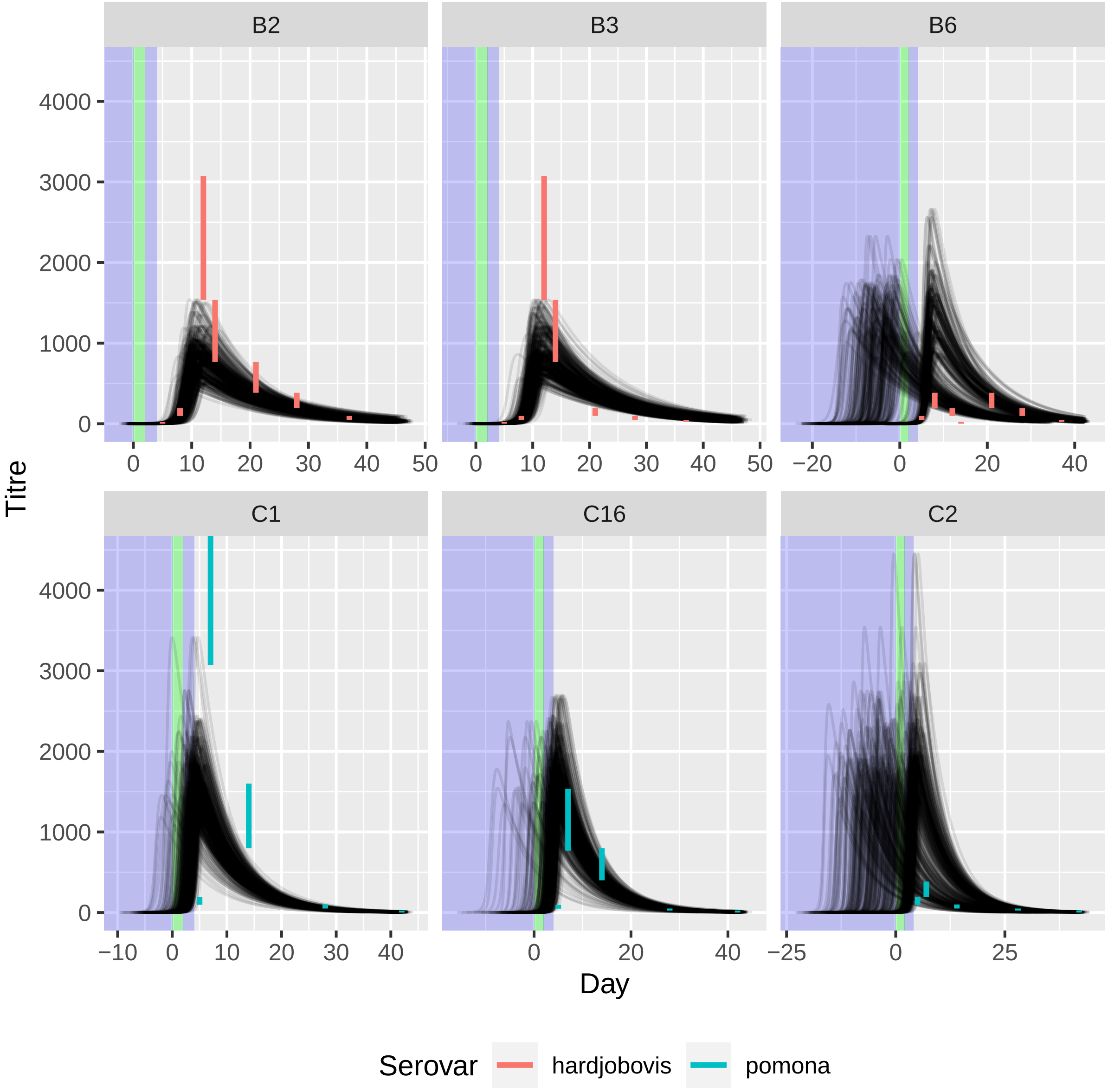


Figure 4.11: Sample of fitted curves from τ_0 models for 6 individuals. The observed data are shown as vertical line segments coloured by serovar. Individuals B6 and C2 exhibit some bi-modality in estimated times of infection. The true time of infection is within the green rectangle. The prior for τ_0 excludes any times after the first observed titre value, therefore, it excludes values outside the green and blue regions. The results show that the predicted curves fit the data well, including those in both posterior modes for individuals B6 and C2.

to generalise the results, however, we note that the observed differences are consistent with published serovar effects. In particular, serovars causing more severe infections (Pomona and Copenhageni) have been found to produce higher titres than Hardjobovis infections (Carter et al., 1982; Faine and World Health Organization, 1982; Ayanegui-Alcerreca, 2006; Heuer et al., 2012) as Hardjobovis is likely a maintenance host in sheep (Heuer et al., 2012). To our knowledge, there are no other comparable data sets for leptospirosis in sheep, so it is not possible to compare the resulting fitted titre curves with independent data. Vallée (2016) investigated titre patterns over nearly 4 years in naturally infected sheep on New Zealand farms. The testing intervals were 2-6 months, and true time(s) of challenge unknown. Vallée (2016) showed that predicted log titres were strongly dependent on age (likely due in part to the presence of maternal antibodies and environmental conditions). Maximum titres were estimated to occur 5-10 months after maternal antibodies had waned, and antibodies had a half life of over 6 months. Due to the irregular and infrequent testing, shorter term patterns are unclear. However, they showed a quicker decrease for Pomona infected animals than Hardjobovis infected ones, which is consistent with our results. The maximum log titre observed for Hardjobovis infected animals was around 5, whilst Pomona infected animals peaked at around 3. This contradicts our results, which showed less difference between serovars (and Pomona being higher on average) and much higher peak average peak values (6.8 for Hardjobovis and 7.4 for Pomona in our dataset). However, it is challenging to estimate maximum titre values, particularly with infrequent sampling, as they occur for such a short period of time. Therefore, the maximum log titre estimates in Vallée (2016) may be underestimates. Alternatively, our maximum observed titres may be much larger due to increased initial pathogen dose, or because antibody dynamics are different for naturally versus experimentally challenged individuals.

Although we have focused this chapter on modelling titres, our model (perhaps with some modifications) is appropriate for many other measures of immune response such as optical densities (OD), or direct antibody concentration measurements. This main requirement is that the measured variable is strongly correlated with the true antibody concentration. We could also extend the model to use observed pathogen concentration measurements rather than antibody concentration data. The set of ordinary differential equations we used were adequate to fit the Massey data well, but could easily be extended to accommodate different temporal patterns of immune response for different animal species, humans or pathogen types or other covariates. Additional future work includes fitting models with random effects on the ODE parameters, extending the model to allow for multiple challenge times, jointly estimating τ_0 with the ODE parameters and applying the model to other data sets.

4.6 Conclusion

Our model estimates sensibly fitted titre curves (including sensible peak antibody titre values and times) and is able to accurately predict the time of infection from a single test individual. This implies that time of infection can be estimated at an individual level given enough prior information about the antibody decay pattern. Although we cannot generalise the biological results given the limitations of the data, it demonstrates that the model is valid and works well on real world data.

4.7 Supporting Information

4.7.1 MCMC fitting details

The functional form of the system of ordinary differential equations (ODE) used in this chapter restricts the fitted B-cell and pathogen counts to be positive. The data suggests that the counts get near to zero at times far from the peak. This can cause numerical issues for the ODE solvers implemented in Stan causing occasional negative fitted values when the true value is very close to zero. There are a number of ways of addressing this issue, including adding a very small "fudge factor" to the fitted values, increasing tolerances (causing much slower fitting), changing priors to pull fitted values away from zero (not appropriate for this model) and rewriting the equations such that they solve for the concentrations on the log scale.

We chose the final option as our relatively simple equations are easily log transformed. The general formula for log transforming an ordinary differential equation is as follows:

$$\frac{\mathsf{d}y}{\mathsf{d}t} = f\left(y, t\right) \tag{4.24}$$

$$\frac{\mathsf{d}\log y}{\mathsf{d}t} = \frac{1}{y}\frac{\mathsf{d}y}{\mathsf{d}t} = \frac{f\left(y, t\right)}{y} \tag{4.25}$$

The log transformed ODE equations for our model (see equation 4.10) become

$$\frac{\mathsf{d}C^*}{\mathsf{d}t} = \frac{\beta CA - \rho C}{C} = \beta \exp A^* - \rho \tag{4.26}$$

$$\frac{\mathsf{d}A^*}{\mathsf{d}t} = \frac{\gamma A - \delta CA}{A} = \gamma - \delta \exp C^* \tag{4.27}$$

where C^* and A^* are the log number of antibody cells and pathogens respectively.

Bibliography

Lotka, A.J., 1910. Contribution to the Theory of Periodic Reaction. *J. phys. chem.*

Kiesel, G.K. and Dacres, W.G., 1959. A study of Leptospira pomona bacterin in cattle. *The cornell veterinarian.*

Strother, H.L., 1974. Host animal efficacy studies using multivalent leptospria bacterin. *Procedures of u.s. animal health a.*

Higgins, R.J., Harbourne, J.F., Little, T.W., and Stevens, A.E., 1980. Mastitis and abortion in dairy cattle associated with Leptospira of the serotype hardjo. *Veterinary record.*

Pearson, J.K., Mackie, D.P., and Ellis, W.A., 1980. Milk drop syndrome resulting from Leptospira hardjo. *Veterinary record.*

Carter, E.M., Cordes, D.O., Holland, J.T.S., Lewis, S.F., and Lake, D.E., 1982. Leptospirosis: II. investigation of clinical disease in dairy cattle in the Waikato district of New Zealand. *New Zealand Veterinary Journal* [Online], 30(9). PMID: 16030901, pp.136–140. eprint: `https://doi.org/10.1080/00480169.1982.34915`. Available from: `https://doi.org/10.1080/00480169.1982.34915`.

Cordes, D.O., Carter, M.E., Townsend, K.G., Lewis, S.F., and Holland, J.T.S., 1982. Leptospirosis: i. clinical investigation of the infection in dairy cattle in the Waikato district of New Zealand. *New Zealand Veterinary Journal.*

Faine, S. and World Health Organization, 1982. *Guidelines for the control of leptospirosis / edited by s. faine.* World Health Organization.

Cousins, D.V., Ellis, T.M., Parkinson, J., and McGlashan, C.H., 1989. Evidence for sheep as amaintenance host for Leptospira interrogans serovar hardjo. *Veterinary record.*

Lupidi, R., Cinco, M., Balanzin, D., Delprete, E., and Varaldo, P.E., 1991. Serological follow-up of patients involved in a localized outbreak of leptospirosis. *Journal of clinical microbiology* [Online], 29(4), pp.805–809. eprint: `https://jcm.asm.org/content/29/4/805.full.pdf`. Available from: `https://jcm.asm.org/content/29/4/805`.

Ellis, W.A., 1994. Leptospirosis as a cause of reproductive failure. *Veterinary clinics north america:food animal practice.*

Gerritsen, M.J., Koopmans, M.J., Peterse, D., and Olyhoek, T., 1994. Sheep as maintenance host for Leptospira interrogans serovar hardjo subtype hardjobovis. *American journal of veterinary research.*

Fierz, W., 1998. Basic problems of serological laboratory diagnosis. *Methods in molecular medicine* [Online], 13 (), pp.443–71. Available from: `https://doi.org/10.1385/0-89603-485-2:443`.

Langoni, H., de Souza, L.C., da Silva, A.V., Luvizotto, M.C., Paes, A.C., and Lucheis, S.B., 1999. Incidenceof leptospiral abortion in Brazilian dairy cattle. *Preventive veterinary medicine*.

Franklin, G.M. and Grady, C., 2001. The Ethical Challenge of Infection-Inducing Challenge Experiments. *Clinical infectious diseases* [Online], 33(7) (), pp.1028–1033. eprint: `https://academic.oup.com/cid/article-pdf/33/7/1028/1184345/33-7-1028.pdf`. Available from: `https://doi.org/10.1086/322664`.

Levett, P.N., 2001. Leptospirosis. *Clinical microbiology reviews* [Online], 14(2), pp.296–326. eprint: `https://cmr.asm.org/content/14/2/296.full.pdf`. Available from: `https://doi.org/10.1128/CMR.14.2.296-326.2001`.

Magajevski, F.S., Girio, R.J.S., Mathias, L.A., Myashiro, S., Genovez, M.A., and Scarcelli, E.P., 2005. Detection of Leptospira spp. in the semen and urine of bulls serologically reactive to Leptospira interrogans serovar hardjo. en. *Brazilian Journal of Microbiology* [Online], 36 (), pp.43–45. Available from: `http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1517-83822005000100009&nrm=iso`.

McBride, A., Athanazio, D., Reis, M., and Ko, A.I., 2005. Leptospirosis. *Current opinion on infectious diseases*.

Ayanegui-Alcerreca, M.A., 2006. *Epidemiology and control of leptospirosis in farmed deer in New Zealand*. PhD thesis. Massey University.

Ayanegui-Alcerreca, M.A., Wilson, P.R., Mackintosh, C.G., Collins-Emerson, J.M., Heuer, C., Midwinter, A.C., and Castillo-Alcala, F., 2007. Leptospirosis in farmed deer in New Zealand: a review. *New zealand veterinary journal*.

Ko, A.I., Goarant, C., and Picardeau, M., 2009. Leptospira: the dawn of the molecular genetics era for an emerging zoonotic pathogen. *Nat. rev. microbiol.*, 7(10), pp.736–747.

Chaplin, D.D., 2010. Overview of the immune response. *The journal of allergy and clinical immunology*.

Hartskeerl, R.A., Collares-Pereira, M., and Ellis, W.A., 2011. Emergence, control and re-emerging leptospirosis: dynamics of infection in the changing world. *Clinical microbiology and infection*.

Cauchemez, S., Horby, P., Fox, A., Mai, L., Thanh, L., Thai, P., Hoa, L., Hien, N., and Ferguson, N.M., 2012. Influenza infection rates, measurement errors and the interpretation of paired serology. *PLoS pathogens* [Online], 8(12). Available from: `https://doi.org/10.1371/journal.ppat.1003061`.

Heuer, C., Benschop, J., Stringer, L., Collins-Emerson, J., Sanhueza, J., and Wilson, P., 2012. *Leptospirosis in New Zealand - best practice recommendations for the use of vaccines to prevent human exposure.* (Technical report). A Report by Massey University Prepared for the Zealand Veterinary Association.

Subharat, S., Wilson, P.R., Heuer, C., and Collins-Emerson, J.M., 2012. Growth response and shedding of Leptospira spp. in urine following vaccination for leptospirosis in young farmed deer. *New Zealand Veterinary Journal.*

IESR, 2013. *Notifiable and other diseases in New Zealand: annual report 2012-2012.* (Technical report). Porirua, New Zealand: The Institute of Environmental Science and Research Ltd.

Musso, D. and La Scola, B., 2013. Laboratory diagnosis of leptospirosis: a challenge. *Journal of microbiology, immunology and infection* [Online], 46(4), pp.245–252. Available from: `https://doi.org/https://doi.org/10.1016/j.jmii.2013.03.001`.

de Graaf, W.F., Kretzschmar, M.E.E., Teunis, P.F.M., and Diekmann, O., 2014. A two-phase within-host model for immune response and its application to serological profiles of pertussis. *Epidemics* [Online], 9. Available from: `http://www.sciencedirect.com/science/article/pii/S1755436514000371`.

Fang, F., 2014. *Leptospirosis diagnostics and exposure at the human and animal interface in New Zealand.* PhD thesis. Massey University.

Goris, M.G.A. and Hartskeerl, R.A., 2014. Leptospirosis serodiagnosis by the Microscopic Agglutination Test. *Current protocols in microbiology* [Online], 32(1). eprint: `https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/9780471729259.mc12e05s32`. Available from: `https://doi.org/10.1002/9780471729259.mc12e05s32`.

Costa, F., Hagan, J., Calcagno, J., Kane, M., Torgerson, P., Martinez-Silveira, M.S., Stein, C., Abela-Ridder, B., and Ko, A., 2015. Global morbidity and mortality of leptospirosis: a systematic review. *Plos neglected tropical diseases.*

Costa, F., Hagan, J.E., Calcagno, J., Kane, M., Torgerson, P., Martinez-Silveira, M.S., Stein, C., Abela-Ridder, B., and Ko, A.I., 2015. Global Morbidity and Mortality of Leptospirosis: A Systematic Review. *PLoS Neglected Tropical Diseases*, 9(9).

Costa, F., Wunder, E.A., De Oliveira, D., Bisht, V., Rodrigues, G., Reis, M.G., Ko, A.I., Begon, M., and Childs, J.E., 2015. Patterns in Leptospira Shedding in Norway Rats (Rattus norvegicus) from Brazilian Slum Communities at High Risk of Disease Transmission. *PLoS Neglected Tropical Diseases*, 9.

Haake, D. and Levett, P., 2015. Leptospirosis in humans. *Current topics in microbial immunology.*

Health Intelligence Team. and Health Group., 2015. *Notifiable Diseases in New Zealand: Annual Report 2015.* (Technical report). Porirua, New Zealand: The Institute of Environmental Science and Research Ltd.

Kucharski, A.J., Lessler, J., Read, J.M., Zhu, H., Jiang, C.Q., Guan, Y., Cummings, D.A.T., and Riley, S., 2015. Estimating the life course of influenza A (H3N2) antibody responses from cross-sectional data. *PLoS biology* [Online], 13(3) (), pp.1–16. Available from: `https://doi.org/10.1371/journal.pbio.1002082`.

Niloofa, R., Fernando, N., Silva, N.L. de, Karunanayake, L., Wickramasinghe, H., Dikmadugoda, N., Premawansa, G., Wickramasinghe, R., Silva, H.J. de, Premawansa, S., Rajapakse, S., and Handunnetti, S., 2015. Diagnosis of Leptospirosis: Comparison between Microscopic Agglutination Test, IgM-ELISA and IgM Rapid Immunochromatography Test. *PLoS ONE* [Online], 10(6), pp.1–12. Available from: `https://doi.org/10.1371/journal.pone.0129236`.

Torgerson, P.R., Hagan, J.E., Costa, F., Calcagno, J., Kane, M., Martinez-Silveira, M.S., Goris, M.G., Stein, C., Ko, A.I., and Abela-Ridder, B., 2015. Global Burden of Leptospirosis: Estimated in Terms of Disability Adjusted Life Years. *PLoS Neglected Tropical Diseases*, 9(10).

World Health Organization, 2015. *Who estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015* [Online]. available on the WHO web site (www.who.int) or can be purchased from WHO Press, World Health Organization, 20 Avenue Appia, 1211 Geneva 27, Switzerland. Available from: `http://apps.who.int/iris/bitstream/10665/199350/1/9789241565165_eng.pdf?ua=1`.

Borremans, B., Hens, N., Beutels, P., Leirs, H., and Reijniers, J., 2016. Estimating time of infection using prior serological and individual information can greatly improve incidence estimation of human and wildlife infections. *PLoS computational biology* [Online], 12(5). Available from: `https://doi.org/10.1371/journal.pcbi.1004882`.

Frigolett, J.M.S., 2016. *Occupational leptospirosis in New Zealand.* PhD thesis. Institute of Veterinary, Animal and Biomedical Sciences, Massey University.

Nicholson, L.B., 2016. The immune system. *Essays in biochemistry.*

Vallée, E., 2016. *Epidemiology and production effects of leptospirosis in New Zealand sheep.* PhD thesis. Massey University.

Zhao, X., Siegel, K., Chen, M.I., and Cook, A.R., 2017. Rethinking thresholds for serological evidence of influenza virus infection. *Influenza and other respiratory viruses* [Online], 11(3). eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/irv.12452`. Available from: `https://doi.org/10.1111/irv.12452`.

Antia, A., Ahmed, H., Handel, A., Carlson, N.E., Amanna, I.J., Antia, R., and Slifka, M., 2018. Heterogeneity and longevity of antibody memory to viruses and vaccines. *PLoS biology* [Online], 16(8). Available from: https://doi.org/10.1371/journal.pbio.2006601.

Kucharski, A., Lessler, J., Cummings, D., and Riley, S., 2018. Timescales of influenza A/H3N2 antibody dynamics. *PLoS biology* [Online], 16(8). Available from: https://doi.org/10.1371/journal.pbio.2004974.

World Heath Organisation, 2018. *Global Health Estimates 2016 Summary Tables.* (Technical report). World Heath Organisation.

Hay, J.A., Minter, A., Ainslie, K., Lessler, J., Kucharski, A.J., and Riley, S., 2019. Serosolver: an open source tool to infer epidemiological and immunological dynamics from serological data. *Biorxiv* [Online]. eprint: https://www.biorxiv.org/content/early/2019/08/08/730069.full.pdf. Available from: https://doi.org/10.1101/730069.

Stan Development Team, 2020. *RStan: the R interface to Stan* [Online]. R package version 2.19.2. Available from: http://mc-stan.org/.

Owers, K. and Diggle, P., n.d. *Accounting for interval-censored antibody titre decay substantially increases seroincidence in a longitudonal cohort study of leptospirosis*. Submitted to: American Journal of Epidemiology. Unpublished.

Chapter 5

Discussion

Advanced statistical models are a key tool in developing interventions to reduce disease incidence, particularly in low resource settings. Epidemiological studies typically include complexities such as multi-level dependency structures, correlated covariates, missing and/ or censored data and unexplained overdispersion. The data are typically generated from a complex biological system which is only partially observed (with high noise). These challenges are integral parts of the 3 projects in this thesis, and we address them using Bayesian models. The following sections will discuss these challenges in the context of each project and propose future work.

5.1 Chapter overviews

Chapter 2 develops a novel class of Bayesian non-parametric source attribution model, which classifies strain types by differential epidemiological behaviour and accurately quantifies uncertainty. Our `R` package `sourceR` provides the first standard software for source attribution and is designed for use by epidemiologists and public health decision makers. Our model provides significant advancements over previous source attribution models and identifies sub-types that have unusually high virulence, pathogenicity and/or survivability which alongside further genetic sequencing could be used to predict the danger posed by emerging strains. The model is particularly important for foodborne zoonotic diseases such as campylobacteriosis, but can also be used in other settings such as estimating transmission pathways of diseases such as leptospirosis (which can spread directly though contact with animal carriers and through environmental sources).

Chapter 3 expands the scope from estimating risk from known sources of infection (in Chapter 2) to consider a wide range of potential risk factors for contracting disease in complex heterogeneous environments. This chapter developed a methodological approach to incorporate highly heterogeneous spatio-temporally varying covariates (with uncertainty), into traditional epidemiological risk models using marginalisation. This modelling strategy relies on accurate disease status labels for study participants. The criteria commonly used to diagnose an individual as infected, using serology data, is not particularly robust. Therefore, it is important to improve methodologies for estimating infection status using serology data.

In Chapter 4, the challenges of estimating infection status were addressed. We developed a mechanistic model for antibody concentration changes after challenge, and use it to estimate the time of infection, using noisy, censored serology data with small sample sizes. This is also the first study to model short term titre trends post challenge with leptospirosis in sheep. We identified consistent differences between serovars, a result which provides room for hypotheses on the observed differential immune responses of an individual, with reference to the infecting serovar. Our model is highly extensible and could be integrated with many epidemiological studies for infectious diseases.

5.2 Limitations and challenges of the data and approaches utilised

The projects in this thesis highlighted some common challenges and limitations of epidemiological studies, including complex correlation and dependency structures; confounding; noisy, missing and censored data; and technical challenges when fitting Bayesian models.

We suggest some future work including extensions and improvements to the developed models, and proposals for further research to address these limitations.

5.2.1 Complex correlation and dependency structures

Our source attribution model, introduced in Chapter 2, could be extended to incorporate spatio-temporal information in a more sophisticated manner than using independent source and type effects for each time and location. Incorporating correlation structures, such as those used in Chapter 3, would allow estimation of changes in source effects and type effects; but most importantly, the proportion of cases attributable to the various sources over time and space if given enough informative data. This could give valuable insight into the processes driving infection, and highlight areas or times when certain sources have an increased risk. It is likely that there is some seasonality and the possibility of source-type interactions; for example, cases attributable to chicken may increase in summer months when chicken is more likely to be inadequately cooked (e.g. on BBQ's) and will spoil more quickly if not safely stored and prepared.

MLST sub-typing is based on similarities in housekeeping genes which are chosen as they mutate slowly. Genes responsible for virulence, survivability, and pathogenicity (type effects) may mutate more quickly, resulting in several sub-strains within an MLST sub type. These strains could have different infection characteristics which would hinder clustering based on type effects. Directly incorporating additional genetic information could allow estimation of genetic relatedness between samples and inform type effect clustering. This would be particularly effective if sequences with associations to survivability, pathogenicity or virulence were known; although, this is rarely the case. Comparison of full sequences between strains that cluster into low and high type effect groups may help to identify these sequences, and could inform estimation of the level of danger posed by a novel subtype based on sequence data. A Dirichlet Process may not be the best approach for modelling the type effects if additional genetic information is available, as it does not account for correlations between sub-types. Incorporating this additional complexity (particularly allowing for interactions) significantly increases the complexity of the model, and severely risks over-fitting; however, with enough data and techniques such as cross validation or shrinkage methods (e.g. Horseshoe prior Carvalho, Polson, and Scott, 2010) the risks of over-fitting could be mitigated.

Chapter 3 highlighted some of the difficulties in quantifying contributions to risk factors in the presence of correlated covariates and when data is highly unbalanced. This is an ongoing problem in epidemiological studies due to the complex biological processes driving infection patterns alongside ethical and practical constraints. Although complex models can go some way to adjusting for these issues, low quality data cannot be "fixed" using advanced modelling strategies. Even with these reservations, we believe that our study has provided practically useful results which are consistent with other published literature. The following paragraphs introduce some of the challenges and limitations of our data and method with respect to these issues.

Practical limitations meant that rat data was collected from a small (randomly selected) sub-area over a period of weeks, rather than all at once. This causes significant confounding between spatial and temporal effects on rat mark distributions. Collecting data from multiple areas per time period, and from each area at multiple times, would likely give enough information to separate out the effects; but would be much more expensive, and was not feasible for our study. Instead, we must assume that important changes to rat density occur at longer time scales than those taken to collect data within a campaign. Additional confounding occurred between time (campaign), environmental damage, and rainfall. Our study included two main time periods (campaigns) which were chosen to fall in the wet and dry periods of the year respectively. The study area experienced extensive environmental damage caused by flooding in campaign 2 (wet season), which was not captured by our covariates. Therefore, we cannot disentangle the rainfall, environmental damage and other season/time effects solely based on this data. However, there is strong literature support for high rainfall being associated with increased cases of leptospirosis, which supports our conclusion that increased rainfall is likely associated with increased risk of leptospirosis in Pau da Lima. Eight campaigns worth of data were collected for both rat distribution and human cases, however, only two campaigns were available when this analysis was performed. Expanding the analysis to include the other six campaigns would provide significantly more information about the risks attributable to each of the potential risk factors, in particular, removing the confounding between campaign and rainfall. The estimated spatial surfaces at the two time points was similar enough to suspect that hotspots are reasonably consistent in Pau da Lima. This means significant gains could be made by estimating a joint spatio-temporal surface, rather than independent surfaces, for each time point. This would have the additional benefit of allowing prediction of the rat prevalence surface at times not present in the data set. If the rat prevalence hotspots are reasonably stable over longer time periods, further work could be done to identify the unknown risk factors for rat prevalence. This could produce more accurate predictions of rat prevalence and leptospirosis risk for regions outside the study area.

The data contained additional correlation structures which complicated model fitting and interpretation. Many of the socio-economic and environmental risk factors for leptospirosis are inextricably linked. Additionally, covariates may affect the risk of leptospirosis directly, or indirectly, by affecting rat prevalence/activity levels. For example, individuals who identify as black in our study typically had a lower household income, lower education level, and are less likely to have a job when compared with other individuals in the area. These individuals were also more likely to be adversely affected by floods as they lived in the least desirable areas, near open sewers and rubbish dumps, at the bottom of valleys. This also potentially increased their exposure to leptospire shedding rats which are estimated to have higher prevalence/ activity near food sources such as rubbish dumps, and their exposure to leptospires which can survive for long periods in moist soil near open sewers. All of these variables potentially contribute to the risk of contracting leptospirosis. These correlations make it extremely difficult to disentangle the individual effect of each variable to disease risk. Instead, we interpret the results with caution, and rely on existing biological and epidemiological leptospirosis knowledge to guide our interpretation. Further studies which directly manipulate targeted covariates may help to disentangle some of these variables, but these studies are typically expensive and more invasive (for example, randomly allocating residents to reside near open sewers).

The large correlations between covariates and complex chains of infection make variable selection challenging, particularly when we expect some continuous covariates to require polynomial or spline terms. I decided to include all collected covariates rather than perform variable selection, as variable selection methods can produce poor outcomes, particularly when there are many correlated variables (Harrell, 2001). We did not consider interactions as none were expected a priori, and inclusion of all possible interactions (even restricting to just two way interactions) would have produced an extremely complicated model and risked over-fitting the data. Cross validation could reduce the risk of over fitting, but it is difficult to design an effective cross validation strategy when there is significant structure in the data and a low sample size.

The leptospirosis challenge data from Chapter 4 experienced severe confounding, due to the experimental design allocating only one serovar per trial. This prevents us from generalising the results outside of the animals used in the study; however, the differences we observe are consistent with other published studies, suggesting that the differences we observe may be caused by Serovar. This was an unavoidable constraint due to the low availability of appropriate data sets for model development (where time of infection is known, and many measurements are made for each individual throughout the infection process).

5.2.2 Missing, censored and unreliable data

There is likely significant under-reporting of campylobacteriosis cases in our source attribution data set, largely due to the non-specific mild symptoms experienced by a large proportion of infected individuals. It is probable that the severity of illness depends, in part, on the sub-species causing the illness. This means that data is missing not at random (MNAR), causing biased estimates of the proportion of cases attributable to each food source. However, it is arguable that the unobserved cases are of less importance, as they are typically less severe than those included in the data set. Hence, if we redefine our response to be "cases of moderate to severe illness" rather than "all cases of illness", the results are likely much less biased. This bias could be addressed by randomly sampling individuals (regardless of disease status) from a population, and testing for evidence of campylobacteriosis challenge. However, this relies on tests having good sensitivity and specificity at widely varying prevalences, and is likely to be very expensive given the relative rarity of the disease. Additionally, many surveys of randomly selected individuals often experience low response rates, and rates would likely be even lower due to the requirements that respondents provide a stool sample.

We estimated rat prevalence/activity as a proxy for the missing rat exposure covariate values for each individual in the Brazilian leptospirosis study. There are many complications and limitations of this approach to consider. Although we know that rats carry leptospires and shed them into the environment, there are many methods by which they could infect a person. For example, individuals may be directly exposed whilst killing rats, they may be exposed to rat urine on tin cans or other food packaging; or they may be exposed to leptospires by touching rat-urine contaminated dirt, water, or other surfaces in the wider environment. Additionally, it is unclear using our method, whether areas identified as having a high probability of observing rat marks are areas of high rat population density, or areas with high rat activity levels. To some extent, both high rat numbers and high rat activity may contribute to an increased risk of disease, as both may lead to increased risk of direct contact and higher urine presence. An additional complication is that individuals may be exposed to rats not only in their home (the geo-located point used in our study), but also anywhere they visit within and outside the study area. This is also a concern when aiming to estimate individuals exposure to environmental covariates, for example, open sewers. Additional geo-located places where individuals spend significant portions of their time, could be added into the model; for example, school, work, or shopping locations. However, this requires detailed knowledge of individuals locations over long time periods which raises some ethical considerations. Additionally, it is not clear how to incorporate data from all locations, especially if many fall outside the study area, and are therefore missing covariate information such as GIS ground cover data. A possible remedy would be to integrate individuals exposure to the entire study area using (likely self identified) weights depending on the proportion of time spent in regions of the study area during each campaign.

The rat models predictive and explanatory power could likely be improved by including additional covariates such as the presence of food sources (e.g. fruit trees). However, this data is difficult to collect at a large number of sites, as it requires manual recording (rather than extraction from a map or photograph as is possible for covariates such as sewer location). Future technological advancements, such as high resolution satellite imagery, combined with image recognition algorithms, may make large scale collection of such data feasible.

Direct manipulation of rat exposure, for individuals in treatment and control areas, would greatly improve estimates of the effect of rat exposure on the risk of leptospirosis because rat exposure would no longer be correlated with other risk factors. A similar experiment was conducted in Pau da Lima where 2 of the 3 valleys were treated with rat poison. Preliminary results (personal communication with Peter Diggle) showed little difference in leptospirosis cases; however, this may be partly due to unforeseen practical challenges. For example, the treatment areas were not able to be intensively poisoned due to safety and consent issues. Many individuals did not consent to rat poison being laid near their home, resulting in large swathes of the treatment area not experiencing the full treatment. Additionally, it is likely that rats from the untreated zones will migrate into the habitat left unclaimed; therefore, poisoning must be intensively maintained over time to keep numbers low. This also has the effect of reducing rat density in the untreated areas, lessening the difference in exposure between treated and untreated individuals. Additionally, high levels of environmental disturbance and damage occurred in the control zone (due to intensive public works beginning) which likely had the effect of reducing rat numbers by destroying their habitat and food sources.

The source attribution method, described in Chapter 2, could be used in this setting to estimate the proportion of leptospirosis cases attributable to sources such as rats, hedgehogs, open sewers, and moist soil. This would require isolation and genetic typing of leptospires collected from potential sources of infection, and from infected individuals, which could be challenging. This approach may be difficult given that animals infect humans through environmental contamination. It may be possible to estimate this transmission pathway by comparing the subtypes found in environmental sources with those found in nearby rat populations, in addition to comparison with types found in human cases.

The MAT methodology likely results in many false negative tests for individuals. This makes it difficult to identify risk factors, as many individuals (in high risk areas) who have the disease will test negative. However, if false negatives occur totally randomly, they should not introduce significant bias other than underestimating the overall disease prevalence. We did not attempt to correct the labels, or estimate which individuals were most likely the false negatives, as this is difficult when there are so few titre measurements for each individual over time.

The titre data in Chapter 4 was generated from a process that is known to be extremely complicated, and is only partially observed. It is challenging to fit this model when observations of pathogen concentrations are structurally missing. This results in some non-identifiability, and when combined with the heavy censoring of titre values, it means initial values are impossible to estimate. This also complicates model interpretation, because parameter estimates depend on these initial values. However, it allows comparison between serovars/trials within the experiment. This provides valuable insights into the possible mechanisms driving the differences in observed titre patterns between these groups.

An additional difficulty was that all individuals within a trial had their titres observed at the same times, which means we have no data for many time periods. This means that we cannot be sure our model is fitting well between observation times, as no data is available to compare it to. This is particularly problematic during times of rapid antibody concentration change, such as near the peak. It is more expensive and time consuming to measure a portion of individuals each day, rather than all individuals on a subset of days; but this method would have provided more information given the same number of animals and MAT tests.

A richer data set (e.g. larger sample size, more closely spaced temporal measurements, lower censoring, and/or less noisy data) could allow more sophisticated exploration of the differences between individual mean fitted curves. The addition of random effects on each of the ODE parameters, could allow us to estimate the variation in effect size for each parameter between individuals in a population, compared with that seen between pathogen types or other covariates. This could provide valuable information about the expected titre patterns for future individuals, and perhaps shed light on differences between immune processes within host. The model could additionally be modified to include a lag phase where titre concentrations do not change for a short initial period. The duration of this period could be estimated given enough informative data. Other advancements could include estimation of time of infection jointly with the ODE parameters (rather than using the cut model). This would be a significant improvement as it would remove the need for an initial data set with known infection times to be available. However, this would require more informative data, by some combination of more individuals, more measurements per individual over time, lower levels of censoring, and perhaps, more informative priors or other restrictions such as enforcing monotonicity on either side of the peak.

Application of this methodology to new populations and pathogens would likely require changes to the ODE equations in some instances, as not all disease processes show the same type of titre pattern over time. The model is flexible enough to easily modify the ODE equations as required for new applications. The mean function could also be replaced with a non-ODE based curve, such as a spline or polynomial; however this makes it harder to infer possible causes of difference in observed patterns between groups using the model. In particular, the ODE model could be extended to incorporate multiple challenges at known times, or to allow estimation of additional challenge times (using change point methods) if given sufficient informative data. A future study of humans in the Pau da Lima area, that employs a more regular titre testing regime, and is carried out in conjunction with regular rat tracking, could enable far more accurate estimation of leptospirosis prevalence and of the relative threats posed by the various risk factors. Titre data could also be incorporated into traditional infectious disease models (such as Susceptible-Infected-Removed models), which estimate times of infection for all individuals, to enable inference on transmission patterns and pathways, identify risk factors, and predict future spread.

5.2.3 High noise, high uncertainty and small sample size

Due to the complex biological data generating processes involved in each study, and the limited information from the data available, the results from our 3 studies had high uncertainty and wide credible intervals for many parameters of interest. This limits our ability to identify interventions with a high probability of success. This is a challenge common to many epidemiological studies, but is preferable to overconfidence which can prove expensive and undermine public trust in intervention strategies. Before implementing expensive intervention strategies, potential interventions can (and should) be tested using pilot studies to more accurately estimate their potential effect. For example, the risk of leptospirosis attributable to rat exposure was difficult to estimate using our data, due to the complex correlations and interactions between rats and other risk factors; and the unreliability of the disease diagnostic tests. A well designed pilot study could be implemented to directly assess the effect of a potential rat reduction intervention. Although a pilot study of this type may greatly improve estimates of the rat effect, they are more intrusive for study participants, raise ethical issues around withholding potential treatments from at-risk individuals, are generally more expensive to run, and only improve estimates for a the risk factors that are directly manipulated. Therefore, although the study design used in Chapter 3 resulted in significant limitations and challenges, it was an essential first step in risk factor identification and quantification, and generated valuable suggestions for future intervention studies.

The titre model in Chapter 4 was challenging to fit due to the small sample size (both the number of measurements per individual and the number of individuals), heavy censoring, and large noise. This resulted in wide credible intervals, due to the uncertainty associated with so little informative data. The wide credible intervals (and high noise variance estimate) were also partially due to the simplicity of the model. It is expected that different individuals exposed to the same pathogen will all respond slightly differently; however, it is hard to detect individual level differences with so little data. We chose to implement a simple, multiplicative, individual level random effect to account for this dependency, however, more data may show that different adjustments (such as random effects on some or all of the ODE parameters) may be more suitable. Another simplification, was the assumption in the model of a single challenge time (on day 0), when we knew that the animals in the Massey dataset were challenged daily for the first 3 days of the trial. We do not attempt incorporate this into the model because there is not sufficient information in the data to warrant complicating the model further (particularly as no titre measurements were done on days 1 to 4).

When the data are not sufficiently informative to identify the timing of the peak, the titre model may return a bimodal estimated time of infection distribution. This can make estimation of time of infection more challenging, particularly when quantifying uncertainty limits (which may require splitting the uncertainty interval into two sub-intervals). The likelihood of getting a bimodal time of infection estimate depends on the amount of data per individuals, how informative each individuals data is (including whether it includes the peak), and how unusual the individuals titre pattern is. As was observed in our data, it is more challenging to estimate accurate times of infection for individuals with unusual titre patterns. A larger data set would be able to account for unusual titre patterns using some extensions; however, in our data set unusual patterns are indistinguishable from noise.

Challenge data sets allow identification and quantification of any bias in the results, as the true time of infection is known. The Pomona challenged individuals had a bias of 2-3 days in the estimated times of infection. The fitted curves fit the data well, and the bias was small. Given it is so consistent, it would be possible to adjust for this for future animals (if it was shown to occur in other more generalisable data sets). This observed bias is small compared to the overall time course of the titre patterns. Narrower censoring intervals, and more titre observations during the early phase of challenge, would likely provide enough information to the model to more accurately estimate the time of infection.

5.2.4 MCMC sampling

Bayesian modelling is extremely flexible, interpretable, and allows incorporation of prior knowledge, which can help with identifiability issues in complex models. However, model fitting can be challenging, particularly when off the shelf samplers (such as PyMC3 and Stan) are not suitable.

Although Dirichlet Processes can be fitted using off the shelf samplers (such as PyMC3), at the time this model was being coded they only supported the stick breaking construction, whereas our model used the chinese restaurant process construction. Additionally, the Hald model (similar to my model) had displayed fitting issues when coded in WinBUGS. Therefore, I decided to hand code the fitting algorithms in R, using adaptive Metropolis-Hastings samplers. I gained a much deeper understanding of Bayesian modelling, MCMC methods, and designing modular code through this endeavour.

The models in Chapter 3 were fitted using PyMC3. This significantly reduced the coding overhead compared to the source attribution model. However, it was still challenging to implement as PyMC3 does not natively support cut models or pseudo marginal samplers. Therefore, we manually implemented a cut model which required fitting hundreds of models for the human case data and combining their results. This is much less efficient than a pseudo marginal sampler because there is a burn-in period associated with every model. Although this increased run time significantly, it decreased the time taken to code the models, which resulted in an overall more efficient modelling process.

In Chapter 4 I originally coded the model in R, using a heavily modified version of the code used in Chapter 2, as ODE solvers were not integrated into any common Bayesian inference programs (such as PyMC3, WinBUGS or Stan). This model was ill-suited to simple Metropolis-Hastings samplers due to the strong posterior correlations, was very slow, and had poor mixing. The model was reimplemented using the development version of Stan, after they implemented an ODE solver allowing variable time vectors in July, 2019. The NUTS algorithm significantly improved sampling efficiency, and consequently run time. Even with this speed up, the cut model implementation resulted in long run times due to the need to discard burn-in iterations for so many models. It may be possible to jointly infer the time of infection with the ODE parameters by increasing the volume of informative data, using strong informative priors, or implementing other constraints.

Another sampling difficulty caused by these strong correlations, is a restriction on the allowable starting values for the sampler. Many combinations of starting values produce ODE equations that cannot be solved, causing the model to fail to run. Therefore, the model must be started using carefully chosen values rather than totally randomly selected values. This makes it challenging for the model to sample from any possible alternative modes, and hinders our ability to judge convergence.

5.3 Conclusions

In conclusion, this thesis outlined 3 approaches to modelling complex epidemiological data to advance understanding of the hidden biological and environmental processes that drive infectious disease progression and spread. These approaches were developed as part of 3 projects, each of which highlighted common challenges, limitations, and successes. The models developed in this thesis were developed to be readily generalisable to other data types, pathogens, environments, and host species. Common challenges include small data sets, large noise, missing data, and complex dependency structures. These challenges were addressed by developing novel zoonotic disease models (Chapters 2 and 4), and by developing data collection and modelling strategies to inform practical interventions (Chapter 3). Further exploration of these methods could prove exceedingly helpful in assessing various epidemiological challenges globally, and further stresses the challenges and importance of thorough study design and implementation.

Bibliography

Harrell, F., 2001. *Regression modelling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer-Verlag New York.

Carvalho, C.M., Polson, N.G., and Scott, J.G., 2010. The horseshoe estimator for sparse signals. *Biometrika* [Online], 97(2), pp.465–480. Available from: `http://www.jstor.org/stable/25734098`.