

Methods and applications of Bayesian  
spatio-temporal statistics for HIV  
surveillance

**Imperial College  
London**

Adam Howes

Department of Mathematics

Imperial College London

In partial fulfillment of the requirements for the degree of

*Doctor of Philosophy*

September 2023

# Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

# Statement of Originality

This thesis, and the work presented in it, is work that I conducted myself. In all cases where I describe others' work, I provide appropriate references.

*For  $\sum_i u_i$  subject to side constraints*

# Acknowledgements

Thanks to Jeff Eaton and Seth Flaxman for supervision of this research; staff and students of the StatML CDT at Imperial and Oxford; members of the HIV Inference Group at Imperial; members of the Machine Learning and Global Health Network; the Bill & Melinda Gates Foundation and EPSRC for funding this PhD; Mike McLaren, Kevin Esvelt, the Nucleic Acid Observatory team, and the Sculpting Evolution lab for hosting my visit to MIT; Alex Stringer for hosting my visit to Waterloo; the Effective Altruism community; my friends and family.

Adam Howes  
Imperial College London  
2023

# Abstract

Progress towards ending AIDS as a public health threat by 2030 is faltering. Disease burden is unevenly distributed. Effective public health response and prioritised prevention requires accurate, timely, high-resolution estimates of epidemic and demographic indicators. Thoughtful statistical modelling is required to overcome significant data challenges. In this thesis, I develop and apply Bayesian spatio-temporal methods for HIV surveillance.

# Contents

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>x</b>
<b>List of Abbreviations</b>	<b>xi</b>
<b>List of Notations</b>	<b>xii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Chapter overview . . . . .	2
<b>2 Background</b>	<b>3</b>
2.1 The HIV/AIDS epidemic . . . . .	3
2.2 Bayesian spatio-temporal statistics . . . . .	5
<b>3 Spatial structure</b>	<b>8</b>
3.1 Background . . . . .	9
3.2 Models based on adjacency . . . . .	9
3.3 Models using kernels . . . . .	9
3.4 Simulation study . . . . .	9
3.5 HIV prevalence study . . . . .	9
3.6 Discussion . . . . .	9
<b>4 A model for risk group proportions</b>	<b>10</b>
4.1 Background . . . . .	10
4.2 Data . . . . .	11
4.3 Model for risk group proportions . . . . .	11
4.4 Prevalence and incidence by risk group . . . . .	18
4.5 Discussion . . . . .	20

## *Contents*

<b>5</b>	<b>Fast approximate Bayesian inference</b>	<b>21</b>
5.1	Background . . . . .	21
5.2	The Naomi model . . . . .	21
5.3	Adaptive Gauss-Hermite quadrature . . . . .	21
5.4	Malawi case-study . . . . .	21
5.5	Discussion . . . . .	21
<b>6</b>	<b>Future work and conclusions</b>	<b>22</b>
6.1	Future work . . . . .	22
6.2	Conclusions . . . . .	22
 <b>Appendices</b>		
<b>A</b>	<b>The First Appendix</b>	<b>26</b>
<b>Works Cited</b>		<b>27</b>



# List of Figures

2.1	Overall picture. . . . .	4
4.1	Risk grid. . . . .	11
4.2	Proportion of FSW by age group (including the age groups 30-34, 35-39, 40-44 and 45-49) as produced by the disaggregation procedure.	18

## List of Tables

# List of Abbreviations

<b>HIV</b>	Human Immunodeficiency Virus.
<b>AIDS</b>	Acquired Immune Deficiency Syndrome.
<b>PEPFAR</b>	President’s Emergency Plan for AIDS Relief.
<b>HIV</b>	Demographic and Health Surveys.
<b>AIS</b>	AIDS Indicator Survey.
<b>MCMC</b>	Markov Chain Monte Carlo.
<b>INLA</b>	Integrated Nested Laplace Approximation.
<b>GP</b>	Gaussian Process.
<b>CAR</b>	Conditionally Auto-regressive.
<b>ANC</b>	Antenatal Clinic.
<b>ART</b>	Antiretroviral Therapy.
<b>UNAIDS</b>	United Nations Joint Programme on HIV/AIDS.
<b>CDC</b>	Centers for Disease Control and Prevention.
<b>UAT</b>	Unlinked Anonymous Testing.
<b>PMTCT</b>	Prevention of Mother-to-Child Transmission.
<b>PLHIV</b>	People Living with HIV.
<b>MPES</b>	Multi-parameter Evidence Synthesis.
<b>VI</b>	Variational Inference.
<b>SAE</b>	Small Area Estimation.
<b>GMRF</b>	Gaussian Markov Random Field.
<b>HMC</b>	Hamiltonian Monte Carlo.
<b>PrEP</b>	Pre-Exposure Prophylaxis
<b>PEP</b>	Post-Exposure Prophylaxis

# List of Notations

$\rho$	.....	HIV prevalence.
$\alpha$	.....	ART coverage.
$\mathcal{S}$	.....	Spatial study region $\mathcal{S} \subseteq \mathbb{R}^2$ .
$s \in \mathcal{S}$	.....	Point location.
$\mathcal{T}$	.....	Temporal study period $\mathcal{T} \subseteq \mathbb{R}$ .
$t \in \mathcal{T}$	.....	Time.

# 1

## Introduction

This thesis is about applied and methodological Bayesian statistics. It is Bayesian in the sense that I use probability models to draw conclusions from data. It is applied and methodological in the sense that I am concerned with real world questions and the means to answer them.

The real world questions relate to surveillance of HIV in sub-Saharan Africa. Though important progress has been made, over forty years since the first reported cases millions of people remain impacted by HIV each year. Quantifying the epidemic using statistics is an important part of the public health response, and the path towards disease control and elimination.

The data in this thesis are related to people. Usually, people answering survey questions or using healthcare facilities. Because the data have positions in space and time, statisticians describe it as spatio-temporal (Cressie and Wikle 2015). While there is a great diversity of spatio-temporal data, there are important and distinctive commonalities which make their shared study worthwhile.

Computation is an essential part of modern statistical practice. Each project chapter is accompanied by computer code, hosted on GitHub.

## **1.1 Chapter overview**

- Chapter 2: I start by reviewing the required background for the rest of the thesis. This background concerns the HIV/AIDS epidemic and Bayesian spatio-temporal statistics.
- Chapter 3: The prevailing model for spatial structure used in small-area estimation (Besag et al. 1991) was designed for analysis of a grid of pixels. The districts of a country are not a grid. I evaluate the practical consequences of this this concern (Howes, Eaton, et al. 2023+).
- Chapter 4: Adolescent girls and young women are a demographic group at disproportionate risk of HIV infection. The Global AIDS Strategy suggests prioritising interventions on the basis of behaviour to prevent the most new infections. I estimate the size of behavioural risk groups across priority countries to enable implementation of the strategy, and assess the potential benefits (Howes, Risher, et al. 2023).
- Chapter 5: The Naomi small-area estimation model (Eaton, Dwyer-Lindgren, et al. 2021) is used across countries to estimate HIV indicators. I develop an approximate Bayesian inference method, designed with Naomi in mind, which combines adaptive Gauss-Hermite quadrature with principal components analysis, and apply it to data in Malawi (Howes, Stringer, et al. 2023+).
- Chapter 6: I discuss avenues for future work, and my conclusions regarding the research.

# 2

## Background

### 2.1 The HIV/AIDS epidemic

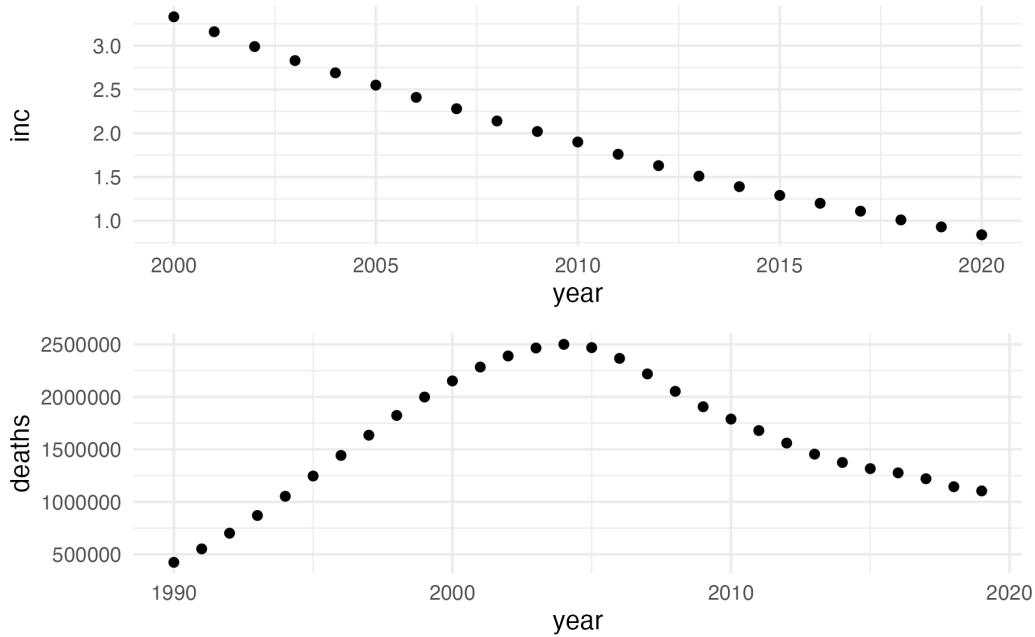
HIV/AIDS remains a source of substantial disease burden. In 2021 there were thirty-eight million people living with HIV, six hundred fifty thousand AIDS-related deaths, and one million, five hundred thousand people newly infected with HIV (UNAIDS 2021).

A major global effort has been made to address the epidemic. Significant progress has been made, both in reducing the number of new HIV cases and decreasing the number of AIDS related deaths (Figure 2.1). Roll out of antiretroviral therapy (ART) has been a key tool. Other interventions include condoms, pre-exposure prophylaxis (PrEP) and PEP, education, economic empowerment, VMMC.

Disease burden is not evenly distributed in space. The region most affected is Sub-Saharan Africa. Within sub-Saharan Africa, there is significant geographic heterogeneity.

Disease burden is not evenly distributed across populations. Disproportionately impacted groups are sometimes referred to as key populations. Key populations include men who have sex with men, female sex workers, people who inject drugs,

## Background



**Figure 2.1:** Overall picture.

transgender people, and incarcerated people. Key populations are often marginalised, and face legal and social issues.

In sub-Saharan Africa, the epidemic is not as concentrated in key populations as in other contexts. Large demographic groups at higher risk include adolescent girls and young women.

HIV interventions can be prioritised. Precision public health aims to get the right interventions, to the right population, in the right place, at the right time. Methods for prevention prioritisation include geographic, demographic, key population services, risk screening, and individual-level risk characteristics.

Surveillance is used is conducted to track epidemic trends, identify at-risk populations, find drivers of transmission, and evaluate the impact of prevention and treatment programs. HIV prevalence is the proportion of the population who have HIV. HIV incidence is the rate of new HIV infections. ART coverage is the proportion of people living with HIV who are on ART.

There are significant data related difficulties associated with furnishing these estimates. These include sparsity in space and time, survey bias, conflicting



## *Background*

information sources, hard to reach populations, changing demographics. These data limitations foreground the importance of synthesising multiple sources of information to obtain estimates. Doing so increases the difficulty and complexity of the statistical modelling required.

Aims for HIV response going forward, and surveillance capabilities are needed to meet them. Phasing out of nationally-representative household surveys for HIV.

## **2.2 Bayesian spatio-temporal statistics**

### **2.2.1 Bayesian statistics**

Bayesian statistics is a paradigm for learning from data. I provide a brief overview in this section, and recommend McElreath (2020) or Gelman et al. (2013) for a more complete introduction.

At it's best, the Bayesian framework allows the analyst focus their attention on modelling the data at hand. This is achieved by the construction of a generative model for the observed data  $y$  together with parameters  $\vartheta$

$$(y, \vartheta) \sim p(y, \vartheta). \tag{2.1}$$

This model is generative in the sense that one could simulate pairs  $(y, \vartheta)$  from it. If these pairs differ too greatly from what the analyst would expect, then the generative model may be refined. This is what is known as a prior predictive check.

Computation of the posterior distribution

$$p(\vartheta | y) = \frac{p(y | \vartheta)p(\vartheta)}{p(y)}, \tag{2.2}$$

usually proceeds using approximate Bayesian inference methods. Markov chain Monte Carlo (MCMC) is the most popular approach for computing Equation 2.2, and proceeds by simulating samples from a Markov chain with stationary distribution equal to the distribution of interest. Variational Bayes approaches assume the posterior distribution belongs to some class and use optimisation to choose the best member of that class. Particular properties of spatio-temporal

## Background

models make integrated nested Laplace approximations, if feasible, often the best option Empirical Bayes approaches, like Template Model Builder (Osgood-Zimmerman and Wakefield 2021).

### 2.2.2 Spatio-temporal statistics

In spatio-temporal statistics the data we observe are indexed by spatial or temporal location. Commonly used independent and identically distributed (IID) assumptions on observations are rarely suitable in this setting because we expect there to be spatio-temporal correlation structure.

Split the parameters  $\vartheta = (x, \theta)$ . Call  $x$  the latent field. Call  $\theta$  the hyperparameters. Often, the latent field is assumed to be jointly multivariate Gaussian.

### 2.2.3 Latent Gaussian and extended latent Gaussian models

Latent Gaussian models (Rue et al. 2009) are of the form

$$\begin{aligned} y_i &\sim p(y_i | \eta_i, \theta_1), \quad i \in [n] \\ \mu_i &= \mathbb{E}(y_i | \eta_i) = g(\eta_i), \\ \eta_i &= \beta_0 + \sum_{l=1}^p \beta_l z_{li} + \sum_{k=1}^r f_k(u_{ki}), \end{aligned}$$

where  $[n] = \{1, \dots, n\}$ . The response variable is  $y = (y)_{i \in [n]}$  with likelihood  $p(y | \eta, \theta_1) = \prod_{i=1}^n p(y_i | \eta_i, \theta_1)$ , where  $\eta = (\eta)_{i \in [n]}$ . Each response has conditional mean  $\mu_i$  with inverse link function  $g : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\mu_i = g(\eta_i)$ . The vector  $\theta_1 \in \mathbb{R}^s$ , with  $s_1$  assumed small, are additional parameters of the likelihood. The structured additive predictor  $\eta_i$  may include an intercept  $\beta_0$ , linear effects  $\beta_j$  of the covariates  $z_{ji}$ , and unknown functions  $f_k(\cdot)$  of the covariates  $u_{ki}$ . The parameters  $\beta_0, \{\beta_j\}, \{f_k(\cdot)\}$  are each assigned Gaussian priors. It is convenient to collect these parameters into a vector  $x \in \mathbb{R}^N$  called the latent field such that  $x \sim \mathcal{N}(0, Q(\theta_2)^{-1})$  where  $\theta_2 \in \mathbb{R}^{s_2}$  are further parameters, again with  $s_2$  assumed small. Let  $\theta = (\theta_1, \theta_2) \in \mathbb{R}^s$  with  $m = s_1 + s_2$  be all hyperparameters, with prior  $p(\theta)$ . Common examples of latent Gaussian models include the following.

## *Background*

Many of the leading-edge models used in small-area estimation fall outside the latent Gaussian model class. Examples include disaggregation models, evidence synthesis models (Eaton, Bajaj, et al. 2019; Eaton, Dwyer-Lindgren, et al. 2021), attendance models, risk group models. However, many of these models do fit into the class of extended latent Gaussian models (Stringer et al. 2021). By allowing many-to-one link functions, extended latent Gaussian models facilitate modelling of non-linearities.

### **2.2.4 Survey methods**

# 3

## Spatial structure

Code for the analysis in this chapter is available from `athowes/beyond-borders`

and supported by the R package `arealutils`. Include an edited version of the

corresponding paper [here](#).

## **3.1 Background**

### **3.1.1 Areal and point data**

### **3.1.2 Spatial random effects**

## **3.2 Models based on adjacency**

### **3.2.1 The Besag model**

### **3.2.2 The BYM2 model**

## **3.3 Models using kernels**

### **3.3.1 The centroid kernel model**

### **3.3.2 The integrated kernel model**

## **3.4 Simulation study**

### **3.4.1 Synthetic data-sets**

### **3.4.2 Inferential models**

Priors

Kernel details

### **3.4.3 Inference algorithms**

### **3.4.4 Model assessment**

Continuous ranked probability score

### **3.4.5 Results**

## **3.5 HIV prevalence study**

### **3.5.1 Results**

## **3.6 Discussion**

### **3.6.1 Limitations**

### **3.6.2 Conclusion**

# 4

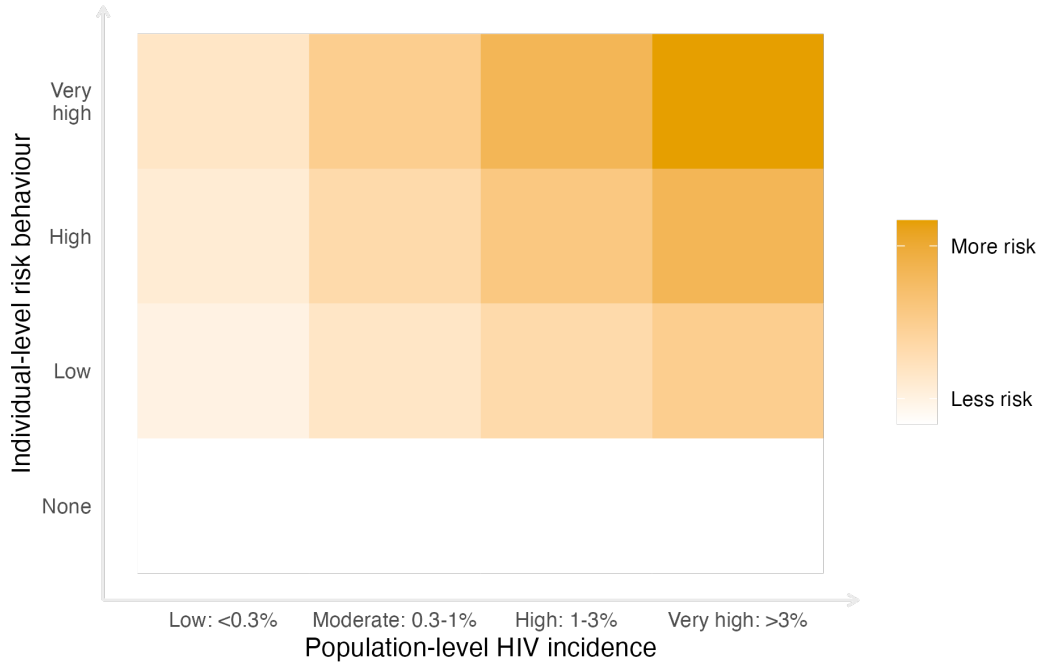
## A model for risk group proportions

In this chapter I describe an application of Bayesian spatio-temporal statistics to small-area estimation of HIV risk group proportions. This work was done in collaboration with colleagues from the MRC Centre for Global Infectious Disease Analysis and UNAIDS. My primary role was to develop the statistical model, building upon an earlier version of the analysis conducted by Kathryn Risher. The results are described in Howes, Risher, et al. (2023). Kathryn has created a spreadsheet tool which is beginning to be used by countries to guide policy. Code for the analysis in this chapter is available from `athowes/multi-agyw` and supported by the R package `multi.utils`.

### 4.1 Background

Risk of acquiring HIV infection is not equal for all individuals. Moreover, variation in risk is systematic across demographic characteristics. Adolescent girls and young women (AGYW) – that is, females aged 15-29 – are one such demographic group at particular risk.

The Global AIDS Strategy 2021-2026 proposed stratifying HIV prevention packages to AGYW based on local population-level HIV incidence and individual-



**Figure 4.1:** Risk grid.

level sexual risk behaviour. The idea is that prioritisation based on both of these factors is more efficient than on only or one the other.

Implementing the strategy requires estimates.

## 4.2 Data

I used household survey data.

## 4.3 Model for risk group proportions

I found that it was not appropriate to use the surveys without a specific transactional sex question on an equal footing to the other surveys. For this reason, I took a two-stage modelling approach to estimate the four risk group proportions. In particular, index the four risk groups as  $k \in \{1, 2, 3, 4\}$ , and denote being in either the third or fourth risk group by  $k = 3^+$ . First, using all the surveys, I used a multinomial logistic regression model to model the proportion of AGYW in the risk groups  $k \in \{1, 2, 3^+\}$ . Then, using only those surveys with a specific transactional

sex question, I fit a logistic regression model to estimate the proportion of those in the  $k = 3^+$  risk group that were in the  $k = 3$  and  $k = 4$  risk groups respectively.

### 4.3.1 Spatio-temporal multinomial logistic regression

Let  $i \in \{1, \dots, n\}$  denote districts which partition the 13 studied AGYW priority countries  $c[i] \in \{1, \dots, 13\}$ . Consider the years 1999-2018 denoted as  $t \in \{1, \dots, T\}$ , and age groups  $a \in \{15-19, 20-24, 25-29\}$ . Let  $p_{itak} > 0$  with  $\sum_{k=1}^{3^+} p_{itak} = 1$ , be the probability of membership of risk group  $k$  for an individual in district  $i$ , at time  $t$ , in age group  $a$ .

#### Multinomial logistic regression

A baseline category multinomial logistic regression model is specified by

$$\mathbf{y}_{ita} = (y_{ita1}, \dots, y_{ita3^+})^\top \sim \text{Multinomial}(m_{ita}; p_{ita1}, \dots, p_{ita3^+}), \quad (4.1)$$

$$\log(p_{itak}/p_{ita1}) = \eta_k, k = 2, 3^+, \quad (4.2)$$

where the number in risk group  $k$  is  $y_{itak}$ , the sample size is  $m_{ita} = \sum_{k=1}^{3^+} y_{itak}$ ,  $p_{itak} > 0$  is the probability of membership of the  $k$ th risk group, and  $k = 1$  is the baseline category.

#### The multinomial-Poisson transformation

The multinomial-Poisson transformation reframes a given multinomial logistic regression model as an equivalent Poisson log-linear model of the form

$$y_{itak} \sim \text{Poisson}(\kappa_{itak}), \quad (4.3)$$

$$\log(\kappa_{itak}) = \eta_{itak}, \quad (4.4)$$

for certain choice of the linear predictor  $\eta_{itak}$ . The basis of the transformation is that, conditional on their sum, Poisson counts are jointly multinomially distributed (McCullagh and Nelder 1989) as follows

$$\mathbf{y}_{ita} | m_{ita} \sim \text{Multinomial}\left(m_{ita}; \frac{\kappa_{ita1}}{\kappa_{ita}}, \dots, \frac{\kappa_{ita3^+}}{\kappa_{ita}}\right), \quad (4.5)$$



### *A model for risk group proportions*

where  $\kappa_{ita} = \sum_{k=1}^{3+} \kappa_{itak}$  such that category probabilities are obtained by the softmax function

$$p_{itak} = \frac{\exp(\eta_{itak})}{\sum_{k=1}^{3+} \exp(\eta_{itak})} = \frac{\kappa_{itak}}{\sum_{k=1}^{3+} \kappa_{itak}} = \frac{\kappa_{itak}}{\kappa_{ita}}. \quad (4.6)$$

In the equivalent model, the sample sizes  $m_{ita} = \sum_k y_{itak}$  are treated as random, rather than fixed as they would be in the multinomial logistic regression model, taking a Poisson distribution

$$m_{ita} \sim \text{Poisson}(\kappa_{ita}). \quad (4.7)$$

In the equivalent model, the joint distribution of  $p(\mathbf{y}_{ita}, m_{ita}) = p(\mathbf{y}_{ita} | m_{ita})p(m_{ita})$  is

$$p(\mathbf{y}_{ita}, m_{ita}) = \exp(-\kappa_{ita}) \frac{(\kappa_{ita})^{m_{ita}}}{m_{ita}!} \times \frac{m_{ita}!}{\prod_k y_{itak}!} \prod_k \left( \frac{\kappa_{itak}}{\kappa_{ita}} \right)^{y_{itak}} \quad (4.8)$$

$$= \prod_k \left( \frac{\exp(-\kappa_{itak}) (\kappa_{itak})^{y_{itak}}}{y_{itak}!} \right) \quad (4.9)$$

$$= \prod_k \text{Poisson}(y_{itak} | \kappa_{itak}). \quad (4.10)$$

corresponding to the product of independent Poisson likelihoods as in Equation 4.3. This model, including random sample sizes, is equivalent to the multinomial logistic regression only when these normalisation constants are recovered exactly. To ensure that this is the case, one approach is to include observation-specific random effects  $\theta_{ita}$  in the equation for the linear predictor. Multiplying each of  $\{\kappa_{itak}\}_{k=1}^{3+}$  by  $\exp(\theta_{ita})$  has no effect on the category probabilities, but does provide the necessary flexibility for  $\kappa_{ita}$  to recover  $m_{ita}$  exactly. Although in theory an improper prior  $\theta_{ita} \propto 1$  should be used, in practise, by keeping  $\eta_{ita}$  otherwise small using appropriate constraints, so that arbitrarily large values of  $\theta_{ita}$  are not required, it is sufficient (and practically preferable for inference) to instead use a vague prior.

### **Model specifications**

I considered four models for  $\eta_{ita}$  of the form

$$\eta_{ita} = \theta_{ita} + \beta_k + \zeta_{c[i]k} + \alpha_{ac[i]k} + \phi_{ik} + \gamma_{tk}.$$

### *A model for risk group proportions*

Observation random effects  $\theta_{ita} \sim \mathcal{N}(0, 1000^2)$  were included in all models we considered. To capture country-specific proportion estimates for each category, we included category random effects  $\beta_k \sim \mathcal{N}(0, \tau_\beta^{-1})$  and country-category random effects  $\zeta_{ck} \sim \mathcal{N}(0, \tau_\zeta^{-1})$ . Heterogeneity in risk group proportions by age was allowed by including age-country-category random effects  $\alpha_{ack} \sim \mathcal{N}(0, \tau_\alpha^{-1})$ . I considered two specifications, independent and identically distributed (IID) and Besag (Besag et al. 1991), for the space-category  $\phi_{ik}$  random effects (Section ??) and two specifications, IID and first order autoregressive (AR1), for the year-category  $\gamma_{tk}$  random effects (Section ??). All random effect precision parameters  $\tau \in \{\tau_\beta, \tau_\zeta, \tau_\alpha, \tau_\phi, \tau_\gamma\}$  were given independent penalised complexity (PC) priors (Simpson et al. 2017) with base model  $\sigma = 0$  given by  $p(\tau) = 0.5\nu\tau^{-3/2} \exp(-\nu\tau^{-1/2})$  where  $\nu = -\ln(0.01)/2.5$  such that  $\mathbb{P}(\sigma > 2.5) = 0.01$ .

### **Spatial random effects**

The specifications we considered were IID

$$\phi_{ik} \sim \mathcal{N}(0, \tau_\phi^{-1}),$$

and Besag grouped by category

$$\boldsymbol{\phi} = (\phi_{11}, \dots, \phi_{n1}, \dots, \phi_{13+}, \dots, \phi_{n3+})^\top \sim \mathcal{N}(\mathbf{0}, (\tau_\phi \mathbf{R}_\phi^*)^-),$$

where the scaled structure matrix  $\mathbf{R}_\phi^* = \mathbf{R}_b^* \otimes \mathbf{I}$  is given by the Kronecker product of the scaled Besag structure matrix  $\mathbf{R}_b^*$  and the identity matrix  $\mathbf{I}$ , and  $-$  denotes the generalised matrix inverse. Scaling of the structure matrix to have generalised variance one ensures interpretable priors may be placed on the precision parameter ([sorbye2014scaling](#)). We followed the further recommendations of [freni2018note](#) with regard to disconnected adjacency graphs, singletons and constraints. The Besag structure matrix  $\mathbf{R}_b$  is obtained by the precision matrix of the random effects  $\mathbf{b} = (b_1, \dots, b_n)^\top$  with full conditionals

$$b_i | \mathbf{b}_{-i} \sim \mathcal{N}\left(\frac{\sum_{j:j \sim i} b_j}{n_{\delta i}}, \frac{1}{n_{\delta i}}\right), \quad (4.11)$$

where  $j \sim i$  if the districts  $A_i$  and  $A_j$  are adjacent, and  $n_{\delta i}$  is the number of districts adjacent to  $A_i$ .

In preliminary testing, we excluded spatial random effects from the model, but found that this negatively effected performance. We also tested using the BYM2 model (Simpson et al. 2017) in place of the Besag, but found that the proportion parameter posteriors tended to be highly peaked at the value one. For simplicity and to avoid numerical issues, by using Besag random effects we decided to fix this proportion to one.

### Temporal random effects

The specifications we considered were IID

$$\phi_{tk} \sim \mathcal{N}(0, \tau_\phi^{-1}),$$

and AR1 grouped by category

$$\boldsymbol{\gamma} = (\gamma_{11}, \dots, \gamma_{13+}, \dots, \gamma_{T1}, \dots, \gamma_{T3+})^\top \sim \mathcal{N}(\mathbf{0}, (\tau_\phi \mathbf{R}_\gamma^*)^-),$$

where the scaled structure matrix  $\mathbf{R}_\gamma^* = \mathbf{R}_r^* \otimes \mathbf{I}$  is given by the Kronecker product of a scaled AR1 structure matrix  $\mathbf{R}_r^*$  and the identity matrix  $\mathbf{I}$ . The AR1 structure matrix  $\mathbf{R}_r$  is obtained by precision matrix of the random effects  $\mathbf{r} = (r_1, \dots, r_T)^\top$  specified by

$$r_1 \sim \left(0, \frac{1}{1 - \rho^2}\right), \tag{4.12}$$

$$r_t = \rho r_{t-1} + \epsilon_t, \quad t = 2, \dots, T, \tag{4.13}$$

where  $\epsilon_t \sim \mathcal{N}(0, 1)$  and  $|\rho| < 1$ . For the lag-one correlation parameter  $\rho$ , we used the PC prior, as derived by [sorbye2017penalised](#), with base model  $\rho = 1$  and condition  $\mathbb{P}(\rho > 0) = 0.75$ . We chose the base model  $\rho = 1$  corresponding to no change in behaviour over time, rather than the alternative  $\rho = 0$  corresponding to no correlation in behaviour over time, as we judged the former to be more plausible a priori.

## Constraints

To ensure interpretable posterior inferences of random effect contribution, we applied sum-to-zero constraints such that none of the category interaction random effects altered overall category probabilities. For the space-year-category random effects, we applied analogous sum-to-zero constraints to maintain roles of the space-category and year-category random effects. Together, these were:

1. Category  $\sum_k \beta_k = 0$
2. Country  $\sum_c \zeta_{ck} = 0, \forall k$
3. Age-country  $\sum_a \alpha_{ack} = 0, \forall c, k,$
4. Spatial  $\sum_i \phi_{ik} = 0, \forall k$
5. Temporal  $\sum_t \gamma_{tk} = 0, \forall k$

## Survey weighted likelihood

We included surveys which use a complex design, in which each individual has an unequal probability of being included in the sample. For example the DHS often employs a two-stage cluster design, first taking an urban rural stratified sample of enumeration areas, before selecting households from each enumeration area using systematic sampling (**measure2012sampling**).

To account for this aspect of survey design, we use a weighted pseudo-likelihood where the observed counts  $y$  are replaced by effective counts  $y^*$  calculated using the survey weights  $w_j$  of all individuals  $j$  in the corresponding strata. We multiplied direct estimates produced using the **survey** package (**JSSv009i08**) by the Kish effective sample size (**kish1965survey**)

$$m^* = \frac{\left(\sum_j w_j\right)^2}{\sum_j w_j^2} \quad (4.14)$$

to obtain  $y^*$ . These counts may not be integers, and as such the Poisson likelihood we used in Equation 4.3 is not appropriate. Instead, we used a generalised Poisson pseudo-likelihood  $y^* \sim \text{xPoisson}(\kappa)$ , given by

$$p(y^*) = \frac{\kappa^{y^*}}{[y^*!]} \exp(-\kappa), \quad (4.15)$$

as implemented by `family = "xPoisson"` in R-INLA, which accepts non-integer input.

## **Model selection**

### **4.3.2 Spatial logistic regression**

#### **Model specifications**

#### **Survey weighted likelihood**

#### **Model selection**

### **4.3.3 Coverage assessment**

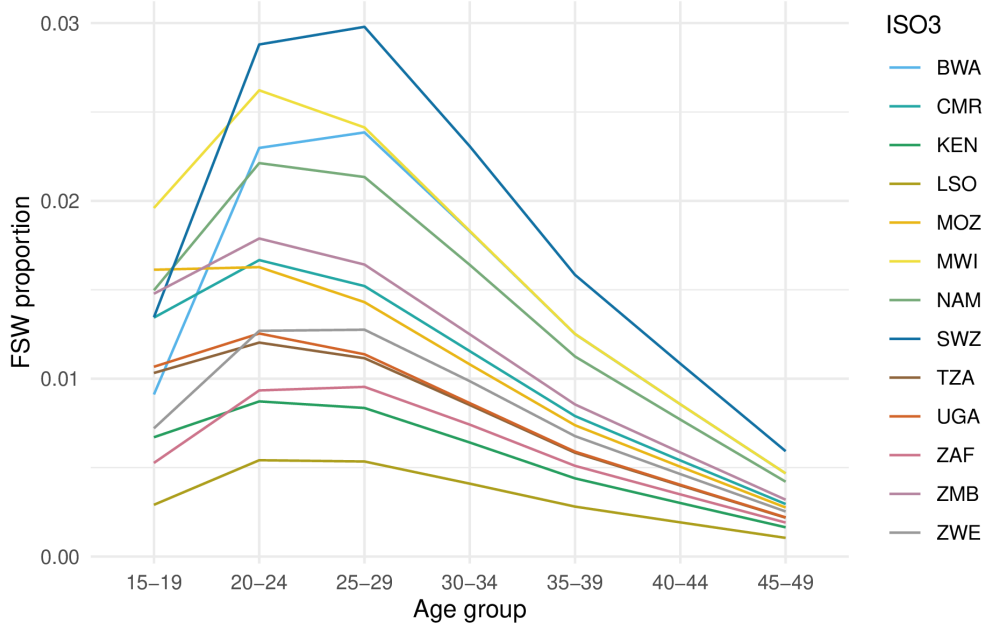
### **4.3.4 Female sex worker population size adjustment**

Responding “yes” to the survey question “have you had sex in return for gifts, cash or anything else in the past 12 months” is not considered sufficient to constitute sex work. In recognition of this, I adjusted the estimates obtained based on the survey to match FSW population size estimates obtained via alternative methods.

Stevens et al. (2022) used a Bayesian meta-analysis of key population specific data sources to estimate adult (15-49) FSW population size by country. I disaggregated these estimates by age according to the following method. First, I calculated the total sexually debuted population in each age group, in each country. To describe the distribution of age at first sex, I used skew logistic distributions (Nguyen and Eaton 2022) with cumulative distribution function given by

$$F(x) = (1 + \exp(\kappa_c(\mu_c - x)))^{-\gamma_c}, \quad (4.16)$$

where  $\kappa_c, \mu_c, \gamma_c > 0$  are country-specific shape, shape and skewness parameters respectively. Next, I used the assumed Gamma( $\alpha = 10.4, \beta = 0.36$ ) FSW age distribution in South Africa from the Thembisa model (Johnson and Dorrington 2020) to calculate the implied ratio between the number of FSW and the sexually debuted population in each age group. I assumed these ratios in South Africa were applicable to every country, allowing calculation of the number of FSW by age group in all 13 countries. The results obtained are shown in Figure 4.2.



**Figure 4.2:** Proportion of FSW by age group (including the age groups 30-34, 35-39, 40-44 and 45-49) as produced by the disaggregation procedure.

### 4.3.5 Results

Coverage assessment

Variance decomposition

Estimates

## 4.4 Prevalence and incidence by risk group

### 4.4.1 Disaggregation of Naomi estimates

I calculated HIV incidence  $\lambda_{iak}$  and number of new HIV infections  $I_{iak}$  stratified according to district, age group and risk group by linear disaggregation

$$I_{ia} = \sum_k I_{iak} = \sum_k \lambda_{iak} N_{iak} \quad (4.17)$$

$$= 0 + \lambda_{ia2} N_{ia2} + \lambda_{ia3} N_{ia3} + \lambda_{ia4} N_{ia4} \quad (4.18)$$

$$= \lambda_{ia2} (N_{ia2} + \text{RR}_3 N_{ia3} + \text{RR}_4 (\lambda_{ia}) N_{ia4}). \quad (4.19)$$

Risk group specific HIV incidence estimates are then given by

$$\lambda_{ia1} = 0, \quad (4.20)$$

$$\lambda_{ia2} = I_{ia} / (N_{ia2} + \text{RR}_3 N_{ia3} + \text{RR}_4(\lambda_{ia}) N_{ia4}), \quad (4.21)$$

$$\lambda_{ia3} = \text{RR}_3 \lambda_{ia2}, \quad (4.22)$$

$$\lambda_{ia4} = \text{RR}_4(\lambda_{ia}) \lambda_{ia2}. \quad (4.23)$$

which we evaluated using Naomi model estimates of the number of new HIV infections  $I_{ia} = \lambda_{ia} N_{ia}$ , HIV infection risk ratios  $\{\text{RR}_3, \text{RR}_4(\lambda_{ia})\}$ , and risk group population sizes as above. The risk ratio  $\text{RR}_4(\lambda_{ia})$  was defined as a function of general population incidence. The number of new HIV infections are then  $I_{iak} = \lambda_{iak} N_{iak}$ .

#### 4.4.2 Expected new infections reached

I calculated the number of new infections that would be reached prioritising according to each possible stratification of the population—that is for all  $2^3 = 8$  possible combinations of stratification by location, age, and risk group. As an illustration, for stratification just by age, we aggregated the number of new HIV infections and HIV incidence as such

$$I_a = \sum_{ik} I_{iak}, \quad (4.24)$$

$$\lambda_a = I_a / \sum_{ik} N_{iak}. \quad (4.25)$$

Under this stratification, individuals in each age group  $a$  are prioritised according to the highest HIV incidence  $\lambda_a$ . By cumulatively summing the expected infections, for each fraction of the total population reached we calculated the fraction of total expected new infections that would be reached.

This analysis was relatively simple. More involved analyses might consider prioritisation of a hypothetical intervention which has some, possibly varying, probability of preventing HIV acquisition, as well as the costs associated to its roll-out.

## **4.5 Discussion**

### **Distribution of risk**

- Connection to phylogenetic results from BDI
- About transmission rather than incidence
- Only age-sex structured not age-sex-behaviour
- Does not undermine my work

### **Community engagement**

- CSO engagement
- Problem in Malawi with FSW

#### **4.5.1 Limitations**

#### **4.5.2 Conclusion**



# 5

## Fast approximate Bayesian inference

In this chapter I describe a novel Bayesian inference method I developed with the aim of facilitating fast and accurate inference for the Naomi small-area estimation model. This work builds on that of others, including Rue et al. (2009), Kristensen et al. (2016) and Stringer et al. (2021). I began working on this project in 2020, but did not make significant progress until I read Alex Stringer’s work. We later began collaborating, including Alex supervising my visit to the University of Waterloo in 2022.

Code for the analysis in this chapter is available from `athowes/naomi-aghq` and supported by the R package `inf.utils`. Include an edited version of the corresponding paper [here](#).

### 5.1 Background

### 5.2 The Naomi model

### 5.3 Adaptive Gauss-Hermite quadrature

### 5.4 Malawi case-study

### 5.5 Discussion

# 6

## Future work and conclusions

### 6.1 Future work

Avenues for future work include:

1. Extending the risk group model described in Chapter 4 to include all adults 15-49. This may involve modelling of age-stratified sexual partnerships (Wolock et al. 2021). Such a model would likely fall out of the scope of **R-INLA**, but would be possible to write with **TMB** and therefore amenable to the methods discussed in 5.
2. Evaluating the accuracy of **aghq** with Laplace marginals for a greater variety of extended latent Gaussian models.

### 6.2 Conclusions

Chapter 3 is interesting because:

- I designed experiments to thoroughly compare models for spatial structure using tools for model assessment such as proper scoring rules and posterior predictive checks.

Chapter 4 is interesting because:

## *Conclusions*

- I estimated HIV risk group proportions for AGYW, enabling countries to prioritise their delivery of HIV prevention services.
- I analysed the number of new infections that might be reached under a variety of risk stratification strategies.
- I used **R-INLA** to specify multinomial spatio-temporal models via the Poisson-multinomial transformation. This includes complex two- and three-way Kronecker product interactions defined using the `group` and `replicate` options.

Chapter 5 is interesting because:

- I developed a novel Bayesian inference method, motivated by a challenging and practically important problem in HIV inference.
- The method enables integrated nested Laplace approximations to be fit to and studied on a wider class of models than was previously possible.
- My implementation of the method was straightforward, building on the **TMB** and **aghq** packages, and described completely and accessibly in Howes, Stringer, et al. (2023+).

My final conclusions are:

- Modelling complex data, more often than not, pushes the boundaries of the statistical toolkit available.
- A challenge I encountered was the difficulty of implementing identical models across multiple frameworks with the aim of studying the inference method. Or, of a similarly fraught nature, comparing different models implemented in different frameworks with the aim of studying model differences. The frequently asked questions section of the **R-INLA** website (Rue 2023) notes that “the devil is in the details”. I have resolved this challenge by using a given **TMB** model template to fit models using multiple inference methodologies. The benefits of such an ecosystem of packages are noted by Stringer (2021). I particularly highlight the benefit of enabling analysts to easily vary their

## *Conclusions*

choice of inference method based on the stage of model development that they are in.

- To the best of my abilities, I have written this thesis, and the work described within it, in keeping with the principles of open science. I hope that doing so allows my work to be scrutinised, and optimistically built upon. This would not have been possible without a range of tools from the R ecosystem such as `rmarkdown` and `rticles`, as well as those developed within the MRC Centre for Global Infectious Disease Analysis such as `orderly` and `didehpc`.

# Appendices



# The First Appendix

## Works Cited

- Besag, Julian, Jeremy York, and Annie Mollié (1991). “Bayesian image restoration, with two applications in spatial statistics”. In: *Annals of the Institute of Statistical Mathematics* 43.1, pp. 1–20.
- Cressie, Noel and Christopher K Wikle (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.
- Eaton, Jeffrey W, Sumali Bajaj, et al. (2019). “Joint small-area estimation of HIV prevalence, ART coverage and HIV incidence”. In: *Working paper*.
- Eaton, Jeffrey W, Laura Dwyer-Lindgren, et al. (2021). “Naomi: A New Modelling Tool for Estimating HIV Epidemic Indicators at the District Level in Sub-Saharan Africa”. In.
- Gelman, Andrew et al. (2013). *Bayesian data analysis*. CRC press.
- Howes, Adam, Jeffrey W. Eaton, and Seth R. Flaxman (2023+). “Beyond borders: evaluating the suitability of spatial adjacency for small-area estimation”. In.
- Howes, Adam, Kathryn A. Risher, et al. (Apr. 2023). “Spatio-temporal estimates of HIV risk group proportions for adolescent girls and young women across 13 priority countries in sub-Saharan Africa”. In: *PLOS Global Public Health* 3.4, pp. 1–14. DOI: 10.1371/journal.pgph.0001731. URL: <https://doi.org/10.1371/journal.pgph.0001731>.
- Howes, Adam, Alex Stringer, et al. (2023+). “Fast approximate Bayesian inference of HIV indicators using PCA adaptive Gauss-Hermite quadrature”. In.
- Johnson, L and RE Dorrington (2020). “Thembisa version 4.3: A model for evaluating the impact of HIV/AIDS in South Africa”. In: *View Article*.
- Kristensen, Kasper et al. (2016). “TMB: Automatic Differentiation and Laplace Approximation”. In: *Journal of Statistical Software* 70.i05.
- McCullagh, Peter and John A Nelder (1989). *Generalized linear models*. Routledge.
- McElreath, Richard (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.
- Nguyen, Van K  nh and Jeffrey W. Eaton (2022). “Trends and country-level variation in age at first sex in sub-Saharan Africa among birth cohorts entering adulthood between 1985 and 2020”. In: *BMC Public Health* 22.1, p. 1120. DOI: 10.1186/s12889-022-13451-y. URL: <https://doi.org/10.1186/s12889-022-13451-y>.
- Osgood-Zimmerman, Aaron and Jon Wakefield (2021). *A Statistical Introduction to Template Model Builder: A Flexible Tool for Spatial Modeling*. arXiv: 2103.09929 [stat.ME].
- Rue, Havard (2023). “‘R-INLA’ Project - FAQ”. Accessed 23/01/2023. URL: <https://www.r-inla.org/faq>.
- Rue, H  vard, Sara Martino, and Nicolas Chopin (2009). “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace

## Works Cited

- approximations”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 319–392.
- Simpson, Daniel et al. (2017). “Penalising model component complexity: A principled, practical approach to constructing priors”. In: *Statistical Science* 32.1, pp. 1–28.
- Stevens, Oliver et al. (2022). “Estimating key population size, HIV prevalence, and ART coverage for sub-Saharan Africa at the national level”. In.
- Stringer, Alex (2021). “Implementing Approximate Bayesian Inference Using Adaptive Quadrature”. Statistics Graduate Student Research Day 2021, The Fields Institute for Research in Mathematical Sciences. URL:  
<http://www.fields.utoronto.ca/talks/Implementing-Approximate-Bayesian-Inference-Using-Adaptive-Quadrature>.
- Stringer, Alex, Patrick Brown, and Jamie Stafford (2021). “Fast, Scalable Approximations to Posterior Distributions in Extended Latent Gaussian Models”. In: *arXiv preprint arXiv:2103.07425*.
- UNAIDS (2021). *Global AIDS Update 2021*.  
<https://www.unaids.org/en/resources/documents/2021/global-aids-update>. Accessed: June 2023.
- Wolock, Timothy M et al. (June 2021). “Evaluating distributional regression strategies for modelling self-reported sexual age-mixing”. In: *eLife* 10. Ed. by Eduardo Franco, Talía Malagón, and Adam Akullian, e68318. DOI: 10.7554/eLife.68318. URL:  
<https://doi.org/10.7554/eLife.68318>.