Methods and applications of Bayesian spatio-temporal statistics for prioritised HIV prevention

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

Statement of Originality

Putting aside the parts I didn't do, I did this work. Me. Adam Howes.

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; 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

List of Figures

List of Tables

List of Abbreviations

HIV . . . . . Human Immunodeficiency Virus.

AIDS . . . . . Acquired Immune Deficiency Syndrome.

PEPFAR . . . President's Emergency Plan for AIDS Relief.

HIV . . . . . Demographic and Health Surveys.

AIS . . . . . AIDS Indicator Survey.

MCMC . . . . Markov Chain Monte Carlo.

INLA . . . . . Integrated Nested Laplace Approximation.

GP . . . . . Gaussian Process.

CAR . . . . . Conditionally Auto-regressive.

ANC . . . . . Antenatal Clinic.

ART . . . . . Antiretroviral Therapy.

UNAIDS . . . United Nations Joint Programme on HIV/AIDS.

CDC . . . . . Centers for Disease Control and Prevention.

UAT . . . . . Unlinked Anonymous Testing.

PMTCT . . . Prevention of Mother-to-Child Transmission.

PLHIV . . . . People Living with HIV.

MPES . . . . . Multi-parameter Evidence Synthesis.

VI . . . . . Variational Inference.

SAE . . . . . Small Area Estimation.

GMRF . . . . Gaussian Markov Random Field.

HMC . . . . . Hamiltonian Monte Carlo.

List of Notations

ρ	HIV prevalence.
α	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 Background

1.1 Small-area estimation

Small-area estimation methods aim to estimate population indicators for subgroups, typically in situations where direct estimates perform poorly due to data limitations. These subgroups may often correspond to small geographic areas. Small-area estimation methods have been used in a wide range of fields. The Small-Area Health Statistics Unit (SASHU) at Imperial College London was set-up to monitor health around point sources of environmental pollution in response to the Sellafield enquiry into the increased incidence of childhood leukemia leukaemia near a nuclear reprocessing plant (Elliott et al. 1992). The research of SASHU has a focus on ratios of observed events to expected events, and testing hypothesis about hot-spots.

1.2 The HIV/AIDS epidemic

```
plhiv2022 <- 38000000
deaths2022 <- 700000
infections2022 <- 1700000
```

According to latest estimates, in 2022 thirty-eight million people are living with HIV, there were seven hundred thousand AIDS-related deaths, and there were

one million, seven hundred thousand people newly infected with HIV. 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. Sub-Saharan Africa is the most affected region. Within sub-Saharan Africa, disease burden is unevenly distributed in space and across communities and individuals. Key populations include men who have sex with men, female sex workers, people who inject drugs, transgender people, incarcerated people. Larger demographic groups of higher risk include adolescent girls and young women. Key HIV indicators are HIV prevalence, HIV incidence, coverage of ART and other interventions. Key interventions are ART, condoms, PrEP and PEP, education, economic empowerment, VMMC.

There are significant data related difficulties associated with furnishing these estimates. These include sparsity in space and time, survey bias, conflicting information sources, hard to reach populations, changing demographies. 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.

Methods for prevention prioritisation include geographic, demographic, key population services, risk screening, individual-level risk characteristics. Are there differences in effectiveness of treatments for different groups.

The population strategy (Rose 2001) is based on reducing risk factors across an entire population. The individual strategy focuses on prevention in high-risk individuals.

1.3 Bayesian spatio-temporal statistics

Bayesian statistics is a statistical paradigm which, at its best, lets the analyst focus their attention on modelling the data at hand. In particular, the primary concern

is construction of a generative model for the observed data y

$$(y, \vartheta) \sim p(y, \vartheta).$$

Given a generative model, computation of the posterior distribution

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

proceeds using approximate Bayesian inference methods. Markov chain Monte Carlo (MCMC) is the most popular approach, 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 models make integrated nested Laplace approximations, if feasible, often the best option Empirical Bayes approaches, like Template Model Builder (Osgood-Zimmerman and Wakefield 2021).

In spatio-temporal statistics the data we observe are indexed by spatial or temporal location. The independent and identically distributed (IID) assumptions commonly used for observations are rarely suitable in this setting because we expect there to be spatio-temporal structure. Split the parameters $\vartheta = (x, \theta)$. Call x the latent field. Call θ the hyperparameters. Often, the latent field is assumed to be jointly multivariate Gaussian.

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

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

where $[n] = \{1, \ldots, n\}$. 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 μ_i with inverse link function $g : \mathbb{R} \to \mathbb{R}$ 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 η_i may include an intercept β_0, linear effects β_j

of the covariates z_{ji}, and unknown functions $f_k(\cdot)$ of the covariates u_{ki}. The parameters β_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.

Many of the cutting-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

Understanding models for spatial structure

Code for the analysis in this chapter is available from `athowes/areal-comparison` and supported by the R package `arealutils`. Include an edited version of the corresponding paper here.

2.1 Background

2.1.1 Areal and point data

2.1.2 Spatial random effects

2.2 Models based on adjacency

2.2.1 The Besag model

2.2.2 The BYM2 model

2.3 Models using kernels

2.3.1 The centroid kernel model

2.3.2 The integrated kernel model

2.4 Simulation study

2.4.1 Synthetic data-sets

2.4.2 Inferential models

Priors

Kernel details

2.4.3 Inference algorithms

2.4.4 Model assessment

Continuous ranked probability score

2.4.5 Results

2.5 HIV prevalence study

2.5.1 Results

2.6 Discussion

2.6.1 Limitations

2.6.2 Conclusion

3

A multinomial spatio-temporal 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 role in the work was to develop the statistical model, building upon an earlier version of the analysis conducted by Kathryn Risher. The results are described in Howes et al. (2023), and are now beginning to be used by countries to guide policy via a spreadsheet tool created by Kathryn. Code for the analysis in this chapter is available from `athowes/multi-agyw` and supported by the R package `multi.utils`.

3.1 Background

The risk of acquiring HIV infection is not equal for all individuals.

3.2 Data

I used the following household surveys.

3.3 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 as the other surveys. For this reason, I took a two-stage modelling approach to estimating the four risk group proportions. In particular, let the four risk groups be $k \in \{1, 2, 3, 4\}$, 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.

3.3.1 Spatio-temporal multinomial logistic regression

As notation, let $i \in \{1, \ldots, n\}$ denote subnational units which partition the 13 studied AGYW priority countries $c[i] \in \{1, \ldots, 13\}$. Consider the years 1999-2018 denoted as $t \in \{1, \ldots, T\}$, and age groups $a \in \{15\text{–}19, 20\text{–}24, 25\text{–}29\}$.

The multinomial-Poisson transformation

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

$$y_{itak} \sim \text{Poisson}(\kappa_{itak}), \tag{3.1}$$

$$\log(\kappa_{itak}) = \eta_{itak}, \tag{3.2}$$

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

$$\mathbf{y}_{ita} \,|\, m_{ita} \sim \text{Multinomial}\left(m_{ita}; \frac{\kappa_{ita1}}{\kappa_{ita}}, \ldots, \frac{\kappa_{ita3^+}}{\kappa_{ita}}\right), \tag{3.3}$$

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

$$p_{itak} = \frac{\exp(\eta_{itak})}{\sum_{k=1}^{3^+} \exp(\eta_{itak})} = \frac{\kappa_{itak}}{\sum_{k=1}^{3^+} \kappa_{itak}} = \frac{\kappa_{itak}}{\kappa_{ita}}. \tag{3.4}$$

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

$$m_{ita} \sim \text{Poisson}\left(\kappa_{ita}\right). \tag{3.5}$$

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

$$p(\mathbf{y}_{ita}, m_{ita}) = \exp(-\kappa_{ita})\frac{(\kappa_{ita})^{m_{ita}}}{m_{ita}!} \times \frac{m_{ita}!}{\prod_k y_{itak}!}\prod_k \left(\frac{\kappa_{itak}}{\kappa_{ita}}\right)^{y_{itak}} \tag{3.6}$$

$$= \prod_k \left(\frac{\exp(-\kappa_{itak})\,(\kappa_{itak})^{y_{itak}}}{y_{itak}!}\right) \tag{3.7}$$

$$= \prod_k \text{Poisson}\left(y_{itak} \,|\, \kappa_{itak}\right). \tag{3.8}$$

corresponding to the product of independent Poisson likelihoods as in Equation 3.1. 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 θ_{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 κ_{ita} to recover m_{ita} exactly. Although in theory an improper prior $\theta_{ita} \propto 1$ should be used, in practise, by keeping η_{ita} otherwise small using appropriate constraints, so that arbitrarily large values of θ_{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 η_{ita} of the form

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

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 ϕ_{ik} random effects (Section **??**) and two specifications, IID and first order autoregressive (AR1), for the year-category γ_{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 $p(\tau) = 0.5\nu\tau^{-3/2}\exp\left(-\nu\tau^{-1/2}\right)$ 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

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

where the scaled structure matrix $\mathbf{R}_\phi^\star = \mathbf{R}_b^\star \otimes \mathbf{I}$ is given by the Kronecker product of the scaled Besag structure matrix $\mathbf{R}_b^\star$ 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 $\mathbf{b} = (b_1, \ldots, b_n)^\top$ with full conditionals

$$b_i \,|\, \mathbf{b}_{-i} \sim \mathcal{N}\left(\frac{\sum_{j:j\sim i} b_j}{n_{\delta i}}, \frac{1}{n_{\delta i}}\right), \tag{3.9}$$

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

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

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

$$r_1 \sim \left(0, \frac{1}{1-\rho^2}\right), \tag{3.10}$$

$$r_t = \rho r_{t-1} + \epsilon_t, \quad t = 2, \ldots, T, \tag{3.11}$$

where $\epsilon_t \sim \mathcal{N}(0, 1)$ and $|\rho| < 1$. For the lag-one correlation parameter ρ, 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 ennumeration areas, before selecting households from each ennumeration 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 $y^\star$ 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**)

$$m^\star = \frac{\left(\sum_j w_j\right)^2}{\sum_j {w_j}^2} \tag{3.12}$$

to obtain $y^\star$. These counts may not be integers, and as such the Poisson likelihood we used in Equation 3.1 is not appropriate. Instead, we used a generalised Poisson pseudo-likelihood $y^\star \sim \text{xPoisson}(\kappa)$, given by

$$p(y^\star) = \frac{\kappa^{y^\star}}{\lfloor y^\star ! \rfloor} \exp\left(-\kappa\right), \tag{3.13}$$

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

Model selection

3.3.2 Spatial logistic regression

Model specifications

Survey weighted likelihood

Model selection

3.3.3 Coverage assessment

3.3.4 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

$$F(x) = (1 + \exp(\kappa_c(\mu_c - x)))^{-\gamma_c}, \tag{3.14}$$

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 3.1.

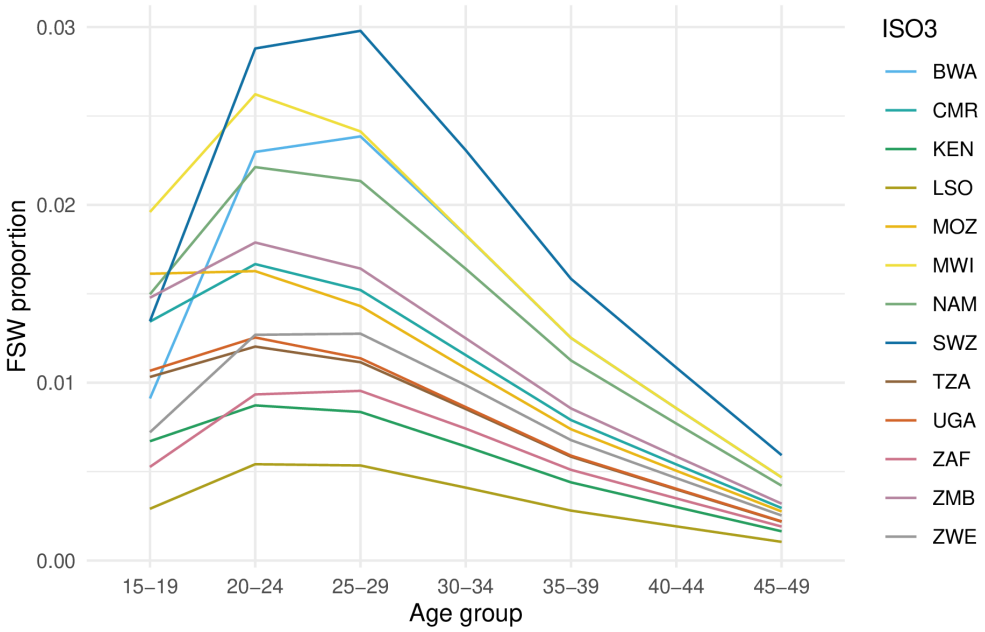

Figure 3.1: Proportion of FSW by age group (including the age groups 30-34, 35-39, 40-44 and 45-49) as produced by the disaggregation procedure.

3.3.5 Results

Coverage assessment

Variance decomposition

Estimates

3.4 Calculation of prevalence and incidence stratified by risk group

3.4.1 Disaggregation of Naomi estimates

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

$$I_{ia} = \sum_k I_{iak} = \sum_k \lambda_{iak} N_{iak} \tag{3.15}$$

$$= 0 + \lambda_{ia2} N_{ia2} + \lambda_{ia3} N_{ia3} + \lambda_{ia4} N_{ia4} \tag{3.16}$$

$$= \lambda_{ia2} \left(N_{ia2} + \mathrm{RR}_3 N_{ia3} + \mathrm{RR}_4(\lambda_{ia}) N_{ia4} \right). \tag{3.17}$$

Risk group specific HIV incidence estimates are then given by

$$\lambda_{ia1} = 0, \tag{3.18}$$

$$\lambda_{ia2} = I_{ia} / \left(N_{ia2} + \mathrm{RR}_3 N_{ia3} + \mathrm{RR}_4(\lambda_{ia}) N_{ia4} \right), \tag{3.19}$$

$$\lambda_{ia3} = \mathrm{RR}_3 \lambda_{ia2}, \tag{3.20}$$

$$\lambda_{ia4} = \mathrm{RR}_4(\lambda_{ia}) \lambda_{ia2}. \tag{3.21}$$

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

3.4.2 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

$$I_a = \sum_{ik} I_{iak}, \tag{3.22}$$

$$\lambda_a = I_a / \sum_{ik} N_{iak}. \tag{3.23}$$

Under this stratification, individuals in each age group a are prioritised according to the highest HIV incidence λ_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.

3.5 Discussion

3.5.1 Limitations

3.5.2 Conclusion

4

Fast, approximate inference for the Naomi model

Code for the analysis in this chapter is available from `athowes/elgm-inf` and supported by the R package `inf.utils`. Include an edited version of the corresponding paper here.

5 Future work and conclusions

5.1 Future work

Avenues for future work include:

1. Extending the risk group model described in Chapter 3 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 may be possible using `aghq` with Laplace marginals as described in Chapter 4.
2. Evaluating the accuracy of `aghq` with Laplace marginals for a greater variety of extended latent Gaussian models.

5.2 Conclusions

The spatial structure chapter 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.

The risk group chapter 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.

The fast, approximate inference chapter 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 **howes2023integrated**.

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: empirical Bayes with Gaussian marginals (Kristensen et al. 2016), AGHQ with Gaussian marginals (Stringer 2021b), AGHQ with Laplace marginals (**howes2023integrated**), and HMC using NUTS (Monnahan and Kristensen

2018). The benefits of such a ecosystem of packages are noted by Stringer (2021a). I would particularly highlight the benefit of enabling analysts to easily vary their choice of inference method based on the stage of model development that they are in.

- I have aimed to write 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 like `orderly` and `didehpc`.

Appendices

A

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.

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.

Elliott, Paul et al. (1992). "The Small Area Health Statistics Unit: a national facility for investigating health around point sources of environmental pollution in the United Kingdom." In: *Journal of Epidemiology & Community Health* 46.4, pp. 345–349.

Howes, Adam 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.

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.

Monnahan, Cole C and Kasper Kristensen (2018). "No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages". In: *PloS one* 13.5, e0197954.

Nguyen, Van Kính 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].

Rose, Geoffrey (2001). "Sick individuals and sick populations". In: *International Journal of Epidemiology* 30.3, pp. 427–432.

Rue, Havard (2023). "'R-INLA' Project - FAQ". Accessed 23/01/2023. URL: https://www.r-inla.org/faq.

Rue, Håvard, Sara Martino, and Nicolas Chopin (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace 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 (2021a). "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.

— (2021b). "Implementing Approximate Bayesian Inference using Adaptive Quadrature: the aghq Package". In: *arXiv preprint arXiv:2101.04468*.

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*.

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.