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

Imperial College
London

Adam Howes

Imperial College London

A thesis submitted for the degree of

Doctor of Philosophy

2023

For $\sum_i u_i$

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; 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; the Bill & Melinda Gates Foundation and EPSRC for funding this PhD; my friends and family for their support.

Adam Howes
Imperial College London
2023

Abstract

HIV remains a large problem. 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

Appendices

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 (Howes, Risher, et al. 2022). Code for the analysis in this chapter is available from `athowes/multi-agyw` and supported by the R package `multi.utils`.

3.1 Background

3.2 Data

3.3 Model for risk group proportions

To estimate the proportion of AGYW in each risk group, I took a two-stage modelling approach.

3.3.1 Spatio-temporal multinomial logistic-regression

The multinomial-Poisson transformation

Alternative model specifications

3.3.2 Spatial logistic regression

Alternative model specifications

3.3.3 Female sex worker population size adjustment

To estimate the number of FSW by age group and country, we disaggregated country-specific estimates of adult (15-49) FSW population size from Stevens et al. (2022) by age group. First, we calculated the total sexually debuted population in each age group, in each country. To describe the distribution of age at first sex, we used skew logistic distributions (Nguyen and Eaton 2022) with cumulative distribution function given by

$$F(x) = (1 + \exp(\kappa_c(\mu_c - x)))^{-\gamma_c}\,, \tag{3.1}$$

where $\kappa_c, \mu_c, \gamma_c > 0$ are country-specific shape, shape and skewness parameters respectively. Next, we 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. We assumed these ratios in South Africa were applicable to every country to calculate the number of FSW by age group in all 13 countries.

3.3.4 Results

Coverage assessment

Variance decomposition

Estimates

3.4 Calculation of prevalence and incidence stratified by risk group

3.4.1 Disaggregation of Naomi estimates

3.4.2 Expected new infections reached

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:

- 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 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: empirical Bayes with Gaussian marginals (Kristensen et al. 2016), AGHQ with Gaussian marginals (Stringer 2021b), AGHQ with Laplace marginals (Howes,

Stringer, et al. 2023), 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

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, Kathryn A Risher, et al. (2022). "Spatio-temporal estimates of HIV risk group proportions for adolescent girls and young women across 13 priority countries in sub-Saharan Africa". In: *medRxiv*.

Howes, Adam, Alex Stringer, et al. (2023). "Integrated nested Laplace approximations for extended latent Gaussian models with application to the Naomi HIV model". In: *arXiv*.

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.

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.

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`.