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 lab at Imperial; Mike McLaren, Kevin Esvelt 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

Disease surveillance alters policy. Complex statistical models and quantification of uncertainty are required.

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.

Background

0.1 Disease surveillance

- Small-area estimation in health, epidemiology and environment
  - The Small-Area Health Statistics Unit at Imperial 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). This research has a focus on ratios of observed events to expected events, and testing hypothesis about hot-spots

0.2 HIV/AIDS

- The HIV/AIDS epidemic has a large disease burden
- The disease burden is unevenly distributed in space and across communities
- Surveillance techniques and statistical models have been used to respond to the epidemic
- Aims for HIV response going forward, and surveillance capabilities are needed to meet them
- Data difficulties including sparsity in space and time, survey bias, conflicting information sources, hard to reach populations, demography
- Phasing out of nationally-representative household surveys for HIV. Importance of relying on multiple sources of information: evidence synthesis, Naomi, multivariate models. Bayesian survey design
- Why isn't case-based surveillance included yet?

  - There aren't individual linked databases and patient records have to be consolidated
  - Passive case-based surveillance
  - Post-hoc matching and create a case-based surveillance record

- Key HIV indicators are HIV prevalence, HIV incidence, ART coverage and coverage of other interventions such as PrEP, PEP
- Drivers of transmission
- Possible interventions are ART, condoms, PrEP and PEP, education, economic empowerment, VMMC
- Geographic priorisation versus demographic priorisation: hotspots, key populations, screening and individual level risk characteristics
- Adolescent girls and young women identified as a key demographic, stratification by sexual risk
- Interventions more likely to be demographic specific rather than geographic specific so if majority of difference in effectiveness depends on intervention type then demographic targeting may be more priority
- The population strategy of Geoffrey Rose

0.3 Statistics

- Definition of a latent Gaussian model (Rue et al. 2009)
  - Common examples
- Examples of models used in HIV inference which are close to being latent Gaussian models, but aren't, and hence can't be fit using INLA
  - Disaggregation models
  - Evidence synthesis models like Naomi (Eaton, Dwyer-Lindgren, et al. 2021; Eaton, Bajaj, et al. 2019)
  - Compartmental models
  - ART attendance models

- Multinomial models like for district-level risk factors
  - Multinomial logistic regression

- Other complex models from ecology that can't currently be fit using INLA
- Definition of extended latent Gaussian models (Stringer et al. 2021)
  - Many-to-one is not an issue for `R-INLA`, the latent field is implemented as a concatenation of many vectors already. For example, for $\eta_i = \beta_0 + \phi_i$ with $i = 1, \ldots, n$ the latent field is $(\eta_1, \ldots, \eta_n, \beta_0, \phi_1, \ldots, \phi_n)^\top$ of dimension $2n + 1$
  - For additive models, the only non-linearity is in the link function
- Approximate Bayesian inference methods
  - Markov chain Monte Carlo
  - Variational Bayes
  - Laplace approximation and integrated nested Laplace approximation
  - Empirical Bayes
- Particular properties of spatio-temporal models (and LGMs) which make INLA, if feasible, often the best option
- The increasing popularity of empirical Bayes approaches, like Template Model Builder (Osgood-Zimmerman and Wakefield 2021)
- Adaptive Gauss Hermite quadrature (AGHQ), like the central composite design (CCD) and grid strategies, is one way to choose the hyper-parameter integration points in the integrated nested Laplace approximation (INLA)
- Finn Lindgren is working on a method for non-linear predictors, called the iterative INLA method
  - More slides here
- Thesis work of Follestad that stayed as a preprint
- How does the ecological fallacy relate to aggregated output models

Understanding models for spatial structure in small-area estimation

The repository for this work is `athowes/areal-comparison`. Include an edited version of the corresponding paper here.

Spatio-temporal estimates of HIV risk group proportions for adolescent girls and young women across 13 priority countries in sub-Saharan Africa

The repository for this work is `athowes/multi-agyw`. Include an edited version of the corresponding paper here.

Simplifying Integrated nested Laplace approximation with adaptive Gaussian Hermite quadrature

The repository for this work is `athowes/elgm-inf`. Include an edited version of the corresponding paper here.

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.

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, 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.

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*.