Mapping malaria prevalence in Kenya by reconciling changes in administrative boundaries using MCMC and deep learning

Elizaveta Semenova[1], Swapnil Mishra[2], Samir Bhatt[2], Seth Flaxman[1], H. Juliette T. Unwin[3]

[1]University of Oxford, [2]University of Copenhagen, [3]Imperial College London

{elizaveta.semenova, seth.flaxman}@cs.ox.ac.uk, {samir.bhatt, swapnil.mishra}@sund.ku.dk, h.unwin@imperial.ac.uk

Background

- In Kenya approximately 75% of the population is still at risk of malaria in 2022.
- District level disease mapping remains a fundamental surveillance tool for analysing spatial disease distribution.
- Administrative boundaries changed in Kenya in 2010 from 69 districts to 47 districts, presenting a methodological challenge.
- Hierarchical Bayesian models are the current state-of-the-art approach to mapping but are computationally intensive since they capture spatial correlations by Gaussian process (GP) priors.

Contributions

We present a practical solution which

- relies on a novel methodology combining deep learning and fully Bayesian inference [1, 2],
- uses continuous representation of spatial processes and aggregates them to the district level,
- learns the aggregated GP priors with a variational autoencoder (VAE) and uses the learnt representation for inference.

We demonstrate that it is faster and more efficient than state-of-the-art Bayesian models estimated via Markov Chain Monte Carlo (MCMC).

Malaria prevalence data and administrative boundaries

We used DHS 2015 survey containing information on locations of clusters and test positivity to calculate district-specific prevalence.

Method

Gaussian process over a fine grid

GP prior $f(.)$ realisations are drawn over a fine artificial grid $\{g_1, ...g_n\}$ covering the study domain as a multivariate normal distribution:

$$f = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix} \sim \text{MVN}(0, \Sigma), \quad \Sigma_{jk} = \sigma^2 \exp\left(-\frac{d_{jk}^2}{2l^2}\right),$$

where $f_j = f(g_j)$, $d_{jk} = \text{dist}(g_j, g_k)$ and σ^2, l are hyperparameters.

Aggregation.

We aggregate the continuous process $f(.)$ to the district level (polygons $p_1, ..., p_K$):

$$f_{\text{GP-aggr}}^{p_i} = \int_{p_i} f(s)ds \approx c \sum_{g_j \in p_i} f_j.$$

We can, therefore, construct a vector where each entry represents spatial random effect at a district:

$$f_{\text{GP-aggr}} = \begin{pmatrix} f_{\text{GP-aggr}}^{p_1} \\ \vdots \\ f_{\text{GP-aggr}}^{p_K} \end{pmatrix}$$

The vector $f_{\text{GP-aggr}}$ collects values both old and new administrative boundaries, e.g. via stacking.

VAE-Encoding.

We encode $f_{\text{GP-aggr}}$ jointly for old and new boundaries with a VAE using a lower-dimensional representation with independent Gaussian components $z_1, ..., z_d, d < K$ as $f_{\text{VAE-aggr}}$, using the technique from [1].

Model.

Malaria prevalence $\theta_i, i \in 1, \dots K$ is inferred using the Negative Binomial distribution

$$\begin{cases} n_i^{\text{pos}} & \sim \text{NegBin}(n_i^{\text{tests}}, \theta_i), \\ \text{logit}(\theta_i) & = b_0 + f_{\text{aggr}}^{p_i}. \end{cases}$$

where n_i^{tests} and n_i^{pos} are the number of total and positive RDT tests, correspondingly.

Inference.

At the inference stage, we use the encoded spatial random effect $f_{\text{VAE-aggr}}$ instead of $f_{\text{GP-aggr}}$.

Results: speed and efficiency

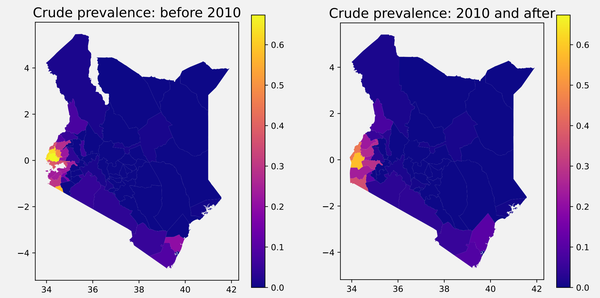
Comparison of MCMC for models with $f_{\text{GP-aggr}}$ and $f_{\text{VAE-aggr}}$ using 200 warm-up steps and 1000 iterations:

Model of the spatial random effect	Elapsed time	Average effective sample size of the random effect
GP-aggr	**15h***	129
VAE-aggr	**5s**	231

Table 1:Model comparison.

* GP-aggr model has not achieved full convergence after this time, i.e. R-hat of the length-scale is 1.42.

Results: maps produced by the two models



References

[1] Elizaveta Semenova, Yidan Xu, Adam Howes, Theo Rashid, Samir Bhatt, Swapnil Mishra, and Seth Flaxman. Priorvae: encoding spatial priors with variational autoencoders for small-area estimation. *Journal of the Royal Society Interface*, 19(191):20220094, 2022.

[2] Swapnil Mishra, Seth Flaxman, Tresnia Berah, Mikko Pakkanen, Harrison Zhu, and Samir Bhatt. pi vae: Encoding stochastic process priors with variational autoencoders. *Forthcoming in Statistics Computing 2022*, 2020.