Combining data from multiple spatially referenced surveys: geostatistical analysis of childhood malaria in Chikhwawa District, Malawi

Emanuele Giorgi [1] Sanie S. S. Sesay [2] Dianne J. Terlouw[2]
Peter J. Diggle[1]

[1]Medical School, Lancaster University, Lancaster, UK

[2]Liverpool School of Tropical Medicine, Liverpool, UK

Leahurst, University of Liverpool, 4-5 November 2013

Overview

- Motivation: quality and temporal variation across surveys.
- Data: two Malaria Indicator Surveys and an Easy Access Group study in Chikwawa District, Southern Malawi.
- Methods: a tri-variate generalised linear geostatistical model.
- Monte Carlo maximum likelihood.
- Results: estimation and prediction.
- Discussion.

Randomised and Convenience sampling

- Ideally every subject is drawn from the target population and is chosen randomly, such that each individual has the same probability of being chosen (**random sample**).
- **Convenience sampling** is a non-probability sampling technique where subjects are selected because of their convenient accessibility and proximity to the researcher.
- Convenience sampling is tempting in resource-poor settings because it is more economical but it can potentially produce bias estimates.

Temporal variation

- When surveys are **repeated over time**, it may be of interest to estimate changes over time.
- If surveys are correlated, there may be some in gain in the use of a joint model.
- A joint model is usually more advantageous if surveys do not use the same sampling locations.
- Surveys that are conducted further apart in time may not may provide two independent pieces of information.

The problem

1. How to combine data in a joint model?
2. How to account for bias from non-randomised surveys?
3. Is there any gain in spatial predictions from a joint model?

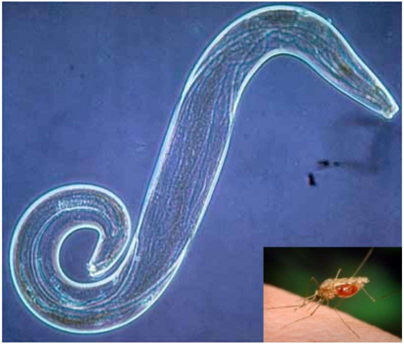
Some (non-spatial) answers from the literature: Moriarity and Schoren (2001); Elliot and Davis (2005); Lohr and Rao (2006); Manzi, Spiegelhalter, Turner, Flowers and Thompson (2011); Hedt and Pagano (2011).

Malaria and Malawi

Republic of Malawi

Capital: Lilongwe
Official language: English, Chichewa
Population: 14.901.000

Malaria and Malawi



Republic of Malawi

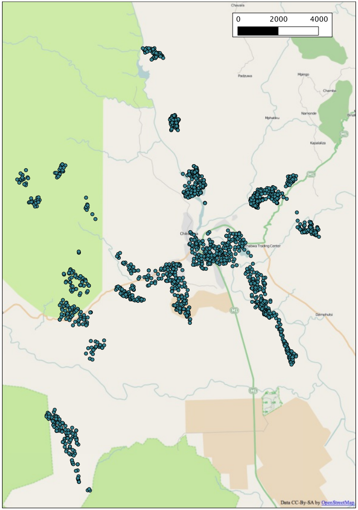
Capital: Lilongwe
Official language: English, Chichewa
Population: 14.901.000

Malaria and Malawi

Republic of Malawi

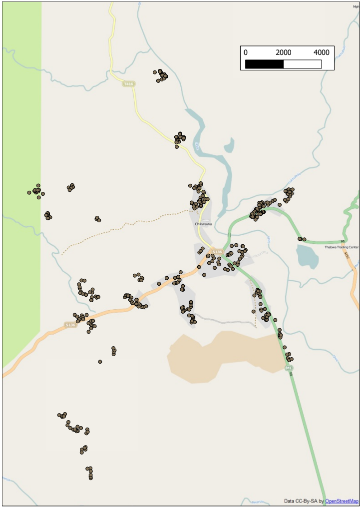
Capital: Lilongwe
Official language: English, Chichewa
Population: 14.901.000

Three surveys: 1st Malaria Indicator Survey (May 2010 - April 2011)



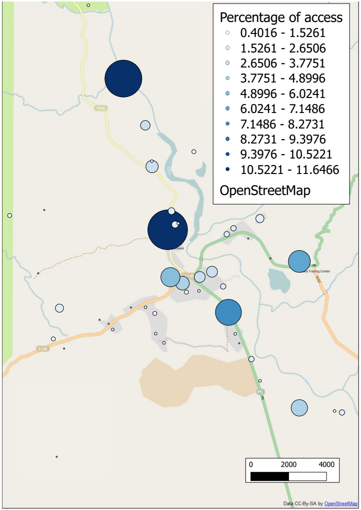
- Target population: children under 5 years.
- Sampling scheme: (first step) selection of 7 or 8 villages every month, and each village twice a year (rainy and dry season); (second step); random selection of households based on a household list.

Three surveys: 2nd Malaria Indicator Survey (May 2011 - April 2012)



- Target population: all groups of age.
- Sampling scheme: (first step) selection of 7 or 8 villages every month, and each village twice a year (rainy and dry season); (second step); random selection of households based on a spin-the-bottle approach.

Three surveys: Convenience survey (May 2011 - April 2012)



- Target population: children under 5 years.
- Sampling scheme: enrolment of children who came to the hospital for their childhood vaccines.

Model: a tri-variate GLGM (1)

$$Y_{ij} = \text{``number of positive RDTs for the } j\text{-th household in the } i\text{-th survey''}$$

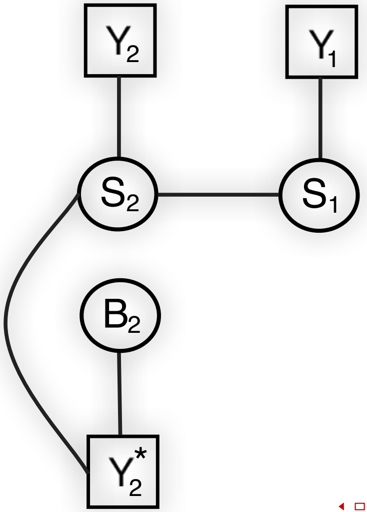
$$Y_{ij} \mid \text{random effect} \sim \text{Binomial}(n_{ij}, p_{ij})$$

Two main sources of heterogeneity across the three surveys:

- temporal variation of the underlying prevalence;
- quality-variation between randomised and convenience surveys.

Spatial covariates: Insecticide Treated Net (ITN), Indoor Residual Spraying (IRS), Rainy season (RS), Distance from closest waterway (DW), Socio-Economic-Status (SES).

Model: a tri-variate GLGM (2)



Model: a tri-variate GLGM (3)

- 1st Malaria Indicator Survey (May 2010 - April 2011)

$$\log\{p_{1j}/(1-p_{1j})\} = d(x_{1j})^\top \beta + S_1(x_{ij}) + Z_{1j}, j = 1, \ldots, 475.$$

- 2nd Malaria Indicator Survey (May 2011 - April 2012)

$$\begin{aligned} \log\{p_{2j}/(1-p_{2j})\} &= d(x_{2j})^\top \beta + S_2(x_{2j}) + Z_{2j}, j = 1, \ldots, 425. \\ \text{cor}(S_1(x), S_2(x)) &= \alpha \in (-1, 1). \end{aligned}$$

- Convenience survey (May 2011 - April 2012)

$$\begin{aligned} \log\{p_{3j}/(1-p_{3j})\} &= d(x_{3j})^\top \beta + S_2(x_{3j}) + \beta^* SES_{3j} + B(x_{3j}) + Z_{3j}, \\ & \quad j = 1, \ldots, 249. \end{aligned}$$

$S_i(x)$ and $B(x)$ are stationary isotropic RGPs with exponential correlation for $i = 1, 2$; Z_{ij} is Gaussian noise. We use Monte Carlo Maximum Likelihood to fit the model to the data.

Monte Carlo maximum likelihood

Let T and θ denote the vector of the random effects and model parameters, respectively, for given data y.

- **Likelihood function**

$$L(\theta) = f_Y(y;\theta) = \int f_T(t;\theta) f_{Y|T}(y|t)\, dt \qquad (1)$$

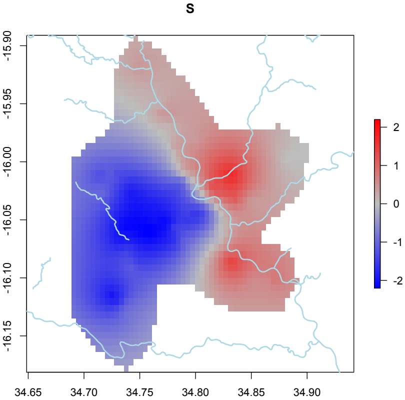
Let $\tilde{f}(y,t) = f_T(t,\theta_0) f_{Y|T}(y|t)$, then (1) can be expressed as

$$\begin{aligned} L(\theta) = \int \frac{f_T(t;\theta) f_{Y|T}(y|t)}{\tilde{f}(y,t)} \tilde{f}(y,t)\, dt \quad &\propto \quad \int \frac{f_T(t;\theta)}{f_T(t;\theta_0)} \tilde{f}_{T|Y}(t|y)\, dt \\ &= \quad \tilde{E}_{T|Y}\left[\frac{f_T(t;\theta)}{f_T(t;\theta_0)}\right] \end{aligned}$$

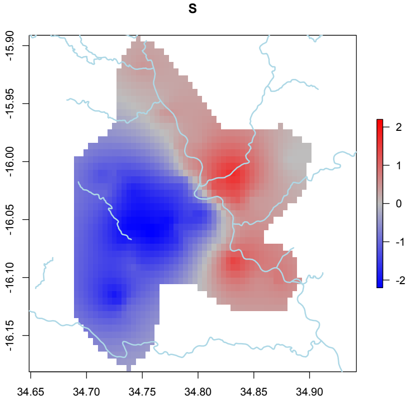
Parameter estimates

Term	Estimate	2.5%	97.5%
Intercept	0.130	-1.741	1.027
ITN	-0.840	-1.004	-0.680
IRS	-0.485	-0.666	-0.305
RS	0.550	0.348	0.761
DW	-0.282	-0.742	0.175
SES	-0.150	-0.238	-0.065
SES (spatial bias)	-0.160	-0.239	0.072
α	0.859	0.483	0.924

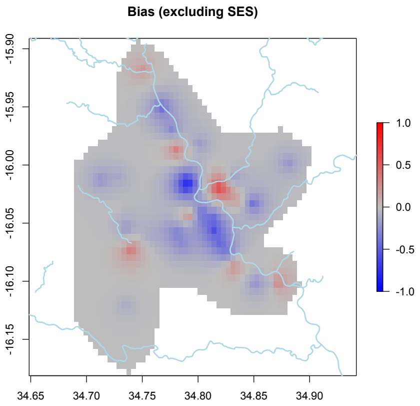
Prediction of $S_2(x)$



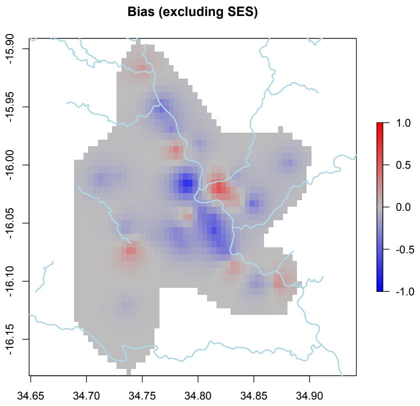
Prediction of $S_2(x)$

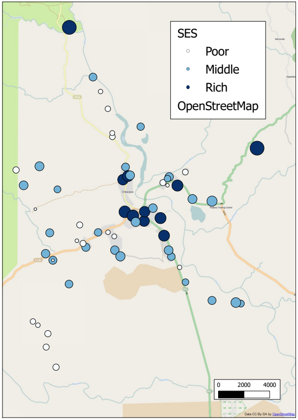


Prediction of $B(x)$

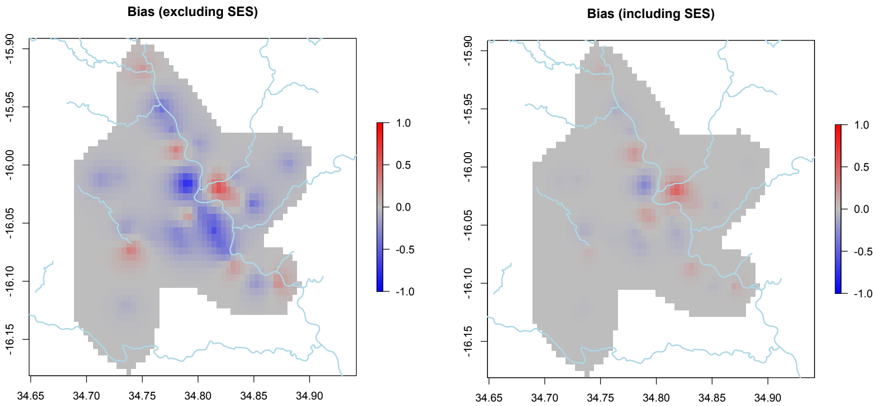


Prediction of $B(x)$

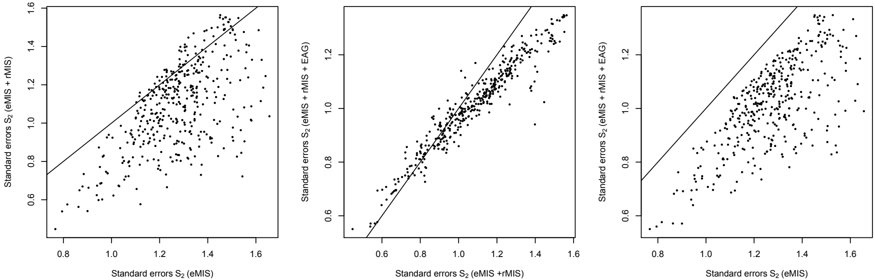




Prediction of $B(x)$



Prediction (3)



Further research

Potential applications in disease control are:

- development of computational procedures to inform improved prospective data collection for efficient hybrid sampling approaches;
- more accurate local spatio-temporal risk stratification maps that can inform more targeted control efforts.
- it is time to rethink about convenience sampling.
- how do we test the gold-standard assumption?
- the method is applicable under any general form of biased sampling but further validation is needed.

Thanks for the attention