Binomial geostatistical models

Peter Diggle & Emanuele Giorgi

Lancaster Medical School, Lancaster University, Lancaster, UK

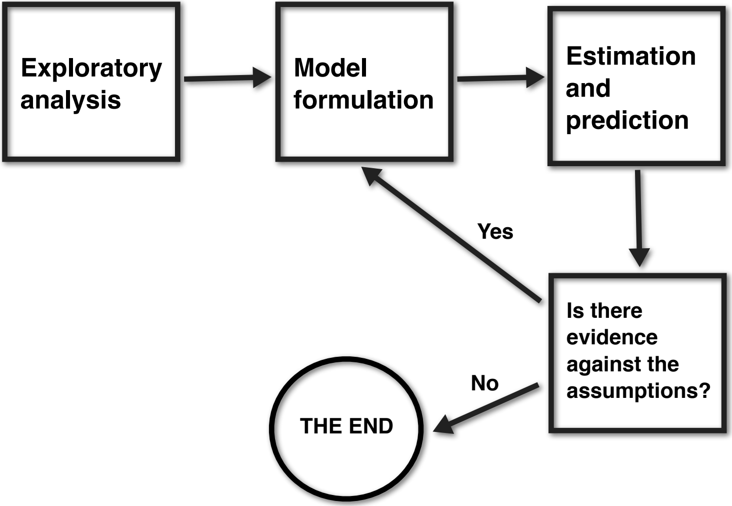
Lancaster University

Model-based geostatistics: geospatial statistical methods for public health applications, 5-9 October 2015

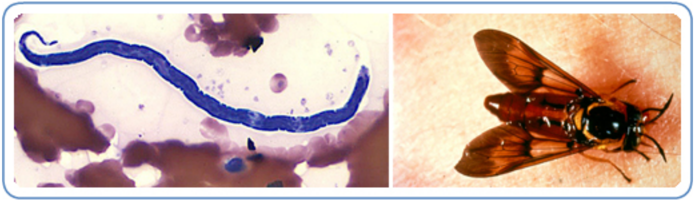
Overview

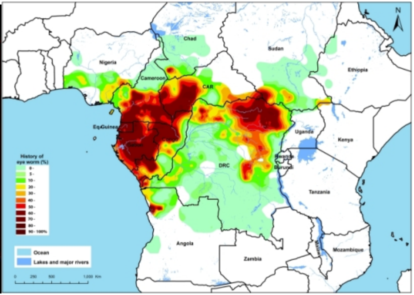
1. Exploratory analysis of prevalence data.
2. Linear geostatstical model based on logit-transformations of prevalence.
3. Binomial geostatistical models.
4. Parameter estimation: likelihood-based and Bayesian inference
5. Combining data from multiple surveys.

Simplified scheme of a statistical analysis

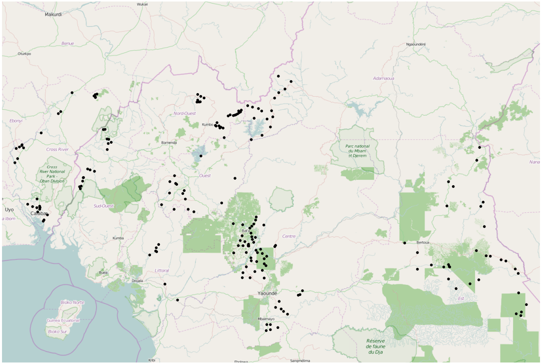


Loa loa




Chad
Sudan
Nigeria
Ethiopia
Cameroon
CAR
Eq.Guinea
Uganda
Kenya
Rwanda
Burundi
DRC
Tanzania
Angola
Zambia
Mozambique
History of eye worm (%)
0 -
5 -
10 -
20 -
30 -
40 -
50 -
60 -
70 -
80 -
90 - 100%
Ocean
Lakes and major rivers
0 250 500 1,000 Km
N

Loa loa in Cameroon and Nigeria



Epidemiological and geostistical questions

- What are the main risk factors of Loa loa?
- How do we identify Loa-loa hotspots?

A non-spatial model for prevalence

- Y_i =number of Loa loa cases in the i-village.
- m_i =number of examined people at location x_i.
- p_i =probability of being infected with Loa loa.

A non-spatial model for prevalence

- Y_i =number of Loa loa cases in the i-village.
- m_i =number of examined people at location x_i.
- p_i =probability of being infected with Loa loa.

What is a natural model for the data?

A non-spatial model for prevalence

- Y_i =number of Loa loa cases in the i-village.
- m_i =number of examined people at location x_i.
- p_i =probability of being infected with Loa loa.

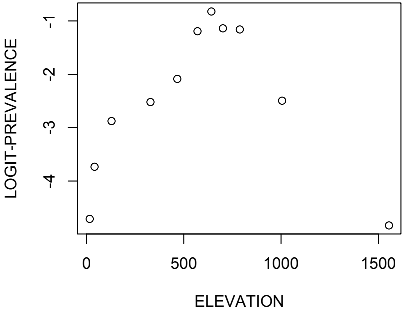
What is a natural model for the data?

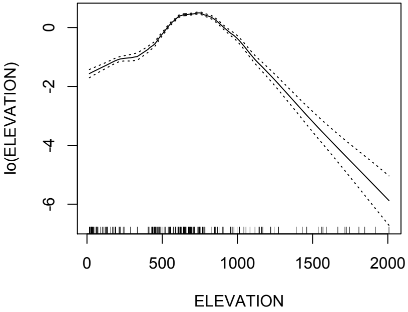
- $Y_i \sim$ Binomial(m_i, p_i)
- $\log\{p_i/(1-p_i)\} = d_i^\top \beta$

The empirical logit-transformation

$$\text{Empirical logit} = Z_i = \log\left\{\frac{Y_i + 0.5}{m_i - Y_i + 0.5}\right\}$$

Exploring the association between elevation and Loa loa risk





Non-spatial linear model

- $Z_i = \beta_0 + \beta_1 \text{elev}(x_i) + \beta_2 \text{elev}^2(x_i) + U_i$.
- $U_i \sim N(0, \tau^2)$ i.i.d. for all i.

Non-spatial linear model

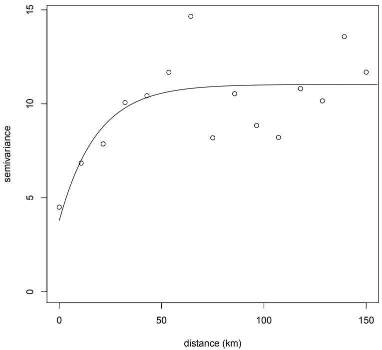
- $Z_i = \beta_0 + \beta_1 \text{elev}(x_i) + \beta_2 \text{elev}^2(x_i) + U_i$.
- $U_i \sim N(0, \tau^2)$ i.i.d. for all i.

Is there evidence against the assumptions?

Non-spatial linear model

- $Z_i = \beta_0 + \beta_1 \text{elev}(x_i) + \beta_2 \text{elev}^2(x_i) + U_i$.
- $U_i \sim N(0, \tau^2)$ i.i.d. for all i.

Is there evidence against the assumptions?



Non-spatial vs spatial linear model

- $Z_i = \beta_0 + \beta_1 \textsf{elev}(x_i) + \beta_2 \textsf{elev}^2(x_i) + S(x_i) + U_i$.
- $S(x) \sim \textsf{GP}(0, \sigma^2, \rho(\cdot;\phi))$.
- $\rho(u;\phi) = \exp\{-u/\phi\}$.
- $U_i \sim N(0, \tau^2)$ i.i.d. for all i.

Non-spatial vs spatial linear model

- $Z_i = \beta_0 + \beta_1 \text{elev}(x_i) + \beta_2 \text{elev}^2(x_i) + S(x_i) + U_i.$
- $S(x) \sim \text{GP}(0, \sigma^2, \rho(\cdot;\phi)).$
- $\rho(u;\phi) = \exp\{-u/\phi\}.$
- $U_i \sim N(0, \tau^2)$ i.i.d. for all i.

	Non-spatial LM		Spatial LM	
	Estimate	Std. Error	Estimate	StdErr
β_0	-3.689	0.199	-2.324	0.741
$\beta_1 \times 10^3$	5.963	0.539	1.362	1.395
$\beta_2 \times 10^6$	-4.066	0.311	-1.736	0.604
σ^2	-	-	1.867	0.279
ϕ	-	-	137.116	0.005
τ^2	1.163	0.786	0.403	1.605

Binomial geostatistical models

- $Y_i | S(x_i) \sim \text{Binomial}(m_i, p_i)$.
- $\log\{p_i/(1-p_i)\} = d_i^\top \beta + S(x_i) + U_i$
- $S(x) \sim \text{GP}(0, \sigma^2, \rho(\cdot; \phi))$.
- $\rho(u; \phi) = \exp\{-u/\phi\}$.
- $U_i \sim N(0, \tau^2)$ i.i.d. for all i.

Binomial geostatistical models

- $Y_i | S(x_i) \sim \text{Binomial}(m_i, p_i)$.
- $\log\{p_i/(1-p_i)\} = d_i^\top \beta + S(x_i) + U_i$
- $S(x) \sim \text{GP}(0, \sigma^2, \rho(\cdot; \phi))$.
- $\rho(u; \phi) = \exp\{-u/\phi\}$.
- $U_i \sim N(0, \tau^2)$ i.i.d. for all i.

Warning: the likelihood is not available in closed form.

$$L(\theta) = f(y; \theta) = \int_{\mathbb{R}^n} f(y, S; \theta)\, dS$$

A likelihood-based solution: the Monte Carlo maximum likelihood method

$$L(\theta) = \int_{\mathbb{R}^n} f(y, S; \theta)\, dS$$

A likelihood-based solution: the Monte Carlo maximum likelihood method

$$
\begin{aligned}
L(\theta) &= \int_{\mathbb{R}^n} f(y, S; \theta)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(y, S; \theta)}{f(y, S; \theta_0)} f(y, S; \theta_0)\, dS
\end{aligned}
$$

A likelihood-based solution: the Monte Carlo maximum likelihood method

$$
\begin{aligned}
L(\theta) &= \int_{\mathbb{R}^n} f(y, S; \theta)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(y, S; \theta)}{f(y, S; \theta_0)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta) f(y|S)}{f(S; \theta_0) f(y|S)} f(y, S; \theta_0)\, dS
\end{aligned}
$$

A likelihood-based solution: the Monte Carlo maximum likelihood method

$$
\begin{aligned}
L(\theta) &= \int_{\mathbb{R}^n} f(y, S; \theta)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(y, S; \theta)}{f(y, S; \theta_0)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta) f(y|S)}{f(S; \theta_0) f(y|S)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta)}{f(S; \theta_0)} f(y, S; \theta_0)\, dS
\end{aligned}
$$

A likelihood-based solution: the Monte Carlo maximum likelihood method

$$
\begin{aligned}
L(\theta) &= \int_{\mathbb{R}^n} f(y, S; \theta)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(y, S; \theta)}{f(y, S; \theta_0)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta) f(y|S)}{f(S; \theta_0) f(y|S)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta)}{f(S; \theta_0)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta)}{f(S; \theta_0)} f(y; \theta_0) f(S|y; \theta_0)\, dS
\end{aligned}
$$

$$
\begin{aligned}
L(\theta) &= \int_{\mathbb{R}^n} f(y, S; \theta)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(y, S; \theta)}{f(y, S; \theta_0)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta) f(y|S)}{f(S; \theta_0) f(y|S)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta)}{f(S; \theta_0)} f(y, S; \theta_0)\, dS \\
&= \int_{\mathbb{R}^n} \frac{f(S; \theta)}{f(S; \theta_0)} f(y; \theta_0) f(S|y; \theta_0)\, dS \\
&\propto \int_{\mathbb{R}^n} \frac{f(S; \theta)}{f(S; \theta_0)} f(S|y; \theta_0)\, dS = E_{S|y}\left[\frac{f(S; \theta)}{f(S; \theta_0)}\right]
\end{aligned}
$$

The MCML algorithm

1. Initialize θ_0.

The MCML algorithm

1. Initialize θ_0.
2. Simulate N samples, say $S_{(j)}$, from $S|y$ under θ_0 to approximate

$$L(\theta) = E_{S|y}\left[\frac{f(S;\theta)}{f(S;\theta_0)}\right] \approx L_N(\theta) = \frac{1}{N}\sum_{j=1}^{N}\frac{f(S_{(j)};\theta)}{f(S_{(j)};\theta_0)}.$$

The MCML algorithm

1. Initialize θ_0.
2. Simulate N samples, say $S_{(j)}$, from $S|y$ under θ_0 to approximate

$$L(\theta) = E_{S|y}\left[\frac{f(S;\theta)}{f(S;\theta_0)}\right] \approx L_N(\theta) = \frac{1}{N}\sum_{j=1}^{N}\frac{f(S_{(j)};\theta)}{f(S_{(j)};\theta_0)}.$$

3. Maximize $L_N(\theta)$ with respect to θ to obtain the MCML estimates, say $\hat{\theta}_N$.

The MCML algorithm

1. Initialize θ_0.
2. Simulate N samples, say $S_{(j)}$, from $S|y$ under θ_0 to approximate

$$L(\theta) = E_{S|y}\left[\frac{f(S;\theta)}{f(S;\theta_0)}\right] \approx L_N(\theta) = \frac{1}{N}\sum_{j=1}^{N}\frac{f(S_{(j)};\theta)}{f(S_{(j)};\theta_0)}.$$

3. Maximize $L_N(\theta)$ with respect to θ to obtain the MCML estimates, say $\hat{\theta}_N$.
4. Set $\theta_0 = \hat{\theta}_N$ and reiterate 1, 2 and 3 until convergence, e.g. until $L_N(\hat{\theta}_N) < 1$.

Bayesian inference

- $\theta^\top = (\beta, \sigma^2, \phi, \tau^2)$
- $\overbrace{f(\theta, S|y)}^{\text{posterior}} = \overbrace{f(\theta)}^{\text{prior}} f(S|\theta) f(y|S)$
- $f(\theta) = f(\beta) f(\sigma^2) f(\phi)\, f(\tau^2).$

Bayesian inference

- $\theta^\top = (\beta, \sigma^2, \phi, \tau^2)$
- $\overbrace{f(\theta, S|y)}^{\text{posterior}} = \overbrace{f(\theta)}^{\text{prior}} f(S|\theta) f(y|S)$
- $f(\theta) = f(\beta) f(\sigma^2) f(\phi)\ f(\tau^2)$.

Sampling from the posterior using Markov chain Monte Carlo algorithms

- In `PrevMap`, the three blocks (σ^2, ϕ, τ^2), β and S are updated separately.

Bayesian inference

- $\theta^\top = (\beta, \sigma^2, \phi, \tau^2)$
- $\overbrace{f(\theta, S|y)}^{\text{posterior}} = \overbrace{f(\theta)}^{\text{prior}} f(S|\theta) f(y|S)$
- $f(\theta) = f(\beta) f(\sigma^2) f(\phi)\ f(\tau^2)$.

Sampling from the posterior using Markov chain Monte Carlo algorithms

- In `PrevMap`, the three blocks (σ^2, ϕ, τ^2), β and S are updated separately.
  1. Random walk Metropolis Hastings for σ^2, ϕ and τ^2.
  2. Gibbs update for β.
  3. Hamiltonian Monte Carlo for S.

Bayesian inference

- $\theta^\top = (\beta, \sigma^2, \phi, \tau^2)$
- $\overbrace{f(\theta, S|y)}^{\text{posterior}} = \overbrace{f(\theta)}^{\text{prior}} f(S|\theta) f(y|S)$
- $f(\theta) = f(\beta) f(\sigma^2) f(\phi)\ f(\tau^2).$

Sampling from the posterior using Markov chain Monte Carlo algorithms

- In `PrevMap`, the three blocks (σ^2, ϕ, τ^2), β and S are updated separately.
  1. Random walk Metropolis Hastings for σ^2, ϕ and τ^2.
  2. Gibbs update for β.
  3. Hamiltonian Monte Carlo for S.
- Other packages: `geoRglm`, `spBayes`, `geostatsp`, `geoBayes`.

Non-spatial vs spatial binomial model

- **Priors specification:** $\beta \sim N(0, 10^3 I)$, $\log \sigma^2 \sim N(0.2, 0.3)$, $\log \phi \sim N(4, 0.4)$, $\log \tau^2 \sim N(-2.5, 1)$.

Non-spatial vs spatial binomial model

- **Priors specification:** $\beta \sim N(0, 10^3 I)$, $\log \sigma^2 \sim N(0.2, 0.3)$, $\log \phi \sim N(4, 0.4)$, $\log \tau^2 \sim N(-2.5, 1)$.

	Non-spatial		Spatial (MCML)		Spatial (Bayes)	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
β_0	-3.815	0.082	-3.093	0.543	-3.100	0.566
$\beta_1 \times 10^3$	0.789	0.024	0.439	0.134	0.445	0.133
$\beta_2 \times 10^6$	-5.719	0.186	-3.472	0.725	-3.537	0.725
σ^2	-	-	1.241	0.254	1.288	0.263
ϕ	-	-	60.767	0.007	64.925	17.666
τ^2	-	-	0.086	8.242	0.088	0.040
$\log \sigma^2$	-	-	0.216	0.315	0.233	0.201
$\log \phi$	-	-	4.107	0.423	4.137	0.268
$\log \tau^2$	-	-	-2.453	0.897	-2.542	0.493

Additional topics

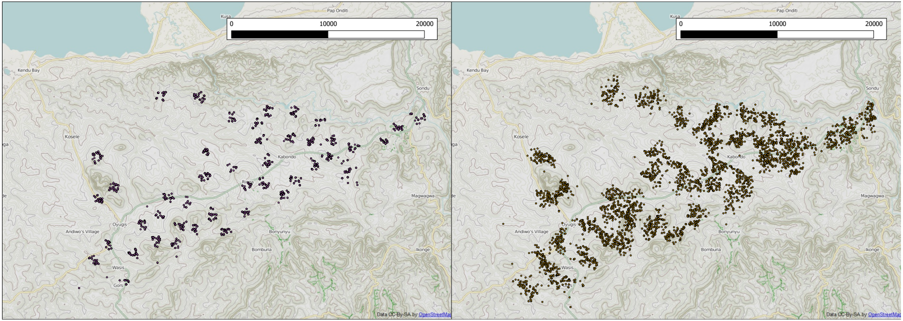
- **Combining data from multiple surveys.**
  Questions. How to combine data from biased convenience surveys with gold-standard prevalence surveys? What is the gain in doing so?

Additional topics

- **Combining data from multiple surveys.**
  Questions. How to combine data from biased convenience surveys with gold-standard prevalence surveys? What is the gain in doing so?

- **Spatially structured zero-inflation.**
  Question. Are zero reported cases a manifestation of binomial sampling error or a consequence of the community being disease-free?

Community and school surveys in Nyanza Province, Kenya



- **Community survey:** 1430 individuals; 740 compounds.
- **School survey:** 4852 pupils (46 schools); 3791 compounds.

A model for the data

- **Community survey.**

$$\log\{p_{ij}/(1 - p_{ij})\} = d_{ij}^{\top}\beta + S(x_i) + Z_i.$$

A model for the data

- **Community survey.**

$$\log\{p_{ij}/(1-p_{ij})\} = d_{ij}^{\top}\beta + S(x_i) + Z_i.$$

- **School survey.**

$$\log\{p_{ij}/(1-p_{ij})\} = d_{ij}^{\top}(\beta+\delta) + S(x_i) + B(x_i) + Z_i.$$

	Term
β_0	Intercept
β_1	Age in years
β_2	District (=1 if ``Rachuonyo''; =0 otherwise)
β_3	Socio-economic status (score from 1 to 5)
δ_0	Survey indicator, 1 if ``school,'' 0 if ``community'' (bias term)
δ_1	Age in years (bias term)

Results

	Estimate	95% Confidence interval
β_0	-1.412	(-2.303, -0.521)
β_1	-0.141	(-0.174, -0.109)
β_2	2.006	(1.228, 2.785)
β_3	-0.121	(-0.169, -0.072)
δ_0	-0.761	(-1.354, -0.167)
δ_1	0.094	(0.046, 0.142)

To be continued in the next lecture.

Bibliography

- Diggle, P. J., Giorgi, E. (2015). *Model-based geostatistics for prevalence mapping in low-resource settings.* Under review. Available at `http://arxiv.org/abs/1505.06891`
- Giorgi, E., Diggle, P. J. (2015). *PrevMap: and R package for prevalence mapping.* Under review.
- Giorgi, E., Sanie, S. S. S., Terloouw, D. J., Diggle, P. J. (2015). *Combining data from multiple spatially referenced prevalence surveys using generalized linear geostatistical models.* JRSS A, 178:445-464.