Spatio-temporal analysis

Dr Emanuele Giorgi
Lancaster University
e.giorgi@lancaster.ac.uk

Toowoomba 21-25 October 2019

Overview

Overview

- Repeated cross-sectional surveys.

Overview

- Repeated cross-sectional surveys.
- Analysing reported cases to health facilities.

Geostatistical problems and geostatistical methods

- **Geostatistical problems:** The object of scientific interest is to make inference an spatially continuous surface (e.g. disease risk) or some of its properties (e.g. regional prevalence) in a geographical area of interest.
- **Model-based geostatistics:** Likelihood-based methods of inference that use spatially sparse noisy data to make inference on a spatially continuous surface of interest.

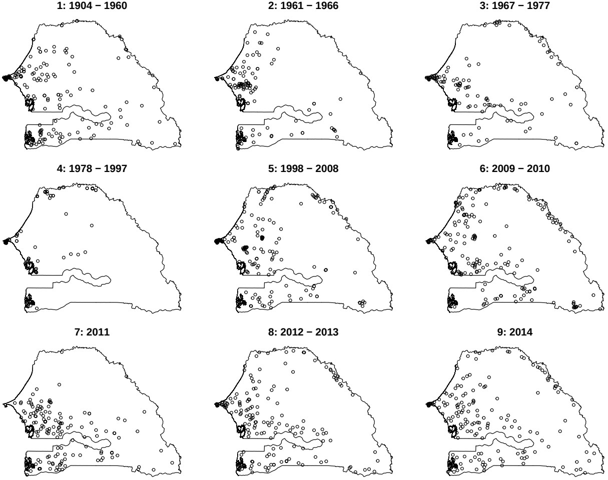
Geostatistical problems and geostatistical methods

- **Geostatistical problems:** The object of scientific interest is to make inference an spatially continuous surface (e.g. disease risk) or some of its properties (e.g. regional prevalence) in a geographical area of interest.
- **Model-based geostatistics:** Likelihood-based methods of inference that use spatially sparse noisy data to make inference on a spatially continuous surface of interest.
- Geostatistical problems are **NOT** dependent on the data-format (e.g. point-level or areal-level data).

Geostatistical problems: some examples

- Spatio-temporal interpolation of disease risk.
- Identification of disease hotspots (i.e. areas where disease risk exceeds policy relevant thresholds).
- Identification of low transmission areas (i.e. areas where disease risk is below policy relevant thresholds).
- Estimating the association between risk factors and disease risk metrics.
- Accounting for the excess of zero-reported cases (spatially structured zero-inflation).
- Combining data from randomised survey and convenience samples.

The data


1: 1904 – 1960
2: 1961 – 1966
3: 1967 – 1977
4: 1978 – 1997
5: 1998 – 2008
6: 2009 – 2010
7: 2011
8: 2012 – 2013
9: 2014

The standard model for prevalence

- $\mathcal{D} = \{(x_i, t_i, y_i, n_i) : x_i \in A, i = 1, \ldots, N^* = \sum_{i=1}^{N} n_i\}$

The standard model for prevalence

- $\mathcal{D} = \{(x_i, t_i, y_i, n_i) : x_i \in A, i = 1, \ldots, N^* = \sum_{i=1}^{N} n_i\}$
- $d(x, t) =$explanatory variables.

The standard model for prevalence

- $\mathcal{D} = \{(x_i, t_i, y_i, n_i) : x_i \in A, i = 1, \ldots, N^* = \sum_{i=1}^{N} n_i\}$
- $d(x, t) =$explanatory variables.
- $S(x, t) =$unexplained spatial variation.

The standard model for prevalence

- $\mathcal{D} = \{(x_i, t_i, y_i, n_i) : x_i \in A, i = 1, \ldots, N^* = \sum_{i=1}^{N} n_i\}$
- $d(x, t) =$explanatory variables.
- $S(x, t) =$unexplained spatial variation.
- $Z(x, t) =$unexplained unstructured variation.

The standard model for prevalence

- $\mathcal{D} = \{(x_i, t_i, y_i, n_i) : x_i \in A, i = 1, \ldots, N^* = \sum_{i=1}^{N} n_i\}$
- $d(x, t) =$explanatory variables.
- $S(x, t) =$unexplained spatial variation.
- $Z(x, t) =$unexplained unstructured variation.
- A model for the data.

$$\log\left\{\frac{p(x_i, t_i)}{1 - p(x_i, t_i)}\right\} = d(x_i, t_i)^\top \beta + S(x_i, t_i) + Z(x_i, t_i). \quad (1)$$

The standard model for prevalence

- $\mathcal{D} = \{(x_i, t_i, y_i, n_i) : x_i \in A, i = 1, \ldots, N^* = \sum_{i=1}^{N} n_i\}$
- $d(x, t)$ =explanatory variables.
- $S(x, t)$ =unexplained spatial variation.
- $Z(x, t)$ =unexplained unstructured variation.
- A model for the data.

$$\log\left\{\frac{p(x_i, t_i)}{1 - p(x_i, t_i)}\right\} = d(x_i, t_i)^\top \beta + S(x_i, t_i) + Z(x_i, t_i). \quad (1)$$

- $E[S(x, t)] = E[Z(x, t)] = 0$; $V[S(x, t)] = \sigma^2$; $V[Z(x, t)] = \tau^2$ and

$$\mathrm{corr}(S(x, t), S(x', t')) = \rho(x, x', t, t'; \theta).$$

The standard model for prevalence

- $\mathcal{D} = \{(x_i, t_i, y_i, n_i) : x_i \in A, i = 1, \ldots, N^* = \sum_{i=1}^{N} n_i\}$
- $d(x, t) =$explanatory variables.
- $S(x, t) =$unexplained spatial variation.
- $Z(x, t) =$unexplained unstructured variation.
- A model for the data.

$$\log\left\{\frac{p(x_i, t_i)}{1 - p(x_i, t_i)}\right\} = d(x_i, t_i)^\top \beta + S(x_i, t_i) + Z(x_i, t_i). \qquad (1)$$

- $E[S(x, t)] = E[Z(x, t)] = 0$; $V[S(x, t)] = \sigma^2$; $V[Z(x, t)] = \tau^2$ and

$$\text{corr}(S(x, t), S(x', t')) = \rho(x, x', t, t'; \theta).$$

- **Question.** What can we do with this model?

Exploring the residual variation

- $W(x, t) = S(x, t) + Z(x, t)$

Exploring the residual variation

- $W(x, t) = S(x, t) + Z(x, t)$
- $u = \|x - x'\|$, $v = |t - t'|$.

Exploring the residual variation

- $W(x,t) = S(x,t) + Z(x,t)$
- $u = \|x - x'\|$, $v = |t - t'|$.
- The theoretical STV

$$\begin{aligned}\gamma(u,v;\theta) &= E[\{W(x,t) - W(x',t')\}^2] \\ &= \tau^2 + \sigma^2[1 - \rho(u,v;\theta)]. \qquad (2)\end{aligned}$$

- The empirical STV

$$\tilde{\gamma}(u,v) = \frac{1}{2|N(u,v)|} \sum_{(i,j)\in N(u,v)} \{\tilde{Z}(x_i,t_i) - \tilde{Z}(x_j,t_j)\}^2. \qquad (3)$$

A test for spatio-temporal correlation

- $\tilde{Z}(x, t) =$point estimate from a non-spatial binomial mixed model.

A test for spatio-temporal correlation

- $\tilde{Z}(x, t)$ =point estimate from a non-spatial binomial mixed model.

- The algorithm.
  - (*i*) permute the order of the data, including $\tilde{Z}(x_i, t_i)$, while holding (x_i, t_i) fixed;

A test for spatio-temporal correlation

- $\tilde{Z}(x, t) =$point estimate from a non-spatial binomial mixed model.

- The algorithm.
  (*i*) permute the order of the data, including $\tilde{Z}(x_i, t_i)$, while holding (x_i, t_i) fixed;

  (*ii*) compute the empirical semi-variogram for $\tilde{Z}(x_i, t_i)$;

A test for spatio-temporal correlation

- $\tilde{Z}(x, t)$ =point estimate from a non-spatial binomial mixed model.

- The algorithm.
  - (*i*) permute the order of the data, including $\tilde{Z}(x_i, t_i)$, while holding (x_i, t_i) fixed;
  - (*ii*) compute the empirical semi-variogram for $\tilde{Z}(x_i, t_i)$;
  - (*iii*) repeat(*i*) and (*ii*) a large enough number of times, say B;

A test for spatio-temporal correlation

- $\tilde{Z}(x, t) =$point estimate from a non-spatial binomial mixed model.

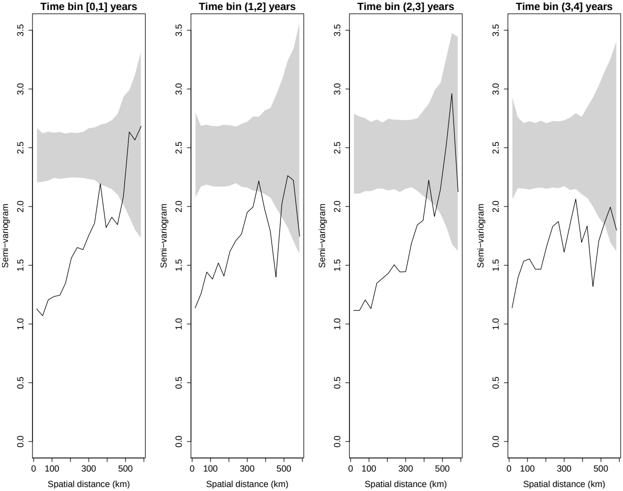
- The algorithm.
  (*i*) permute the order of the data, including $\tilde{Z}(x_i, t_i)$, while holding (x_i, t_i) fixed;

  (*ii*) compute the empirical semi-variogram for $\tilde{Z}(x_i, t_i)$;

  (*iii*) repeat(*i*) and (*ii*) a large enough number of times, say B;

  (*iv*) use the resulting B empirical semi-variograms to generate 95% tolerance intervals at each of the pre-defined distance bins.

A test for spatio-temporal correlation



The covariance model

- The Gneiting's spatio-temporal correlation function.

$$\rho(u, v; \theta) = \frac{1}{(1 + v/\psi)^{\delta+1}} \exp\left\{-\frac{u/\phi}{(1 + v/\psi)^{\xi/2}}\right\}. \qquad (4)$$

In the application $\delta = \xi = 0$.

The covariance model

- The Gneiting's spatio-temporal correlation function.

$$\rho(u, v; \theta) = \frac{1}{(1 + v/\psi)^{\delta+1}} \exp\left\{-\frac{u/\phi}{(1 + v/\psi)^{\xi/2}}\right\}. \qquad (4)$$

  In the application $\delta = \xi = 0$.
- How to choose a covariance model?

The covariance model

- The Gneiting's spatio-temporal correlation function.

$$\rho(u, v; \theta) = \frac{1}{(1 + v/\psi)^{\delta+1}} \exp\left\{-\frac{u/\phi}{(1 + v/\psi)^{\xi/2}}\right\}. \qquad (4)$$

  In the application $\delta = \xi = 0$.
- How to choose a covariance model?
- Guiding principle: **parsimony**, e.g. find the simplest model that can explain the variation of the data.

Parameter estimation

- $[\cdot]$= "the distribution of $\cdot$".

Parameter estimation

- $[\cdot]$= "the distribution of $\cdot$".
- $\lambda^\top = (\beta^\top, \sigma^2, \theta^\top)$.

Parameter estimation

- $[\cdot]$="the distribution of $\cdot$".
- $\lambda^\top = (\beta^\top, \sigma^2, \theta^\top)$.
- The likelihood function.

$$L(\lambda) = [y|\lambda] = \int [W, y|\lambda] \, dW.$$

- Bayesian inference.

$$[\lambda|y] = \frac{[\lambda][y|\lambda]}{[y]} = \frac{[\lambda]L(\lambda)}{[y]}.$$

Parameter estimation

- $[\cdot]$="the distribution of $\cdot$".
- $\lambda^\top = (\beta^\top, \sigma^2, \theta^\top)$.
- The likelihood function.

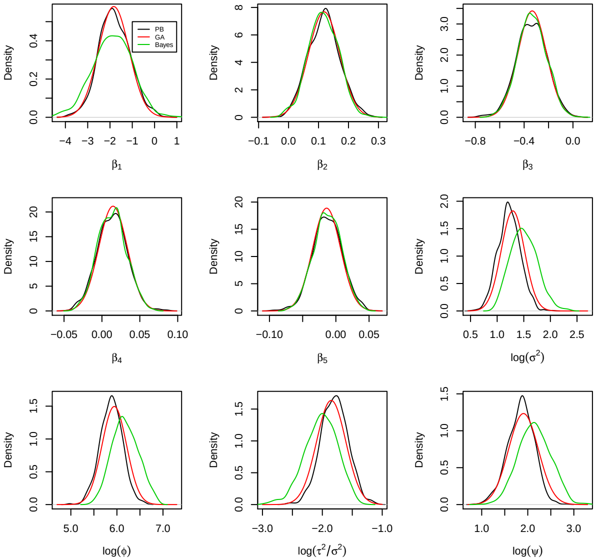
$$L(\lambda) = [y|\lambda] = \int [W, y|\lambda] \, dW.$$

- Bayesian inference.

$$[\lambda|y] = \frac{[\lambda][y|\lambda]}{[y]} = \frac{[\lambda]L(\lambda)}{[y]}.$$

- Prior distributions.
  - $\beta \sim N(0, 10^4)$
  - $\sigma^2 \sim \text{Uniform}(0, 20)$
  - $\phi \sim \text{Uniform}(0, 1000)$
  - $\tau^2/\sigma^2 \sim \text{Uniform}(0, 20)$
  - $\psi \sim \text{Uniform}(0, 20)$

Estimation: Likelihood-based vs Bayesian inference



Validation of the covariance model

- Replace (i) with

  (a) Simulate $W(x_i, t_i)$ at observed locations x_i and times t_i, for $i = 1, \ldots, n$, from its marginal multivariate distribution under the assumed model.

  (b) Conditionally on the simulated values of $W(x_i, t_i)$, simulate binomial data y_i from (1).

  (c) Compute a point estimates $\tilde{Z}(x_i, t_i)$ using the simulated data.

Validation of the covariance model

- Replace (i) with

  (a) Simulate $W(x_i, t_i)$ at observed locations x_i and times t_i, for $i = 1, \ldots, n$, from its marginal multivariate distribution under the assumed model.

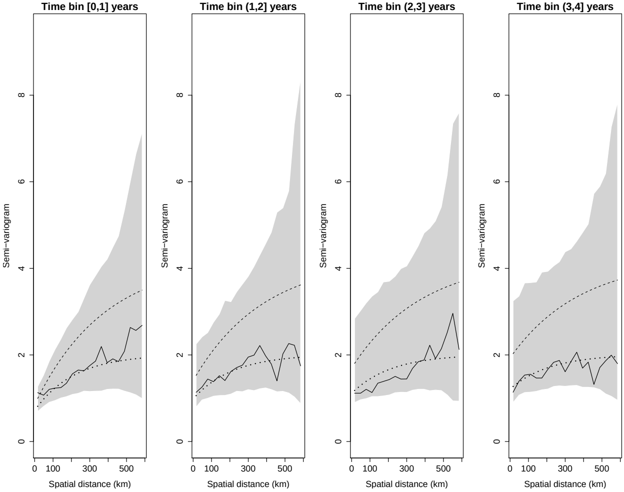
  (b) Conditionally on the simulated values of $W(x_i, t_i)$, simulate binomial data y_i from (1).

  (c) Compute a point estimates $\tilde{Z}(x_i, t_i)$ using the simulated data.

- Test statistics.

$$T = \sum_{k=1}^{K} |N(u_k, t_k)| [\tilde{\gamma}(u_k, v_k) - \gamma(u_k, v_k; \theta)]^2. \tag{5}$$

Validation of the covariance model



Spatio-temporal prediction

- $W^* = \{W(x,t) : x \in \mathcal{X}, t \in \mathcal{T}\}$.

Spatio-temporal prediction

- $W^* = \{W(x,t) : x \in \mathcal{X}, t \in \mathcal{T}\}$.
- Likelihood-based inference (plug-in).

$$[W^*|y,\lambda] = \int [W^*, W|y,\lambda]\, dW$$

Spatio-temporal prediction

- $W^* = \{W(x,t) : x \in \mathcal{X}, t \in \mathcal{T}\}$.
- Likelihood-based inference (plug-in).

$$[W^*|y,\lambda] = \int [W^*, W|y,\lambda] \, dW$$

- Bayesian inference.

$$[W^*|y] = \int \int [\lambda|y][W|y,\lambda][W^*|W,\lambda] \, dW \, d\lambda,$$

Spatio-temporal prediction

- $W^* = \{W(x, t) : x \in \mathcal{X}, t \in \mathcal{T}\}$.
- Likelihood-based inference (plug-in).

$$[W^*|y, \lambda] = \int [W^*, W|y, \lambda]\, dW$$

- Bayesian inference.

$$[W^*|y] = \int \int [\lambda|y][W|y, \lambda][W^*|W, \lambda]\, dW\, d\lambda,$$

- Likelihood-based inference (accounting for parameter uncertainty).

$$[W^*|y] = \int \int [\hat{\Lambda}|y][W|y, \hat{\Lambda}][W^*|W, \hat{\Lambda}]\, dW\, d\hat{\Lambda},$$

Targets of prediction

Targets of prediction

- Exceedance probability surface.

$$\mathcal{R}_l(W^*) = \{r(x,t) = P(p(x,t) > l|y)\}. \quad (6)$$

Targets of prediction

- Exceedance probability surface.

$$\mathcal{R}_l(W^*) = \{r(x,t) = P(p(x,t) > l|y)\}. \tag{6}$$

- District-wide average prevalence.

$$p_t(D) = \frac{1}{|D|}\int_D p(x,t)\,dx, \tag{7}$$

Targets of prediction

- Exceedance probability surface.

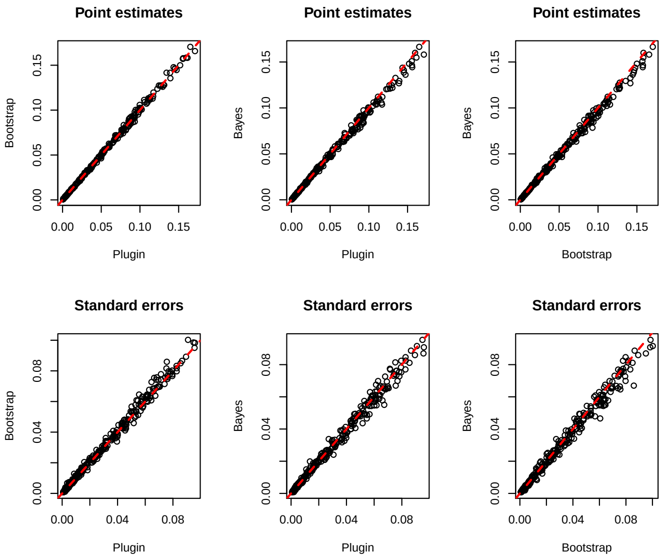
$$\mathcal{R}_l(W^*) = \{r(x,t) = P(p(x,t) > l|y)\}. \tag{6}$$

- District-wide average prevalence.

$$p_t(D) = \frac{1}{|D|}\int_D p(x,t)\,dx, \tag{7}$$

`http://fhm-chicas-apps.lancs.ac.uk/shiny/users/giorgi/mapMalariaSEN/`

Prediction: Likelihood-based vs Bayesian inference



What is a disease outbreak?

- **WHO:** *"A disease outbreak is the occurrence of cases of disease in excess of what* **would normally be expected** *in a defined community, geographical area or season. An outbreak may occur in a restricted geographical area, or may extend over several countries. It may last for a few days or weeks, or for several years."*

What is a disease outbreak?

- **WHO:** *"A disease outbreak is the occurrence of cases of disease in excess of what* **would normally be expected** *in a defined community, geographical area or season. An outbreak may occur in a restricted geographical area, or may extend over several countries. It may last for a few days or weeks, or for several years."*

- A disease outbreak is the occurrence of cases of disease in excess **of a policy-relevant risk threshold** in a defined community, geographical area or season.

Data

- **Outcome:** Weekly reports of confirmed cases from 142 districts in Mozambique from the first week of 2010 through the ninth week of 2017.

Data

- **Outcome:** Weekly reports of confirmed cases from 142 districts in Mozambique from the first week of 2010 through the ninth week of 2017.
- **Population:** WorldPop + 2007 Census Data.

Data

- **Outcome:** Weekly reports of confirmed cases from 142 districts in Mozambique from the first week of 2010 through the ninth week of 2017.
- **Population:** WorldPop + 2007 Census Data.
- **Climatic variables:** FLDAS data are available from NASA at daily temporal resolution and 0.1 degree spatial resolution.
  - total rainfall (mm)
  - average weekly temperature ($^\circ$C)
  - average weekly relative humidity (%)
  - average weekly surface barometric pressure (hPa)
  - average weekly saturation vapour pressure deficit (mmHg)

Model formulation

- i=district; t=week

Model formulation

- i=district; t=week
- Y_{it}=random variable of the number of reported cases

Model formulation

- i=district; t=week
- Y_{it}=random variable of the number of reported cases
- d_{it}=vector of explanatory (climatic) variables

Model formulation

- i=district; t=week
- Y_{it}=random variable of the number of reported cases
- d_{it}=vector of explanatory (climatic) variables
- m_{it}=total number of under 5 children

Model formulation

- i=district; t=week
- Y_{it}=random variable of the number of reported cases
- d_{it}=vector of explanatory (climatic) variables
- m_{it}=total number of under 5 children
- $Y_{it} | S_{it} \sim Pois(m_{it}\lambda_{it})$, such that

$$\log\{\lambda_{it}\} = d_{it}^\top \beta + S_{it}$$

Model formulation

- i=district; t=week
- Y_{it}=random variable of the number of reported cases
- d_{it}=vector of explanatory (climatic) variables
- m_{it}=total number of under 5 children
- $Y_{it}|S_{it} \sim Pois(m_{it}\lambda_{it})$, such that

$$\log\{\lambda_{it}\} = d_{it}^{\top}\beta + S_{it}$$

Objective

Spatio-temporal mapping of exceedance probabilities

$$\mathrm{Prob}\{\exp\{S_{it}\} > l|\text{data}\}$$

at a future week t.

Modelling the residual spatio-temporal variation

$$S_{it} = \frac{1}{|\mathcal{D}_i|} \int_{\mathcal{D}_i} S(x, t) \, dx$$

Modelling the residual spatio-temporal variation

$$S_{it} = \frac{1}{|\mathcal{D}_i|} \int_{\mathcal{D}_i} S(x, t) \, dx$$

- Spatial covariance structure

$$\mathrm{cov}\{S_{it}, S_{jt}\} = \frac{\sigma^2}{|\mathcal{D}_i||\mathcal{D}_j|} \int_{\mathcal{D}_i} \int_{\mathcal{D}_j} \exp\{-\|x - x'\|/\phi\} \, dx dx'$$

Modelling the residual spatio-temporal variation

$$S_{it} = \frac{1}{|\mathcal{D}_i|} \int_{\mathcal{D}_i} S(x, t)\, dx$$

- Spatial covariance structure

$$\mathrm{cov}\{S_{it}, S_{jt}\} = \frac{\sigma^2}{|\mathcal{D}_i||\mathcal{D}_j|} \int_{\mathcal{D}_i} \int_{\mathcal{D}_j} \exp\{-\|x - x'\|/\phi\}\, dx dx'$$

- $S_t = (S_{1,t}, \ldots, S_{142,t})$.

Modelling the residual spatio-temporal variation

$$S_{it} = \frac{1}{|\mathcal{D}_i|} \int_{\mathcal{D}_i} S(x,t)\, dx$$

- Spatial covariance structure

$$\mathrm{cov}\{S_{it}, S_{jt}\} = \frac{\sigma^2}{|\mathcal{D}_i||\mathcal{D}_j|} \int_{\mathcal{D}_i} \int_{\mathcal{D}_j} \exp\{-\|x - x'\|/\phi\}\, dx dx'$$

- $S_t = (S_{1,t}, \ldots, S_{142,t})$.
- Temporal covariance structure

$$S_t = \rho S_{t-1} + W_t.$$

Models comparison and validation

Models comparison and validation

- $\mathcal{M}_1 : \log\{\lambda_{it}\} = d_{it}^\top \beta$

Models comparison and validation

- $\mathcal{M}_1 : \log\{\lambda_{it}\} = d_{it}^\top \beta$
- $\mathcal{M}_2 : \log\{\lambda_{it}\} = \alpha + S_{it}$

Models comparison and validation

- $\mathcal{M}_1 : \log\{\lambda_{it}\} = d_{it}^\top \beta$
- $\mathcal{M}_2 : \log\{\lambda_{it}\} = \alpha + S_{it}$
- $\mathcal{M}_3 : \log\{\lambda_{it}\} = d_{it}^\top \beta + S_{it}$

Models comparison and validation

- $\mathcal{M}_1 : \log\{\lambda_{it}\} = d_{it}^\top \beta$
- $\mathcal{M}_2 : \log\{\lambda_{it}\} = \alpha + S_{it}$
- $\mathcal{M}_3 : \log\{\lambda_{it}\} = d_{it}^\top \beta + S_{it}$
- Root-mean-square-error (RMSE)

$$\mathrm{RMSE}_t = \sqrt{\frac{1}{142}\sum_{i=1}^{142}(\lambda_{it}^{emp} - \hat{\lambda}_{it})^2}$$

Models comparison and validation

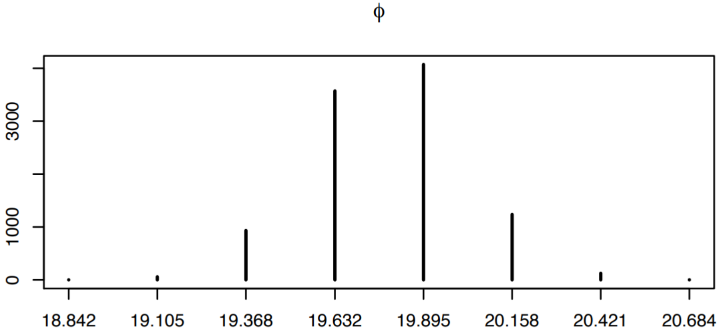
- $\mathcal{M}_1 : \log\{\lambda_{it}\} = d_{it}^\top \beta$
- $\mathcal{M}_2 : \log\{\lambda_{it}\} = \alpha + S_{it}$
- $\mathcal{M}_3 : \log\{\lambda_{it}\} = d_{it}^\top \beta + S_{it}$
- Root-mean-square-error (RMSE)

$$\mathrm{RMSE}_t = \sqrt{\frac{1}{142}\sum_{i=1}^{142}(\lambda_{it}^{emp} - \hat{\lambda}_{it})^2}$$

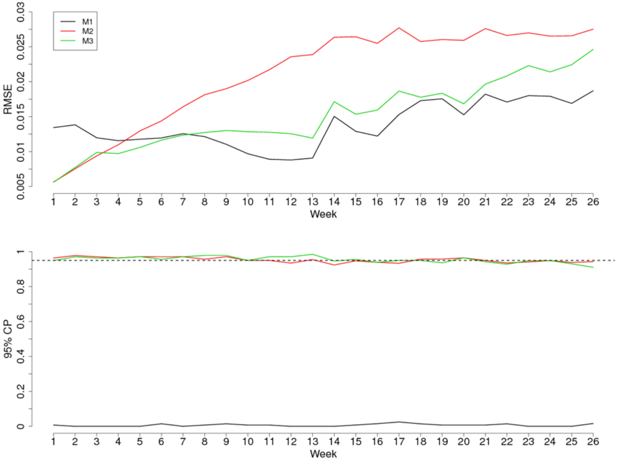
- 95% coverage probabilities (CI)

$$\mathrm{CI}_t = \frac{1}{142}\sum_{i=1}^{142} I(\lambda_{it}^{0.025} < \lambda_{it}^{emp} < \lambda_{it}^{0.975})$$

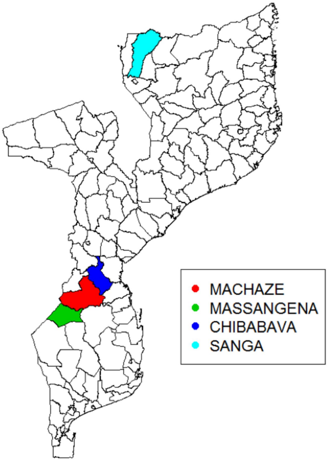
Results: scale of spatial correlation



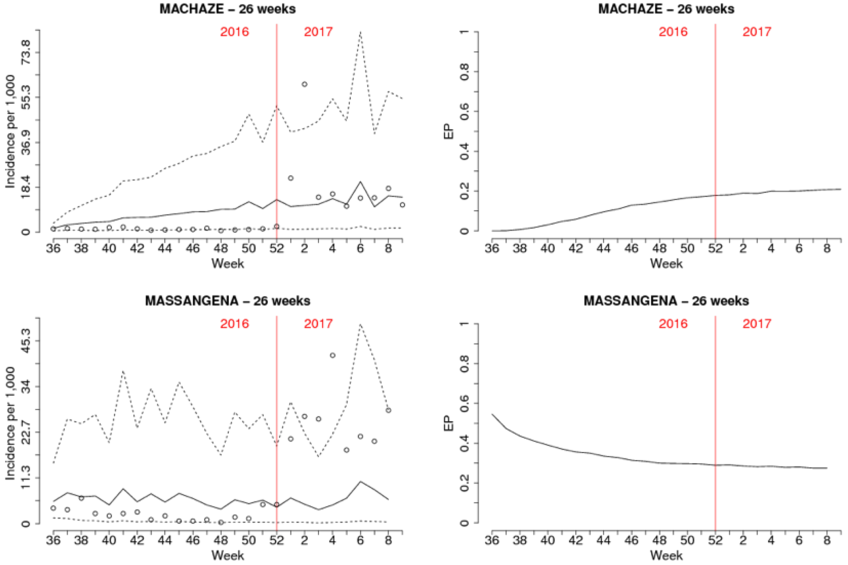
Results: validation



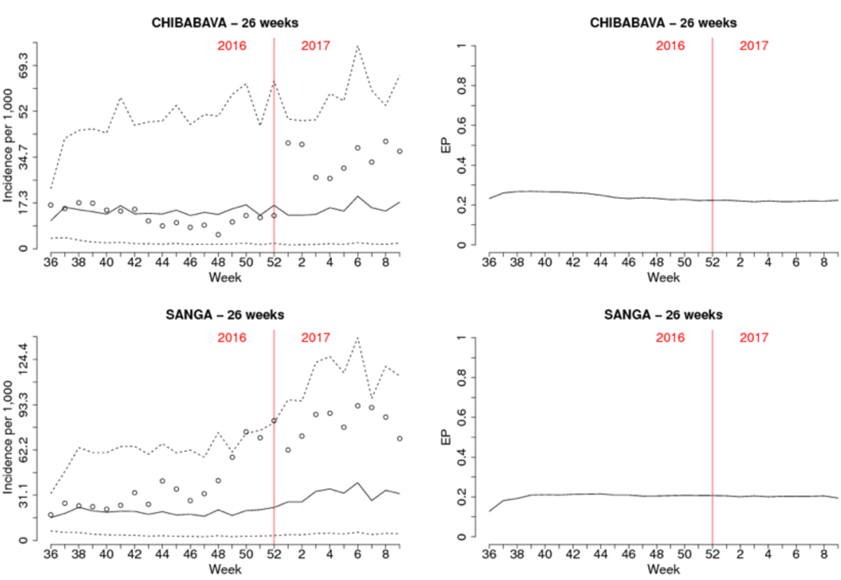
Results: exceedance probabilities



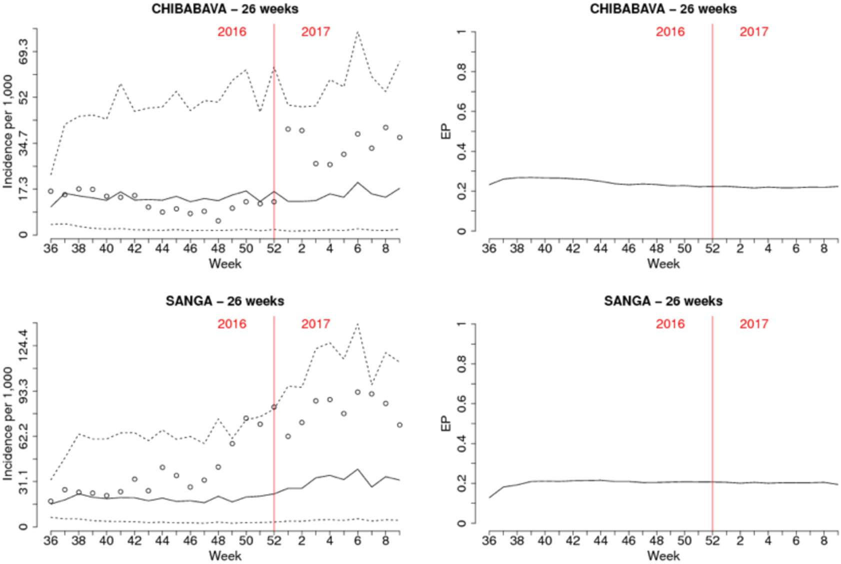
Results: exceedance probabilities



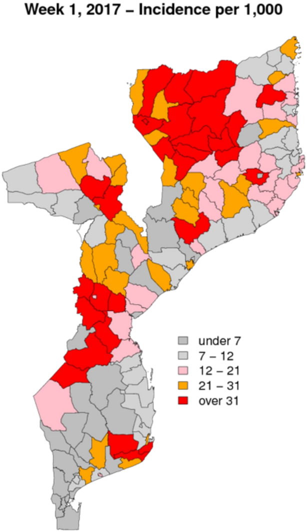
Results: exceedance probabilities

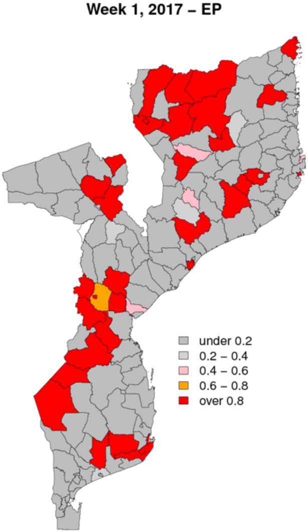


Results: exceedance probabilities

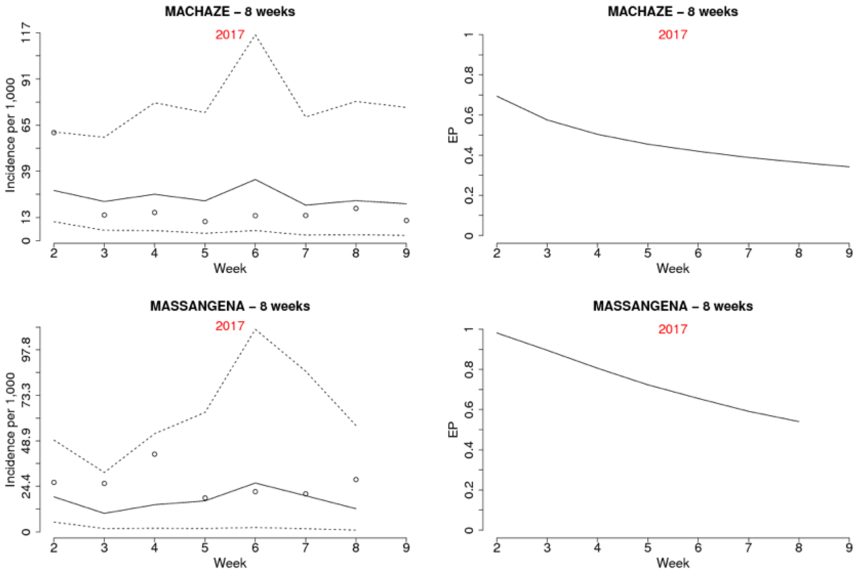


Results: incidence vs exceedance probabilities

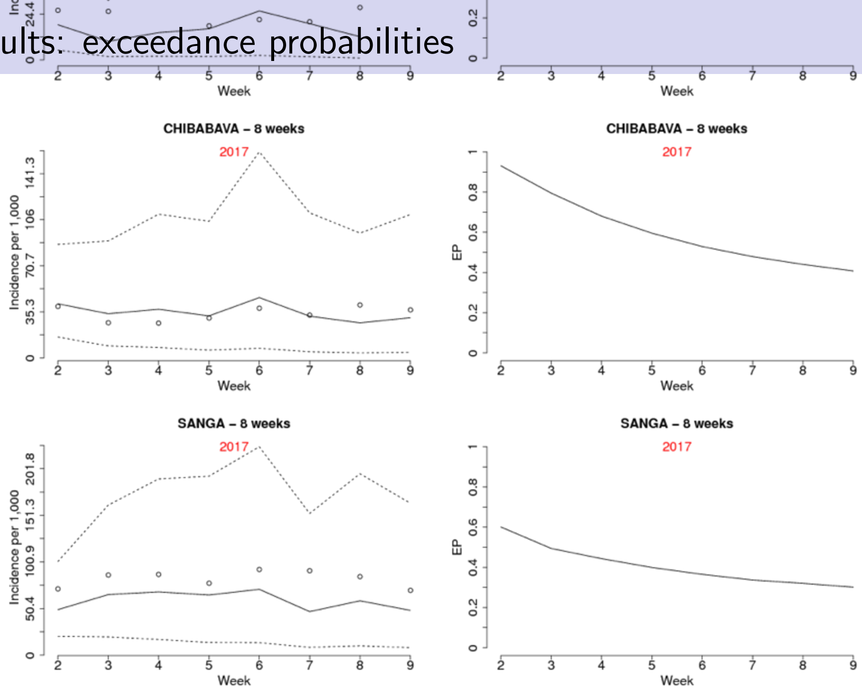




Results: exceedance probabilities



Results: exceedance probabilities



Conclusions

Conclusions

- We need to account for unmeasured risk factors (S_{it}).

Conclusions

- We need to account for unmeasured risk factors (S_{it}).
- Climatic variables provide information on the long term trends; residual spatio-temporal random effects on short term.

Conclusions

- We need to account for unmeasured risk factors (S_{it}).
- Climatic variables provide information on the long term trends; residual spatio-temporal random effects on short term.
- Exceedance probabilities are a useful tool for detecting outbreaks within at least 4 weeks.

Conclusions

- We need to account for unmeasured risk factors (S_{it}).
- Climatic variables provide information on the long term trends; residual spatio-temporal random effects on short term.
- Exceedance probabilities are a useful tool for detecting outbreaks within at least 4 weeks.
- Outstanding issue: how to identify relative risk thresholds.

Conclusions

- We need to account for unmeasured risk factors (S_{it}).
- Climatic variables provide information on the long term trends; residual spatio-temporal random effects on short term.
- Exceedance probabilities are a useful tool for detecting outbreaks within at least 4 weeks.
- Outstanding issue: how to identify relative risk thresholds.
- Web interface: `http://www.lancaster.ac.uk/staff/haleac/malaria/`(by Dr Alison Hale)

Conclusions

- We need to account for unmeasured risk factors (S_{it}).
- Climatic variables provide information on the long term trends; residual spatio-temporal random effects on short term.
- Exceedance probabilities are a useful tool for detecting outbreaks within at least 4 weeks.
- Outstanding issue: how to identify relative risk thresholds.
- Web interface: `http://www.lancaster.ac.uk/staff/haleac/malaria/`(by Dr Alison Hale)