Modeling dependent survival data through random effects with spatial correlation at the subject level

Application to malaria data analysis

Estelle Kuhn

INRAE, MaIAGE

Joint work with A.Oodally (INRAE, MaIAGE), Klara Goethals, Luc Duchateau (Faculty of Veterinary Medicine, Ghent University)

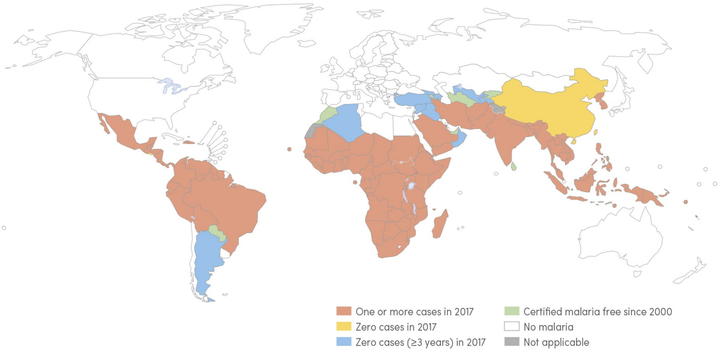
Some informations on malaria disease

- eliminated in the 1950s in America and in the 1970s in Europe

- 219 million of malaria cases worldwide in 2017
- 435,000 deaths in 2017
- 61% are children under the age of 5
- no significant progress towards a decrease in the number of malaria cases worldwide between 2015-2017

Malaria in the world

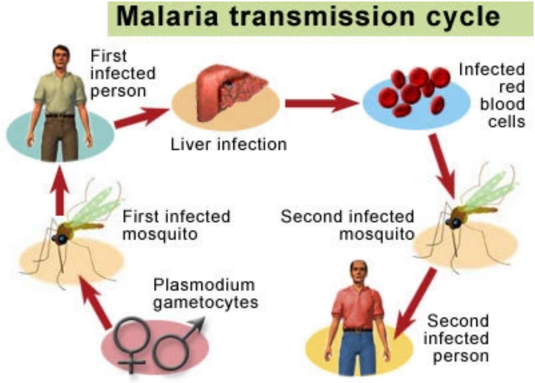
Countries with indigenous cases in 2000 and their status by 2017 Countries with zero indigenous cases over at least the past 3 consecutive years are considered to be malaria free. All countries in the WHO European Region reported zero indigenous cases in 2016 and again in 2017. In 2017, both China and El Salvador reported zero indigenous cases. *Source: WHO database.*



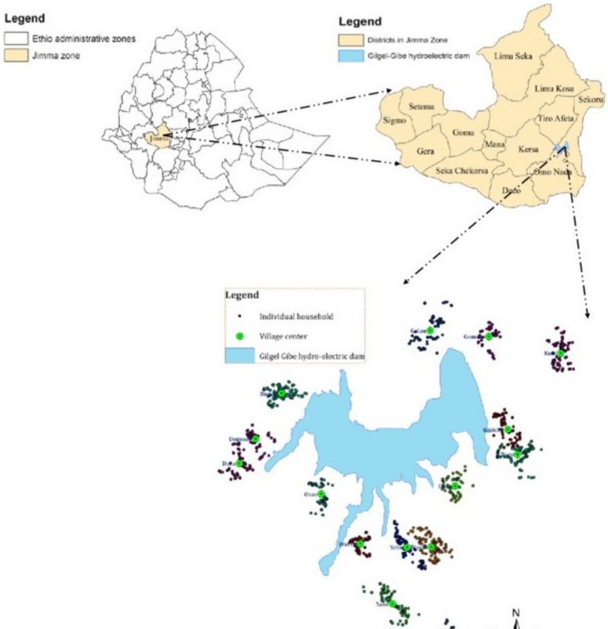
One or more cases in 2017
Zero cases in 2017
Zero cases (≥3 years) in 2017
Certified malaria free since 2000
No malaria
Not applicable

WHO: World Health Organization.

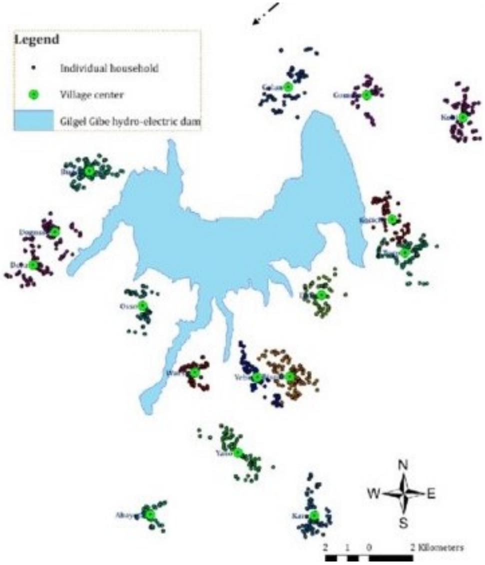
Propagation of malaria



Focus on the Jimma zone



16 villages in Gilgel Gibe dam region



Objectives

⟹ quantify the effect of the dam on malaria propagation

⟹ propose a model taking into account distance to dam and distance between individuals

Outline:

1. Context of survival analysis
2. General frailty model
3. Spatially correlated frailty model
4. Estimation procedure
5. Numerical experiments
6. Perspectives

Context of survival analysis

Consider an event of interest in a population of individuals

Denote by T_i the time to event of individual i

Examples :

- time to infection
- time to death
- time to flowering date
- time till recovering a job after unemployment
- ...

Survival function, hazard, Cox model (Cox (1972))

Survival function of individual i: $S_i(t) = P(T_i \geq t)$.

Hazard of individual i:

$$\lambda_i(t) = \lim_{dt \to 0^+} \frac{P(t \leq T_i < t + dt | T_i \geq t)}{dt}$$

Thus

$$S_i(t) = \exp\left(-\int_0^t \lambda_i(s) ds\right)$$

Cox model:

$$\lambda_i(t) = \lambda_0(t) \exp(X_i' \beta)$$

with λ_0 unknown baseline,
X_i covariates vector of individual i,
β unknown parameters vector of interest.

Proportional hazards assumption in Cox model

$S_i(t) = \exp\left(-\int_0^t \lambda_i(s)ds\right)$ and $\lambda_i(t) = \lambda_0(t)\exp(X_i'\beta)$

leads to $\log(-\log S_i(t)) = X_i'\beta + \log\left(\int_0^t \lambda_0(s)ds\right)$

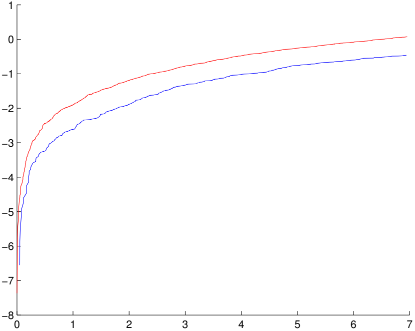


Figure: Plot of $t \to \log(-\log(S(t))$ in red for group 1 and in blue ofr group 2.

Frailty model (Vaupel et al. (1979))

Idea : model heterogeneity in population through random effects

Examples :

- clinical study in several centers
- crop on several parcels and several environmental conditions
- ...

Denote by i the indice of the group
and by j the indice of the individual,

Model the hazard by: $\lambda_{ij}(t|b_i) = \lambda_0(t)\exp(X'_{ij}\beta + b_i)$

with X_{ij} covariates of individual j of group i,
λ_0 unknown baseline function,
β unknown parameters vector
b_i random effect of group i.

Spatially correlated univariate frailty model

$\Longrightarrow$ introduce spatial correlation in the frailty term at subject level

Model the hazard as follows:

$$\lambda_i(t|b_i) = \lambda_0(t) \;\; \exp(X_i^t \beta + b_i) \qquad (\mathcal{M}_1)$$

where λ_0 is the baseline hazard function,
X_i covariates vector
β the vector of the unknown regression parameters,
b_i the frailty term of subject i

and model the frailty vector $\mathbf{b} = (b_i)_{1 \leq i \leq N}$ as follows:

$$\mathbf{b} \sim \mathcal{N}(0, \sigma^2 \Sigma(\rho)).$$

with σ^2 scaling factor
and $\Sigma(\rho)$ correlation matrix parameterized by $\rho > 0$

Different correlation structures and baseline function

$\Longrightarrow$ consider two usual different correlation structures following Li and Ryan (2002)

$$\Sigma_{\text{exp}}(\rho) = \exp(-\rho D)$$

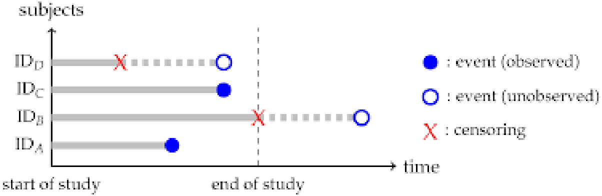
$$\Sigma_{\text{pol}}(\rho) = \frac{1}{1 + D^{\rho}}$$

with $D = (d_{ii'}) \in \mathcal{M}_N(\mathbb{R}^+)$
and $d_{ii'}$ the distance between subject i and subject $i^{'}$.

$\Longrightarrow$ consider usual parametric baseline hazard function parametrized by α

$\Longrightarrow$ Model parameters are $\theta = (\alpha, \beta, \sigma^2, \rho)$.

Censoring in survival analysis



- T_i time to event
- C_i censoring time

$\Longrightarrow (T_i)$ and (C_i) non observed

Available observations:

- $Y_i = T_i \wedge C_i$ censored observation
- $\Delta_i = \mathbb{1}_{T_i \leq C_i}$ censoring indicator

Maximum marginal likelihood estimation

$\theta = (\alpha, \beta, \sigma^2, \rho)$

Complete likelihood expression:

$$L_{\text{comp}}(\theta; \mathbf{Y}, \boldsymbol{\Delta}, \mathbf{b}) = \prod_{i=1}^{N} \left(\frac{(\lambda_0(Y_i) \exp(X_i^t \beta + b_i))^{\Delta_i}}{\exp(\Lambda_0(Y_i) \exp(X_i^t \beta + b_i))} \right) f_{\sigma^2 \boldsymbol{\Sigma}(\rho)}(\mathbf{b})$$

where $\Lambda_0(Y_i) = \int_0^{Y_i} \lambda_0(t)dt$ is the cumulative hazard function

Marginal likelihood expression:

$$L_{\text{marg}}(\theta; \mathbf{Y}, \boldsymbol{\Delta}) = \int L_{\text{comp}}(\theta; \mathbf{Y}, \boldsymbol{\Delta}, \mathbf{b}) d\mathbf{b}$$

Maximum marginal likelihood estimate:

$$\hat{\theta} = \text{argmax } L_{\text{marg}}(\theta; \mathbf{Y}, \boldsymbol{\Delta}).$$

Estimation in frailty models

- approximated likelihood criteria
  - penalized likelihood (McGilchrist et al. (1991))
  - partial likelihood (Nielsen et al. (1992))
  - partial penalized likelihood (Therneau et al. (2000))
  - complete penalized likelihood (Rondeau et al. (2003))
- bayesian (Ducrocq et Casella (1996))
- exact likelihood
  - EM algorithm
  - Monte Carlo EM (*prop.* Wei et al. (1990), *frailty.* Ripatti et al. (2002)) *theory* Fort et Moulines (2003))
    $\Longrightarrow$ long computation times
  - EM-Laplace (Abrahantes et Burzykowski (2005)
    $\Longrightarrow$ no convergence property

EM algorithm (Dempster et al. (1977))

$\Longrightarrow$ deal with estimation in latent variable model

Iteration k:

Step E : compute

$$Q(\theta|\theta_k) = \mathrm{E}(\log L_{\text{comp}}(\mathbf{Y}, \boldsymbol{\Delta}, \mathbf{b}; \theta)|\mathbf{Y}, \boldsymbol{\Delta}, \theta_k)$$

Step M : update

$$\theta_{k+1} = \arg\max Q(\theta|\theta_k)\,.$$

$\Longrightarrow$ convergence toward a critical point of marginal likelihood

$\Longrightarrow$ drawback: step E may be intractable

Stochastic approximation EM with MCMC method

(Delyon et al. (1999), Kuhn et al. (2004),
Allassonnière et al. (2010))

Iteration k:

Simulation step : $\mathbf{b}^{k+1} \sim \Pi_{\theta_k}(\mathbf{b}^k, \cdot)$ with Π_θ transition kernel of ergodic Markov chain having as stationnary distribution the posterior distribution $\pi_\theta(\mathbf{b}|\mathbf{Y}, \mathbf{\Delta})$.

Stochastic approximation step :

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k \left(\log L_{\text{comp}}(\mathbf{Y}, \mathbf{\Delta}, \mathbf{b}^{k+1}; \theta) - Q_k(\theta) \right),$$

with $(\gamma_k)_k$ positive step sizes s.t. $\sum \gamma_k = \infty$,
$\sum \gamma_k^2 < \infty$

Update step :

$$\theta_{k+1} = \arg\max Q_{k+1}(\theta)$$

$\Longrightarrow$ a.s. convergence toward a critical point of marginal likelihood

Some heuristic

$$Q_{k+1}(\theta) = Q_k(\theta) + \gamma_k \left[\log L_{\text{comp}}(\mathbf{Y}, \boldsymbol{\Delta}, \mathbf{b}^{k+1}; \theta) - Q_k(\theta) \right]$$

$$\frac{Q_{k+1}(\theta) - Q_k(\theta)}{\gamma_k} = \{ E[\log L_{\text{comp}}(\mathbf{Y}, \boldsymbol{\Delta}, \mathbf{b}; \theta) | \mathbf{Y}, \boldsymbol{\Delta}; \theta] - Q_k(\theta) \}$$
$$+ \left\{ \log L_{\text{comp}}(\mathbf{Y}, \boldsymbol{\Delta}, \mathbf{b}^{k+1}; \theta) - E[\log L_{\text{comp}}(\mathbf{Y}, \boldsymbol{\Delta}, \mathbf{b}; \theta) | \mathbf{Y}, \boldsymbol{\Delta}; \theta] \right\}$$

$$\frac{Q_{k+1}(\theta) - Q_k(\theta)}{\gamma_k} \approx \{ E[\log L_{\text{comp}}(\mathbf{Y}, \boldsymbol{\Delta}, \mathbf{b}; \theta) | \mathbf{Y}, \boldsymbol{\Delta}; \theta] - Q_k(\theta) \} + e_k$$

with e_k little centered perturbation.

Simulation study

$\Longrightarrow$ mimic the malaria data (Getachew et al. (2013))

Model ($\mathcal{M}_1$)

$$\lambda_i(t|b_i) = \lambda_0(t)\ \exp(X_i^t\beta + b_i) \ \text{ and } \ \mathbf{b} \sim \mathcal{N}(0, \sigma^2 \exp(-\rho D))$$

- $N = 300$ subjects
- D is chosen by taking subsets of size 300 of the real malaria distance matrix.
- piecewise constant baseline $\sum_{m=1}^{3} \lambda_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t)$ with $(\tau_0, \tau_1, \tau_2, \tau_3) = (0, 0.2, 0.8, +\infty)$, $(\lambda_1, \lambda_2, \lambda_3) = (2, 0.5, 1)$
- $X_i \overset{iid}{\sim} \mathcal{B}(0.5)$
- $\beta = (2, 3)$
- $(\sigma^2, \rho) = (1.5, 1)$
- 3 different censoring settings: no censoring, moderate censoring (40%) and heavy censoring (60%)

$\Longrightarrow$ use random scan Gibbs sampler to face the high dimension of frailty vector

Effect of censoring rate

Table: Mean of the parameter estimates and empirical standard error in parentheses estimated in model $(\mathcal{M}_1)$ from 100 repetitions with data generated under model $(\mathcal{M}_1)$. The number of subjects N is fixed at 300.

Parameters	True	No censoring	40 % censoring	60 % censoring
h_3	1	0.957 (0.447)	1.089 (0.611)	1.209 (0.884)
β_2	3	2.969 (0.210)	3.010 (0.254)	3.061 (0.340)
σ^2	1.5	1.554 (0.444)	1.642 (0.463)	1.654 (0.552)
ρ	1	0.977 (0.277)	1.051 (0.318)	1.072 (0.322)

Robustness to misspecification of the correlation structure

$\Longrightarrow$ evaluate effects of misspecification with respect to correlation structure

Let introduce model $(\mathcal{M}_2)$ defined by:

$$\lambda_i(t|b_i) = \sum_{m=1}^{3} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t) \exp(X_i^t \beta + b_i)$$

$$\mathbf{b} \sim \mathcal{N}(0, \sigma^2 \Sigma_{\text{pol}}(\rho))$$

with

$$\Sigma_{\text{pol}}(\rho) = \frac{1}{1 + D^\rho}$$

Robustness to misspecification of correlation structure

Table: Mean of the parameter estimates and empirical standard error in parentheses estimated in model ($\mathcal{M}_2$) from 100 repetitions with data generated under model ($\mathcal{M}_1$). The number of subjects N is fixed at 300.

Parameters	True	No censoring	40 % censoring	60 % censoring
h_1	2	2.276 (1.600)	2.298 (1.642)	2.556 (2.223)
β_2	3	3.098 (0.223)	3.045 (0.276)	3.086 (0.333)
σ^2	1.5	1.932 (0.495)	1.805 (0.566)	1.946 (0.590)
ρ	1	0.817 (0.164)	0.748 (0.168)	0.648 (0.167)

Comparison with other models without spatial correlation structure

Consider the proportional hazards model ($\mathcal{M}_3$):

$$\lambda_i(t|b_i) = \sum_{m=1}^{3} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t) \exp(X_i^t \beta) \qquad (\mathcal{M}_3)$$

and the univariate frailty model ($\mathcal{M}_4$):

$$\lambda_i(t|b_i) = \sum_{m=1}^{3} h_m \mathbb{1}_{[\tau_{m-1},\tau_m[}(t) \exp(X_i^t \beta + b_i) \qquad (\mathcal{M}_4)$$

$$b_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

Results of comparisons with other models

Table: Mean of the parameter estimates and empirical standard error in parentheses estimated in model ($\mathcal{M}_3$) and model ($\mathcal{M}_4$) from 100 repetitions with data generated under model ($\mathcal{M}_1$). The number of subjects N is fixed at 300.

Parameters	True	Prop. hazards model	Univ. frailty model
h_1	2	2.583 (0.721)	2.172 (0.688)
h_2	0.5	0.351 (0.128)	0.455 (0.194)
h_3	1	0.298 (0.115)	0.757 (0.342)
β_1	2	1.555 (0.210)	1.874 (0.250)
β_2	3	2.299 (0.269)	2.835 (0.294)
σ^2	1.5	×	0.988 (0.270)

Analysing the Gilgel Gibe time to malaria data set

Oodally et al. (2020)

- 2037 children
- 16 different villages
- 4 covariates: distance to the dam, sex, structure of the roof of the household, age $(3-7, 7<)$

Results

- correlation structure Σ_{pol} chosen by comparing AIC
- reject null hypothesis H_0:$\rho = \infty$ using likelihood ratio test
- no significant effect for distance to dam
- significant effect for children older than 7 years
  higher malaria risk of 42%

Comparisons with other models

Hazard rate estimates based on univariate spatially correlated frailty model with correlation structure $\Sigma_{\text{pol}}(\rho)$

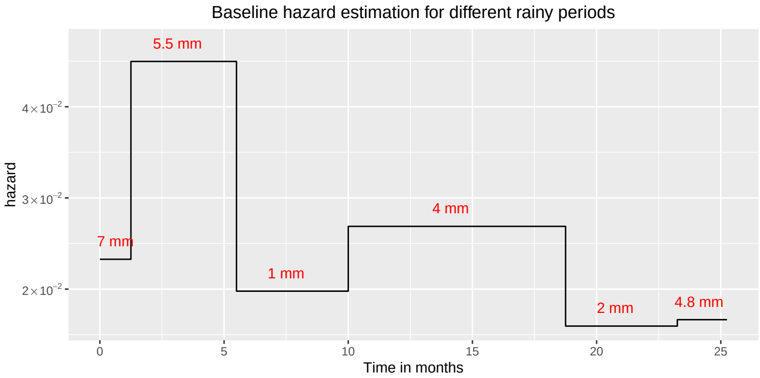


Figure: Hazard rates estimates. Average daily rainfall within different time periods annotated in red.

Plot of the correlation as a function of the distance

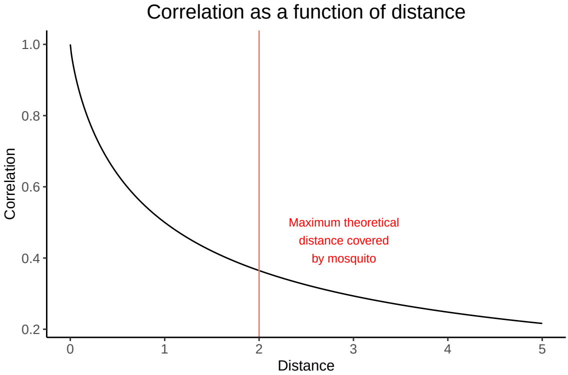


Figure: Correlation values $\Sigma_{\text{pol}}(\hat{\rho})$ as a function of distance.

Conclusion and perspectives

Conclusion:

- spatially correlated univariate frailty model
- convergent estimation algorithm
- analysis of malaria data using a model taking into account distance to dam and distance between individual without confounding effect

Perspective:

- more complex correlation structure
- partial likelihood criteria

Bibliographie

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society 34*, 187—-220.

Getachew, Y., P. Janssen, D. Yewhalaw, N. Speybroeck, and L. Duchateau (2013). Coping with time and space in modelling malaria incidence: a comparison of survival and count regression models. *Statistics in medicine 32*(18), 3224–3233.

Li, Y. and L. Ryan (2002). Modeling spatial survival data using semiparametric frailty models. *Biometrics 58*(2), 287–297.

Oodally, A., E. Kuhn, K. Goethals, and L. Duchateau (2020). Modeling dependent survival data through random effects with spatial correlation at the subject level. *Arxiv*.

Vaupel, J., K. Manton, and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*, 439—-454.