Bayesian (Hierarchical) Spatial modeling in epidemiology

Thomas N O Achia[1]

[1]School of Mathematics, Statistics and Computer Science
University of KwaZulu Natal

Strathmore University, 2012

Outline

Outline

Motivating example

Kitweru	Rongo
Kanga	Kodero Bara

Region	Observed (O_i)	Total (N_i)
Rongo	600	10000
Kitweru	600	1000
Kodero Bara	600	6000
Kanga	600	8000
	2400	25000

$$\hat{\theta} = \frac{\sum_i O_i}{\sum_i N_i} = \frac{2400}{25000} = 0.096$$

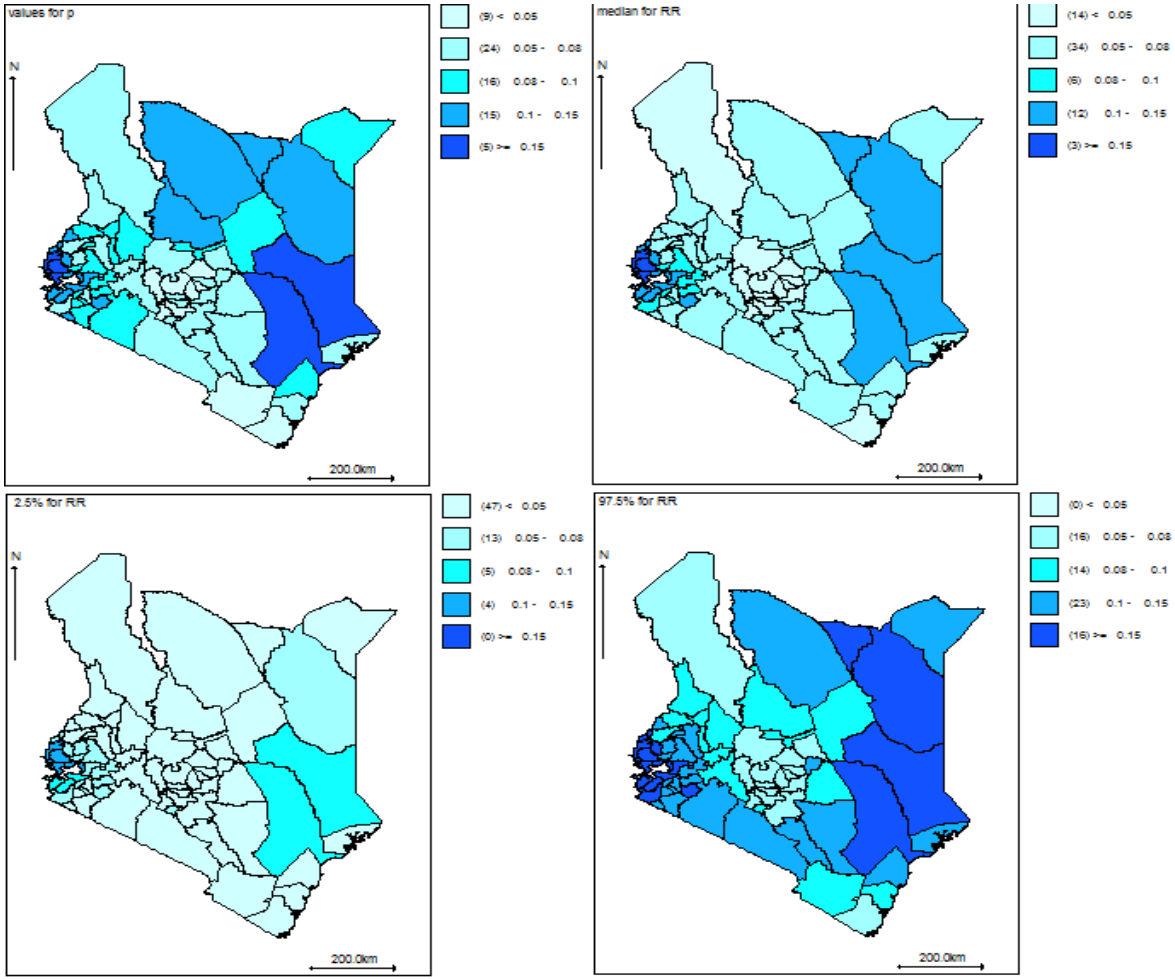
96 deaths per 1,000 live births

Region	Observed (O_i)	Total (N_i)	Expected (E_i)	SMR (θ)
Rongo	600	10000	960	0.625
Kitweru	600	1000	96	6.25
Kodero Bara	600	6000	576	1.04
Kanga	600	8000	768	0.78125

$$E_i = rN_+$$

$$SMR_i = \frac{O_i}{E_i}$$

Map of infant mortality in Kenya



Spatial statistics-Opportunity to contribute

- *Spatial statistical analysis has never been in the mainstream of statistical theory. However, there is a growing interest both for epidemiologic studies, and in analyzing disease processes.*
- *...since spatial statistics is often out of the mainstream of statistical theory, it is often also out of the mainstream of statistical software.* [Waller and Gotway, 2004]

Issues

- Areal referenced data
- Maps often show raw SMRs (O_i/E_i)-Crude SMR maps
- Naive use of disease mapping can be Misleading
- For small areas and/or rare diseases, risk estimates are very unstable
- *Borrowing strength* and developing (Choropleth maps) risk maps
- Estimation techniques and Inference for:
  - Gaussian data (Spatial error, spatial lag and spatial durbin models)-MLE, REML and
  - non-Gaussian data (Bayesian MCMC methods and the Integrated Nested Laplace Approximation, INLA)

The standard GLMM

Let y_{ij} be a random variable from the exponential dispersion family and so the stochastic part of the model is

$$f(y_{ij}|\beta, \mathbf{u}, \phi) = \exp\left(\frac{y_{ij} - \psi(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right)$$

The deterministic part of the model

$$g(E(\mathbf{y}|\mathbf{u})) = \mathbf{X}\beta + \mathbf{Z}\mathbf{u}$$

where $\mathbf{u} \sim MVN(\mathbf{0}, \Sigma_u)$

Parameter estimation in the frequentist paradygm

- The likelihood function is given as

$$\begin{aligned} L(\beta, \mathbf{U}, \phi; \mathbf{y}) &= \prod_{i=1}^{N} f(y_i|\mathbf{U}, \beta, \phi) \\ &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f(y_i|\mathbf{U}, \beta, \phi) f(u_i|\mathbf{U}) du_i \end{aligned} \tag{1.1}$$

- cannot be evaluated analytically so ML-estimation leads to problems. Exceptions are normal-normal models and poisson-gamma models. For numerical integration we have the problem of high-dimensional integration

Hierarchical spatial modeling

Let $y_1, \ldots, y_n$ denote response from n areal units and θ denote the risk or parameter of interest.

$$\underbrace{f(\theta|\mathbf{y})}_{posterior} \propto \underbrace{f(y|\theta)}_{Likelihood} \underbrace{f(\theta)}_{Prior}$$

$$\underbrace{f(\beta, \phi, \psi|\mathbf{y})}_{Posterior} \propto \underbrace{f(\mathbf{y}|\beta, \phi, \psi)}_{Likelihood} \underbrace{f(\beta|\phi, \psi)}_{Prior} \underbrace{f(\phi)f(\psi)}_{Hyperpriors}$$

Disease mapping with count data

1. The Poisson-gamma model with a two-level hierarchy

$$\underbrace{y_i|\theta, E_i \sim Poisson(\theta E_i)}_{\text{Stage 1-Likelihood}}, \quad \underbrace{\theta \sim Gamma(\alpha, \beta)}_{\text{Stage 2-Prior distribution}} .$$

2. The Poisson-gamma model with a three-level hierarchy

$$\underbrace{y_i|\theta, E_i \sim Poisson(\theta E_i)}_{\text{Stage 1-Likelihood}},$$

$$\underbrace{\theta|\alpha, \beta \sim Gamma(\alpha, \beta)}_{\text{Stage 2-Prior distribution}},$$

$$\underbrace{\alpha|\nu \sim h_\alpha(\nu) \text{ and } \beta|\rho \sim h_\beta(\rho)}_{\text{Stage 3-Hyperprior distribution}}$$

Focus areas

- Hierarchical Spatial modeling (Small area estimation problems)
- Geostatistics
- Spatiotemporal modeling
- Spatial survival modeling
- Joint disease modeling
- **Areas of emphasis**: Disease mapping; poverty mapping; animal movement; public health and epidemiological problem (in Africa)

Examples of Standard spatial models for disease mapping

- Interest: Estimating risk of disease (θ);
  Response variable: $y_i \sim$Poisson(θE_i)
- Univariate (response) spatial methods

GLM	GLMM $1-$UH model
$\log(y_i) = \log(E_i) + \mathbf{X}\beta$	$\log(y_i) = \log(E_i) + \mathbf{X}\beta + v_i$
GLMM-CAR	**GLMM-Besag, York and Mollie**
$\log(y_i) = \log(E_{ij}) + \mathbf{X}\beta + u_i$	$\log(y_i) = \log(E_i) + \mathbf{X}\beta + u_i + v_i$

- Spatiotemporal methods

$$\log(y_{ij}) = \log(E_i) + \mathbf{X}_{ij}\beta + u_i + v_i + \tau_j + \eta_{ij}$$

Outline

Models for small area estimation problems

GLM	GLMM 1−UH model
$\text{logit}(p_i) = \mathbf{X}\beta$	$\text{logit}(p_i) = \mathbf{X}\beta + v_i$
GLMM-CAR	GLMM-Besag, York and Mollie
$\text{logit}(p_i) = \mathbf{X}\beta + u_i$	$\text{logit}(p_i) = \mathbf{X}\beta + u_i + v_i$

Bayesian logistic regression model

- Let y_i be the number of infant deaths out of n_i live births cases in district $i, i = 1, \ldots, n$. Assume $y_i \sim Bin(n_i, \pi_i)$
- The Bayesian logistic regression model

$$\text{logit}(\pi_i) = \mathbf{X}_i'\beta$$

- The posterior distribution is given as

$$f(\beta|\mathbf{y}) \propto f(\mathbf{y}|\beta)f(\beta) = \prod_{i=1}^{n} f(y_i|\beta) \prod_{i=1}^{p} f(\beta_i),$$

$$\propto \underbrace{\prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}}_{Likelihood} \times \underbrace{\prod_{j=1}^{p} e^{-\frac{1}{2c}\beta_j^2}}_{Normal\ \ priors}. \quad (2.1)$$

The Conditionally autoregressive (CAR) Model

$$\text{logit}(\pi_i) = \mathbf{X}_i'\beta + \underbrace{u_i}_{\substack{\textit{spatially structured} \\ \textit{random effect}}}$$

where $u_i|\mathbf{u}_{-i} \sim N(\frac{\phi}{n_i}\sum_{j\in\mathcal{N}_i} u_j, \frac{1}{\tau^2 n_i})$.

The UH Model

- The logistic regression model with spatial unstructured random effects is given as

$$\text{logit}(\pi_i) = \mathbf{X}_i'\beta + \underbrace{v_i}_{\substack{\textit{spatially unstructured} \\ \textit{random effect}}}$$

where $v_i \sim N(0, \sigma^2)$.

Besag, York and Mollie model

- Let y_i be the number of infant deaths out of n_i live births cases in district $i, i = 1, \ldots, n$. Assume $y_i \sim Bin(n_i, \pi_i)$
- The convolution model

$$\text{logit}(\pi_i) = \mathbf{X}_i'\beta + \underbrace{u_i}_{\substack{\textit{spatially structured} \\ \textit{random effect}}} + \underbrace{v_i}_{\substack{\textit{spatially unstructured} \\ \textit{random effect}}}$$

where $v_i \sim N(0, \sigma^2)$ and $u_i|\mathbf{u}_{-i} \sim N(\frac{\phi}{n_i}\sum_{j\in\mathcal{N}_i} u_j, \frac{1}{\tau^2 n_i})$.

Parameters estimation and Inference

The posterior distribution is given as

$$f(\mathbf{u},\mathbf{v},\kappa,\lambda,\boldsymbol{\beta}|\mathbf{y}) \propto \prod_{i=1}^{n} f(y_i|u,v,\kappa,\lambda,\beta) \prod_{i=1}^{p} f(\beta_i) \prod_{i=1}^{n} f(u_i|\kappa) f(\kappa) \prod_{i=1}^{n} f(v_i|\lambda) f(\lambda),$$

Expressible as the Hierarchical model

$$f(u,v,\kappa,\lambda|\mathbf{y}) \propto \underbrace{\prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}}_{\textit{Likelihood}} \times \underbrace{\frac{1}{\kappa^{n/2}} \times \exp -\frac{1}{2\kappa} \sum_{i\neq j}(u_i-u_j)^2}_{\text{structured CAR prior}}$$

$$\times \underbrace{\frac{1}{\lambda^{n/2}} \exp -\frac{1}{2\lambda} \sum_{i=1}^{n} v_i^2}_{\text{unstructured exchangable prior}} \times \underbrace{\prod_{j=1}^{p} e^{-\frac{1}{2c}\beta_j^2}}_{\textit{Normal priors}} \times \underbrace{e^{-\kappa/0.005}\kappa^{0.05} \times e^{-\lambda/0.005}\lambda^{0.05}}_{\textit{hyperpriors}}$$

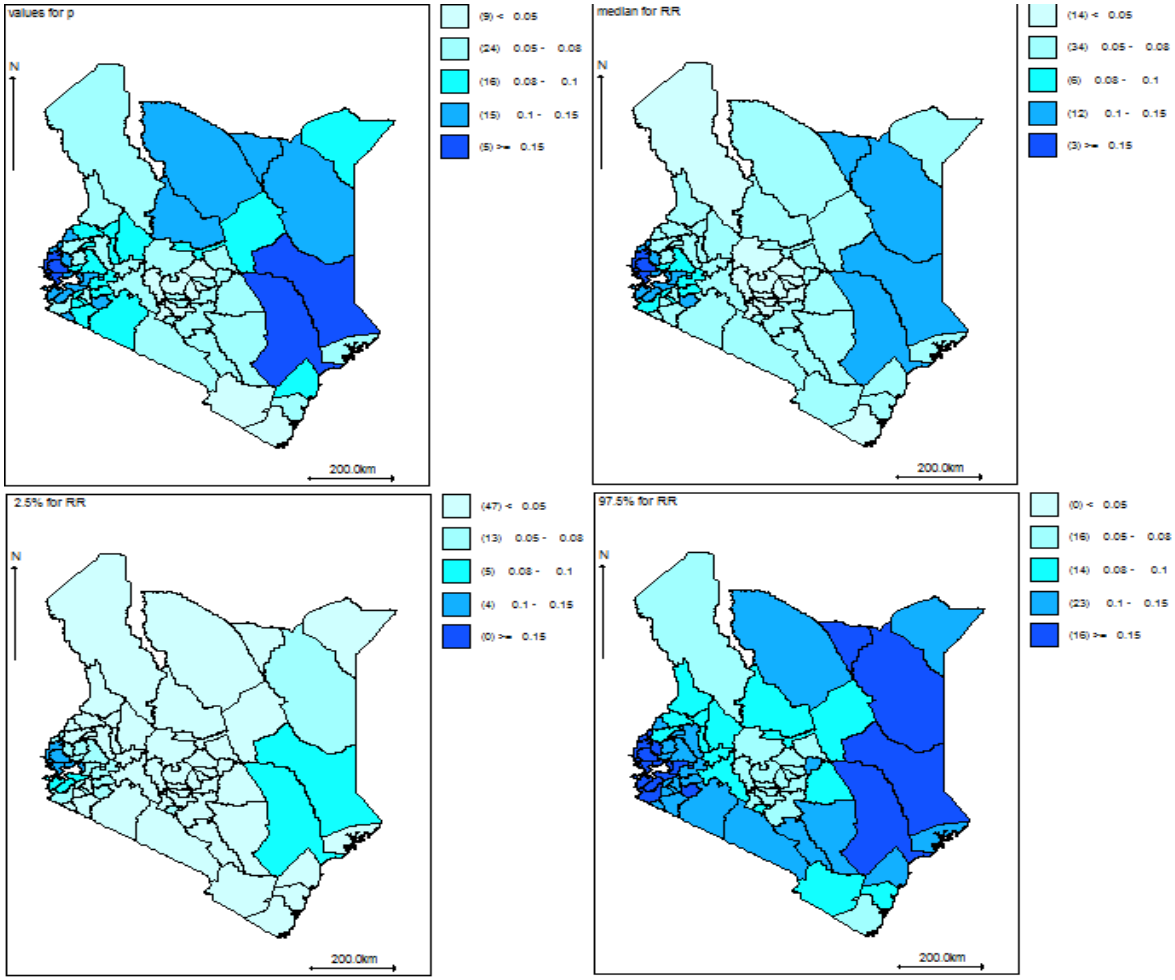
CAR model

- The spatial modeling was carried out in WinBUGS 1.4.3 [Spiegelhalter et al., 2002], using Markov Chain Monte Carlo (MCMC) simulation methodology.
- For all the models, 100,000 iterations were carried out, discarding the first 10,000 samples and storing every tenth sample, giving 9,000 samples. These were then summarized to get the relevant estimates.
- Then, following the approach given in [Kleinschmidt et al., 2006], the distribution of

$$\hat{p}_i = \frac{\exp\left(\mathbf{X}_i'\beta + u_i + v_i\right)}{1 + \exp\left(\mathbf{X}_i'\beta + u_i + v_i\right)} \Rightarrow RR_i = \frac{\hat{p}_i}{\hat{p}}$$

the corresponding 2.5 and 97.5 percentiles were mapped as estimates of an approximate 95% credible interval for the

Map of infant mortality in Kenya



Outline

Issues

- *Focus*: Point referenced data; Smooth maps of disease and ecological phenomena
- Standard Kriging techniques not well developed for non-Gaussian data;
- *Kriging* assumes a Gaussian random field and a covariance function are known exactly-may not be case in practice [Pilz and Spöck, 2008].
- Poorly estimated theoretical variograms=incorrectly fitted theoretical models.
- Lognormal [Dowd, 1982], disjunctive [Rivoirard, 1994], tranGaussian [Christensen et al., 2001]and Bayesian transformed tranGaussian Kriging [De Oliveira et al., 1997];
- Current Bayesian MCMC approaches are costly in computing time

The Bayesian geostatistical linear model

- Model based geostatistics [Diggle et al., 1998]. Assuming $y_i \sim bin(n_i, p_i)$, with

$$logit(p_i) = \beta_0 + \mathbf{Z}_i'\beta + S(\mathbf{x}_i),$$

$S = \{S(\mathbf{x}_1), \ldots, S(\mathbf{x}_n)\}$, $\mu(\mathbf{x}) = E[S(\mathbf{x})]$,
$\gamma(h) = Cov\{S(\mathbf{x}), S(\mathbf{x}+\mathbf{h})\}$ and $\rho(\boldsymbol{\theta}) = \sigma^2 \exp(-d_{ij}/\phi)$,
$\phi-$rate of correlation decay parameter, and $d_{ij} = ||x_i - x_j||$.

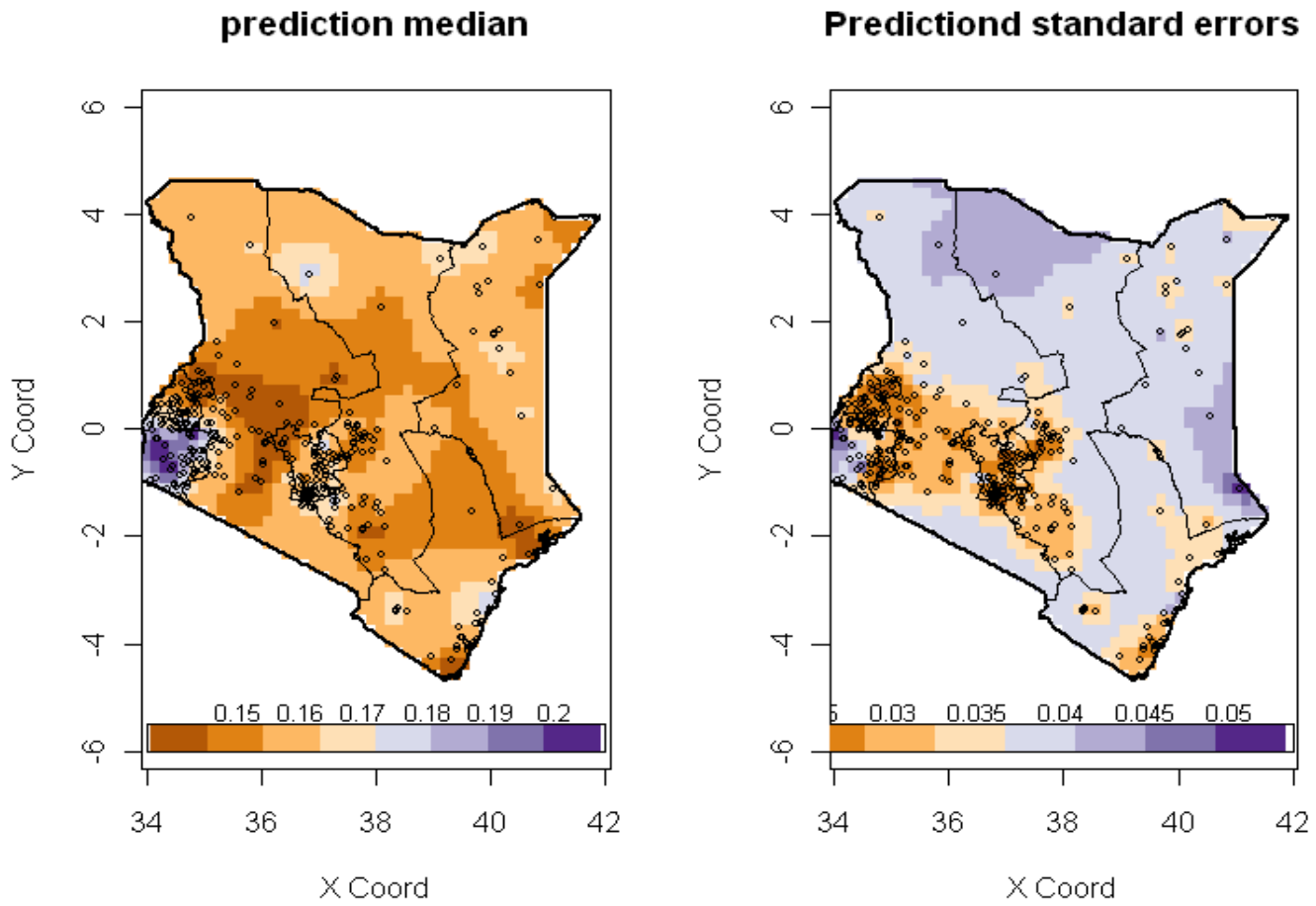
- Estimation. The joint posterior distribution is given as

$$f(\beta, S, \sigma^2, \phi|\mathbf{Y}) \propto f(\beta, S; \mathbf{Y}) f(S|\sigma^2, \phi) f(\sigma^2) f(\phi),$$

- Prediction. $\mathbf{Y}_0$ given $\hat{\beta}, \hat{S}, \hat{\sigma}^2, \hat{\phi}$ is given by

$$f(\mathbf{Y}_0|\hat{\beta}, \hat{S}, \hat{\sigma}^2, \hat{\phi}) = \int f(\mathbf{Y}_0|\hat{\beta}, S_0) f(S_0|\hat{S}, \hat{\sigma}^2, \hat{\phi}) dS_0$$

Smooth map of infant mortality in Kenya



Currents trends on spatial methods for relative risk estimation

- Univariate (response) spatial methods
  - [Clayton and Kaldor, 1987, Besag et al., 1991, Marshall, 1991, Kazembe et al., 2006, Kazembe, 2007]
- Spatiotemporal methods
  - [Bernardinelli et al., 1995, CARLIN and LOUIS, 2000, Böhning et al., 2000, Knorr-Held et al., 1998, Knorr-Held and Richardson, 2003]
- Spatial survival & Longitudinal methods
  - [Oleson et al., 2008, Manda et al., 2012b, Banerjee and Carlin, 2003, Banerjee et al., 2003, Carlin and Hodges, 1999]
- Joint disease mapping
  - [Manda et al., 2012a, Kazembe et al., 2009]

Current work

1. Geospatial modeling of Infant mortality in Kenya
2. Bayesian semiparametric modeling and mapping of self-reported fevers using routine data in Kenya
3. Bayesian spatiotemporal modeling and mapping of TB in Kenya
4. Bayesian spatiotemporal modeling and mapping of malaria in Angola
5. Joint modeling and mapping of HIV, HSV-2 and STI's in Kenya
6. Spatiotemporal modeling of Elephant movement data from the Kruger National Park

S. Banerjee and B.P. Carlin. Semiparametric spatio-temporal frailty modeling. *Environmetrics*, 14(5):523–535, 2003.

S. Banerjee, M.M. Wall, and B.P. Carlin. Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics*, 4(1): 123–142, 2003.

L. Bernardinelli, D. Clayton, C. Pascutto, C. Montomoli, M. Ghislandi, and M. Songini. Bayesian analysis of spacetime variation in disease risk. *Statistics in Medicine*, 14(21-22):2433–2443, 1995.

J. Besag, J. York, and A. Mollie. Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann Inst Stat Math.*, 43:1–59, 1991.

D. Böhning, E. Dietz, and P. Schlattmann. Space-time mixture modelling of public health data. *Statistics in medicine*, 19(17-18):2333–2344, 2000.

B. CARLIN and T.A. LOUIS. Bayes and empirical bayes methods for data analysis (2000). *Recherche*, 67:02, 2000.

B.P. Carlin and J.S. Hodges. Hierarchical proportional hazards regression models for highly stratified data. *Biometrics*, 55(4):1162–1170, 1999.

O.F. Christensen, P.J. Diggle, and P.J. Ribeiro. Analysing positive-valued spatial data: the transformed gaussian model. *GeoENV IIIGeostatistics for Environmental Applications*, pages 287–298, 2001.

D. Clayton and J. Kaldor. Empirical bayes estimates of age-standardized relative risk for use in disease mapping. *Biometrics*, 43:671–681, 1987.

V. De Oliveira, B. Kedem, and D.A. Short. Bayesian prediction of transformed gaussian random fields. *Journal of the American Statistical Association*, pages 1422–1433, 1997.

P.J. Diggle, R. A. Moyeed, and J. A. Tawn. Model-based geostatistics. *Applied Statistics*, 47:299–350, 1998.

PA Dowd. Lognormal krigingthe general case. *Mathematical Geology*, 14(5):475–499, 1982.

L.N. Kazembe. Spatial modelling and risk factors of malaria incidence in northern malawi. *Acta tropica*, 102(2):126–137, 2007.

L.N. Kazembe, I. Kleinschmidt, and B.L. Sharp. Patterns of malaria-related hospital admissions and mortality among malawian children: an example of spatial modelling of hospital register data. *Malaria journal*, 5(1):93, 2006.

L.N. Kazembe, A.S. Muula, and C. Simoonga. Joint spatial modelling of common morbidities of childhood fever and diarrhoea in malawi. *Health & Place*, 15(1): 165–172, 2009.

I. Kleinschmidt, A. Ramkissoon, N. Morris, Z. Mabude, B. Curtis, and M. Beksinska. Mapping indicators of sexually transmitted infection services in the south african public health sector. *Tropical Medicine & International Health*, 11(7):1047–1057, 2006.

L. Knorr-Held and S. Richardson. A hierarchical model for space–time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(2):169–183, 2003.

L. Knorr-Held, J. Besag, et al. Modelling risk from a disease in time and space. *Statistics in medicine*, 17(18):2045–2060, 1998.

S.O.M. Manda, R.G. Feltbower, and M.S. Gilthorpe. A multivariate random frailty effects model for multiple spatially dependent survival data. *Modern Methods for Epidemiology*, pages 157–172, 2012a.

S.O.M. Manda, C.J. Lombard, T. Mosala, et al. Divergent spatial patterns in the prevalence of the human immunodeficiency virus (hiv) and syphilis in south african pregnant women. *Geospatial Health*, 6(2):221–231, 2012b.

R.J. Marshall. Mapping disease and mortality rates using empirical bayes estimators. *Applied Statistics*, pages 283–294, 1991.

J.J. Oleson, B.J. Smith, and H. Kim. Joint spatio-temporal modeling of low incidence cancers sharing common risk factors. *Journal of Data Science*, 6:105–123, 2008.

J. Pilz and G. Spöck. Why do we need and how should we implement bayesian kriging methods. *Stochastic Environmental Research and Risk Assessment*, 22(5):621–632, 2008.

J. Rivoirard. *Introduction to disjunctive kriging and non-linear geostatistics*. Clarendon Press, 1994.

S. D. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. D. Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. ISSN 1369-7412. URL `http://yaroslav.hopto.org/papers/speigelhalter-bayesian-model-complexity.pdf`.

L.A. Waller and C.A. Gotway. *Applied spatial statistics for public health data*, volume 368. Wiley-Interscience, 2004.

- THANK YOU
- *Questions Comments*