

Case-study: Rongelap Island

This case-study illustrates a model-based geostatistical analysis combining:

- a Poisson log-linear model for the sampling distribution of the observations, conditional on a latent Gaussian process which represents spatial variation in the level of contamination
- Bayesian prediction of non-linear functionals of the latent process
- MCMC implementation

Details are in Diggle, Moyeed and Tawn (1998).

Radiological survey of Rongelap Island

- **Rongelap Island**

- approximately 2500 miles south-west of Hawaii
- contaminated by nuclear weapons testing during 1950's
- evacuated in 1985
- now safe for re-settlement?

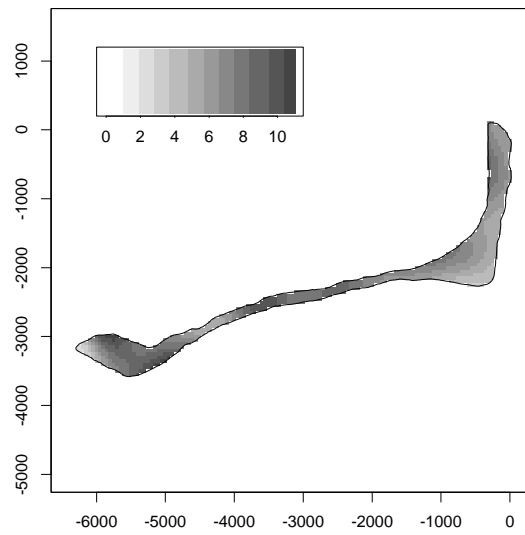
- **The statistical problem**

- field-survey of ^{137}Cs measurements
- estimate spatial variation in ^{137}Cs radioactivity
- compare with agreed safe limits

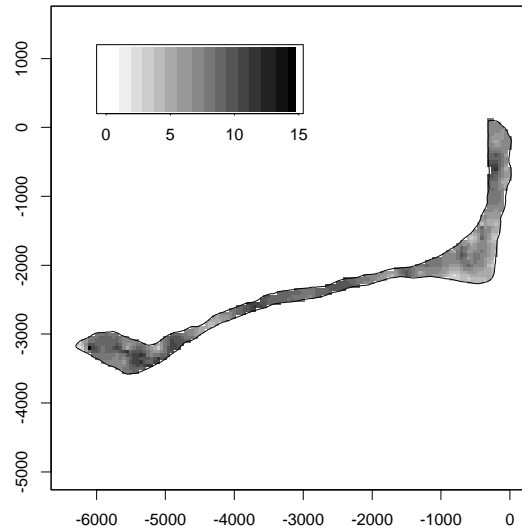
Poisson Model for Rongelap Data

- Basic measurements are nett counts Y_i over time-intervals t_i at locations x_i ($i = 1, \dots, n$)
- Suggests following model:
 - $S(x) : x \in R^2$ stationary Gaussian process (local radioactivity)
 - $Y_i | \{S(\cdot)\} \sim \text{Poisson}(\mu_i)$
 - $\mu_i = t_i \lambda(x_i) = t_i \exp\{S(x_i)\}$.
- Aims:
 - predict $\lambda(x)$ over whole island
 - $\max \lambda(x)$
 - $\arg(\max \lambda(x))$

Predicted radioactivity surface using log-Gaussian kriging



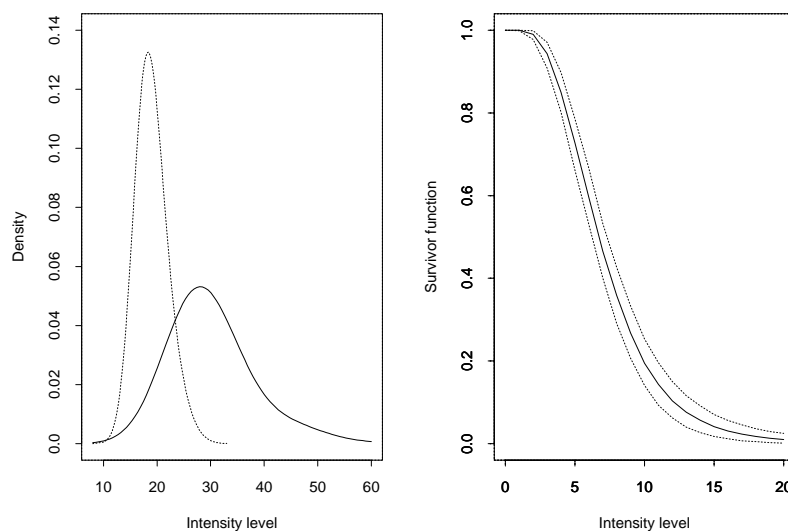
Predicted radioactivity surface using Poisson log-linear model with latent Gaussian process



- The two maps above show the difference between:
 - log-Gaussian kriging of observed counts per unit time
 - log-linear analysis of observed counts
- the principal visual difference is in the extent of spatial smoothing of the data, which in turn stems from the different treatments of the nugget variance

Bayesian prediction of non-linear functionals of the radioactivity surface

The left-hand panel shows the predictive distribution of maximum radioactivity, contrasting the effects of allowing for (solid line) or ignoring (dotted line) parameter uncertainty; the right-hand panel shows 95% pointwise credible intervals for the proportion of the island over which radioactivity exceeds a given threshold.



- The two panels of the above diagram illustrate Bayesian prediction of non-linear functionals of the latent Gaussian process in the Poisson log-linear model
- the left-hand panel contrasts posterior distributions of the maximum radioactivity based on:
 - (i) the fully Bayesian analysis incorporating the effects of parameter uncertainty in addition to uncertainty in the latent process (solid line)
 - (ii) fixing the model parameters at their estimated values, ie allowing for uncertainty only in the latent process
- the right-hand panel gives posterior estimates with 95% point-wise credible intervals for the proportion of the island over which radioactivity exceeds a given threshold (dotted line).

Case-study: Gambia malaria

- In this example, the spatial variation is of secondary scientific importance.
- The primary scientific interest is to describe how the prevalence of malarial parasites depends on explanatory variables measured:
 - on villages
 - on individual children
- There is a particular scientific interest in whether a vegetation index derived from satellite data is a useful predictor of malaria prevalence, as this would help health workers to decide how to make best use of scarce resources.

Data-structure

- 2039 children in 65 villages
- test each child for presence/absence of malaria parasites

Covariate information at child level:

- age (days)
- sex (F/M)
- use of mosquito net (none, untreated, treated)

Covariate information at village level:

- location
- vegetation index, from satellite data
- presence/absence of public health centre

Logistic regression model

Logistic model for presence/absence in each child:

- $Y_{ij} = 0/1$ for absence/presence of malaria parasites in j th child in i th village
- f_{ij} = child-specific covariates
- w_i = village-specific covariate
- $\text{logit}P(Y_{ij} = 1|S(\cdot)) = f'_{ij}\beta_1 + w'_i\beta_2 + S(x_i)$

Is it reasonable to assume conditionally independent infections within same village?

If not, we might wish to extend the model to allow for non-spatial extra-binomial variation:

- $U_i \sim N(0, \nu^2)$
- $\text{logit}P(Y_{ij} = 1|S(\cdot), U) = f'_{ij}\beta_1 + w'_i\beta_2 + U_i + S(x_i)$

Exploratory analysis

- fit standard logistic linear model, ignoring $S(x)$ and/or U

- compute for each village:

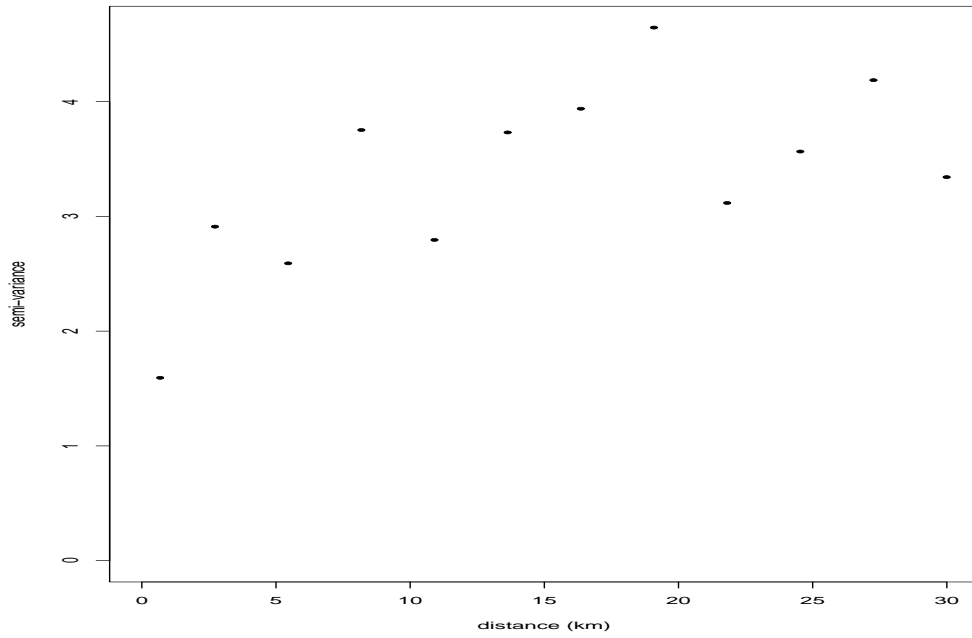
$$N_i = \sum_{j=1}^{n_i} Y_{ij}$$

$$\mu_i = \sum_{j=1}^{n_i} \hat{P}_{ij}$$

$$\sigma_i^2 = \sum_{j=1}^{n_i} \hat{P}_{ij}(1 - \hat{P}_{ij})$$

- compute village-residuals, $r_i = (N_i - \mu_i)/\sigma_i$
- apply conventional geostatistics to derived data r_i
- variogram indicates residual spatial structure

Variogram of residuals



Model-based geostatistical analysis

α = intercept term in linear predictor

β_1 = regression coefficient for age

β_2 = regression coefficient for bed-net use

β_3 = regression coefficient for treated bed-net

β_4 = regression coefficient for green-ness index

β_5 = regression coefficient for presence of public health centre in village

ν^2 = variance of non-spatial random effects U_i

σ^2 = variance of spatial process $S(x)$

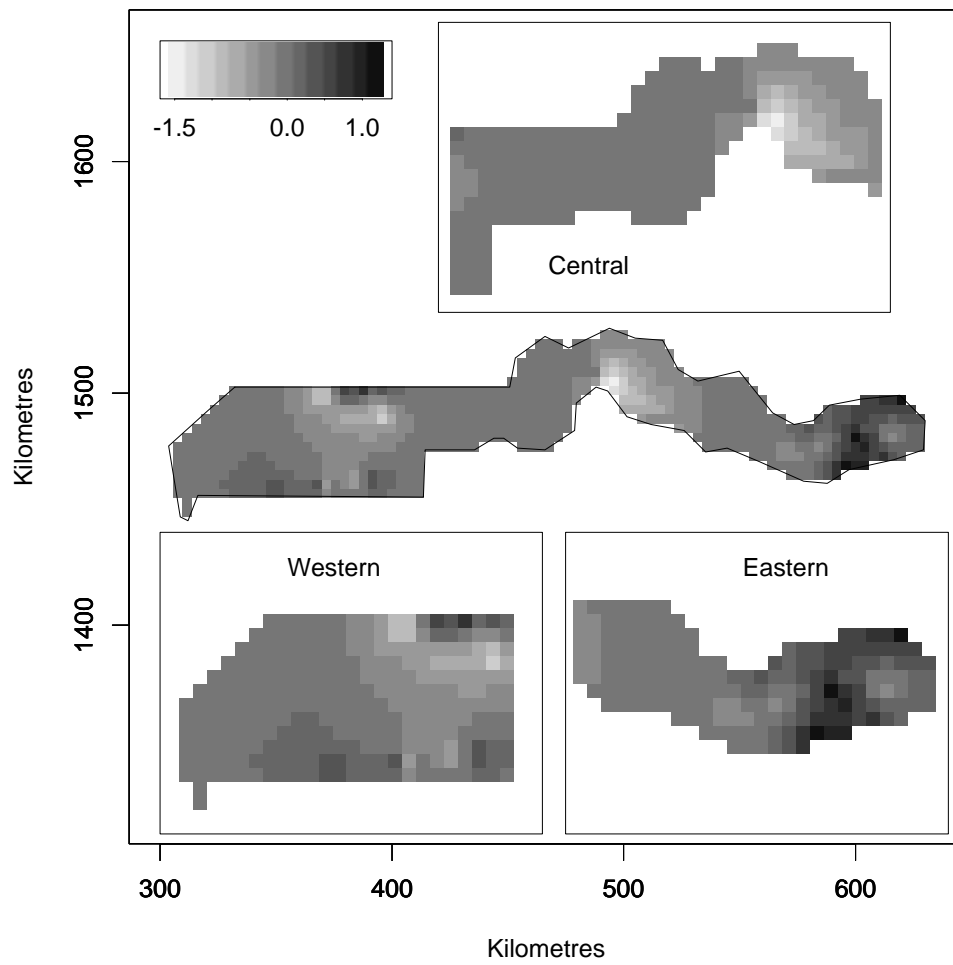
ϕ = rate of decay of spatial correlation with distance

κ = shape parameter for Matérn correlation function

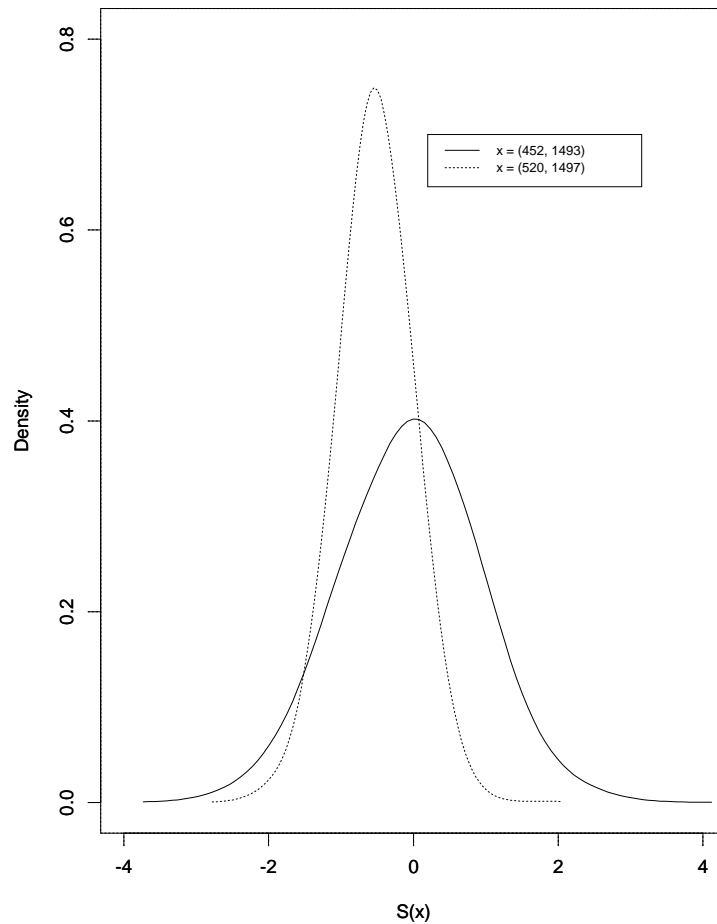
Param.	2.5% Qt.	97.5% Qt.	Mean	Median
α	-4.232073	1.114734	-1.664353	-1.696228
β_1	0.000442	0.000918	0.000677	0.000676
β_2	-0.684407	-0.083811	-0.383750	-0.385772
β_3	-0.778149	0.054543	-0.355655	-0.355632
β_4	-0.039706	0.071505	0.018833	0.020079
β_5	-0.791741	0.180737	-0.324738	-0.322760
ν^2	0.000002	0.515847	0.117876	0.018630
σ^2	0.240826	1.662284	0.793031	0.740790
ϕ	1.242164	53.351207	11.653717	7.032258
κ	0.150735	1.955524	0.935064	0.830548

- note concentration of posterior for ν^2 close to zero

Map of the predicted surface $\hat{S}(x)$ (posterior mean)

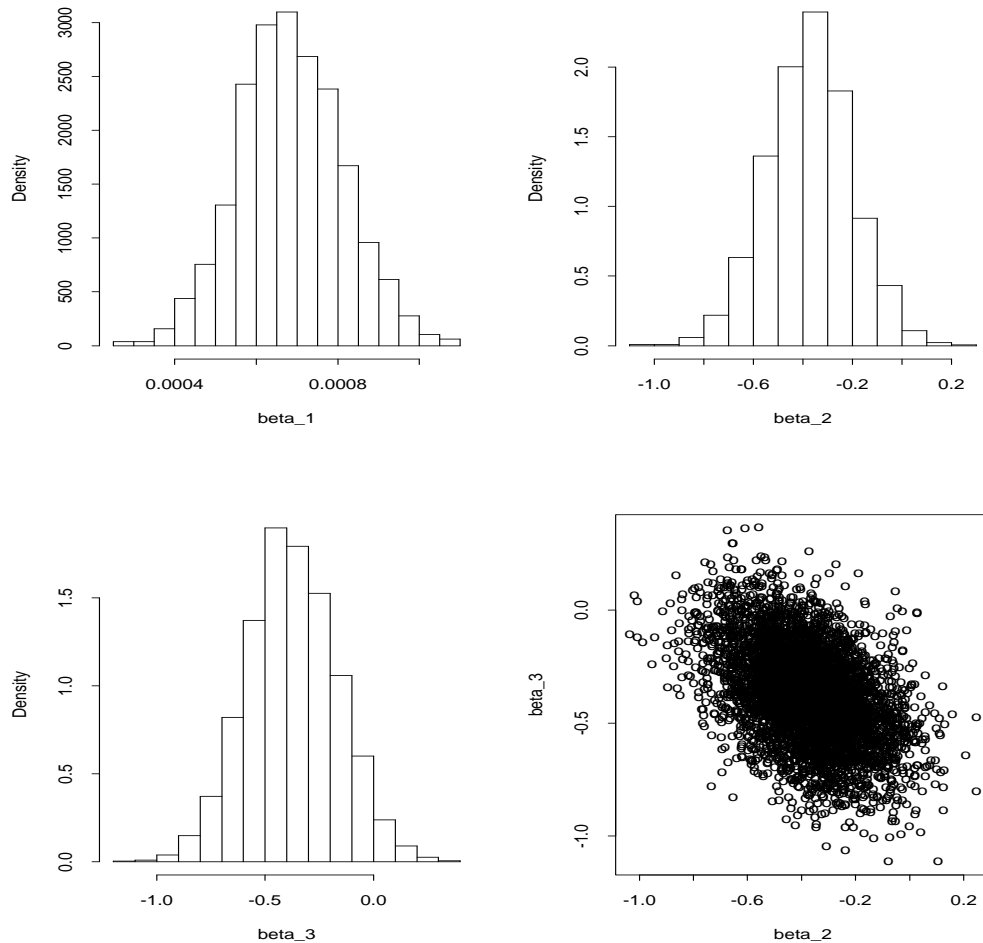


Posterior density estimates for $S(x)$ at two selected locations.



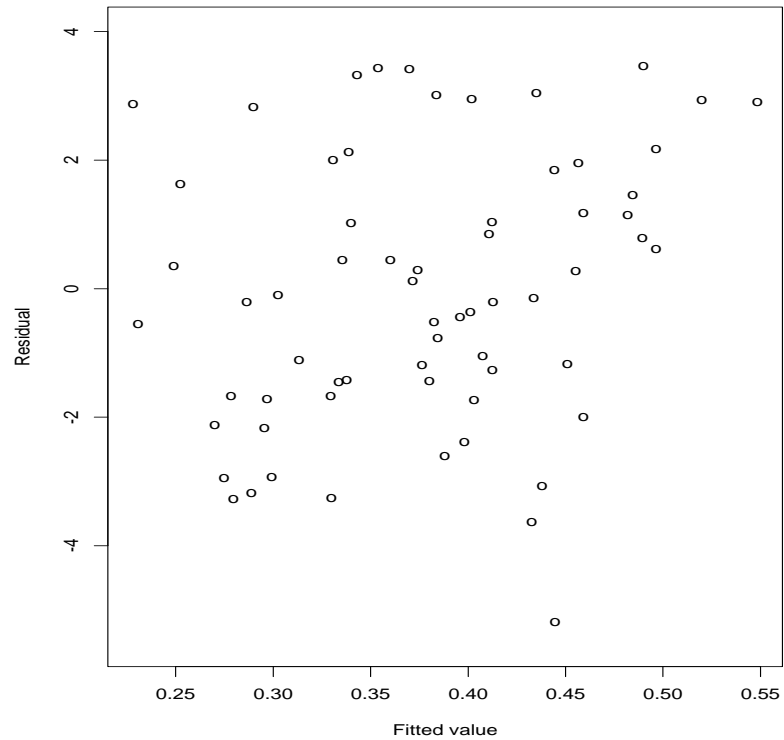
- solid curve – remote location (452, 1493),
- dashed curve – location (520, 1497), close to observed sites in central region.

Empirical posterior distributions for regression parameters



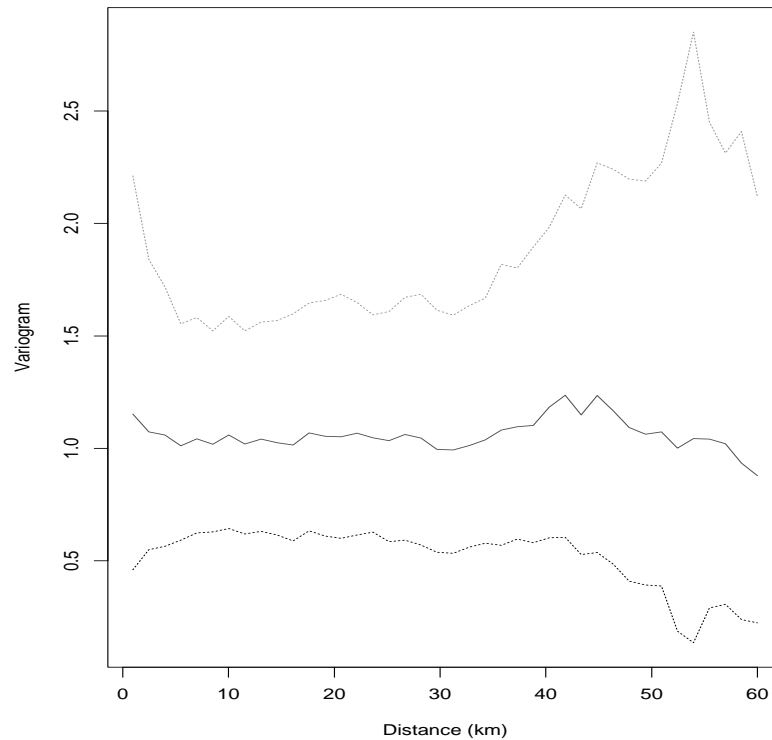
- β_1 = effect of age
- β_2 = effect of untreated bed-nets
- β_3 = additional effect of treated bed-nets

Goodness-of-fit for Gambia malaria model



Village-level residuals against fitted values.

- $r_{ij} = (Y_{ij} - \hat{p}_{ij}) / \sqrt{\{\hat{p}_{ij}(1 - \hat{p}_{ij})\}}$
- $r_i = \sum r_{ij} / \sqrt{n_i}$
- intended to check adequacy of model for p_{ij}



Standardised residual empirical variogram plot (village-level data and pointwise 95% posterior intervals constructed from simulated realisations of fitted model).

- $r_{ij} = (Y_{ij} - \hat{p}_{ij}^*) / \sqrt{\{\hat{p}_{ij}^*(1 - \hat{p}_{ij}^*)\}}$
- $r_i = \sum r_{ij} / \sqrt{n_i}$
- $\text{logit}p_{ij}^* = \hat{\alpha} + f'_{ij}\hat{\beta} + \hat{S}(x_i)$
- intended to check adequacy of model for $S(x)$