

# Geostatistics, Preferential Sampling and Environmental Monitoring

**Peter J Diggle**

*Department of Medicine, Lancaster University  
and*

*Department of Biostatistics, Johns Hopkins University*

In collaboration with:

Ciprian Crainiceanu (Johns Hopkins University, USA)

Raquel Menezes (University of Minho, Portugal)

Barry Rowlingson (Lancaster University, UK)

Ting-Li Su (Lancaster University, UK)

September 2007

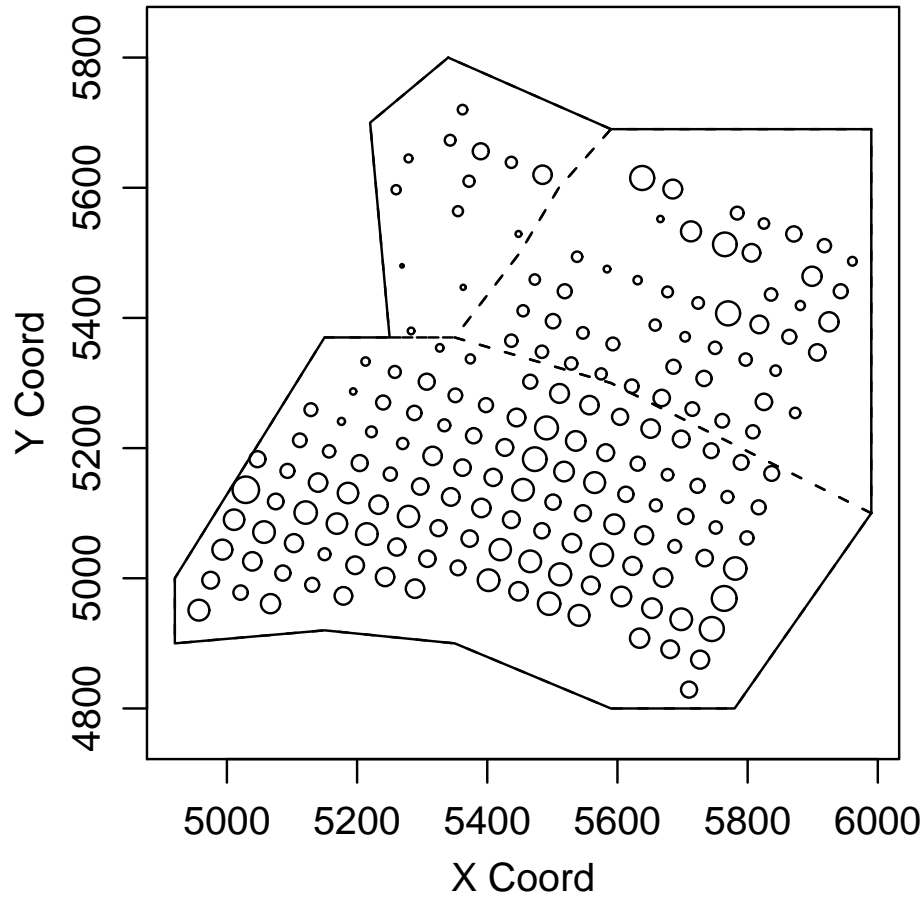
# Outline

- **Model-based geostatistics**
  - modelling and likelihood-based inference
  - case study: tropical disease prevalence mapping
- **Preferential sampling**
  - definitions
  - likelihood-based inference
  - two (incomplete) environmental monitoring applications
- **Discussion**

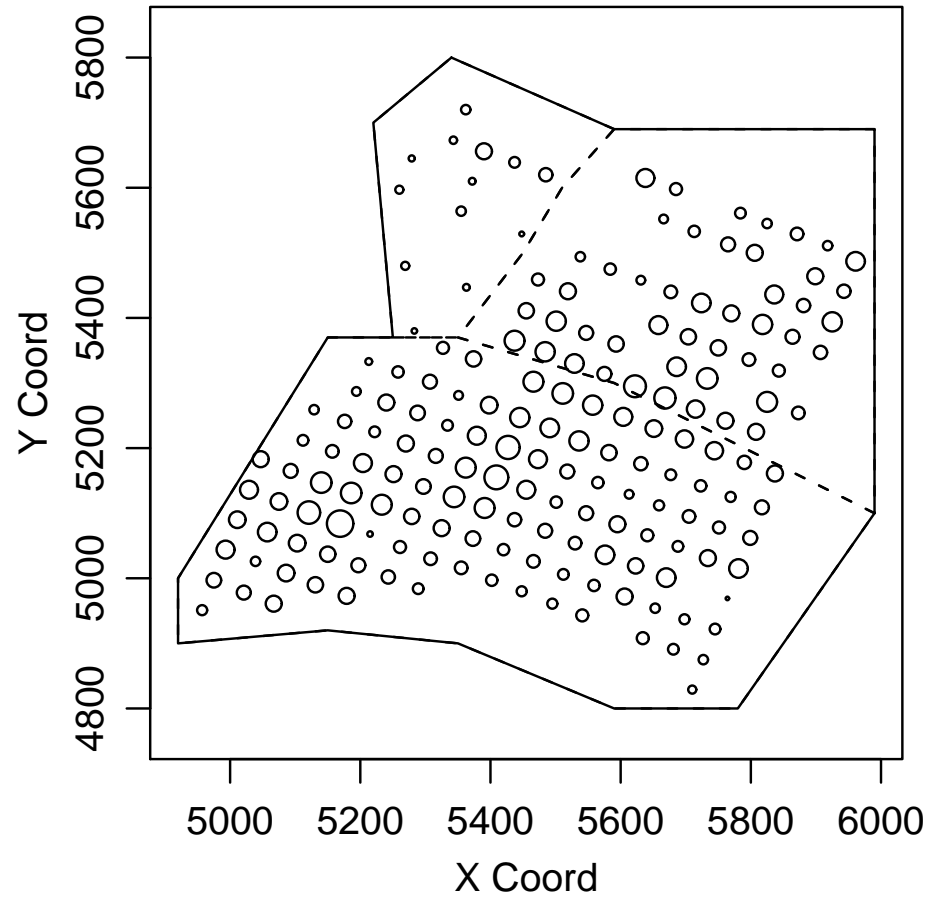
# Geostatistics

- traditionally, a self-contained methodology for spatial prediction, developed at École des Mines, Fontainebleau, France
- nowadays, that part of spatial statistics which is concerned with data obtained by spatially discrete sampling of a spatially continuous process

# Soil chemistry data



**calcium**



**magnesium**

# Model-based Geostatistics

- the application of general principles of statistical modelling and inference to geostatistical problems
- **Example:** kriging as minimum mean square error prediction under Gaussian modelling assumptions

# Gaussian geostatistics

## Model

- Stationary Gaussian process  $S(x) : x \in \mathbb{R}^2$ 
  - $\mathbf{E}[S(x)] = \mu$
  - $\text{Cov}\{S(x), S(x')\} = \sigma^2 \rho(\|x - x'\|)$
- Mutually independent  $Y_i | S(\cdot) \sim \mathbf{N}(S(x), \tau^2)$

Point predictor  $\hat{S}(x) = \mathbf{E}[S(x) | Y]$

- linear in  $Y = (Y_1, \dots, Y_n)$ ;
- interpolates  $Y$  if  $\tau^2 = 0$ .

## Parameter uncertainty

- traditionally ignored (plug-in prediction)
- Bayesian paradigm accommodates easily

# African Programme for Onchocerciasis Control

- “river blindness” – an endemic disease in wet tropical regions
- donation programme of mass treatment with ivermectin
- approximately 30 million treatments to date
- serious adverse reactions experienced by some patients highly co-infected with *Loa loa* parasites
- precautionary measures put in place before mass treatment in areas of high *Loa loa* prevalence

<http://www.who.int/pbd/blindness/onchocerciasis/en/>

# The *Loa loa* prediction problem

## Ground-truth survey data

- random sample of subjects in each of a number of villages
- blood-samples test positive/negative for *Loa loa*

## Environmental data (satellite images)

- measured on regular grid to cover region of interest
- elevation, green-ness of vegetation

## Objectives

- predict local prevalence throughout study-region (Cameroon)
- compute local exceedance probabilities,

$$P(\text{prevalence} > 0.2 | \text{data})$$

# Loa loa: a generalised linear model

- Latent spatial process

$$S(x) \sim \text{SGP}\{0, \sigma^2, \rho(u)\}$$

$$\rho(u) = \exp(-|u|/\phi)$$

- Linear predictor

$d(x)$  = environmental variables at location  $x$

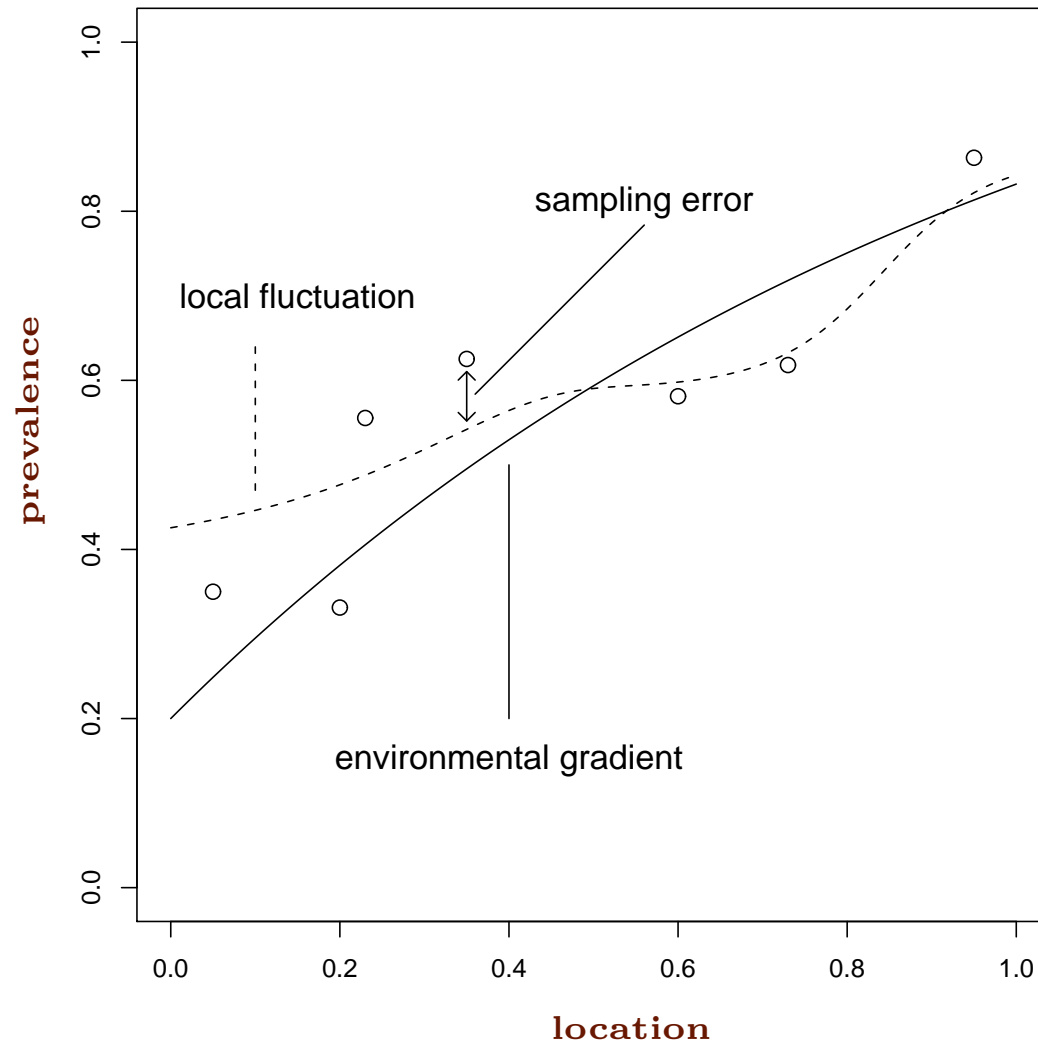
$$\eta(x) = d(x)' \beta + S(x)$$

$$p(x) = \log[\eta(x) / \{1 - \eta(x)\}]$$

- Conditional distribution for positive proportion  $Y_i/n_i$

$$Y_i | S(\cdot) \sim \text{Bin}\{n_i, p(x_i)\}$$

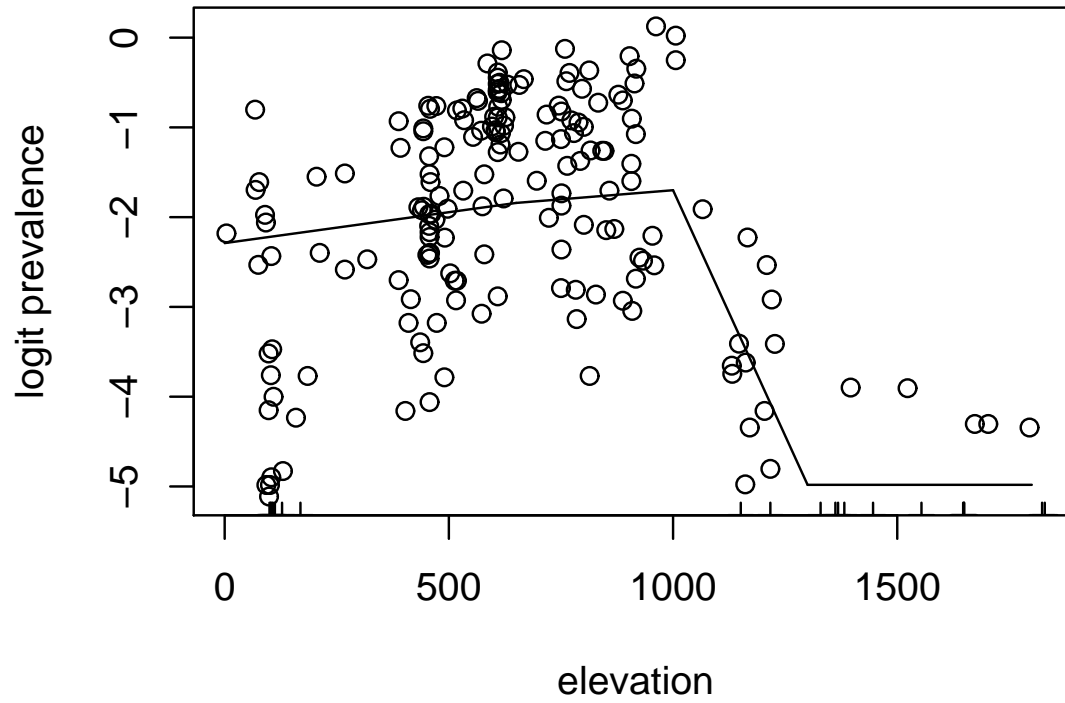
# Schematic representation of *Loa loa* model



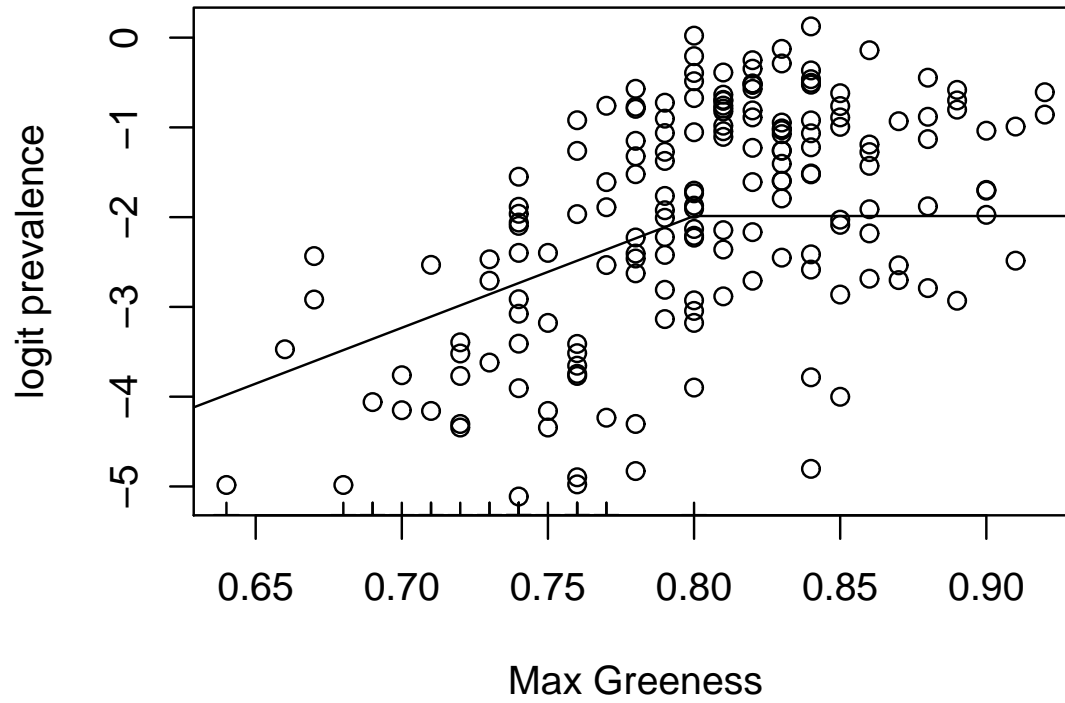
# The modelling strategy

- use relationship between environmental variables and ground-truth prevalence to construct preliminary predictions via logistic regression
- use local deviations from regression model to estimate smooth residual spatial variation
- Bayesian paradigm for quantification of uncertainty in resulting model-based predictions

# logit prevalence vs elevation

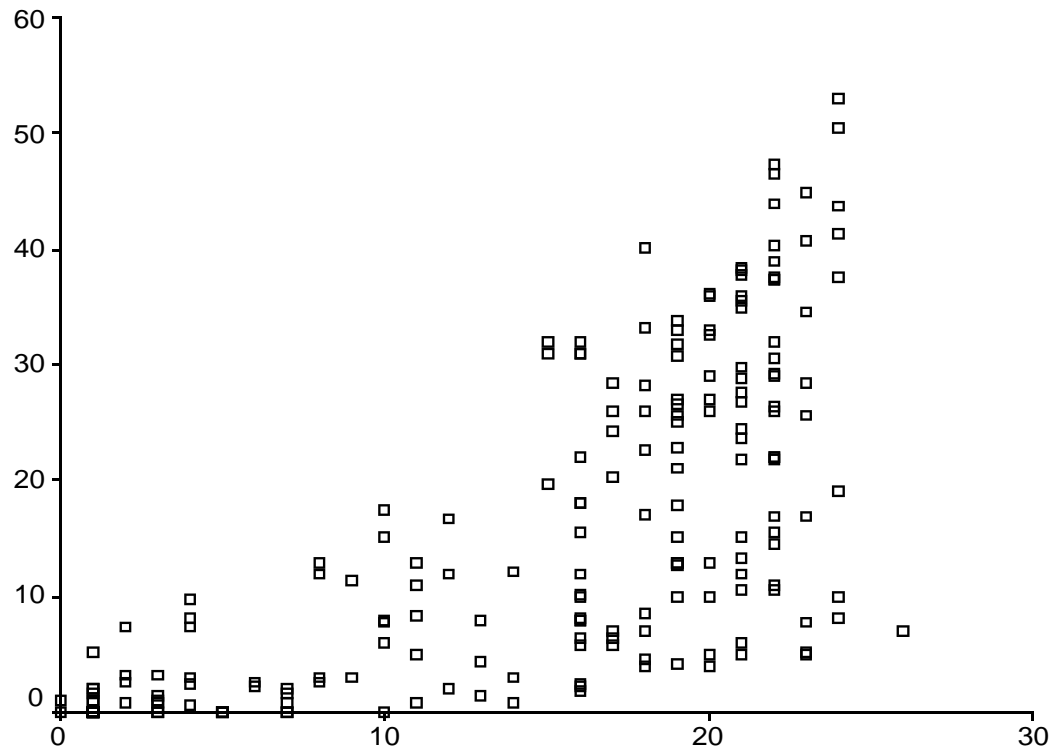


# logit prevalence vs max NDVI

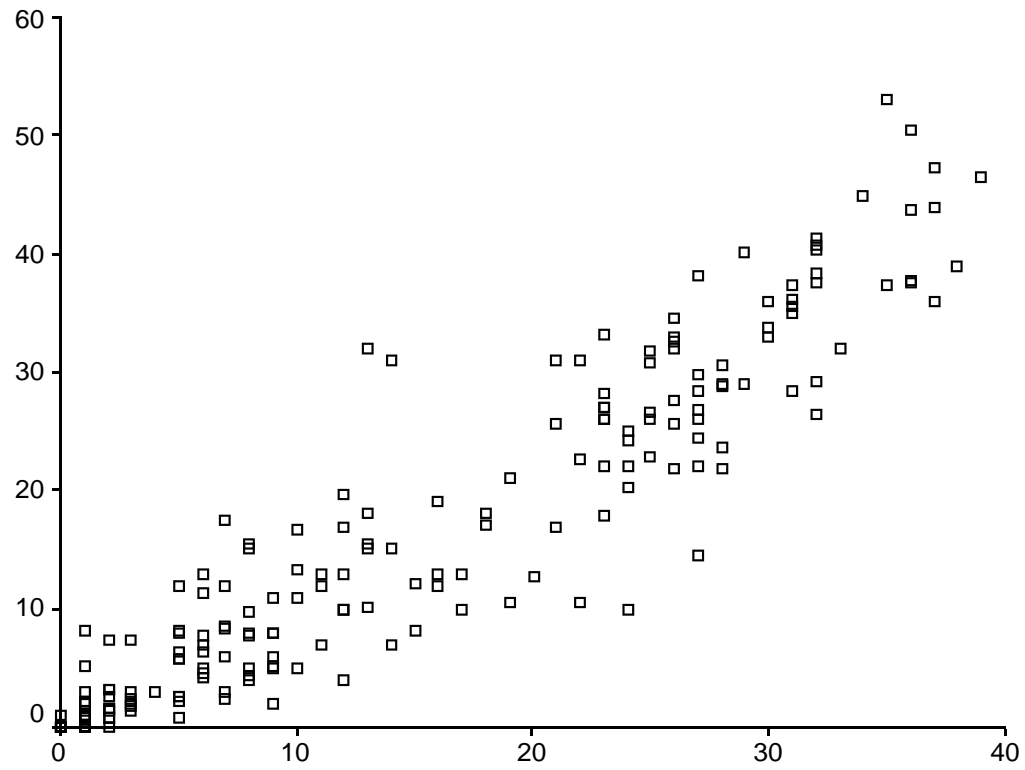


# Comparing non-spatial and spatial predictions in Cameroon

## Non-spatial



# Spatial



# Probabilistic prediction in Cameroon

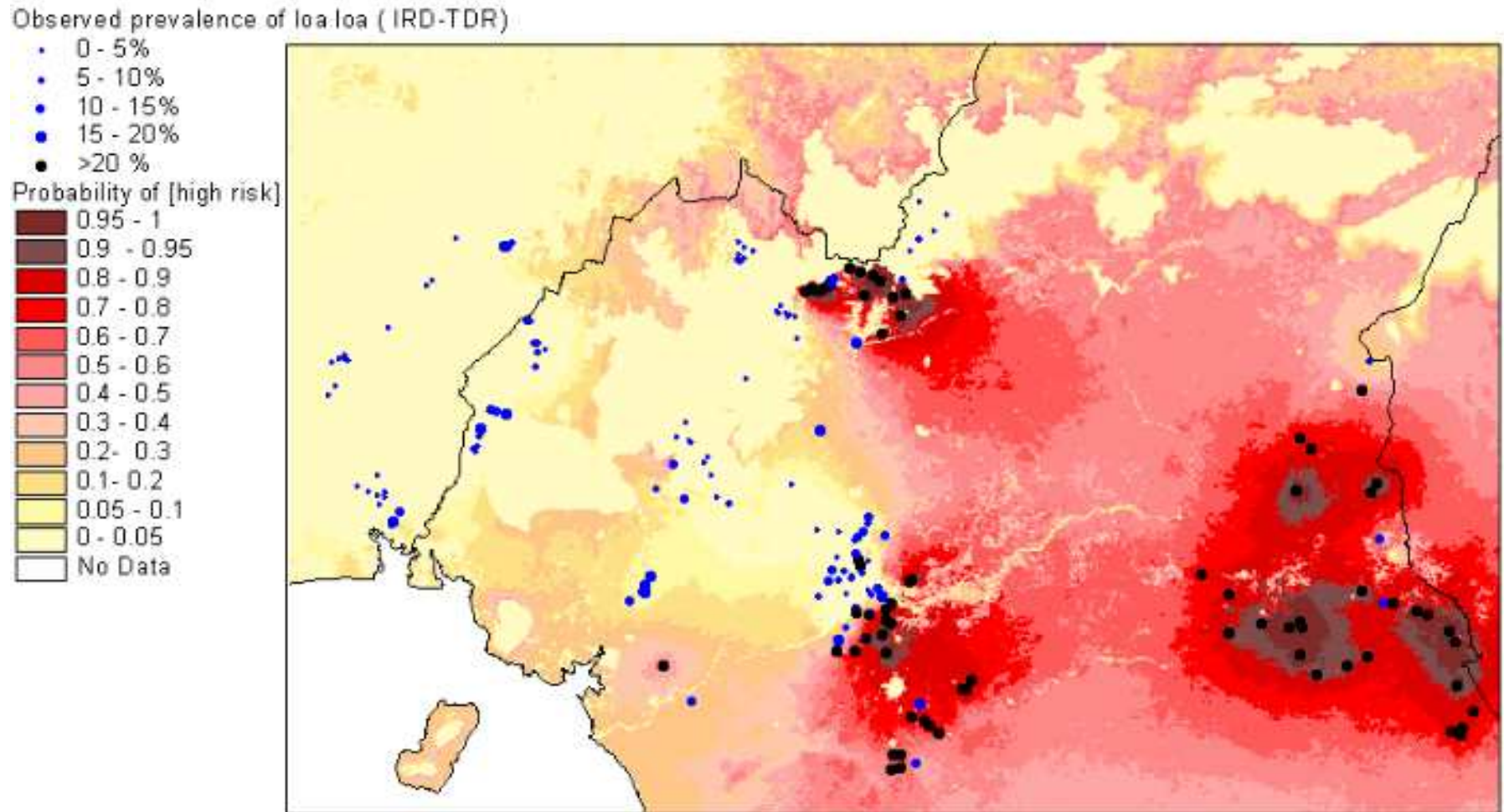


Figure 6: *PCM for [high risk] in Cameroon based on  $ERM_r$  with ground truth data.*

# Geostatistics re-visited

locations  $X$       signal  $S$       measurements  $Y$

- Conventional geostatistical model:  $[S, Y] = [S][Y|S]$
- What if  $X$  is stochastic?

Usual implicit assumption:  $[X, S, Y] = [X][S][Y|S]$

Hence, can ignore  $[X]$  for likelihood-based inference about  $[S, Y]$ .

$$L(\theta) = \int [S][Y|S]dS$$

# Marked point processes

locations  $X$       marks  $Y$

- $X$  is a point process
- $Y$  need only be defined at points of  $X$
- natural factorisation of  $[X, Y]$  depends on scientific context

$$[X, Y] = [X][Y|X] = [Y][X|Y]$$

# Preferential sampling

locations  $X$       signal  $S$       measurements  $Y$

- Conventional model:

$$[X, S, Y] = [S][X][Y|S] \quad (1)$$

- Preferential sampling model:

$$[X, S, Y] = [S][X|S][Y|S, X] \quad (2)$$

Under model (2), typically  $[Y|S, X] = [Y|S_0]$  where  $S_0 = S(X)$  denotes the values of  $S$  at the points of  $X$

# An idealised model for preferential sampling

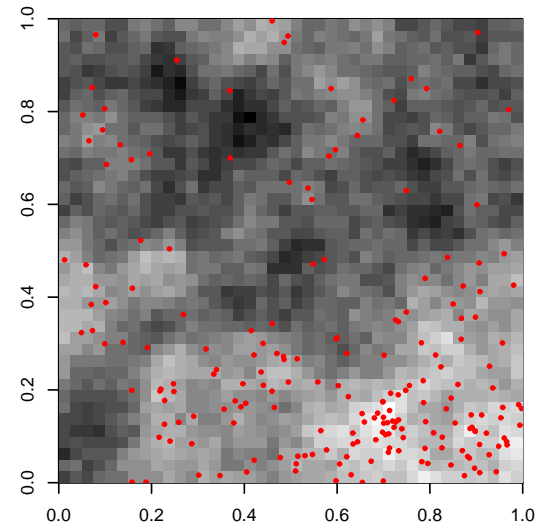
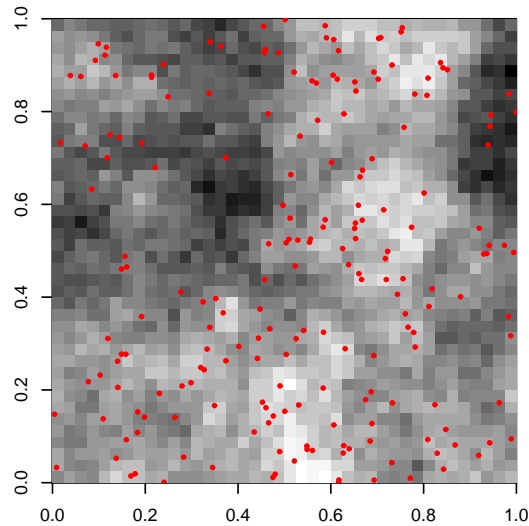
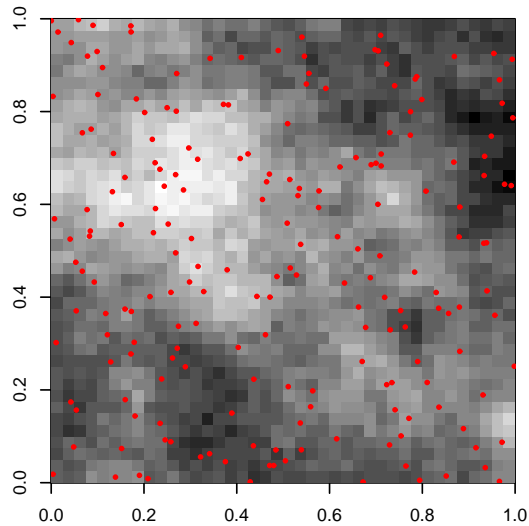
$$[X, S, Y] = [S][X|S][Y|S, X]$$

- $[S] = \text{SGP}(\mu, \sigma^2, \rho)$  (stationary Gaussian process)
- $[X|S] =$  inhomogenous Poisson process with intensity

$$\lambda(x) = \exp\{\alpha + \beta S(x)\}$$

- $[Y|S, X] = \prod_{i=1}^n [Y_i|S(X_i)]$
- $[Y_i|S(X_i)] = \text{N}(S(X_i), \tau^2)$

# Simulation of preferential sampling model



$\beta = 0.0, 0.25, 0.5$

# Impact of preferential sampling on spatial prediction

- target for prediction is  $S(x)$ ,  $x = (0.5, 0.5)$
- 100 data-locations on unit square
- three sampling designs

---

	Sampling design		
	uniform	clustered	preferential
bias	(-0.081, 0.059)	(-0.082, 0.186)	(1.290, 1.578)
MSE	(0.268, 0.354)	(0.948, 1.300)	(2.967, 3.729)

---

# Likelihood inference (crude Monte Carlo)

$$[X, S, Y] = [S][X|S][Y|S, X]$$

- data are  $X$  and  $Y$ , likelihood is

$$L(\theta) = \int [X, S, Y] dS = \mathbf{E}_S [[X|S][Y|S, X]]$$

- evaluate expectation by Monte Carlo,

$$L_{MC}(\theta) = m^{-1} \sum_{j=1}^m [X|S_j][Y|S_j, X],$$

using anti-thetic pairs,  $S_{2j} = -S_{2j-1}$

# An importance sampler

Re-write likelihood as

$$L(\theta) = \int [X|S][Y|X, S] \frac{[S|Y]}{[S|Y]} [S] dS$$

- $[S] = [S_0][S_1|S_0]$
- $[S|Y] = [S_0|Y][S_1|S_0, Y] = [S_0|Y][S_1|S_0]$
- $[Y|X, S] = [Y|S_0]$

$\Rightarrow$

$$\begin{aligned} L(\theta) &= \int [X|S] \frac{[Y|S_0]}{[S_0|Y]} [S_0][S|Y] dS \\ &= \mathbf{E}_{S|Y} \left[ [X|S] \frac{[Y|S_0]}{[S_0|Y]} [S_0] \right] \end{aligned}$$

## An importance sampler (continued)

- simulate  $S_j \sim [S|Y]$  (anti-thetic pairs)
- if  $Y$  is measured without error, set  $[Y|S_{0j}]/[S_{0j}|Y] = 1$

Monte Carlo approximation is:

$$L_{MC}(\theta) = m^{-1} \sum_{j=1}^m \left[ [X|S_j] \frac{[Y|S_{0j}]}{[S_{0j}|Y]} [S_{0j}] \right]$$

# Practical solutions to weak identifiability

1. explanatory variables  $U$  to break dependence between  $S$  and  $X$
2. strong Bayesian priors
3. two-stage sampling

# Ozone monitoring in California

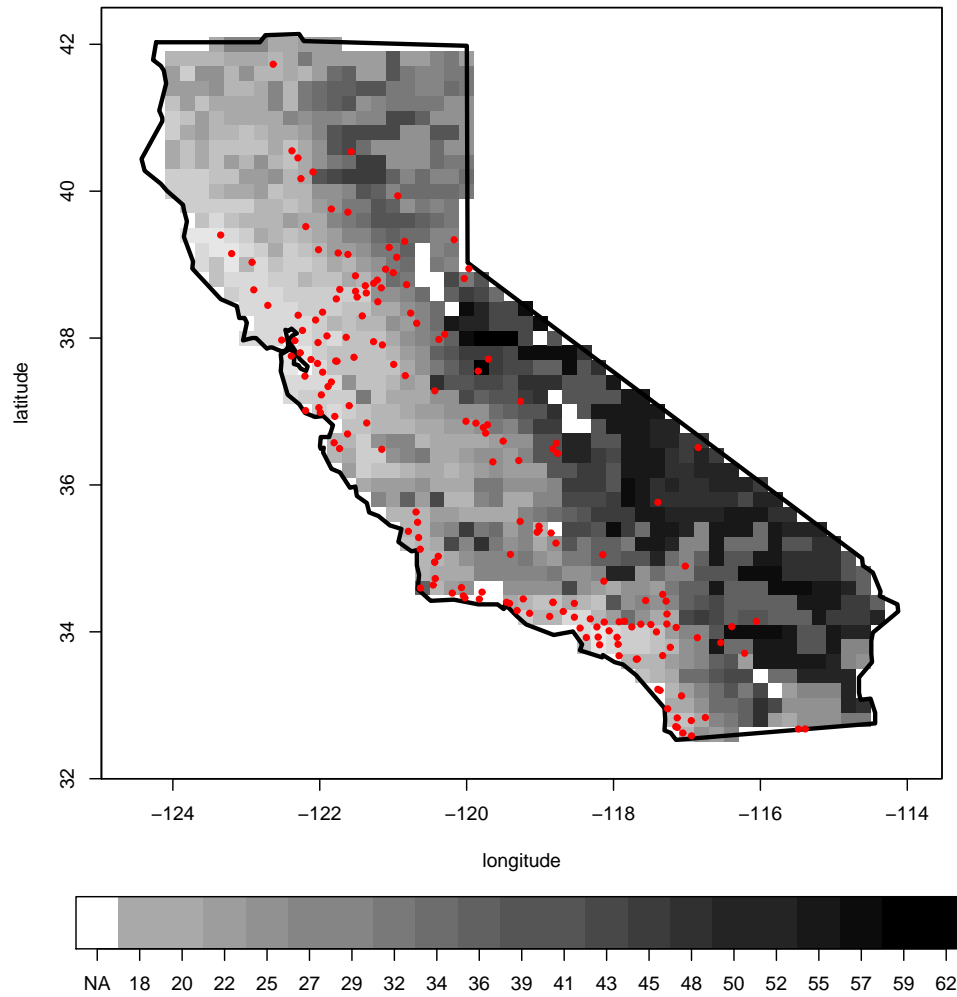
## Data:

- yearly averages of  $O_3$  from 178 monitoring locations throughout California
- census information for each of 1709 zip-codes

## Objective:

- estimate spatial average of  $O_3$  in designated sub-regions

# California ozone monitoring data



# Ozone monitoring in California (continued)

## Preferential sampling?

- highly non-uniform spatial distribution of monitors, negatively associated with levels of pollution
- may be able to allow for this if demographic and/or socio-economic factors are associated both with levels of pollution and with intensity of monitoring

# Ozone monitoring in California (continued)

## Modelling assumption

- dependence induced by latent variables  $U$ ,

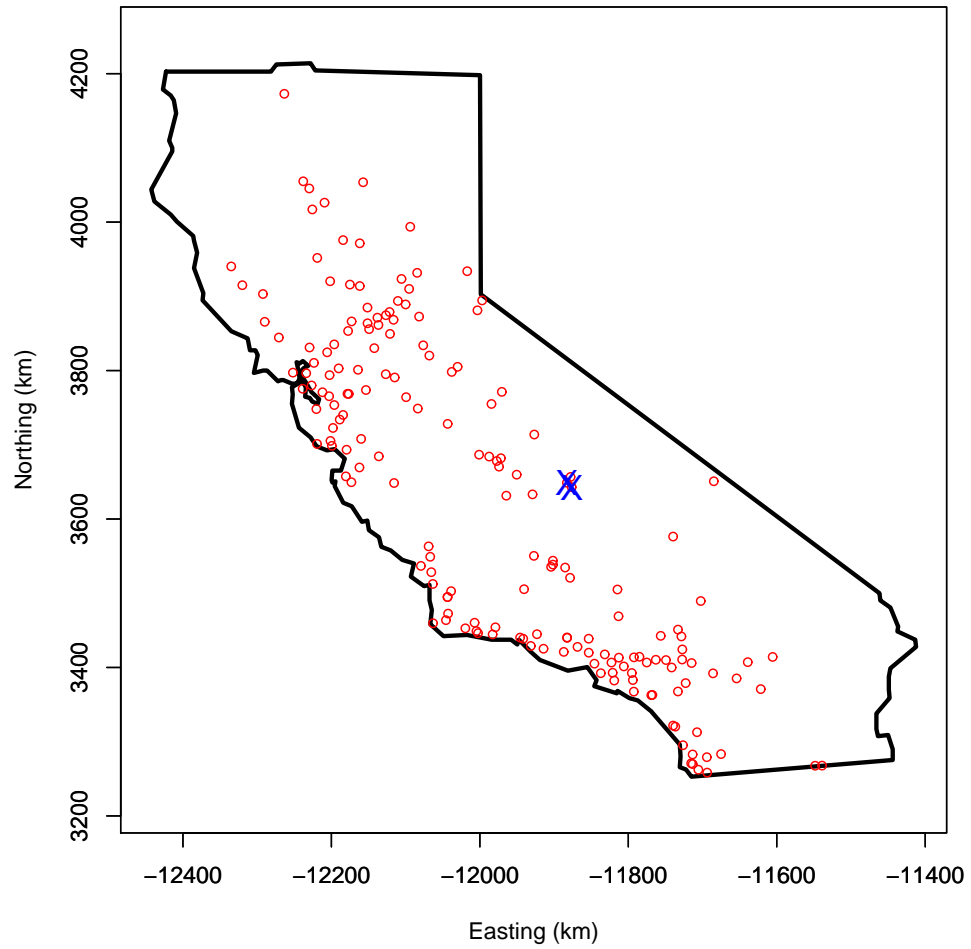
$$[X, S, Y] = \int [X|U][S|U][Y|S, U][U]dU$$

- if  $U$  observed:
  - use conditional likelihood,

$$[X, S, Y|U] = [X|U][S|U][Y|S, U]$$

- and ignore term  $[X|U]$  for inference about  $S$

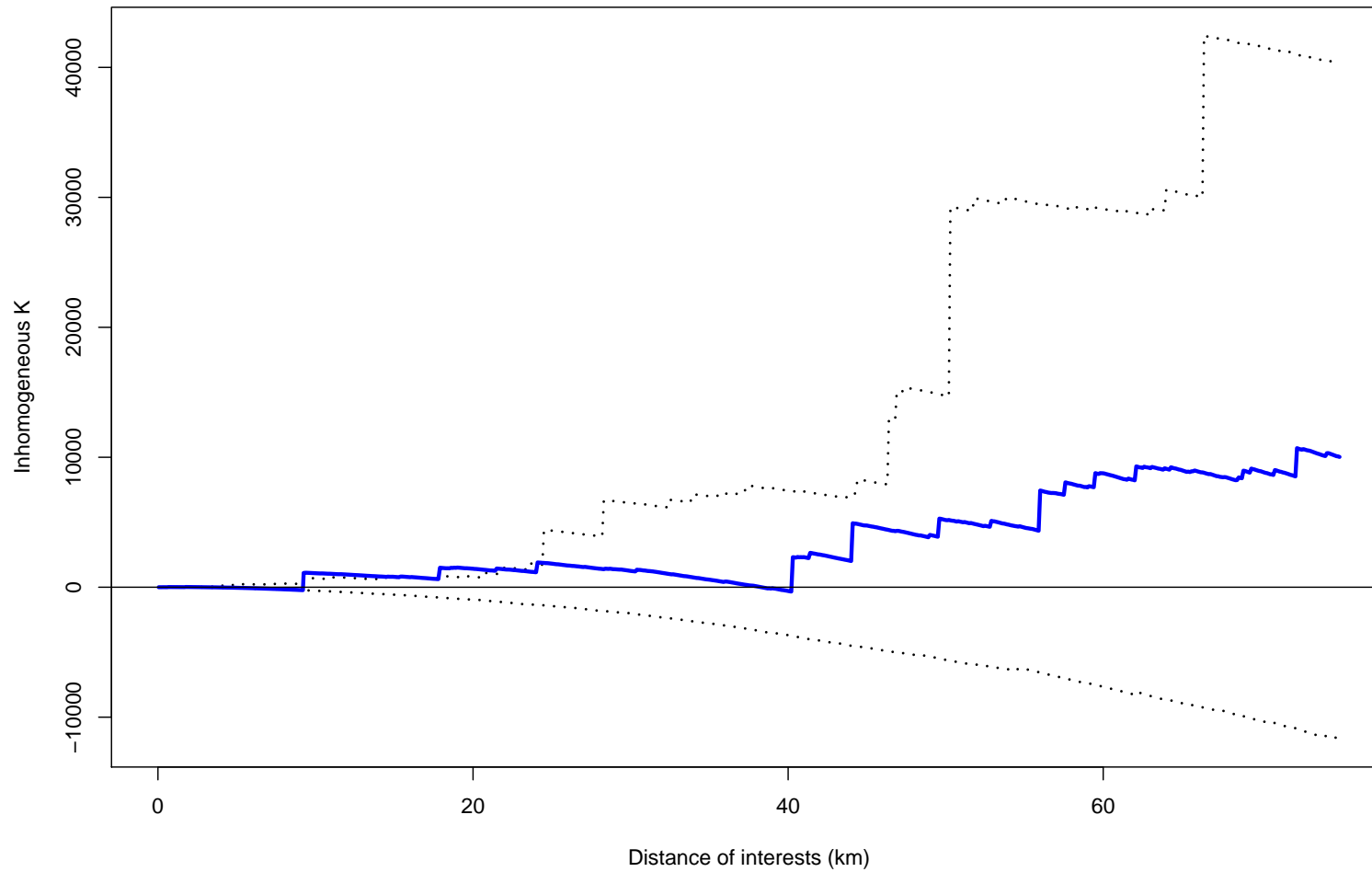
# California ozone monitors: outlier?



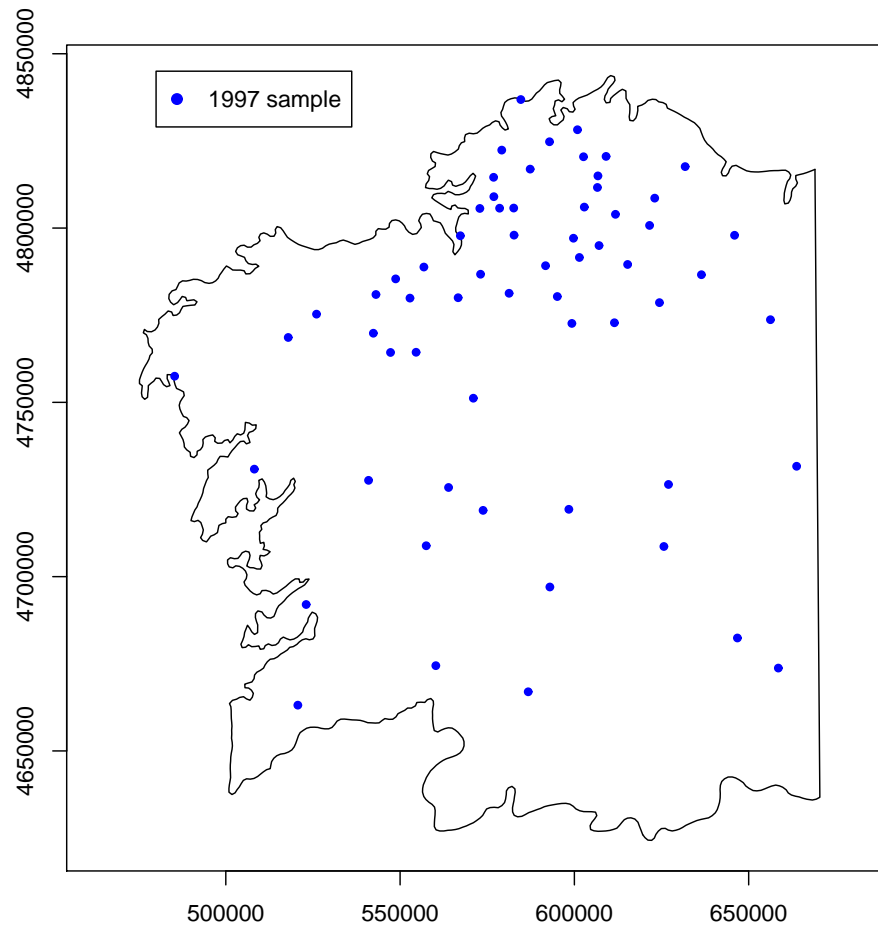
# Analysis of California ozone monitor locations

- monitor intensity associated with:
  - population density (positive)
  - percentage College-educated (positive)
  - median family income (negative)
- good fit to inhomogeneous Poisson process model (after removal of one outlier)

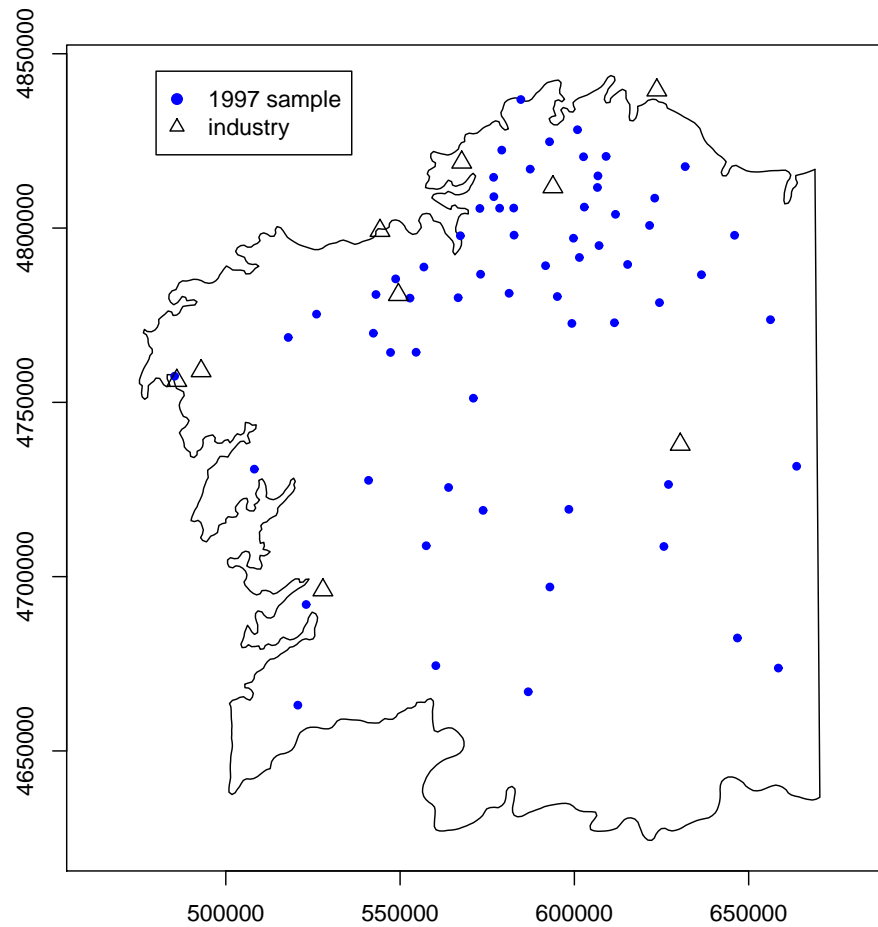
# California ozone monitors: fit to Poisson model



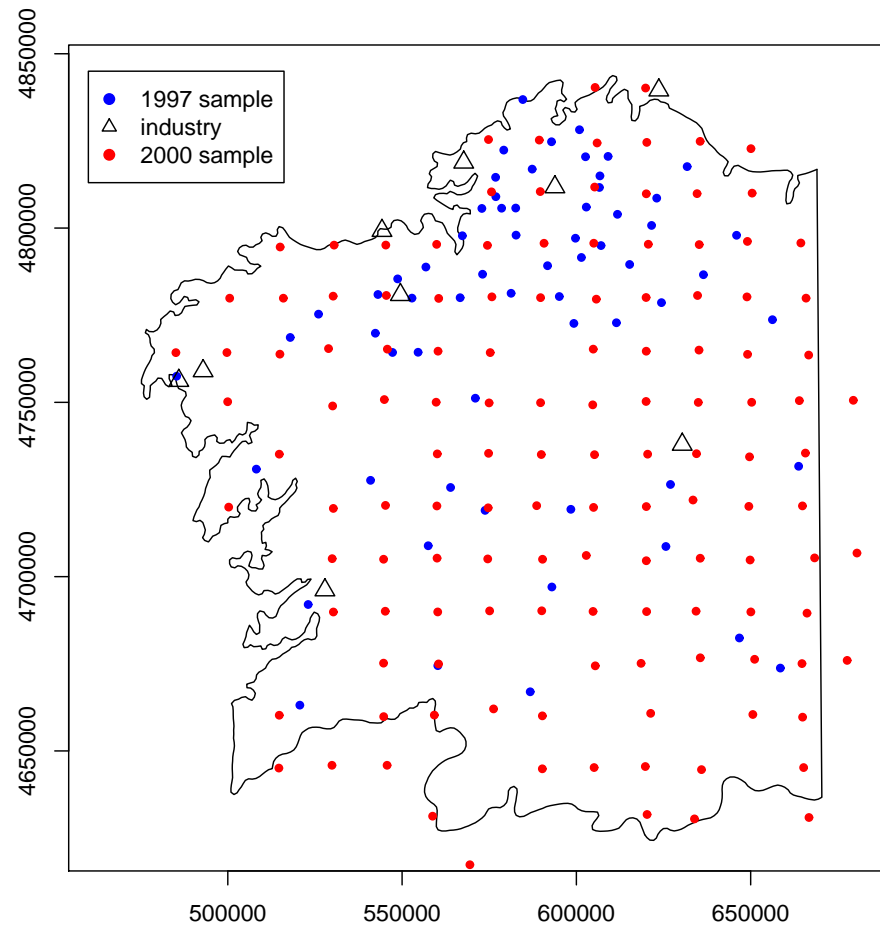
# Heavy metal bio-monitoring in Galicia



# Heavy metal bio-monitoring in Galicia



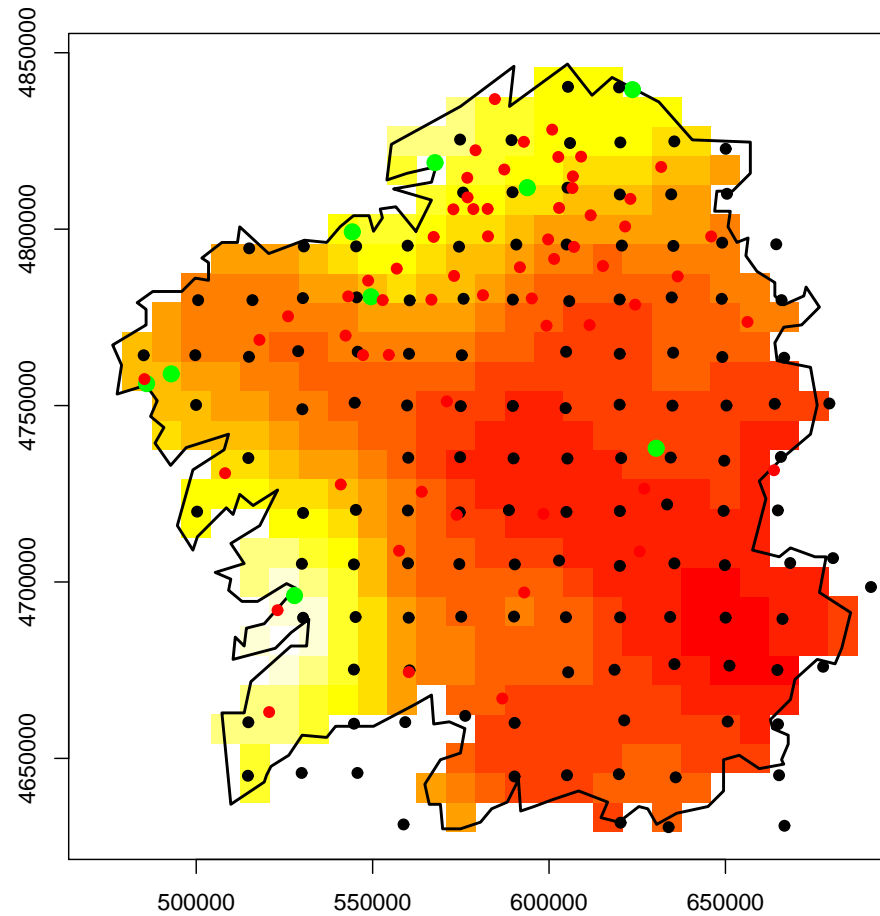
# Heavy metal bio-monitoring in Galicia



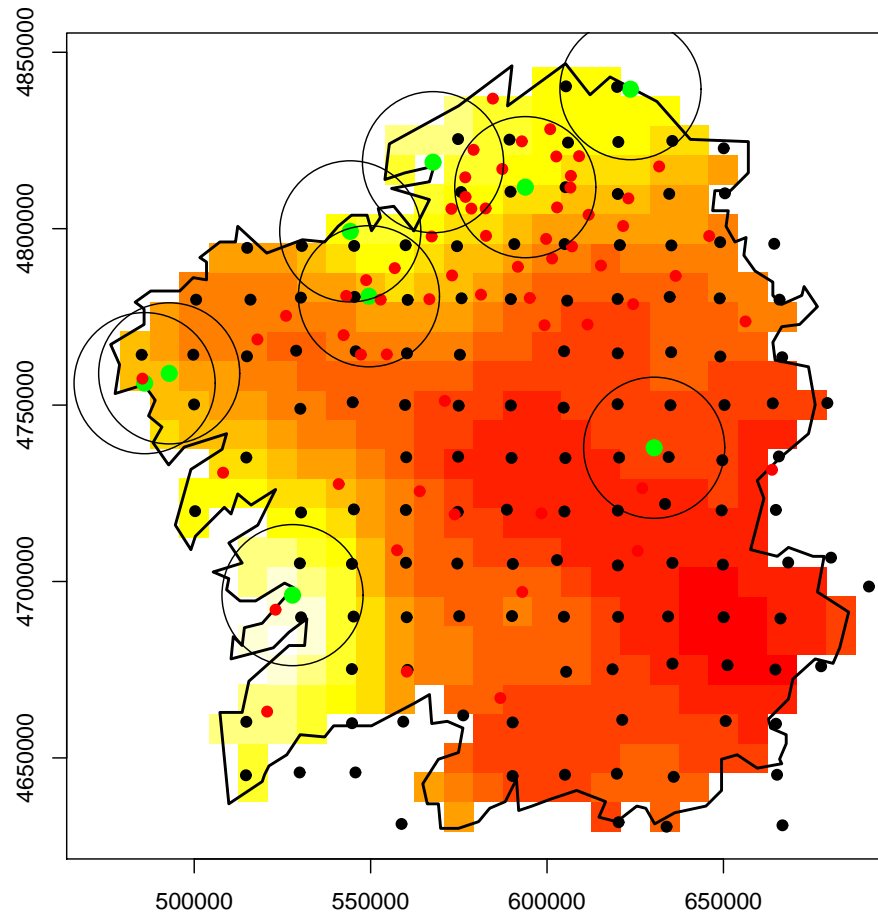
# Heavy metal bio-monitoring in Galicia

- 1997 sampling design is good for monitoring effects of industrial activity
- but would lead to potential biased estimates of residual spatial variation
- 2000 sampling design is good for fitting model of residual spatial variation
- assuming stability of pollution levels over time, possible analysis strategy is:
  - use 2000 data, or sub-set thereof, to model spatial variation
  - holding spatial correlation parameters fixed, use 1997 data to model point-source effects of industrial locations.

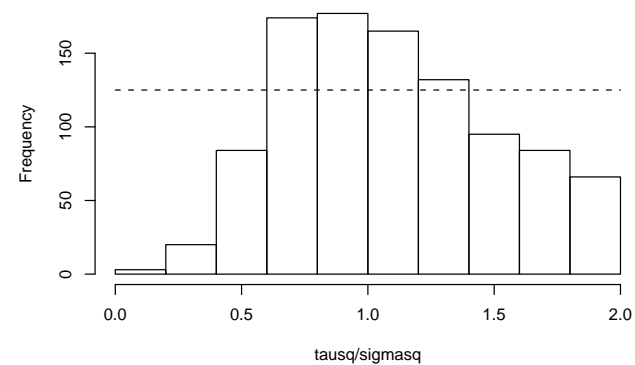
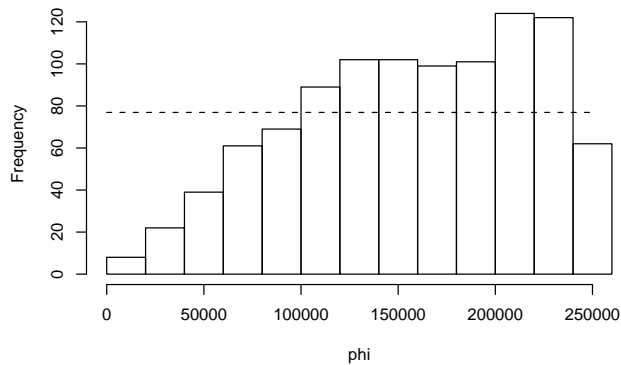
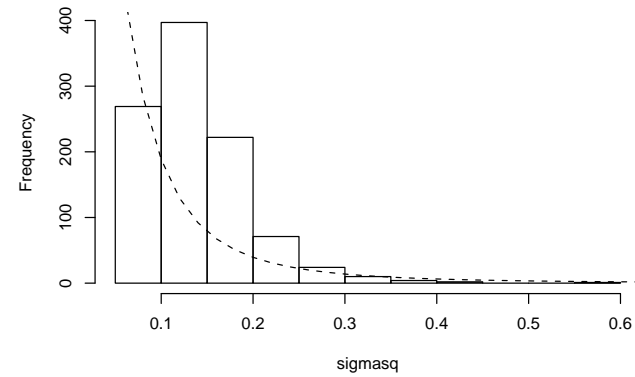
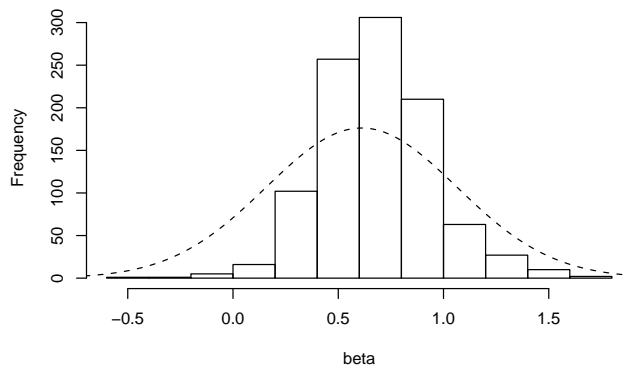
# Galicia: 2000 predictions (posterior mean)



# Galicia: excision of areas close to industry



# Galicia: posteriors from analysis of 2000 data



$$\mathbf{E}[S(x)] = \mu_0 \quad V(u) = \tau^2 + \sigma^2 \{1 - \exp(-u/\phi)\}$$

## Galicia: analysis of 1997 data

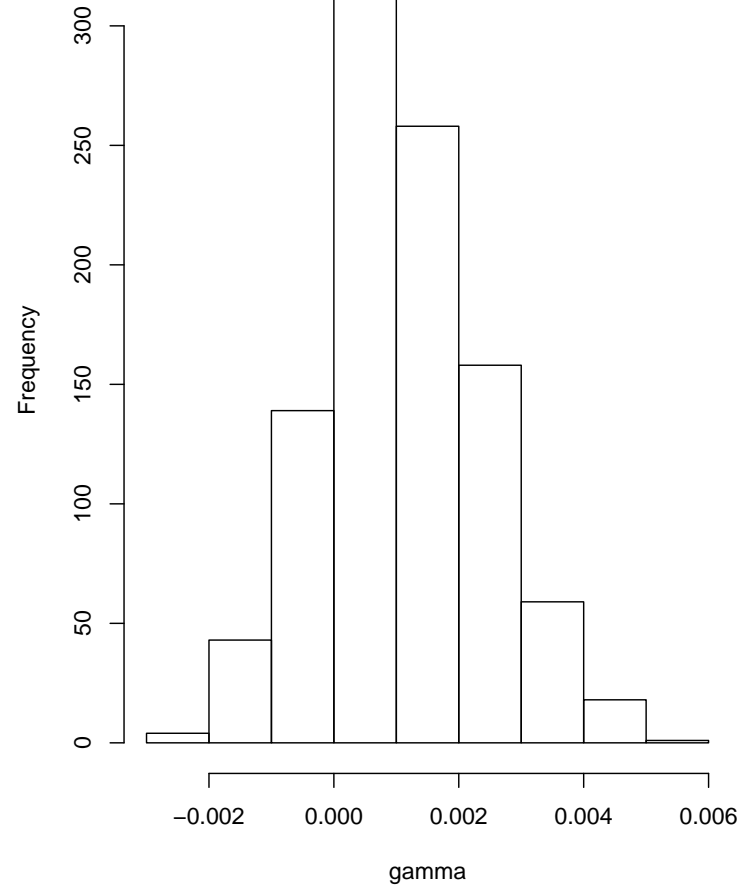
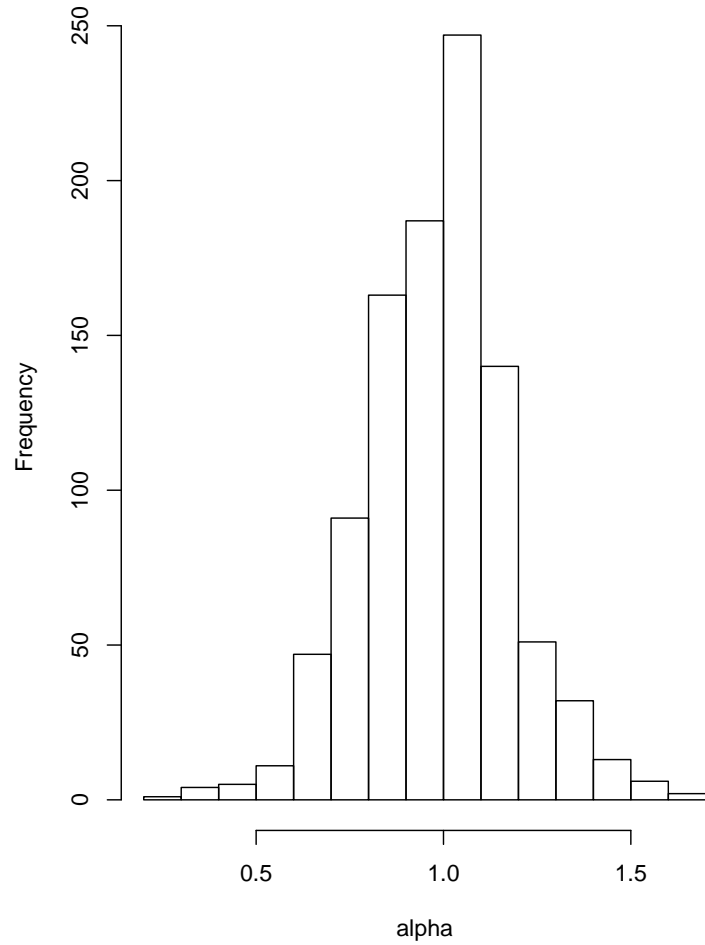
- introduce distance to nearest industry as explanatory variable,

$$\mu(x) = \mu_0 + \alpha + \gamma d(x)$$

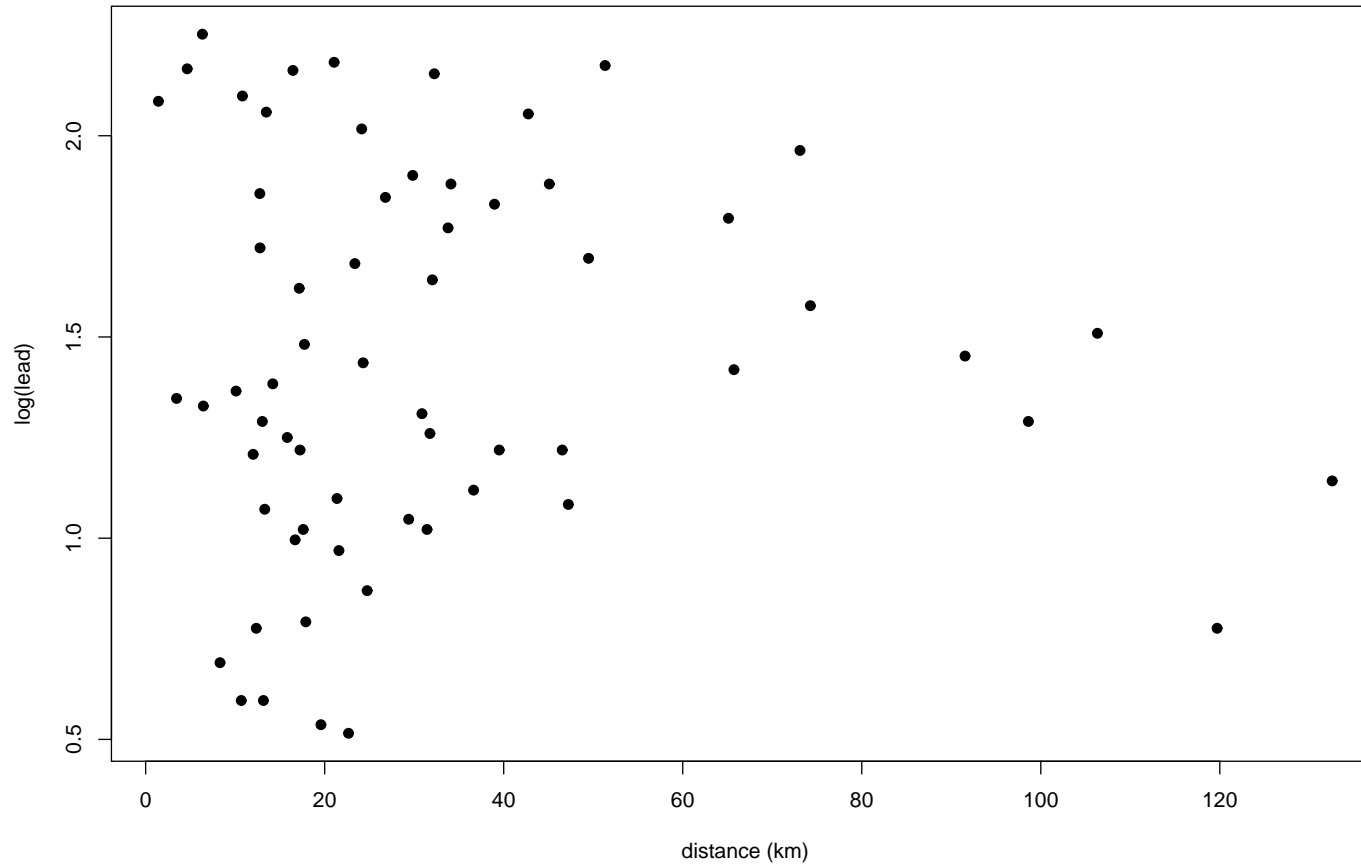
- for  $\beta$  and spatial covariance parameters, use posteriors from 2000 analysis
- resulting posterior mean and SD for  $(\alpha, \gamma)$

	$\alpha$	$\gamma$
mean	0.601	-0.000561
SD	0.179	0.000609

# Galicia: posteriors for $(\alpha, \gamma)$



# Galicia: a cautionary note



**Suggests missing explanatory variable(s)?**

## Closing remarks

- preferential sampling is widespread in practice, but almost universally ignored
- its effects may or may not be innocuous
- model parameters may be poorly identified, hence
- reliance on formal likelihood-based inference for a single data-set may be unwise
- different pragmatic analysis strategies may be needed for different applications