Model-based Geostatistics

Peter J Diggle

Lancaster University and Johns Hopkins University School of Public Health

and

Paulo Justiniano Ribeiro, Jr

Department of Statistics, Universidade Federal do Paraná

Outline

1. Introduction - motivating examples
2. Linear models
3. Bayesian inference
4. Generalised linear models
5. Geostatistical design
6. Geostatistics and marked point processes

Diggle and Ribeiro (2007). Model-based Geostatistics.
New York : Springer.

Section 1

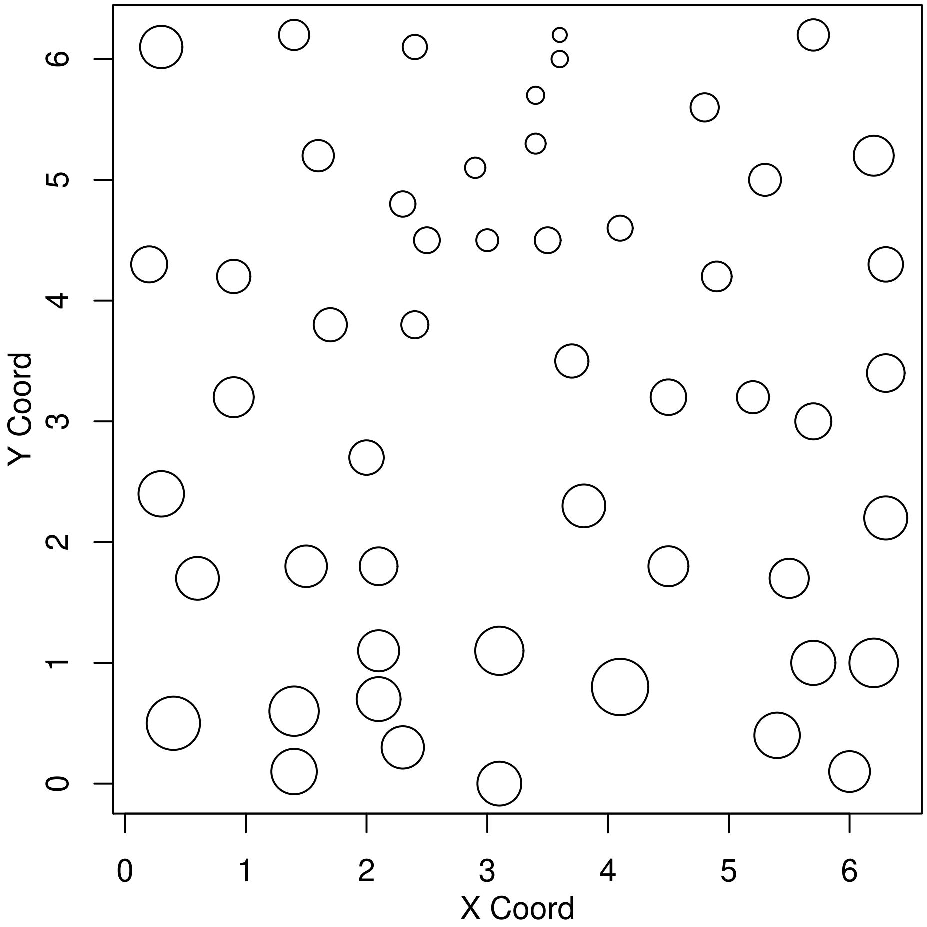
Introduction - motivating examples

Geostatistics

- traditionally, a self-contained methodology for spatial prediction, developed at École des Mines, Fontainebleau, France

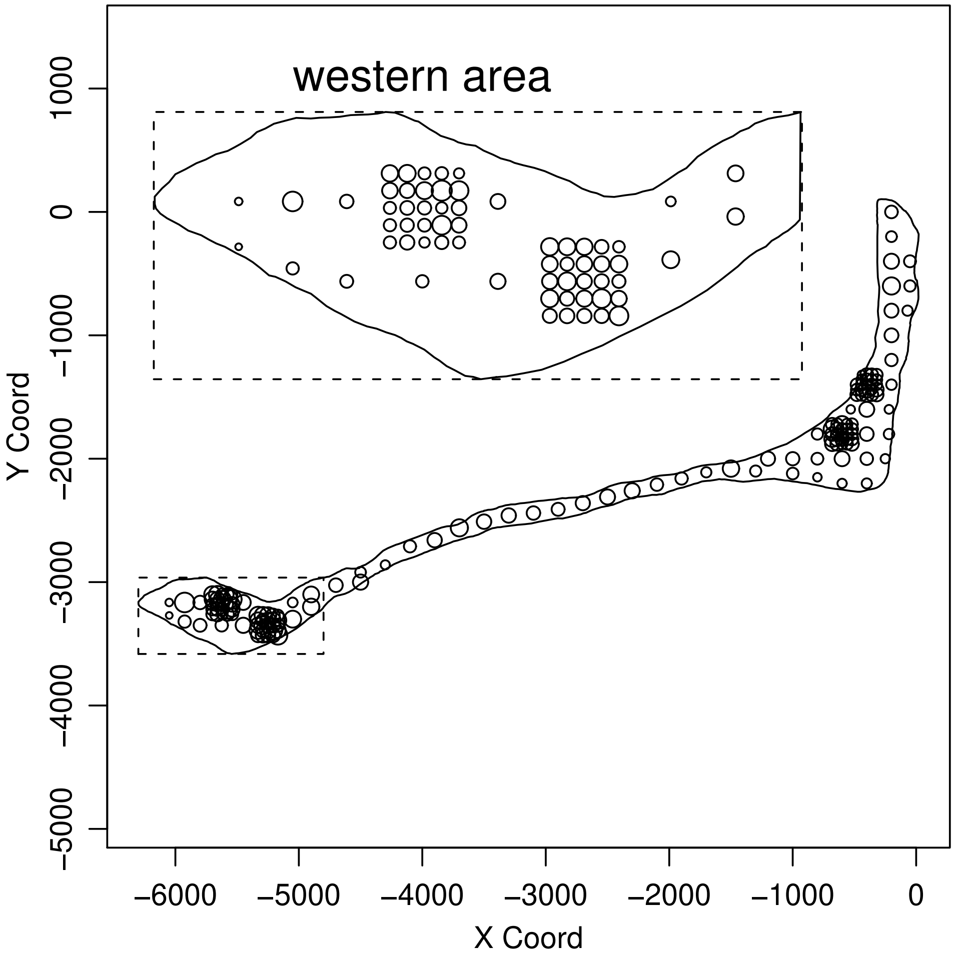
- nowadays, that part of spatial statistics which is concerned with data obtained by spatially discrete sampling of a spatially continuous process

Example 1.1: Measured surface elevations

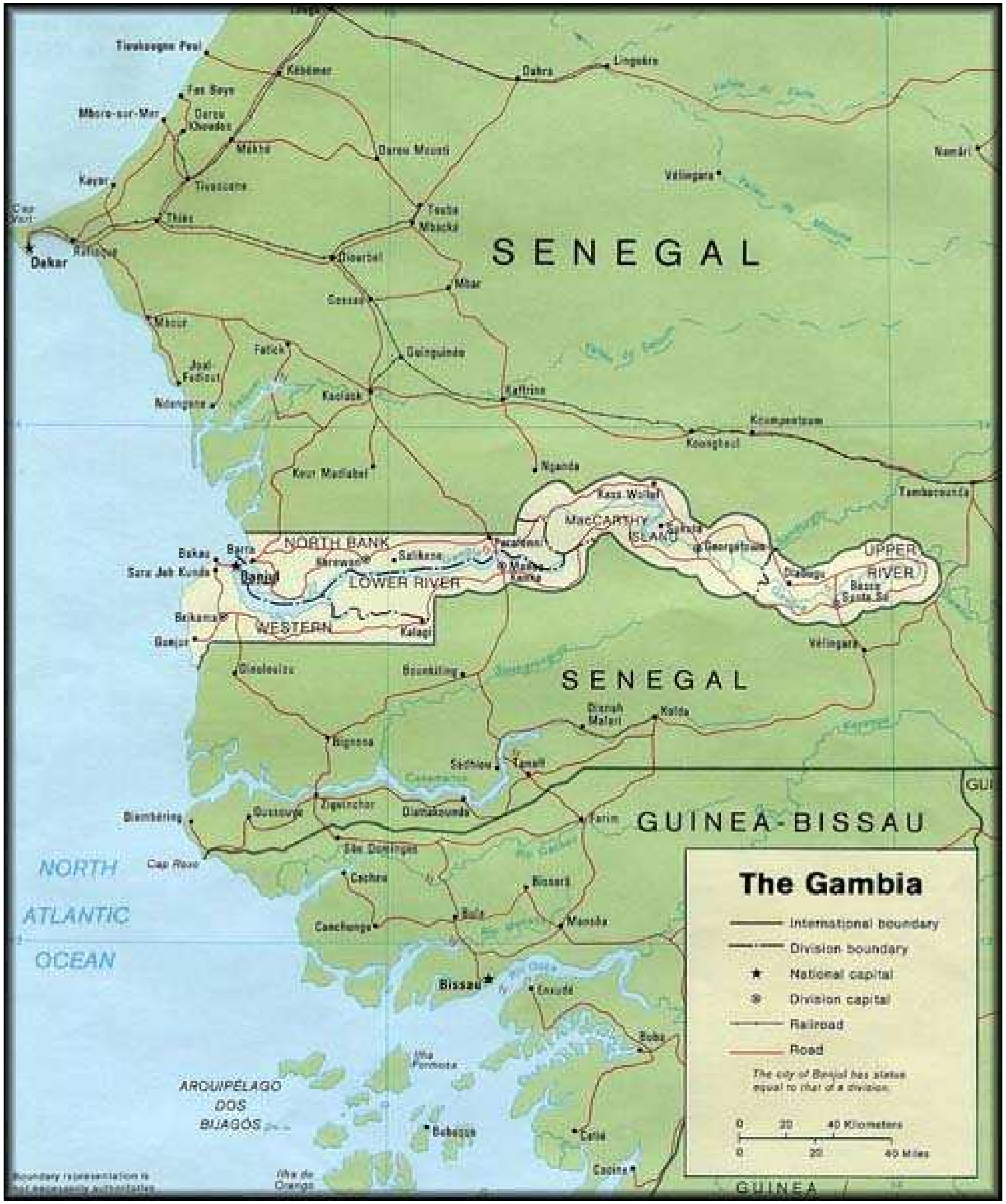


No explanatory variables?

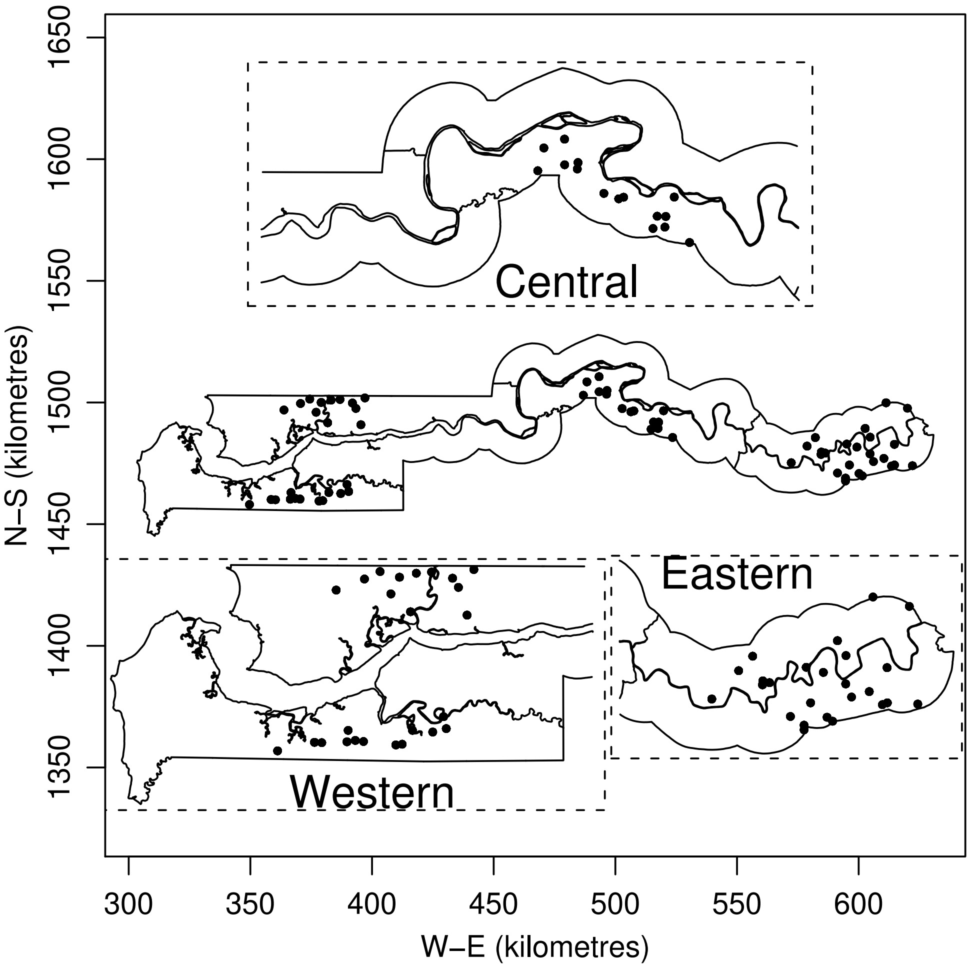
Example 1.2: Residual contamination from nuclear weapons testing



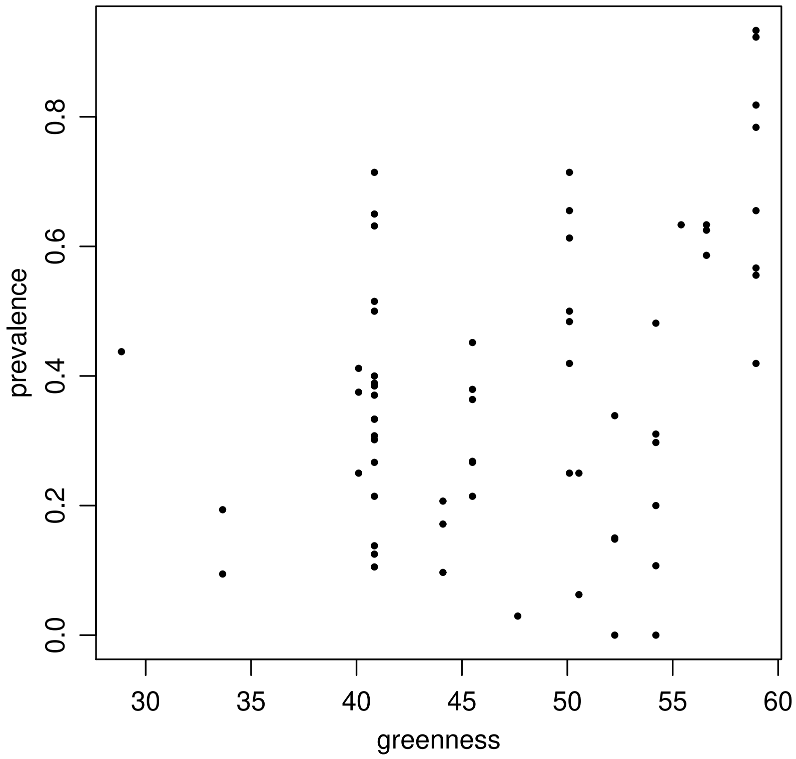
Sunset at Georgetown


SENEGAL
SENEGAL
GUINEA-BISSAU
GUINEA
NORTH ATLANTIC OCEAN
ARQUIPÉLAGO DOS BIJAGÓS
Dakar
Rufisque
Thiès
Tivaouane
Kébémer
Louga
Dahra
Linguère
Diourbel
Touba
Mbacké
Mbar
Gossas
Fatick
Kaolack
Kaffrine
Koungheul
Koumpentoum
Tambacounda
Nganda
Keur Madiabel
Mbour
Joal-Fadiout
Ndangane
Kayar
Mboro-sur-Mer
Fas Boye
Darou Khoudos
Mékhé
Tiwaouane Peul
Darou Mounti
Vélingara
Namâri
Guinguinéo
NORTH BANK
LOWER RIVER
WESTERN
MacCARTHY ISLAND
UPPER RIVER
Banjul
Barra
Bakau
Sara Jeb Kunda
Brikama
Gunjur
Kerewan
Salikene
Farafenni
Mansa Konko
Kaiaf
Kalagi
Kuntaur
Kass Wollof
Georgetown
Diabugu
Basse Santa Su
Bansang
Vélingara
Kolda
Diaroumé
Bounkiling
Diouloulou
Bignona
Sédhiou
Tanaff
Ziguinchor
Oussouye
Diattakounda
Elinkine
Cap Roxo
Farim
São Domingos
Cacheu
Bissora
Bula
Canchungo
Mansôa
Bissau
Enxudé
Buba
Bubaque
Catió
Cacine
Ilha Formosa
Ilha de Orango
Boundary representation is not necessarily authoritative.
The Gambia
International boundary
Division boundary
★ National capital
⊛ Division capital
Railroad
Road
The city of Banjul has status equal to that of a division.
0 20 40 Kilometers
0 20 40 Miles

Example 1.3: Childhood malaria in Gambia

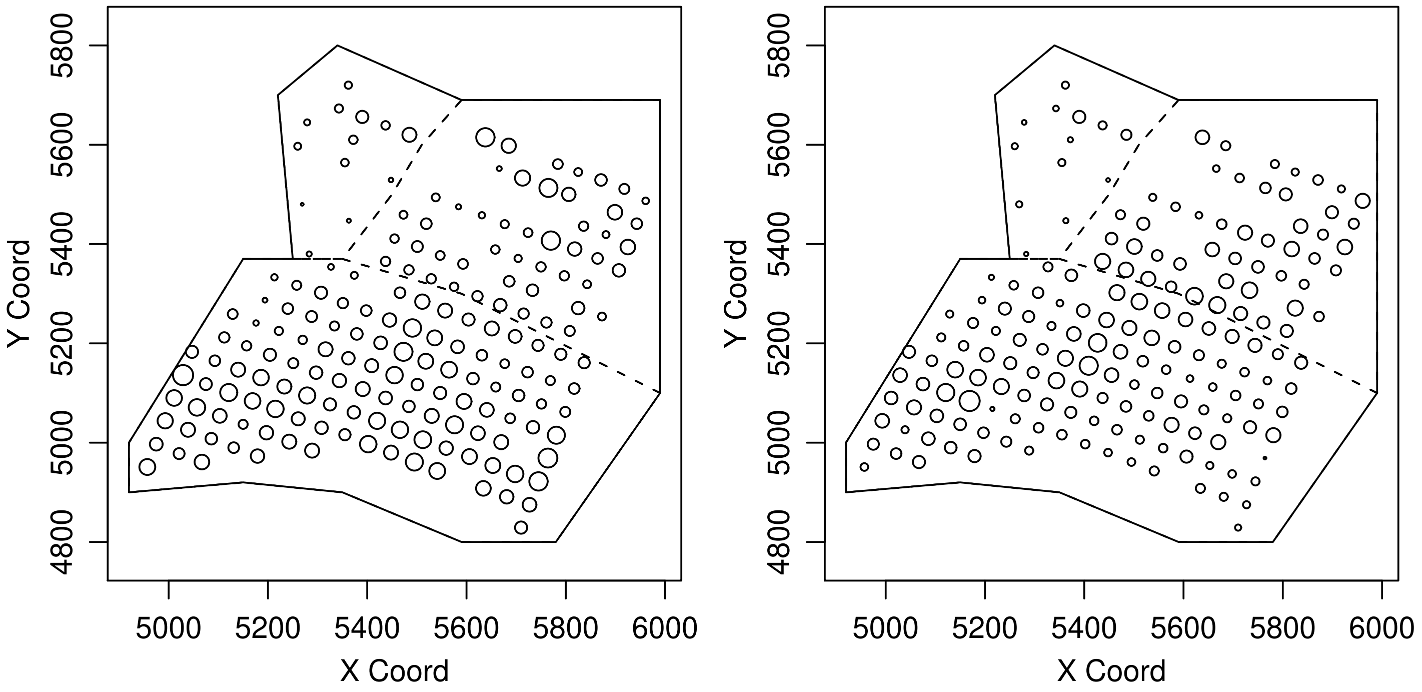


Example 1.3: continued



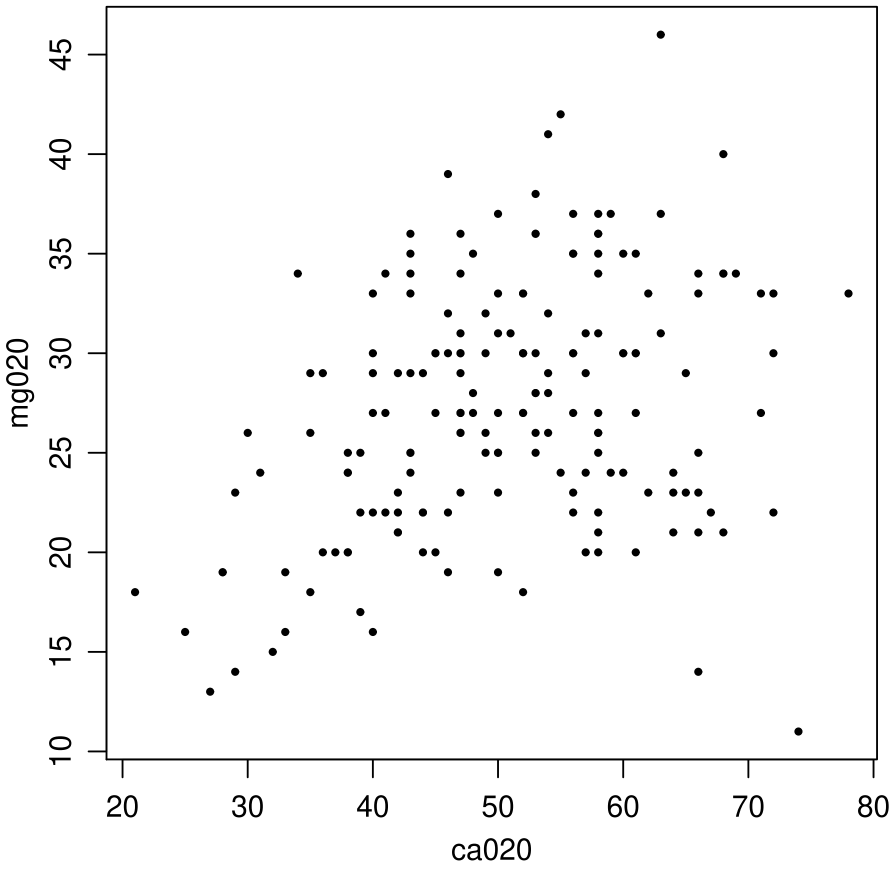
Correlation between prevalence and green-ness of vegetation

Example 1.4: Soil data



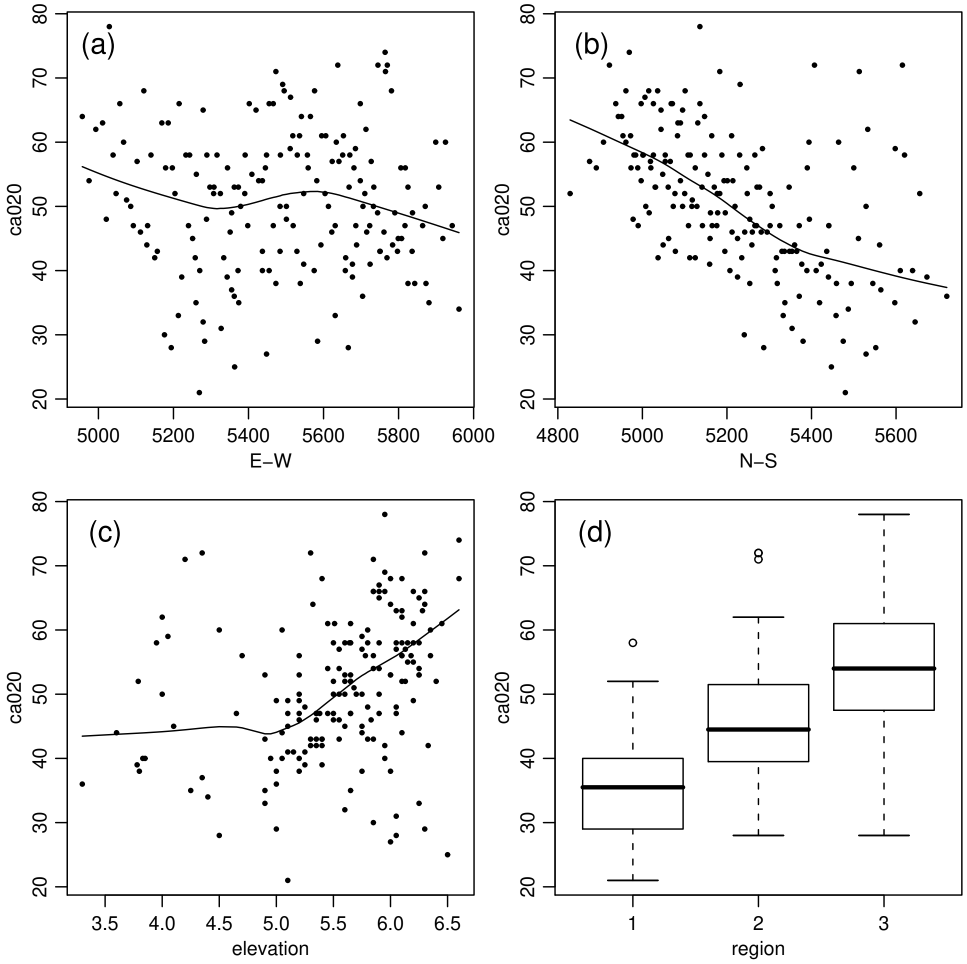
Ca (left-panel) and Mg (right-panel) concentrations

Example 1.4: Continued



Correlation between local Ca and Mg concentrations.

Example 1.4: Continued



Covariate relationships for Ca concentrations.

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**

Section 2

Linear models

Notation

- $Y = \{Y_i : i = 1, ..., n\}$ is the measurement data
- $\{x_i : i = 1, ..., n\}$ is the sampling design (note lower case)
- $Y = \{Y(x) : x \in A\}$ is the measurement process
- $S^* = \{S(x) : x \in A\}$ is the signal process
- $T = \mathcal{F}(S)$ is the target for prediction
- $[S^*, Y] = [S^*][Y|S^*]$ is the geostatistical model

Gaussian model-based geostatistics

Model specification:

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

Minimum mean square error prediction

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

- $\hat{T} = t(Y)$ is a point predictor
- $\text{MSE}(\hat{T}) = \text{E}[(\hat{T} - T)^2]$

Theorem: $MSE(\hat{T})$ takes its minimum value when $\hat{T} = \text{E}(T|Y)$.

Proof uses result that for any predictor $\tilde{T}$,

$$\text{E}[(T - \tilde{T})^2] = \text{E}_Y[\text{Var}_T(T|Y)] + \text{E}_Y\{[\text{E}_T(T|Y) - \tilde{T}]^2\}$$

Immediate corollary is that

$$\text{E}[(T - \hat{T})^2] = \text{E}_Y[\text{Var}(T|Y)] \approx \text{Var}(T|Y)$$

Simple and ordinary kriging

Recall Gaussian model:

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

Gaussian model implies

$$Y \sim \mathrm{MVN}(\mu 1, \sigma^2 V)$$

$$V = R + (\tau^2/\sigma^2) I \qquad R_{ij} = \rho(\|x_i - x_j\|)$$

Target for prediction is $T = S(x)$, write $r = (r_1, ..., r_n)$ where

$$r_i = \rho(\|x - x_i\|)$$

Standard results on multivariate Normal then give $[T|Y]$ as multivariate Gaussian with mean and variance

$$\hat{T} = \mu + r' V^{-1} (Y - \mu 1) \tag{1}$$

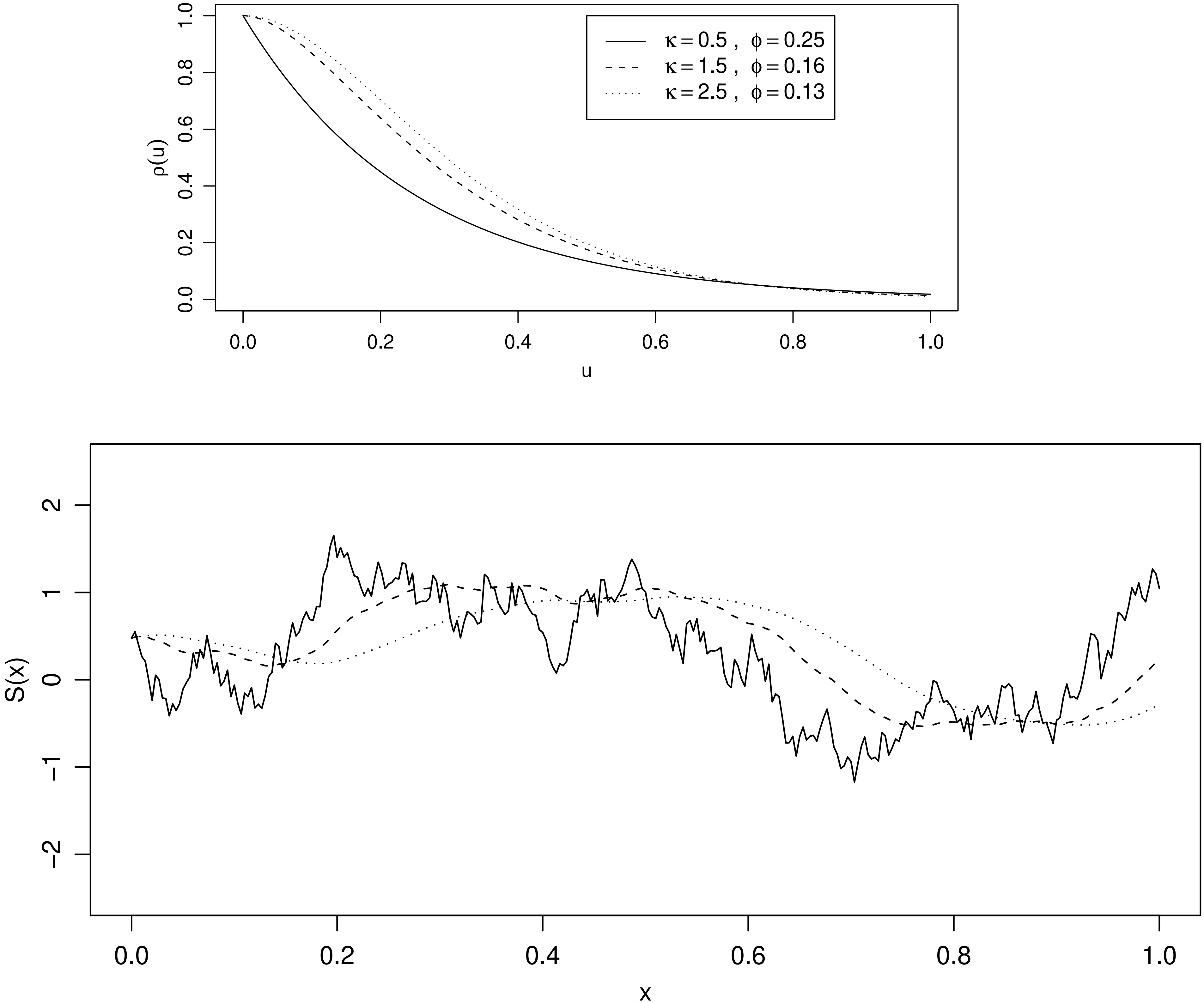
$$\mathrm{Var}(T|Y) = \sigma^2 (1 - r' V^{-1} r). \tag{2}$$

Simple kriging: $\hat{\mu} = \bar{Y}$ Ordinary kriging: $\hat{\mu} = (1' V^{-1} 1)^{-1} 1' V^{-1} Y$

The Matérn family of correlation functions

$$\rho(u) = 2^{\kappa-1}(u/\phi)^{\kappa} K_{\kappa}(u/\phi)$$

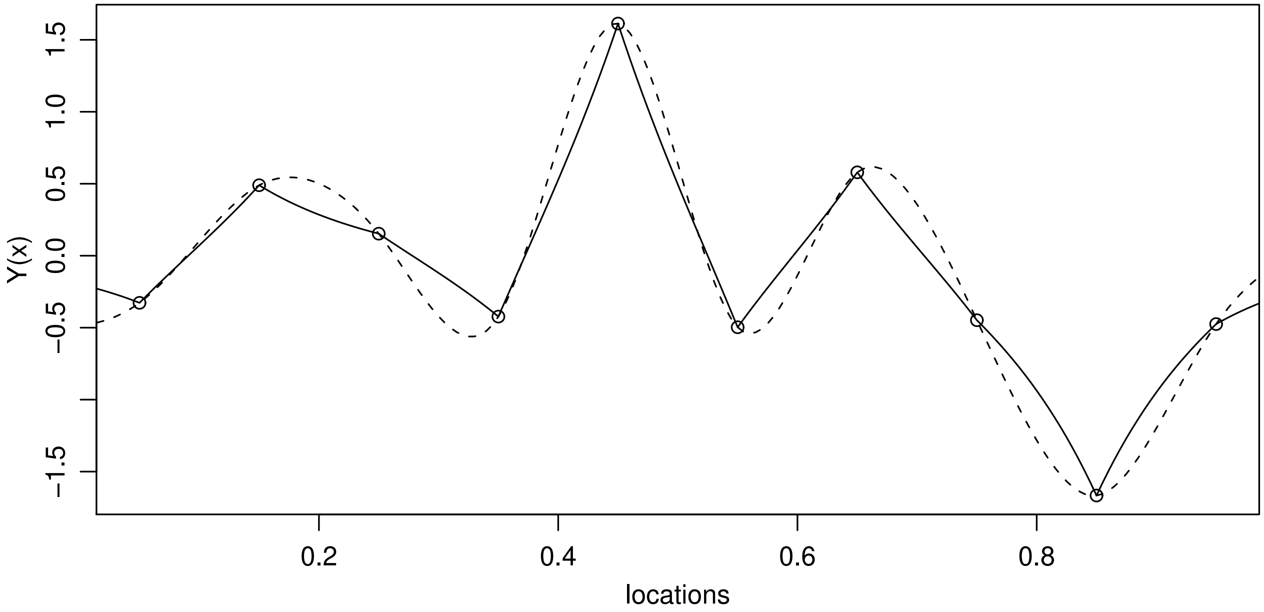
- parameters $\kappa > 0$ and $\phi > 0$
- $K_{\kappa}(\cdot)$: modified Bessel function of order κ
- $\kappa = 0.5$ gives $\rho(u) = \exp\{-u/\phi\}$
- $\kappa \to \infty$ gives $\rho(u) = \exp\{-(u/\phi)^2\}$
- κ and ϕ are not orthogonal:
  - helpful re-parametrisation: $\alpha = 2\phi\sqrt{\kappa}$
  - but estimation of κ is difficult



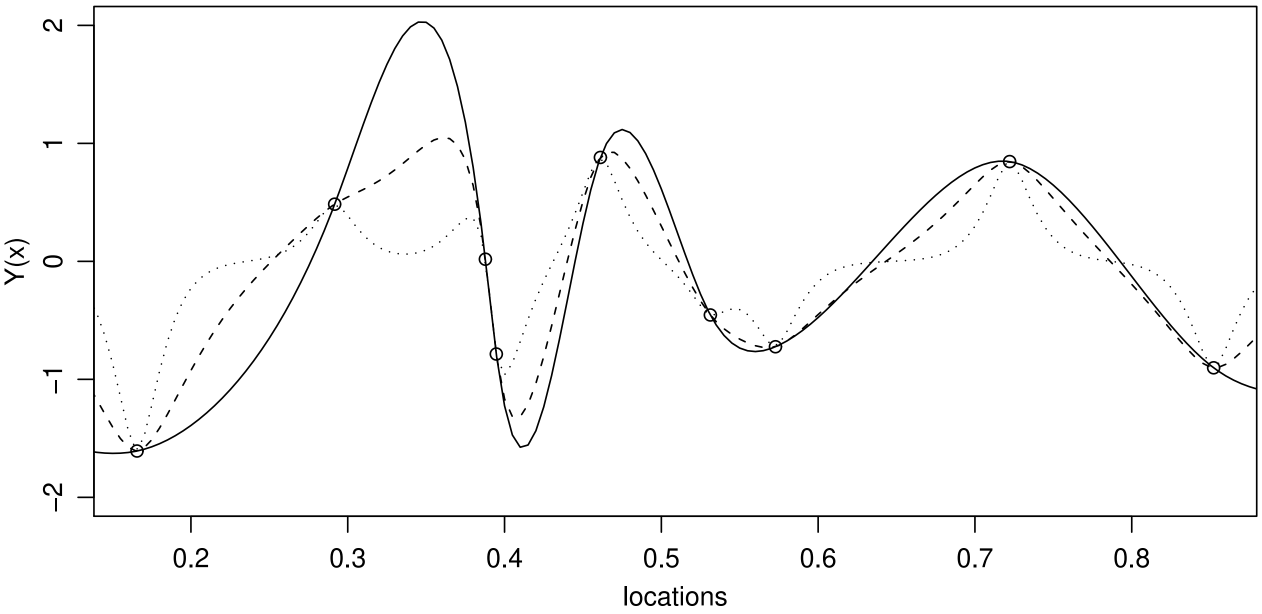
ρ(u)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
u
κ = 0.5 , ϕ = 0.25
κ = 1.5 , ϕ = 0.16
κ = 2.5 , ϕ = 0.13
S(x)
2
1
0
−1
−2
0.0
0.2
0.4
0.6
0.8
1.0
x

Simple kriging: three examples

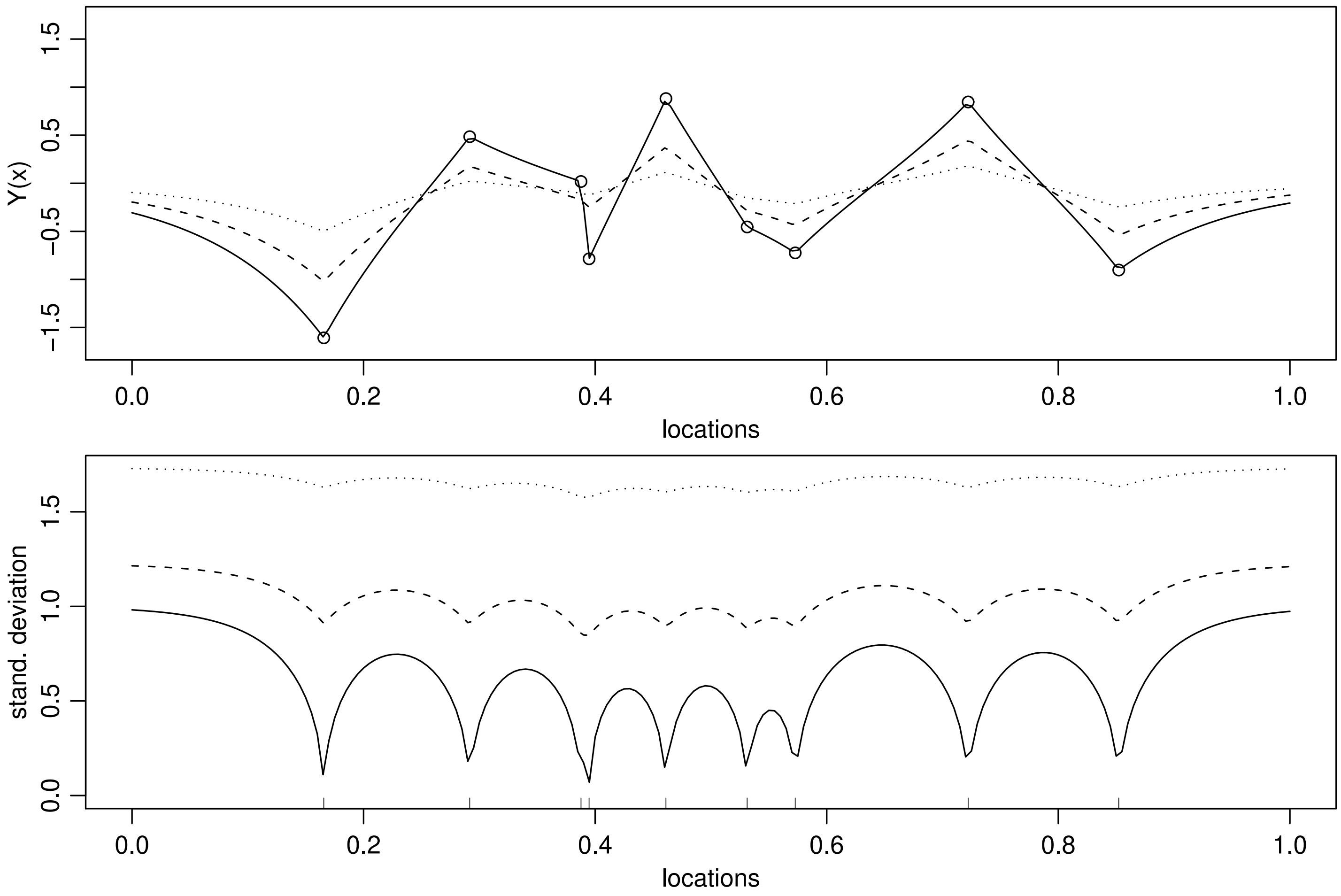
1. Varying κ (smoothness of $S(x)$)



2. Varying ϕ (range of spatial correlation



3. Varying τ^2/σ^2 (noise-to-dignal ratio)



Predicting non-linear functionals

- minimum mean square error prediction is not invariant under non-linear transformation

- the complete answer to a prediction problem is the predictive distribution, $[T|Y]$

- Recommended strategy:

  - draw repeated samples from $[S^*|Y]$ (conditional simulation)

  - calculate required summaries (examples to follow)

Theoretical variograms

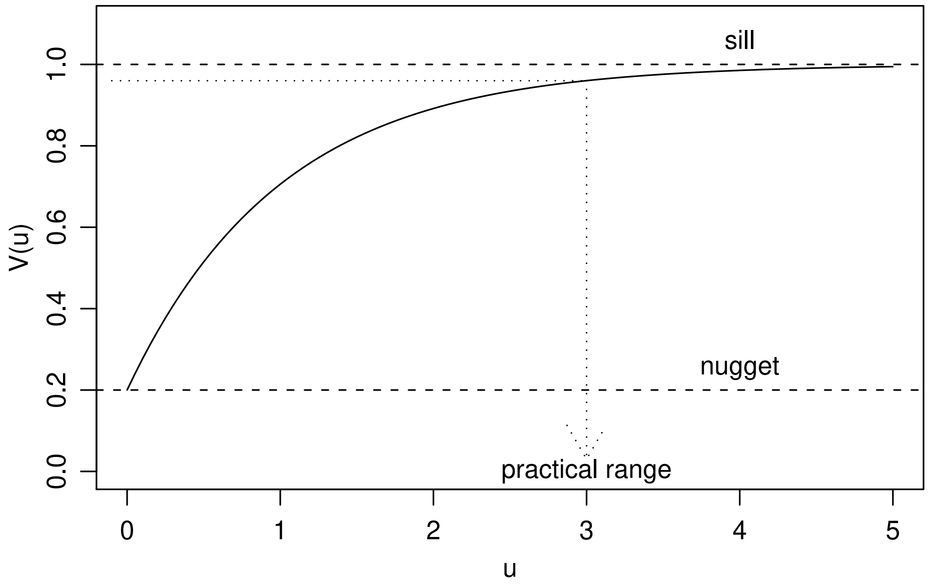
- the variogram of a process $Y(x)$ is the function

$$V(x, x') = \frac{1}{2}\mathrm{Var}\{Y(x) - Y(x')\}$$

- for the spatial Gaussian model, with $u = ||x - x'||$,

$$V(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}$$

- provides a summary of the basic structural parameters of the spatial Gaussian process



- **the nugget variance:** $\boldsymbol{\tau^2}$

- **the sill:** $\boldsymbol{\sigma^2 = \mathrm{Var}\{S(x)\}}$

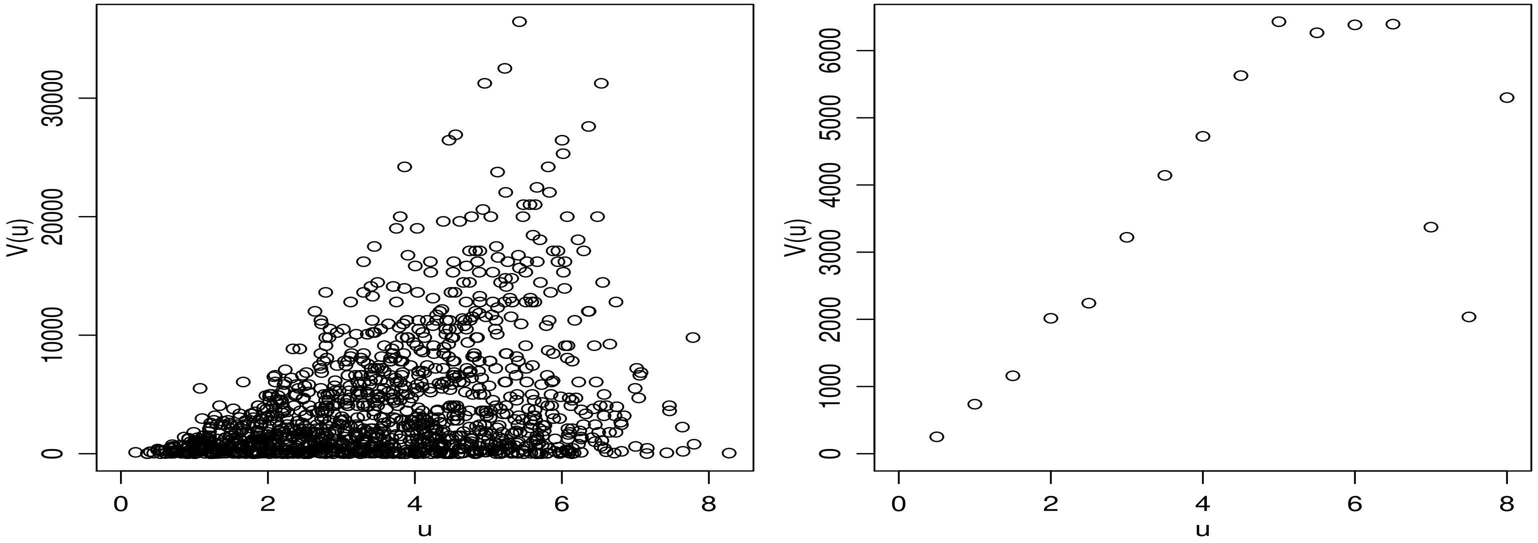
- **the practical range:** $\boldsymbol{\phi}$**, such** $\boldsymbol{\rho(u) = \rho(u/\phi)}$

Empirical variograms

$$u_{ij} = \|x_i - x)j\| \qquad v_{ij} = 0.5[y(x_i) - y(x_j)]^2$$

- the variogram cloud is a scatterplot of the points (u_{ij}, v_{ij})
- the empirical variogram smooths the variogram cloud by averaging within bins: $u - h/2 \leq u_{ij} < u + h/2$
- for a process with non-constant mean (covariates), use residuals $r(x_i) = y(x_i) - \hat{\mu}(x_i)$ to compute v_{ij}

Limitations of $\hat{V}(u)$



1. $v_{ij} \sim V(u_{ij})\chi_1^2$

2. the v_{ij} are correlated

Consequences:

- variogram cloud is unstable, pointwise and in overall shape

- binning addresses point 1, but not point 2

Parameter estimation using the variogram

- fitting a theoretical variogram function to the empirical variogram provides estimates of the model parameters.

- weighted least squares criterion:

$$W(\theta) = \sum_k n_k \{[\bar{V}_k - V(u_k;\theta)]\}^2$$

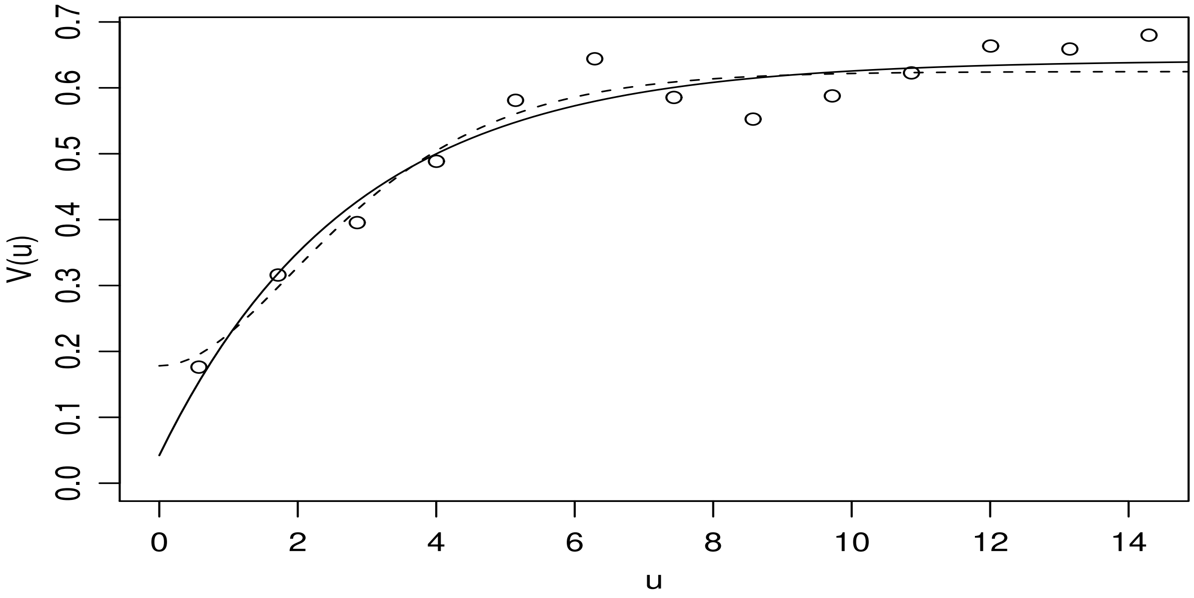
  where θ denotes vector of covariance parameters and $\bar{V}_k$ is average of n_k variogram ordinates v_{ij}.

- need to choose upper limit for u (arbitrary?)

- variations include:
  - fitting models to the variogram cloud
  - other estimators for the empirical variogram
  - different proposals for weights

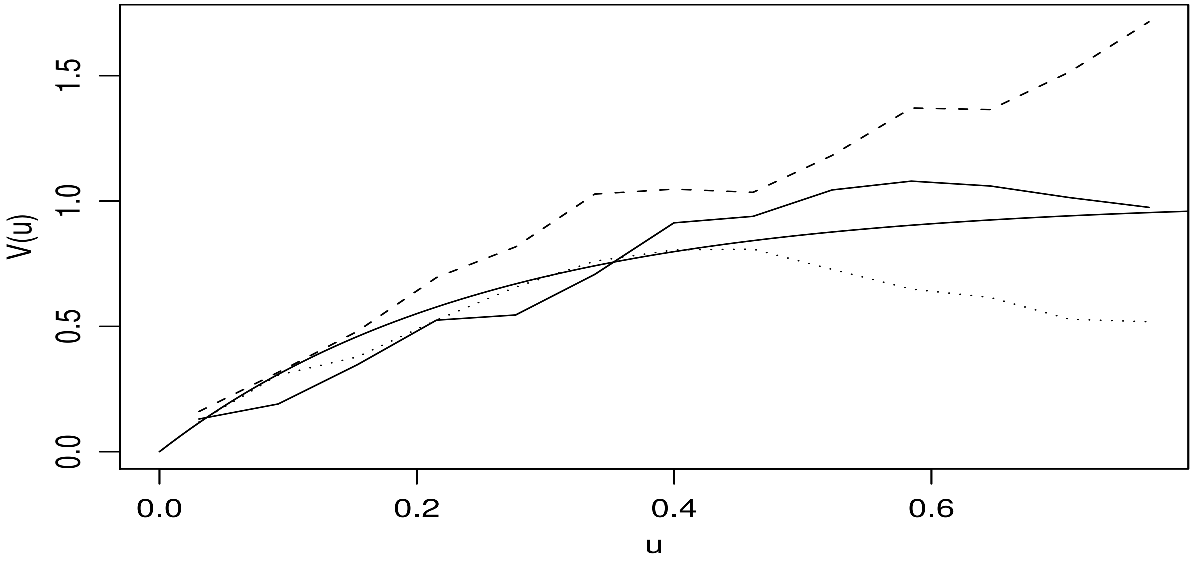
Comments on variogram fitting

1. Can give equally good fits for different extrapolations at origin.



2. Correlation between variogram points induces smoothness.

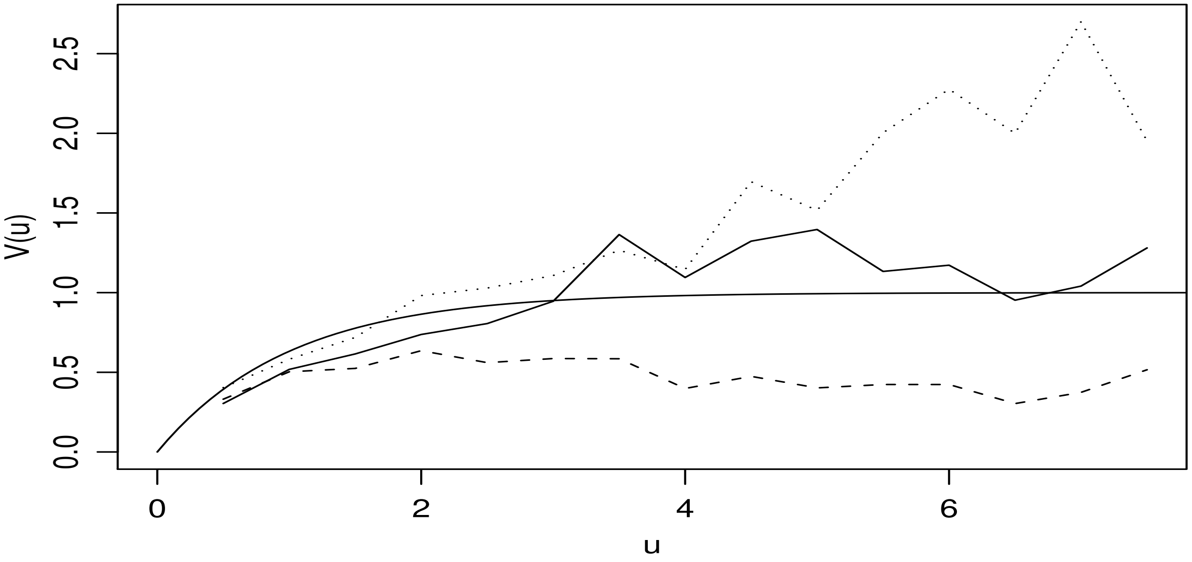
Empirical variograms for three simulations from the same model.



3. **Fit is highly sensitive to specification of the mean.**

Illustration with linear trend surface:

- **solid smooth line: theoretical variogram;**
- **dotted line: from data;**
- **solid line: from true residuals;**
- **dashed line : from estimated residuals.**



Parameter estimation: maximum likelihood

$$Y \sim \mathrm{MVN}(\mu 1, \sigma^2 R + \tau^2 I)$$

R is the $n \times n$ matrix with $(i, j)^{th}$ element $\rho(u_{ij})$ where $u_{ij} = ||x_i - x_j||$, Euclidean distance between x_i and x_j.

Or more generally:

$$\mu(x_i) = \sum_{j=1}^{k} f_k(x_i)\beta_k$$

where $d_k(x_i)$ is a vector of covariates at location x_i, hence

$$Y \sim \mathrm{MVN}(D\beta, \sigma^2 R + \tau^2 I)$$

Gaussian log-likelihood function:

$$\begin{aligned} L(\beta, \tau, \sigma, \phi, \kappa) \propto \quad & -0.5\{\log|(\sigma^2 R + \tau^2 I)| + \\ & (y - D\beta)'(\sigma^2 R + \tau^2 I)^{-1}(y - D\beta)\}. \end{aligned}$$

- write $\nu^2 = \tau^2/\sigma^2$, hence $\sigma^2 V = \sigma^2(R + \nu^2 I)$
- log-likelihood function is maximised for

$$\begin{aligned} \hat{\beta}(V) &= (D'V^{-1}D)^{-1}D'V^{-1}y \\ \hat{\sigma}^2 &= n^{-1}(y - D\hat{\beta})'V^{-1}(y - D\hat{\beta}) \end{aligned}$$

- substitute $(\hat{\beta}, \hat{\sigma^2})$ to give reduced maximisation problem

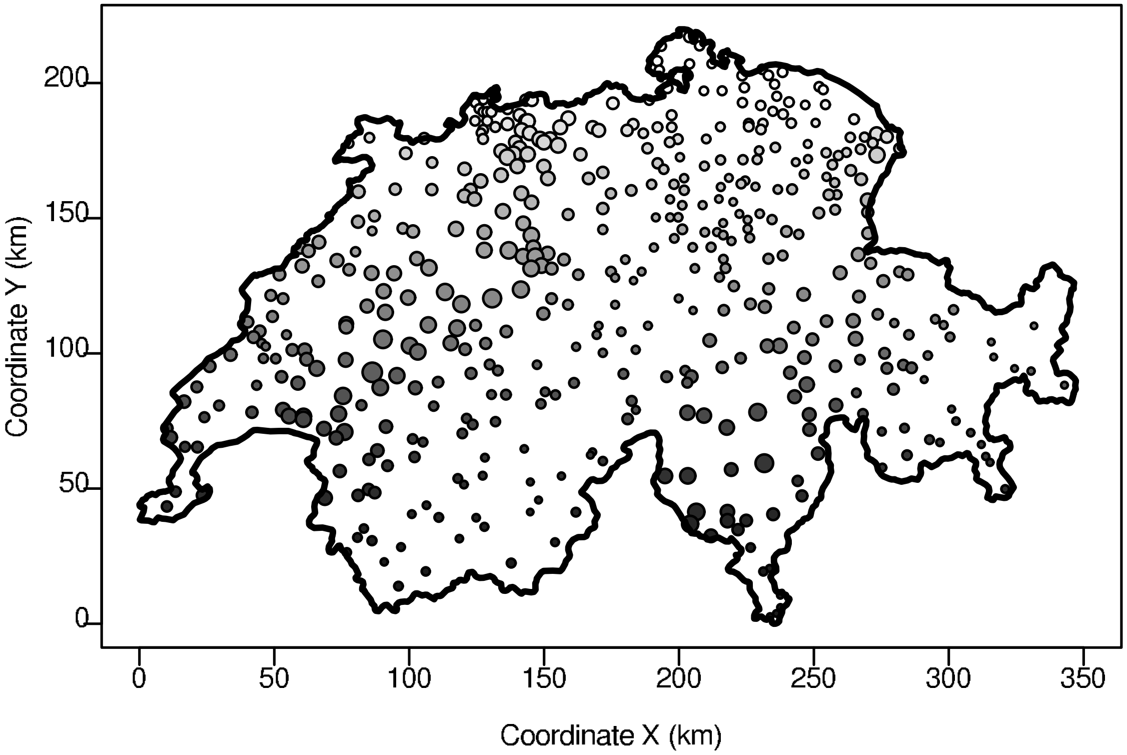
$$L^*(\tau_r, \phi, \kappa) \propto -0.5\{n \log|\hat{\sigma^2}| + \log|(R + \nu^2 I)|\}$$

- usually just consider κ in a discrete set $\{0.5, 1, 2, 3, ..., N\}$

Comments on maximum likelihood

- likelihood-based methods preferable to variogram-based methods
- restricted maximum likelihood is widely recommended but in our experience is sensitive to mis-specification of the mean model.
- in spatial models, distinction between $\mu(x)$ and $S(x)$ is not sharp.
- composite likelihood treats contributions from pairs (Y_i, Y_j) as if independent
- approximate likelihoods useful for handling large data-sets
- examining profile likelihoods is advisable, to check for poorly identified parameters

Swiss rainfall data



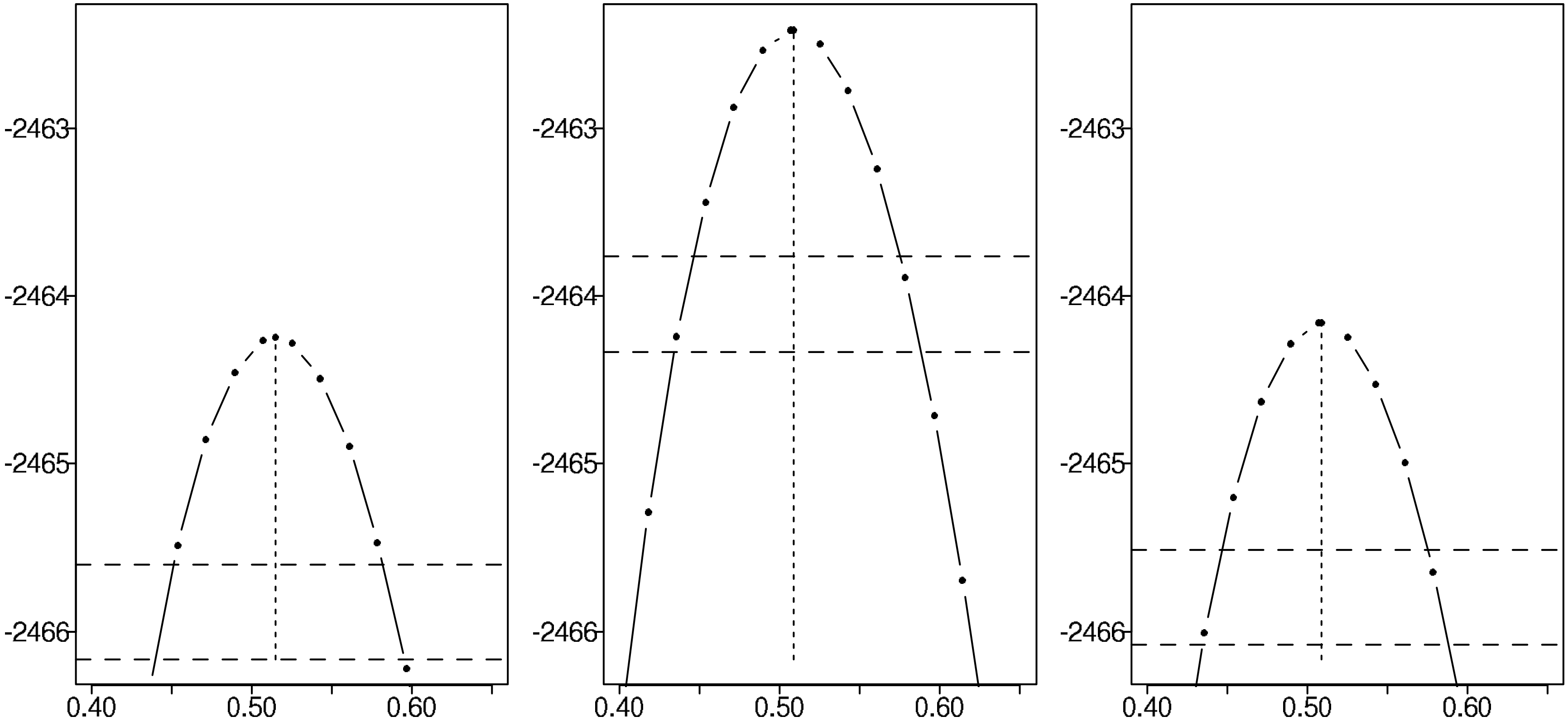
Swiss rainfall: trans-Gaussian model

$$Y_i^* = h_\lambda(Y) = \begin{cases} \frac{(y_i)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y_i) & \text{if } \lambda = 0 \end{cases}$$

$$\begin{aligned} \ell(\beta, \theta, \lambda) &= -\frac{1}{2}\{\log|\sigma^2 V| + (h_\lambda(y) - D\beta)'\{\sigma^2 V\}^{-1}(h_\lambda(y) - D\beta)\} \\ &\quad + \sum_{i=1}^{n} \log\left((y_i)^{\lambda - 1}\right) \end{aligned}$$

Swiss rainfall: profile log-likelihoods for λ

Left panel: $\kappa = 0.5$ **Centre panel: $\kappa = 1$** **Right panel: $\kappa = 2$**

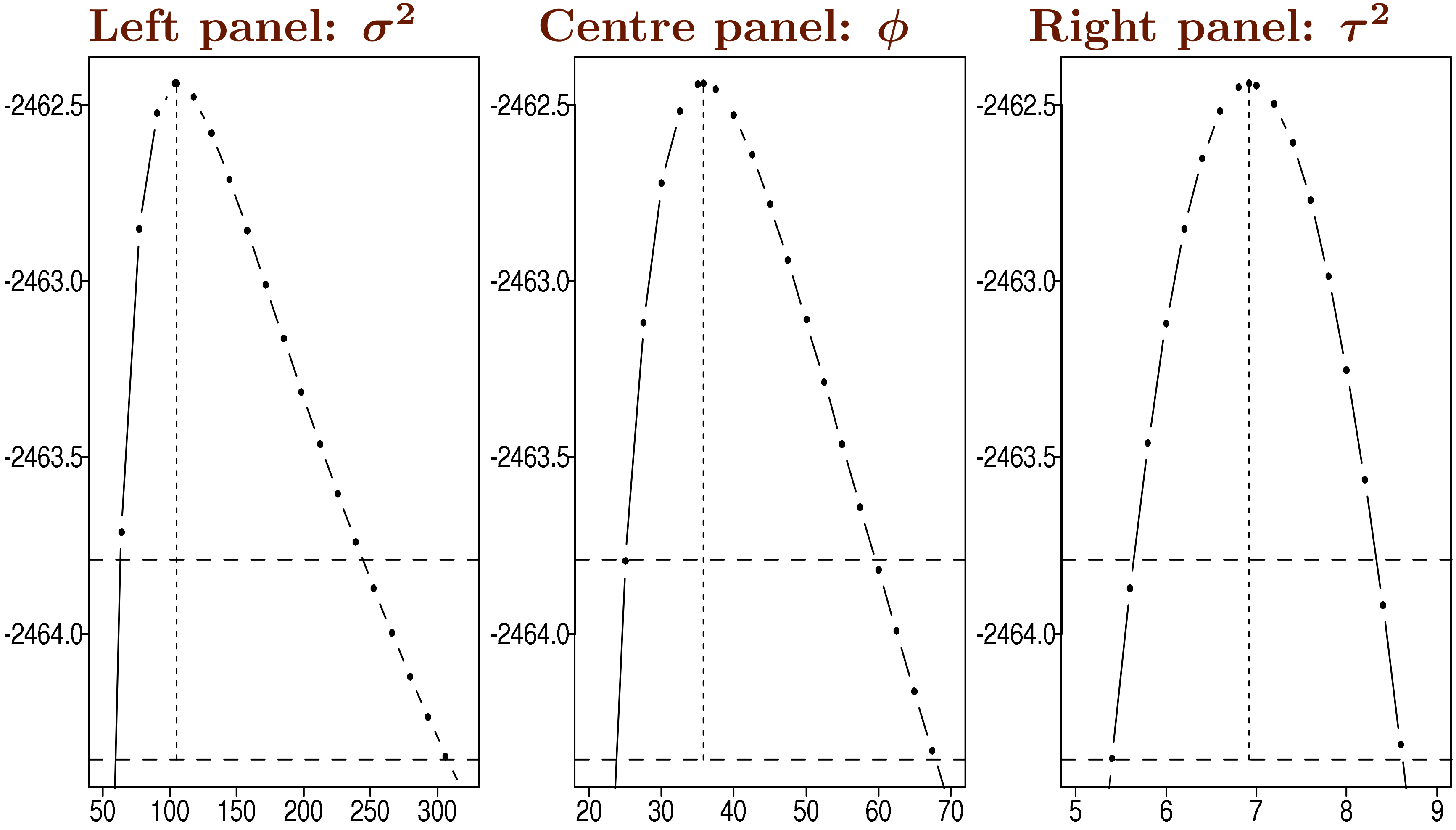


Swiss rainfall: MLE's ($\lambda = 0.5$)

κ	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	$\log \hat{L}$
0.5	18.36	118.82	87.97	2.48	-2464.315
1	20.13	105.06	35.79	6.92	-2462.438
2	21.36	88.58	17.73	8.72	-2464.185

Likelihood criterion favours $\kappa = 1$

Swiss rainfall: profile log-likelihoods $(\lambda = 0.5,\ \kappa = 1)$

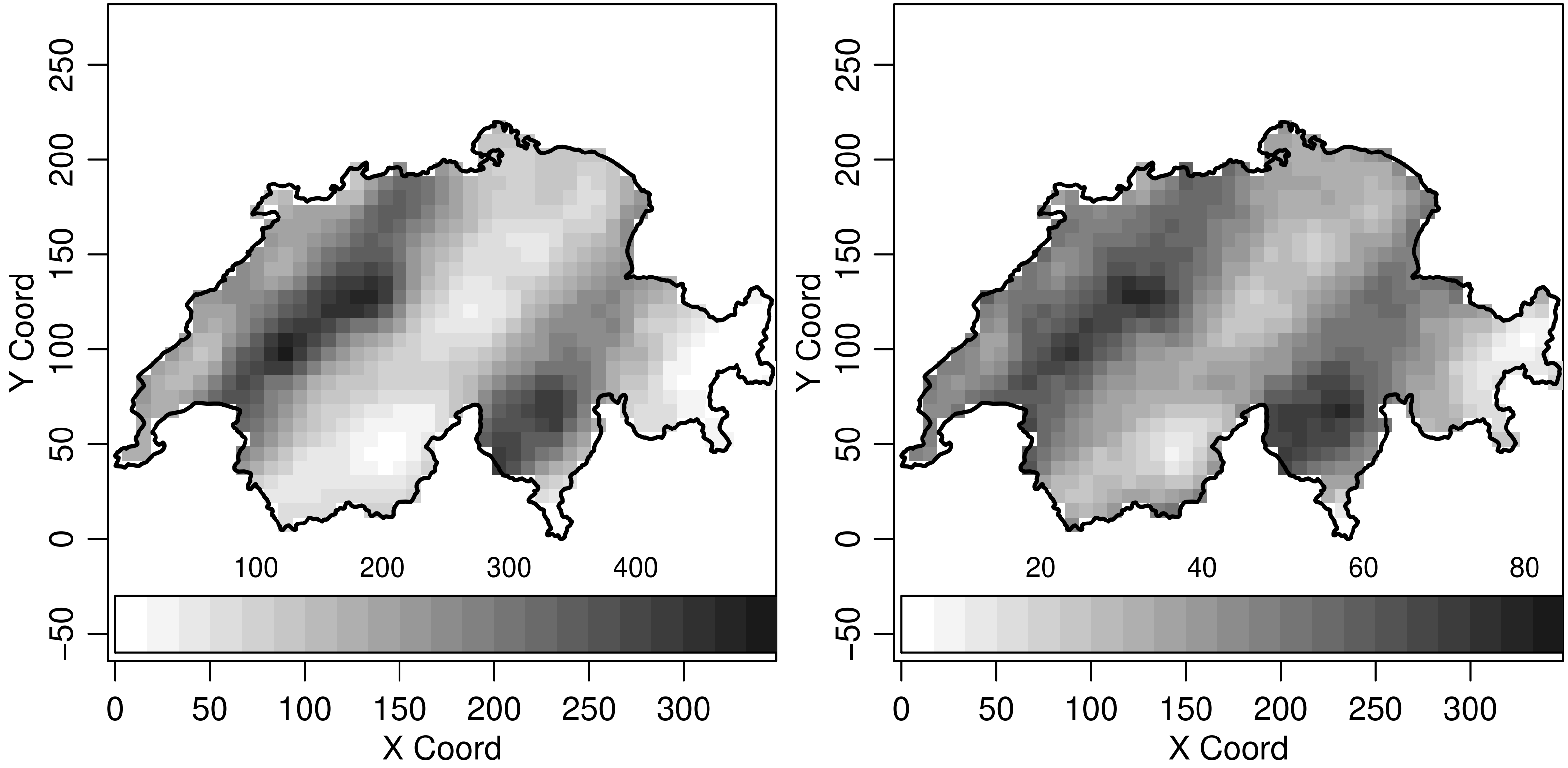


Left panel: σ^2

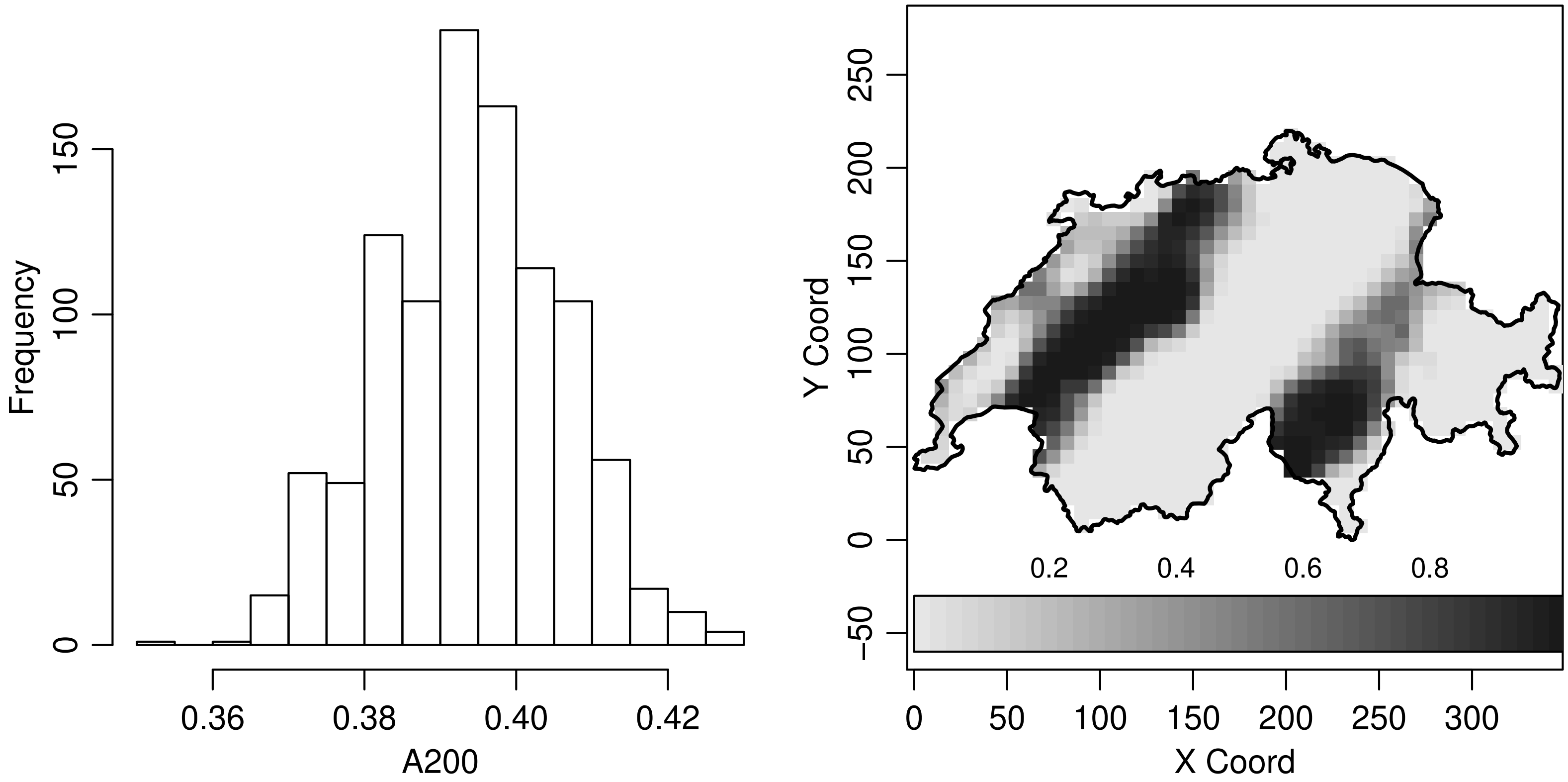
Centre panel: ϕ

Right panel: τ^2

Swiss rainfall: plug-in predictions and prediction variances



Swiss rainfall: non-linear prediction



Left-panel: plug-in prediction for proportion of total area with rainfall exceeding 200 (= 20mm)

Right-panel: plug-in predictive map of $P(\text{rainfall} > 250|Y)$

Section 3

Bayesian inference

Basics

Model specification

$$[\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\theta}] = [\boldsymbol{\theta}][\boldsymbol{S}|\boldsymbol{\theta}][\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{\theta}]$$

Parameter estimation

- **integration gives**

$$[\boldsymbol{Y}, \boldsymbol{\theta}] = \int [\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\theta}] d\boldsymbol{S}$$

- **Bayes' Theorem gives posterior distribution**

$$[\boldsymbol{\theta}|\boldsymbol{Y}] = [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}]/[\boldsymbol{Y}]$$

- **where** $[\boldsymbol{Y}] = \int [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}] d\boldsymbol{\theta}$

Prediction: $S \longrightarrow S^*$

- expand model specification to

$$[Y, S^*, \theta] = [\theta][S|\theta][Y|S,\theta][S^*|S,\theta]$$

- plug-in predictive distribution is

$$[S^*|Y, \hat{\theta}]$$

- Bayesian predictive distribution is

$$[S^*|Y] = \int [S^*|Y,\theta][\theta|Y]d\theta$$

- for any target $T = t(S^*)$, required predictive distribution $[T|Y]$ follows

Notes

- likelihood function is central to both classical and Bayesian inference

- Bayesian prediction is a weighted average of plug-in predictions, with different plug-in values of θ weighted according to their conditional probabilities given the observed data.

- Bayesian prediction is usually more conservative than plug-in prediction

Bayesian computation

1. Evaluating the integral which defines $[S^*|Y]$ is often difficult

2. Markov Chain Monte Carlo methods are widely used

3. but for geostatistical problems, reliable implementation of MCMC is not straightforward (no natural Markovian structure)

4. for the Gaussian model, direct simulation is available

Gaussian models: known (σ^2, ϕ)

$$Y \sim \mathrm{N}(D\beta, \sigma^2 R(\phi))$$

- choose conjugate prior $\beta \sim \mathrm{N}\left(m_\beta \ ; \ \sigma^2 V_\beta\right)$
- posterior for β is $\left[\beta | Y, \sigma^2, \phi\right] \sim \mathrm{N}\left(\hat{\beta}, \sigma^2 \, V_{\hat{\beta}}\right)$

$$\begin{aligned} \hat{\beta} &= (V_\beta^{-1} + D' R^{-1} D)^{-1} (V_\beta^{-1} m_\beta + D' R^{-1} y) \\ V_{\hat{\beta}} &= \sigma^2 \, (V_\beta^{-1} + D' R^{-1} D)^{-1}) \end{aligned}$$

- predictive distribution for S^* is

$$p(S^* | Y, \sigma^2, \phi) \ = \ \int p(S^* | Y, \beta, \sigma^2, \phi) \, p(\beta | Y, \sigma^2, \phi) \, d\beta.$$

Notes

- **mean and variance of predictive distribution can be written explicitly (but not given here)**

- **predictive mean compromises between prior and weighted average of $\boldsymbol{Y}$**

- **predictive variance has three components:**

  - **a priori variance,**
  - **minus information in data**
  - **plus uncertainty in $\boldsymbol{\beta}$**

- **limiting case $V_{\boldsymbol{\beta}} \to \infty$ corresponds to ordinary kriging.**

Gaussian models: unknown (σ^2, ϕ)

Convenient choice of prior is:

$$[\beta|\sigma^2, \phi] \sim N\left(m_b, \sigma^2 V_b\right) \quad [\sigma^2|\phi] \sim \chi^2_{ScI}\left(n_\sigma, S^2_\sigma\right) \quad [\phi] \sim \text{arbitrary}$$

- results in explicit expression for $[\beta, \sigma^2|Y, \phi]$ and computable expression for $[\phi|Y]$, depending on choice of prior for ϕ
- in practice, use arbitrary discrete prior for ϕ and combine posteriors conditional on ϕ by weighted averaging

Algorithm 1:

1. choose lower and upper bounds for ϕ according to the particular application, and assign a discrete uniform prior for ϕ on a set of values spanning the chosen range

2. compute posterior $[\phi|Y]$ on this discrete support set

3. sample ϕ from posterior, $[\phi|Y]$

4. attach sampled value of ϕ to conditional posterior, $[\beta, \sigma^2|y, \phi]$, and sample (β, σ^2) from this distribution

5. repeat steps (3) and (4) as many times as required; resulting sample of triplets (β, σ^2, ϕ) is a sample from joint posterior distribution, $[\beta, \sigma^2, \phi|Y]$

Predictive distribution for S^* given ϕ is tractable, hence write

$$p(S^*|Y) \quad = \quad \int p(S^*|Y,\phi)\, p(\phi|y)\, d\phi.$$

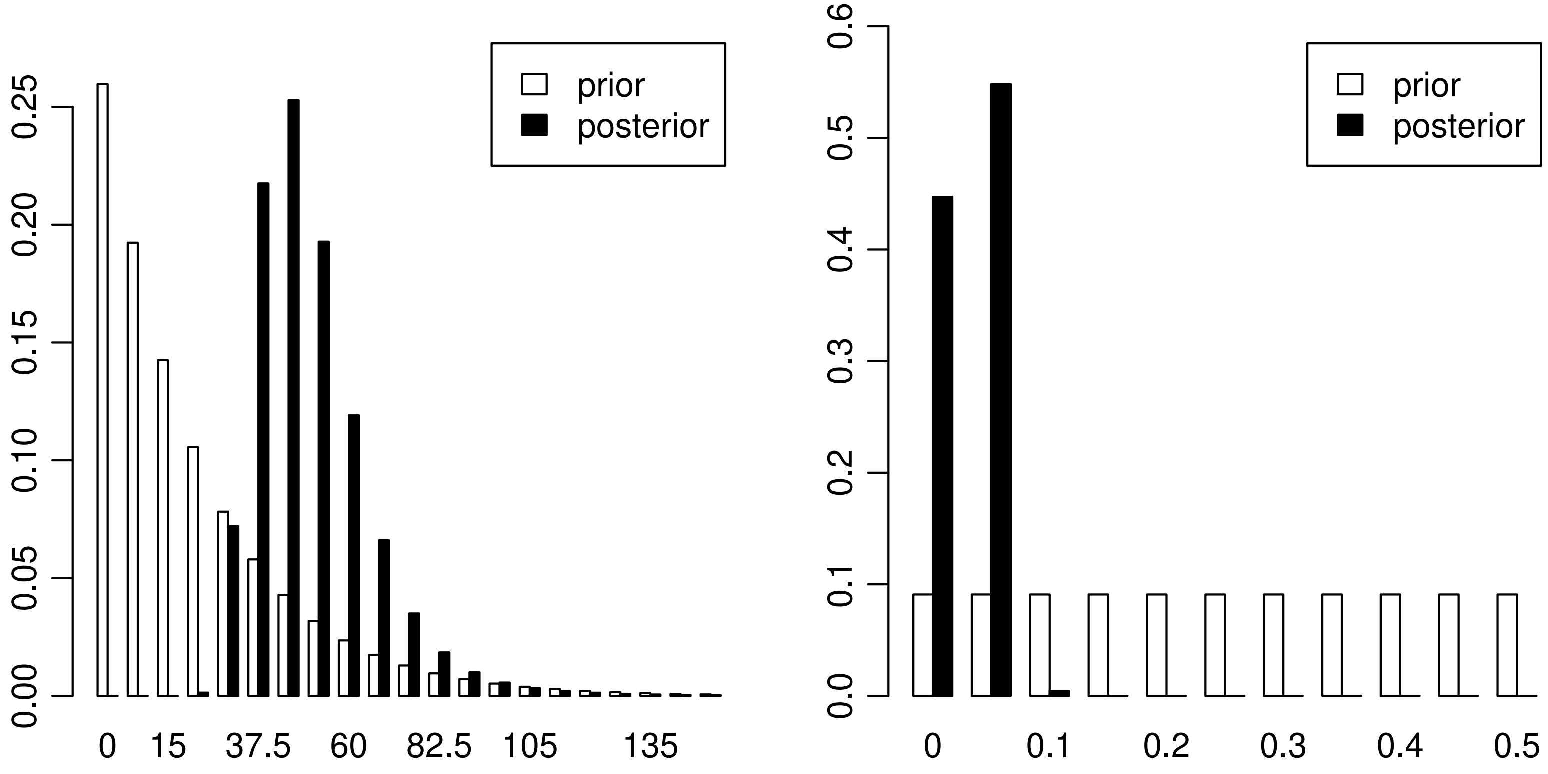
Algorithm 2:

1. **discretise $[\phi|Y]$, as in Algorithm 1.**
2. **compute posterior $[\phi|Y]$**
3. **sample ϕ from posterior $[\phi|Y]$**
4. **attach sampled value of ϕ to $[S^*|y,\phi]$ and sample from this to obtain realisations from $[S^*|Y]$**
5. **repeat steps (3) and (4) as required**

Note: Extends immediately to multivariate ϕ (but may be computationally awkward)

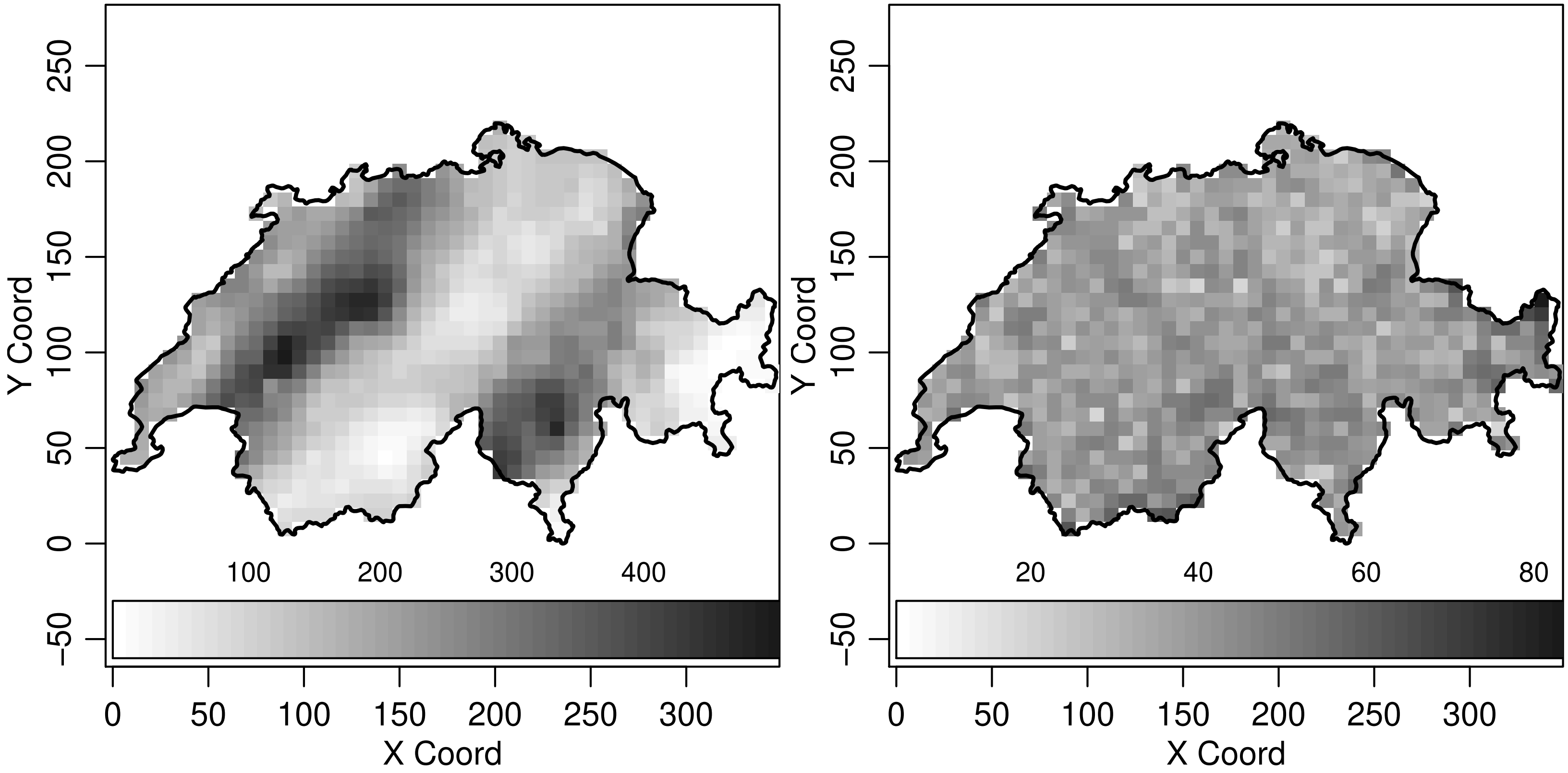
Swiss rainfall

Priors/posteriors for ϕ (left) and ν^2 (right)



Swiss rainfall

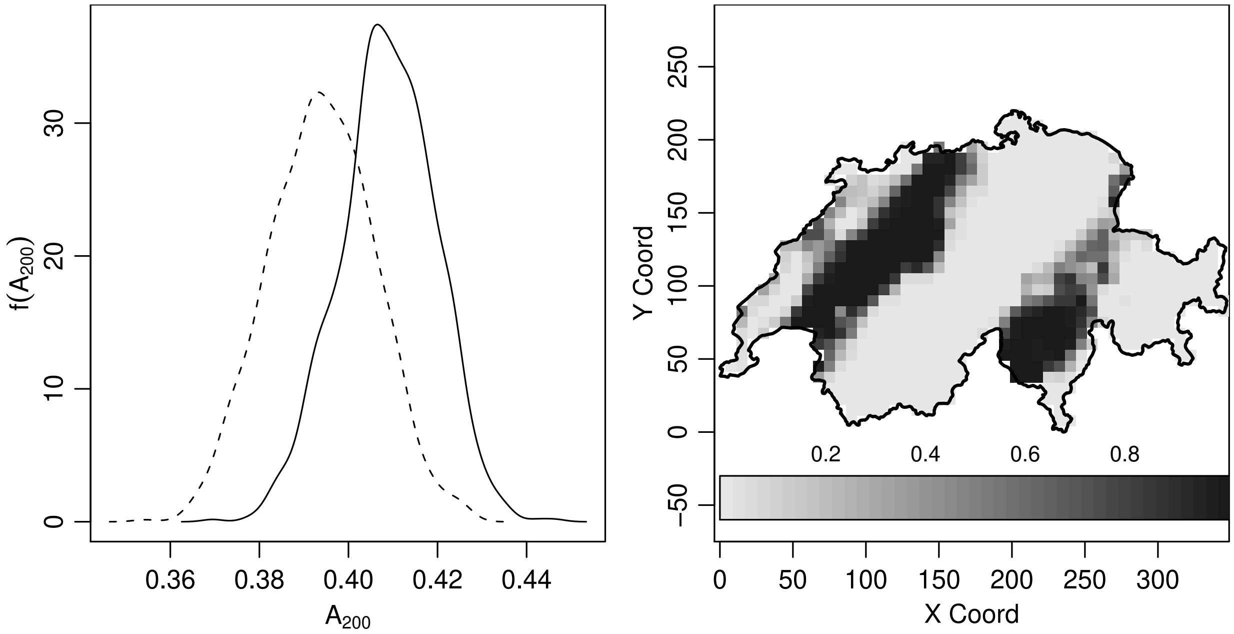
Mean (left-panel) and variance (right-panel) of predictive distribution



Swiss rainfall: posterior means and 95% credible intervals

parameter	estimate	95% interval
β	144.35	[53.08, 224.28]
σ^2	13662.15	[8713.18, 27116.35]
ϕ	49.97	[30, 82.5]
ν^2	0.03	[0, 0.05]

Swiss rainfall: non-linear prediction



Left-panel: **Bayesian (solid) and plug-in (dashed) prediction for proportion of total area with rainfall exceeding 200 (= 20mm)**

Right-panel: **Bayesian predictive map of $P(\text{rainfall} > 250|Y)$**

Section 4

Generalized linear models

Generalized linear geostatistical model

- **Latent spatial process**

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

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

- **Linear predictor**

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

- **Link function**

  $\mathrm{E}[Y_i] = \mu_i = h\{\eta(x_i)\}$

- **Conditional distribution for $Y_i : i = 1, ..., n$**

  $Y_i|S(\cdot) \sim f(y;\eta)$ mutually independent

GLGM

- usually just a single realisation is available, in contrast with GLMM for longitudinal data analysis

- GLM approach is most appealing when there is a natural sampling mechanism, for example Poisson model for counts or logistic-linear models for proportions

- transformed Gaussian models may be more useful for non-Gaussian continuous respones

- theoretical variograms can be derived but are less natural as summary statistics than in Gaussian case

- but empirical variograms of GLM residuals can still be useful for exploratory analysis

A binomial logistic-linear model

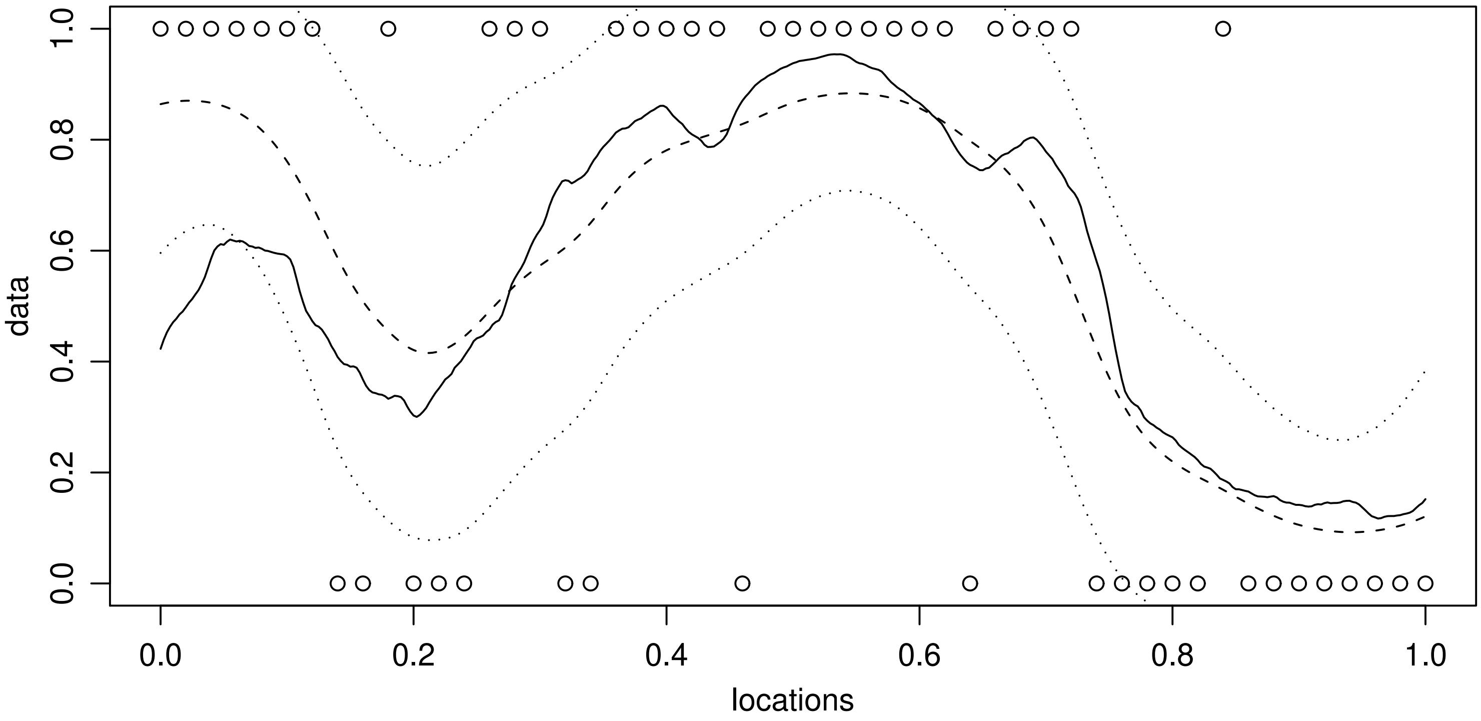
- $S(\cdot) \sim$ zero-mean Gaussian process
- $[Y(x_i) \mid S(x_i)] \sim \text{Bin}(n_i; p_i)$
- $h(p_i) = \log\{p_i/(1-p_i)\} = \sum_{j=1}^{k} d_{ij}\beta_j + S(x_i)$
- model can be expanded by adding uncorrelated random effects Z_i,

  $$h(p_i) = \sum_{j=1}^{k} d_{ij}\beta_j + S(x_i) + Z_i$$

  to distinguish between two forms of the nugget effect:

  - binomial variation is analogue of measurement error
  - Z_i is analogue of short-range spatial variation

Simulation of a binary logistic-linear model



- **data contain little information about $S(x)$**

- **leads to wide prediction intervals for $p(x)$**

- **more informative for binomial responses with large n_i**

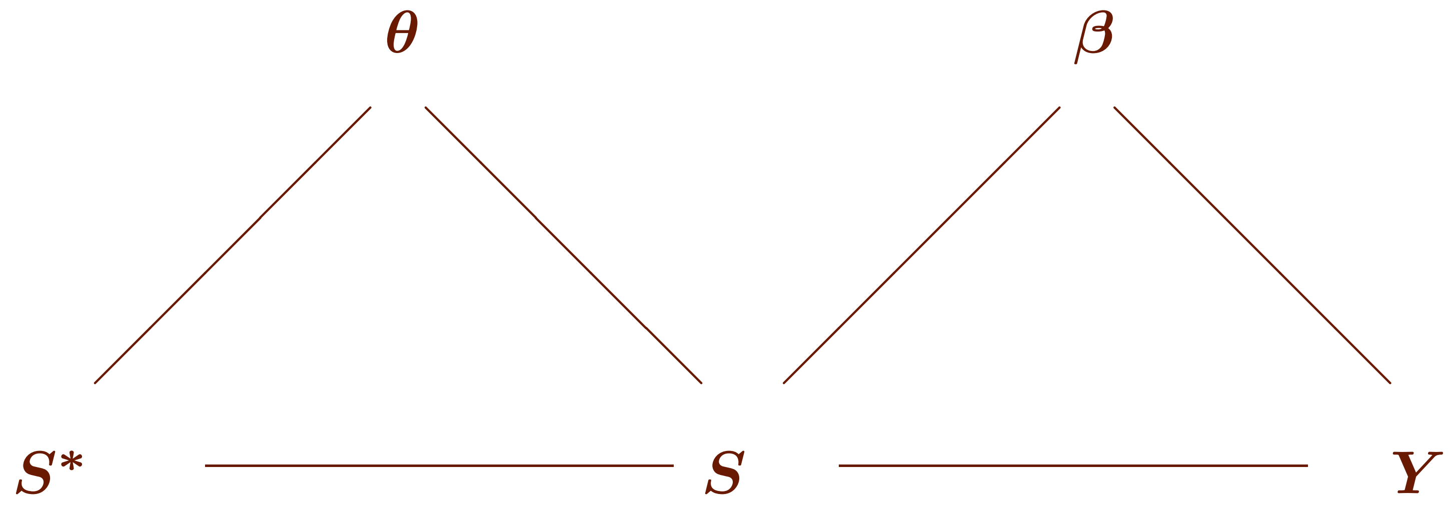
Inference

- Likelihood function

$$L(\theta) = \int_{\mathbb{R}^n} \prod_{i}^{n} f(y_i; h^{-1}(s_i)) f(s \mid \theta) ds_1, \ldots, ds_n$$

- involves high-dimensional integration
- MCMC algorithms exploit conditional independence structure

Conditional independence graph



- only need vertex S^* at prediction stage
- corresponding DAG would delete edge between S and β

Prediction with known parameters

- simulate $s(1), \ldots, s(m)$ from $[S|y]$ (using MCMC).
- simulate $s^*(j)$ from $[S^*|s(j)]$, $j = 1, \ldots, m$ (multivariate Gaussian)
- approximate $\mathrm{E}[T(S^*)|y]$ by $\frac{1}{m}\sum_{j=1}^{m} T(s^*(j))$
- if possible reduce Monte Carlo error by
  - calculating $\mathrm{E}[T(S^*)|s(j)]$ directly
  - estimating $\mathrm{E}[T(S^*)|y]$ by $\frac{1}{m}\sum_{j=1}^{m} \mathrm{E}[T(S^*)|s(j)]$

MCMC for conditional simulation

- **Let $S = D'\beta + \Sigma^{1/2}\Gamma$, $\Gamma \sim N_n(0, I)$.**
- **Conditional density: $f(\gamma|y) \propto f(y|\gamma)f(\gamma)$**

Langevin-Hastings algorithm

- **Proposal: γ' from a $N_n(\xi(\gamma), hI)$,**

$$\xi(\gamma) = \gamma + \frac{h}{2}\nabla \log f(\gamma \mid y)$$

- **Example: Poisson-log-linear spatial model:**

  $\nabla \log f(\gamma|y) = -\gamma + (\Sigma^{1/2})'(y - \exp(s)), \quad s = \Sigma^{1/2}\gamma.$

- **expression generalises to other generalised linear spatial models**
- **MCMC output $\gamma(1), \ldots, \gamma(m)$, hence sample $s(m) = \Sigma^{1/2}\gamma(m)$ from $[S|y]$.**

MCMC for Bayesian inference

Posterior:

- **update Γ from $[\Gamma|y, \beta, \log\sigma), \log(\phi)]$ (Langevin-Hastings))**
- **update β from $[\beta|\Gamma, \log(\sigma), \log(\phi)]$ (RW-Metropolis)**
- **update $\log(\sigma)$ from $[\log(\sigma)|\Gamma, \beta, \log(\phi)]$ (RW-Metropolis)**
- **update $\log(\phi)$ from $[\log(\phi)|\Gamma, \beta, \log(\sigma)]$ (RW-Metropolis)**

Predictive:

- **simulate $(s(j), \beta(j), \sigma^2(j), \phi(j))$, $j = 1, \ldots, m$ (MCMC)**
- **simulate $s^*(j)$ from $[S^*|s(j), \beta(j), \sigma^2(j), \phi(j)]$, $j = 1, \ldots, m$ (multivariate Gaussian)**

Comments

- above is not necessarily the most efficient algorithm available
- discrete prior for ϕ reduces computing time
- can thin MCMC output if storage is a limiting factor
- similar algorithms can be developed for MCMC maximum likelihood estimation

Some computational resources

- geoR package:
  http://www.est.ufpr.br/geoR

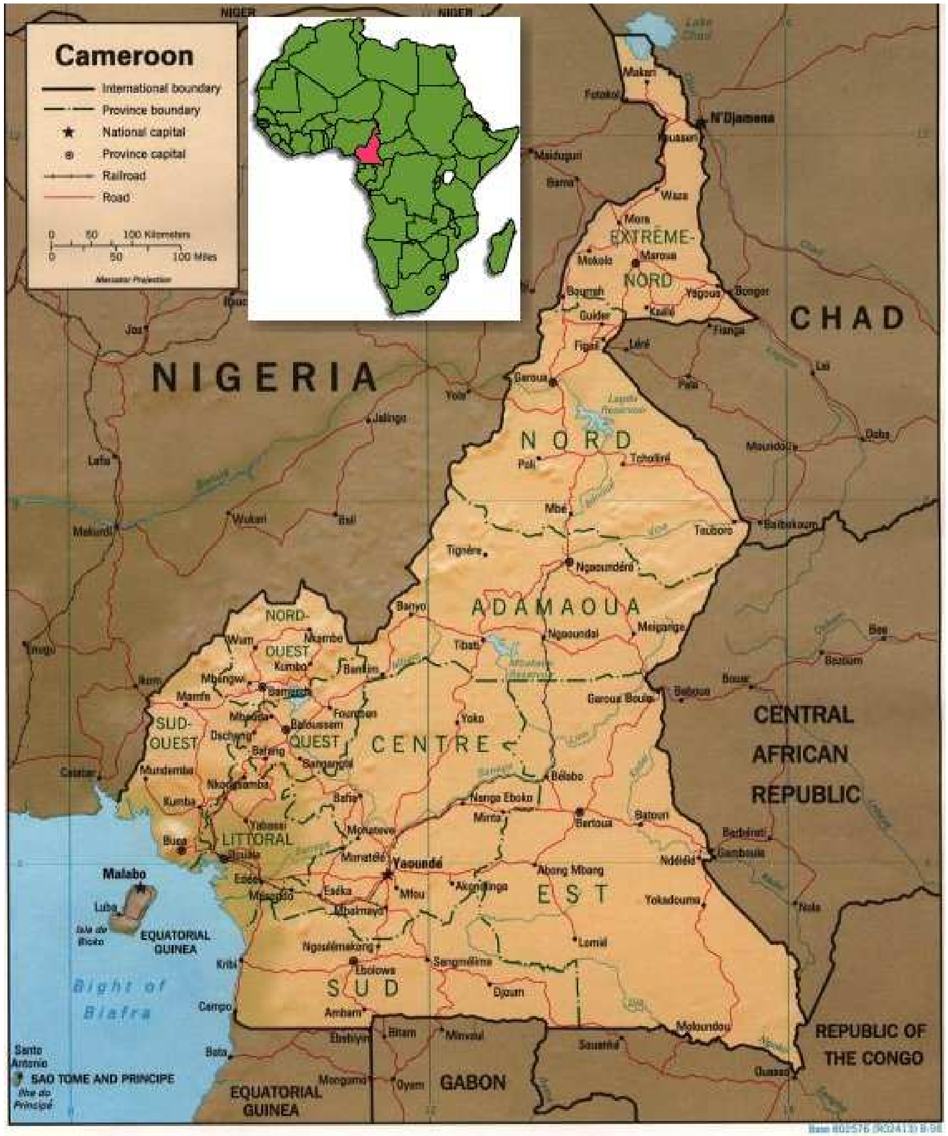
- geoRglm package:
  http://www.est.ufpr.br/geoRglm

- R-project:
  http://www.R-project.org

- CRAN spatial task view:
  http://cran.r-project.org/src/contrib/Views/Spatial.html

- AI-Geostats web-site:
  http://www.ai-geostats.org

- and more ...


Cameroon
International boundary
Province boundary
National capital
Province capital
Railroad
Road
0 50 100 Kilometers
0 50 100 Miles
Mercator Projection
NIGER
NIGER
Lake Chad
N'Djamena
Maiduguri
Bama
CHAD
NIGERIA
Jos
Yola
Jalingo
Lafia
Makurdi
Wukari
Bali
Enugu
Ikom
Calabar
EXTRÊME-NORD
Makari
Fotokol
Kousseri
Waza
Mora
Mokolo
Maroua
Yagoua
Bongor
Kaélé
Guider
Figuil
Léré
Fianga
Pala
Lai
Garoua
Lagdo Reservoir
NORD
Poli
Tcholliré
Moundou
Doba
Mbé
Touboro
Baïbokoum
Tignère
Ngaoundéré
ADAMAOUA
Banyo
Tibati
Ngaoundal
Meiganga
NORD-OUEST
Wum
Nkambe
Kumbo
Banso
Bamenda
Mbengwi
Mamfe
SUD-OUEST
OUEST
Bafoussam
Foumban
Dschang
Bafang
Bangangté
Mundemba
Kumba
Nkongsamba
Buea
Limbe
Douala
Edéa
LITTORAL
CENTRE
Yoko
Bélabo
Bafia
Nanga Eboko
Minta
Monatélé
Yaoundé
Mbalmayo
Mfou
Eséka
Akonolinga
Bertoua
Batouri
Abong Mbang
EST
Garoua Boulaï
Bouar
Bozoum
CENTRAL AFRICAN REPUBLIC
Berbérati
Gamboula
Nola
Ndélélé
Yokadouma
Lomié
Moloundou
Ouesso
REPUBLIC OF THE CONGO
Sangmélima
Ebolowa
Ngoulemakong
Djoum
SUD
Ambam
Kribi
Campo
Bata
Ebebiyin
Bitam
Minvoul
Souanké
Mongomo
Oyem
GABON
EQUATORIAL GUINEA
Malabo
Luba
Isla de Bioko
EQUATORIAL GUINEA
Bight of Biafra
Santo Antonio
SAO TOME AND PRINCIPE
Ilha do Principe
Base 802576 (R02413) 8-98

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,

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

Loa loa: a generalised linear model

- **Latent spatial process**

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

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

- **Linear predictor**

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

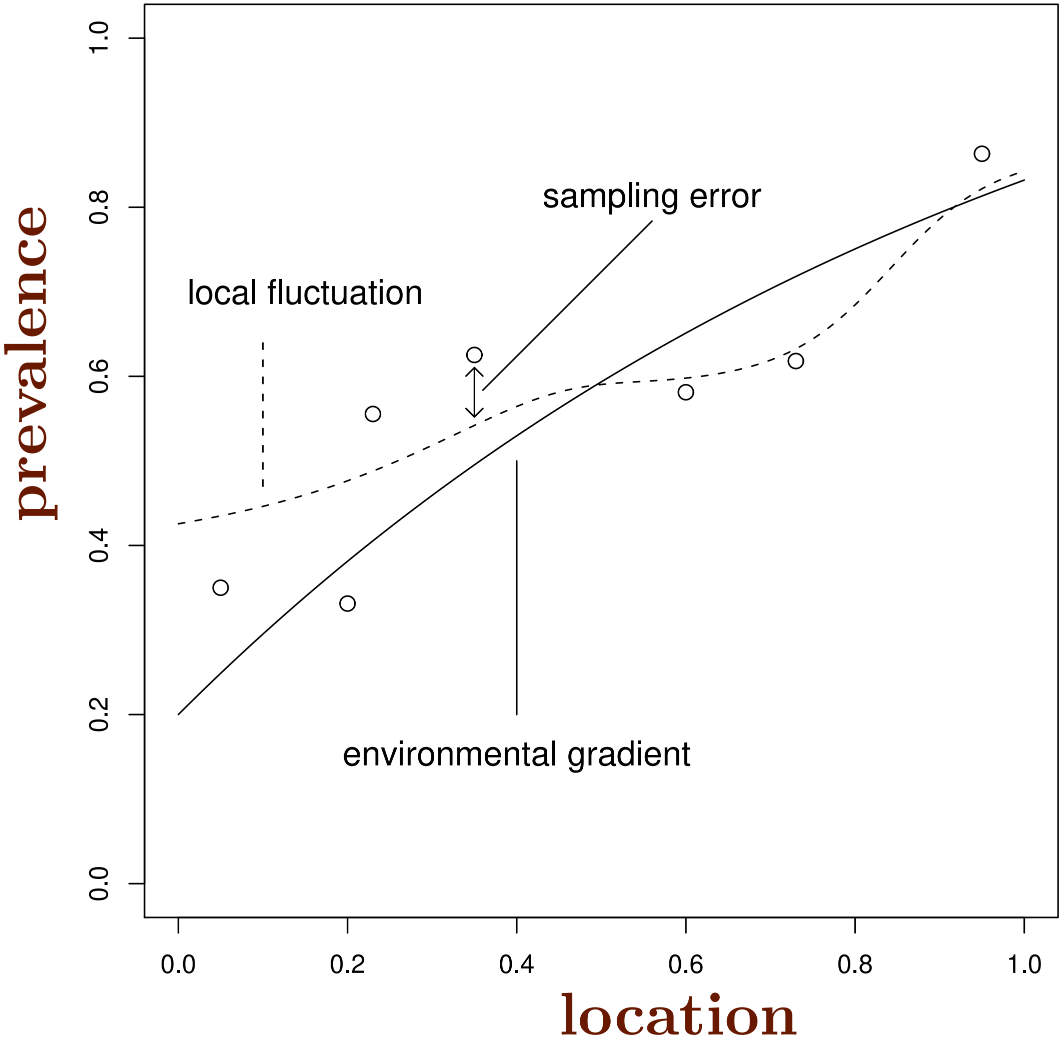
  $\eta(x) = d(x)'\beta + S(x)$

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

- **Error distribution**

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

Schematic representation of Loa loa model



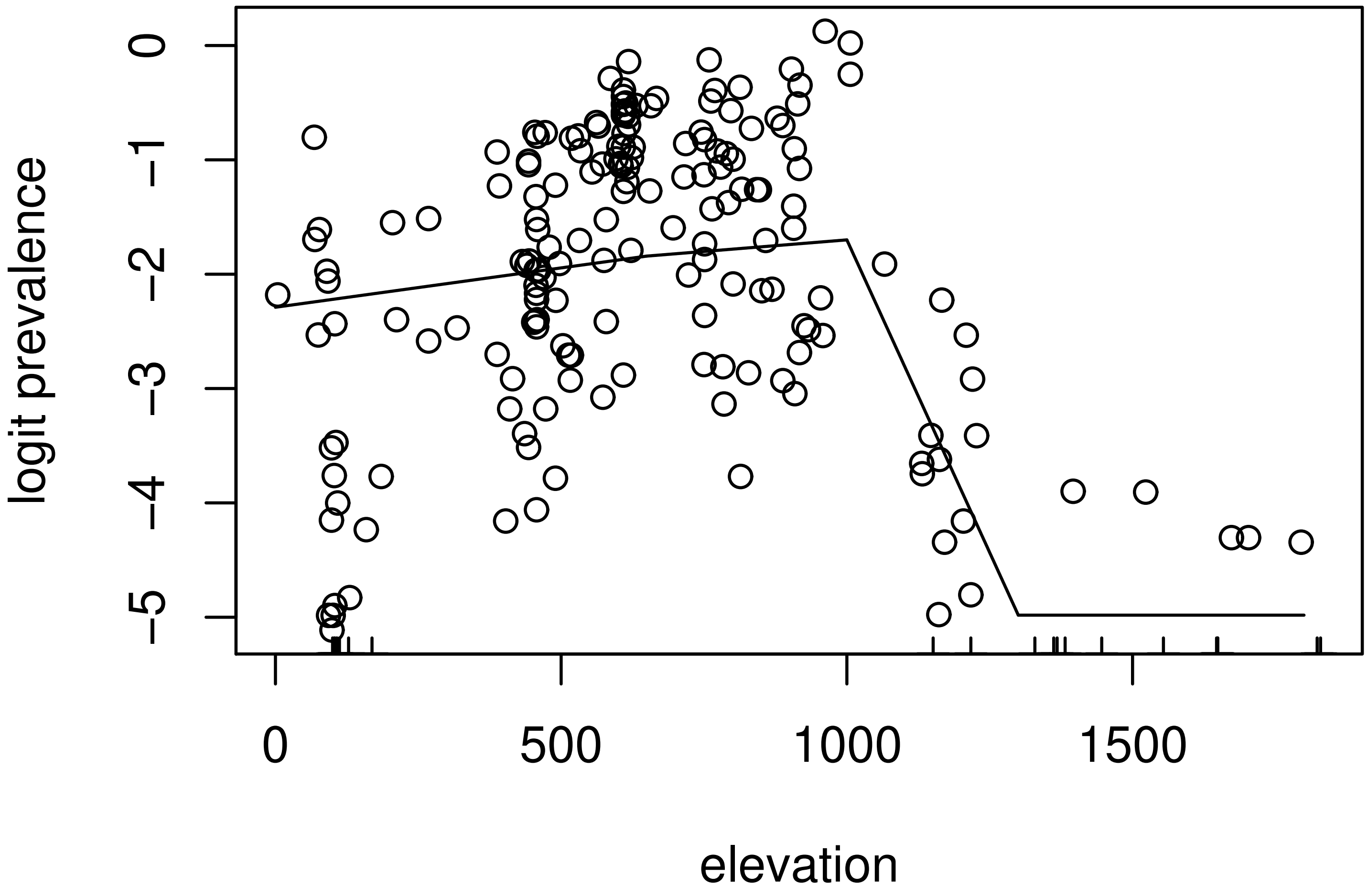
The modelling strategy

- use relationship between environmental variables and ground-truth prevalence to construct preliminary predictions via logistic regression

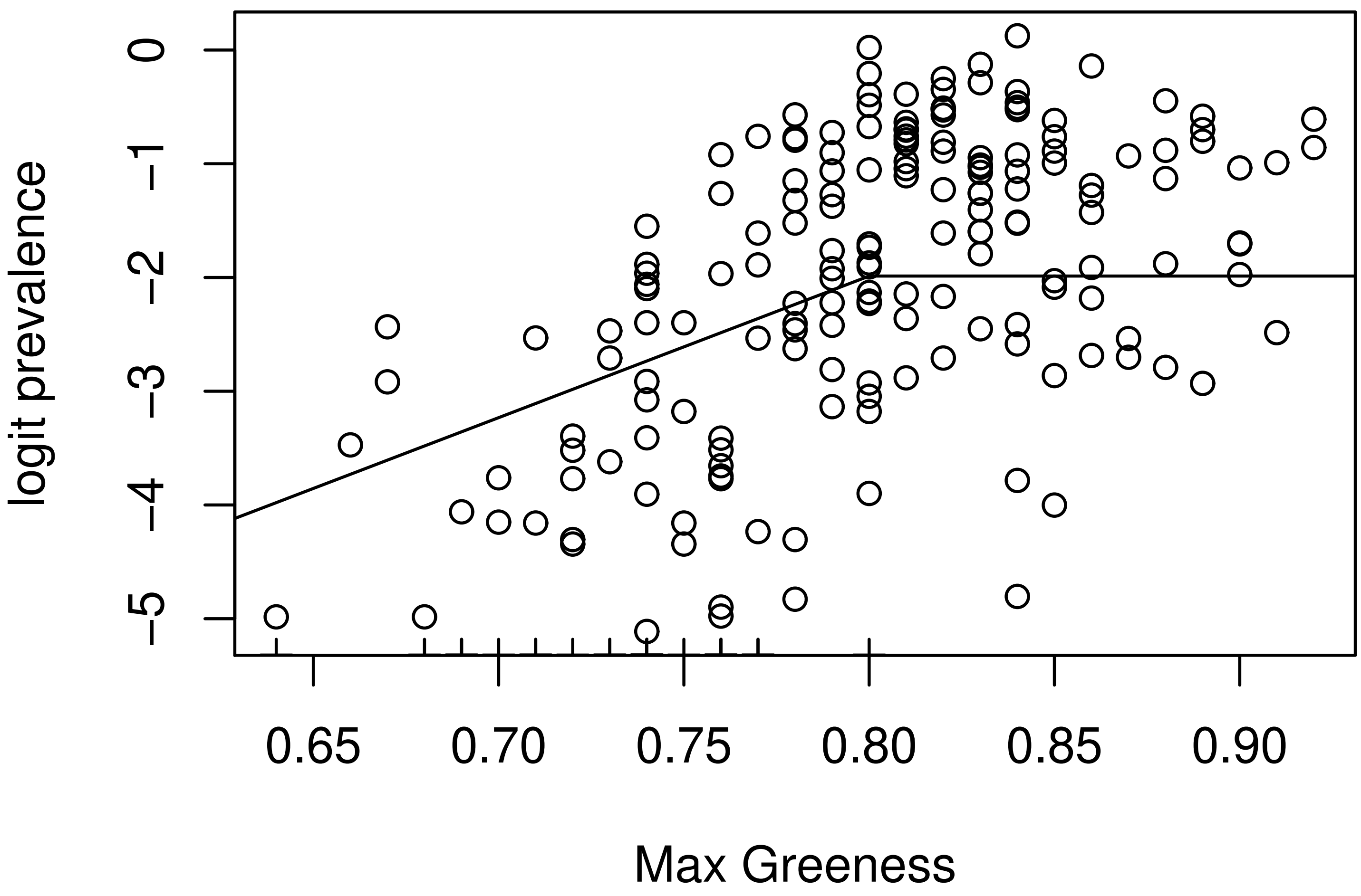
- use local deviations from regression model to estimate smooth residual spatial variation

- Bayesian paradigm for quantification of uncertainty in resulting model-based predictions

logit prevalence vs elevation

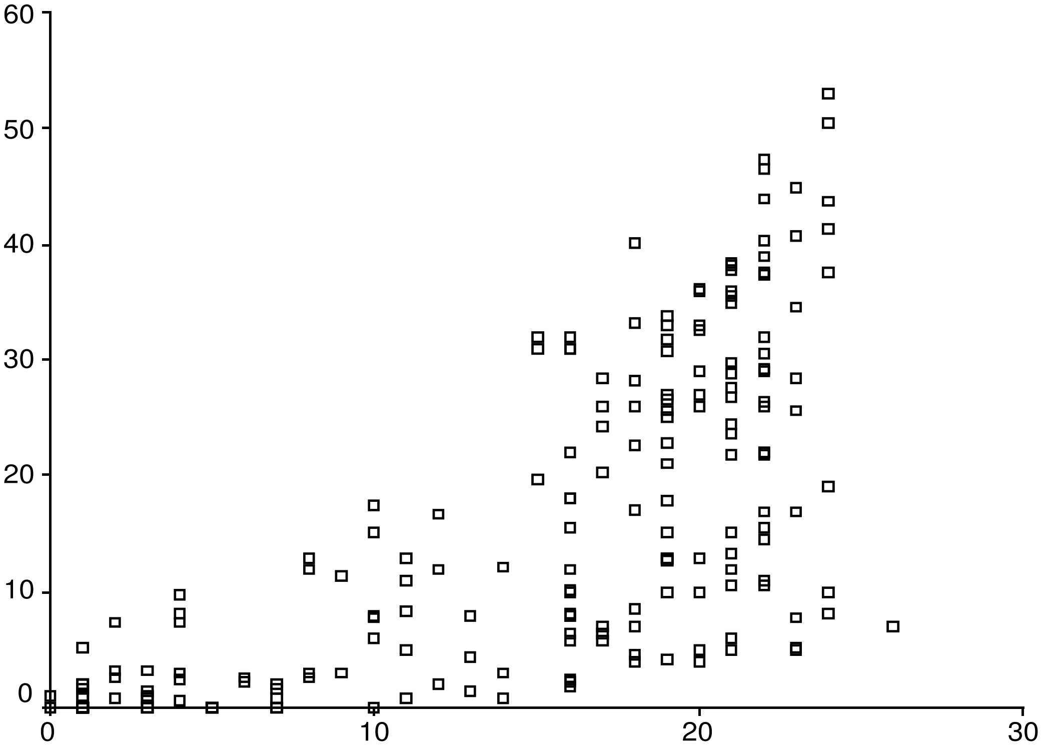


logit prevalence vs MAX = max NDVI

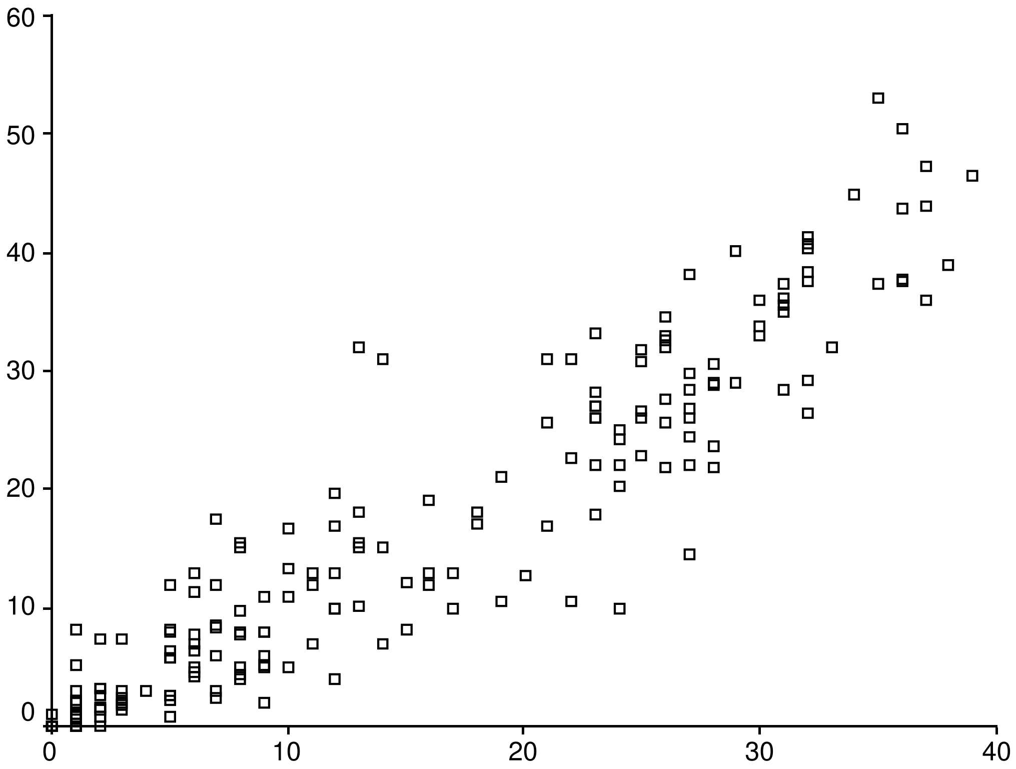


Comparing non-spatial and spatial predictions in Cameroon

Non-spatial



Spatial



Probabilistic prediction in Cameroon

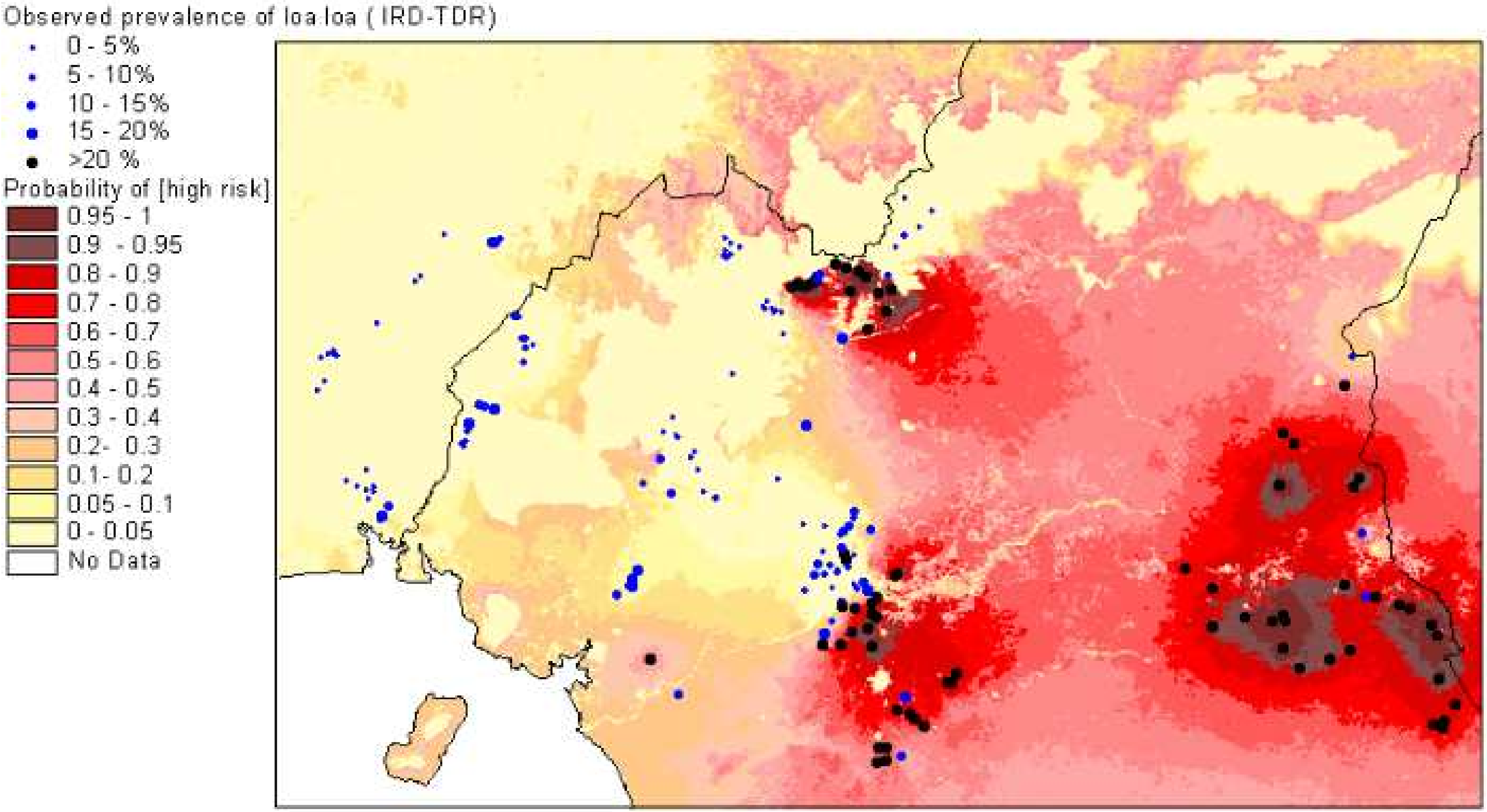


Figure 6: PCM for [high risk] in Cameroon based on ERMr with ground truth data.

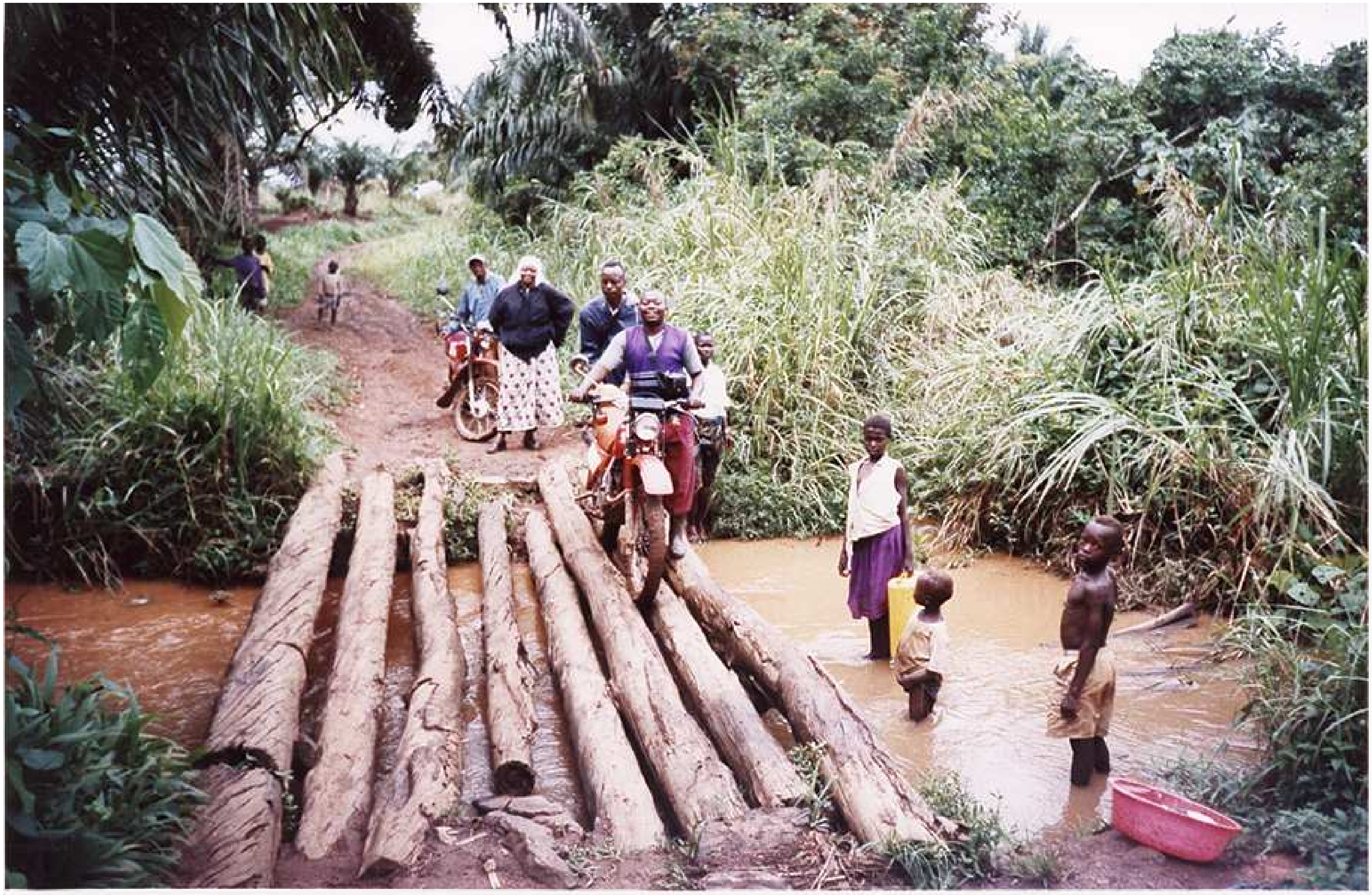
Next Steps

- analysis confirms value of local ground-truth prevalence data

- in some areas, need more ground-truth data to reduce predictive uncertainty

- but parasitological surveys are expensive

Field-work is difficult!



RAPLOA

- a cheaper alternative to parasitological sampling:
  - have you ever experienced eye-worm?
  - did it look like this photograph?
  - did it go away within a week?
- RAPLOA data to be collected:
  - in sample of villages previously surveyed parasitologically (to calibrate parasitology vs RAPLOA estimates)
  - in villages not surveyed parasitologically (to reduce local uncertainty)
- bivariate model needed for combined analysis of parasitological and RAPLOA prevalence estimates

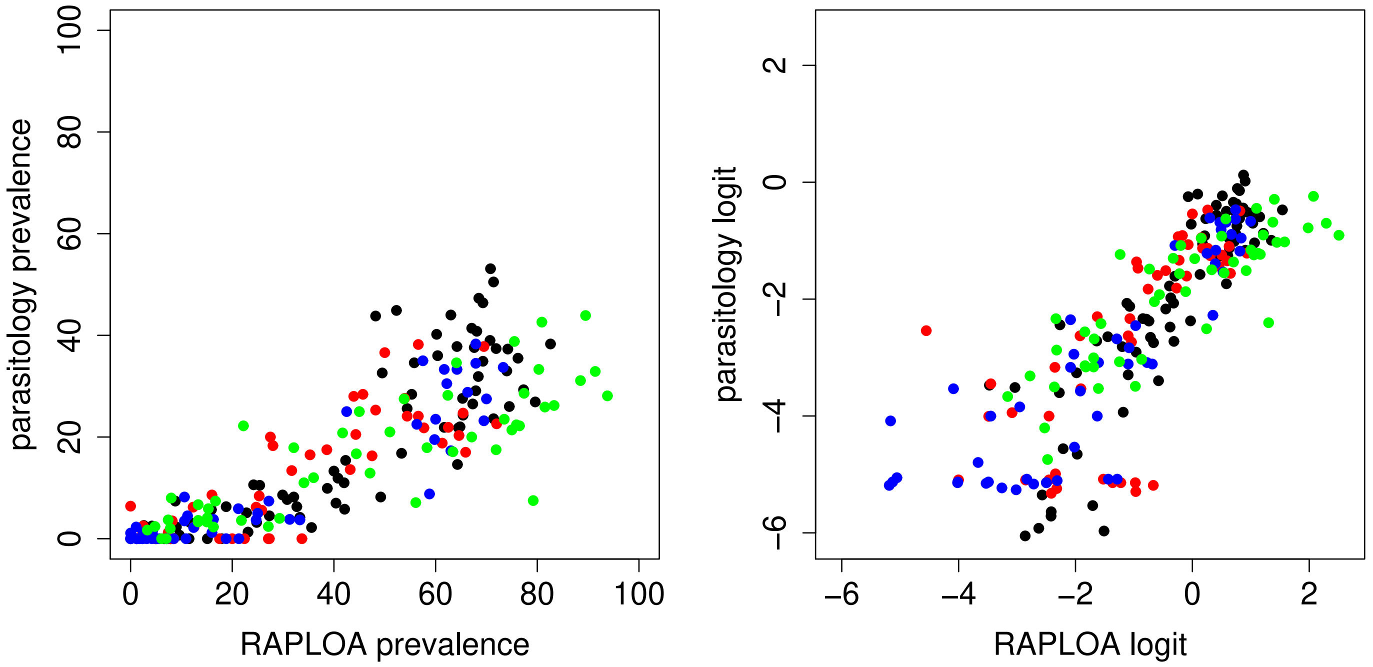
REPORT OF A MULTI-CENTRE STUDY

Rapid Assessment Procedures for Loiasis

UNDP/World Bank/WHO Special Programme for Research & Training in Tropical Disease (TDR)

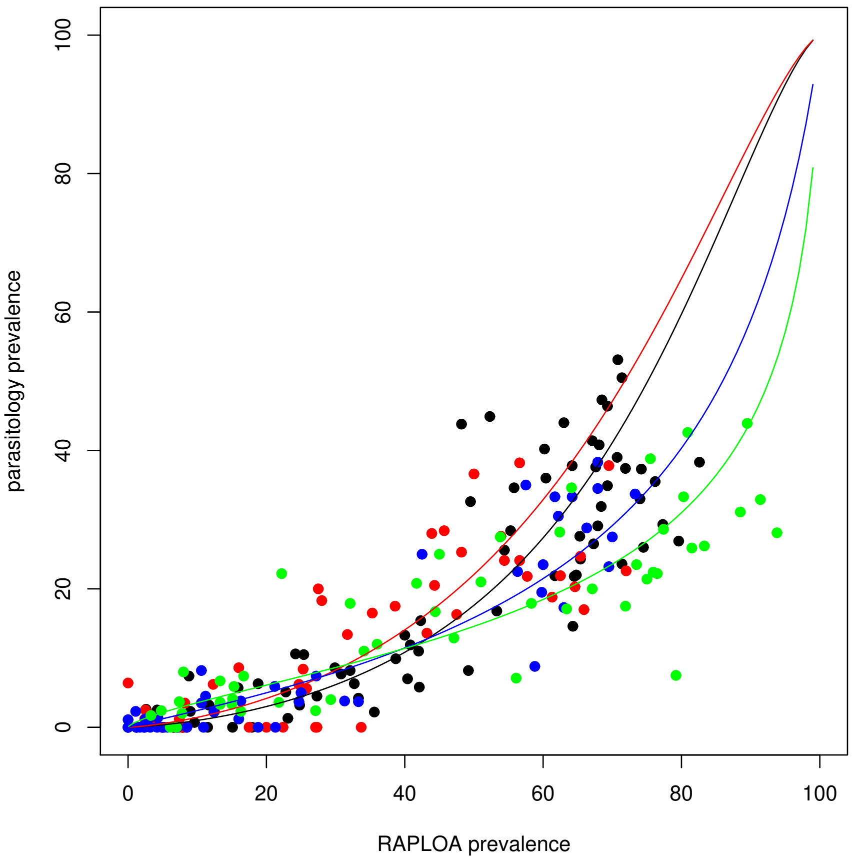
TDR/IDE/RP/RAPL/01.1

RAPLOA calibration



Empirical logit transformation linearises relationship
Colour-coding corresponds to four surveys in different regions

RAPLOA calibration (ctd)



Fit linear functional relationship on logit scale and back-transform

Parasitology/RAPLOA bivariate model

- **treat prevalence estimates as conditionally independent binomial responses**

- **with bivariate latent Gaussian process $\{S_1(x), S_2(x)\}$ in linear predictor**

- **to ease computation, write joint distribution as**

  $$[S_1(x), S_2(x)] = [S_1(x)][S_2(x)|S_1(x)]$$

  with low-rank spline representation of $S_1(x)$

Lecture 3

Geostatistical design; geostatistics and marked point processes

Section 5

Geostatistical design

Geostatistical design

- Retrospective

  Add to, or delete from, an existing set of measurement locations $x_i \in A : i = 1, ..., n$.

- Prospective

  Choose optimal positions for a new set of measurement locations $x_i \in A : i = 1, ..., n$.

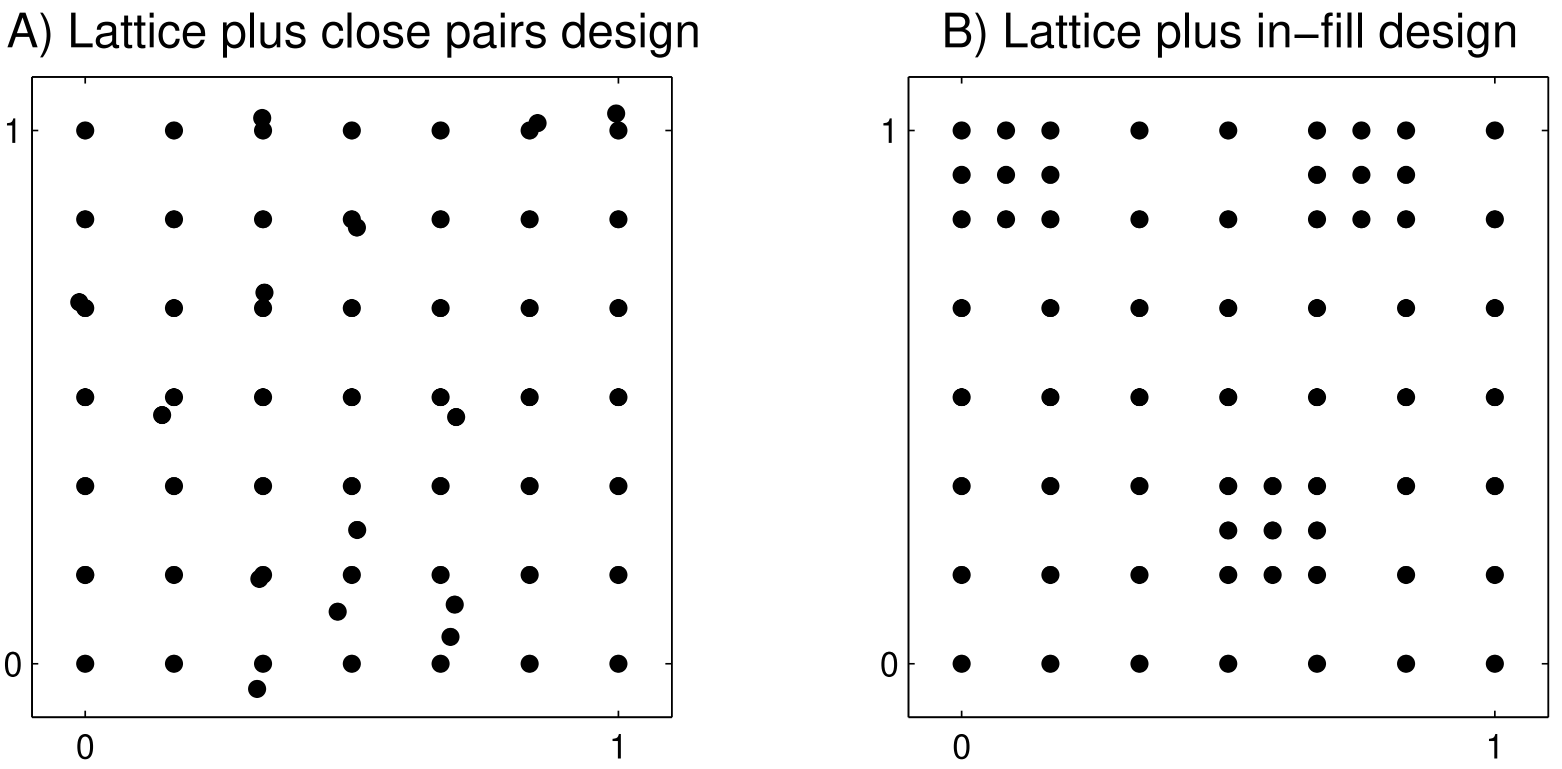
Naive design folklore

- **Spatial correlation decreases with increasing distance.**
- **Therefore, close pairs of points are wasteful.**
- **Therefore, spatially regular designs are a good thing.**

Less naive design folklore

- Spatial correlation decreases with increasing distance.
- Therefore, close pairs of points are wasteful if you know the correct model.
- But in practice, at best, you need to estimate unknown model parameters.
- And to estimate model parameters, you need your design to include a wide range of inter-point distances.
- Therefore, spatially regular designs should be tempered by the inclusion of some close pairs of points.

Examples of compromise designs



A Bayesian design criterion

Assume goal is prediction of $S(x)$ for all $x \in A$.

$$[S|Y] = \int [S|Y,\theta][\theta|Y]d\theta$$

For retrospective design, minimise

$$\bar{v} = \int_A \mathrm{Var}\{S(x)|Y\}dx$$

For prospective design, minimise

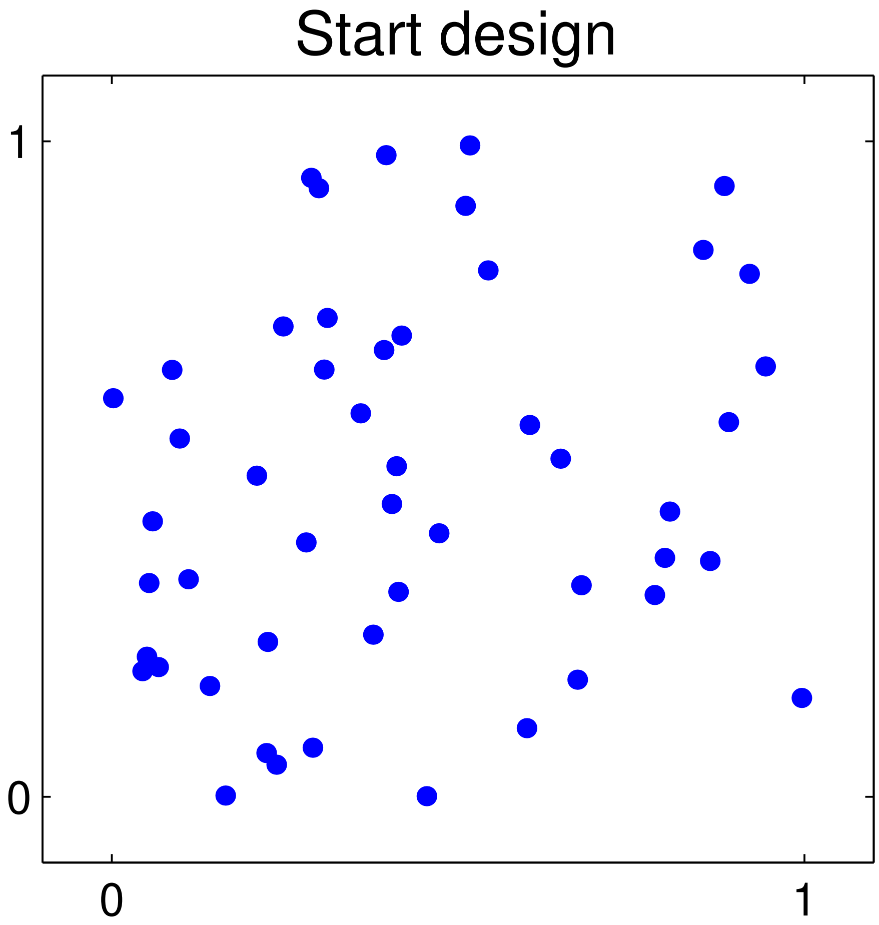
$$\mathrm{E}(\bar{v}) = \int_y \int_A \mathrm{Var}\{S(x)|y\}f(y)dy$$

where $f(y)$ corresponds to

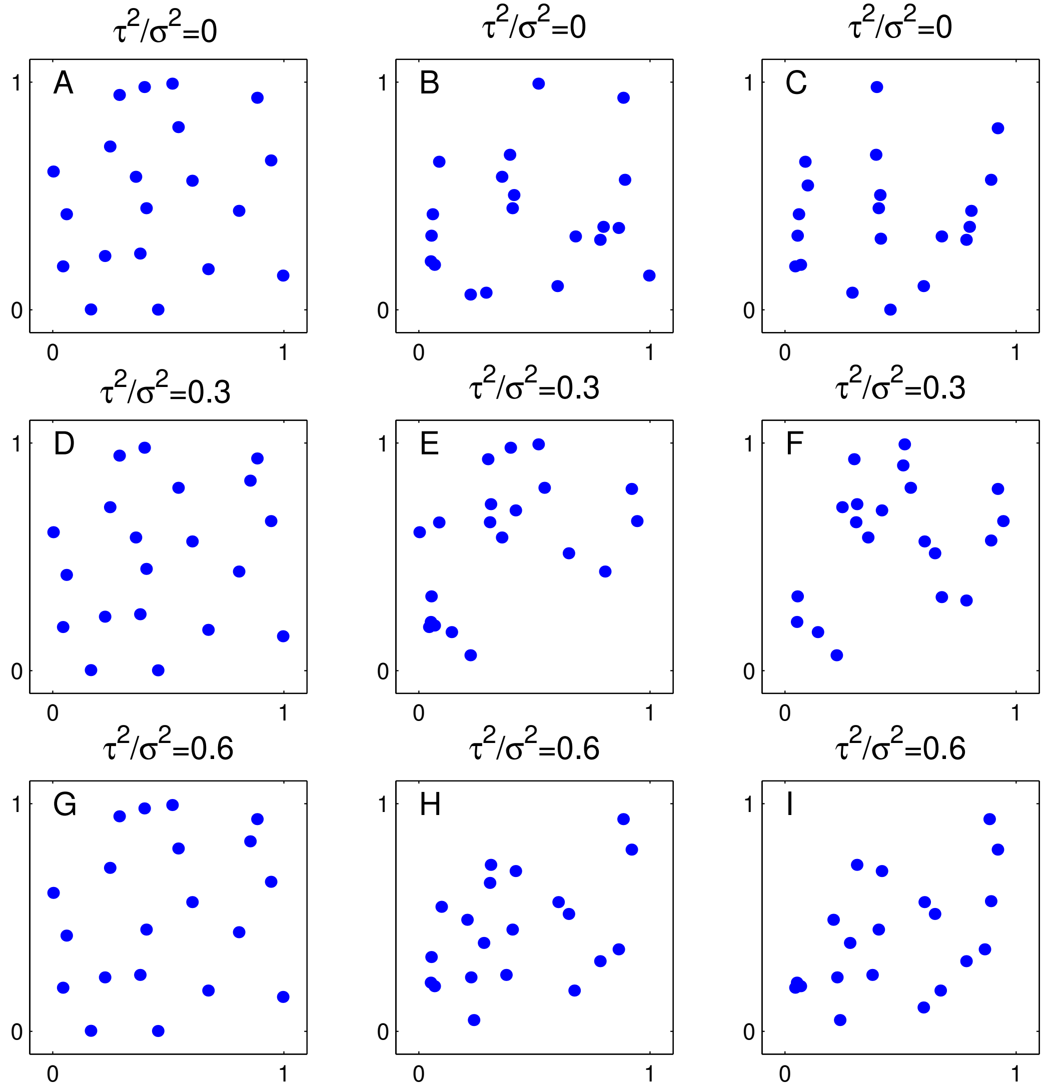
$$[Y] = \int [Y|\theta][\theta]d\theta$$

Results

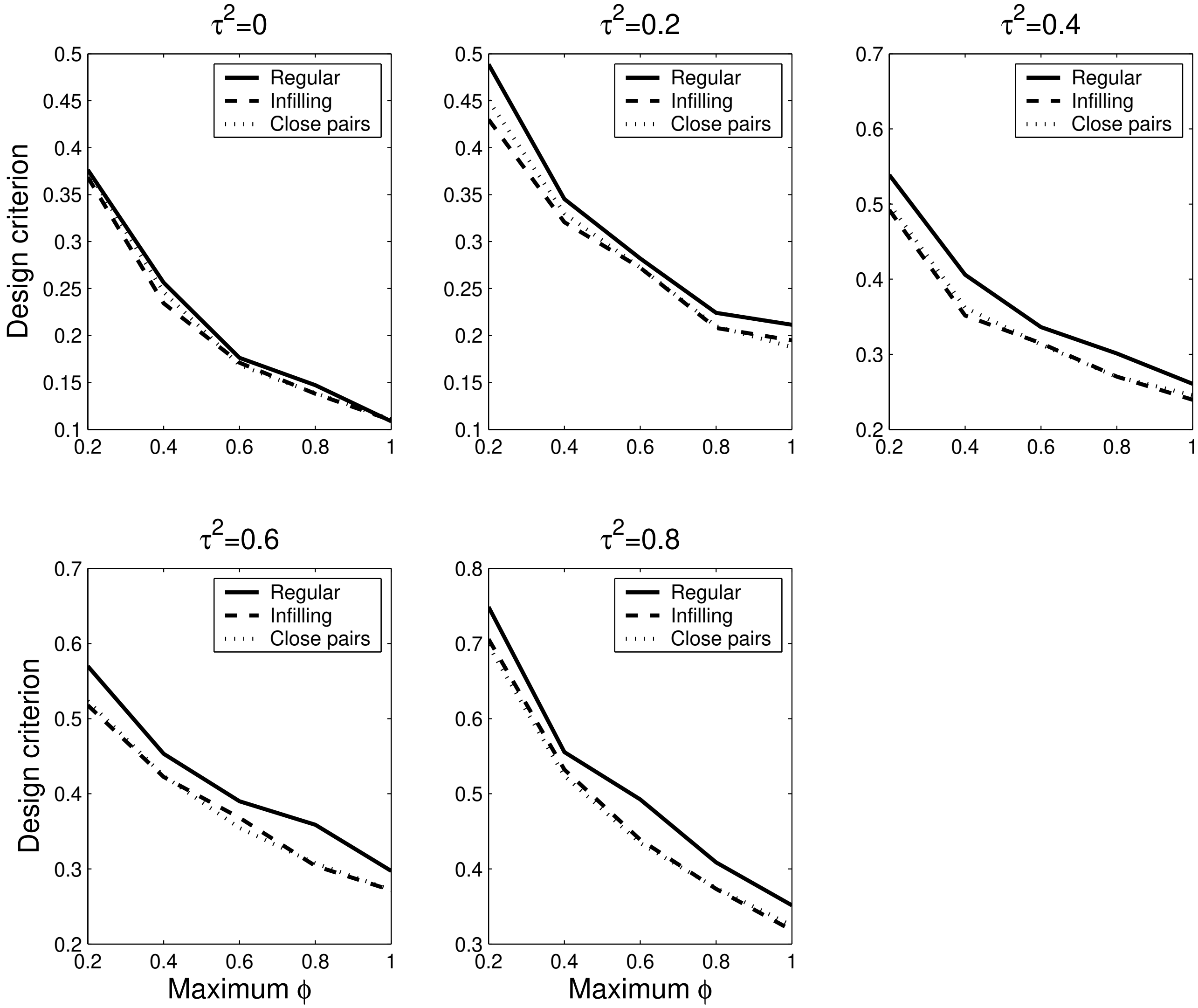
Retrospective: deletion of points from a monitoring network



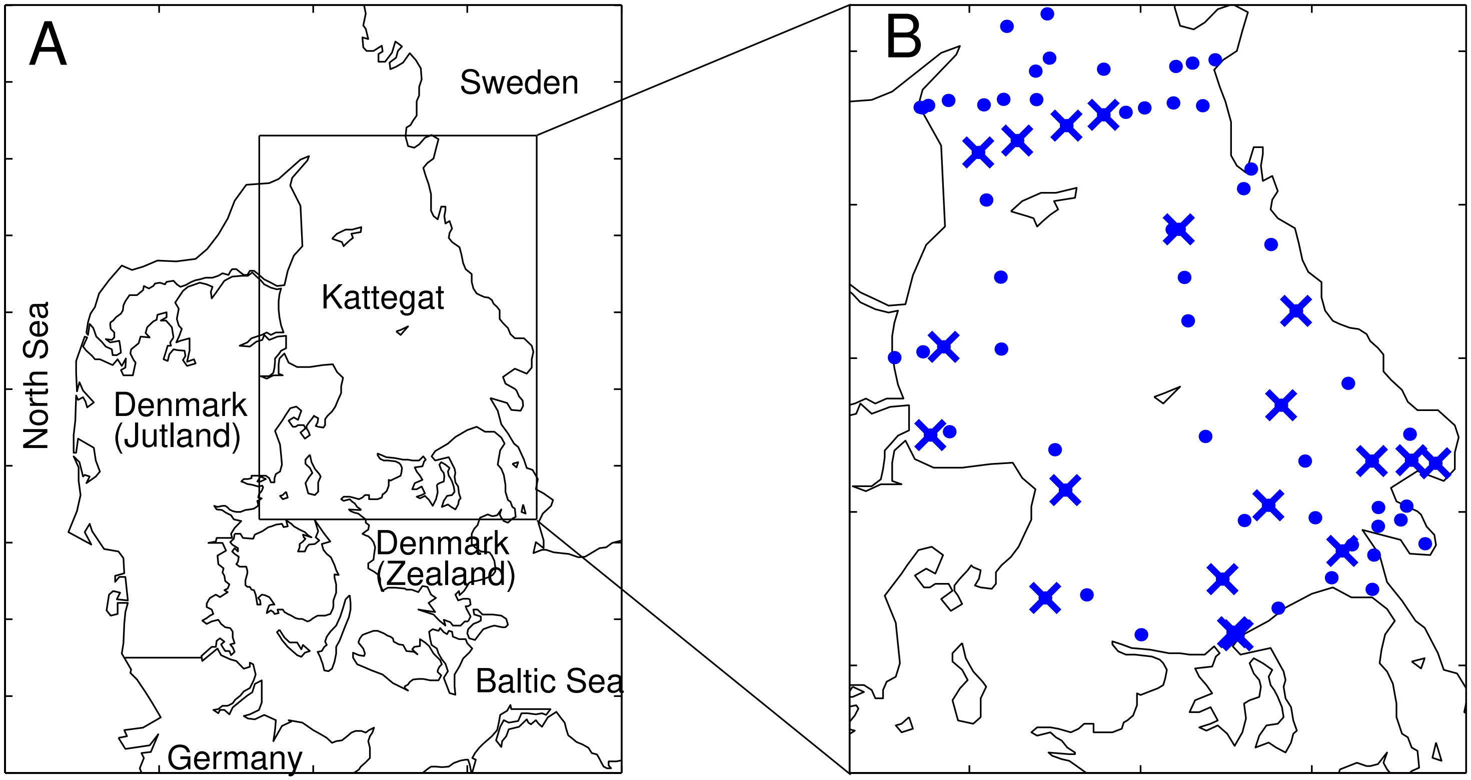
Selected final designs



Prospective: regular lattice vs compromise designs



Monitoring salinity in the Kattegat basin



Solid dots are locations deleted for reduced design.

Further remarks on geostatistical design

1. Conceptually more complex problems include:

   (a) design when some sub-areas are more interesting than others;

   (b) design for best prediction of non-linear functionals of $S(\cdot)$;

   (c) multi-stage designs (see below).

2. Theoretically optimal designs may not be realistic (eg Loa loa mapping problem)

3. Goal here is not optimal design, but to suggest constructions for good, general-purpose designs.

Section 6

Geostatistics and marked point processes

The geostatistical model re-visited

locations $\boldsymbol{X}$ signal $\boldsymbol{S}$ measurements $\boldsymbol{Y}$

- **Conventional geostatistical model:** $[\boldsymbol{S},\boldsymbol{Y}] = [\boldsymbol{S}][\boldsymbol{Y}|\boldsymbol{S}]$
- **What if $\boldsymbol{X}$ is stochastic?**

  Usual implicit assumption: $[\boldsymbol{X},\boldsymbol{S},\boldsymbol{Y}] = [\boldsymbol{X}][\boldsymbol{S}][\boldsymbol{Y}|\boldsymbol{S}]$

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

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

Marked point processes

locations $\boldsymbol{X}$ marks $\boldsymbol{Y}$

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

$$[\boldsymbol{X}, \boldsymbol{Y}] = [\boldsymbol{X}][\boldsymbol{Y}|\boldsymbol{X}] = [\boldsymbol{Y}][\boldsymbol{X}|\boldsymbol{Y}]$$

Preferential sampling

locations $\boldsymbol{X}$ signal $\boldsymbol{S}$ measurements $\boldsymbol{Y}$

- **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

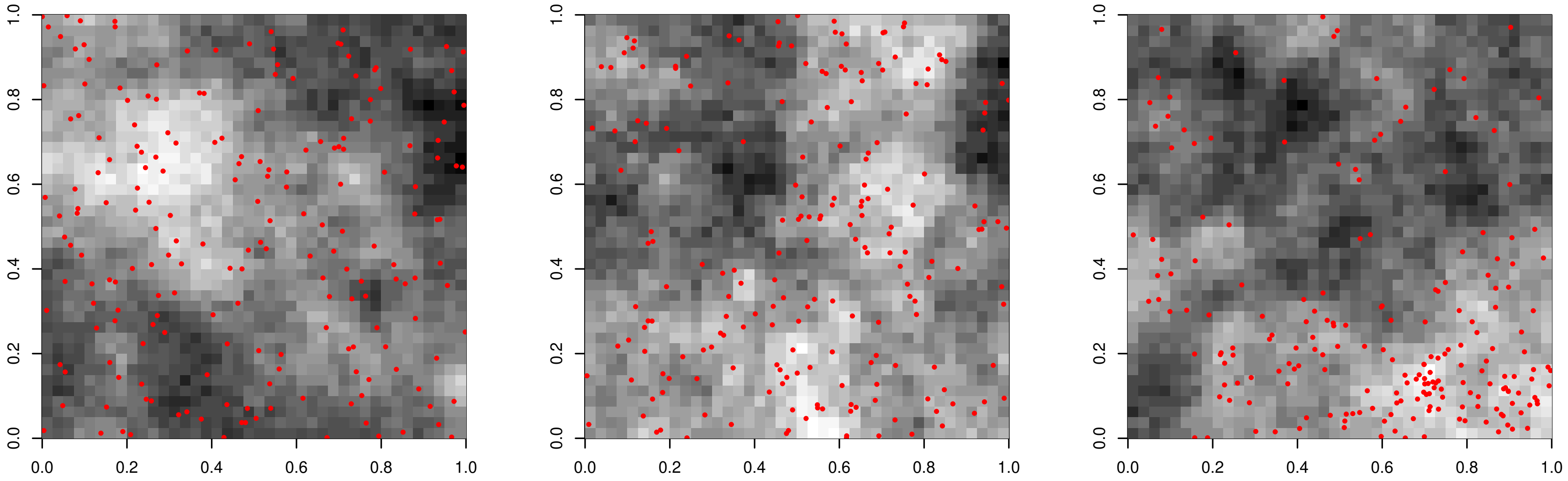
$$[\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{Y}] = [\boldsymbol{S}][\boldsymbol{X}|\boldsymbol{S}][\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{X}]$$

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

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

- $[\boldsymbol{Y}|\boldsymbol{S}, \boldsymbol{X}] = \prod_{i=1}^{n}[\boldsymbol{Y}_i|\boldsymbol{S}(\boldsymbol{X}_i)]$
- $[\boldsymbol{Y}_i|\boldsymbol{S}(\boldsymbol{X}_i)] = \mathrm{N}(\boldsymbol{S}(\boldsymbol{X}_i), \boldsymbol{\tau}^2)$

Simulation of preferential sampling model



$\boldsymbol{\beta = 0.0, 0.25, 0.5}$

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

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

- 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

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

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

$\Rightarrow$

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

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:

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

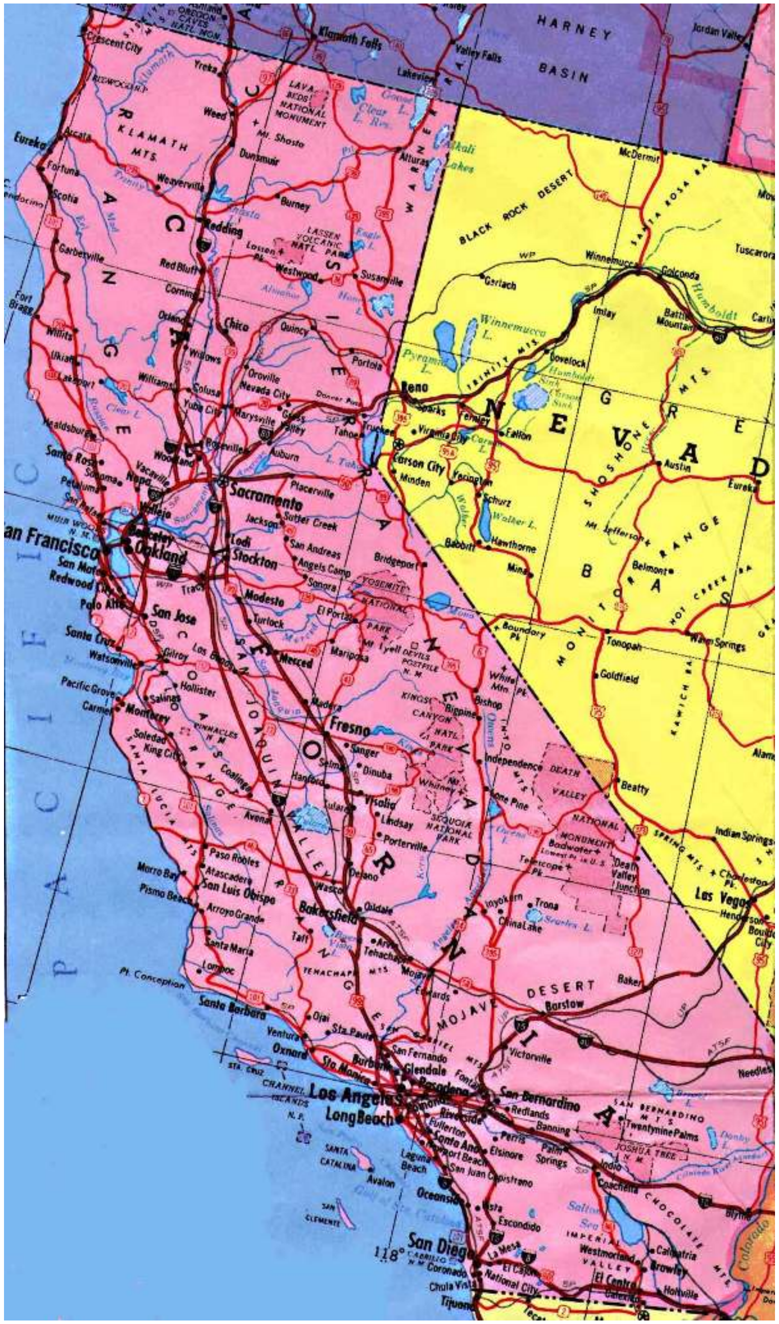
Practical solutions to weak identifability

1. **explanatory variables U to break dependence between S and X**

2. **strong Bayesian priors**

3. **two-stage sampling**

TIRES
REGULAR
PREMIUM
QUAKER STAT
MOTOR OI
CERTIFIED
GUARANTE


HARNEY BASIN
Jordan Valley
Crescent City
Klamath Falls
Valley Falls
Lakeview
Yreka
LAVA BEDS NATIONAL MONUMENT
Clear L. Res.
Goose L.
Weed
Mt. Shasta
KLAMATH MTS.
Eureka
Arcata
Dunsmuir
Alturas
Alkali Lakes
McDermitt
Fortuna
Scotia
Weaverville
Burney
BLACK ROCK DESERT
Redding
LASSEN VOLCANIC NAT'L PARK
Eagle L.
Garberville
Red Bluff
Westwood
Susanville
Winnemucca
Golconda
Tuscarora
Gerlach
Corning
Humboldt
Fort Bragg
Chico
Quincy
Imlay
Battle Mountain
Winnemucca L.
Willits
Orland
Willows
Portola
Pyramid L.
Lovelock
Ukiah
Oroville
Williams
Colusa
Nevada City
Reno
Sparks
Fernley
Lakeport
Clear L.
Yuba City
Marysville
Grass Valley
Truckee
Tahoe
Virginia City
Fallon
Healdsburg
Woodland
Roseville
Auburn
Carson City
Santa Rosa
Napa
Vacaville
Minden
Yerington
Austin
Eureka
Petaluma
Sonoma
L. Tahoe
Sacramento
Placerville
Schurz
Walker L.
San Rafael
Vallejo
Jackson
Sutter Creek
Hawthorne
MUIR WOODS N. M.
Berkeley
Lodi
Stockton
San Andreas
Mt. Jefferson
San Francisco
Oakland
Angels Camp
Bridgeport
Mina
San Mateo
Redwood City
Tracy
Modesto
Sonora
YOSEMITE NATIONAL PARK
Mono L.
Belmont
Palo Alto
San Jose
Turlock
El Portal
Boundary Pk.
Tonopah
Warm Springs
Santa Cruz
Los Banos
Merced
Mt. Lyell
DEVILS POSTPILE N. M.
Watsonville
Gilroy
Mariposa
White Mtn. Pk.
Goldfield
Pacific Grove
Hollister
Madera
Bishop
Carmel
Monterey
Salinas
KINGS CANYON NAT'L PARK
Big Pine
Soledad
PINNACLES N. M.
Fresno
King City
Sanger
Selma
Dinuba
Independence
DEATH VALLEY NATIONAL MONUMENT
Coalinga
Hanford
Visalia
Mt. Whitney
Lone Pine
Beatty
Avenal
Tulare
Lindsay
SEQUOIA NATIONAL PARK
Owens L.
Porterville
Badwater
Lowest Pt. in U.S.
Telescope Pk.
Death Valley Junction
Indian Springs
Paso Robles
Delano
Charleston Pk.
Morro Bay
Atascadero
Wasco
Las Vegas
Pismo Beach
San Luis Obispo
Arroyo Grande
Bakersfield
Oildale
Inyokern
Trona
Searles L.
China Lake
Henderson
Boulder City
Taft
Arvin
Tehachapi
Santa Maria
TEHACHAPI MTS.
Lompoc
Mojave
Pt. Conception
Santa Barbara
Edwards
Baker
MOJAVE DESERT
Barstow
Ojai
Ventura
Sta Paula
Oxnard
Victorville
Needles
San Fernando
Burbank
Glendale
Sta Monica
Pasadena
Los Angeles
San Bernardino
Redlands
STA. CRUZ
CHANNEL ISLANDS N. M.
Long Beach
Riverside
Banning
SAN BERNARDINO MTS.
Twentynine Palms
JOSHUA TREE N. M.
Fullerton
Perris
Santa Ana
Elsinore
Palm Springs
Newport Beach
SANTA CATALINA
Avalon
Laguna Beach
San Juan Capistrano
Indio
Coachella
CHOCOLATE MTS.
Oceanside
SAN CLEMENTE
Salton Sea
Blythe
Vista
Escondido
IMPERIAL VALLEY
118°
San Diego
La Mesa
Westmorland
Calipatria
Brawley
El Cajon
Coronado
National City
El Centro
Calexico
Holtville
Chula Vista
Tijuana
Colorado
PACIFIC
CALIFORNIA
COAST RANGES
SIERRA NEVADA
SAN JOAQUIN VALLEY
SANTA LUCIA RANGE
DIABLO RANGE
NEVADA
GREAT BASIN
SHOSHONE MTS.
TOIYABE RANGE
MONITOR RANGE
HOT CREEK RANGE
KAWICH RANGE
SPRING MTS.
WARNER MTS.
TRINITY MTS.
SANTA ROSA RA.

Ozone monitoring in California

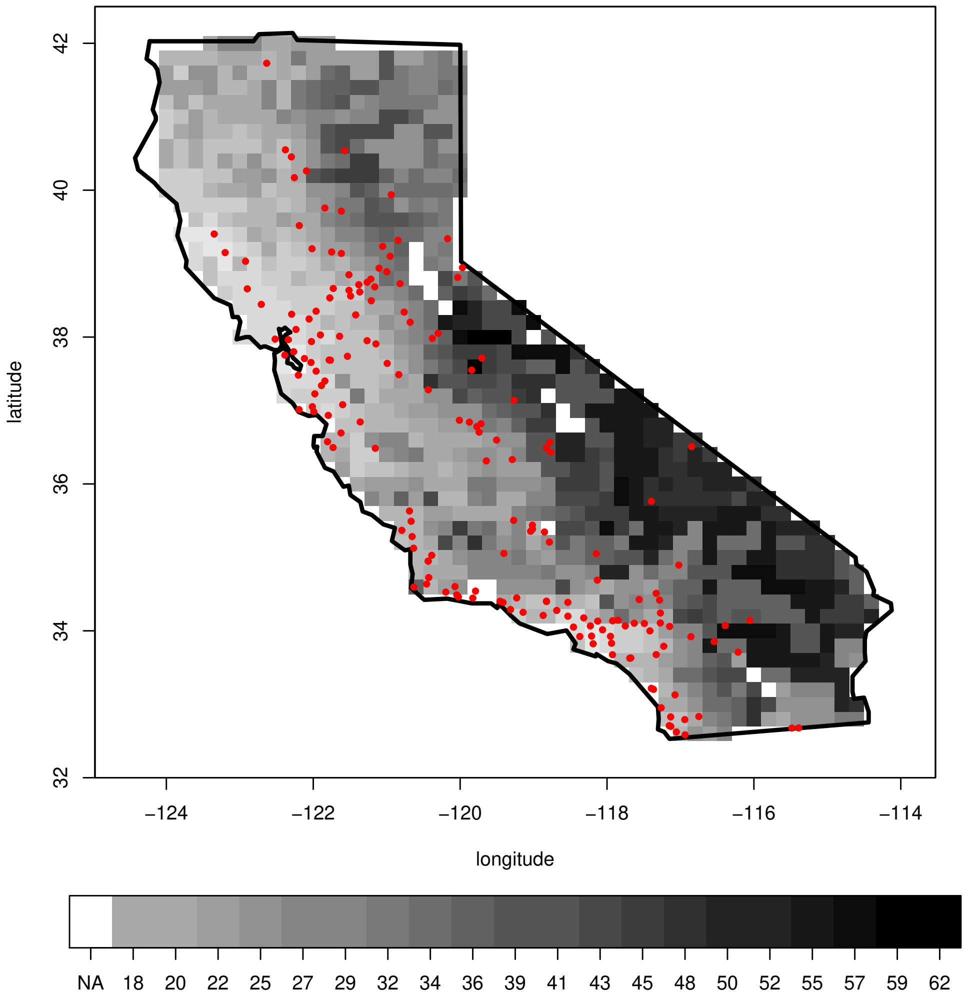
Data:

- 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 $\boldsymbol{U}$,

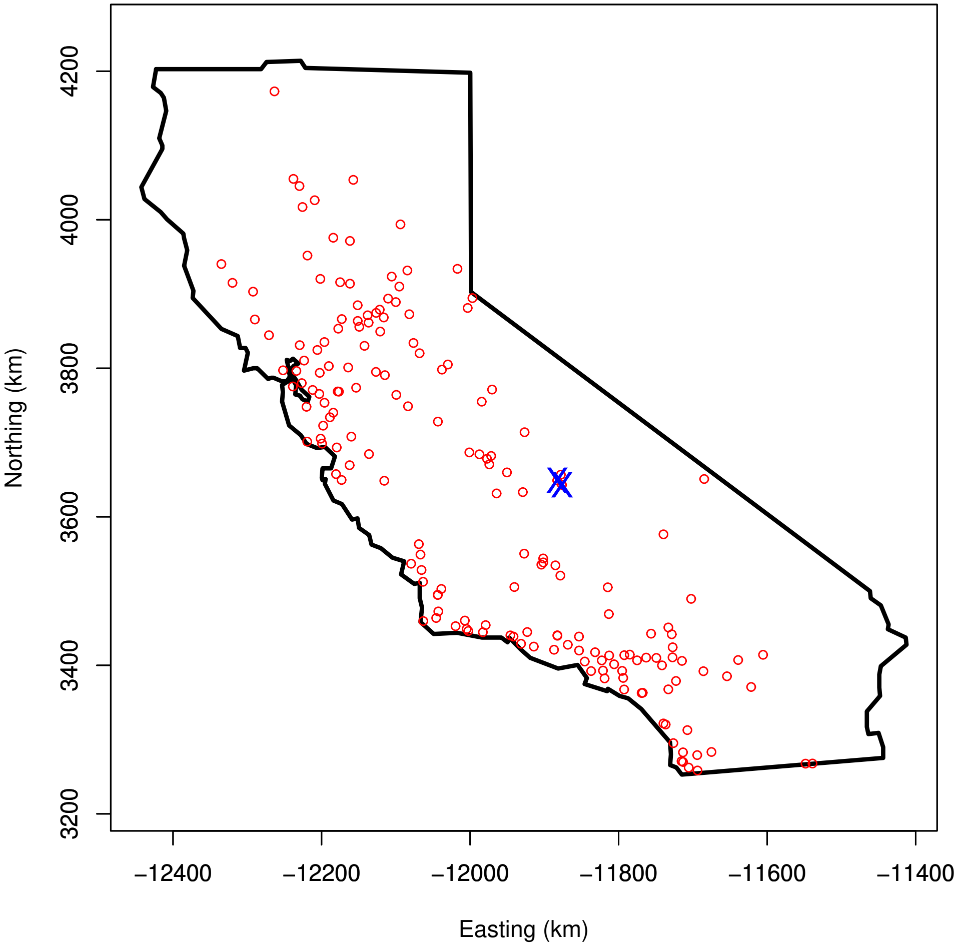
$$[\boldsymbol{X},\boldsymbol{S},\boldsymbol{Y}] = \int [\boldsymbol{X}|\boldsymbol{U}][\boldsymbol{S}|\boldsymbol{U}][\boldsymbol{Y}|\boldsymbol{S},\boldsymbol{U}][\boldsymbol{U}]d\boldsymbol{U}$$

- if $\boldsymbol{U}$ observed:
  - use conditional likelihood,

$$[\boldsymbol{X},\boldsymbol{S},\boldsymbol{Y}|\boldsymbol{U}] = [\boldsymbol{X}|\boldsymbol{U}][\boldsymbol{S}|\boldsymbol{U}][\boldsymbol{Y}|\boldsymbol{S},\boldsymbol{U}]$$

  - and ignore term $[\boldsymbol{X}|\boldsymbol{U}]$ for inference about $\boldsymbol{S}$

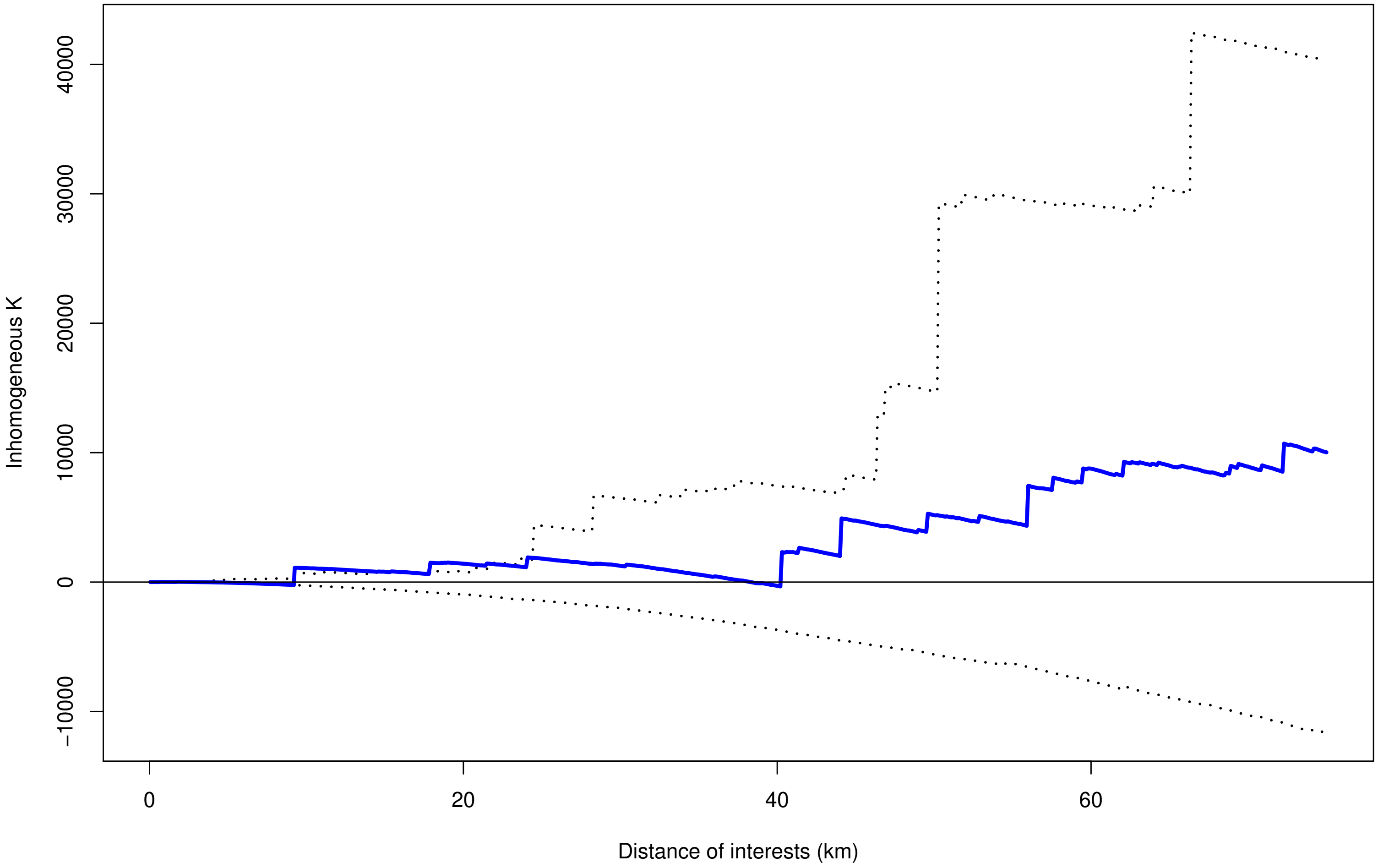
California ozone monitors: outlier?

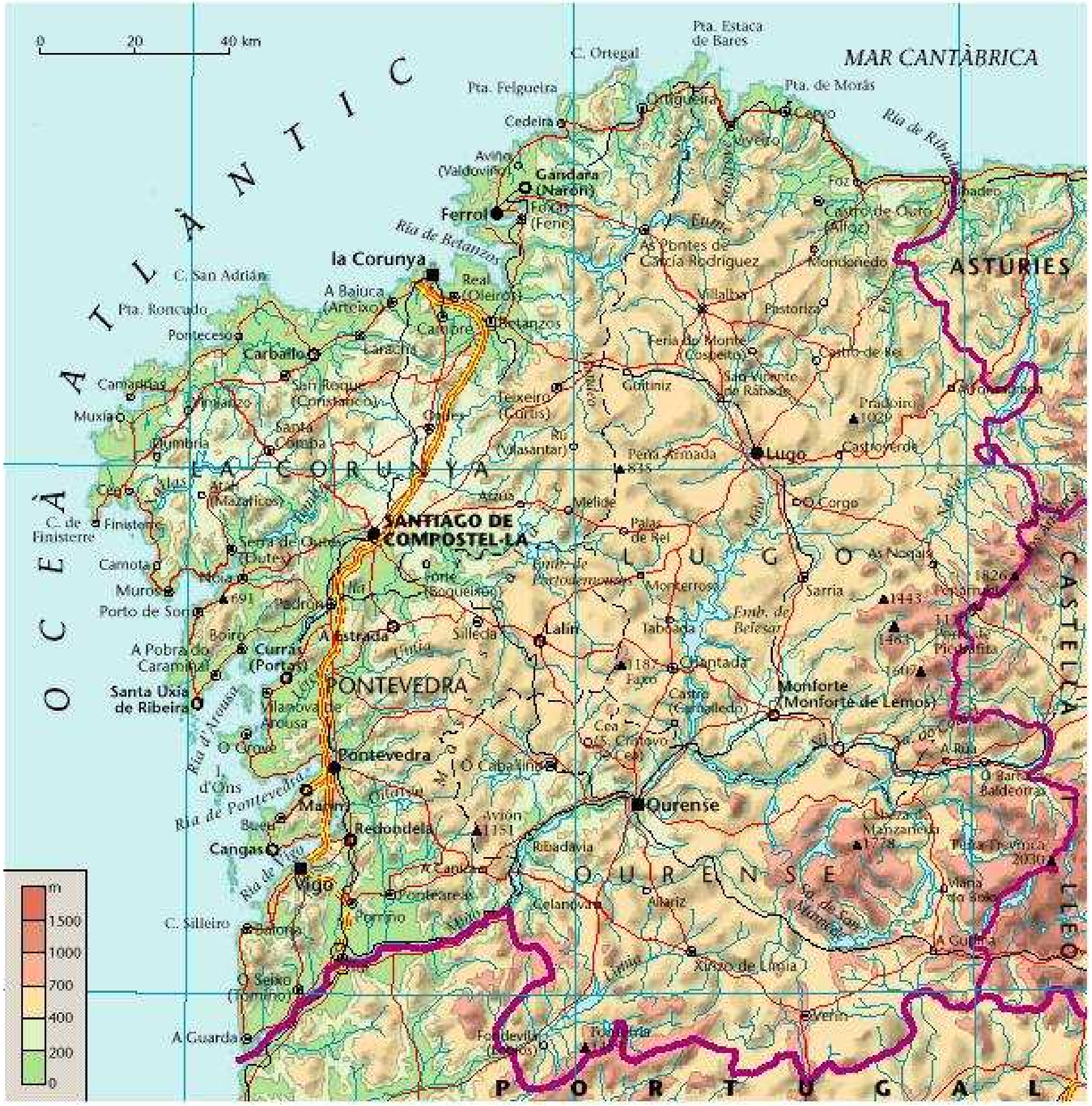


Analysis of California ozone monitor locations

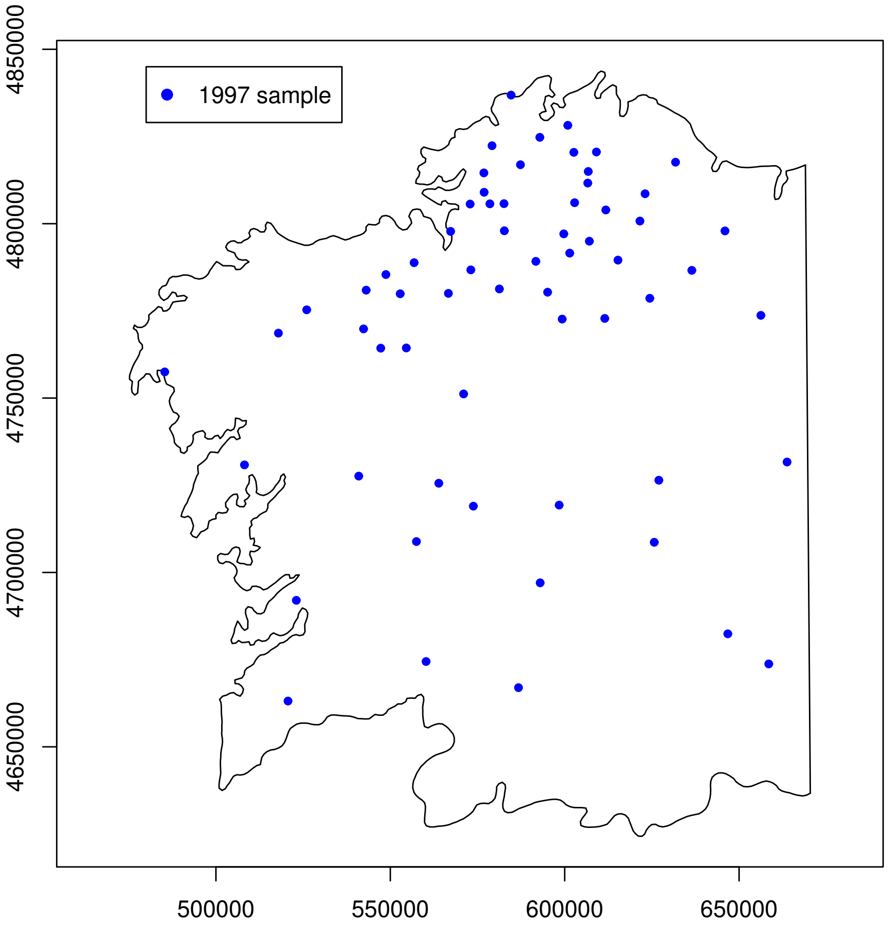
- monitor intensity associated with:
  - population density (positive)
  - percentage College-educated (positive)
  - median family income (negative)
- good fit to inhomogeneous Poisson process model (after removal of one outlier)

California ozone monitors: fit to Poisson model

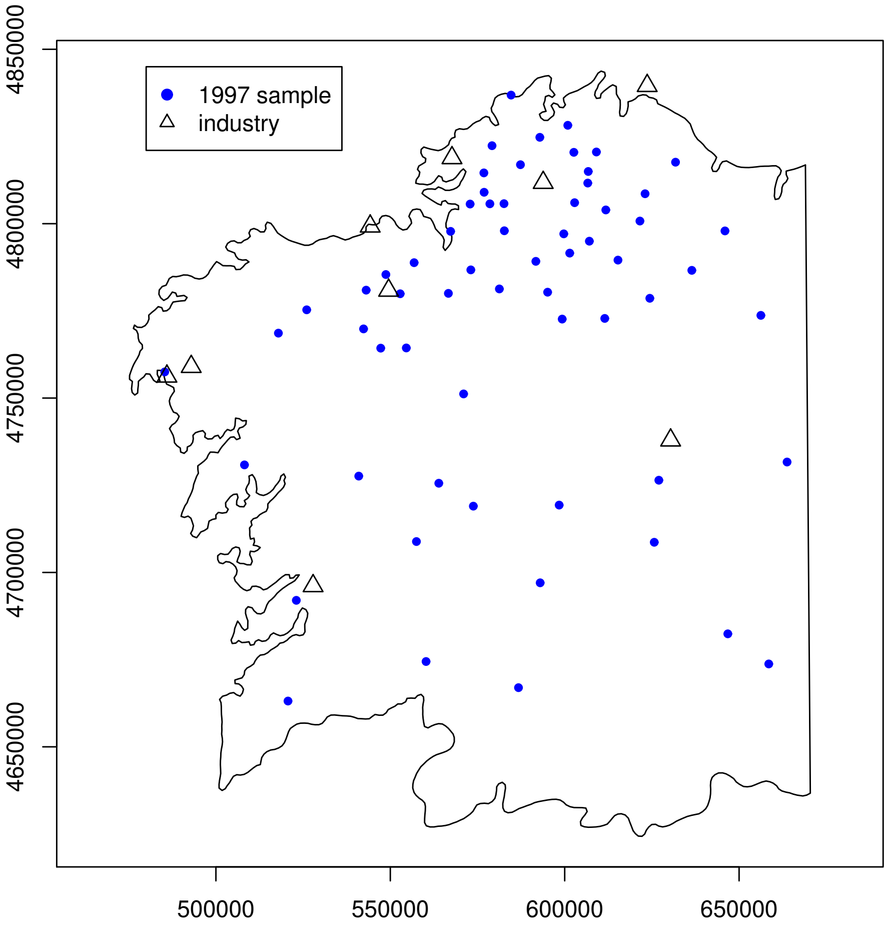



0
20
40 km
OCEÀ ATLÀNTIC
MAR CANTÀBRICA
C. Ortegal
Pta. Estaca de Bares
Pta. Felgueira
Pta. de Morás
Ortigueira
Cariño
Viveiro
Cedeira
Aviño (Valdoviño)
Gándara (Narón)
Ferrol
Foz
Ribadeo
Ría de Ribadeo
Fene
Ría de Betanzos
Castro de Oro (Alfoz)
As Pontes de García Rodríguez
Mondoñedo
ASTÚRIES
la Corunya
C. San Adrián
Pta. Roncudo
A Bajuca (Arteixo)
Real (Oleiros)
Betanzos
Cambre
Ponteceso
Carballo
Laracha
Villalba
Pastoriza
Feria do Monte Cospeito
Castro de Rei
Camariñas
Muxía
San Roque (Coristanco)
Teixeiro (Curtis)
Guitiriz
San Vicente de Rábade
Pradoiro 1020
Santa Comba
Ordes
Dumbría
Rú (Vilasantar)
Pena Armada
Lugo
Castroverde
Cee
LA CORUNYA
Arzúa
Melide
Palas de Rei
O Corgo
C. de Finisterre
Finisterre
SANTIAGO DE COMPOSTEL·LA
Emb. de Portodemouros
LUGO
As Nogais
Camota
Serra de Outes
Muros
Noia
Ponte (Boqueixón)
Monterroso
Sarria
1826
Porto de Son
Padrón
Emb. de Belesar
1443
Piedrafita
Boiro
A Estrada
Silleda
Lalín
Taboada
1463
A Pobra do Caramiñal
Curras (Porto)
Chantada
Santa Uxía de Ribeira
PONTEVEDRA
Faro
Carballedo
Monforte (Monforte de Lemos)
Vilanova de Arousa
Ría d'Arousa
O Grove
Cea
Ribadavia
Pontevedra
O Carballiño
A Rúa
O Barco
I. d'Ons
Marín
Ourense
Ría de Pontevedra
Bueu
Redondela
Avión 1151
Manzaneda 1778
Pena Trevinca 2030
Cangas
Ría de Vigo
Vigo
OURENSE
Maceda
C. Silleiro
Baiona
Ponteareas
Porriño
Celanova
Allariz
Serra de San Mamede
A Gudiña
CASTELLA I LLEÓ
O Seixo (Tomiño)
Xinzo de Limia
Verín
A Guarda
Fondevila
PORTUGAL
m
1500
1000
700
400
200
0

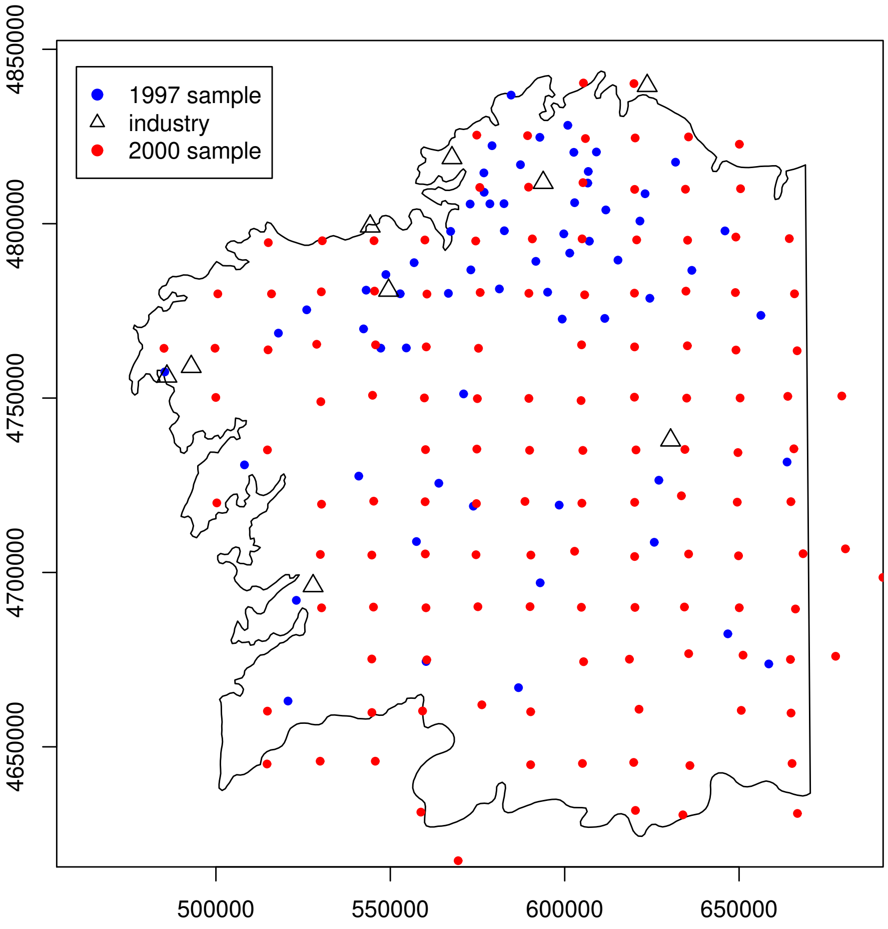
Heavy metal bio-monitoring in Galicia



Heavy metal bio-monitoring in Galicia



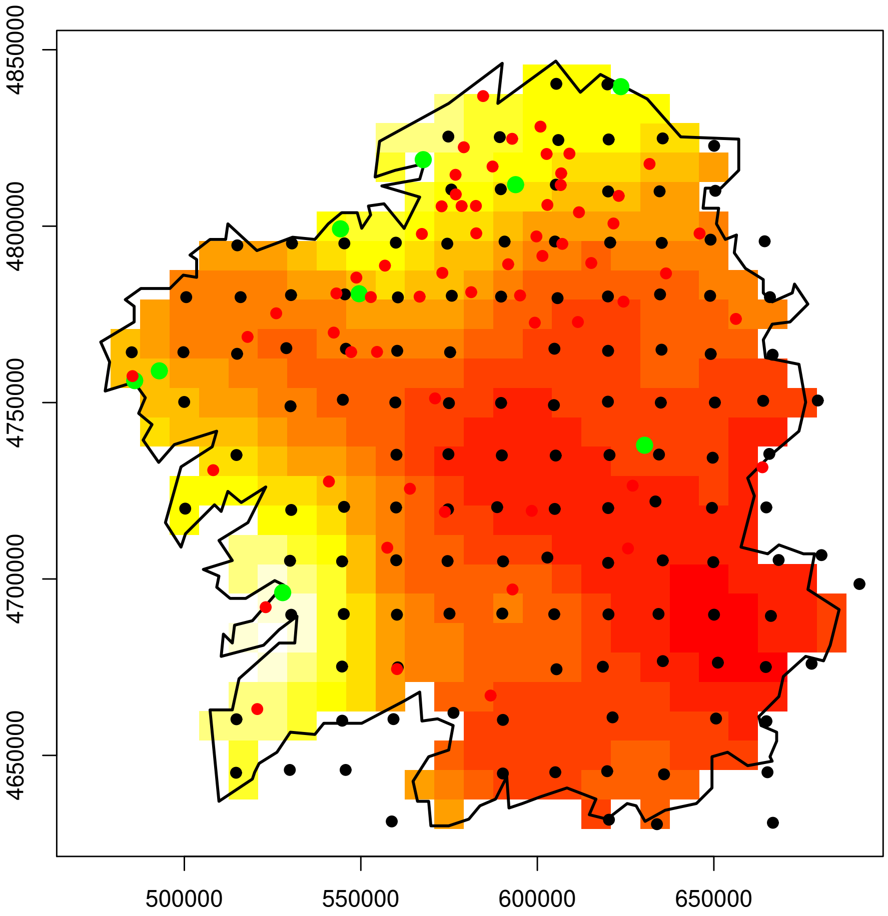
Heavy metal bio-monitoring in Galicia



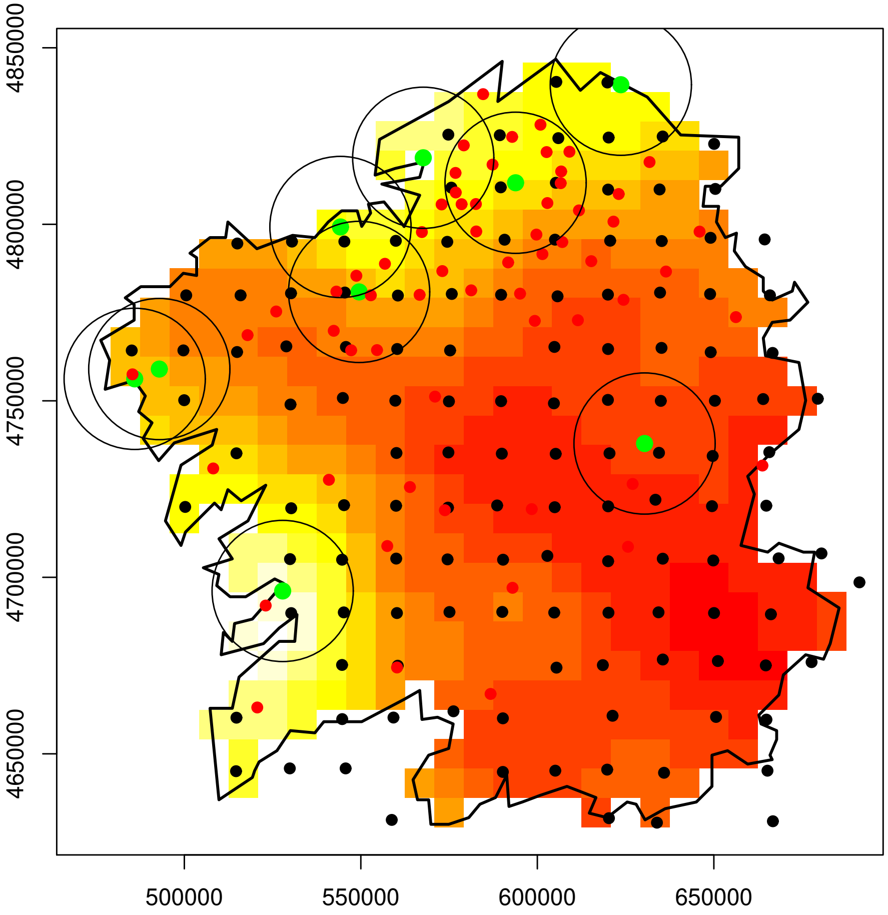
Heavy metal bio-monitoring in Galicia

- 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.

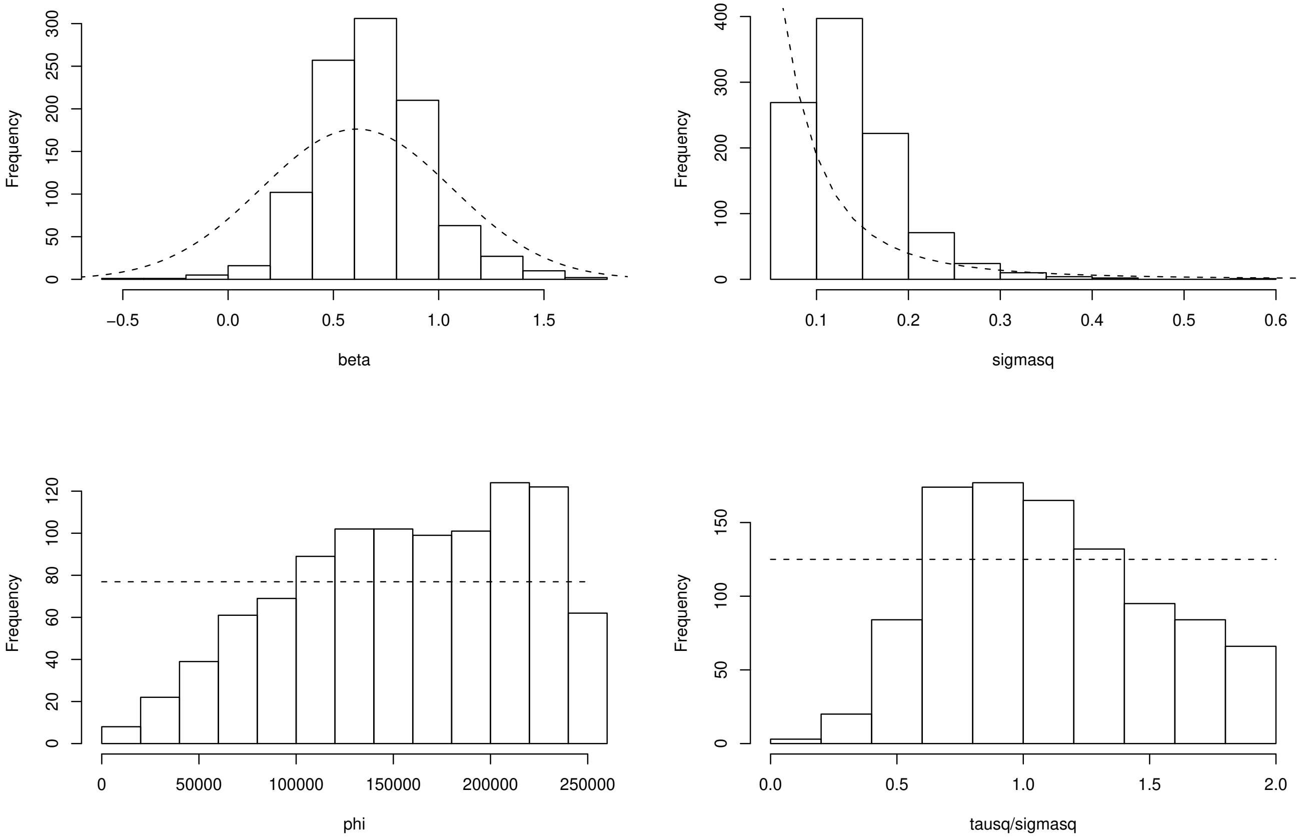
Galicia: 2000 predictions (posterior mean)



Galicia: excision of areas close to industry



Galicia: posteriors from analysis of 2000 data



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

Galicia: analysis of 1997 data

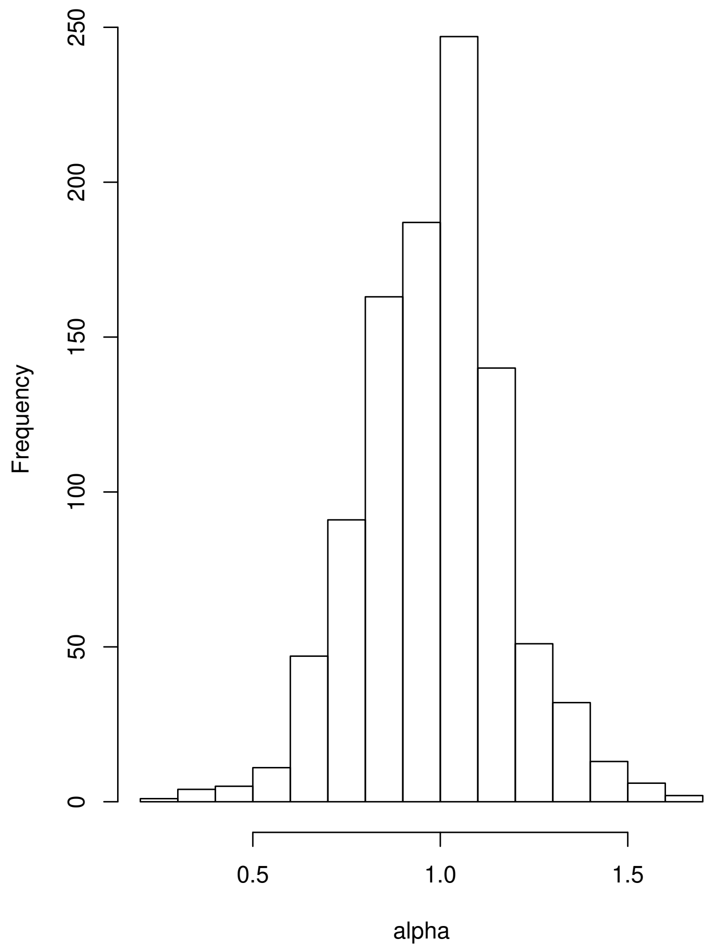
- introduce distance to nearest industry as explanatory variable,
$$\mu(x) = \mu_0 + \alpha + \gamma d(x)$$

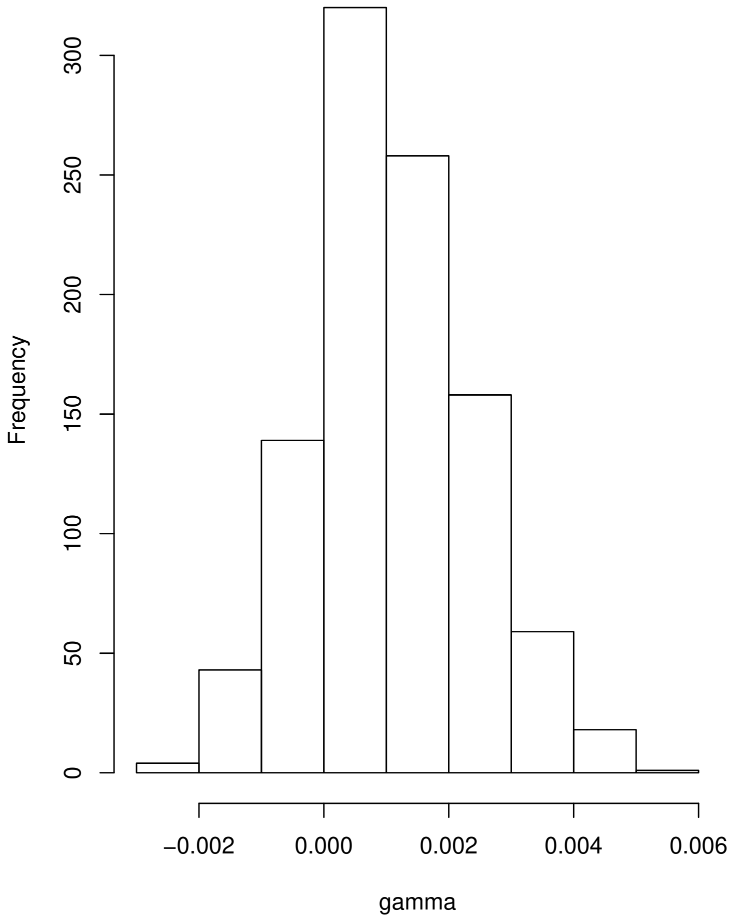
- for β and spatial covariance parameters, use posteriors from 2000 analysis

- resulting posterior mean and SD for (α, γ)

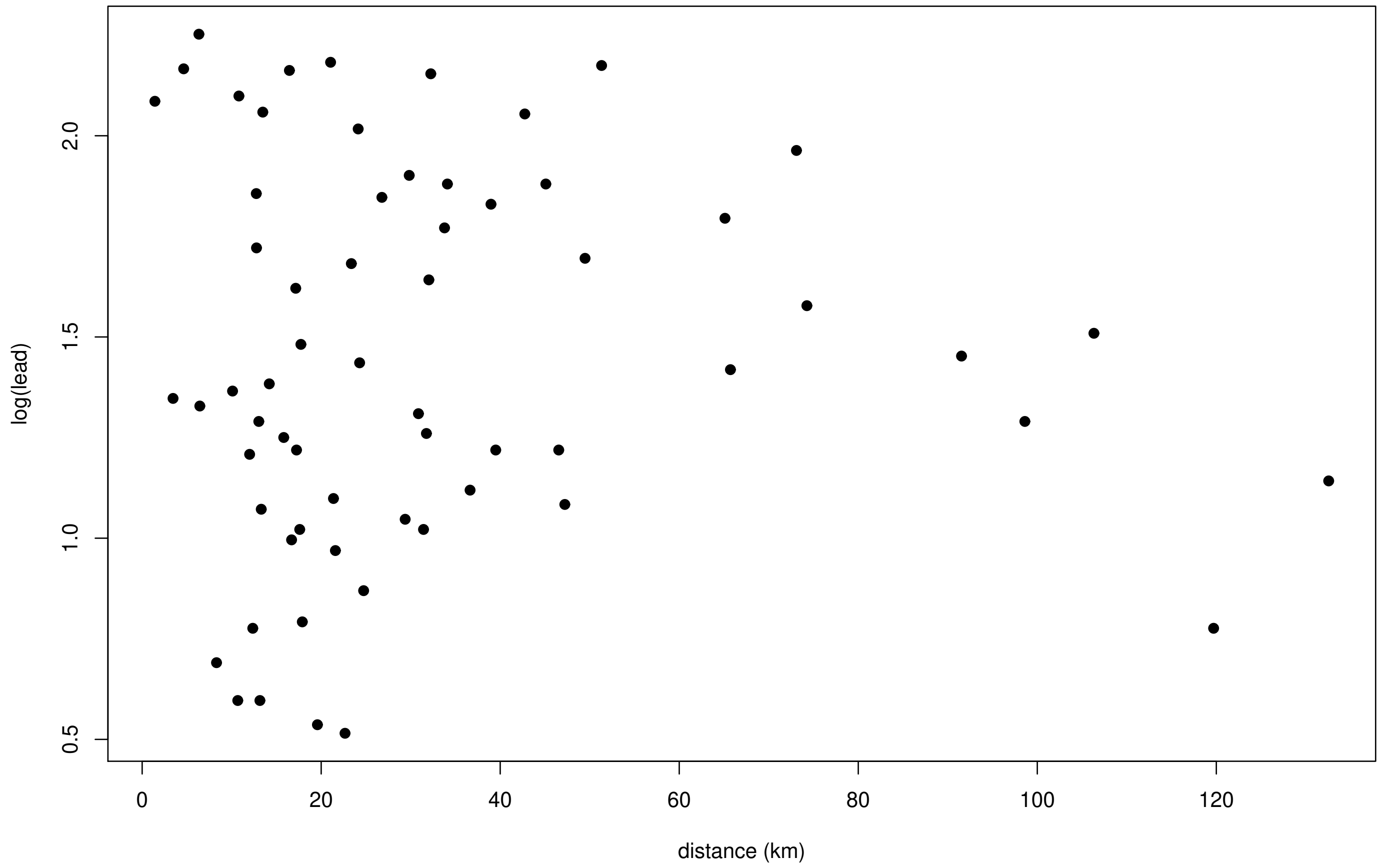
	α	γ
mean	0.601	-0.000561
SD	0.179	0.000609

Galicia: posteriors for (α, γ)





Galicia: a cautionary note



Suggests missing explanatory variable(s)?

Closing remarks on preferential sampling

- preferential sampling is widespread in practice, but almost universally ignored

- its effects may or may not be innocuous

- model parameters may be poorly identifed, 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

Closing remarks on model-based geostatistics

- Parameter uncertainty can have a material impact on prediction.

- Bayesian paradigm deals naturally with parameter uncertainty.

- Implementation through MCMC is not wholly satisfactory:

  – sensitivity to priors?

  – convergence of algorithms?

  – routine implementation on large data-sets?

- **Model-based approach clarifies distinctions between:**
  - **the substantive problem;**
  - **formulation of an appropriate model;**
  - **inference within the chosen model;**
  - **diagnostic checking and re-formulation.**
- **Areas of current research include:**
  - **preferential sampling**
  - **computational issues around large data-sets**
  - **multivariate models**
  - **spatio-temporal models**

- **Analyse problems, not data:**
  - **what is the scientific question?**
  - **what data will best allow us to answer the question?**
  - **what is a reasonable model to impose on the data?**
  - **inference: avoid** *ad hoc* **methods if possible**
  - **fit, reflect, re-formulate as necessary**
  - **answer the question.**