Métodos Geoestatísticos: aplicações e desafios

ou

geoestatística do ponto de vista de modelagem estatística

Paulo Justiniano Ribeiro Junior

Depto de Estatística
Universidade Federal do Paraná
&
Dept Maths & Stats
Lancaster University, UK

9^o SEAGRO & 46^a RBRAS
ESALQ/USP, Piracicaba - SP
10 Julho 2001

Agradecimentos

- Comissão Organizadora
- Depto Exatas, ESALQ/USP
- CAPES
- Prof. Peter Diggle (Lancaster University)

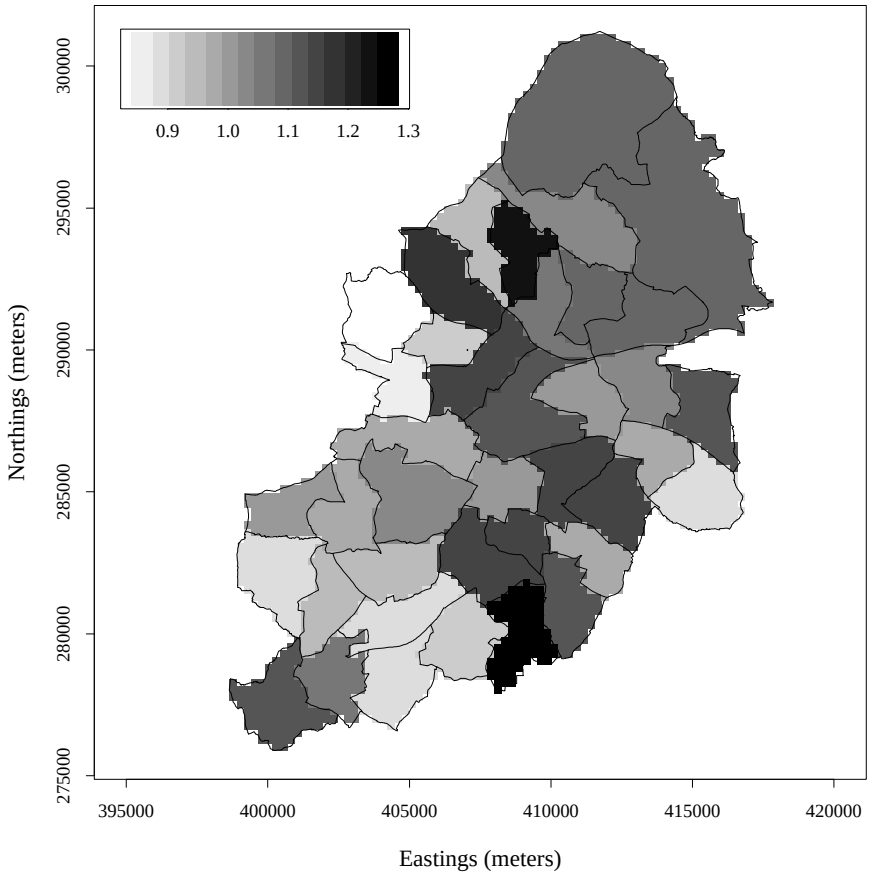
Objetivos

- desmistificar a geoestatística
- mostrar conexões entre geoestatística e modelos usuais em estatística
- provar que problemas geoestatísticos podem ser expressos em linguagem comum à de areas como modelos lineares, lineares generalizados, mixtos, etc
- chamar atenção dos pesquisadores para conexões e oportunidades de integração e trabalho conjunto
- despertar atenção de jovens e/ou potenciais pesquisadores para métodos de estatística espacial
- enfatizar oportunidades nesta área de pesquisa que vem experimentando enorme desenvolvimento e interesse em todo o mundo

1. Estatística Espacial: Exemplos Básicos

(a) Taxas de câncer por regiões administrativas

tons de cinza correspondem à variação estimada do risco relativo de câncer coloretal em 36 zonas eleitorais da cidade de Birmingham, UK.

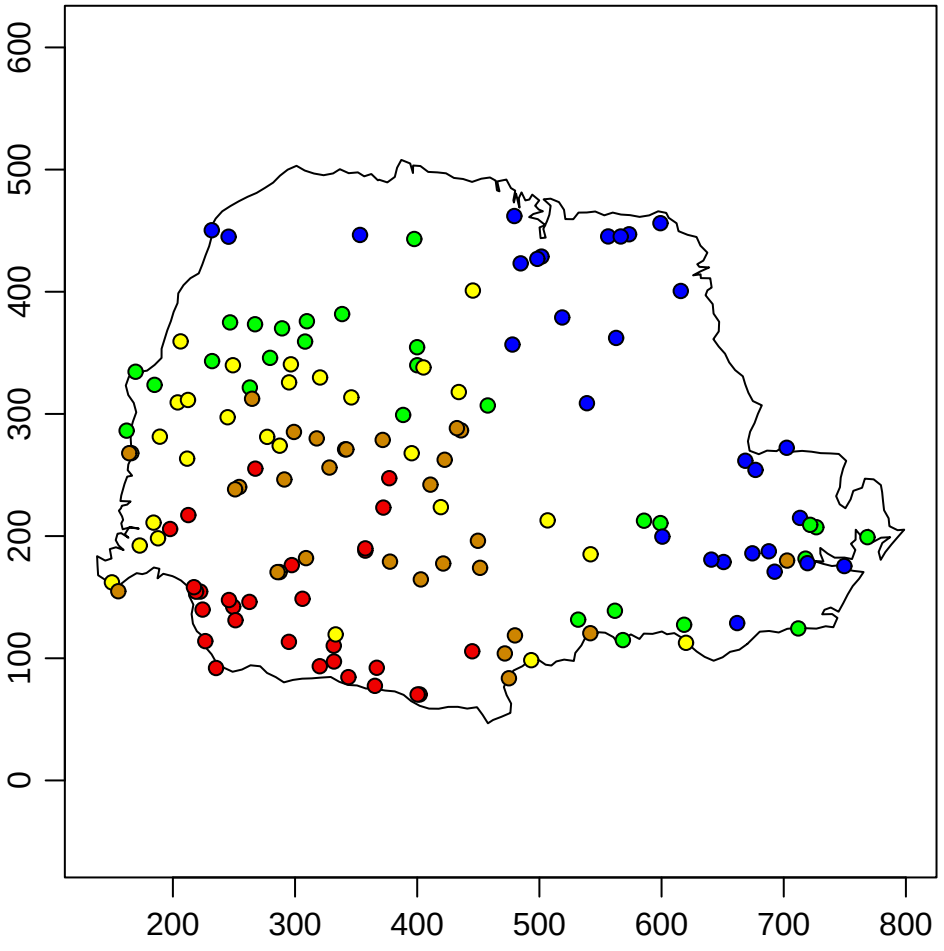


(b) **Precipitação no Estado do Paraná**

Medidas de chuva em 143 postos meteorológicos.

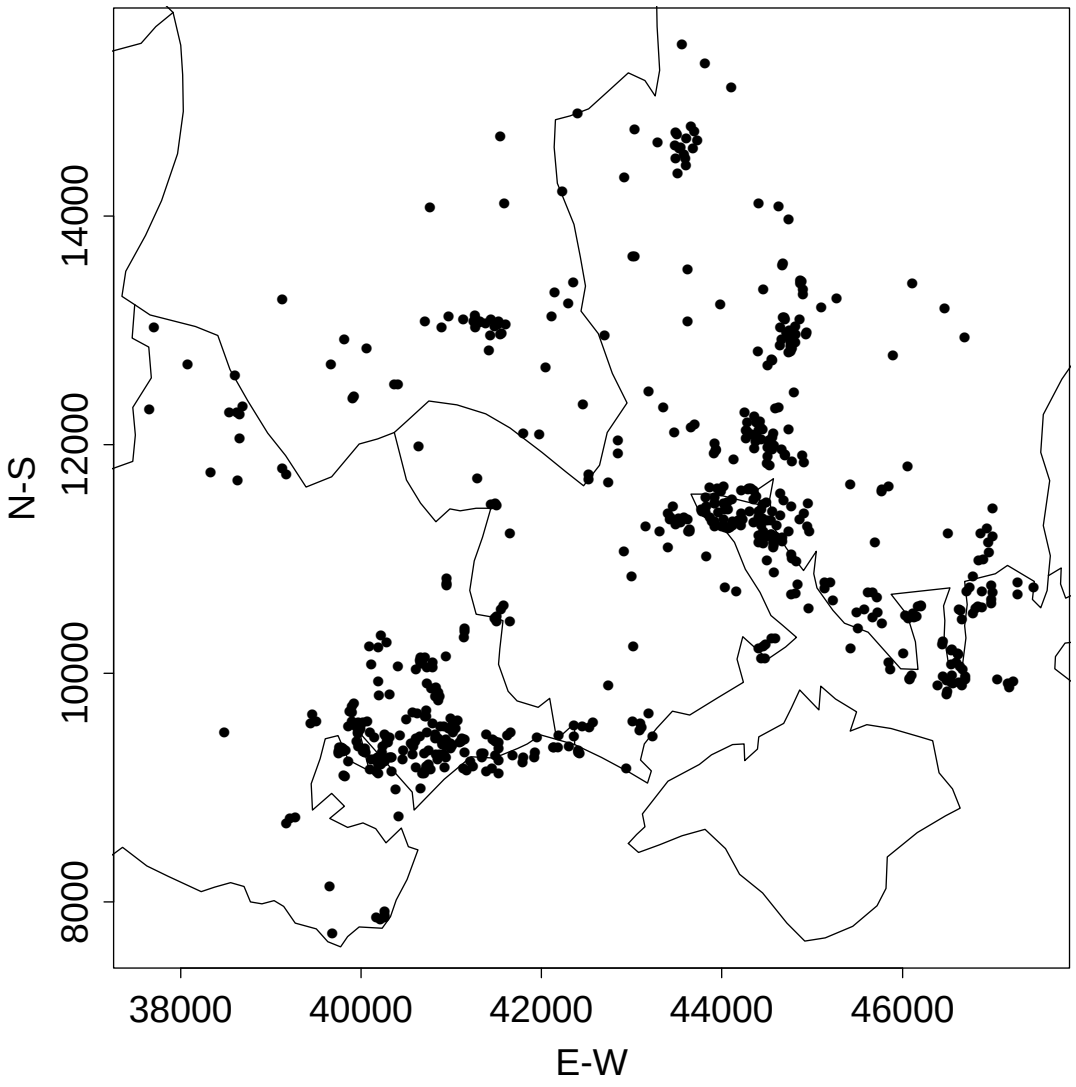
Médias históricas para o período de Maio-Junho (estação seca).

Maiores detalhes: tese de Jacinta L. Zamboti (2001).



(c) **Infecções bacterianas no sul da Inglaterra**

Localizações das residências de 651 casos notificados num per íodo de 1 ano na região central do sul da Inglaterra.



(a) **Variação espacial discreta**
Estrutura básica. $Y_i : i = 1, ..., n$

- raramente ocorre naturalmente
- útil como estratégia pragmática
- modelos são tipicamente definidos indiretamente a partir de condicionais $[Y_i | Y_j, \forall j \neq i]$

(b) **Variação espacial contínua**
Estrutura básica. $Y(x) : x \in \mathbf{R}^2$

- dados $(y_i, x_i) : i = 1, ..., n$, localizações x_i podem ser:
  - não estocástica (ex. grade cobrindo a região em estudo A) ou estocástica, *porém independente do processo* $Y(x)$

(c) **Processo pontual espacial**
Estrutura básica. Conjunto contável de pontos $x_i \in \mathbf{R}^2$, generados estocásticamente.

- às vezes dados são agregados em regiões

Estatística espacial é a seleção de métodos estatísticos nos quais a localização espacial tem papel explícito na análise dos dados.

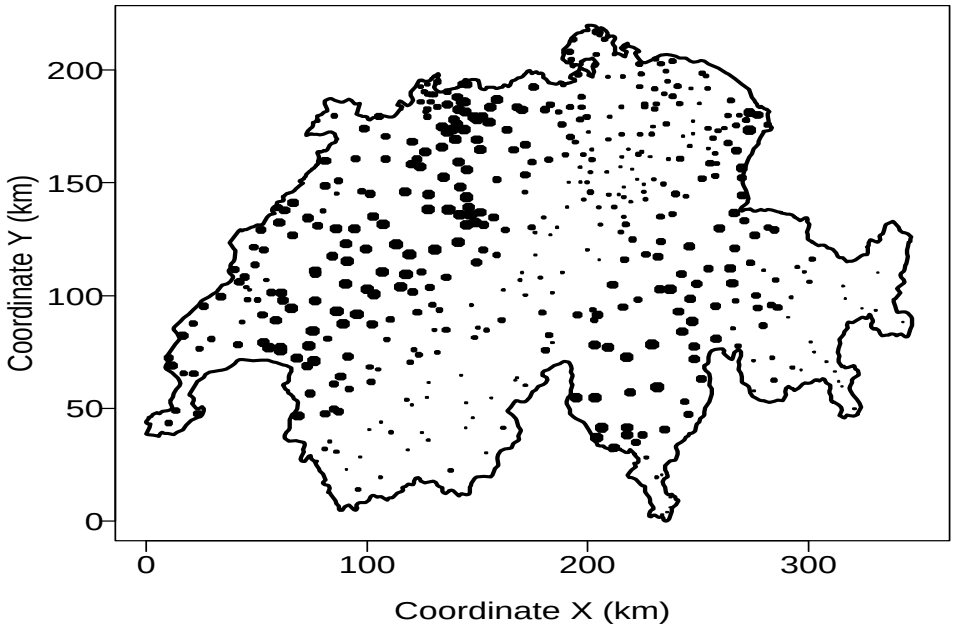
Dois temas estratégicos

- não confundir *formato dos dados* com o *processo subjacente*.

- a escolha do modelo pode ser influenciada pelos objetivos científicos do estudo

2. Outros Exemplos de Problemas Geoestatísticos

(a) Dados de chuva na Suíça



Localizações com tamanhos dos pontos proporcionais aos valores observados de precipitação

- 467 postos na Suíça
- medidas diárias de chuva em 8 de Maio de 1986
- dados do projeto:
  Spatial Interpolation Comparison 97
  `ftp://ftp.geog.uwo.ca/SIC97/`.

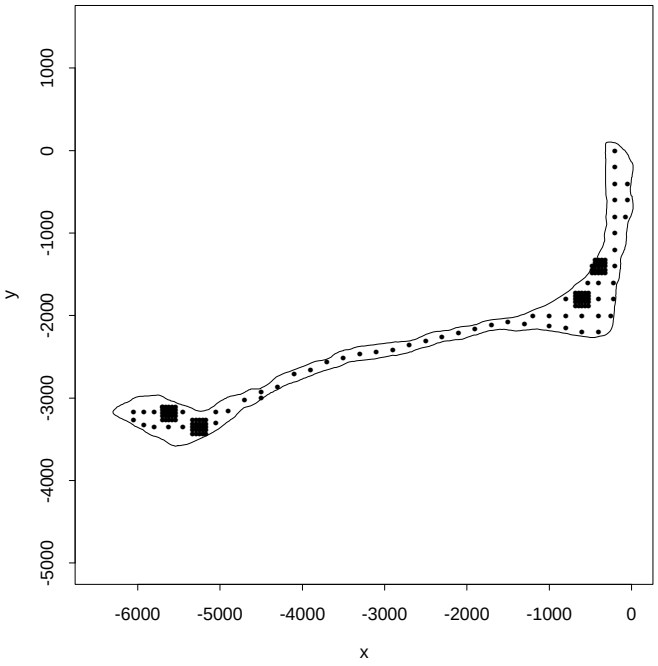
(b) **Ilha de Rongelap**

- estudo do resíduo de contaminação decorrente de testes de armas nucleares durante a década de 50

- ilha evacuada em 1985. Segura para re-ocupação ?

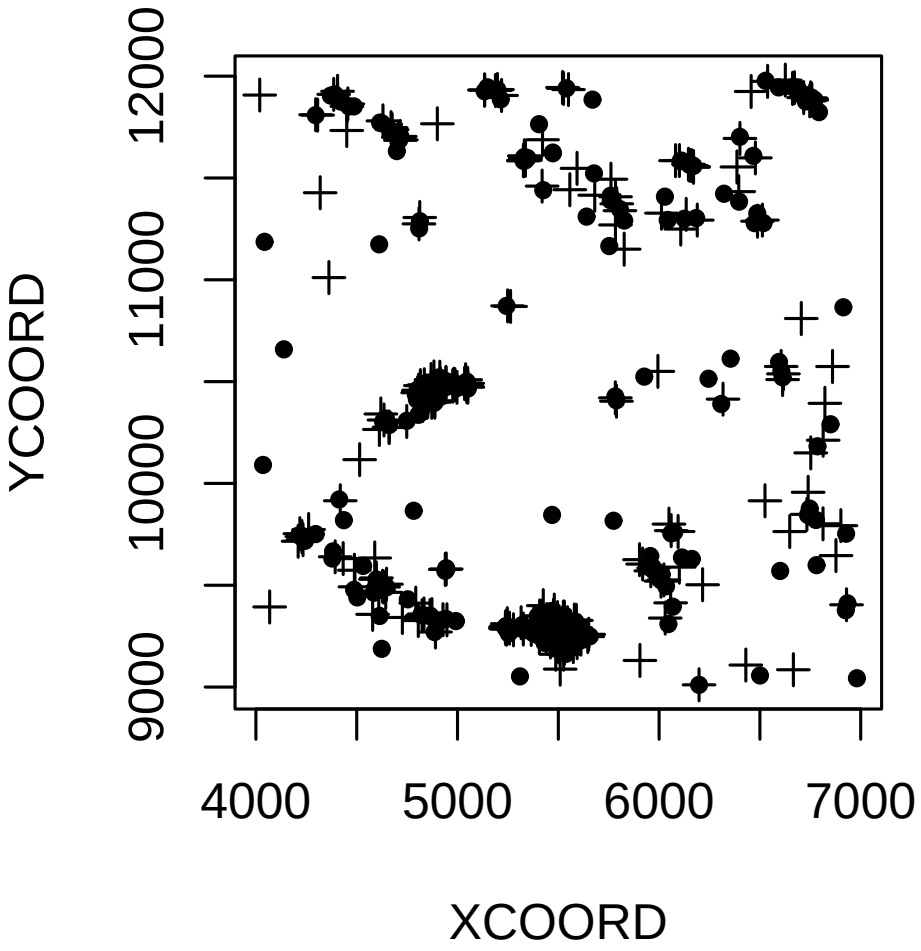
- pesquisa produz medidas com ruído Y_i de concentração de césio radioativo

- particular interesse em níveis máximos de concentração de césio



(c) **Espécies de líquens**

- fatores associados a distribuição espacial da presença de líquens em troncos de ávores
- resposta 0/1: presença ou ausência
- covariáveis: diâmetro, umidade, sombreamento, cobertura do tronco, viva



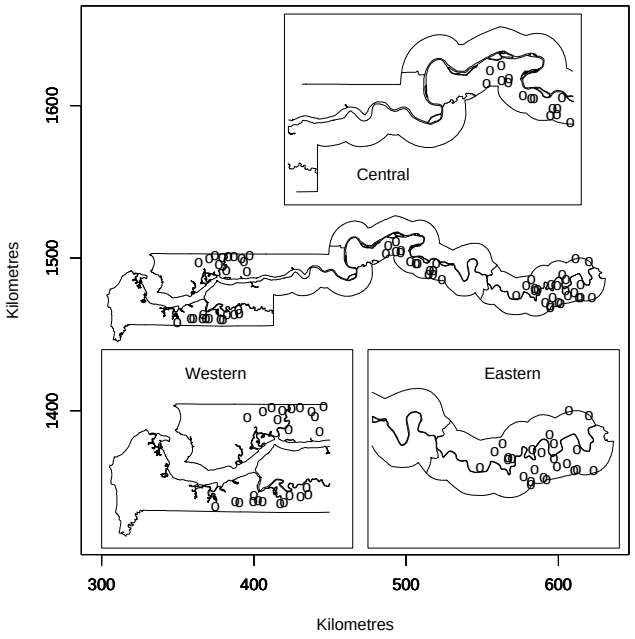
(d) Malária em Gâmbia

- na vila i, dado $Y_{ij} = 0/1$ denota ausência ou presença de malária no sangue da criança j

- covariáveis ao nível de vilas:
  - localização (coordenadas), presença de centro de saúde, índice de vegetação derivado de satélite

- covariáveis ao nível de crianças:
  - idade, uso e tratamento de mosquiteiro

- interesses: efeito das covariáveis e padrão espacial da variação residual



3. Características Principais dos Problemas Geoestatísticos

- dados consistem em **respostas** Y_i associadas com **localizações** x_i

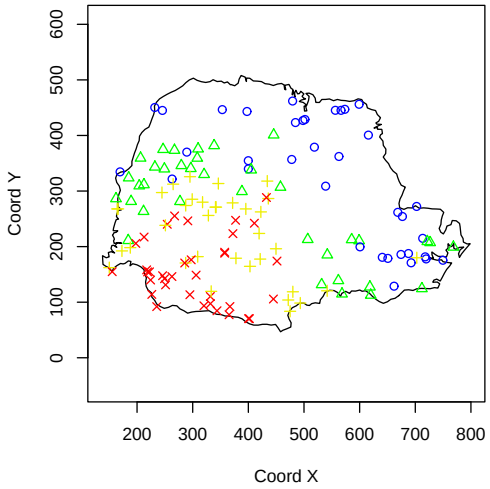
- em princípio, Y pode ser determinado em qualquer localização x dentro da região espacialmente contínua A

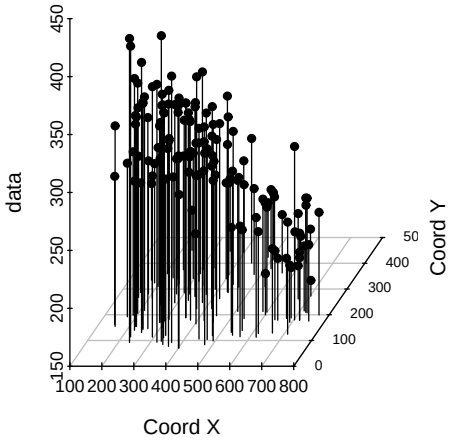
- assume-se que $\{Y(x) : x \in A\}$ é um processo estocástico

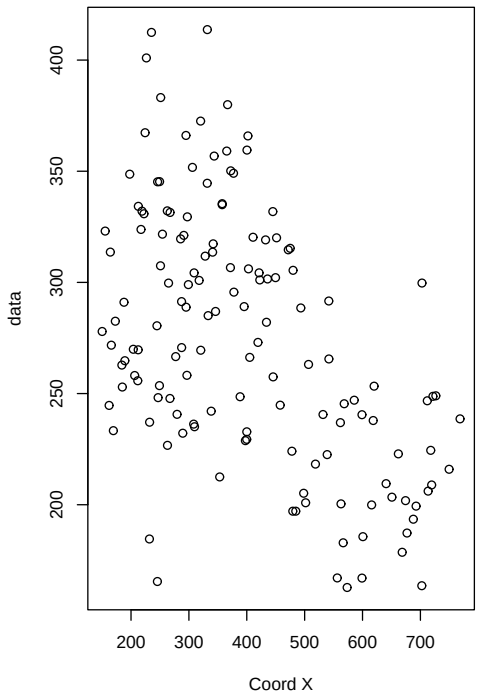
- x_i é tipicamente fixo. Se as localizações x_i são geradas por um processo estocástico pontual, assume-se que este processo é independente de $Y(x)$

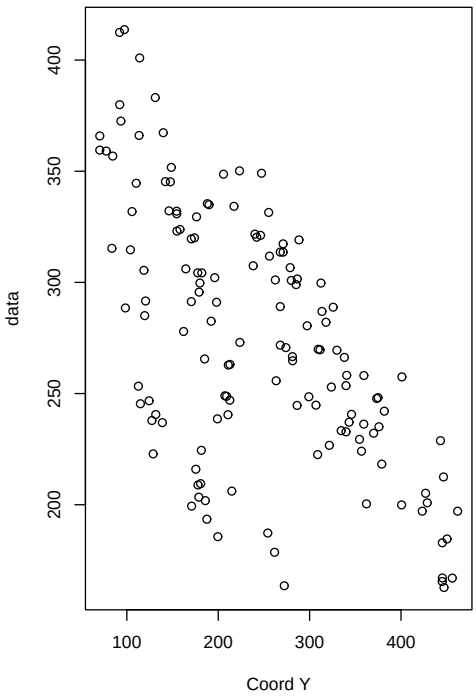
- objetivos científicos incluem a predição de um ou mais funcionais de processo (sem ruído) $\{S(x) : x \in A\}$

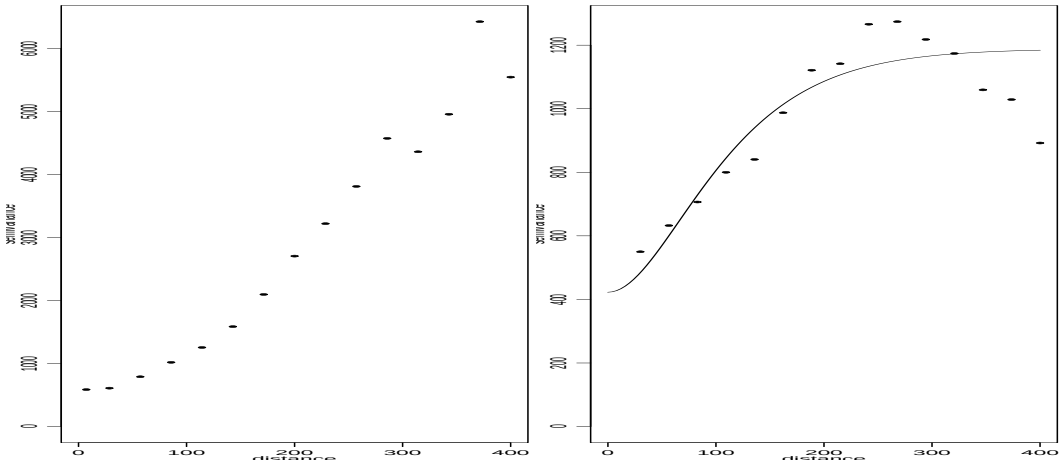
4. Exemplo: chuva no Paraná



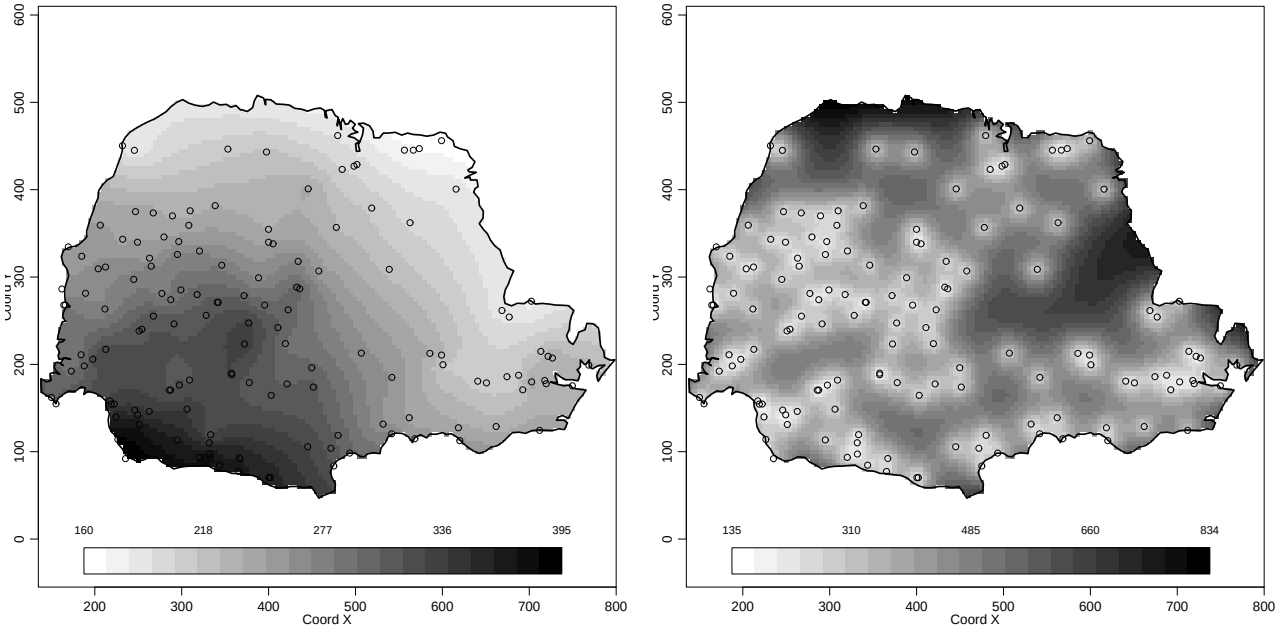








variogramas para dados originais (esquerda) e após retirada de tendência, com modelo ajustado (direita).



Krigagem: mapas de valores preditos (esquerda) e variâncias de predição (direita).

5. História

- Origem: estimação em exploração de reservas minerais (Krige, 1951).

- desenvolvimentos subsequentes independentes do "centro" da comunidade estatística, inicialmente por Matheron e colegas na École des Mines, Fontainebleau, França.

- desenvolvimentos paralelos por Matérn (1946, 1960), Whittle (1954, 1962)

- Ripley (1981): "kriging" em termos de predição de processos estocásticos.

- Significativa cross-fertilização durante 1980's e 1990's (ex o *variograma* é agora uma ferramenta estatística padrao para análise de dados correlacionados no tempo ou espaço.

- Vigorosos debates sobre aspectos práticos ainda persistem:

  – predição *vs* inferência
  – o papel de modelos probabilísticos explícitos

6. Questões Centrais

- **Delineamento**
  - quantas localizações ?
  - quantas medidas?
  - configuração das localizações ?
  - o que deve-se medir em cada localização ?
- **Modelagem**
  - modelo probabilístico para o sinal $[S]$
  - modelo de probabilidade condicional para as medidas, $[Y|S]$
- **Estimação**
  - valores para parâmetros desconhecidos do modelo
  - inferências sobre os parâmetros ou funções destes
- **Predição**
  - avalia-se $[T|Y]$, a distribuição condicional aos dados do objetivo de predição

Geostatística Tradicional:

- evita referência explícita à especificação paramétrica dos modelos
- variogramas como instrumento de inferência (Matheron: "estimação e escolha")
- em geral usa-se estruturas complexas de variogramas
- concentra-se em estimadores lineares
- métodos e paradigmas específicos para:
  - predição pontual (SK, OK, KTE, UK)
  - predição de funcionais não lineares (IK, DK, ...)
  - estimação de densidades preditivas (IK, DK)
  - simulações das preditivas (SGSIM, SISIM, ...)
- *"kriging menu"*

7. Perspectiva histórica - I paradigmas para inferência

(a) Modelos estatísticos:

- redução de dados
- escolha, estimação e predição

(b) Gauss e Legandre

- estudos de astronomia
- erros normais
- discrepância dados e modelo: min. quadrados
- 1^o e 2^o momentos

(c) Fisher e verossimilhança

- uso e interpretação da verossimilhança
- relação com min. quad.: $-2l = \frac{1}{\sigma^2}(y_i - \mu_i)^2$
- máximo, curvaturas, inferência, etc
- Royall, 1997
- pragmatismo e delineamentos

(d) "Model-based" *vs* "design-based"

8. Perspectiva histórica - II
Modelos Lineares Generalizados

- Modelo linear

$$Y = X\beta + \varepsilon$$

- pode ser escrito como:

$$Y \sim N(\mu, \sigma^2)$$
$$\mu = X\beta$$

- e generalizado de 2 formas

$$Y \sim Q(\mu, ...)$$
$$\eta = g(\mu) = X\beta$$

- não mais requer
  - normalidade
  - variância constante
  - preocupação com escala

- verossimilança em destaque

- deviance: $D(\theta) = l(y, y) - l(y, \theta)$

- extensões - modelagem de superdispersão
- modelos mixtos
- inferência Bayesiana

"Às vezes penso que a real diferença entre modelos mixtos e inferência Bayesiana é que o primeiro usa letras romanas e o segundo letras gregas.
(Peter Diggle, na discussão do artigo de Besage & Higdom, 2000)

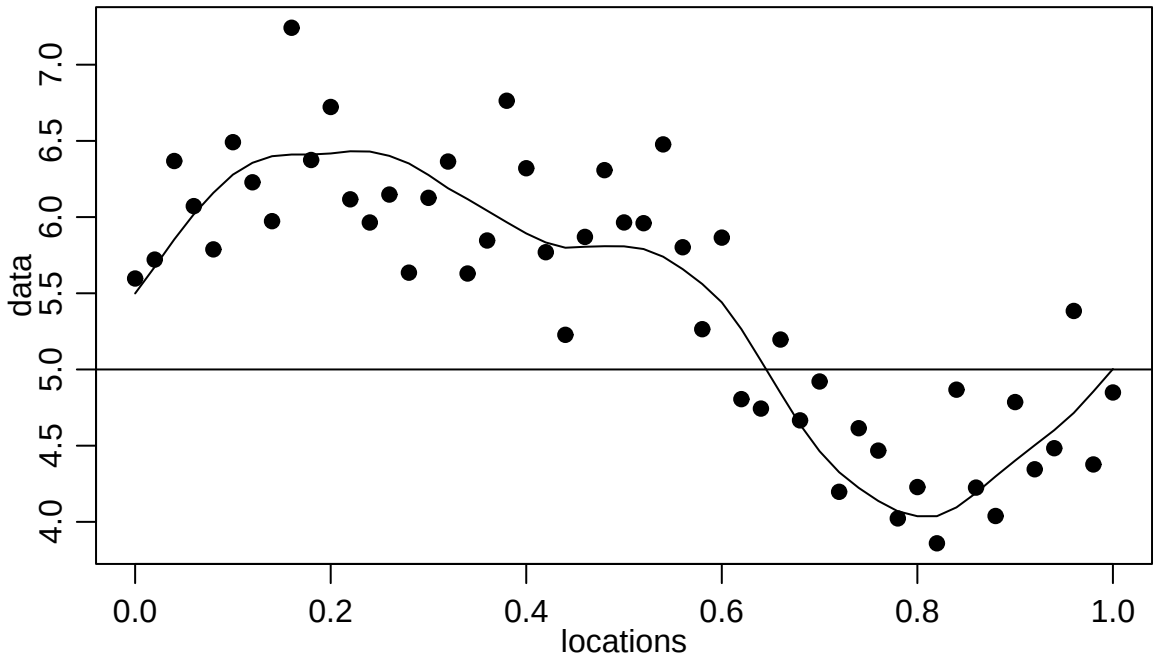
9. "Model-based geostatistics"

- declaração explícita de modelo estocástico
- aplicação de princípios gerais de inferência

Notação

$$(Y_i, x_i) : i = 1, ..., n$$

- $\{x_i : i = 1, ..., n\}$ é o **plano amostral**
- $\{Y(x) : x \in A\}$ é o **processo de medida**
- $\{S(x) : x \in A\}$ é o **processo do sinal**
- $T = \mathcal{F}(S)$ é o **objetivo de predição**
- $[S, Y] = [S][Y|S]$ é o **modelo geoestatístico**



simulação ilustrando os componentes do modelo: dados $Y(x_i)$ (pontos), sinal $S(x)$ (linha curva) e média μ. (linha horizontal).

10. O Modelo Gaussiano

(a) $S(\cdot)$ é um processo Gaussiano estacionário com

i. $\mathrm{E}[S(x)] = \mu$,

ii. $\mathrm{Var}\{S(x)\} = \sigma^2$

iii. $\rho(u) = \mathrm{Corr}\{S(x), S(x-u)\}$;

(b) a distribuição condicional de Y_i dado $S(\cdot)$ é Gaussiana com média $S(x_i)$ e variância τ^2;

(c) $Y_i : i = 1, ..., n$ são mutuamente independentes condicionando em $S(\cdot)$.

Uma formulação equivalente:

$$Y_i = S(x_i) + Z_i : i = 1, ..., n.$$

onde $Z_i : i = 1, ..., n$ são mutuamente independentes e identicamente distribuídos com $Z_i \sim \mathsf{N}(0, \tau^2)$.

Desta forma a distribuição conjunta de Y é multivariada Normal,

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

onde:

$\mathbf{1}$ denota um vetor de 1's com n elementos

I é matrix identidade $n \times n$

R é uma matrix $n \times n$ com $(i, j)^{th}$ elemento $\rho(u_{ij})$ onde
$u_{ij} = ||x_i - x_j||$, é distancia Euclideana entre x_i e x_j.

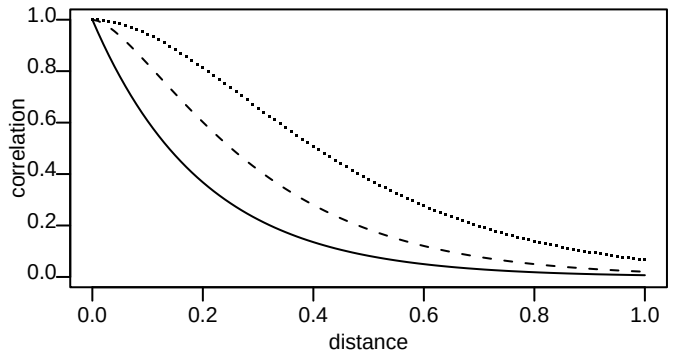
11. Especificação da função de correlação

A família de Matérn

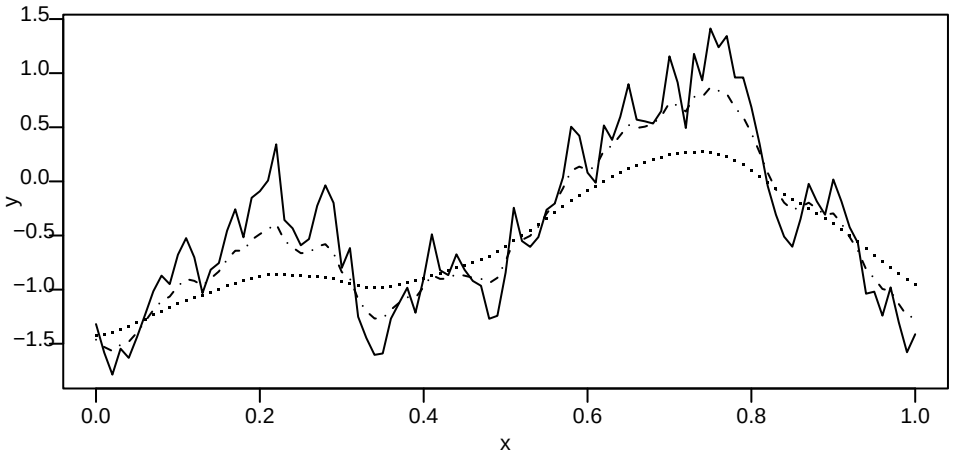
Função de correlação dada por

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

- κ e ϕ são parâmetros
- $K_{\kappa}(\cdot)$ denota função de Bessel de ordem κ
- válida para $\phi > 0$ e $\kappa > 0$.
- $\kappa = 0.5$: *modelo exponencial*
- $\kappa \to \infty$: *modelo Gaussiano*
- $S(x)$ é $\lceil \kappa - 1$ vezes diferenciável



Três exemplos de funções de Matérn com $\phi = 0.2$ and $\kappa = 1$ (linha sólida), $\kappa = 1.5$ (linha interrompida) and $\kappa = 2$ (pontos).



simulações de processos em 1-D com funções de correlação de de Matérn com $\phi = 0.2$ e $\kappa = 0.5$ (linha sólida), $\kappa = 1$ (linha interrompida) and $\kappa = 2$ (linha pontilhada).

12. Propriedades do segundo momento

- o **variograma** de um processo $Y(x)$ é a função

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

- para o modelo linear Gaussiano, com $u = ||x - x'||$,

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

- os paramêtros estruturais básicos são
  - *efeito pepita ("nugget")*: τ^2
  - *patamar ("sill")*: $\tau^2 + \sigma^2 = \text{Var}\{Y(x)\}$
  - *o alcance ("range")*: ϕ, tal que $\rho(u) = \rho_0(u/\phi)$

- implicações práticas:
  - qualquer versão razoável do modelo linear Gaussiano tem pelo menos três parâmetros de covariância
  - um volume da dados substancial pode ser necessário para estimar maior número de parâmetros
  - a família **Matérn** possui um parâmetro extra para determinar a suavidade do processo $S(x)$

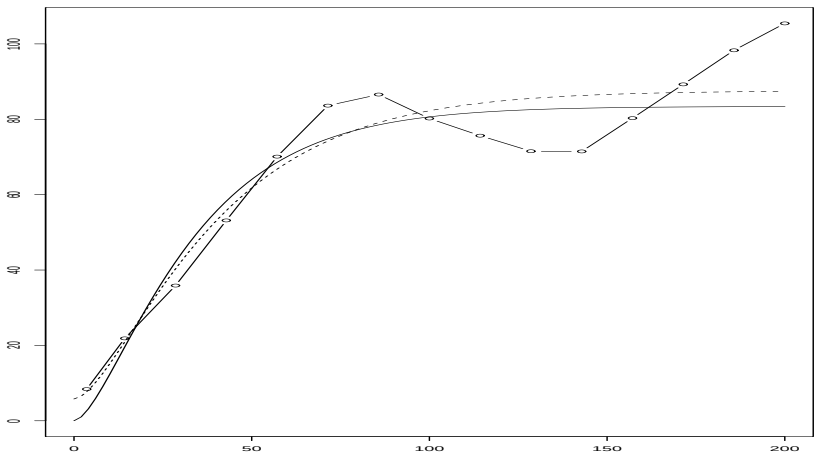
13. Estimação de parâmetros

- **métodos "ad-hoc"** :
  estima-se $\tilde{\theta}$ que minimise

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

  onde $\bar{V}_k$ é a média de n_k ordenadas v_{ij} do variograma

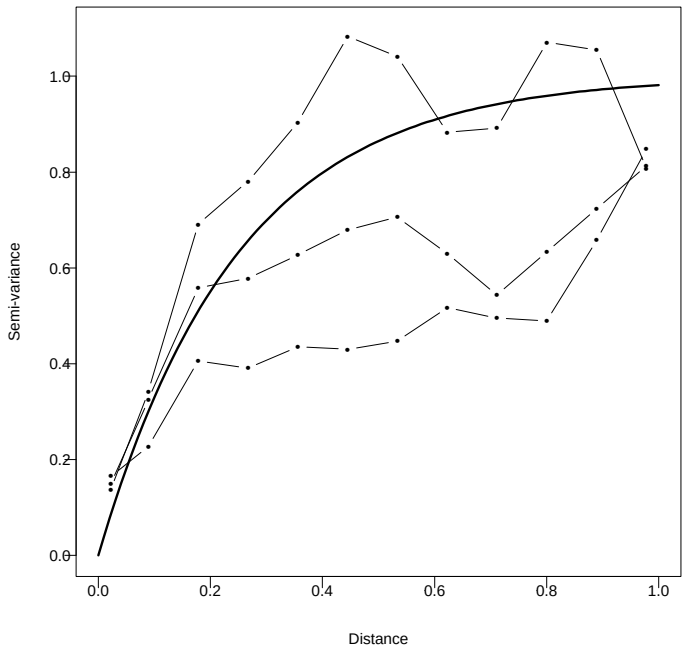
  - corresponde a um sistema de equações de estimação que produz estimativas viciadas de θ,
  - mesmo assim é largamente utilizado na prática
  - potencialmente "perigoso" devido as correlações inerentes aos sucessivos $\bar{V}_k$'s



variograma empírico com estimativas OLS (linha pontilhada) e WLS (linha cheia)

Por que o variograma empírico pode ser inadequado

- sob o modelo linear Gaussiano:
  - $v_{ij} \sim V(u_{ij})\chi^2_1$
  - the v_{ij} são correlacionados
- pode ser instável
- variogramas de dados originais e resíduos podem ser substancialmente diferentes
- estimação baseada em objeto estimado
- seria possivel ajustar o modelo aos **dados** e não ao variograma?



variograma teórico e três simulações

- **Métodos baseados em verossimilhança**:
  tipicamente usados sob pressupostos de normalidade

  – estimativas ótimas sob os pressupostos declarados

  – porém computacionalmente caros e podem nao ser robustos

  – Implementação Bayesiana combinando estimação e predição tem sido cada vez mais aceita (ao menos entre estatísticos!).

14. Estimação via verossimilhança

O modelo Gaussiano é dado por:

$$Y_i|S \sim \mathsf{N}(S(x_i), \tau^2)$$

- $S(x_i) = \mu(x_i) + S_c(x_i)$
- $S_c(\cdot)$ é um processo estocástico Gaussiano com parâmetros de covariância (σ^2, ϕ, κ),
- $\mu(x_i) = F\beta = \sum_{j=1}^{k} f_k(x_i)\beta_k$, onde $f_k(x_i)$ é vetor de covariáveis na localização x_i

Podemos escrever:

$$Y_i = \mu(x_i) + S_c(x_i) + Z_i : i = 1, ..., n$$

e então

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

e a função de verossimilhaça é:

$$L(\beta, \tau, \sigma, \phi, \kappa) \propto -0.5\{\log|(\sigma^2 R + \tau^2 I)| + (y - F\beta)'(\sigma^2 R + \tau^2 I)^{-1}(y - F\beta)\}.$$

para qual maximização (numérica) produz as estimativas de máxima verossimilhança

Detalhes computacionais

- reparametrize $\nu^2 = \frac{\tau^2}{\sigma^2}$ e denote: $\sigma^2 V = \sigma^2(R + \nu^2 I)$

- a log-verossimilhança é maximizada para

$$\widehat{\beta}(V) = (F'V^{-1}F)^{-1}F'V^{-1}y$$

$$\widehat{\sigma}^2 = n^{-1}(y - F\widehat{\beta})'V^{-1}(y - F\widehat{\beta})$$

- substituindo (β, σ^2) por $(\widehat{\beta}, \widehat{\sigma^2})$ em 1 a maximização se reduz à:

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

- Para família de Matérn considere tomar κ em um conjunto discreto $\{0.5, 1, 2, 3, ..., N\}$

15. Predição "plug-in"

Em geral o interesse está em predizer

- o valor da realização do processo $S(\cdot)$ em um ponto
- ou a média de $S(\cdot)$ em uma região

$$T = |B|^{-1} \int_B S(x)dx$$

onde $|B|$ denota a área da região B.

Para o modelo Gaussiano o preditor de minímos quadrados de $T = S(x)$ é:

$$\hat{T} = \mu + \sigma^2 \mathbf{r}'(\tau^2 I + \sigma^2 R)^{-1}(Y - \mu \mathbf{1})$$

e a variância de predição

$$\text{Var}(T|Y) = \sigma^2 - \sigma^2 \mathbf{r}'(\tau^2 I + \sigma^2 R)^{-1}\sigma^2 \mathbf{r}$$

onde os únicos termos desconhecidos são os parâmetros do modelo

A **predição "plug-in"** consiste em substituir os parâmetros por suas estimativas.

Comentários

- ML e krigagem simples
- REML e krigagem ordinária
- Predição *ad-hoc*:

(a) estima β por OLS, $\tilde{\beta} = (F'F)^{-1}F'Y$, e calcula-se resíduos $Z = Y - F\tilde{\beta}$.

(b) calcula-se o variograma empírico (dos resíduos) que é utilizado para formulação do modelo e estimação de parâmetros

(c) reestima-se β por GLS e usa-se modelo ajustado para predição

- papel dos variogramas empíricos
  - diagnóstico (abordagem "model-based")
  - ferramenta de inferência (abordagem tradicional)
- ambas abordagens anexam estimativas dos parametros ao modelo como se fosse valores verdadeiros.

Predição "plug-in"

- usualmente produz boas estimativas pontuais de $T = S(x)$
- em geral sub-estima variância de predição
- pode produzir estimativas inacuradas de outras quantidades objetivo T

16. Modelos Gaussianos transformados

O modelo Gaussiano é claramente inapropriado para distribuições assimétricas

Parâmetro extra λ da transformação Box-Cox introduz certa flexibilidade.

O modelo fica definido da forma:

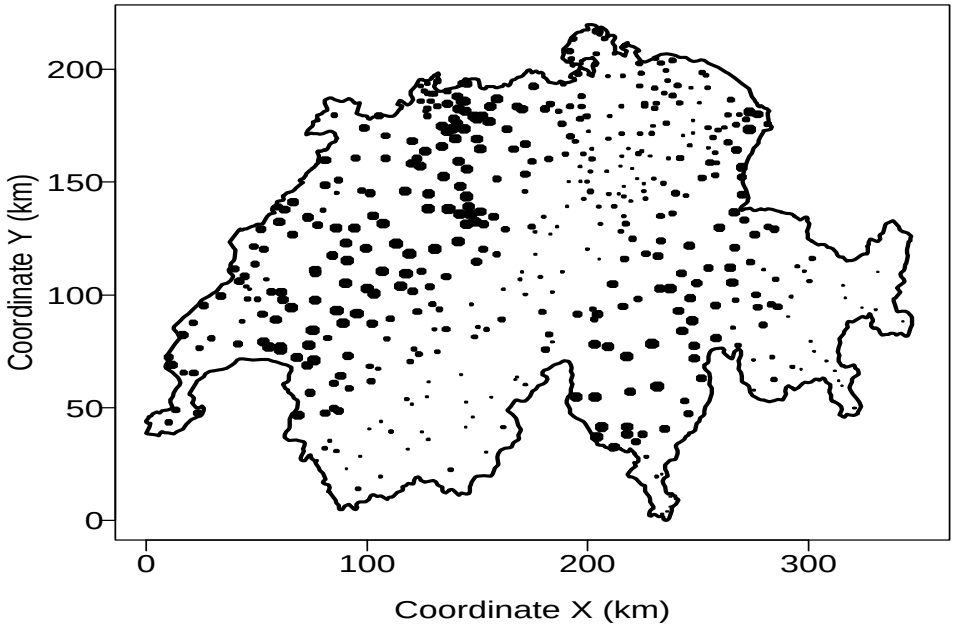
- assume-se $Y^* \sim MVN(F\beta, \sigma^2 V)$
- dados $y = (y_1, ..., y_n)$, são gerados por uma transformação do modelo linear Gaussiano $Y = h_\lambda^{-1}(Y^*)$ tal que:

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

A log-verossimilhança é:

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

17. Estudo de caso: chuva na Suíça



Localizações com tamanho dos pontos proporcional aos valores observados. Distâncias em kilometros

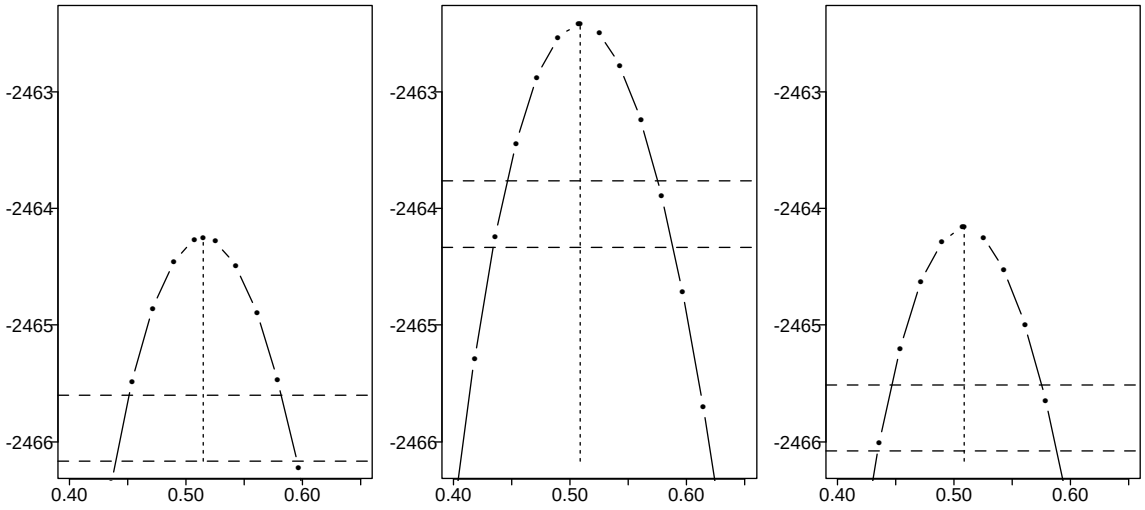
- 467 localizações
- medidas de precipitação em 8 de Maio 1986
- dados são valores inteiros com unidade de medida igual á 1/10 mm
- 5 localizações com valores iguais à zero.

chuva na Suíça (cont.)

Estimação : parâmetros de transformação e suavidade (modelo de Matérn)

κ	$\widehat{\lambda}$	$\log \widehat{L}$
0.5	0.514	-2464.246
1	0.508	-2462.413
2	0.508	-2464.160

Estimativas de MV de $\widehat{\lambda}$ e valores da log-verossimilhança $\log \widehat{L}$ para diferentes valores de κ.



Verossimilhanças prefilhadas para λ. esquerda: $\kappa = 0.5$, meio: $\kappa = 1$, direita: $\kappa = 2$.

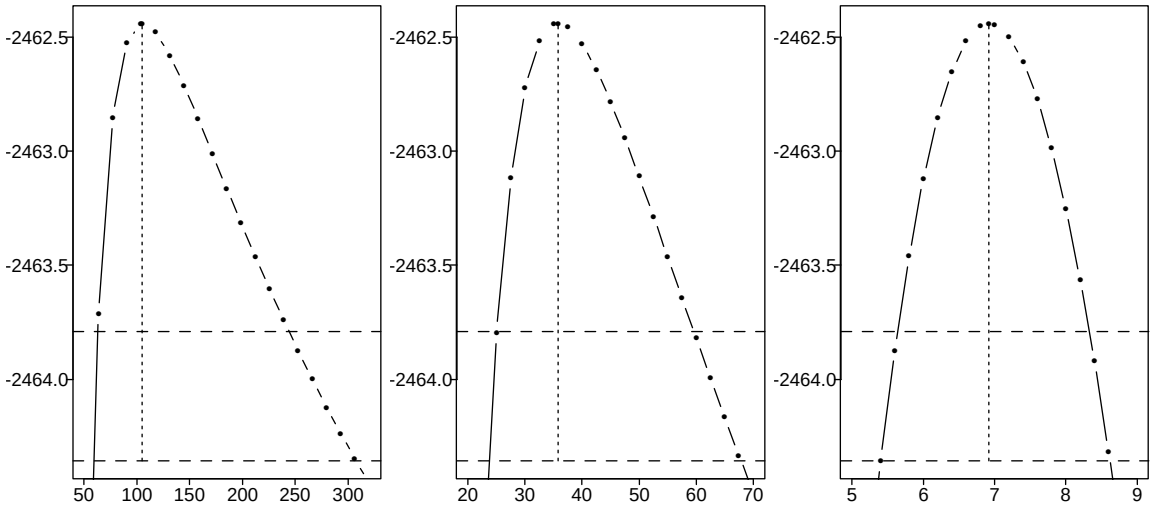
transformação logarítimica ou não - transformação não são indicadas

chuva na Suíça (cont.)

Estimativas para modelo com $\lambda = 0.5$

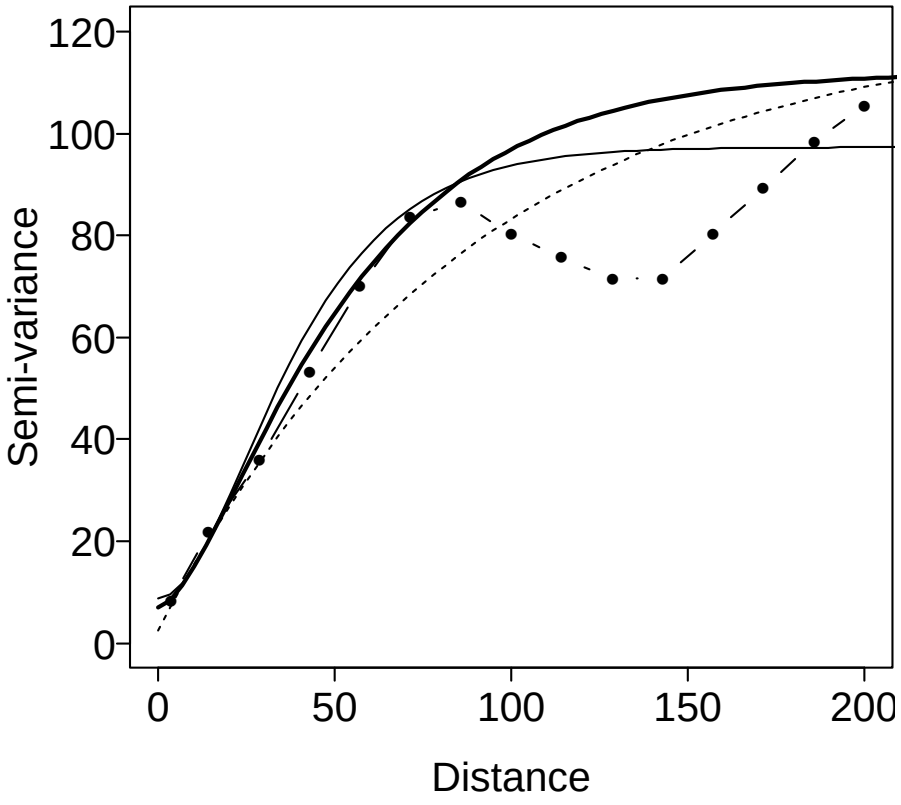
κ	$\widehat{\beta}$	$\widehat{\sigma}^2$	$\widehat{\phi}$	$\widehat{\tau}^2$	$\log \widehat{L}$
0.5	18.36	118.82	87.97	2.48	-2464.315
1	20.13	105.06	35.79	6.92	-2462.438
2	21.36	88.58	17.73	8.72	-2464.185

Maximum likelihood estimates $\widehat{\beta}$, $\widehat{\phi}$, $\widehat{\sigma}$, $\widehat{\tau}$ and the corresponding value of the likelihood function $\log \widehat{L}$ for different values of the Matérn parameter κ, for $\lambda = 0.5$



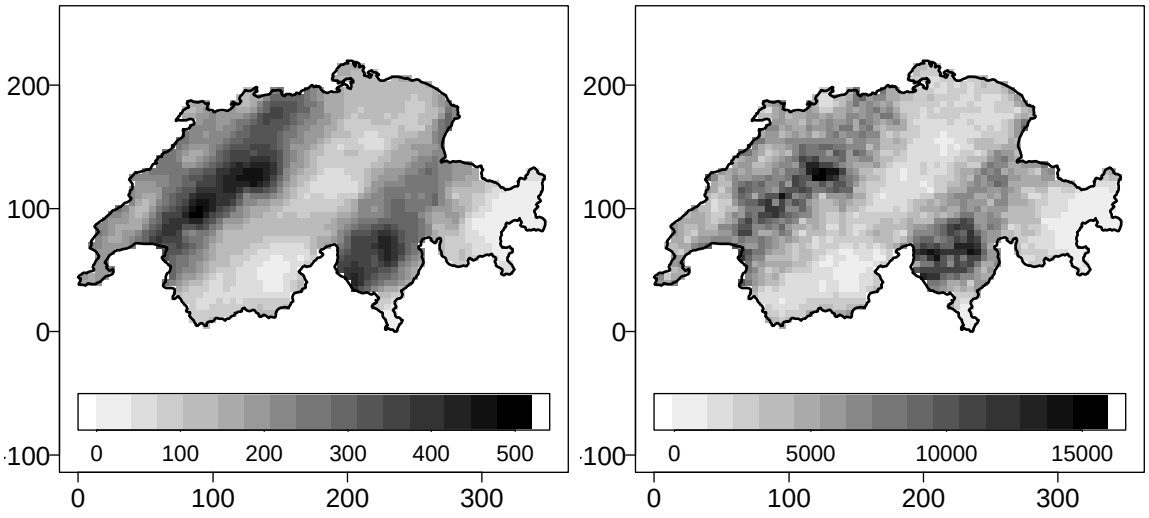
Verossimilhança perfilhada para parâmetros de covariância $\kappa = 1$ and $\lambda = 0.5$. esquerda: σ^2, meio: ϕ, direita: τ^2.

chuva na Suíça (cont.)



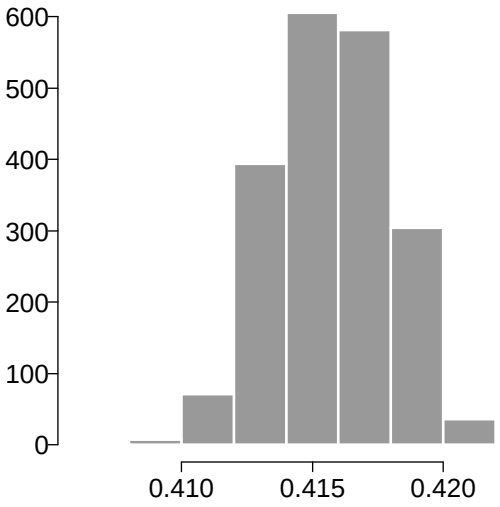
semivariograma empírico para dados transformados e variogramas teóricos com estimativas de MV para $\kappa = 0.5$ (linha interrompida), $\kappa = 1$ (linha grossa), $\kappa = 2$ (linha fina).

chuva na Suíça (cont.)



Mapas com predições (esquerda) e variâncias de predição (direita).

Predição da percentagem da área onde $Y(x) \geq 200$: $\tilde{A}_{200}$ é de 0.4157



Amostras da preditiva de $\tilde{A}_{200}$.

Notas:

Modelo log-Gaussiano é caso particular com $\lambda = 0$.

Inference strategies:

(a) λ como parâmetro aleatório (De Oliveira, Kadem and Short, 1997). Predições tiram médias de vários modelos.

(b) Abordagem alternativa: estima λ e então fixa ao valor estimado (Christensen, Diggle e Ribeiro, 2000):

   i. encontre a *melhor* transformação maximisando a verosimilhança perfilhada de λ

   ii. fixe a transformação , transforme dados

   iii. inferências na escala transformada

   iv. transformação reversa dos resultados

- Estimação por máxima verossimilhança restrita (REML)
- Verossimilhanças perfilhadas
- Modelos anisotrópicos
- Modelos não estacionários
  - Relações funcionais entre médias e variâncias
  - média não constante

    $$\mu(x) = F\beta = \sum_{j=1}^{k} \beta_j f_j(x)$$

    para covariáveis $f_j(x)$.

    Nota: à **krigagem universal** ou **krigagem com tendência externa**
  - variação aleatória não estacionária

    Intrínseca Campos aleatórios de Markov (Besag, York and Molié, 1991).

    Deformações espaciais Sampson and Guttorp, 1992 tentam obter estacionaridade através de transformações não lineares do espaço geográfico x. Ver tese de Alexandra Smith (2001).
  - flexibilidade *vs* identificabilidade

izados

O modelo linear generalizado clássico:

- $Y_i : i = 1, ..., n$ mutuamente independentes, com $\mu_i = \mathsf{E}[Y_i]$
- $h(\mu_i) = \sum_{j=1}^{k} f_{ij}\beta_j$, para uma função de ligação conhecida $h(\cdot)$

O modelo linear generalizado mixto:

- $Y_i : i = 1, ..., n$ mutuamente independentes, com $\mu_i = \mathsf{E}[Y_i]$, condicionados à realização do conjunto de variáveis aleatórias latentes U_i
- $h(\mu_i) = U_i + \sum_{j=1}^{k} f_{ij}\beta_j$, para função de ligação conhecida $h(\cdot)$

O modelo linear generalizado geoestatístico:

- $Y_i : i = 1, ..., n$ mutuamente independentes, com $\mu_i = \mathsf{E}[Y_i]$, condicionados à realização do conjunto de variáveis aleatórias latentes U_i
- $h(\mu_i) = U_i + \sum_{j=1}^{p} z_{ij}\beta_j$, para uma função de ligação conhecida $h(\cdot)$
- $U_i = S(x_i)$ onde $\{S(x) : x \in \mathbf{R}^2\}$ é um processo estocástico espacial

Inferência para o modelo geoestatístico linear generalizado

- avaliação da verossimilhança envolve integração numérica de multidimensional
- métodos aproximados (ex Breslow and Clayton, 1993) tem acurácia duvidosa
- MCMC é possível embora não rotineira

Esquemas para MCMC

- Ingredientes

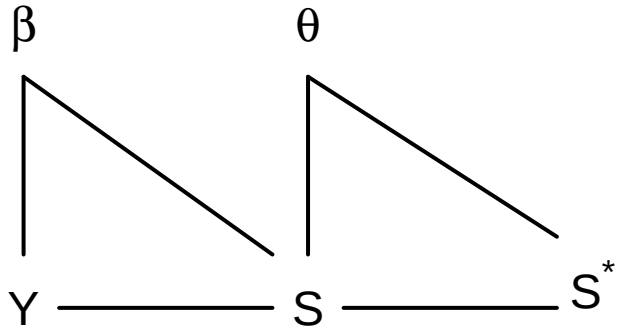
  · Prioris para os parâmetros de regressão β e de covariância θ

  · Dados: $Y = (Y_1, ..., Y_n)$

  · $S = (S(x_1), ..., S(x_n))$

  · $S^* =$ todos outros $S(x)$

- Estrutura de independência condicional



- use resultados das cadeias para contruir declarações à posteriori sobre $[T|Y]$, onde $T = \mathcal{F}(S^*)$

20. Estudo de caso: Ilha Rongelap

- **Ilha Rongelap**
  - aproximadamente a 2500 milhas sudoeste do Hawaii
  - contaminada por testes de armas nucleares em 1950's
  - evacuada em 1985
  - segura para re-assentamento?
- **Problemas estatísticos**
  - delineamento e medidas de campo de ^{137}Cs
  - estimar variação espacial da radiotividade de ^{137}Cs
  - comparação com padrões de segurança

O modelo Poisson

- Medidas básicas são contagens Y_i em intervalos de tempo t_i nas localizações x_i $(i = 1, ..., n)$

- estrutura dos dados sugere o modelo:

  · $S(x) : x \in R^2$ processo estacionário Gaussiano (radioatividade local)

  · $Y_i|\{S(\cdot)\} \sim$ Poisson(μ_i)

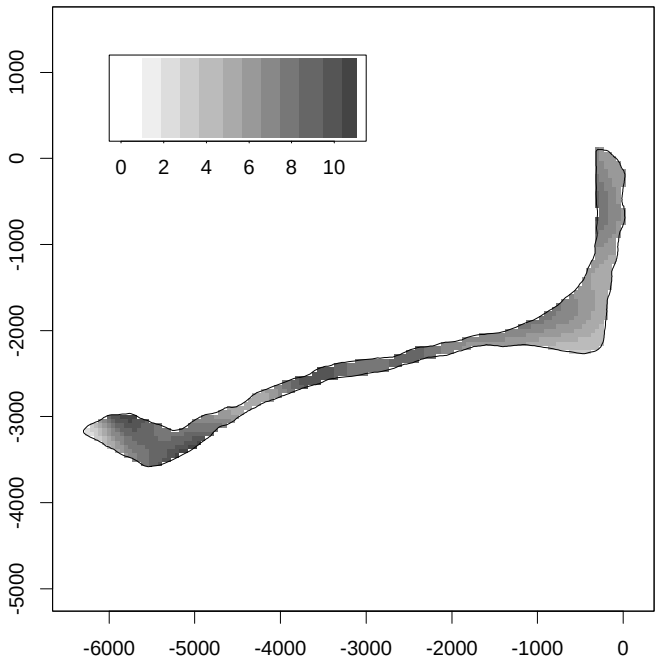
  · $\mu_i = t_i\lambda(x_i) = t_i \exp\{S(x_i)\}$.

- Objetivos:

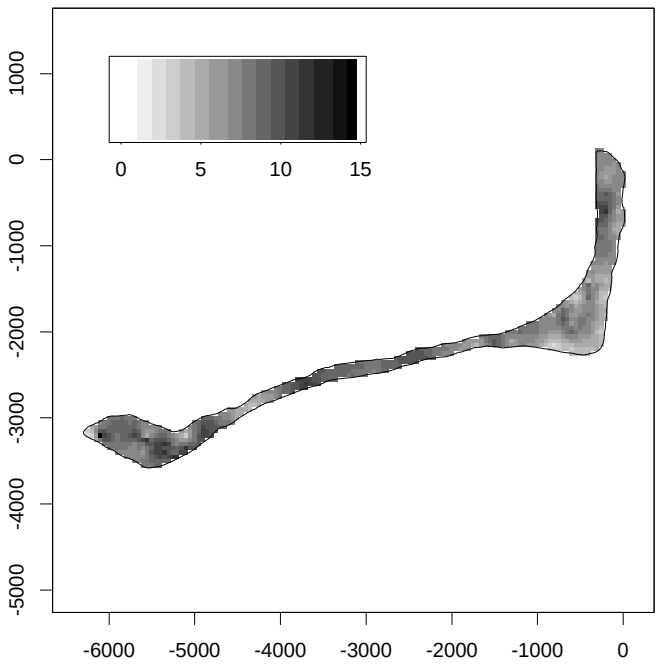
  · predizer $\lambda(x)$ sobre toda ilha

  · max $\lambda(x)$

  · arg(max $\lambda(x)$)

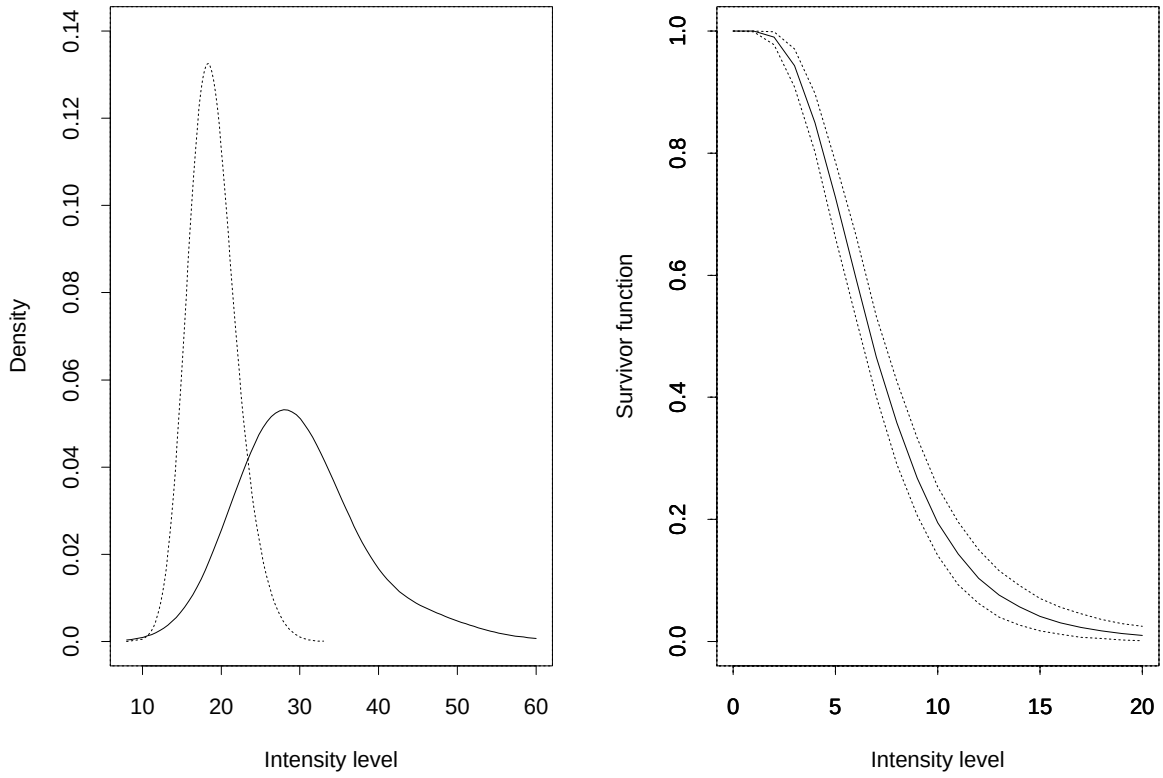


superfície de radiotividade predita utilizando krigagem logarítmica



superfície de radiotividade predita utilizando o modelo log-linear Poisson com processo latentre Gaussiano

Predição Bayesiana de funcionais não lineares da superfície de radiação



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.

21. Estudo de caso: Malária em Gambia

- Neste exemplo a variação espacial é de interesse científico secundário.

- O objetivo primário e descrever a dependência entre a prevalência de parasitas de malária e as covariáveis medidas

  – em vilas

  – em indivíduos

- Particular interesse em saber se o índice de vegetação derivado de medidas de satélite pode ser utilizado como preditor da prevalência de malária.

  Isto ajudaria profissionais de saúde a alocar melhor os recursos que são escarsos.

Estrutura dos dados

- 2039 crianças em 65 vilas
- cada uma testada para presença de parasitas de malária no sangue

Covariáveis das crianças

- idade (dias)
- sexo (F/M)
- uso de mosquiteiro (nenhum, não tratado e tratado)

Covariáveis das vilas:

- localização
- indice de vegetação (satélite)
- presença de centro de saúde na vila

Modelo de regressão logística

- $Y_{ij} = 0/1$ presença ou ausência de parasitas de malária na jth criança da ith vila

- $f_{ij} =$ covariável da criança

- $w_i =$ covariável da vila

- $\text{logit}(P(Y_{ij} = 1|S(\cdot))) = f'_{ij}\beta_1 + w_i{}'\beta_2 + S(x_i)$

É razoável assumir infeções condicionalmente independentes na mesma vila?

Caso não , o modelo deve ser extendido para permitir variabilidade extra-binomial não -espacial

- $U_i \sim \text{N}(0, \nu^2)$

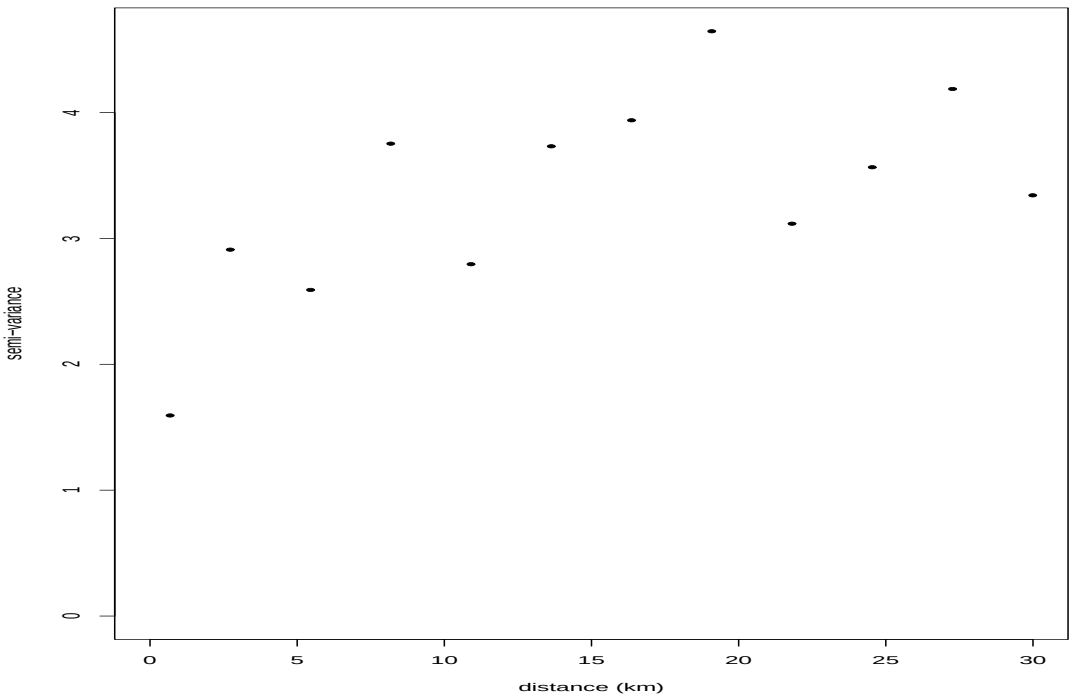
- $\text{logit}P(Y_{ij} = 1|S(\cdot), U) = f'_{ij}\beta_1 + w_i{}'\beta_2 + U_i + S(x_i)$

Análise exploratória

- ajuste modelo logístico padrão sem $S(x)$ e/ou U
- calcule para cada vila:

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

- resíduos de vila, $r_i = (N_i - \mu_i)/\sigma_i$
- derivar dados r_i
- ajuste de parâmetros de covariância



Variograma do resíduos de vilas

Análise "model-based"

$\alpha =$ intercepto

$\beta_1 =$ coeficiente para idade

$\beta_2 =$ coeficiente uso de mosquiteiro

$\beta_3 =$ coeficiente para mosquiteiro tratado

$\beta_3 =$ coeficiente para indice de verde

$\beta_4 =$ coeficiente para presença de centro de saúde

$\nu^2 =$ variância do efeito aleatório não espacial U_i

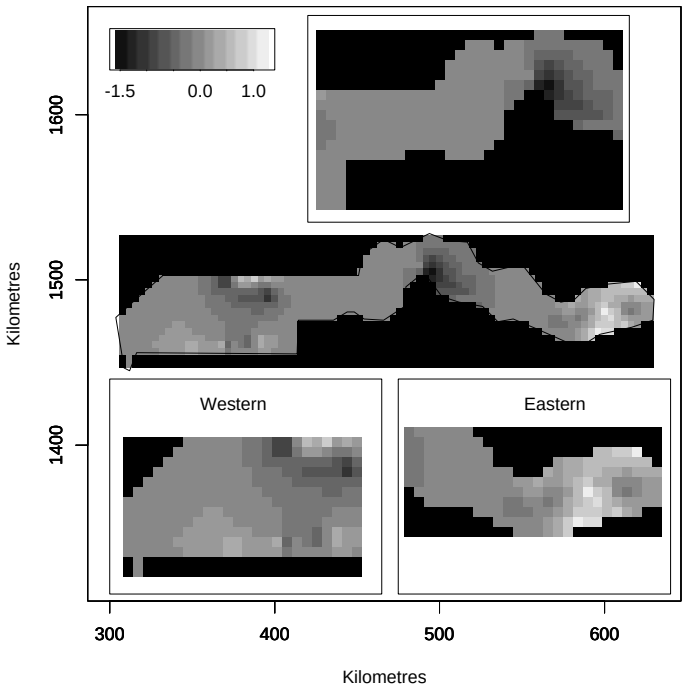
$\sigma^2 =$ variância do preocesso espacial $S(x)$

$\phi =$ parâmetro de decaimento da correlação

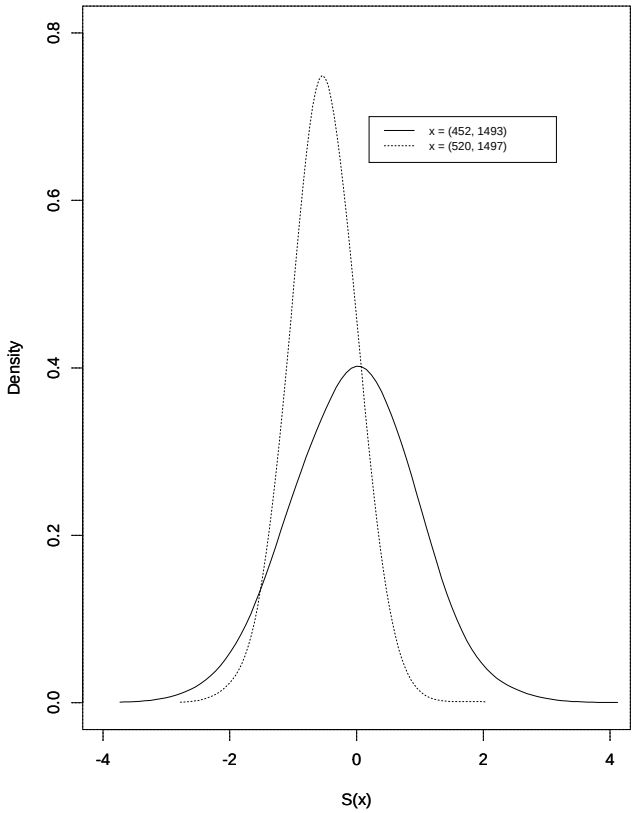
$\kappa =$ parâmetro de suavidade

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

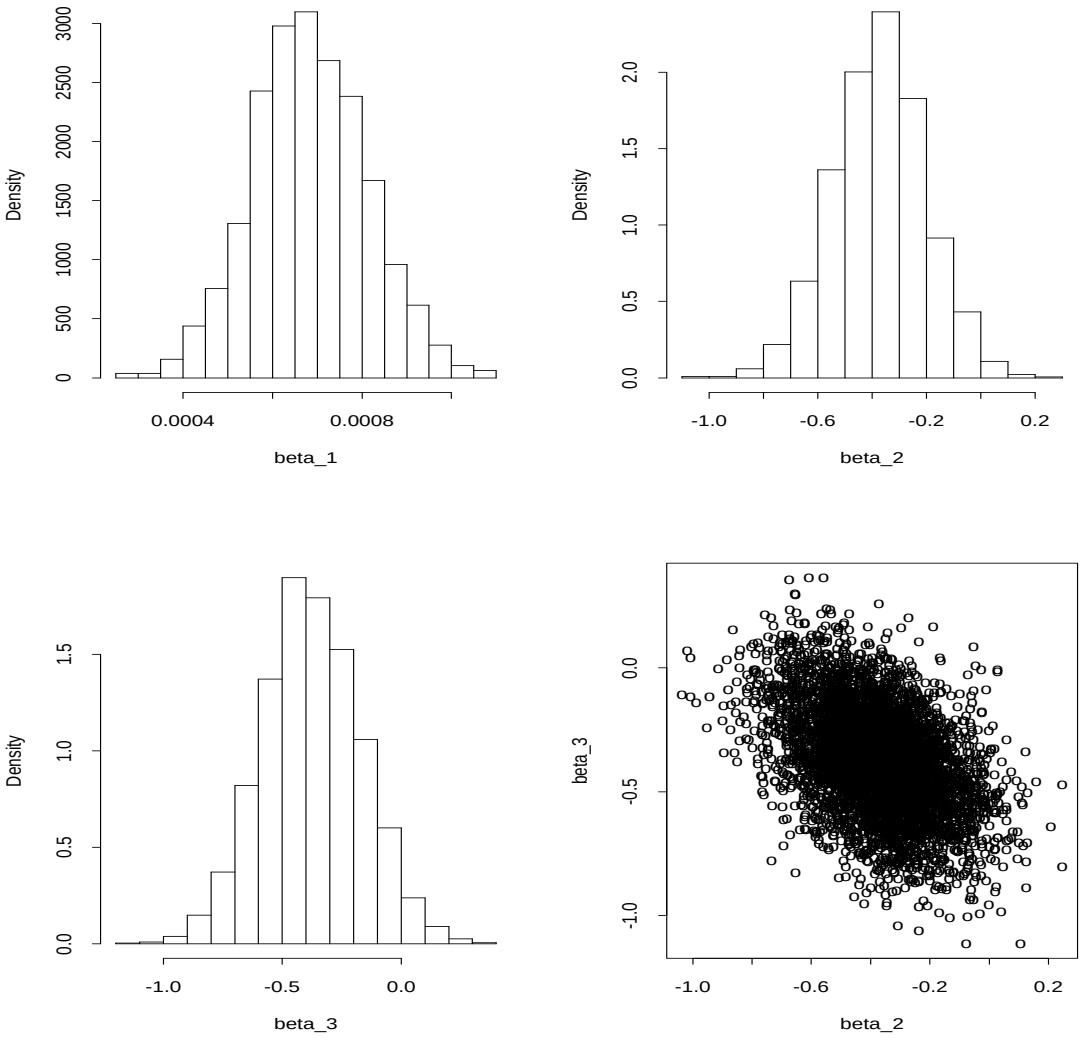
- ν^2 próximo de zero



superfície predita $\widehat{S}(x)$ (média à posteriori)



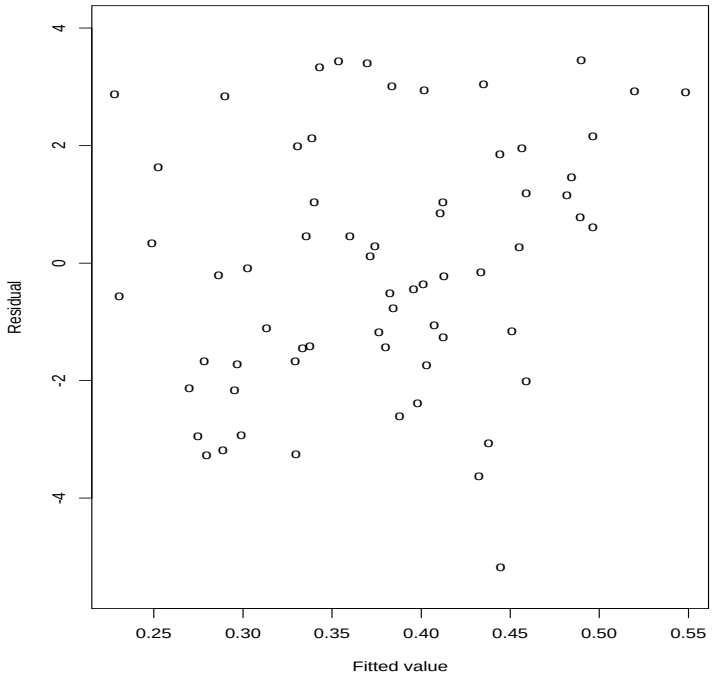
Posterioris para $S(x)$ em dias localizacoes, linha sólida – remota (452, 1493), linha interrompida – central (520, 1497)



posterioris para os parâmetros de regressão

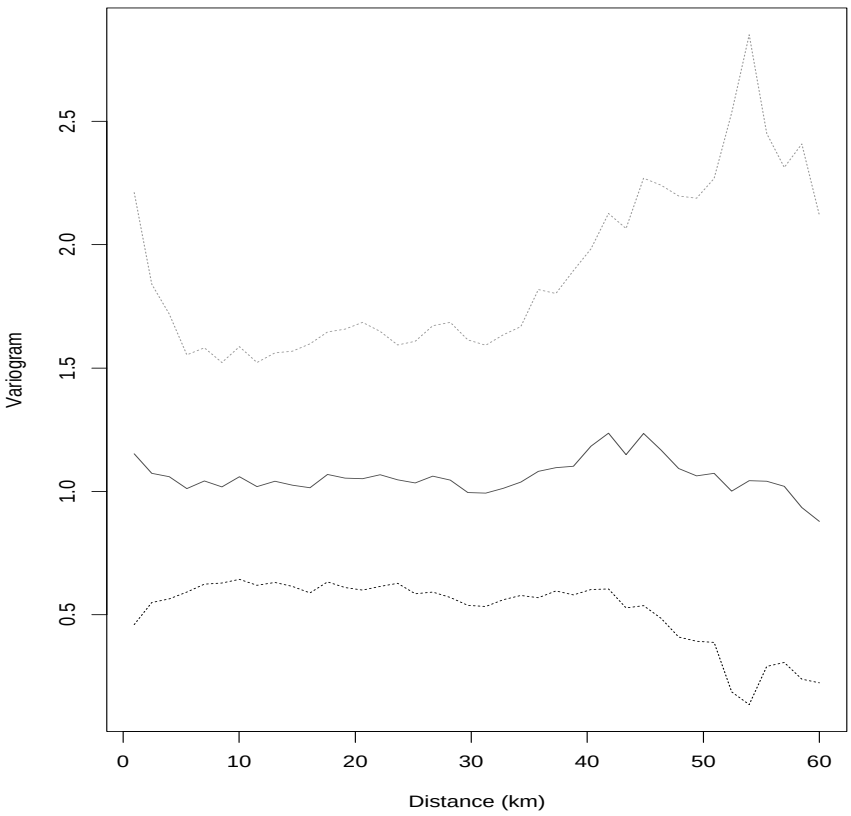
- $\beta_1 =$ efeito de idade
- $\beta_2 =$ efeito de mosquiteiro não tratado
- $\beta_3 =$ efeito adicional de tratamento de mosquiteiro

Qualidade do ajusto do modelo



resíduos de vila *vs* valores ajustados

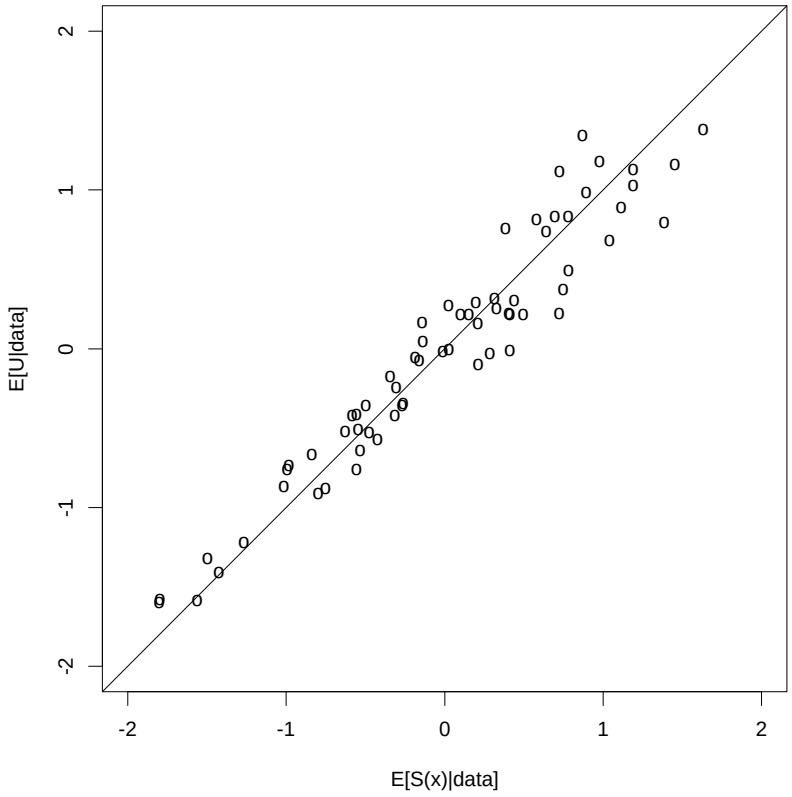
- $r_{ij} = (Y_{ij} - \widehat{p}_{ij})/\sqrt{\{\widehat{p}_{ij}(1 - \widehat{p}_{ij})\}}$
- $r_i = \sum r_{ij}/\sqrt{n_i}$
- checa adequacidade do modelo para p_{ij}



variograma empírico de resíduos padronizados com intervalos de confiança (95%) construídos a partir de simulações do modelo ajustado

- $r_{ij} = (Y_{ij} - \widehat{p}^*_{ij})/\sqrt{\{\widehat{p}^*_{ij}(1 - \widehat{p}^*_{ij})\}}$
- $r_i = \sum r_{ij}/\sqrt{n_i}$
- $\text{logit}(p^*_{ij}) = \widehat{\alpha} + z'_{ij}\widehat{\beta}\widehat{S}(x_i)$
- checa adequacidade do modelo para $S(x)$

O modelo geostatistico é mesmo necessário?



média da posteriori para os efeitos aleatórios $\widehat{U}_i$ de um GLMM não espacial contra médias a posteriori de $\widehat{S}(x_i)$ nas localizações observadas no modelo geoestatístico

- alta correlação evidencia depêndencia espacial

Potenciais para desenvolvimento e co-laboração

- dados censurados
- dados multivariados
- métodos não -lineares
- modelagem espaço-temporal
- amostragem preferencial
- processos não -estacionários
- processos pontuais marcados
- . . .

- todos modelos são errados, mas alguns são úteis

- independentemente do modelo adotado, procedimentos de inferência que respeitam princípios gerais de estatística tendem a ter melhor performance do que procedimentos *ad-hoc*

- ignorar incerteza associada aos parâmetros pode prejuidcar seriamente os intervalos de predição nominais

- o paradigma Bayesiano fornece uma integração tratável para estimação e predição

- entretanto resultados podem ser especialmente sensíveis à elicitação das prioris

- os melhores modelos tendem a ser aqueles desenvolvidos em colaboração entre estatísticos e especialistas no tema em estudo

- NÃO ANALISE DADOS . . .

  . . . ANALISE PROBLEMAS !

22. Programas computacionais

- programa estatístico: **R**
  `www.r-project.org`
- "package" ("library"): **geoR**
  `www.maths.lancs.ac.uk/~ribeiro/geoR.html`
- versões para **Linux**, Windows e Mac
- versão para S-PLUS (baixa prioridade)
- introdução ao pacote:
  `www.maths.lancs.ac.uk/~ribeiro/geoRintro.html`
- literatura relacionada