Introdução a Geoestatística

Paulo Justiniano Ribeiro Junior *

LEG - Laboratório de Estatística e Geoinformação
Departamento de Estatística
Universidade Federal do Paraná

Material de Curso

Piracicaba, SP
Agosto, 2005

Endereço para correspondência: Departamento de Estatística, Universidade Federal do Paraná, E-mail: paulojus@est.ufpr.br

INTRODUÇÃO À ESTATÍSTICA ESPACIAL

1. **Exemplos Básicos de dados espaciais**

2. **Terminologia para estatística espacial**

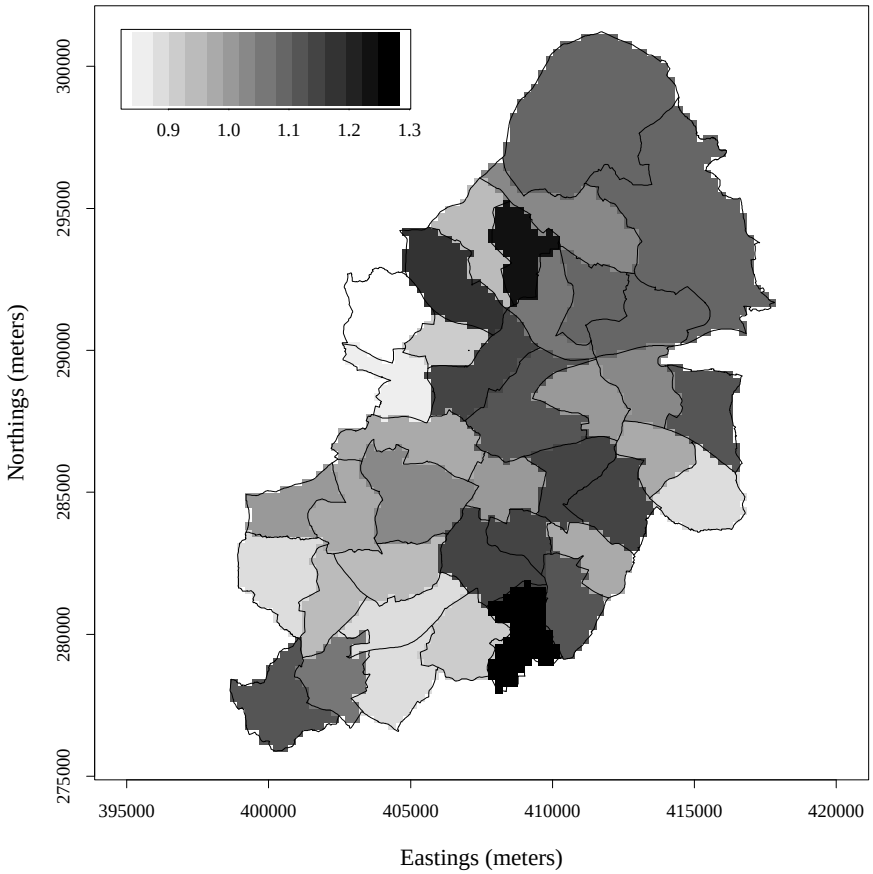
3. **Modelos e problemas em estatística espacial**

4. **Um estudo de caso: doença de citrus**

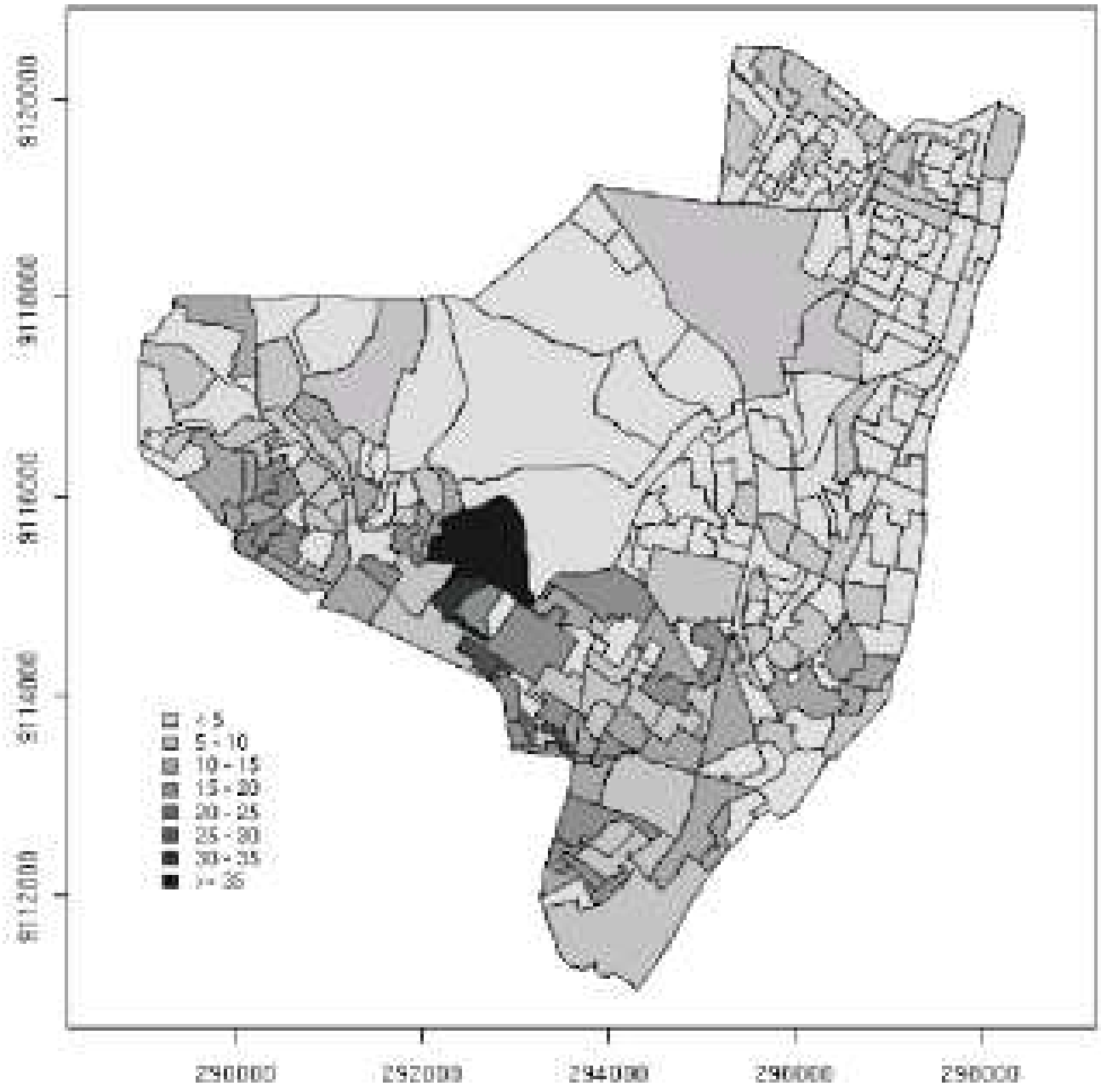
1. Estatística Espacial: Alguns Exemplos Básicos

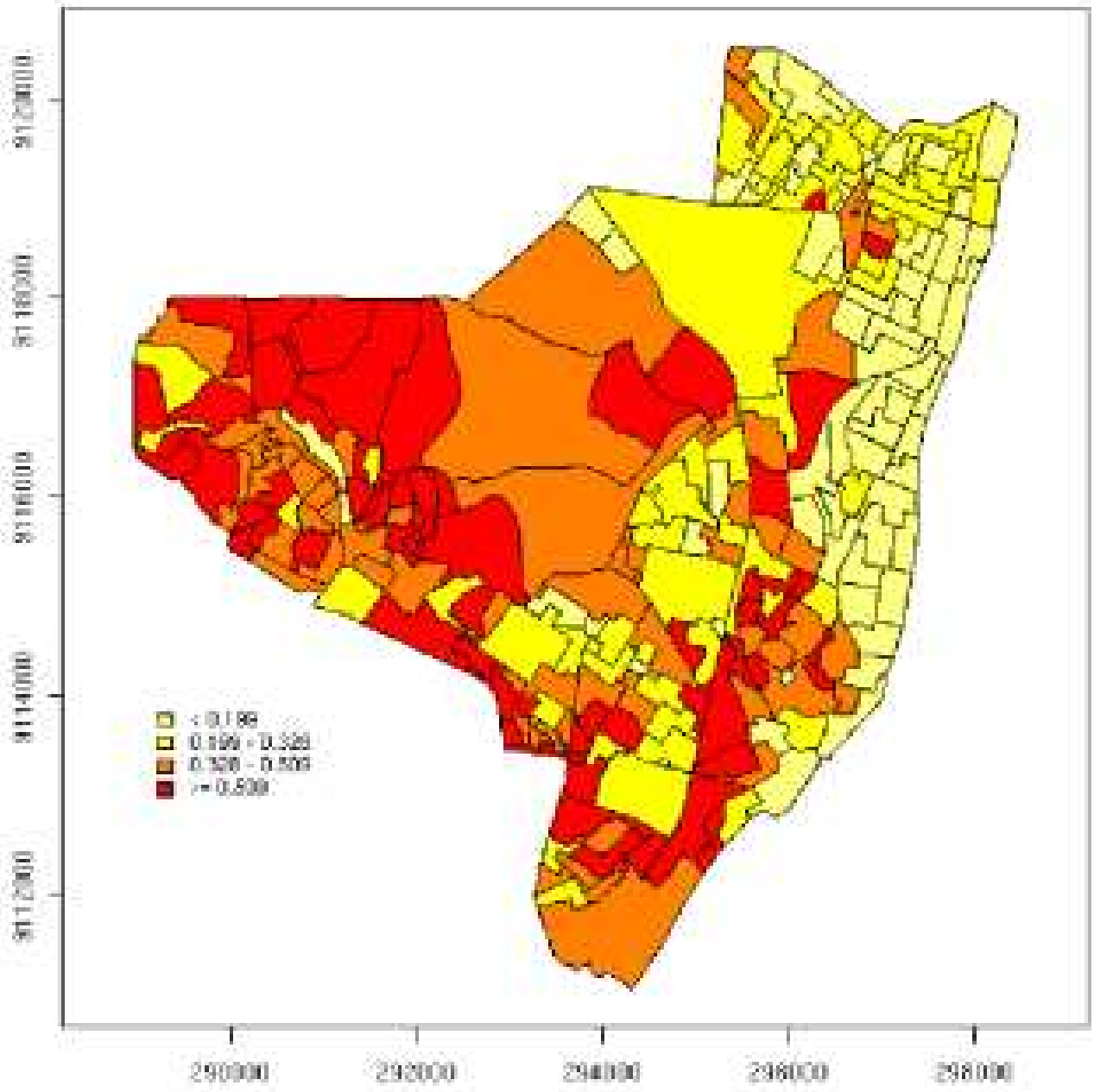
(a) Taxas de câncer por regiões administrativas

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



(b) Hanseníase em Olinda




290000
292000
294000
296000
298000

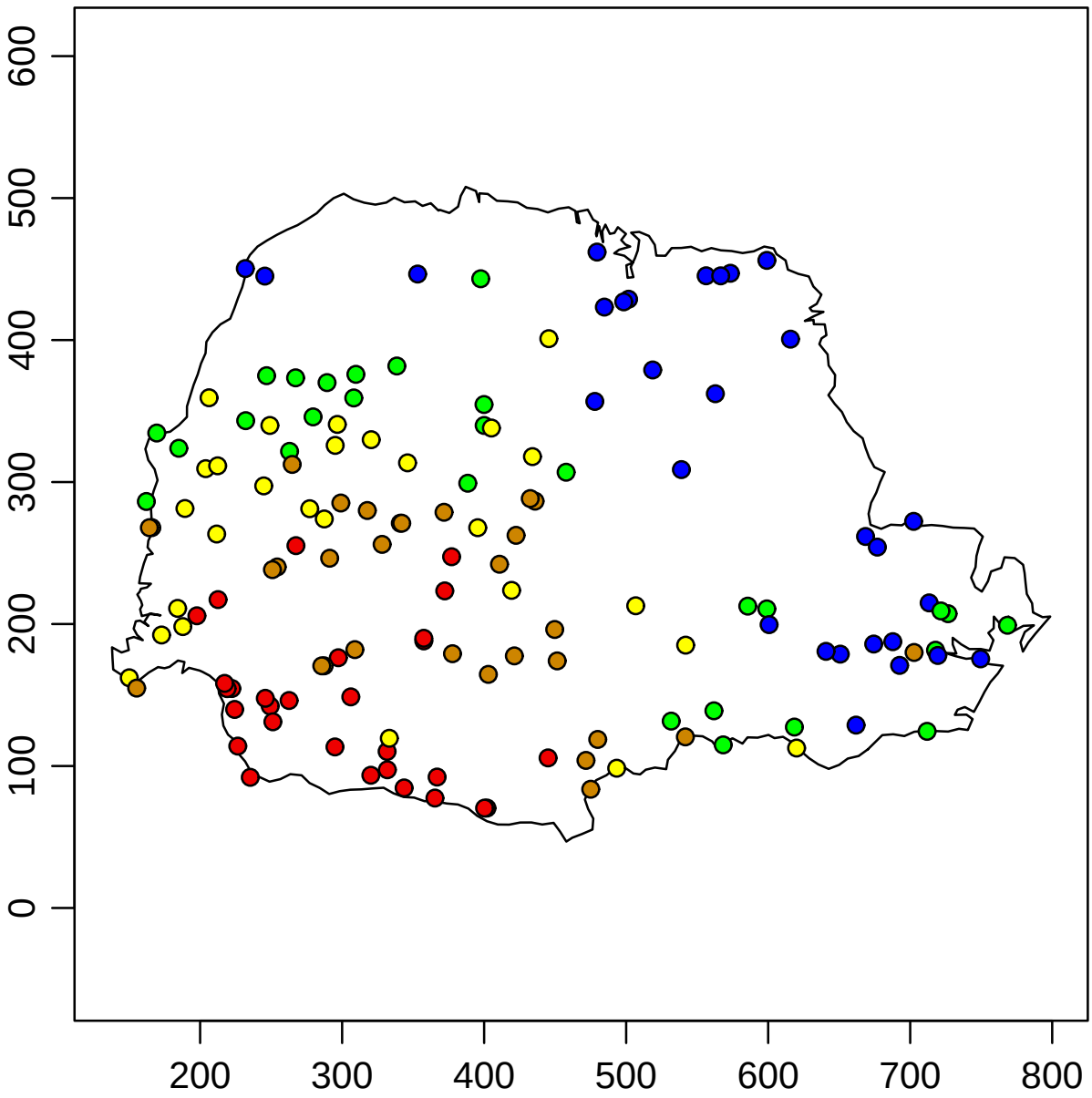
Alguns problemas com estrutura de dados semelhante

- Índices de criminalidade por bairros
- Mortalidade infantil (e/ou outros indicadores) por municípios de um estado
- Experimentos agrícolas de campo
- Análise de imagens

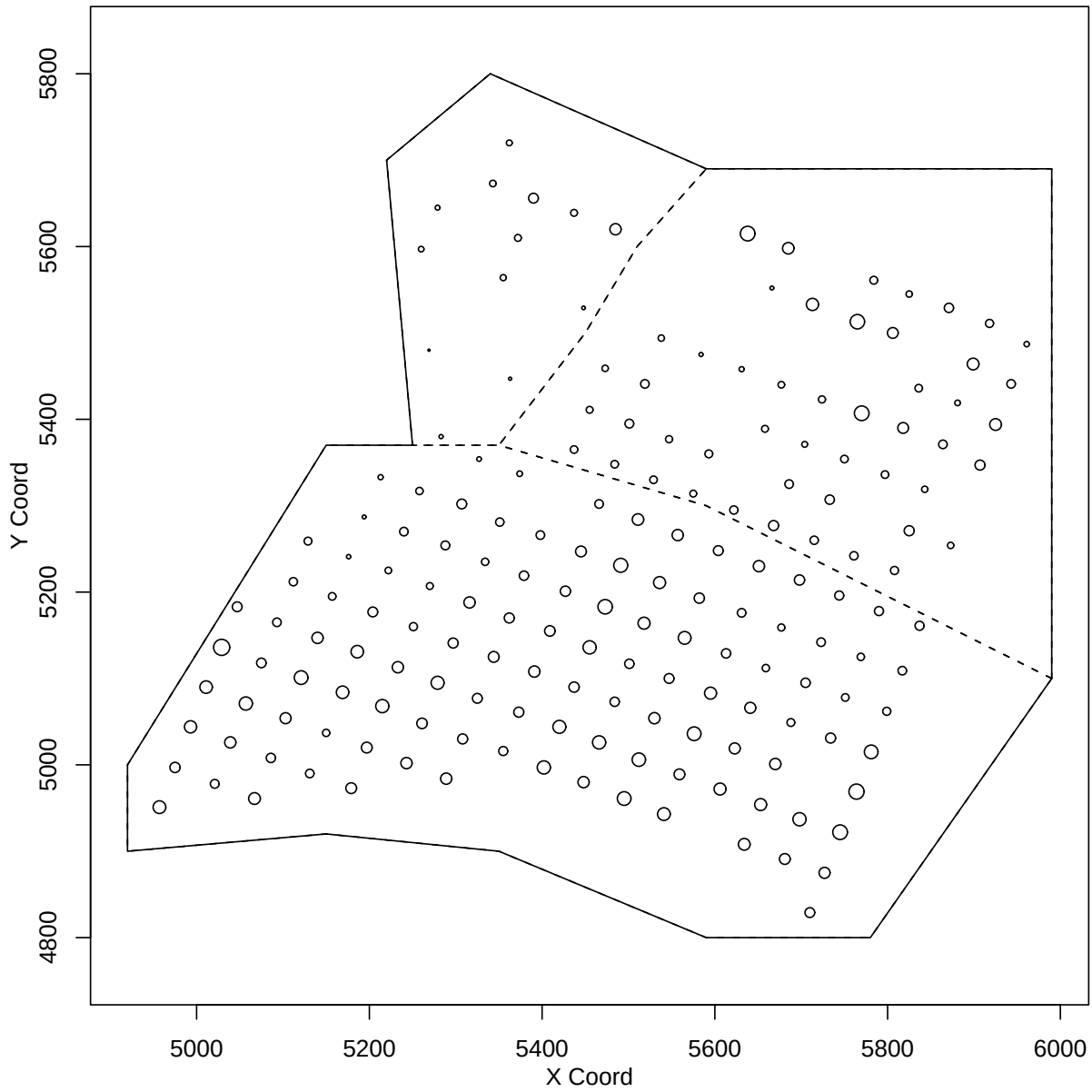
(c) Precipitação no Estado do Paraná

Medidas de chuva em 143 postos meteorológicos.

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



(d) Teores de Cálcio em um solo agrícola

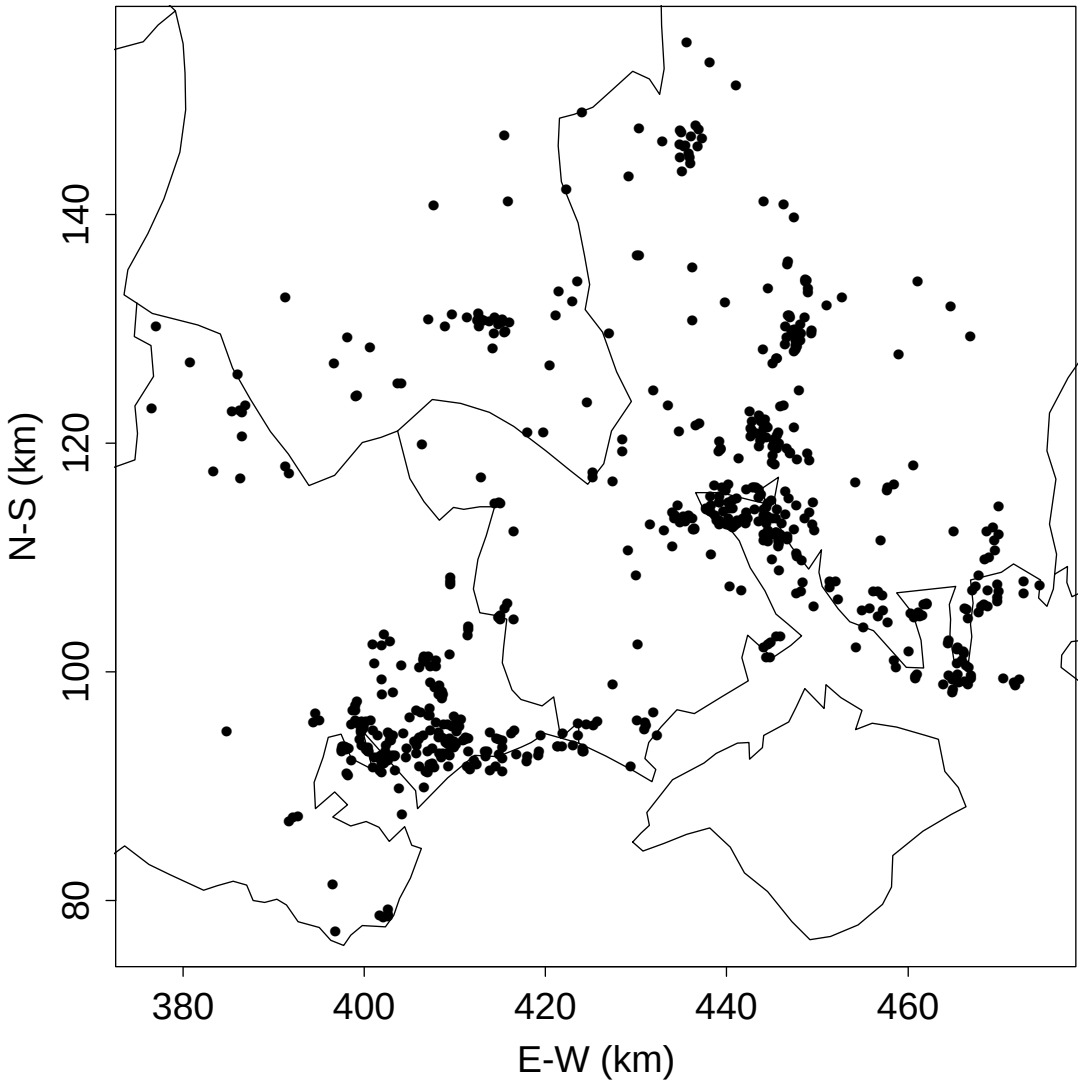


Alguns problemas com estrutura de dados semelhante

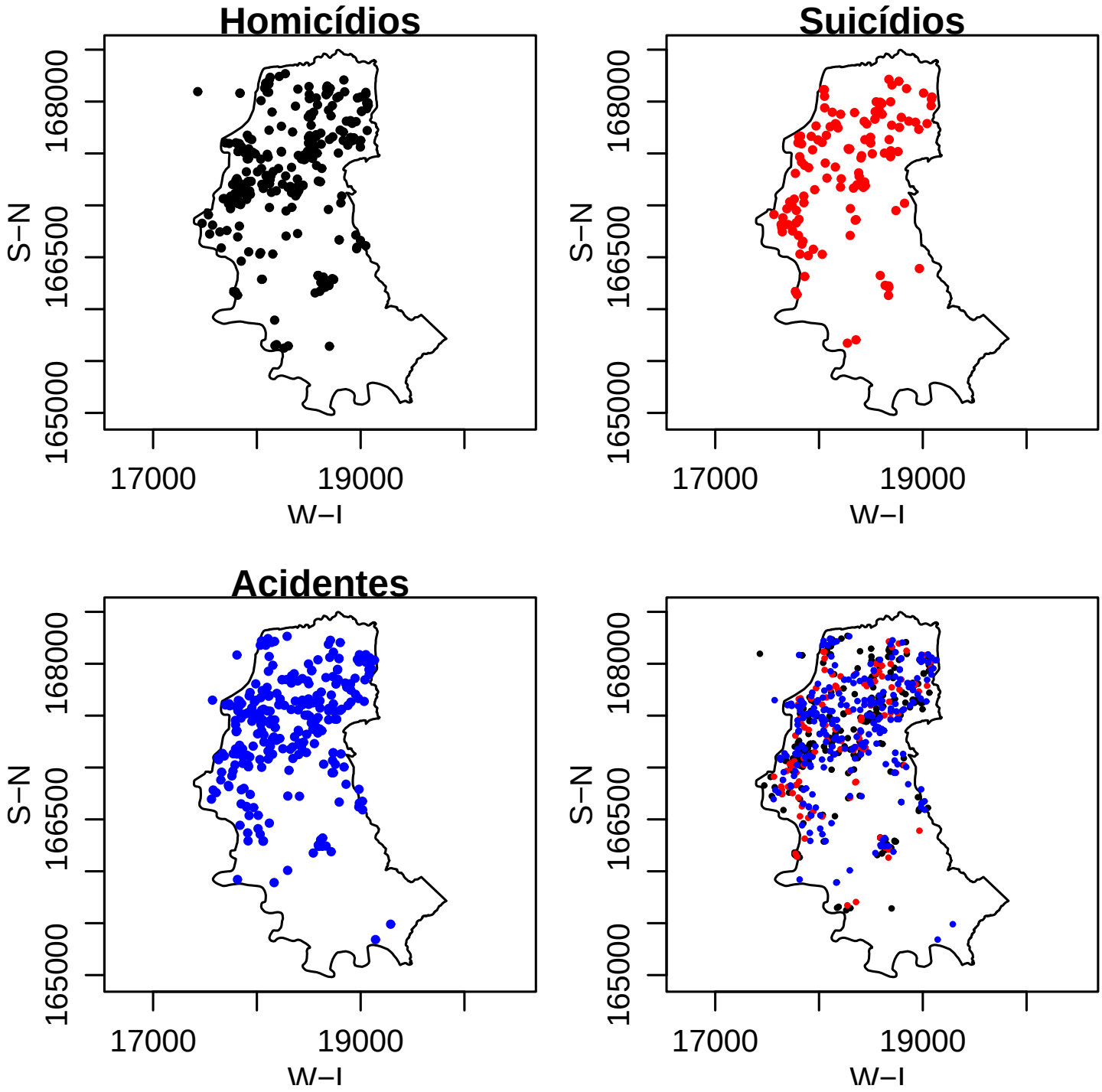
- Teores de elementos minerais em uma jazida
- Níveis de poluição do ar medidos em estações de monitoramento
- Estoque de peixes em uma certa área marítima

(e) Infecções bacterianas no sul da Inglaterra

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



(f) Ocorrências em Porto Alegre



Alguns problemas com estrutura de dados semelhante

- Localização de árvores de certa espécie em uma area de floresta natural

- Pontos de ocorrência de crimes em uma cidade

- Posições de ninhos de certo pássaro em uma região

2. Terminologia e questões para estatística espacial

(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 $[Y_i|Y_j, \forall j \neq i]$

i. Medidas de agregação comumente utilizadas (por ex. *I de Moran*)

ii. Diversas opções de modelos, entre eles:

iii. a. Regressão ponderada geográficamente

iv. b. Modelos CAR (auto-regressivo condicional) e SAR

v. Definição de vizinhança pode depender do problema

(b) Variação espacial contínua

Estrutura básica. $Y(x) : x \in \mathbb{R}^2$

- 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* $Y(x)$

i. em geral (mas nem sempre!) o objetivo é de predição
ii. a predição pode ser do processo subjacente ou um funcional deste
iii. possíveis relações com covariáveis
iv. modelos comumente utilizados podem ser vistos como MLG, com efeito aleatório espacialmente estruturado
v. grande número de métodos/algorítmos "ad-hoc" na literatura de geoestatística

(c) Processo pontual espacial

Estrutura básica. Conjunto contável de pontos $x_i \in \mathbb{R}^2$, gerados estocásticamente.

- às vezes dados são agregados em regiões

i. questão chave é dizer se processo é aleatório, agrupado ou regular
ii. modelagem básica para superfície de intensidade do processo pontual
iii. vários modelos disponíveis, "fácil" de simular, difícil de estimar
iv. conhecimento do processo subjacente pode guiar escolha de modelos
v. alguns pontos importantes: correções de borda, correções para população sob risco, estudos de caso-controle, etc

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

Temas estratégicos

- não confundir *formato dos dados* com o *processo subjacente*.
- a escolha do modelo pode ser influenciada pelos objetivos científicos do estudo
- problemas reais não necessariamente se encaixam em um dos tipos básicos, podem ser abordados de difrentes formas ou conterem elementos de cada um dos tipos. A divisão é puramente didática
- há outras possibilidades tais comos processos pontuais marcados, processos espaço temporais, etc
- Estatística espacial e Sistemas de Informação Geográfica (SIG)
- Estatística espacial e Geoestatística
- Problemas espaço-temporais

3. Estudo de caso – Doença de citrus

Problema consiste em quantificar a importância relativa dos dois mecanismos de reprodução do fungo causador a Pinta Preta dos citrus.

Conjectura-se que o padrão espacial da fase assexuada deve ser agragado enquanto que o da fase sexuada deve ser aleatório.

Análises ilustram o uso de métodos associados a cada um dos "tipos básicos" a um mesmo problema.

4. **Palavras finais**

NÃO ANALISE DADOS . . . . . .

. . . . . ANALISE PROBLEMAS !

PROJETO SAUDAVEL
http://saudavel.dpi.inpe.br

LEG – Laboratório de Estatística de Geoinformação
http://www.est.ufpr.br/leg

Geoestatística

1. Geoestatística: outros exemplos

2. Questões Centrais

3. Modelo Geoestatístico

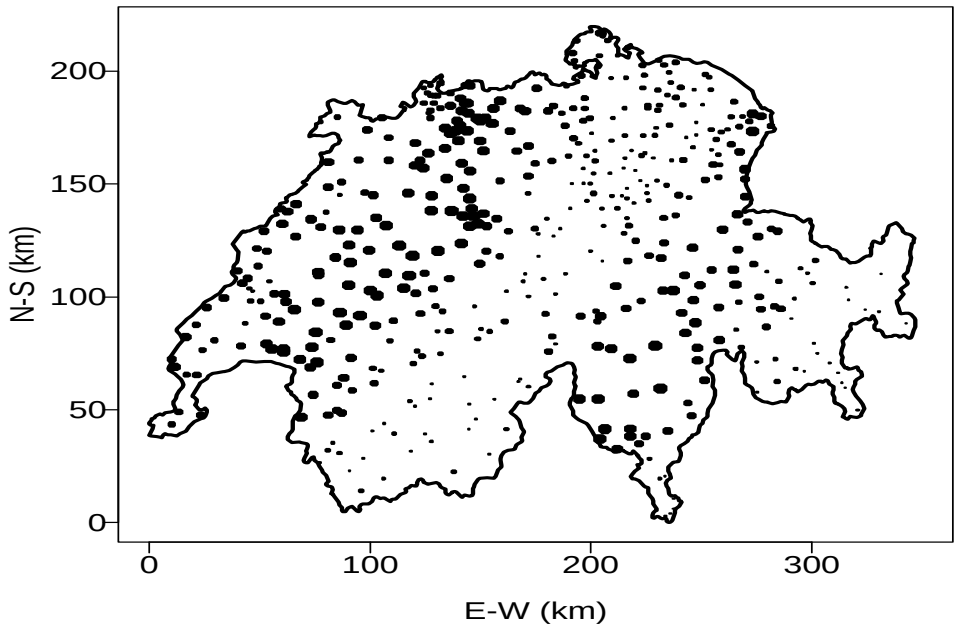
4. O modelo Gaussiano

5. Predição

6. Estudo de Caso

1. Outros Exemplos de Problemas Geoestatísticos

(a) Dados de chuva na Suíça

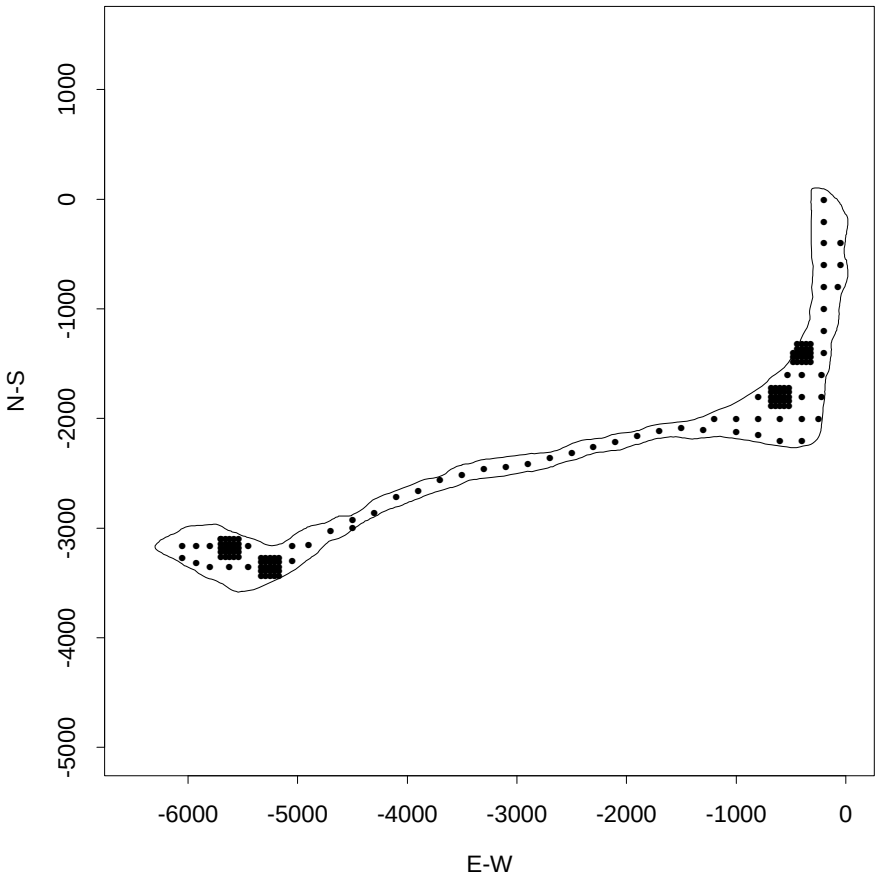


Localizações com tamanhos dos pontos proporcionais aos valores observados de precipitação

- 467 postos na Suíça
- medidas diárias de chuva em 8/05/1986
- dados do projeto:
  Spatial Interpolation Comparison 97
  `ftp://ftp.geog.uwo.ca/SIC97/`.

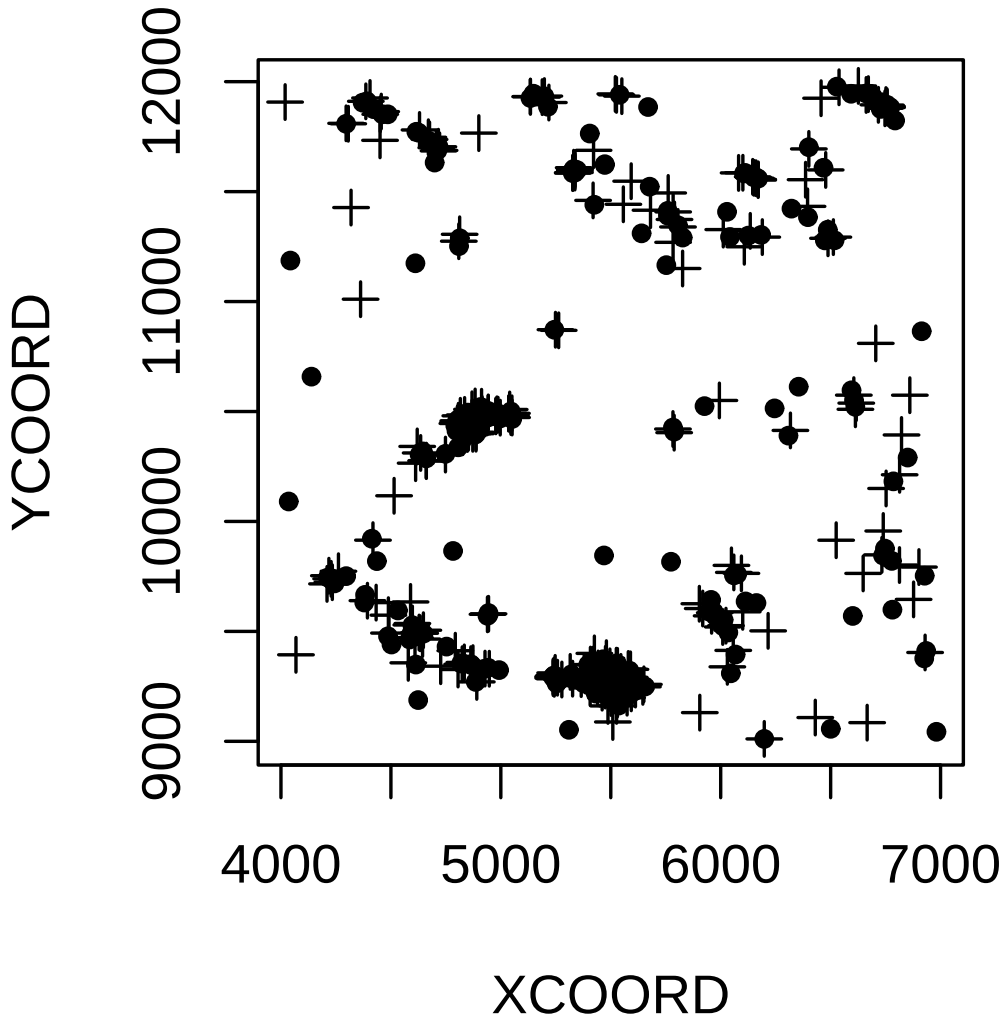
(b) Ilha de Rongelap

- estudo do resíduo de contaminação decorrente de testes de armas nucleares durante a década de 50
- ilha evacuada em 1985. Segura para reocupação
- pesquisa produz medidas com ruído Y_i de concentração de césio radioativo
- particular interesse em níveis máximos de concentração de césio



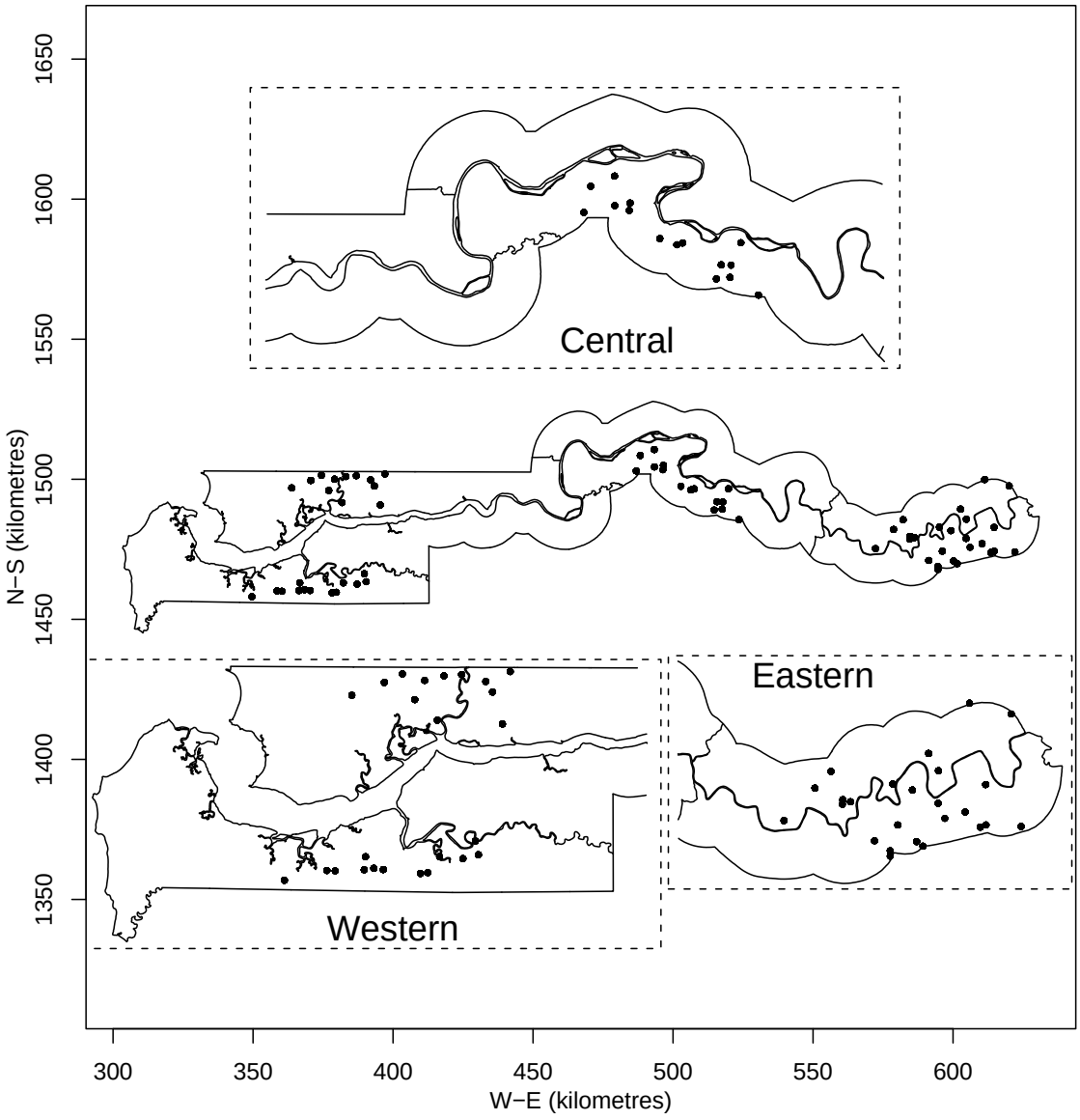
(c) Espécies de líquens

- 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 Gambia

- 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



2. Características Principais dos Problemas Geoestatísticos

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

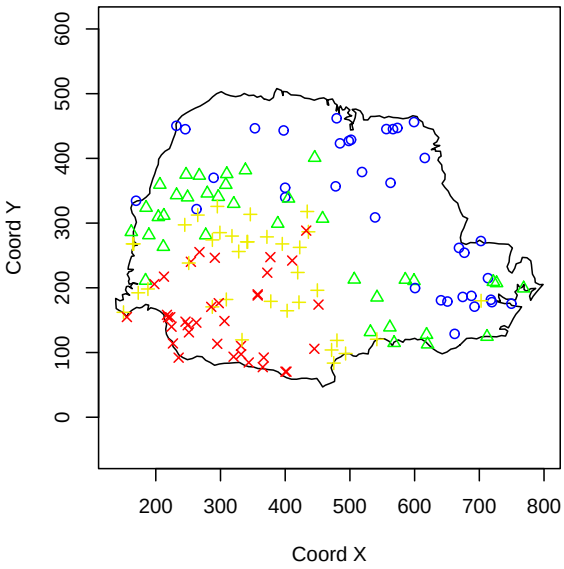
- em princípio, Y pode ser determinado em qualquer localização x dentro da região espacialmente contínua A

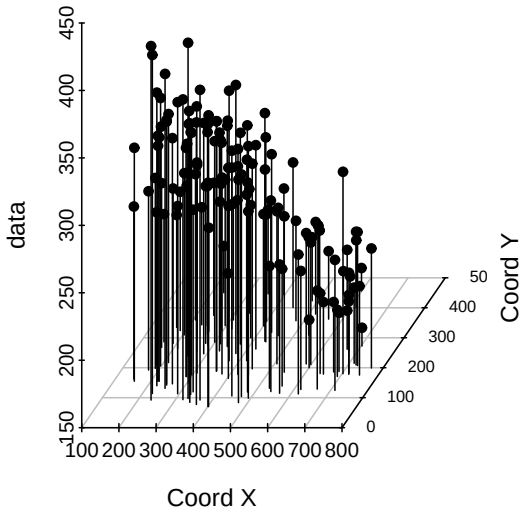
- assume-se que $\{Y(x) : x \in A\}$ é um processo estocástico

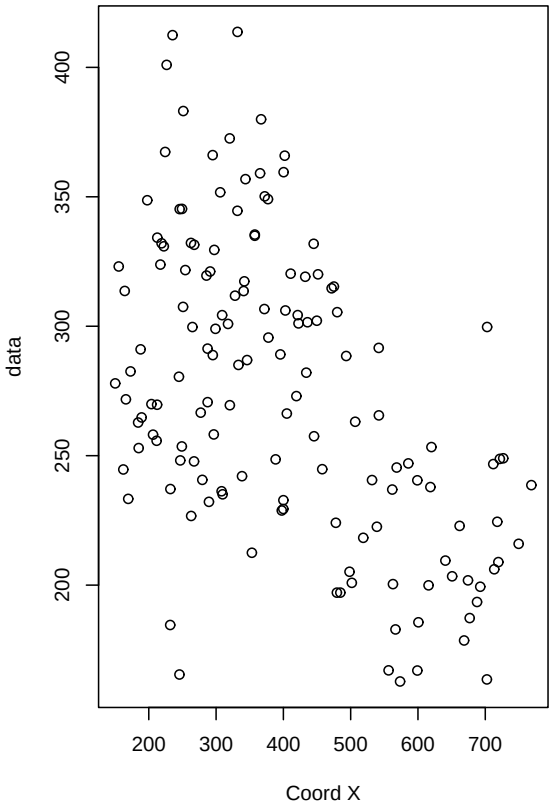
- x_i é tipicamente fixo. Se as localizações x_i são geradas por um processo estocástico pontual, assume-se que este processo é independente de $Y(x)$

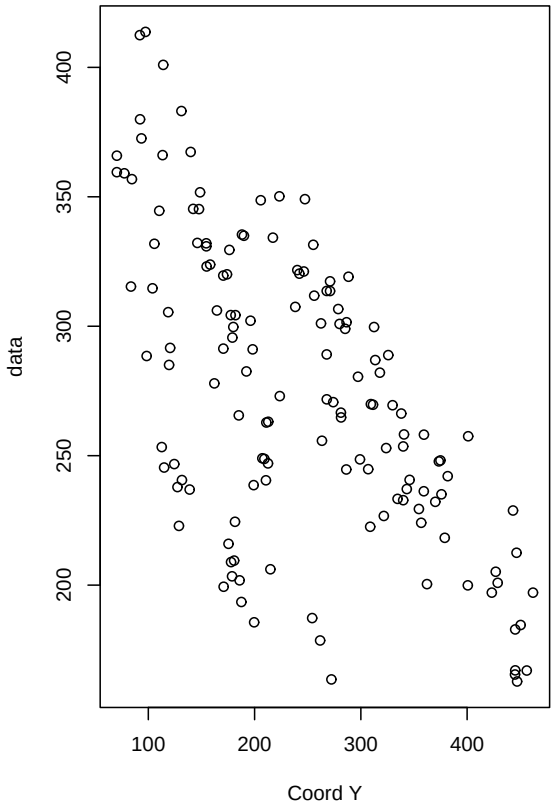
- objetivos científicos incluem a predição de um ou mais funcionais de processo (sem ruído) $\{S(x) : x \in A\}$

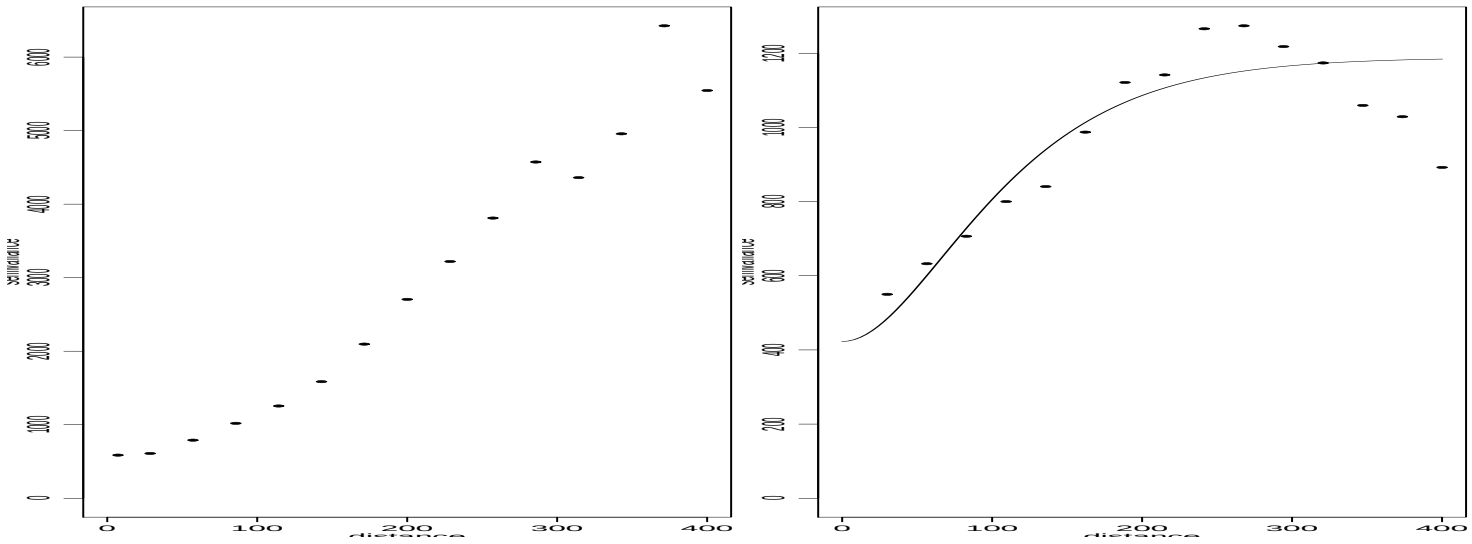
Exemplo básico: chuva no Paraná



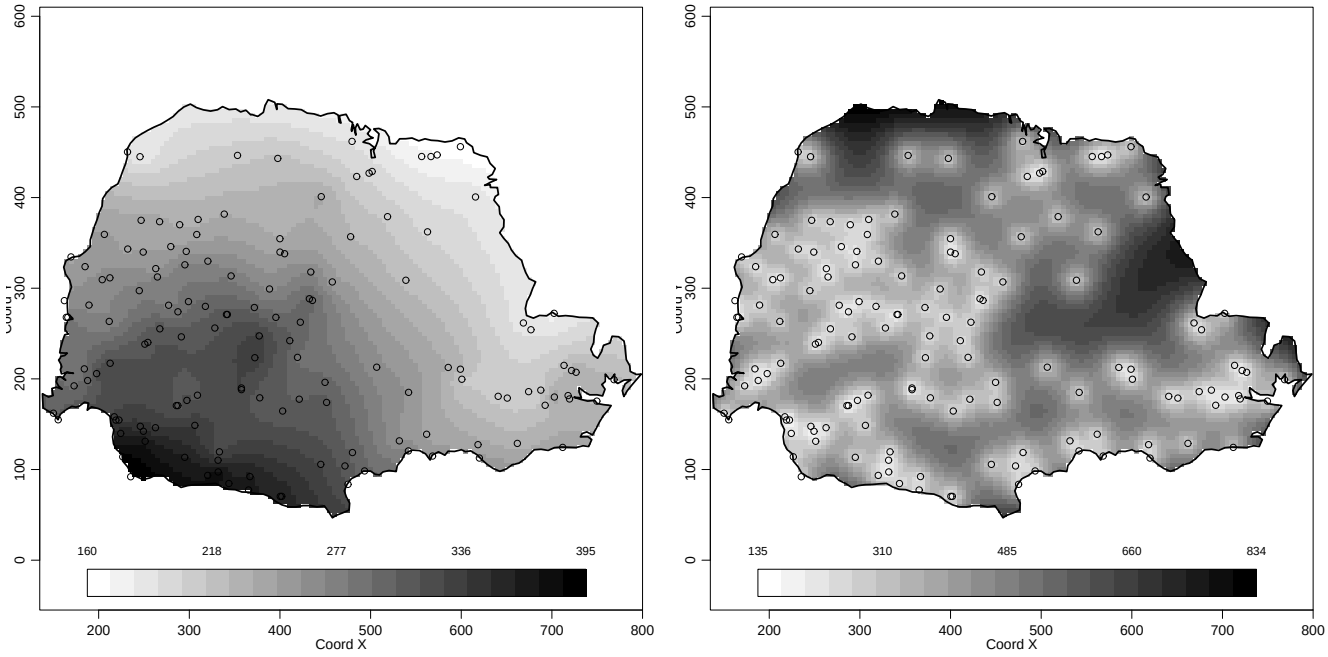








variogramas para dados originais (esquerda) e após retirada de tendência, com modelo ajustado (direita).



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

3. 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

4. "Geoestatística baseada em modelos"

O termo "Geoestatística baseada em modelos" significa que adotamos um enfoque baseado em modelos para esta classe de problemas, o que quer dizer que começamos com um modelo estocástico explícito e derivamos métodos de estimação de parâmetros, interpolação e suavização através da aplicação de princípios gerais de estatística.

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**

Modelo linear generalizado linear clássico

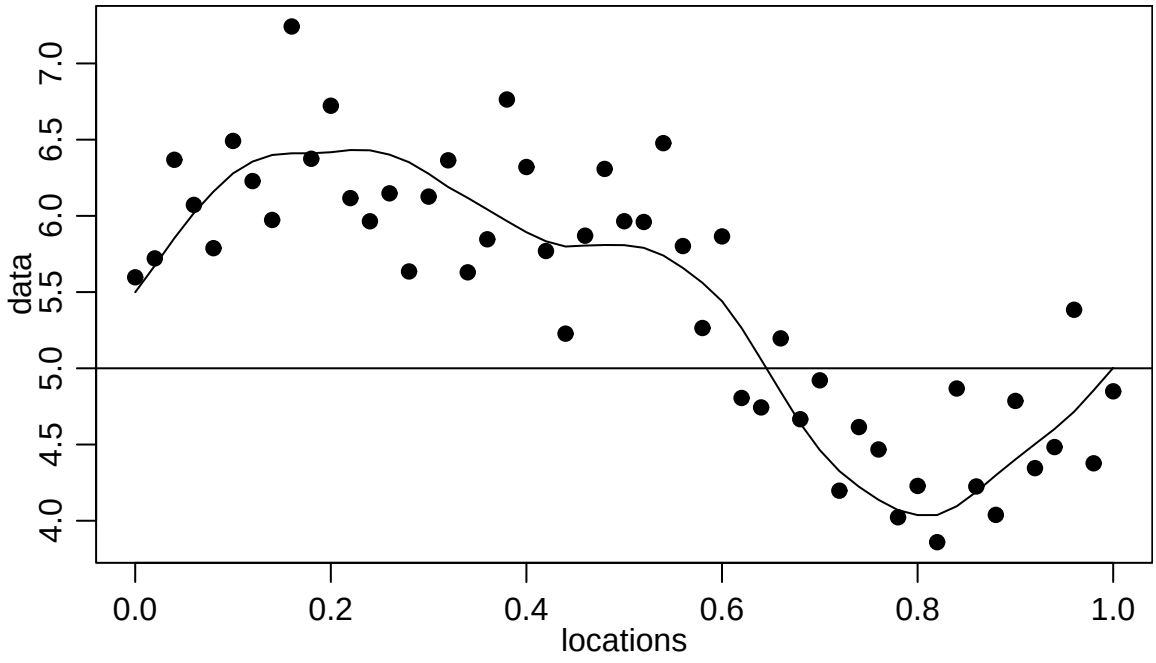
- $Y_i : i = 1, ..., n$
  mutuamente independentes, com $\mu_i = \mathrm{E}[Y_i]$
- $h(\mu_i) = \sum_{j=1}^{k} f_{ij}\beta_j$, com função de ligação conhecida $h(\cdot)$.

Modelo Linear Generalizado Mixto

- $Y_i : i = 1, ..., n$
  mutuamente independentes, com $\mu_i = \mathrm{E}[Y_i]$, conditional às realizações de um conjunto de de variáveis aleatórias latentes U_i,
- $h(\mu_i) = \sum_{j=1}^{k} f_{ij}\beta_j + U_i$,
  para uma função de ligação conhecida $h(\cdot)$.

A modelo espacial (geoestatístico)

- $Y_i : i = 1, ..., n$
  mutuamente independentes, com $\mu_i = \mathrm{E}[Y_i]$, conditional às realizações de um conjunto de de variáveis aleatórias latentes U_i,
- $h(\mu_i) = U_i + \sum_{j=1}^{p} f_{ij}\beta_j$,
  para uma função de ligação conhecida $h(\cdot)$,
- $U_i = S(x_i)$
  onde $\{S(x) : x \in \mathbb{R}^2\}$ é um processo estocástico espacial.
- $h(\mu_i) = \sum_{j=1}^{k} f_{ij}\beta_j + U_i$,



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

5. O Modelo Gaussiano

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

i. $\mathrm{E}[S(x)] = 0$,

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 $\mu + S(x_i)$ e variância τ^2;

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

Uma formulação equivalente para o modelo Gaussiano:

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

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

Desta forma a distribuição conjunta de Y é multivariada Normal,

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

onde:

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.

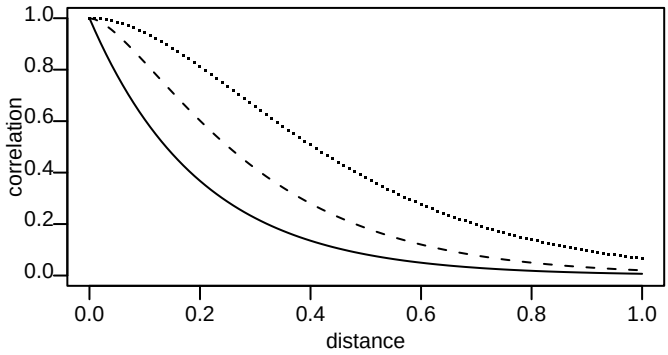
6. Especificação da função de correlação

A família de Matérn

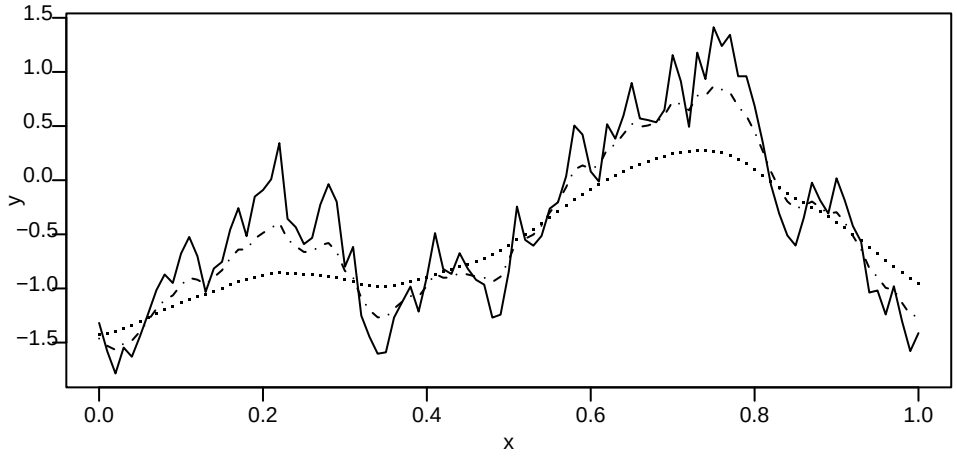
Função de correlação dada por

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

- 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).

7. Extensões do modelo básico

(a) Modelos Gaussianos transformados

- O modelo Gaussiano é claramente inapropriado para distribuições assimétricas.

- Certos dados podem indicar relações entre média e variância, que violam o modelo Gaussiano.

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

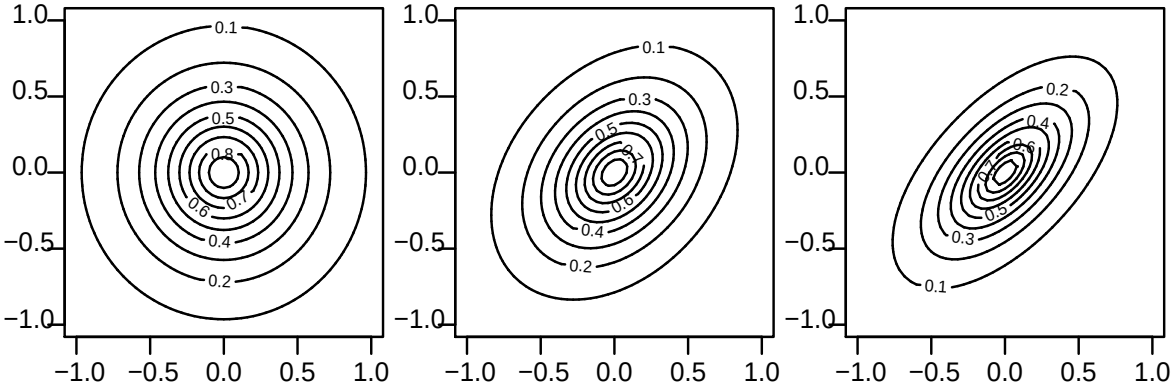
- O modelo fica então definido da forma:
  - assume-se $Y^* \sim MVN(F\beta, \sigma^2 V)$

  - dados $y = (y_1, ..., y_n)$, são gerados por uma transformação do modelo linear Gaussiano $Y = h_\lambda^{-1}(Y^*)$ tal que:

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

(b) Efeitos Direcionais

- Condições ambientais podem induzir efeitos direcionais (vento, formação do solo, etc)
- como consequência a correlação espacial pode variar com a direção

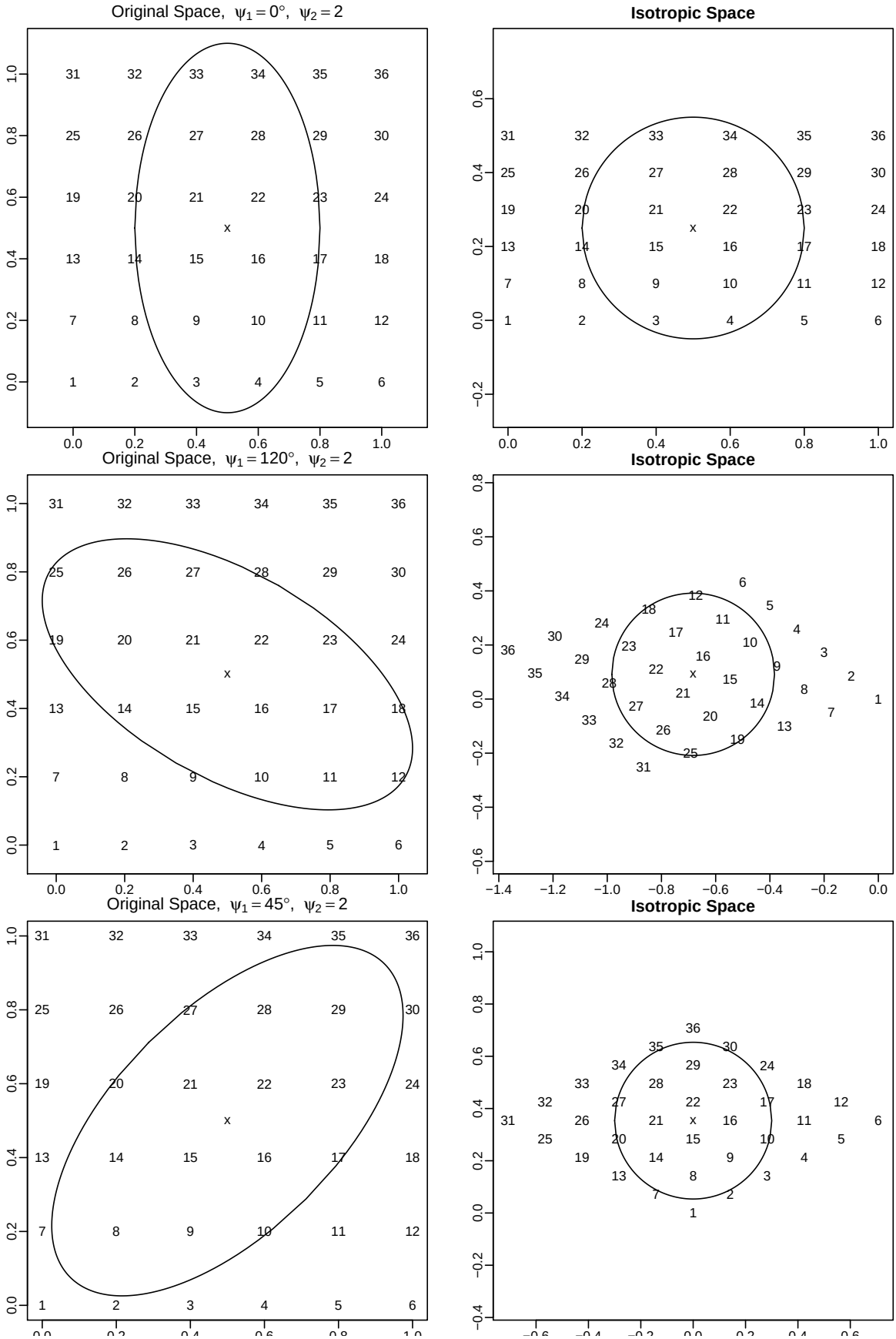


contornos de correlação para modelo isotrópico (esq.) e dois modelos anisotrópicos (centro e dir.)

- *anisotropia geométrica*: possível (e simples) abordagem.
- dois parâmetros extra: *ângulo de anisotropia* ψ_A e *razão de anisotropia* ψ_R.
- rotação e contração/expansão das coordenadas originais:

$$(x_1\prime, x_2\prime) = (x_1, x_2) \begin{pmatrix} \cos(\psi_A) & -\sin(\psi_A) \\ \sin(\psi_A) & \cos(\psi_A) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\psi_R} \end{pmatrix}$$

“Correção” de anisotropia geométrica



(c) Modelos não estacionários

- *Modelos com médias não constantes (ou, incluindo covariáveis)*

  Substituir a média constante μ por

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

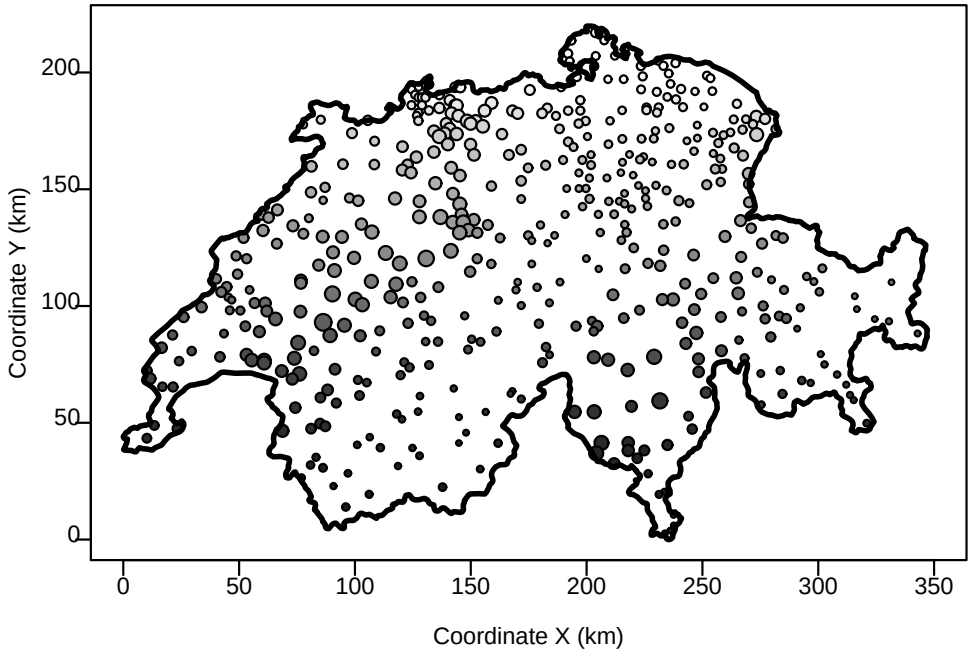
  para medidas $f_j(x)$ das covariáveis (lineares ou não lineares).

- *Variação aleatória não estacionária*

  Variabilidade **intrínsica**: pressuposto mais fraco de estacionaridade (processo com incrementos estacionários, como passeios aleatórios em séries temporais), largamente utilizados como modelo padrão para variação espacial discreta (Besag, York and Molié, 1991).

  Métodos de **deformação espacial** (Sampson and Guttorp, 1992) buscam estacionaridade por transformações (complexas) do espaço geográfico, x.
  É preciso ter em mente o balanço entre a o aumento da flexibilidade de modelos mais gerais contra a sobre-modelagem de dados esparsos, que leva a pobre identificação dos parâmetros.

8. Estudo de caso: chuva na Suíça



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

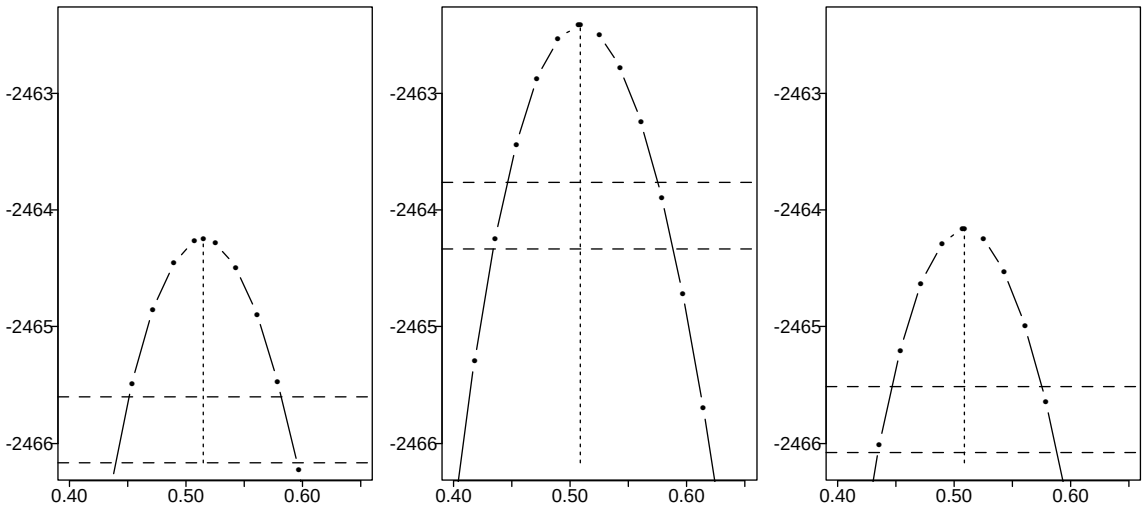
- 467 localizações
- medidas de precipitação em 8 de Maio 1986
- dados são valores inteiros com unidade de medida igual á 1/10 mm
- 5 localizações com valores iguais à zero.

chuva na Suíça (cont.)

Estimação parâmetros de transformação e suavidade (modelo de Matérn)

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

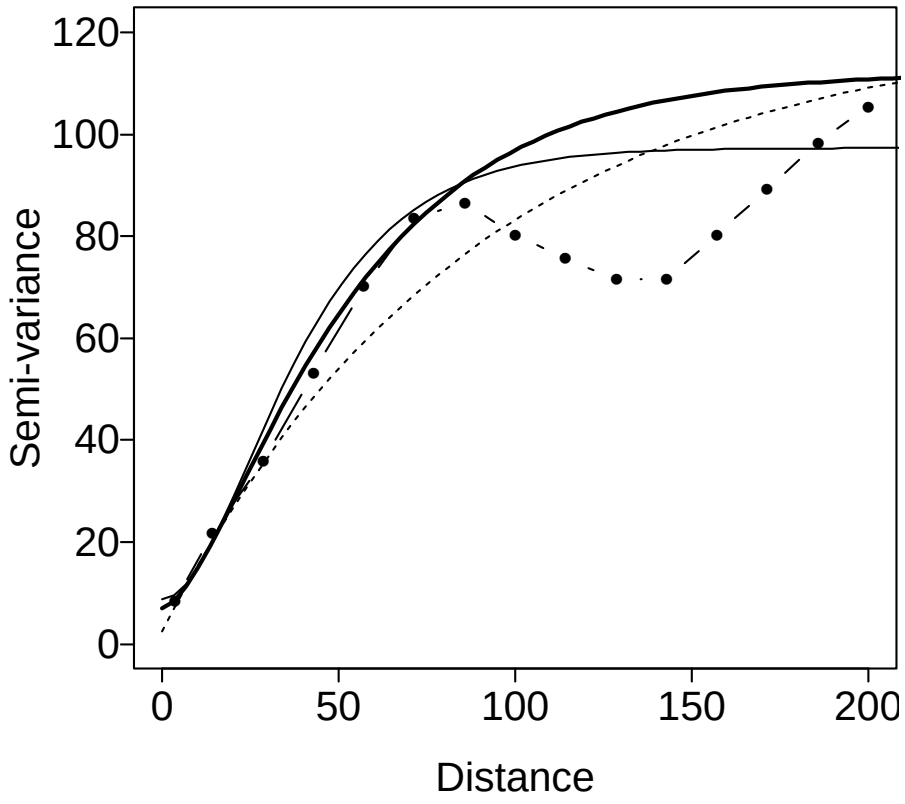
Estimativas de MV de $\hat{\lambda}$ e valores da log-verossimilhança $\log \hat{L}$ para diferentes valores de κ.



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

transformação logarítimica ou não transformação são claramente NÃO indicadas!

chuva na Suíça (cont.)



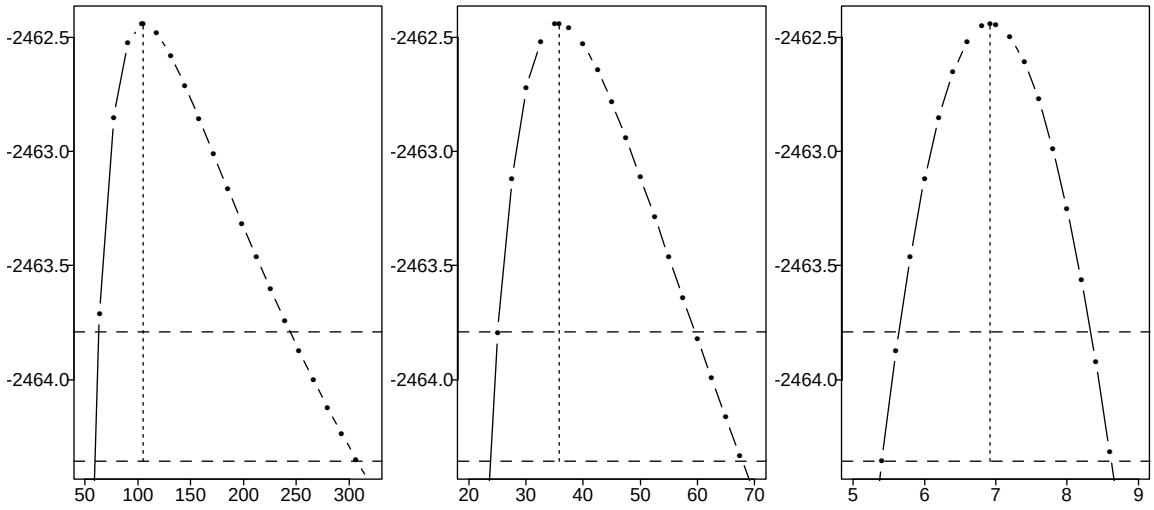
semivariograma empírico para dados transformados e variogramas teóricos com estimativas de MV para $\kappa = 0.5$ (linha interrompida), $\kappa = 1$ (linha grossa), $\kappa = 2$ (linha fina).

chuva na Suíça (cont.)

Estimativas para modelo com $\lambda = 0.5$

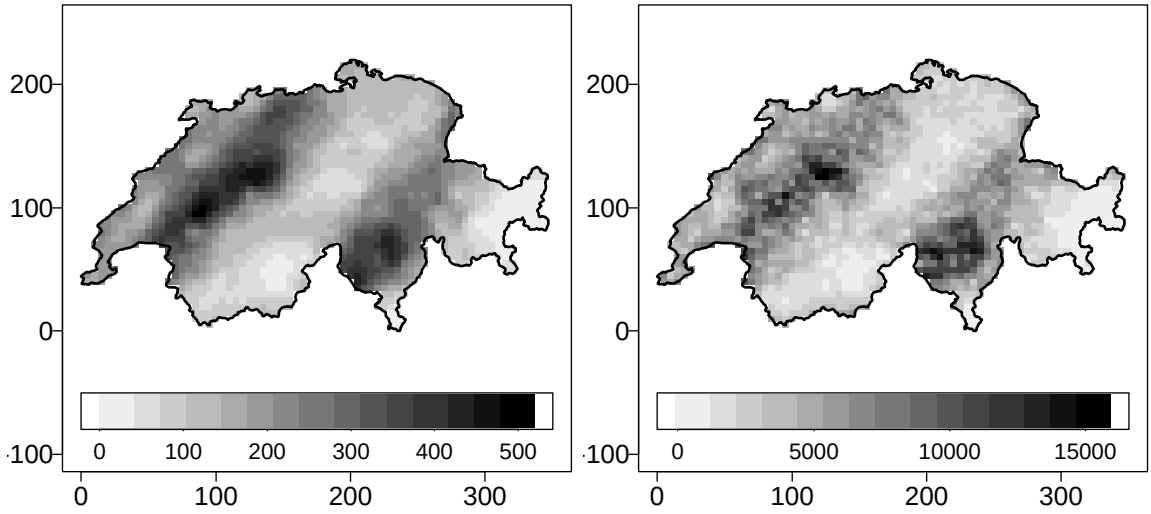
κ	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	$\log \hat{L}$
0.5	18.36	118.82	87.97	2.48	-2464.315
1	20.13	105.06	35.79	6.92	-2462.438
2	21.36	88.58	17.73	8.72	-2464.185

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



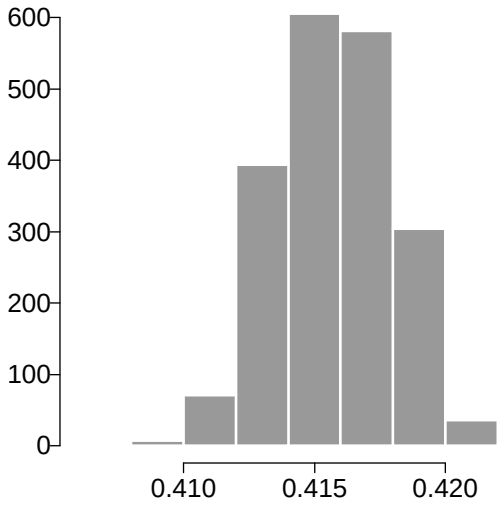
Verossimilhança perfilhada para parâmetros de covariância $\kappa = 1$ and $\lambda = 0.5$. esquerda: σ^2, meio: ϕ, direita: τ^2.

chuva na Suíça (cont.)



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 $\tilde{A}_{200}$.