Estatística Espacial
Geoestatística I

Raquel Menezes

Departamento de Matemática
Universidade do Minho

Setembro de 2023

Dados referentes a pontos - Geoestatística

- O nosso interesse ir-se-á focar na
  - modelação de dados $y_1, ..., y_n$ recolhidos em diferentes localizações $\mathbf{x_1}, ..., \mathbf{x_n}$ (por exemplo, latitude e longitude) na região de estudo $D \subset IR^2$
  - estimação da estrutura de **correlação espacial** subjacente aos dados observados

- Iremos assumir um **domínio espacial contínuo**
  - processo estocástico $Y(\mathbf{x})$ pode ser medido em $\forall \mathbf{x} \in D \subset IR^2$
  - classicamente referido por processo geoestatístico (razões históricas)
  - MAS atualmente cobrindo as diversas áreas de aplicação

Lembrar a importância da AED

- A análise exploratória de dados (AED) deve cobrir quer os aspetos *não espaciais* (e.g. boxplot) quer os *espaciais* (e.g. mapas).

Numa análise exploratória devemos fazer analise espacial e não espacial

- Atenção, as estatísticas pontuais não são suficientes!

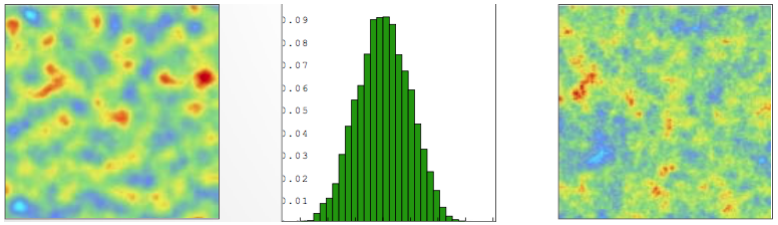


Figure: Dois Campos Aleatórios - CA distintos (paneis da esquerda e da direita), no entanto partilham o mesmo histograma (painel do centro).

Nota: Na terminologia inglesa um Campo Aleatório (CA) diz-se um *Random Field* (RF).

Dados espaciais versus séries temporais

Semelhanças:

- Ambos consideram um conjunto discreto de observações, nos tempos $t_1, \ldots, t_n$ de $Y(t)$ ou nas localizações $\mathbf{x_1}, \ldots, \mathbf{x_n}$ de $Y(\mathbf{x})$, para estimar a estrutura de autocorrelação subjacente ao processo[1]

Diferenças:

- Os dados nas séries temporais têm apenas uma direção, que é do passado para o tempo mais recente, podendo assim ser ordenados na reta do tempo

Aqui no espaço não temos uma ordem associada.

FAC

[1] Nos dados espaciais, em alternativa à função autocorrelação, recorre-se frequentemente à função variograma

Notação

- $Y(\mathbf{x})$ identifica o fenômeno em estudo (por exemplo, abundância de uma espécie ou chuva) que depende de coordenadas espaciais $\mathbf{x} \in D \subset \mathbb{R}^2$, onde D é uma região limitada

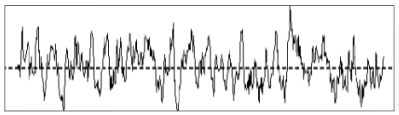
- Se temos as localizações espaciais $\mathbf{x_1}, ..., \mathbf{x_n}$, então $Y(\mathbf{x_1}), ..., Y(\mathbf{x_n})$ identifica dados observados nessas localizações. Essas observações podem ser obtidas a partir de uma ou mais variáveis discretas ou contínuas.

- $Y(\mathbf{x})$ é geralmente definido através da distribuição de dimensão finita

$$F_{\mathbf{x_1},...,\mathbf{x_n}}(y_1, ..., y_n) = P\{Y(\mathbf{x_1}) \leq y_1, ..., Y(\mathbf{x_n}) \leq y_n\}, \quad n \geq 1$$

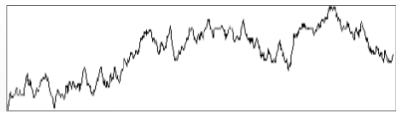
Hipóteses

- $Y(\mathbf{x})$ **estacionário**: invariância sob translações/rotações no espaço.
  Se for feita uma translação (vetor u) a distribuiçao conjunta matém-se.
- $Y(\mathbf{x})$ **estacionário de ordem 2**: os dois primeiros momentos existem e são invariantes sob translações/rotações. Média, variancia e covariancia têm de ser constantes, não podem depender de t
  Os momentos (média, variancia e covariancia) são os mesmos quando fazemos uma translação.
- $Y(\mathbf{x})$ **intrínseco**: os incrementos são estacionários de ordem 2.



Estacionário

Parece ser estacionário na média e na variancia



Intrínseco

Não parece ser estacionário na média

Estacionariedade

- $Y(\mathbf{x})$ é **estrita ou fortemente estacionário** se, para qualquer vector $\mathbf{u} \in \mathbb{R}^2$, temos
  - $F_{\mathbf{x_1},...,\mathbf{x_n}}(y_1, ..., y_n) = F_{\mathbf{x_1}+\mathbf{u},...,\mathbf{x_n}+\mathbf{u}}(y_1, ..., y_n), \quad \forall\ n, \mathbf{u}$

  o que significa que $Y(\mathbf{x})$ permanece invariante quando sujeito a transformações de translação de suas coordenadas

- $Y(\mathbf{x})$ é **estacionário de segunda-ordem** (ou fracamente estacionário), se
  - $\mathrm{E}[Y(\mathbf{x})] = \mu(\mathbf{x}) = \mu \quad \forall\ \mathbf{x} \in D$
  - $\mathrm{Cov}[Y(\mathbf{x_i}), Y(\mathbf{x_j})] = c(\mathbf{x_i} - \mathbf{x_j}), \quad \forall\ \mathbf{x_i}, \mathbf{x_j} \in D$

  $c(.)$ é chamada função de covariância estacionária e $\mu(\mathbf{x})$ é conhecido como tendência do processo

- $Y(\mathbf{x})$ é **intrinsicamente estacionário**, se [2]
  - $\mathrm{E}[Y(\mathbf{x})] = \mu(\mathbf{x}) = \mu \iff \mathrm{E}[Y(\mathbf{x_i}) - Y(\mathbf{x_j})] = 0$
  - $\mathrm{Var}[Y(\mathbf{x_i}) - Y(\mathbf{x_j})] = 2\gamma(\mathbf{x_i} - \mathbf{x_j}), \quad \forall\ \mathbf{x_i}, \mathbf{x_j} \in D$

  $\gamma(.)$ é chamado variograma

[2] Semelhante à série temporal, se Y_t não for estacionário, pode-se considerar as diferenças (de primeira ordem) entre os tempos

Hipóteses

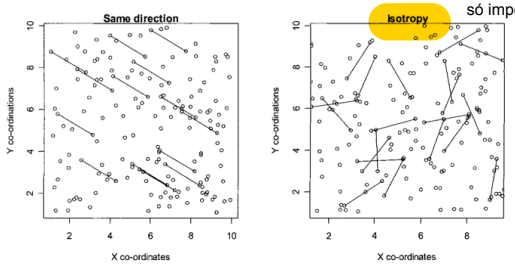
- Um processo $Y(\mathbf{x})$ estacionário também é intrínseco, mas um $Y(\mathbf{x})$ intrínseco nem sempre é estacionário.

- O foco não está na média... mas será que $Y(x)$ apresenta um comportamento sistemático?
  - Exemplo: Seja $Y(x)$ a profundidade do fundo do mar, então $E[Y(x)]$ aumenta regularmente a partir da praia.

- O modelo puramente **intrínseco situa-se entre os casos estacionários e não estacionários**. A escolha do grau de não estacionariedade depende do caso em estudo.

Duas outras propriedades importantes

- $Y(\mathbf{x})$ é **isotrópico**, se permanece invariante quando sujeito a rotações de coordenadas (oposto a anisotrópico). Por exemplo, se o processo aleatório intrínseco $Y(\mathbf{x})$ é isotrópico, então $\mathrm{Var}[Y(\mathbf{x_i}) - Y(\mathbf{x_j})] = 2\gamma(\|\mathbf{x_i} - \mathbf{x_j}\|)$ [3]



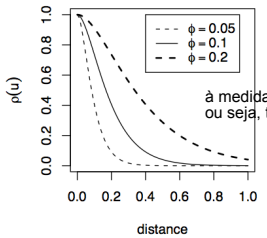
só importa a norma e não a direçao

- $Y(\mathbf{x})$ é **ergódico**, se a média de todas as realizações possíveis for igual à média de uma única realização (permite estimativas de parâmetros com apenas 1 realização)

[3] Como $\|.\|$ denota a norma Euclidiana, então γ apenas depende da distância entre as duas localizações e não da direção do vetor formado pelas duas localizações

As funções de **covariância** e **variograma**

- Se $\mathrm{Var}[Y(\mathbf{x})] = \sigma^2$, então pode-se escrever a função de covariância [4] como $c(u) = \sigma^2\rho(u;\phi)$, onde $\rho(.)$ **é a função de correlação** [5]



à medida que fi aumenta, temos correlação para maiores distancias, ou seja, temos correlação durante mais tempo

- Para $Y(\mathbf{x})$ sem "erro de medição", o variograma pode ser escrito como

$$\gamma(\mathbf{u}) = c(0) - c(u) = \sigma^2 - \sigma^2\rho(u;\phi) = \sigma^2(\mathbf{1} - \rho(\mathbf{u};\phi)) \qquad (1)$$

- O variograma $\gamma(u)$ mede a *desassociação* entre variáveis [6]

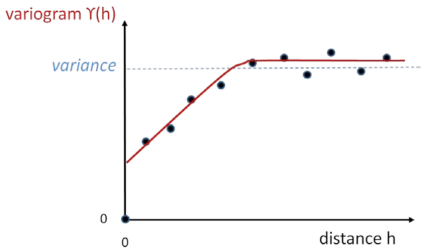
[4] **Assume-se que $Y(\mathbf{x})$ é um processo estacionário e isotrópico de 2ª ordem.**
[5] **Parâmetro ϕ relacionado com a distância além da qual a correlação entre as variáveis é 0 ("raio de influência" ou *range*)**
[6] **Oposto à função de covariância $c(u)$, que mede a *associação* entre variáveis**

Vantagens do variograma (Comparado com a FAC)

- O variograma adapta-se mais facilmente a observações não estacionárias (ex. $\mu(\mathbf{x})$ não constante)
- Para estimar o variograma, nenhuma estimativa de μ é necessária
- A estimação do variograma é mais simples que a estimação da função de covariância

Análise estrutural via o variograma

- A análise estrutural tem como objetivo capturar, descrever e **modelar a maneira como uma variável geo-referenciada é estruturada espacialmente**.
- O variograma mede a **variabilidade média entre dois pontos quaisquer** como função do vetor de distância entre esses pontos.
  1. Primeiro, calculamos o variograma experimentalmente.
  2. Em seguida, ajustamos o variograma e modelamos a variável de interesse.



as coisas mais proximas estao mais associadas, logo as mais perto tb estao menos desassociadas

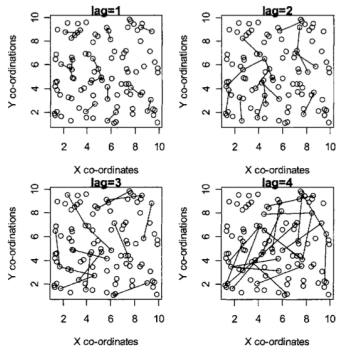
Variograma experimental $\widehat{\gamma}(u)$

$$\gamma(\|\mathbf{x_i}-\mathbf{x_j}\|)=\frac{1}{2}\mathrm{Var}\left[Y(\mathbf{x_i})-Y(\mathbf{x_j})\right]=\frac{1}{2}\mathrm{E}\left[(Y(\mathbf{x_i})-Y(\mathbf{x_j}))^2\right]$$

$\hat{\gamma}(u)$ pode ser obtido aproximando $E[.]$ por uma média amostral

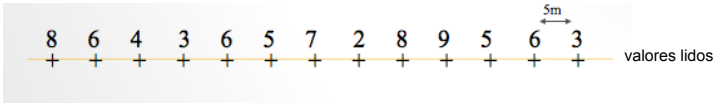
$$\widehat{\gamma}(u)=\frac{1}{2|N(u)|}\sum_{N(u)}\left(Y(\mathbf{x_i})-Y(\mathbf{x_j})\right)^2 \qquad (2)$$

onde $N(u)=\{(\mathbf{x_i},\mathbf{x_j}):\|\mathbf{x_i}-\mathbf{x_j}\|\approx u,\ u\in\mathbb{R}\}$ e $|N(u)|$="n. de pares em N(u)".



Exercício: variograma para uma amostra regular 1-D

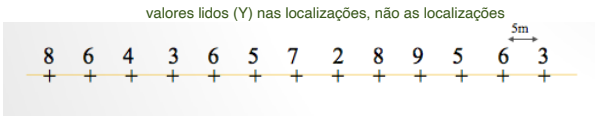
Variável Z definida numa grelha regular em 1-D



- Calcule o variograma experimental para os defasamentos: 5m, 10m e 15m.
- Avalie o variograma experimental da nova variável:

$$Y(x)=Z(x)+3.2$$

Solução: variograma para uma amostra regular 1-D



- Considere a distância de 5m:

$$\hat{\gamma}(5m)=\frac{1}{2\times 12}[(8-6)^2+(6-4)^2+(4-3)^2+\ldots(6-3)^2]=4.625$$

nº de pares distantes 5 km

- Considere a distância de 10m:

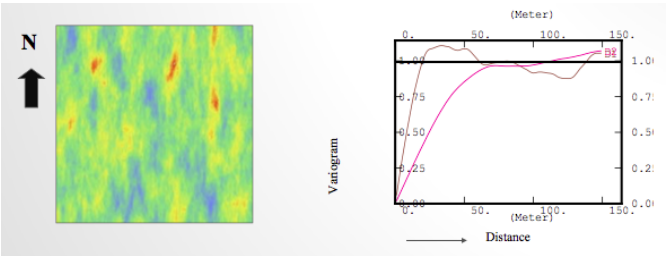
$$\hat{\gamma}(10m)=\frac{1}{2\times 11}[(8-4)^2+(6-3)^2+(4-6)^2+\ldots(5-3)^2]=5.227$$

- Considere a distância de 15m:

$$\hat{\gamma}(15m)=\frac{1}{2\times 10}[(8-3)^2+(6-6)^2+(4-5)^2+\ldots(9-3)^2]=6.000$$

Quanto **maior a distância**, menor o número de pares na linha, e **menos** $\hat{\gamma}(.)$ **é representativo da variabilidade de Z (e de Y)**.

Exemplo – falha de isotropia



- Porque não devemos assumir isotropia, mas sim anisotropia?
  - Cálculo de 2 variogramas, um para a direção E-O e outro para N-S, origina resultados distintos (painel direita)
- Cada variograma direcional usa apenas pares ao longo da direção considerada, com uma **tolerância na direção**.

Variograma direcional versus ou omnidirecional

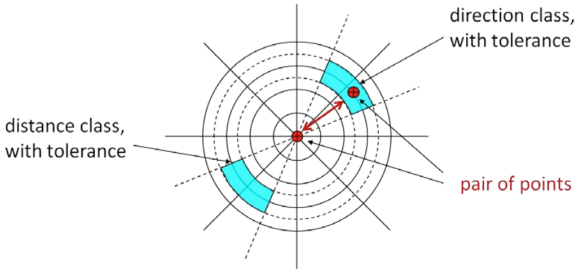


Figure: Cálculo do variograma com **tolerâncias na distância e direção**: cada par de pontos de dados é alocado a uma "categoria" de distância e direção, levando em consideração as tolerâncias.

Por que $\hat{\gamma}(u)$ não deve ser usado para inferência e predição?

O variograma experimental pode não ser definido-negativo condicional, o que **pode levar a valores negativos absurdos para o erro de predição quadrático médio** (considerado um estimador inválido).

Como obter um estimador de variograma válido?

Uma abordagem comum é aproximar $\hat{\gamma}(u)$ por algum modelo teórico $\gamma(u;\theta)$, conhecido por ser válido, capturando a dependência espacial subjacente aos dados disponíveis.

- **Primeiro**, escolhe-se um modelo teórico (por exemplo, exponencial, esférico, ...), normalmente usando ferramentas gráficas, que depende dos parâmetros θ
- **Depois**, recorre-se a:
  - um critério de ajuste clássico (por ex. mínimos quadrados) para completar a especificação de γ final, estimando os parâmetros θ, tal como se segue

$$\hat{\theta} = min_\theta \left\{ \sum_i w_i (\hat{\gamma}(u_i) - \gamma(u_i;\theta))^2 \right\}$$

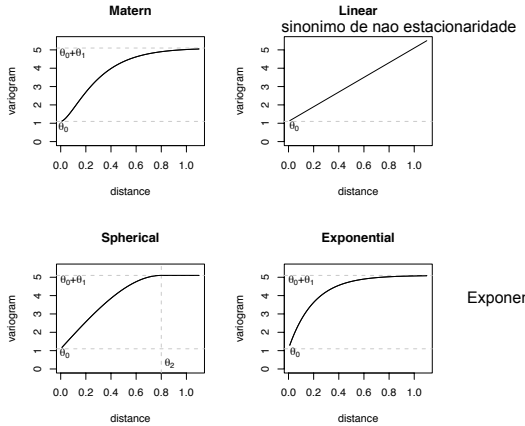
(variograma empirico)

  - ou uma abordagem baseada em métodos de máxima verosimilhança

isotropico teorico

Modelos de variograma isotrópico $\gamma(u;\theta)$

Assumindo-se que $Y(\mathbf{x})$ tem associado um erro de medição

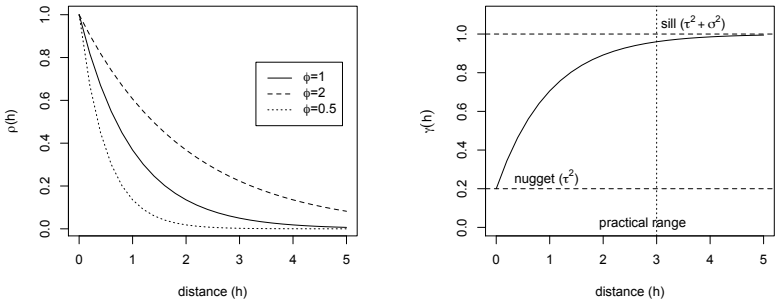


Exponencial é um caso particular da Matern, qdº k=0.5

- $\theta_0 = \tau^2$ é conhecido como variância do erro (*nugget*)
- $\theta_0 + \theta_1 = \tau^2 + \sigma^2$ conhecido como variância total de $Y(\mathbf{x})$ (*sill*) "teto" onde estabiliza
- $\theta_2 = \phi$ conhecido como raio de influência (*range*)

A função de correlação e o variograma

Processo isotrópico e estacionário de segunda ordem

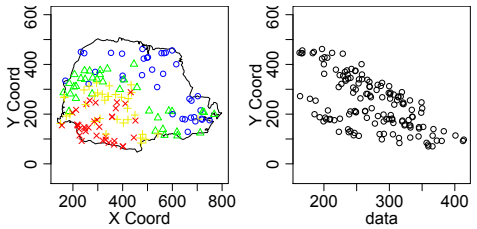


Painel da esquerda: Função de correlação exponencial para diferentes valores de *range* ϕ

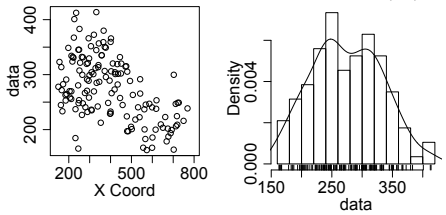
Painel da direita: Representação esquemática de um variograma típico com seus parâmetros estruturais, sendo $\gamma(u) = \tau^2 + \sigma^2(1 - \rho(u;\phi)) = \frac{1}{2}\mathrm{Var}[Y(\mathbf{x}) - Y(\mathbf{x}')]$ e $u = \|\mathbf{x} - \mathbf{x}'\|$

Exemplo 1: Dados de precipitação, Estado do Paraná, Brasil

Em muitas aplicações práticas, é útil considerar um processo Gaussiano espacial com uma função média ou tendência **possivelmente dependendo da localização, i.e.** $\mu(\mathbf{x})$, mas com uma estrutura de covariância estacionária

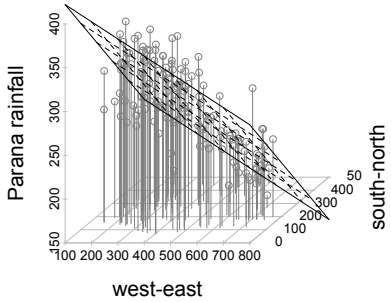


Os vermelhos são acima do Q3 (quartil) os 25% mais altos e os azuis são os 25% mais baixos (Q1)



Exemplo 1: Dados de precipitação - função média $\mu(\mathbf{x})$

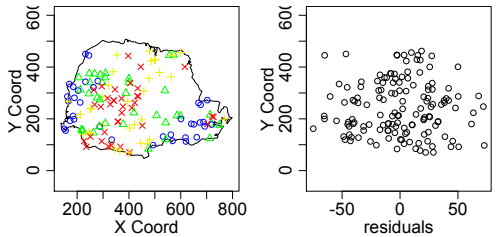
Uma solução possível é especificar $\mu(\mathbf{x})$ como um modelo de regressão, dependendo das próprias coordenadas Fazemos isto para tentar estacionarizar a media



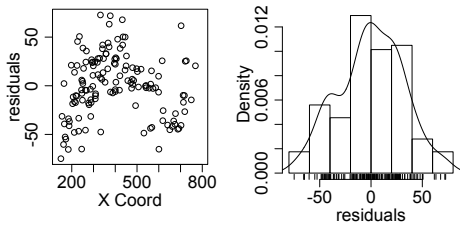
Na presença de variáveis explicativas geo-referenciadas relevantes, é razoável incluí-las em $\mu(\mathbf{x})$. Por exemplo, suponhamos que o objetivo é analisar as temperaturas médias diárias em toda a Argentina, então poderá ser importante considerar a altitude como uma covariável

Exemplo 1: Dados de precipitação - removendo $\mu(\mathbf{x})$

$Y(\mathbf{x}) - \mu(\mathbf{x})$ torna-se um processo estacionário Gaussiano de média zero



Neste caso já nao ha padrao o que é bom



Diferentes escalas de variabilidade

Pode-se escrever o processo aleatório espacial como

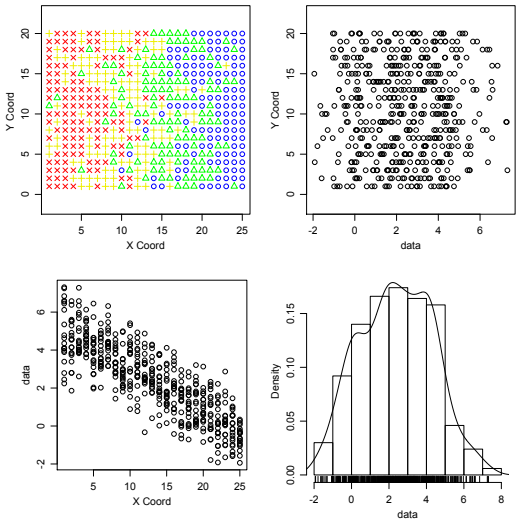
$$Y(\mathbf{x}) = \mu(\mathbf{x}) + S(\mathbf{x}) + \epsilon(\mathbf{x}) = \alpha + \sum_{j=1}^{p} \beta_j X_j(\mathbf{x}) + S(\mathbf{x}) + \epsilon(\mathbf{x})$$

- $\mu(\mathbf{x})$ representa uma tendência determinística, e identifica uma variabilidade de grande-escala. $\mu(\mathbf{x})$ pode envolver variáveis explicativas X_j, eventualmente dependentes da localização $\mathbf{x}$
- $S(\mathbf{x})$ é um processo aleatório com média zero e $\mathrm{Var}[S(\mathbf{x})] = \sigma^2$, que contem a **estrutura de dependência espacial**, com $\mathrm{Corr}[S(\mathbf{x}), S(\mathbf{x}')] = \rho(\|\mathbf{x} - \mathbf{x}'\|; \phi)$, identificando uma variabilidade de pequena-escala
- $\epsilon(\mathbf{x}) \sim N(0, \tau^2)$ é um erro de medição, i.i.d. e independente de $S(\mathbf{x})$[7]

[7] A variância do erro de medição deve ser adicionada ao variograma de $Y(\mathbf{x})$ na equação (1), ficando $\gamma(u) = \tau^2 + \sigma^2(1 - \rho(u; \phi))$.

Exemplo 2: Dados de grãos de trigo (wheat data)

Forte tendência oeste-este (90^0)



Exemplo 2: Trigo – detectando e removendo a tendência

$$\textbf{Grão(x)} = \alpha + \beta_1 \textbf{Longitude(x)} + \beta_2 \textbf{Latitude(x)} + \textbf{S(x)} + \epsilon\textbf{(x)}$$

	estimativa	s.e.	p-valor
α	5.283	0.129	<0.001
β_1	-0.226	0.006	<0.001
β_2	0.004	0.008	0.599