

# Introdução à análise de dados espaciais

Raquel Menezes da Mota Leite

Universidade do Minho

Setembro de 2023



## Análise exploratória de dados

- A análise exploratória de dados (AED) é uma parte integrante da Estatística no âmbito das aplicações, que **deve preceder qualquer abordagem de modelação**. A Estatística Espacial não é exceção.
- No nosso caso, a AED é naturalmente orientada para a investigação preliminar dos **aspectos espaciais** dos dados que são relevantes para nossas suposições de modelação.
- No entanto, **aspectos não espaciais** também podem e **devem ser investigados**.

## Conceito fundamental de “Dependência Espacial”

*Everything is related to everything else, but near things are more related than distant things (first law of geography)*

Waldo Tobler, 1970

## AED e seus principais objetivos

Devemos considerar:

- **Sumários numéricos**, conhecidos como estatísticas descritivas (médias, medianas, quantis, variância, ...)
- **Sumários gráficos** relacionados com a análise preliminar, por meio de visualização de dados

Conheça seus dados!

- distribuições (simétricas, normais, assimétricas)?
- problemas de qualidade dos dados?
- valores discrepantes ou extremos?
- correlações e inter-relações?
- subconjuntos de interesse?
- sugestões de relações funcionais?

## Visualização de dados

Os seres humanos são os melhores identificadores de padrões, portanto, a análise de sumários gráficos pode ser bastante produtiva.

Os **métodos visuais** a serem escolhidos dependerão se temos

- uma, duas ou mais variáveis?
- variáveis categóricas ou quantitativas?
- referência geográfica ou referência temporal?

Os métodos podem incluir

- boxplots ou histogramas
- gráficos de dispersão ou gráficos de barras
- **mapeamento de dados espaciais observados numa região**
- gráfico de uma série temporal ao longo do período observado

5 / 26

## Dados Espaciais

Na estatística espacial, a nossa variável de interesse  $Y(x)$  é definida sobre uma região espacial  $x \in A$ , e existem observações em locais específicos  $x_1, \dots, x_n$ .

Dependendo da **natureza dos dados** e da **agregação espacial** que lhes damos, podemos diferenciar três tipos de dados espaciais:

- 1 Dados referentes a pontos, conhecidos como **dados geoestatísticos**
- 2 Dados referentes a áreas, conhecidos como **dados agregados**
- 3 Dados referentes a processos pontuais

7 / 26

## Estatística Espacial

O termo **Estatística Espacial** é usado para descrever uma ampla gama de **modelos e métodos estatísticos**, destinados à análise de **dados referenciados espacialmente** (Diggle & Ribeiro, 2007).

Na prática, em diversas áreas, como epidemiologia, climatologia, ecologia e ciências ambientais, muitas vezes torna-se necessário analisar dados que são:

- altamente multivariados, com muitos preditores importantes e variáveis resposta
- **referenciados geograficamente** e frequentemente apresentados em forma de mapas
- temporalmente correlacionados, como em estruturas de séries temporais longitudinais ou outras

6 / 26

### 1. Dados referentes a pontos, i.e. **dados geoestatísticos**

**Dados referentes a pontos** (de uma superfície) são constituídos por uma ~~variável aleatória~~ **variável aleatória**  $Y(x)$  **recolhida num conjunto fixo de locais  $x$  sobre um campo espacial contínuo  $S(x)$ .**

- O espaço é tipicamente tratado como **bidimensional**, definido por sua **longitude** e **latitude**, mas também pode incluir altitude ou profundidade para torná-lo tridimensional.
- Exemplos de processos geoestatísticos:
  - ▶  $S(x)$  representa a **superfície de poluição** sobre a cidade de Lisboa, e  $Y(x)$  representa um indicador de poluição medido na estação de monitorização do local  $x$ .
  - ▶ Na pesca,  $S(x)$  pode representar a **superfície de abundância** de uma espécie e  $Y(x)$  a captura desse peixe no local  $x$ .

8 / 26

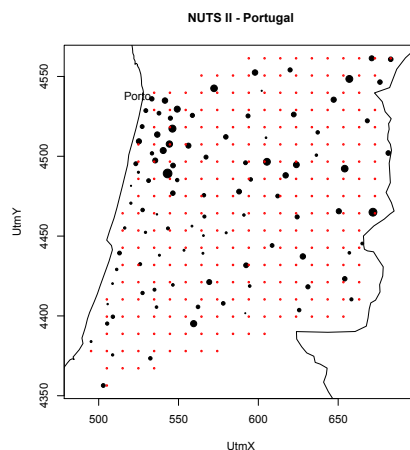
## Dados geoestatísticos

- Eles podem ser referidos como **dados espacialmente contínuos**.
- O termo **contínuo** não **significa** que a variável de interesse seja contínua, mas sim que **pode ser medida em qualquer local na região de estudo**.
- Tais variáveis distribuídas continuamente eram tradicionalmente usadas nas geociências para a análise de concentração de minérios, o que explica o termo **Geoestatística**.
- Hoje em dia, elas são amplamente utilizadas em distintos contextos, desde que a localização geográfica seja usada explicitamente na análise dos dados.  
Exemplos: temperatura da superfície do mar, ou salinidade, ou alguma medida de abundância de peixes, como a concentração de ovos.

9 / 26

## Exemplo de dados geoestatísticos

**Dados de poluição por arsênico  $Y(x)$  com  $x \in A \subset \mathbb{R}^2$**  (Garcia-Soidán e Menezes, 2017).



Nota: Círculos pretos identificam as 98 medições de As, sendo o diâmetro do círculo proporcional ao valor observado. Pontos vermelhos identificam grelha de pontos onde se pretende estimar As.

11 / 26

## Modelação de dados geoestatísticos

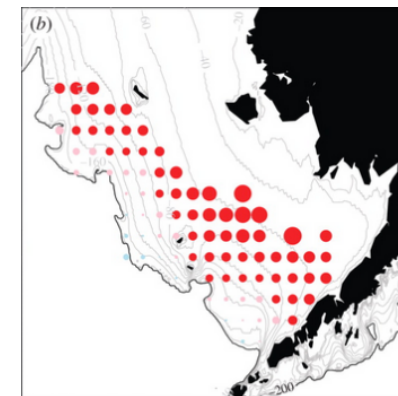
- Ao modelar a nossa variável de interesse, condicionada a eventuais variáveis explicativas, é expectável que os respetivos *resíduos* sejam espacialmente correlacionados.
- O nosso principal objetivo é **inferir a estrutura espacial** subjacente aos nossos dados para melhorar a previsão, usando (por exemplo) técnicas de **krigagem** em locais não amostrados.

Obviamente, em todas as três sub-áreas da Estatística Espacial, a **dimensão espacial pode ser estendida para o domínio espaço-temporal** adicionando a correlação da variável de interesse entre eventos temporais (por exemplo, estudar a abundância de peixes a cada hora, dia, etc).

10 / 26

## Exemplo de dados geoestatísticos

Ciannelli et al. (2012) modelaram a **distribuição de peixes adultos** no Mar de Bering. Medidas  $Y(x)$  são feitas em locais discretos  $x$  do domínio contínuo  $A$ .



12 / 26

## 2. Dados referentes a áreas ou agregados

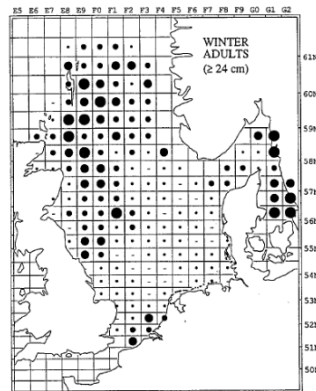
- **Dados referentes a áreas**, ou apenas *dados de área*<sup>1</sup>, representam uma **agregação de observações** sobre uma *unidade de área* predefinida.
- O resultado dessa agregação  $Y(x)$  é definido sobre uma **região discreta**  $A$  com um **número fixo de locais**  $x$ , que poderão identificar os **centróides** das unidades de área.
- Portanto, a região  $A$  é dividida numa **coleção finita de unidades de área** com limites bem definidos.

Na modelação de dados agregados, deve-se ter em conta **se as regiões adjacentes têm semelhanças**, no sentido de que é expectável que áreas próximas tenham mais em comum do que áreas distantes.

<sup>1</sup>Na terminologia inglesa, este tipo de dados são denominados *lattice data* ou *areal data*.

## Exemplo de dados agregados

No âmbito das pescas, o Mar do Norte é discretizado num domínio  $A$ , de acordo com grelha definida pelo ICES<sup>2</sup> (Paradinas, 2017). As observações são então agregadas em cada retângulo da grelha, sendo denotadas por  $Y(x)$ .



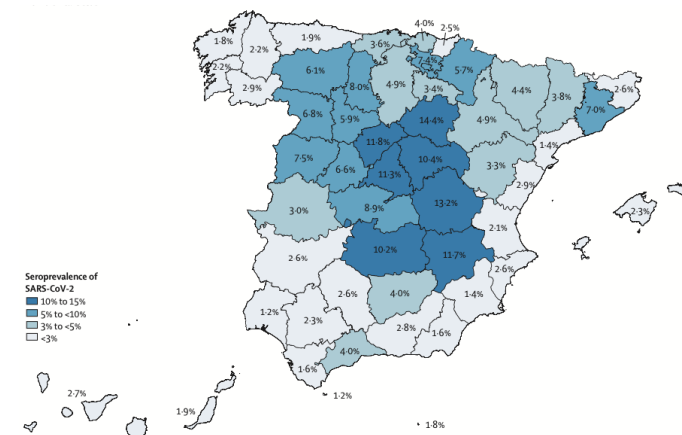
Distribuição, agregada pelos retângulos estatísticos do ICES, do arenque adulto.

<sup>2</sup>International Council for the Exploration of the Sea

## Modelação de dados agregados

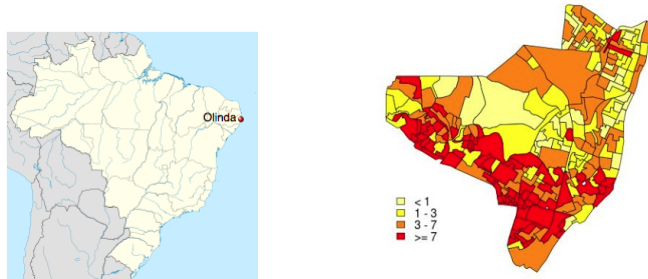
- Modelar dados de área envolve a obtenção de informações de regiões adjacentes.
- A estrutura de modelo mais comum nesses casos é o **modelo condicional autoregressivo** (Besag et al., 1991), mais conhecido como modelo **CAR** ou **BYM** pelas iniciais dos autores.
- Esses modelos consideram correlação espacial autorregressiva por meio de uma estrutura de adjacência das unidades de área.

## Exemplo de dados agregados



Prevalência de SARS-CoV-2 em Espanha, inquérito sorológico de 27 de abril a 11 de maio de 2020, envolvendo 61 075 participantes (Pollán et al., 2020)

## Vigilância de lepra, Olinda NE Brasil (Bailey, 2008)



- Olinda é um município do Estado de Pernambuco, composto (no censo de 1991) por **241 setores** com aproximadamente 350.000 habitantes.
- Dados disponíveis sobre a **incidência de novos casos de lepra por setor no período 1991-1995** (total de 1135 casos), juntamente com as estimativas de população para o período médio (1993) nesses setores.
- Um **indicador simples de privação** também disponível - proporção de chefes de família com renda mensal abaixo de um salário mínimo legal (US\$80).

17 / 26

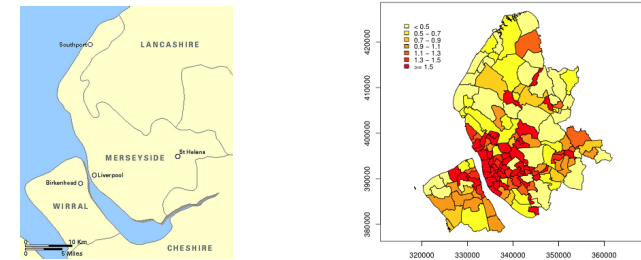
## 3. Dados referentes a Processos Pontuais

- A teoria dos **processos pontuais** surgiu com a **necessidade de modelar como aleatória a localização** de eventos de interesse.
- As primeiras aplicações são da área da ecologia (e.g. locais onde se avista uma ave rara) e ciências florestais (e.g. localização de um tipo de árvores).
- No entanto, o seu leque de aplicações é muito vasto. Podemos considerar a distribuição de embarcações no mar ou a distribuição dos casos conhecidos de uma determinada doença contagiosa.

Nestes vários contextos, o objetivo é **estudar o arranjo espacial do evento de interesse no espaço**, para identificar zonas importantes (e.g. propícias à pesca ou de perigo).

19 / 26

## Cancro da laringe, Noroeste da Inglaterra (Bailey, 2008)

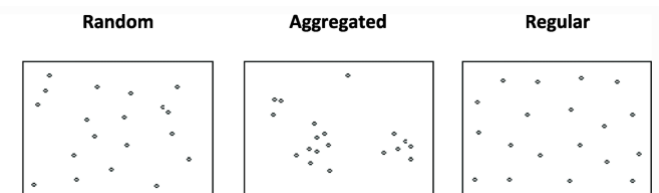


- Dados sobre **876 casos de cancro da laringe** diagnosticados em 1982-1991, em **144 distritos eleitorais** de Mersey e West Lancashire.
  - Disponíveis **números esperados de casos**, com base em taxas de referência nacionais por idade e sexo, e valores da população do censos de 1991.
  - Trabalho de investigação permitiu definir um **indicador da prevalência do tabagismo** em cada distrito ('baixa', 'média' ou 'alta').
  - Disponível **medida anual da poluição do ar** com base no fluxo de tráfego.
- $Y(i) = \text{n}^\circ \text{ de casos no distrito } i, \text{ onde } i=1, \dots, 144$   
 $Z1(i) = \text{indicador de prevalência de tabagismo no distrito } i$   
 $Z2(i) = \text{poluição do ar}$   
 $E_{jk}(i) = \text{n}^\circ \text{ esperado de casos no distrito } i \text{ para o género } j \text{ (M/F) e para a faixa etária } k$

18 / 26

## Modelação de processos pontuais

- Define-se como hipótese de base a **aleatoriedade espacial completa**, i.e. o padrão espacial não apresenta nenhuma estrutura aparente (painel I).
  - Assume-se que o n. de eventos numa região segue **distribuição de Poisson** com média proporcional à área da região e ao n. médio de acontecimentos por unidade de área – **intensidade do processo**.



- Alternativamente, podemos ter um padrão que corresponde a eventos fortemente agregados (painel II), ou um padrão regular (painel III), caso se imponha uma distância mínima entre eventos.

20 / 26

## Modelação de processos pontuais

- A resposta  $Y(x)$  é fixa (1=presença) e os locais  $x$  são gerados aleatoriamente a partir do campo aleatório espacial  $\Lambda$ .
- Temos um **processo pontual marcado**, se associarmos alguma informação ao ponto  $x$  (e.g. se o processo pontual é definido pela ocorrência de uma doença, podemos associar o género do indivíduo).
- Precisamos estudar a **estrutura espacial subjacente**, usando as propriedades topológicas, geométricas ou geográficas dos locais observados.

21 / 26

## Exemplos de dados referentes a processos pontuais

Poderemos considerar um processo pontual para estudar a **distribuição de operações de pesca** no oceano.



Figure: Vista aérea da pesca comercial de arenque no Sitka Sound, Alasca.

$\Lambda$  representa a distribuição de pesca no oceano e  $x$  as localizações onde as embarcações estiveram a pescar num determinado momento.

23 / 26

## Possíveis questões científicas

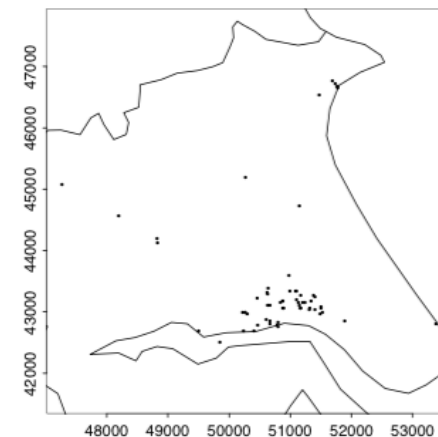
Algumas questões mais relevantes para os dados referentes a processos pontuais:

- A distribuição espacial dos pontos observados é homogênea no espaço?
- Ou temos um processo de agregação?
- E se existir agregação, quais as razões que a poderão justificar ?

22 / 26

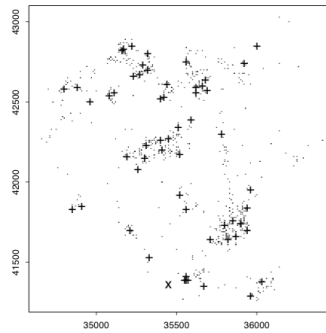
## Leucemia infantil (Cuzick e Edwards, 1990)

Localizações das residências de todos os casos conhecidos de leucemia infantil em Humberside, Inglaterra, durante o período de 1974 a 1982.







24 / 26

## Cancro de pulmão e laringe (Diggle, Gatrell and Lovett, 1990)



O mapa mostra:

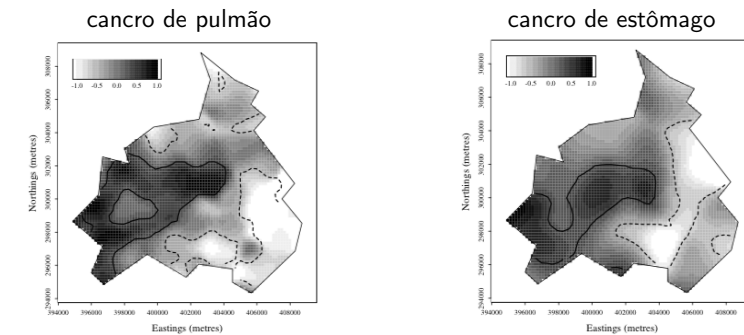
- casos de cancro de pulmão (pontos)
- casos de cancro de laringe (cruzes pequenas)  
- uma incineradora industrial agora desativada (cruz grande  )

### Questões importantes:

- Os casos apresentam uma tendência *surpreendente* de se agruparem?
- O risco de doença varia espacialmente?
- O risco de doença está elevado perto de um local específico?

25 / 26

## Superfícies de risco estimadas (com base na suavização kernel)



Alguns comentários:

- padrão de variação semelhante em ambas as doenças
- linhas sólidas e tracejadas identificam os limites das regiões onde o risco é significativamente mais alto ou mais baixo, respectivamente, do que a média

26 / 26