Departamento de Matemática

Mestrado em Estatística para a Ciência de Dados

Estatística Espacial

Raquel Menezes

Novembro 2024

1 Objetivos genéricos

Este trabalho de Estatística Espacial tem duas componentes de avaliação, associadas à aplicação de dois tipos de **modelos lineares espaciais**: (1) modelos geoestatísticos e (2) modelos referentes a áreas.

Por conseguinte, o aluno (ou o grupo composto por 2 alunos) terá que seleccionar duas bases de dados, tendo em conta que:

(1) **Dados geoestatísticos** são definidos por um processo estocástico $Y(x)$ associado a um conjunto fixo de locais x sobre um campo espacial contínuo $\{S(x) : x \in D \subset I\!R^2\}$. O espaço é tipicamente tratado como **bidimensional**, definido por sua **longitude** e **latitude**

(2) **Dados referentes a áreas**, ou apenas dados de área[1], representam uma **agregação de observações** sobre uma *unidade de área* predefinida (por exemplo, distrito ou freguesia). O resultado dessa agregação $Y(x)$ é definido sobre uma **região discreta** D com um **número fixo de locais** x, que poderão identificar os **centróides** das unidades de área.

Pretende-se comprovar que o aluno/grupo é capaz de realizar a sequência de análise habitual sobre estes dois tipos de dados, recorrendo a ferramentas informáticas adequadas. O aluno/grupo deverá ser capaz de avaliar críticamente os resultados obtidos. Também se verificará a capacidade de utilizar diversas funções de bibliotecas do ambiente R, disponíveis para a análise e modelação de dados espaciais, nomeadamente *geoR*, *spmodel* ou *spdep*.

2 Trabalho proposto

Este trabalho consiste na **análise e modelação** de dados observados numa região, assumindo-se que se tratam de realizações de um processo estocástico espacial $Y(\mathbf{x})$.

PASSO I - O trabalho deverá começar com a apresentação dos dados e uma análise exploratória espacial e não-espacial dos mesmos, incluindo a descrição das principais estatísticas descritivas e principais representações gráficas.

PASSO II -A modelação deverá ser realizada recorrendo-se à regressão linear para dados espacialmente correlacionados, que dependerá se estamos perante dados geoestatísticos ou referentes a áreas. Para além da inferência sobre os parâmetros do modelo, o trabalho a desenvolver deverá considerar **predição espacial**.

2.1 Modelos geoestatísticos

Os tópicos teóricos a cobrir nesta primeira parte do trabalho são:

1. Estimação da tendência espacial
2. Estimação do variograma empírico e teórico
3. Interpolação espacial
4. Diagnóstico do modelo, validação-cruzada

[1] Na terminologia inglesa, este tipo de dados são denominados *lattice data* ou *areal data*.

O aluno/grupo terá de propor uma base de **dados geoestatísticos**, eventualmente procurando em ***sites* públicos da intranet** ou seleccionando dados disponíveis em alguma biblioteca do R, nomeadamente:

- Biblioteca *geoR*
  - *data.frame* "camg" com medições de magnésio em amostras de solo (semelhante ao exemplo apresentado na aula sobre o cálcio, "ca20");
  - *data.frame* "soil250" com conjunto de dados de propriedades químicas do solo;
  - *data.frame* "soja98" com informação sobre produção de soja e outras variáveis em um ensaio de uniformidade;
  - classe *geodata* "wolfcamp" com medições do nível piezométrico feitas no Aquífero Wolfcamp, Texas, EUA.
- Bibliotecas *sp*, *RandomFields*, *spData*, *gstat* ou *spmodel*
  - *data.frame* "meuse" com dados sobre localizações e concentrações de metais pesados da camada superior do solo, recolhidos no rio Meuse.
  - *data.frame* "weather" com erros de previsão de pressão e temperatura no noroeste do Pacífico.
  - ...

2.2 Modelos referentes a áreas

Os tópicos teóricos a cobrir na segunda parte do trabalho são:

- Testes de associação espacial, estatísticas I de Moran e c de Geary
- Modelos auto-regressivos condicionais (CAR) ou simultâneos (SAR)

O aluno/grupo terá de propor uma base de **dados referentes a áreas**, eventualmente procurando em ***sites* públicos da intranet** ou seleccionado dados disponíveis em alguma biblioteca do R, nomeadamente:

- Bibliotecas *spData*, *spdep* ou *spmodel*
  - *data.frame* "auckland" com dados sobre mortalidade infantil em Auckland.
  - *data.frame* "elect80" com dados para os resultados das eleições presidenciais de 1980, cobrindo 3.107 condados dos EUA.
  - *data.frame* "nydata" com dados de leucemia em Nova Iorque.
  - ...

3 Apresentação do trabalho

Sugere-se a entrega de um relatório em formato PDF com os resultados da análise e modelação de dados, onde se incluam todos os comentários e conclusões que considere oportunos para a compreensão do trabalho desenvolvido. Adicionalmente, o aluno/grupo deverá preparar um ficheiro com o código R, utilizado para a análise e modelação de dados.

3.1 Prazos importantes

Até 9 de dezembro, o aluno/grupo deverá enviar informação sobre as 2 bases de dados selecionadas, juntamente com os resultados preliminares da análise exploratória (PASSO I), por *e-mail* para `rmenezes@math.uminho.pt`.

O **trabalho final** (PASSO II) deverá ser enviado também por correio electrónico até 5 de janeiro. A apresentação oral dos trabalhos irá decorrer no dia 9 de janeiro.