



# Análise e Transformação de Dados

## Ficha Prática nº 3 – Parte A

Objetivo: Pretende-se iniciar a análise de séries temporais, efetuando o seu pré-processamento e a sua decomposição em componentes que traduzem os movimentos estruturais e erráticos.

Linguagem de Programação:

- MATLAB

- Python (módulos: *numpy*, *random*, *scipy.interpolate*, *scipy.signal*, *matplotlib.pyplot*)

Exercícios:

1. Uma série temporal é uma sequência temporalmente ordenada de dados. O estudo estatístico de Séries Temporais envolve, em geral, dois aspetos: a) Análise e Modelação da Série Temporal – para descrever a série, verificar as suas características mais relevantes e investigar as possíveis relações com outras séries; b) Previsão da Série Temporal – determinar boas previsões de valores futuros da série, num dado horizonte de previsão, a partir de valores passados da série.

Antes de iniciar a análise de uma série temporal deve-se proceder à sua preparação através do pré-processamento dos dados que envolve, normalmente, as seguintes operações:

- Detecção e regularização do espaçamento dos dados, envolvendo a deteção de dados em falta (por exemplo, identificados pelo valor NaN) e a sua substituição por um valor estimado usando, por exemplo, um método de interpolação ou de extrapolação;
- Detecção e regularização de valores atípicos (*outliers*), envolvendo a sua deteção considerando, por exemplo, o critério  $|x_i - \mu| > 3\sigma$ , sendo  $x_i$  o valor da série no índice  $i$ ,  $\mu$  a média da série e  $\sigma$  o desvio padrão da série, e a sua substituição por um valor adequado. Dependendo do *outlier* ser aditivo ou subtrativo, o valor a usar poderá ser, por exemplo,  $x_i = \mu + 2.5\sigma$  no caso aditivo e  $x_i = \mu - 2.5\sigma$  no caso subtrativo.

De referir que o pré-processamento dos dados é muito importante porque a existência de dados em falta e/ou de *outliers* pode comprometer os procedimentos de análise e de modelação da série temporal, podendo, nomeadamente, induzir uma identificação incorreta do modelo e uma estimação enviesada dos seus parâmetros.

Neste trabalho, pretende-se tratar a série temporal que representa a evolução da temperatura média numa dada localização. No caso, considera-se o *dataset* com os dados de Lisboa de 1980 a 2018. Cada amostra da série corresponde a 1 mês, sendo a primeira referente a janeiro de 1980.

- 1.1 Ler e representar graficamente a série temporal existente no ficheiro de dados “lisbon\_temp\_fmt” (temperaturas com espaçamento temporal em meses).
- 1.2 Verificar a existência de valores não recolhidos/medidos, identificados com NaN (*Not a Number*). Identifique-os, e substitua cada um desses valores por valores que resultam de um processo de extrapolação e represente graficamente a série temporal modificada, comparando-a com a inicial.  
Sugestão:
  - Reconstruir os valores em falta usando uma extrapolação de 3ª ordem, como o método ‘*pchip*’ (**interp1** (MATLAB) | **interp1d** / **PchipInterpolator** (Python)).
- 1.3 Determinar os valores da média (**mean**) e do desvio padrão (**std**) da série temporal. Determinar a correlação (**corrcoef**) entre a temperatura na década de 80 e a temperatura na década de 90. Comentar os resultados.
- 1.4 Verificar a existência de *outliers*. Identifique-os, substitua-os por valores adequados e represente graficamente a série temporal modificada, comparando-a com as anteriores.

2. A análise da série temporal considera, habitualmente, a existência de componentes associadas a movimentos estruturais e a movimentos erráticos: a) tendência (ou tendência-ciclo, quando agrupada com a componente cíclica) – movimento subjacente de longo-prazo que caracteriza a evolução do nível médio da série; b) sazonal – movimentos estritamente periódicos, decorrentes de características ou fatores que influenciam a evolução da série; c) cíclica – movimentos oscilatórios de tipo recorrente; d) errática/irregular – movimentos aleatórios decorrentes de uma multiplicidade de fatores e de natureza imprevisível. Estas quatro componentes podem ser combinadas de forma multiplicativa ou aditiva (forma a considerar neste exercício).

- 2.1 Estimar a componente da tendência para a série temporal que resulta do exercício 1, considerando aproximações polinomiais de grau 0 e 1 e usando a função **detrend**. Calcular a série sem a tendência. Representar graficamente a série temporal em bruto, a componente da tendência e a série temporal sem a tendência de grau 0 e de grau 1.
- 2.2 Estimar a tendência quadrática considerando uma aproximação polinomial de grau 2, usando as funções **polyfit** e **polyval**. Representar graficamente a série temporal em bruto, a componente da tendência e a série temporal sem a tendência quadrática.
- 2.3 Estimar a componente da sazonalidade da série temporal (sem a componente da tendência), assumindo uma sazonalidade anual. Representar graficamente a série temporal em bruto, sem as componentes da tendência e da sazonalidade.
- 2.4 Obter a componente irregular. Representar graficamente a série temporal regularizada, sem a componente irregular e a respetiva componente irregular.