



# Análise e Transformação de Dados

## Ficha Prática nº 3

Objetivo: Pretende-se iniciar a análise de séries temporais, efetuando o seu pré-processamento.

Exercício:

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 detecção 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 detecção 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.

- 1.1 Ler e representar graficamente as duas séries temporais existentes no ficheiro de dados “seriestemp.dat” (ficheiro ASCII com espaçamento temporal em horas).
- 1.2 Verificar a existência de valores não recolhidos/medidos, identificados com NaN (*Not a Number*). Identifique-os, elimine cada um desses valores da respetiva série temporal, substitua-os por valores que resultam de um processo de extrapolação e represente graficamente as séries temporais modificadas, comparando com as anteriores.

Sugestão:

- Reconstruir os valores em falta usando extrapolação com o método ‘*pchip*’ (**interp1**).
- 1.3 Determinar os valores da média (**mean**) e do desvio padrão (**std**) de cada série temporal. Determinar a correlação (**corrcoef**) entre as séries temporais. Comentar os resultados.
  - 1.4 Verificar a existência de *outliers*. Identifique-os, substitua-os por valores adequados e represente graficamente as séries temporais modificadas, comparando-as com as anteriores.