Análise e Transformação de Dados

Mini-Projeto

Objetivo: *No âmbito do mini-projeto pretende-se: i) analisar a série temporal associada ao dataset disponibilizado no inforestudante para cada turma PL, efetuando o seu pré-processamento; ii) decompor a série em componentes que traduzem os movimentos estruturais e erráticos; iii) identificar e estimar o modelo a considerar (AR, ARMA e ARIMA), bem como realizar o teste de diagnóstico do modelo e a sua utilização para previsão de valores futuros da série, seguindo os procedimentos indicados nas fichas práticas 3, 4 e 5.*

Linguagem: *MATLAB* ou Python.

Submissão: Cada grupo deverá submeter através do inforestudante, até ao final do dia 27/Abril/2018, o código realizado (comentado e organizado de forma a que os resultados obtidos possam ser confirmados pela execução do código) e um relatório que apresente os resultados obtidos e a justificação para as opções consideradas.
A submissão poderá ser feita através dum ficheiro .zip contendo os ficheiros com o código devidamente organizado e o relatório em pdf ou através de um *Jupyter Notebook* contendo, de forma integrada, o código em *MATLAB* ou em Python e o relatório.

Observação: A verificação da existência de fraude no trabalho implica para todos os intervenientes a reprovação na disciplina e a aplicação do regulamento disciplinar dos estudantes da UC.

Tarefas:

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 da uma série temporal deve-se proceder à sua preparação através do pré-processamento dos dados que envolve, normalmente, as seguintes operações:
   - Deteção 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;
   - Deteção 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, μ a média e σ o desvio padrão da série (ou de secções 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 a série temporal existente no ficheiro de dados de cada turma PL (ficheiro .csv com o registo do consumo elétrico total diário, em kWh).

1.2 Verificar a existência de valores não recolhidos/medidos, identificados com NaN (*Not a Number*). Identifique-os, elimine cada um desses valores da série temporal, substitua-os por valores que resultam de um processo de extrapolação (ou, se necessário, de interpolação) e represente graficamente a série temporal modificada, comparando com a original.
Sugestão: Considerar a reconstrução dos valores em falta usando extrapolação com o método '*pchip*' (**interp1**) ou com outro método adequado.

1.3 Determinar os valores da média (**mean**) e do desvio padrão (**std**) da série temporal total e por secções (por exemplo, mensais ou trimestrais). 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 a anterior.

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 factores e de natureza imprevisível. Estas quatro componentes podem ser combinadas de forma multiplicativa ou aditiva (forma a considerar neste trabalho).

   2.1 Com base na série temporal que resultou da regularização efetuada (sem valores NaN nem *outliers*), estimar a série temporal sem a componente da tendência, considerando aproximações polinomiais de grau 0 e 1 ou superior e usando a função **detrend**.

   2.2 Obter a componente da tendência paramétrica da série temporal.

   2.3 Representar graficamente a série temporal regularizada, a componente da tendência e a série temporal sem a componente da tendência, considerando a aproximação mais adequada (use as funções **polyfit** e **polyval** para o caso polinomial).

   2.4 Obter a componente da tendência paramétrica da série temporal.

   2.5 Representar graficamente a série temporal regularizada, a componente da tendência e a série temporal sem a tendência.

   2.6 Estimar a componente da sazonalidade da série temporal, considerando uma sazonalidade adequada (periodicidade semanal, mensal, trimestral ou outra).

   2.7 Obter a série temporal sem a componente da sazonalidade.

   2.8 Representar graficamente a série temporal regularizada, sem a componente da sazonalidade e a componente da sazonalidade.

   2.9 Admitindo que a componente cíclica da série temporal é pouco significativa, obter a componente irregular e a série temporal sem a componente irregular.

   2.10 Representar graficamente a série temporal regularizada, sem a componente irregular e a componente irregular.

3. Tendo por base a decomposição da série temporal nas suas componentes tendência, sazonal, cíclica e errática/irregular, as fases seguintes correspondem à determinação do modelo mais adequado para representar o comportamento da série e possibilitar a previsão de valores futuros. Neste trabalho considera-se que a série temporal pode ser descrita por um processo univariado Auto-regressivo (AR), Auto-Regressivo de Médias Móveis (ARMA) ou Auto-Regressivo Integrado de Médias Móveis (ARIMA).

   A escolha do modelo poderá ser estruturada nas seguintes 5 fases:

   Fase 1. **Verificação da estacionaridade da série**: Verificar se a série é estacionária e, caso não o seja, proceder a sucessivas diferenciações até atingir a estacionaridade.

   Fase 2. **Identificação do Modelo**: Determinar os critérios de definição do comportamento da série. Procura-se saber se a série segue um dos processos indicados usando os métodos de Função de Autocorrelação (FAC) e da Função de Autocorrelação Parcial (FACP).

   Fase 3. **Estimação do Modelo**: Estimar os modelos candidatos a serem selecionados após a identificação, procedendo-se à análise dos modelos mais adequados com base em critérios de escolha. Um dos critérios poderá ser o critério do Erro Quadrático Médio.

   Fase 4. **Teste de Diagnóstico**: Consiste em verificar se o modelo descreve adequadamente a série de dados objeto da análise.

   Fase 5. **Previsão**: Consiste em fazer a previsão, isto é, prever os valores futuros da série.

   3.1 Obter as componentes da série temporal que resultaram da decomposição da série temporal regularizada.

   3.2 Verificar a estacionaridade da série regularizada e da componente sazonal usando a função **adftest** do *MATLAB*. Se o resultado desta função for 1, a série deverá ser estacionária.

   3.3 Para identificação do modelo, representar graficamente a Função de Autocorrelação (FAC) e a Função de Autocorrelação Parcial (FACP) da componente sazonal da série, usando as funções **autocorr** e **parcorr**.

   3.4 Criar um objeto **iddata** para a componente sazonal da série, usando a função **iddata** e indicando que o período de amostragem, **Ts**, é de 1 dia.

   3.5 Estimar o modelo **AR** para a componente sazonal da série, considerando o resultado da FACP para definir o valor de ***na*** (histórico da variável a considerar) e definindo uma abordagem adequada (por exemplo, o método dos mínimos quadrados). Para isso, usar a função **arOptions** e **ar** para obter o modelo e a função **polydata** para obter os parâmetros.

   3.6 Para teste de diagnóstico, fazer a sua simulação usando os parâmetros do modelo da componente sazonal e a função **forecast**. Comparar graficamente a componente sazonal com o resultado do modelo.

   3.7 Validar o modelo comparando graficamente a série regularizada com o resultado da combinação aditiva da estimação da componente sazonal e da componente da tendência paramétrica da série. Se necessário, repetir as tarefas 3.5, 3.6 e 3.7 até obter o modelo adequado para a série, considerando como métrica, por exemplo, a soma do quadrado do erro entre os valores medidos e estimados da série.

   3.8 Obter e representar graficamente a previsão da série com o modelo **AR** para um horizonte temporal com uma duração igual ao dobro da original.

3.9 Estimar o modelo **ARMA** para a componente sazonal da série, considerando valores adequados para ***na*** (histórico da variável a considerar) e para ***nc*** (histórico do ruído branco a considerar, tendo em conta o resultado da FAC), definindo um método de procura adequado (por exemplo, a opção automática). Para isso, usar a função **armaxOptions** e **armax** para obter o modelo e a função **polydata** para obter os parâmetros.

3.10 Para teste de diagnóstico, fazer a sua simulação usando os parâmetros do modelo da componente sazonal e a função **forecast**. Comparar graficamente a componente sazonal com o resultado do modelo.

3.11 Validar o modelo comparando graficamente a série regularizada com o resultado da combinação aditiva da estimação da componente sazonal e da componente da tendência paramétrica da série. Se necessário, repetir as tarefas 3.9, 3.10 e 3.11 até obter o modelo adequado para a série, considerando como métrica, por exemplo, a soma do quadrado do erro entre os valores medidos e estimados da série.

3.12 Obter e representar graficamente a previsão da série com o modelo **ARMA** para um horizonte temporal com uma duração igual ao dobro da original.

3.13 Em alternativa às abordagens anteriores, estimar o modelo **ARIMA** da série regularizada, considerando valores adequados para ***p*** (grau do histórico da variável a considerar), ***D*** (número de operações de diferenciação até obter a série estacionária) e para ***q*** (grau do histórico do ruído branco a considerar). Para isso, usar a função **arima** para criar a estrutura do modelo e a função **estimate** para estimar o modelo da série.

3.14 Para teste de diagnóstico, fazer a sua simulação usando a função **simulate**. Comparar graficamente a componente sazonal com o resultado do modelo.

3.15 Validar o modelo comparando graficamente a série regularizada com o resultado da combinação aditiva da estimação da componente sazonal e da componente da tendência paramétrica da série. Se necessário, repetir as tarefas 3.13, 3.14 e 3.15 até obter o modelo adequado para a série, considerando como métrica, por exemplo, a soma do quadrado do erro entre os valores medidos e estimados da série.

3.16 Obter e representar graficamente a previsão da série com o modelo **ARIMA** para um horizonte temporal com uma duração igual ao dobro da original.

3.17 Comparar e comentar os resultados obtidos.