Análise e Transformação de Dados

Mini-Projeto

Licenciatura em Engenharia Informática

Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologias da Universidade de Coimbra

Bruno Grifo, N.º 2014 228 262, bgrifo@student.dei.uc.pt
Rúben Leal, N.º 2011 181 710, rleal@student.dei.uc.pt

Abril 2018

Conteúdo

Introdução

Este projeto teve como objetivo aplicar as técnicas e conhecimentos aprendidos durante a realização de algumas fichas nas aulas práticas num *dataset* novo. O grande desafio foi adaptar as técnicas aprendidas aos novos dados e perceber como abordar os resultados obtidos.

Foi feita a análise de uma série temporal associada a um *dataset*, fornecido pelos docentes, à qual foi feito o pré-processamento, sendo de seguida decomposta nas várias componentes – que traduzem os movimentos estruturais e erráticos – para a determinação de um modelo que represente o comportamento da série e permita a previsão de valores futuros. Foi feita uma análise dos resultados obtidos.

1 Pré-processamento dos dados

O pré-processamento dos dados envolveu a deteção de valores em falta (os valores *NaN*), substituindo cada um deles por valores estimados, usando o método de interpolação *spline*. Seguiu-se a deteção e regularização dos *outliers*.

Quanto aos *outliers*, foi usado como critério $|x_i - \mu| > 3\sigma$, como sugerido no enunciado. Tendo em consideração a possibilidade de erros de leitura ou ruídos no sinal, aplicamos um filtro que apanhasse valores muito atípicos. Substituímos todos estes valores de acordo com o critério $|x_i - \mu| > 2.6\sigma$, considerando que valores inferiores não tiveram origem em erros de leitura ou ruído.

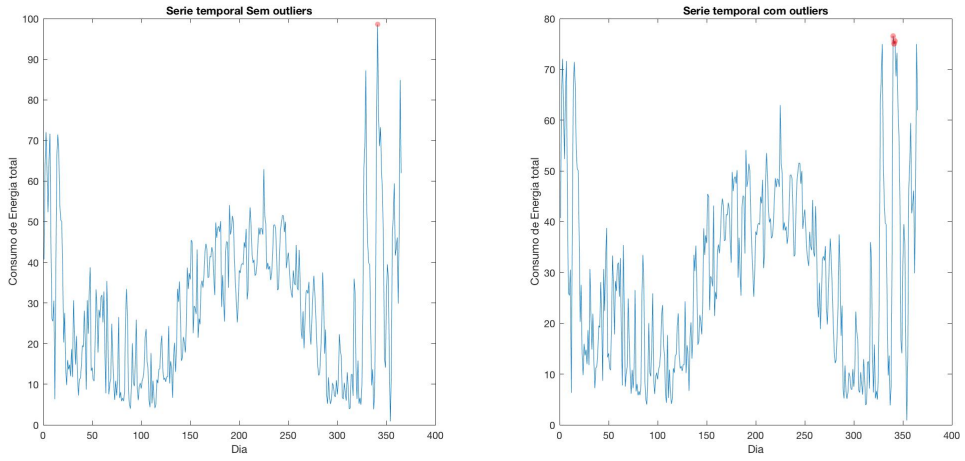


Figura 1: Outliers

2 Componentes da Série Temporal

Foram considerados as componentes associadas a movimentos estruturais e erráticos. Baseando-nos na série temporal sem valores *NaN* nem *outliers* começamos por estimar a mesma sem a componente de tendência, considerando uma aproximação polinomial de terceiro grau. Nesta fase poderíamos ter optado por calcular a tendência do sinal pela Série de Fourier ou através da função *polyfit()*(Calculo dos Coeficientes) do Matlab. Optamos por calcular tendência com a função *polyfit()* utilizando uma aproximação de terceiro grau por ter um comportamento que se aproxima do comportamento do sinal.

A Figura 2 mostra a série temporal regularizada, a série temporal sem tendência e a tendência.

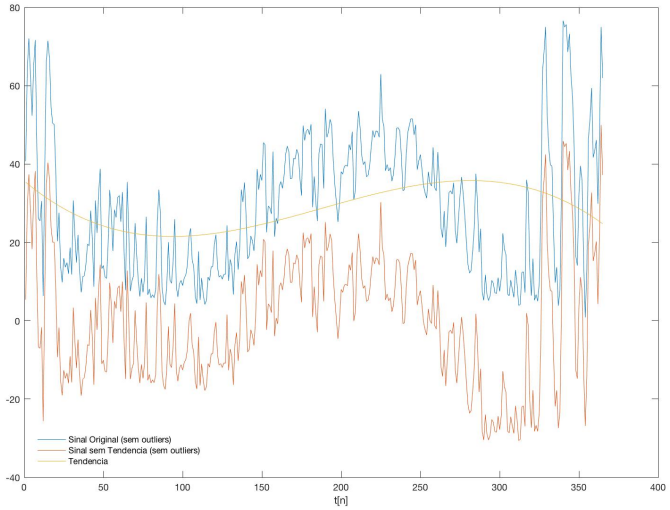


Figura 2: Tendência

De seguida foi feita a estimação da componente da sazonalidade da série. Nesta fase tivemos de decidir o período de sazonalidade (diário, semanal, mensal, trimestral, ...). Optámos por testar a sazonalidade para um período mensal e trimestral, como se pode verificar nas figuras 3 e 4, acabando por escolher sazonalidade mensal visto ter sido a sazonalidade que nos produziu melhores resultados na fase final do projeto.

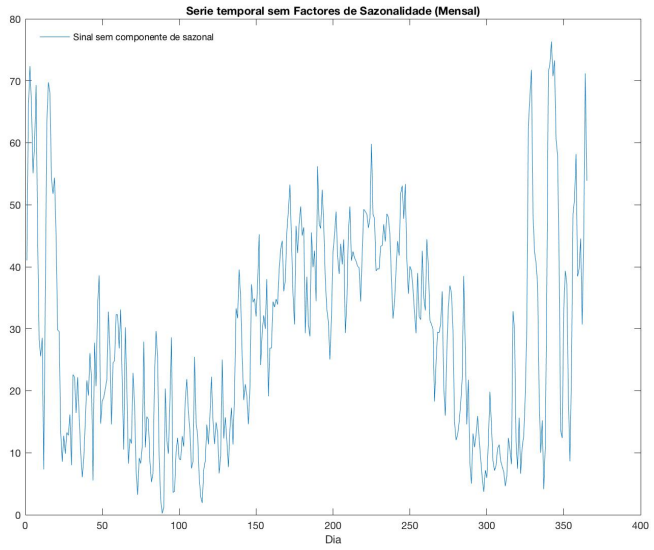


Figura 3: Sazonalidade Mensal

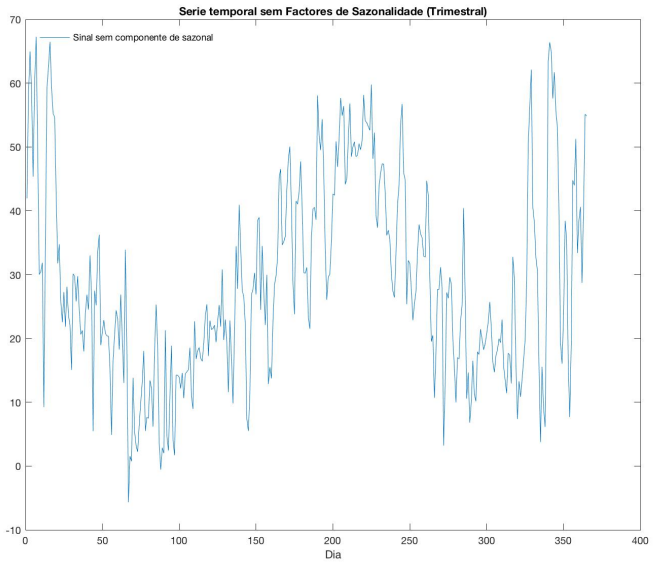


Figura 4: Sazonalidade Trimestral

Por fim, consideramos a componente de irregularidade da série, onde obtemos a componente irregular e a série temporal sem a componente irregular. Para obter a irregularidade da serie temporal, retiramos a componente de sazonalidade e a tendência à serie temporal. As figuras 5 e 6 mostram-nos a componente de irregularidade e o sinal sem a componente de irregularidade, respetivamente.

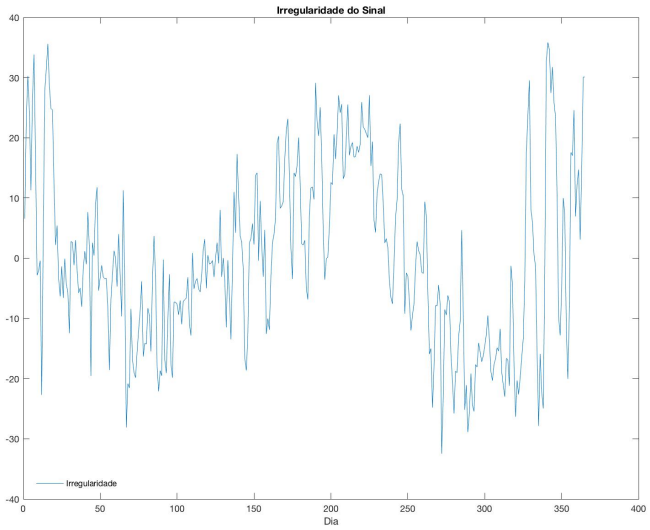


Figura 5: Irregularidade do Sinal

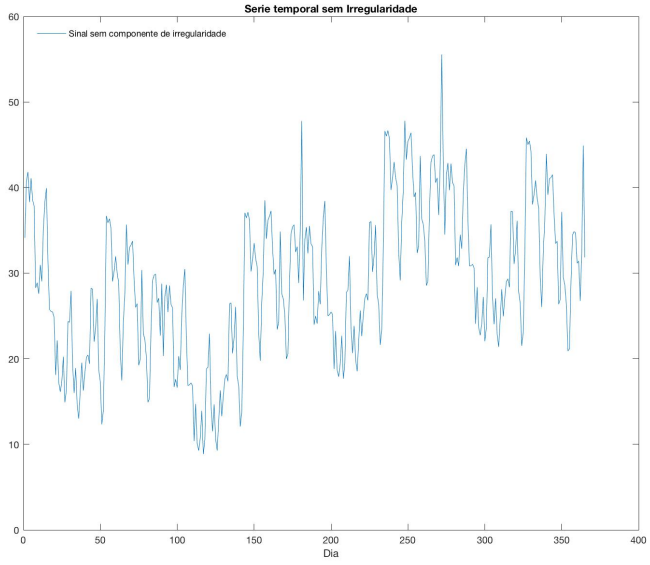


Figura 6: Sinal sem componente de irregularidade

3 Determinação do modelo de representação da série

Depois de verificar a estacionaridade da serie, através da função *adftest()*, procedemos à identificação do modelo. Para determinar os critérios de definição do comportamento da série procuramos saber se a série seguia um dos processos indicados usando os métodos de Função de Autocorrelação e da Função de Autocorrelação Parcial. Nas figuras 7 e 8 podemos observar a correlação dos fatores sazonais para um período de sazonalidade mensal e trimestral, respetivamente.

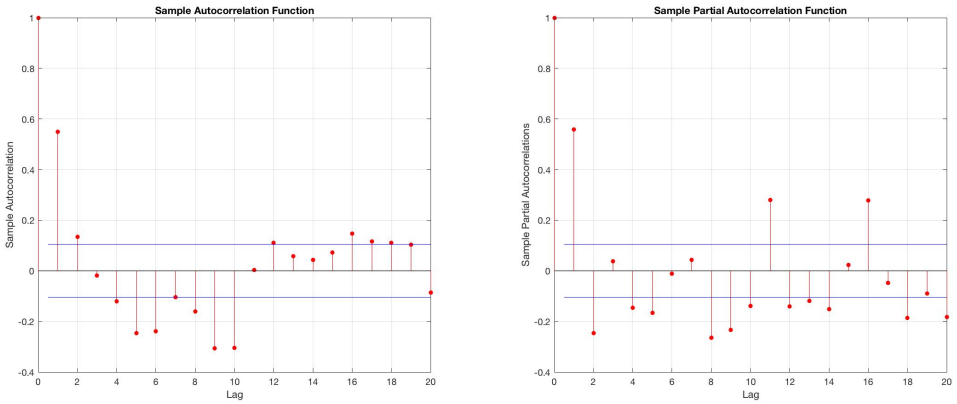


Figura 7: Correlação dos fatores de sazonalidade com período mensal

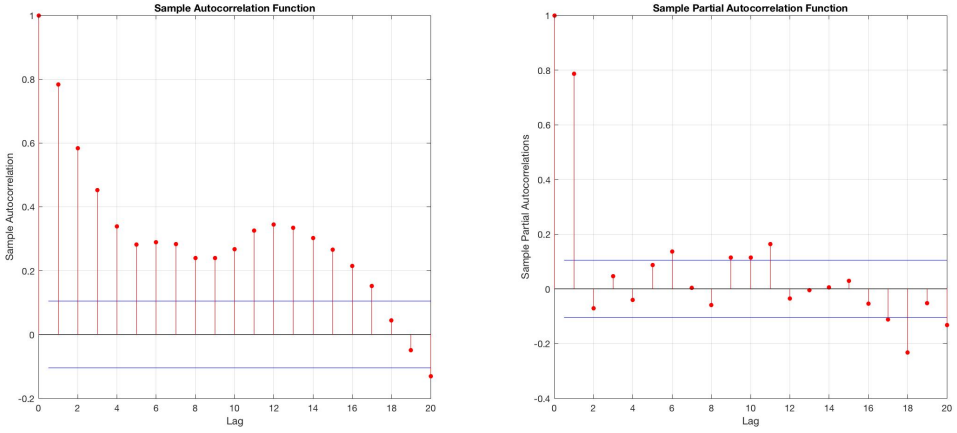


Figura 8: Correlação dos fatores de sazonalidade com período trimestral

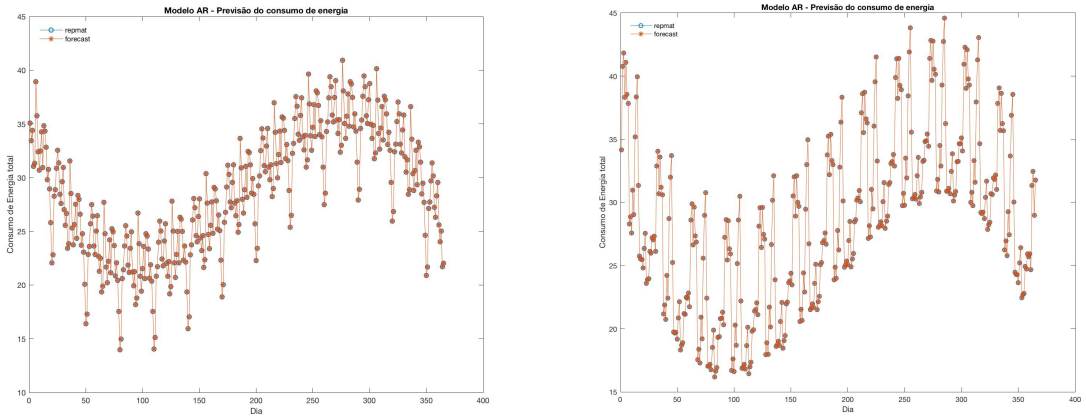
Como pudemos observar na figura 8, a correlação de fatores sazonais num período trimestral teve resultados um pouco altos e por isso mesmo decidimos utilizar um período de sazonalidade mensal na seguintes fases de modelação, visto terem obtido resultados mais promissores nas correlações.

Foi feita a determinação do modelo mais adequado para representar o comportamento da série e possibilitar previsões de valores futuros, tendo por base as componentes descritas na secção 2 – tendência, sazonal, cíclica e irregular. Aqui, considerámos os modelos uni-variados: Auto-regressivo (AR), Auto-Regressivo de Médias Móveis (ARMA) e Auto-Regressivo Integrado de Médias Móveis (ARIMA).

3.1 Modelo AR

Quanto aos parâmetros do modelo AR para a componente sazonal da série, foi feita uma estimação de ordem 20 (i.e. $na = 20$) – visto que no gráfico de correlação parcial os primeiros 20 valores têm uma correlação suficientemente alta para serem considerados no modelo de previsão –, usando como abordagem o método dos mínimos quadradados. Usamos a função *polydata* para obter estes parâmetros, que foram usados para fazer a simulação do modelo em questão.

As figuras 9a e 9b mostra-nos os resultados da previsão dos consumos de energia, durante um ano, utilizando o Modelo AR.



(a) Sazonalidade Mensal

(b) Sazonalidade Trimestral

Figura 9: Previsão dos gastos de energia considerando os períodos sazonais calculados anteriormente.

Na figura 9a, o erro entre os fatores de sazonalidade e o resultado do modelo é de 3.1381×10^5 e o erro entre o sinal original(sem *outliers*) e o resultado do modelo com a componente de tendência é de 4.1020×10^5.

Quanto à figura 9b o erro entre os fatores de sazonalidade e o resultado do modelo é de 3.3091×10^5 e o erro entre o sinal original(sem *outliers*) e o resultado do modelo com a componente de tendência é de 4.0435×10^5.

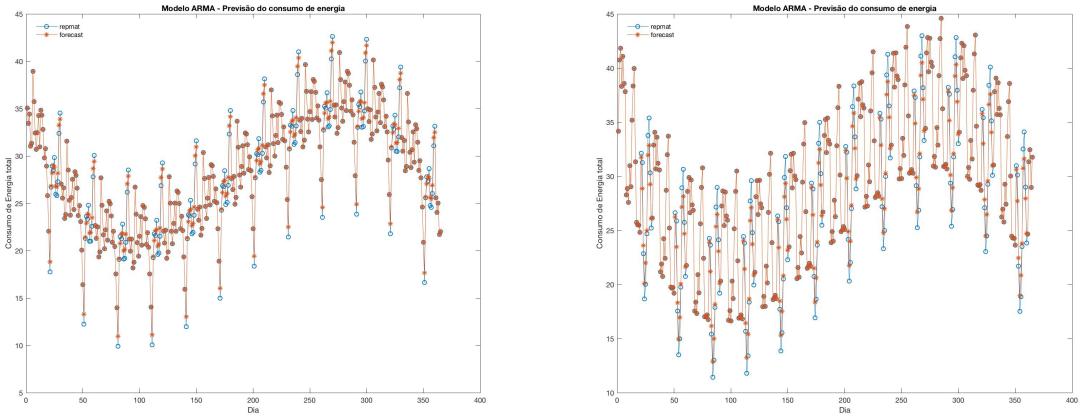
Como podemos verificar a figura 9b apresenta maior dispersão de valores enquanto que a figura

9a, seguindo o mesmo comportamento, apresenta uma menor dispersão de valores. Comparando com o sinal original, a figura 9a apresenta valores visualmente mais parecidos.

3.2 Modelo ARMA

Os parâmetros do modelo ARMA, para a componente sazonal da série, foram estimados tendo em conta $na = 20$ para os dois períodos de sazonalidade, $nc = 10$ para um período de sazonalidade mensal e $nc = 18$ para um período de sazonalidade trimestral, com um método de procura automático. A função *polydata* deu-nos estes parâmetros, que usamos para fazer a previsão da série.

As figuras 10a e 10b mostra-nos os resultados da previsão dos consumos de energia, durante um ano, com um ruído branco (*randn(30,1)*), utilizando o Modelo ARMA.



(a) Sazonalidade Mensal

(b) Sazonalidade Trimestral

Figura 10: Previsão dos gastos de energia considerando os períodos sazonais calculados anteriormente.

Na figura 10a, o erro entre os fatores de sazonalidade e o resultado do modelo é de 3.0673×10^5 e o erro entre o sinal original (sem *outliers*) e o resultado do modelo com a componente de tendência é de 4.0212×10^5.

Quanto à figura 10b o erro entre os fatores de sazonalidade e o resultado do modelo é de 3.4123×10^5 e o erro entre o sinal original (sem *outliers*) e o resultado do modelo com a componente de tendência é de 4.1297×10^5.

Como podemos verificar a figura 10b apresenta maior dispersão de valores enquanto que a figura 10a, seguindo o mesmo comportamento, apresenta uma menor dispersão de valores. Comparando com o sinal original, a figura 10a apresenta valores visualmente mais parecidos.

3.3 Modelo ARIMA

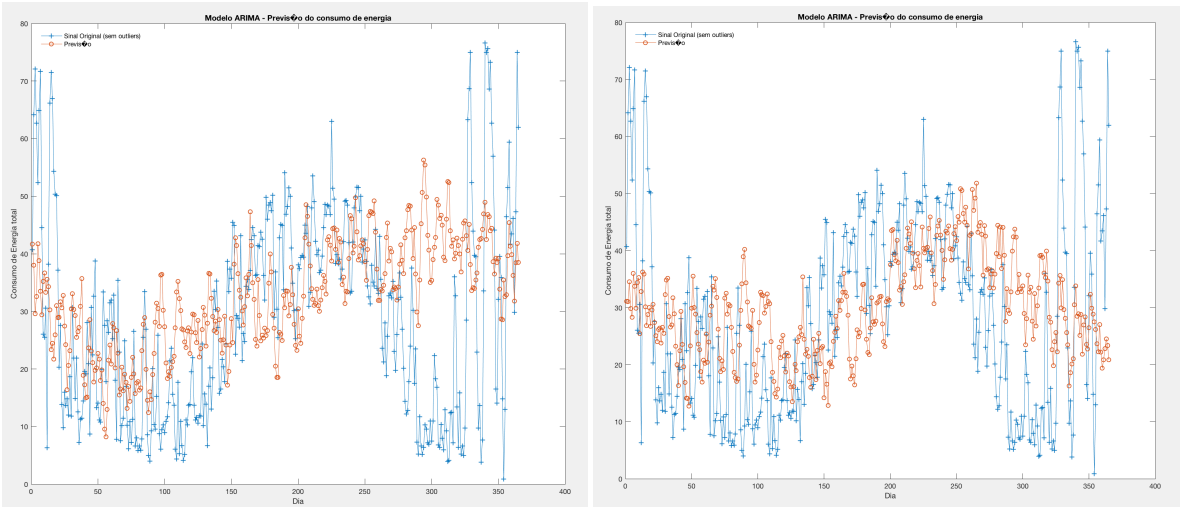
Por fim, estimamos o modelo ARIMA da série regularizada. Para isso usamos a função *arima* com um grau do histórico de 20 ($p = 20$) e uma operação de diferenciação ($D = 1$). Quanto ao grau de

histórico de ruído branco, testamos com vários valores sendo que os que produziram melhores valores foram os de ruído branco de 1 ($q = 1$) e de 3 ($q = 3$). Com a função **estimate**, estimámos o modelo da série.

Para estimar o modelo ARIMA tiramos a média de vários testes para tentar melhorar os resultados produzidos pelo nosso modelo. As figuras 11a, 11b mostram os resultados obtidos para a sazonalidade mensal com 5 testes para $q = 1$ e $q = 3$, respetivamente; 12a, 12b mostram os resultados obtidos para a sazonalidade mensal com 10 testes para $q = 1$ e $q = 3$, respetivamente; 13a, 13b mostram os resultados obtidos para a sazonalidade trimestral com 5 testes para $q = 1$ e $q = 3$, respetivamente; e, por fim, 14a, 14b mostram os resultados obtidos para a sazonalidade trimestral com 10 testes para $q = 1$ e $q = 3$, respetivamente. Cada uma com o respetivo erro na legenda.

De entre todas as figuras, as que mais se aproximam da série original são as figuras 11b e 13a, sendo que também apresentam um erro médio baixo. No entanto, há outras, como a figura 12a, que apesar do erro baixo, visualmente não parece descrever de forma adequada a série em estudo.

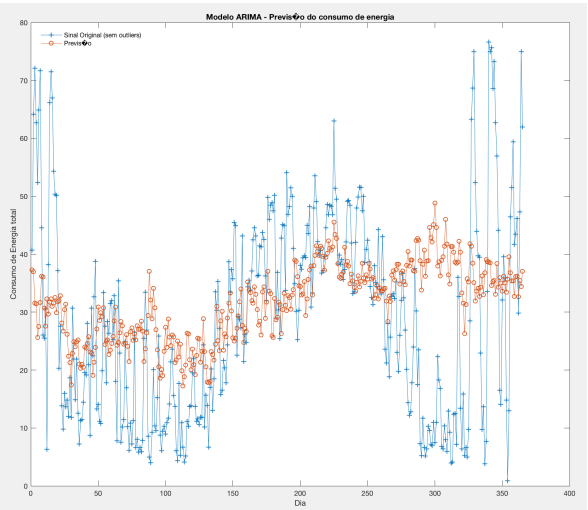
Sazonalidade Mensal:



(a) Erro:1.1960×10^5 para $q = 1$

(b) Erro:1.1158×10^5 para $q = 3$

Figura 11: Sazonalidade Mensal: Média de 5 testes.

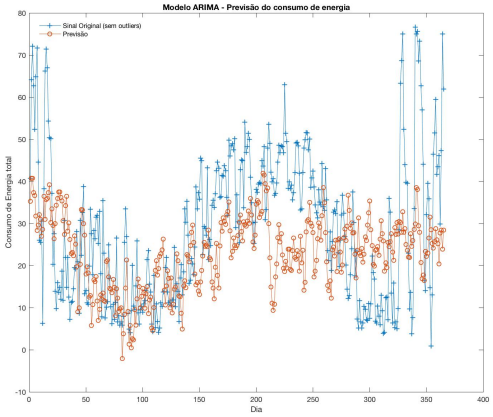


(a) Erro:1.0285×10^5 para $q = 1$

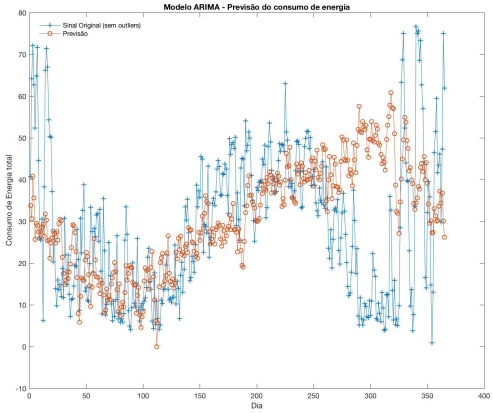
(b) Erro:1.1605×10^5 para $q = 3$

Figura 12: Sazonalidade Mensal: Média de 10 testes.

Sazonalidade Trimestral:

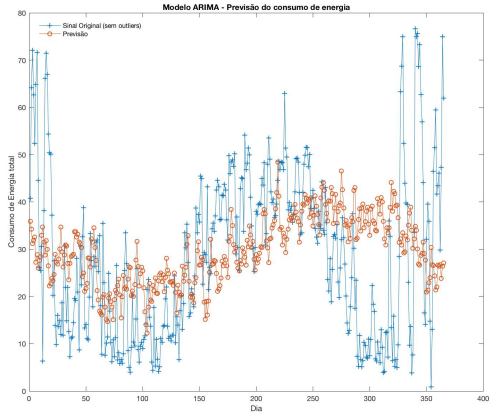


(a) Erro: 1.0596×10^5 para $q = 1$

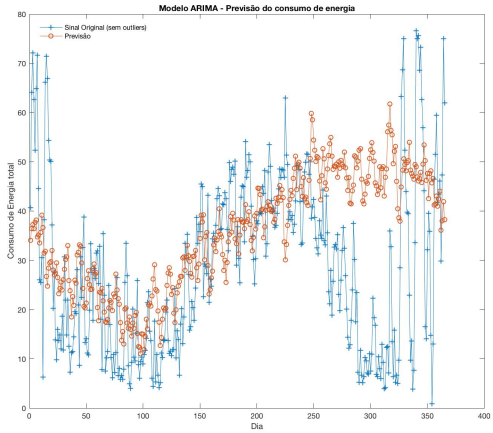


(b) Erro: 1.2768×10^5 para $q = 3$

Figura 13: Sazonalidade Trimestral: Média de 5 testes.



(a) Erro: 1.1079×10^5 para $q = 1$

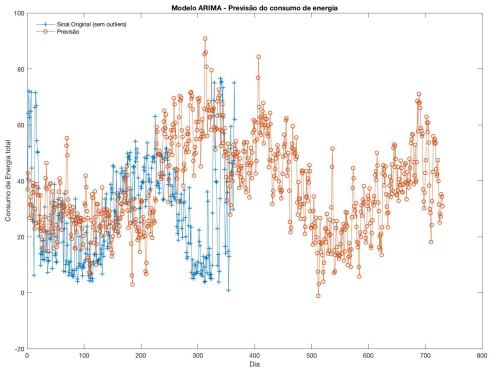


(b) Erro: 1.2913×10^5 para $q = 3$

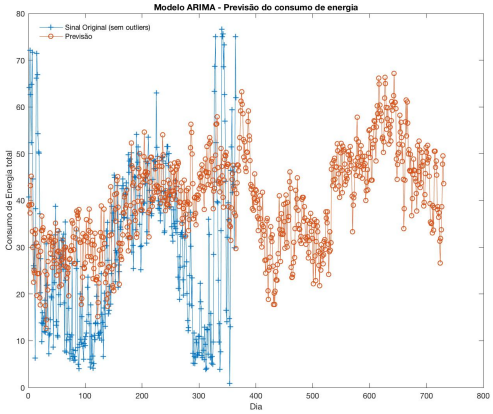
Figura 14: Sazonalidade Trimestral: Média de 10 testes.

4 Previsão para o ano seguinte

Utilizando o Modelo ARIMA e ajustando, conforme necessário, os valores tentámos prever os consumos de energia do próximo ano. Os resultados obtidos são apresentados nas figuras 15a, 15b, 16a e 16b.

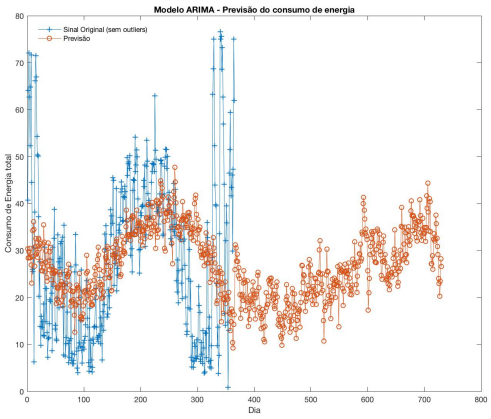


(a) Média de 2 testes para q=1

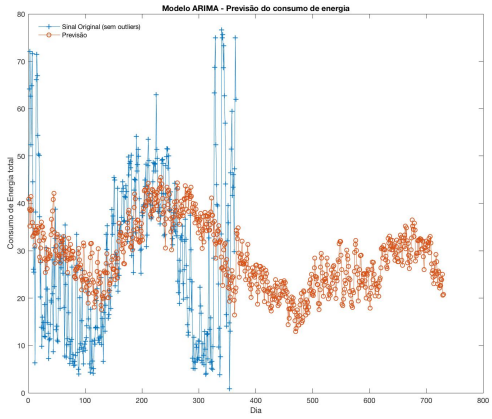


(b) Média de 5 testes para q=1

Figura 15: Previsão dos consumos de energia do próximo ano.



(a) Média de 10 testes para q=1



(b) Média de 15 testes para q=1

Figura 16: Previsão dos consumos de energia do próximo ano.

5 Observações

O grande objetivo desde projeto foi perceber e treinar as técnicas aprendidas nas aulas práticas. Foi-nos dado um *dataset* que nos obrigou a fazer alguns ajustes sobre o código e decisões de como abordar os resultados. Foram sentidas algumas dificuldades visto que com dados reais os resultados nem sempre são os esperados e muitas das vezes um pouco longe dos valores ótimos.