Uma imagem com texto

Descrição gerada automaticamente

**Previsão do Preço diário do Ouro**

**Docente:**

Profª. Doutora Isabel Pereira

**Discentes:**

Ana Rita Cheganças Nº 104633

Nuno Pedrosa Nº 94471

Índice

[Introdução 3](#_Toc107184731)

[Dados e análise exploratória 3](#_Toc107184732)

[*Dataset* 3](#_Toc107184733)

[Exploração dos dados 4](#_Toc107184734)

[Modelos ARIMA/SARIMA 8](#_Toc107184735)

[Separação do *dataset* em treino e teste 8](#_Toc107184736)

[Análise do ACF e PACF 9](#_Toc107184737)

[Obtenção de modelos SARIMA 10](#_Toc107184738)

[Análise residual para Modelos SARIMA 12](#_Toc107184739)

[Modelo *ExponenTial Smoothing* (ETS) 15](#_Toc107184740)

[Análise residual para Modelo ETS 15](#_Toc107184741)

[Previsões 15](#_Toc107184742)

[Discussão de Resultados/Conclusões 19](#_Toc107184743)

[Bibliografia 21](#_Toc107184744)

Índice de Figuras

[Figura 1 - QQPlot para Preço dário do Ouro 4](#_Toc107184745)

[Figura 2 - Histograma para Preço diário do Ouro 5](#_Toc107184746)

[Figura 3 - Gráfico Preço diário do Ouro 5](#_Toc107184747)

[Figura 4 - Comparação dados originais com 1ª's diferenças, log, e 1ª's diferenças do log 6](#_Toc107184748)

[Figura 5 - Modelo aditivo para os dados log 7](#_Toc107184749)

[Figura 6 - Modelo multiplicativo para os dados log 7](#_Toc107184750)

[Figura 7 - Modelo aditivo para os dados 1ª's diferenças de log 8](#_Toc107184751)

[Figura 8 - Modelo multiplicativo para os dados 1ª's diferenças de log 8](#_Toc107184752)

[Figura 9 - Gráfico com dados de treino nas 4 séries temporais obtidas 9](#_Toc107184753)

[Figura 10 - ACF e PACF das séries temporais escolhidas (1ª linha: dataset log; 2º linha: dataset log para 1ªdiferenças 10](#_Toc107184754)

[Figura 11 - Output da função SARIMA do modelo ARIMA(1,1,2) para dados log 11](#_Toc107184755)

[Figura 12 - Output da função SARIMA do modelo ARIMA (3,0,3) para dados log com 1ª's diferenças 12](#_Toc107184756)

[Figura 13 - Gráfico obtido a partir da função SARIMA do modelo ARIMA (1,1,2) para dados log 13](#_Toc107184757)

[Figura 14 - Gráfico da função SARIMA do modelo ARIMA(3,0,3) para dados log e com 1ª's diferenças 14](#_Toc107184758)

# Introdução

O presente relatório foi realizado no âmbito da Unidade Curricular de Séries Temporais no decurso do Mestrado de Ciência de Dados. Este trabalho tem como principal objetivo aplicação dos modelos SARIMA, *Bootstrap*, *forecasting* e *Exponential Smoothing* lecionados em aula. O tema adotado pelo nosso grupo tem por base a previsão dos preços diários do ouro. Este relatório encontra-se dividido em quatro partes distintas. Numa primeira parte do relatório, encontra-se a explicação do *dataset* utilizado e respetiva análise exploratória realizada ao mesmo. Numa segunda parte, apresenta-se a demonstração e explicação dos modelos aplicados. De seguida, apresenta-se as previsões realizadas. Por fim, na quarta parte, poderá encontrar-se a discussão dos resultados obtidos e respetivas conclusões.

O tema escolhido para este trabalho foi o preço do ouro. Este é um material escasso, que não consegue ser gerado pelo ser humano. Assim, o seu valor é completamente alicerçado em torno da sua durabilidade e qualidade, que se mantém com o passar do tempo, desde a antiguidade. Devido a isto, em termos históricos, o ouro foi muito utilizado para moeda de troca. Posteriormente, quando o dinheiro em papel foi introduzido, o ouro passou a ser considerado principalmente como uma reserva de valor. Algo muito importante, pois permite controlar o poder econômico diante de alterações sistêmicas sobre a economia, tanto a nível individual, como a nível dos cofres de governos e de instituições bancárias. Logo, é muito importante ter uma previsão do preço do ouro no futuro para que as reservas deste sejam geridas adequadamente.

# Dados e análise exploratória

## *Dataset*

A escolha do *dataset* para realização deste trabalho, recaiu sobre o preço diário do ouro. Este dataset foi retirado a partir do website [Kaggle](https://www.kaggle.com/datasets/psycon/daily-gold-price-historical-data), disponível no seguinte link: <https://www.kaggle.com/datasets/psycon/daily-gold-price-historical-data>

Este dataset é constituído por 5753 linhas, a primeira linha corresponde a 2000-01-04 e a ultima a 2022-06-22.

Para além disso, tem na sua constituição uma coluna referente à data, outra com o preço de abertura, uma com o preço de fecho, com o preço mais alto, preço mais Elevado, e por fim uma com a moeda de transição. Para o objetivo do nosso trabalho, foram selecionadas apenas duas colunas para serem trabalhadas, nomeadamente a dos preços de fecho e a da data correspondente.

## Exploração dos dados

De forma a percebermos melhor os dados, foi realizada uma análise exploratória aos mesmos. Com isto, conseguimos perceber que os dados apresentavam um preço médio de fecho e uma variância de 1044.9 e 269110.4, respetivamente. O preço mais baixo que o ouro atingiu foi cerca de 255.1, e o valor máximo atingido foi de 2054.6.

A análise do *QQplot* e do Histograma, permitem aferir a normalidade das variáveis a nível de representação gráfica, ou seja, permitem comparar a distribuição dos nossos dados com uma distribuição normal. Desta forma, conseguimos concluir que os nossos dados não contêm uma distribuição normal.

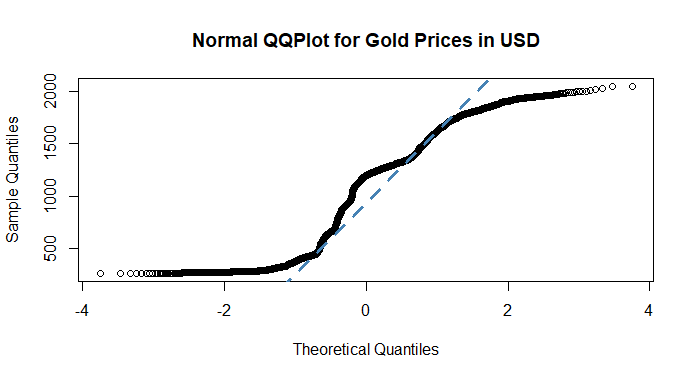


Figura 1 - QQPlot para Preço dário do Ouro

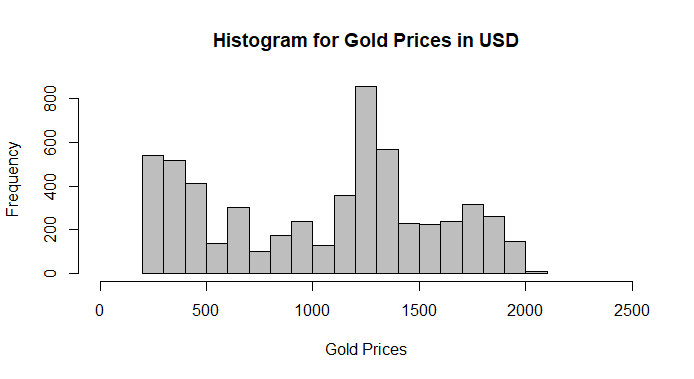


Figura 2 - Histograma para Preço diário do Ouro

Recorremos à visualização gráfica do conjunto completo de dados, de forma a termos uma ideia do tipo de série temporal presente. Assim sendo, a figura 3 mostra-nos que podemos estar perante uma *random walk*.

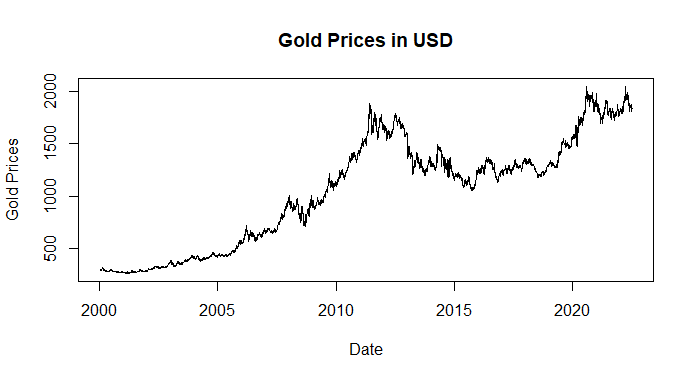


Figura 3 - Gráfico Preço diário do Ouro

Realizámos o cálculo das primeiras diferenças, do logaritmo, e das primeiras diferenças do logaritmo, de forma a melhorar a disposição dos dados. A figura 4 apresenta os resultados obtidos, comparando assim com os dados originais.

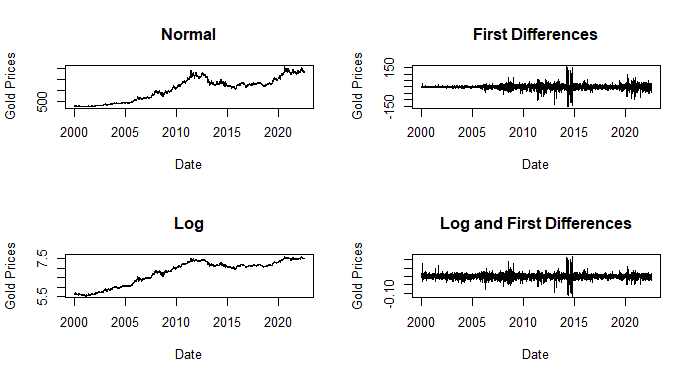


Figura 4 - Comparação dados originais com 1ª's diferenças, log, e 1ª's diferenças do log

Tendo em conta os resultados alcançados, começamos a utilizar o *dataset* com os valores calculados através do logaritmo e o *dataset* das 1ª’s diferenças do logaritmo.

Posteriormente, fomos verificar a estacionariedade de cada *dataset*. Através do teste adf, conseguimos concluir que o *dataset* com os dados do logaritmo, apresentam um *p-value* superior a 5%, o que significa que não se rejeita H0, pelo que não apresentam estacionariedade, o que é um índicio da existência de uma R*andom Walk*. Por outro lado, o *dataset* com os dados das 1ª’s diferenças do logaritmo, apresentam um *p-value* inferior a 5%, o que significa que se rejeita H0, que tem estacionariedade, e que é índicio da existência de uma *White Noise*.

Também fizemos um teste kpss para ver a estacionaridade. Este, interessantemente, considerou ambos os dados com estacionários, uma vez que os p values foram ambos menores que 5%, o que significa que se rejeita H0, que tem estacionariedade. Isto confirma que os dados para com logaritmo e primeiras diferenças são, de facto, estacionários, mas mostra que os dados apenas logaritmizados é estacionária em têndencia. Logo, para tornar a série estritamente estacionária, a tendência precisa de ser removida, algo que foi feito com as primeiras diferenças.

Por último, realizamos uma análise à sazonalidade de cada *dataset*. Existem dois modelos subjacentes a esta, nomeadamente, o modelo aditivo e o multiplicativo. Ambos foram testados para o *dataset* com os dados do logaritmo, e para o *dataset* com os dados das 1ª’s diferenças do logaritmo. Apesar de ambos os modelos terem sido testados, o que melhor se aplica no nosso caso, seria o multiplicativo. Assim sendo, a figura 6, demonstra o modelo multiplicativo para os dados do logaritmo, e conclui-se que não apresentam sazonalidade. Já a figura 8, demonstra o modelo multiplicativo para os dados das 1ª’s diferenças do logaritmo, e conclui-se que não apresentam sazonalidade.

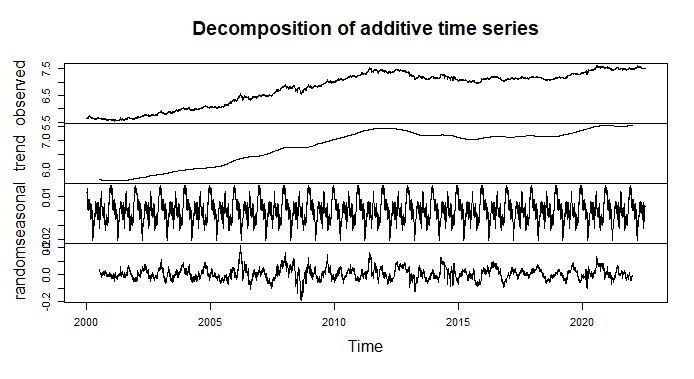


Figura 5 - Modelo aditivo para os dados log

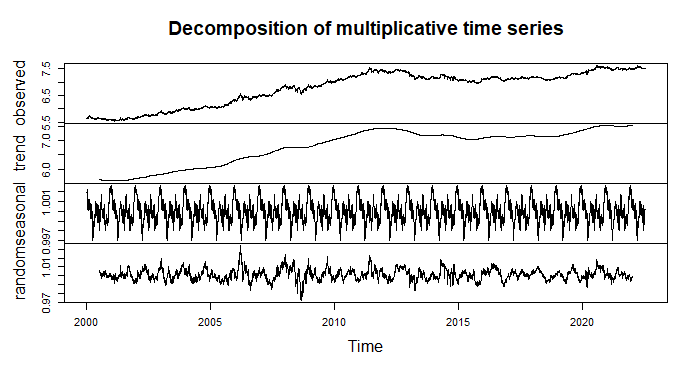


Figura 6 - Modelo multiplicativo para os dados log

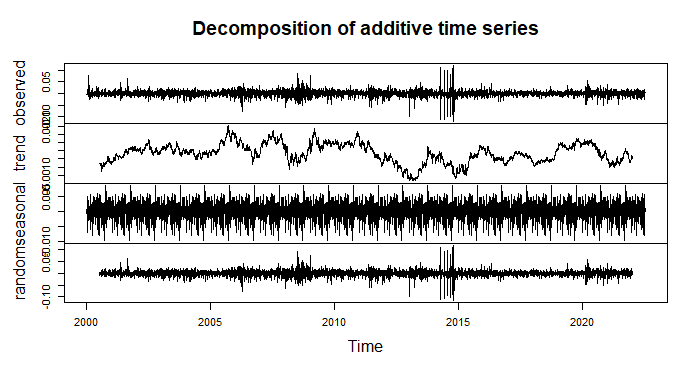


Figura 7 - Modelo aditivo para os dados 1ª's diferenças de log

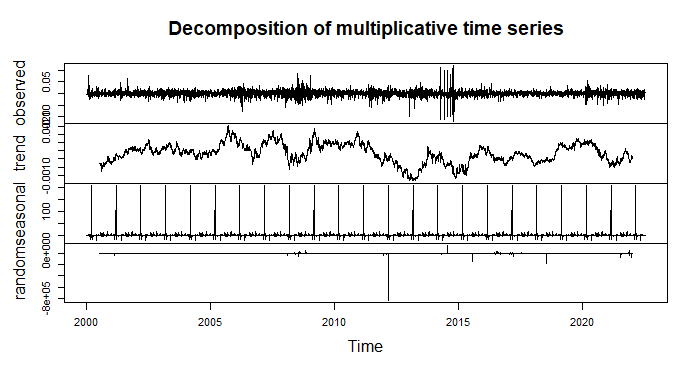


Figura 8 - Modelo multiplicativo para os dados 1ª's diferenças de log

# Modelos ARIMA/SARIMA

## Separação do *dataset* em treino e teste

Primeiro, de modo a tornar possível avaliar a performance do modelo que será obtido, procedeu-se à separação dos dados em dados de treino e dados de teste.

Decidiu-se que se iria usar como dados de teste os últimos 200 valores das séries temporais obtidas anteriormente. Estes valores correspondem a um pouco menos de um ano (nos dados tem-se, aproximadamente, 255 valores por ano), do dia 2021-09-21 ao dia 2022-06-22. Foi decidido este valor, tendo em conta que os nossos dados são muito irregulares, então resultados muito afastados do último valor conhecido podem ser especialmente longe dos dados reais.

Assim, tendo em conta as séries temporais obtidas anteriormente, foram obtidas novas séries temporais apenas com os dados de treino, os resultados estão representados na Figura 9.

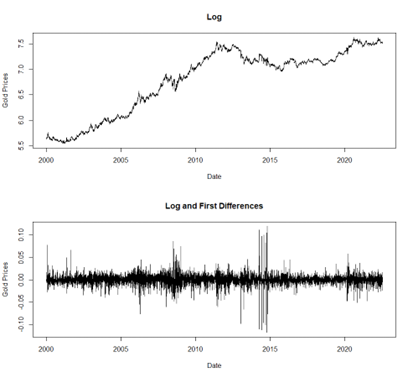


Figura 9 - Gráfico com dados de treino nas 4 séries temporais obtidas

## Análise do ACF e PACF

De seguida, antes de elaborar os modelos, fizemos uma análise do ACF (*Autocorrelation Function*) e do PACF (*Partial Autocorrelation Function*) das séries temporais obtidas. Os resultados estão representados na Figura 10.

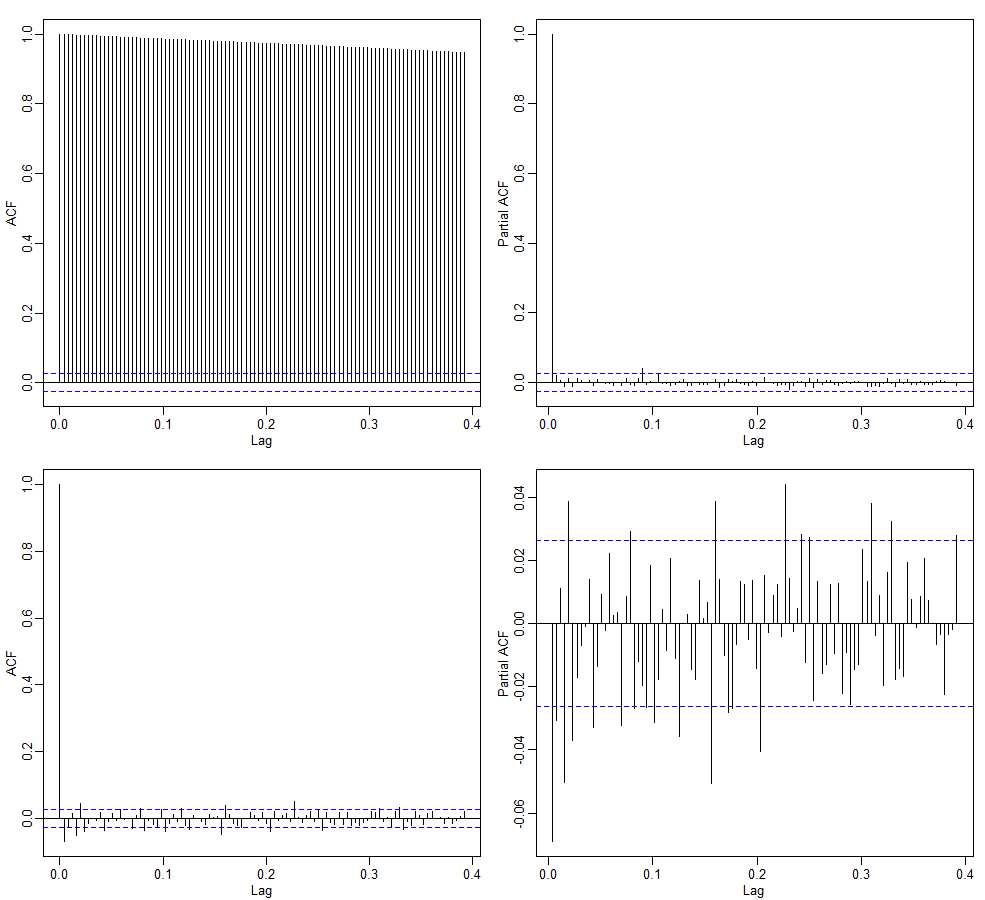


Figura - ACF e PACF das séries temporais escolhidas (1ª linha: dataset log; 2º linha: dataset log para 1ªdiferenças

Ao analisar as ACFs, podemos concluir que os dados log não são estacionários, uma vez que a ACF tende muito lentamente para zero. No entanto os dados log com primeiras diferenças apenas têm um lag diferente de 0 em 1, não evidenciando a não estacionariedade. Estes resultados estão conforme os testes feitos anteriormente. Quando ao PACF, nos dados log, apenas o valor em 1 é significativamente diferente de zero, enquanto que nos dados log com primeiras diferenças nota-se uma tendência lenta para zero.

Uma vez que os nossos modelos serão do tipo ARIMA, não se consegue tirar conclusões concretas sobre os seus parâmetros olhando apenas para as funções ACF e PACF.

## Obtenção de modelos SARIMA

Para se obter modelos adequados aos dados, foi utilizado o método *Box\_Jenkins* para modelos SARIMA.

Numa abordagem inicial, usou-se a função auto arima nos dados de treino, tendo-se obtido duas sugestões de modelo, um modelo ARIMA(2,1,2) com drift para os dados logaritmizados, e um modelo ARIMA(3,0,4) com média diferente de zero para o modelo logaritmizado com primeiras diferenças.

No entanto, decidiu-se fazer uma abordagem mais profunda para procurar os melhores modelos para os nossos dados. Assim, usou-se a função sarima com vários parâmetros diferentes, que faziam sentido, para se encontrar modelos otimizados. Chegou-se a 2 modelos, um para cada tipo de dados:

1. Modelo ARIMA(1,1,2) para dados logaritmizados:

Uma imagem com texto

Descrição gerada automaticamente

Figura 11 - Output da função SARIMA do modelo ARIMA(1,1,2) para dados log

Analisando a Figura 11, observa-se que, no modelo ARIMA(1,1,2), o parâmetro ar1 tem uma estimação de -0.7656, o parâmetro ma1 tem uma estimação de 0.6979, o parâmetro ma2 tem uma estimação de -0.0901 e a constante uma estimação de 0.0003. Todos estes valores são estatisticamente válidos, uma vez que os seus *p-values* são menores que 0.05, sugerindo que se obteve um modelo válido.

Outro aspeto importante são os valores de AIC (*Akaike Information Criteria),* AICc (*Corrected Akaike Information Criteria*) e 2 BIC (*Bayesian Information Criteria*). Estes valores são -5,992552, -5.992551 e -5.986589, respetivamente, que correspondem a quão bem o *fit* do modelo se adequa. Este modelo foi escolhido minimizando estes valores, em relação a outros modelos.

1. Modelo ARIMA(3,0,3) para dados logaritmizados e com primeiras diferenças:

Uma imagem com texto

Descrição gerada automaticamente

Figura 12 - Output da função SARIMA do modelo ARIMA (3,0,3) para dados log com 1ª's diferenças

Analisando a Figura 12, observa-se que, no modelo ARIMA (3,0,3), o parâmetro ar1 tem uma estimação de -0.6040, o parâmetro ar2 tem uma estimação de 0.5219, o parâmetro ar3 tem uma estimação de 0.7444, o parâmetro ma1 tem uma estimação de 0.5476, o parâmetro ma2 tem uma estimação de -0.5743, o parâmetro ma3 tem uma estimação de -0.7172 e a constante uma estimação de 0.0003. Todos estes valores são estatisticamente válidos, uma vez que os seus *p-values* são menores que 0.05, sugerindo que se obteve um modelo válido.

Outro aspeto importante são os valores de AIC (*Akaike Information Criteria*), AICc (*Corrected Akaike Information Criteria*) e 2 BIC (*Bayesian Information Criteria*). Estes valores são -5,99553, -5.995526 e -5.98599, respetivamente, que correspondem a quão bem o *fit* do modelo se adequa. Este modelo foi escolhido minimizando estes valores, em relação a outros modelos.

## Análise residual para Modelos SARIMA

Foi feita uma análise dos resíduos dos modelos para perceber se os modelos são bons.

1. Modelo ARIMA (1,1,2) para dados logaritmizados:

O gráfico dos resíduos, bem como o seu ACF, *QQ Plot* e *p-values* do *Ljung Box* estão representados na Figura 13.

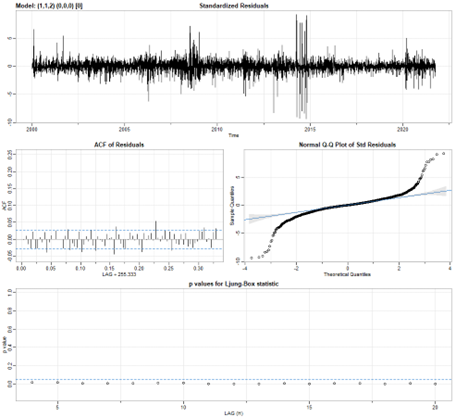


Figura 13 - Gráfico obtido a partir da função SARIMA do modelo ARIMA (1,1,2) para dados log

A nossa análise residual obtida na Figura 13 não é das melhores, uma vez que, na ACF, existem alguns valores que são significativamente diferentes de 0, e nos *p-values* de *Ljung-Box* existem alguns valores muito próximos de 0, evidenciando que alguns dos nossos resíduos são correlacionados. Avaliando o *QQ Plot*, pudemos ver que os nossos valores se desviam muito da linha central, evidenciando que não são explicados por uma distribuição normal. Os resíduos obtidos têm uma média de -2.065831e-06 e uma variância de 0.0001459576, sendo estes valores muito pequenos, sendo um bom sinal.

Posteriormente, foram feitos alguns testes para confirmar certas características dos nossos dados:

* *Box.test* a lag 10 para testar a correlação entre resíduos: obteu-se um *p-value* = 0.04166, logo, não aceitamos a hipótese nula de que os resíduos são ruído branco (nenhuma correlação se ruido branco) a lag 10, o que é um mau sinal. Assim, ainda há informações deixadas nos resíduos que podiam ser usadas no cálculo das previsões.
* *Shapiro* e *Kolmogorov-Smirnov* testes para testar a normalidade: como o *p-value* é aproximadamente 0 para os dois testes, os resíduos não seguem uma distribuição normal (como já evidenciado pelo *QQ plot*).

1. Modelo ARIMA(3,0,3) para dados logaritmizados e com primeiras diferenças:

O gráfico dos resíduos, bem como o seu ACF, *QQ Plot* e *p-values* do *Ljung Box* estão representados na Figura 14.

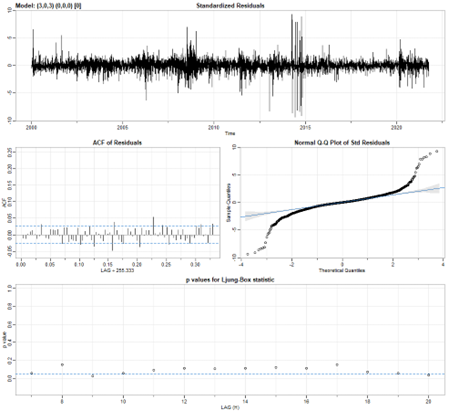


Figura 14 - Gráfico da função SARIMA do modelo ARIMA(3,0,3) para dados log e com 1ª's diferenças

A nossa análise residual obtida na Figura 14 não é das melhores, uma vez que, em semelhança ao obtido no modelo anterior, na ACF, existem alguns valores que são significativamente diferentes de 0, e nos *p-values* de *Ljung-Box* existem alguns valores muito próximos de 0, evidenciando que alguns dos nossos resíduos são correlacionados. Avaliando o *QQ Plot*, também em semelhança com o modelo anterior, pudemos ver que os nossos valores se desviam muito da linha central, evidenciando que não são explicados por uma distribuição normal.

Os resíduos obtidos têm uma média de 2.108876e-06 e uma variância de 0.0001453858, sendo estes valores muito pequenos, o que é um bom sinal.

Posteriormente, tal como no modelo anterior, foram feitos alguns testes para confirmar certas características dos nossos dados:

* *Box.test* a lag 10 para testar a correlação entre resíduos: obteu-se um *p-value* = 0.5185, logo, aceitamos a hipótese nula de que os resíduos são ruído branco (nenhuma correlação) a lag 10, o que é um bom sinal.
* *Shapiro* e *Kolmogorov-Smirnov* testes para testar a normalidade, como o *p-value* é aproximadamente 0 para os dois testes, os resíduos não seguem uma distribuição normal (como já evidenciado pelo *QQ plot* e como aconteceu no outro modelo).

APLICAR TAMBÉM ACF E PACF SOBRE OS RESÍDUOS

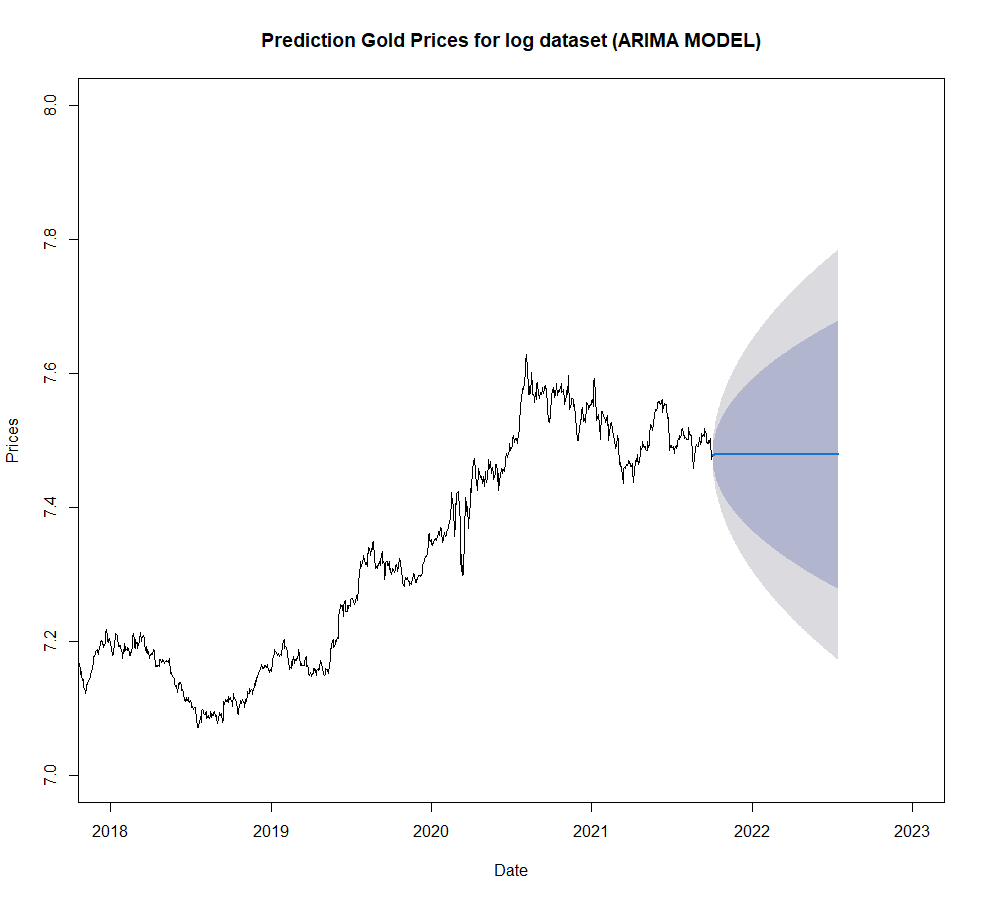
## Modelo *ExponenTial Smoothing* (ETS)

## Análise residual para Modelo ETS

# Previsões

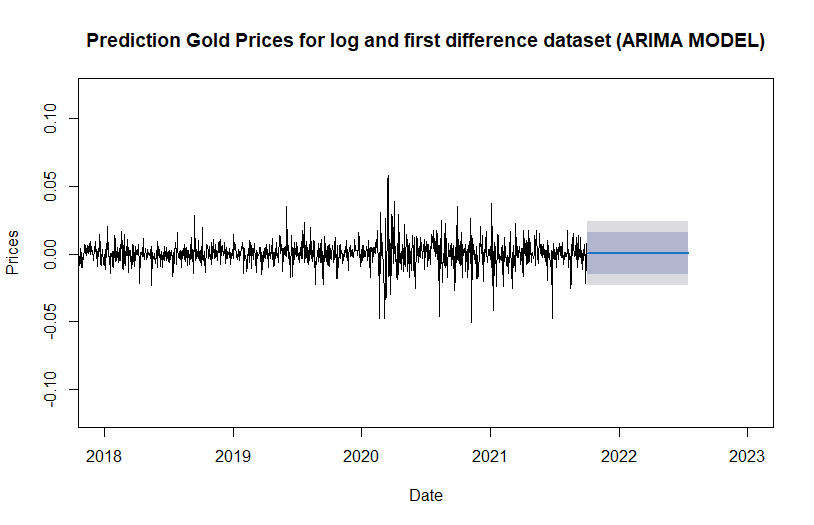
Após aplicação dos modelos expostos nos tópicos anteriores e respetiva escolha dos que apresentavam melhores parâmetros, foram efetuadas previsões. Essas previsões foram realizadas através da biblioteca do R denominada por “*forecast*”. Para além disso, primeiramente foram efetuadas previsões para o *dataset* de treino, e só depois para o de *dataset* de teste. Foram previstos preços para 200 dias. De forma a conseguirmos perceber melhor os resultados de cada previsão, foram analisadas as figuras que se seguem:

1. Previsões efetuadas para modelo ARIMA (1,1,2) - Dados logaritmizados:



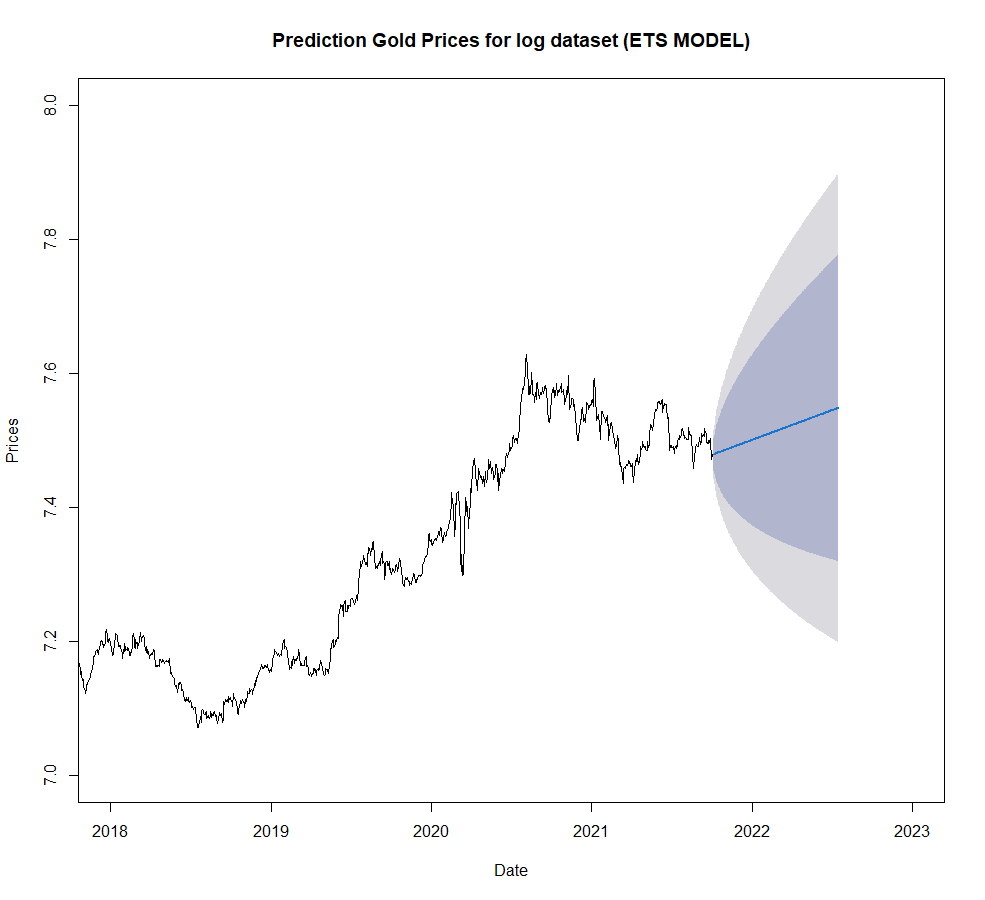
Após aplicação da previsão com o modelo ARIMA e com os dados logaritmizados, recorremos a uma técnica de avaliação denominada accuracy. Através desta, conseguimos visualizar nos dados de treino um RMSE (*Root Mean Square Error*) de 0.0121, e um MASE (*Mean Absolute Standardized Error*) de 0.05473. Já nos dados de teste um RMSE de 0.0571, e um MASE de 0.31897.

1. Previsões efetuadas para modelo ARIMA (3,0,3) – Dados logaritmizados com 1ª’s diferenças:



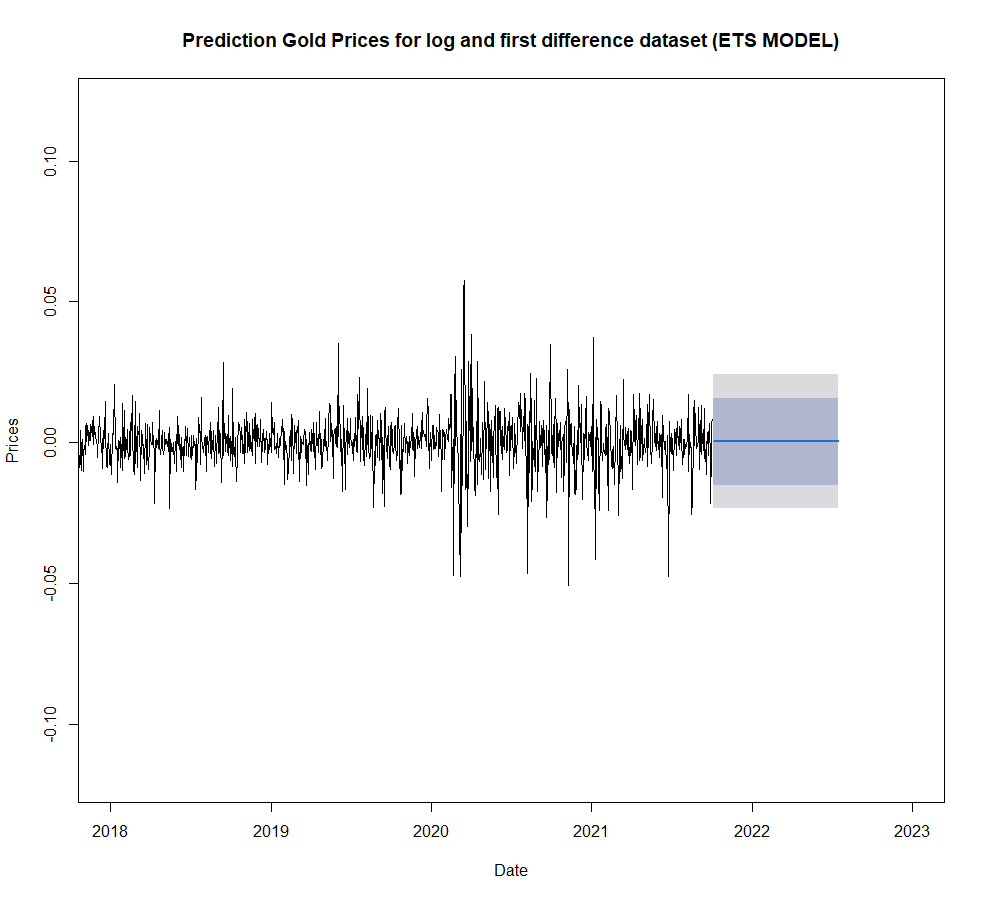
Após aplicação da previsão com o modelo ARIMA e com os dados logaritmizados com as primeiras diferenças, conseguimos visualizar nos dados de treino um RMSE de 0.0120, e um MASE de 0.6614. Já nos dados de teste um RMSE de 0.0092, e um MASE de 0.5807.

1. Previsões efetuadas para modelo ETS - Dados logaritmizados:



Após aplicação da previsão com o modelo ETS e com os dados logaritmizados, conseguimos visualizar nos dados de treino um RMSE de 0.0120, e um MASE de 0.0546. Já nos dados de teste um RMSE de 0.0298, e um MASE de 0.1538.

1. Previsões efetuadas para modelo ETS - Dados logaritmizados com 1ª’s diferenças:



Após aplicação da previsão com o modelo ETS e com os dados logaritmizados com as primeiras diferenças, conseguimos visualizar nos dados de treino um RMSE de 0.0121, e um MASE de 0.6610. Já nos dados de teste um RMSE de 0.00919, e um MASE de 0.5806.

Assim, o melhor modelo e os melhores dados para se efetuar as previsões, tendo como critério a minimização do valor dos erros, será para o modelo ETS com os dados logaritmizados.

# Discussão de Resultados/Conclusões

# Bibliografia

Athanasopoulos, G., & Hyndman, R. (2018). *Forecasting: Principles and Practice 2nd Edition.* Monash University, Australia: Texts.

Hyndman, R., Koehler, A., Ord, J., & Snyder, R. (2011). *Forecasting with Exponential Smoothing: The State Space Approach.*

Pereira, I. (2021/2022). Time Series: Section 3. Modeling SARIMA processes.

Pereira, I. (2021/2022). Time Series: Section 4. Forecast.

Pereira, I. (2021/2022). Time Series: Section 6. Exponential Smoothing Methods.

Shewhart, A., & S.Wilks, S. (2010). *Time Series Application to Finance.*