

DOUGLAS VINÍCIUS GONÇALVES ARAÚJO

**Análise de Séries Temporais: Aplicação do  
modelo ARIMA para previsão do Preço da Soja**

JI-PARANÁ

2022

DOUGLAS VINÍCIUS GONÇALVES ARAÚJO

**Análise de Séries Temporais: Aplicação do modelo  
ARIMA para previsão do Preço da Soja**

Relatório de Pesquisa na disciplina de Séries Temporais, do curso de Bacharel em Estatística da Universidade Federal de Rondônia, *campus* Ji-Paraná, como requisito de avaliação.

UNIVERSIDADE FEDERAL DE RONDÔNIA – UNIR  
DEPARTAMENTO DE MATEMÁTICA E ESTATÍSTICA  
RELATÓRIO DE PESQUISA

JI-PARANÁ

2022

*"Os livros servem para nos lembrar quanto somos estúpidos e tolos. São o guarda pretoriano de César, cochichando enquanto o desfile ruge pela avenida: Lembre-se, César, tu és mortal. A maioria de nós não pode sair correndo por aí, falar com todo mundo, conhecer todas as cidades do mundo, não temos tempo, dinheiro ou tantos amigos assim. As coisas que você está procurando, Montag, estão no mundo, mas a única possibilidade que o sujeito comum terá de ver noventa e nove por cento delas está num livro".*

*- Fahrenheit 451 de Ray Douglas Bradbury*

# Resumo

O objetivo deste trabalho tem como propor a análise do comportamento do preço médio mensal da saca de 60kg (R\$/sc), bem como a previsão aplicada a modelagem ARIMA (Modelo Médio Integrado Auto-Regressivo) da *commodities* de soja no Brasil. Os dados de séries foram coletados no site CEPEA (Centro de Estudos Avançados em Economia Aplicada) da metodologia soja CEPEA/ESALQ - PARANÁ, entre os intervalos de agosto de 1997 a outubro de 2022. Empregamos a metodoliga Box-Jenkins (1976) com o designo de encontrar o melhor ajuste do modelo de série temporal com base histórica dos preços. Ademais, foi utilizado uma função automatizada do pacote "forecast" do Software R para obter o modelo ideal para série. Assim, o modelo **ARIMA**(1,1,0)(2,0,0) foi o melhor para prever o preço médio mensal de soja. Os resultados ilustram um tendência crescente nos preços e na decomposição da série viu-se que há sazonalidade. Assim, este estudo pode fornecer uma abordagem como método de apoio à tomada de decisão para formular a previsão de preço da soja para o mercado brasileiro.

**Palavras-chaves:** Séries Temporais, Modelo ARIMA, Preços da Soja.

# Lista de ilustrações

Figura 1 – Produção de Soja por continentes em 2020 . . . . .	7
Figura 2 – Produção de Soja por ano no Brasil de 1994 a 2020 . . . . .	8
Figura 3 – Família de trajetórias de processos estocásticos. . . . .	10
Figura 4 – Um processo estocástico interpretado como uma família de variáveis. . .	10
Figura 5 – Processo geral de previsão usando um modelo ARIMA. . . . .	12
Figura 6 – Indicador de preço da soja no período de outubro de 1997 até outubro de 2022 . . . . .	14
Figura 7 – Decomposição da série de preço da soja . . . . .	15
Figura 8 – Gráfico de Boxplot, da série distribuída entre os meses (período de 10/1997 a 10/2022) . . . . .	16
Figura 9 – Correlograma da série . . . . .	17
Figura 10 – Gráficos . . . . .	18
Figura 11 – Gráficos de resíduos para o modelo $ARIMA(1,1,0)(2,0,0)$ . . . . .	21
Figura 12 – Previsão para o índice de preço da soja . . . . .	21

# Lista de tabelas

Tabela 1 – Estística Descritiva . . . . .	14
---	----

# Sumário

1	INTRODUÇÃO . . . . .	7
2	SÉRIES TEMPORAIS . . . . .	9
2.1	Processos Estocásticos . . . . .	9
2.2	Modelos ARIMA . . . . .	10
3	METODOLOGIA . . . . .	13
4	RESULTADOS E DISCUSSÕES . . . . .	14
5	CONSIDERAÇÕES FINAIS . . . . .	22
	REFERÊNCIAS . . . . .	23
	APÊNDICES . . . . .	24
	APÊNDICE A – SCRIPT R . . . . .	25

# 1 Introdução

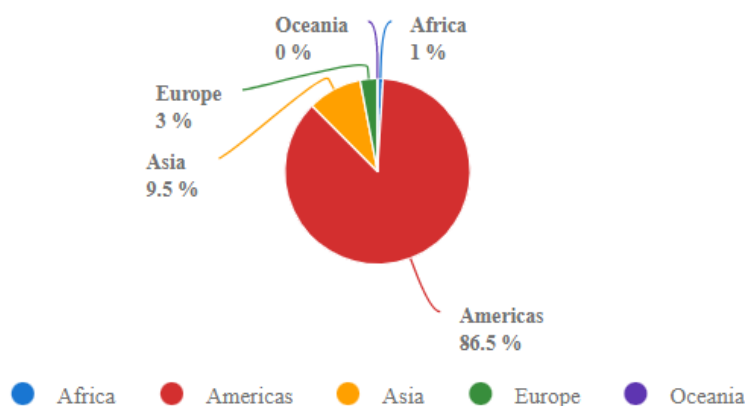
A soja derivou na costa leste da Ásia, nas proximidades do rio Yangtse na China. Antigamente a soja eram plantas rasteiras, bem distinta de hoje que cultivamos que foram modificadas e melhoradas pelo chineses. Apesar de ser conhecida e consumida pela civilização oriental por milhares de anos, ocorreu a sua introdução na Europa no final do século XV.

No Brasil, na década de 60, dois fatores internos que fizeram a produção de soja expandir, fatores estes que influenciaram o cenário mundial da produção de grãos: *i)* com o esforço da produção de suínos e aves assim gerava demanda pelo farelo de soja, já era uma necessidade estratégica; *ii)* na década de 70 houve a explosão do preço da soja no mercado mundial e os agricultores e o governo beneficiavam de uma vantagem competitiva em relação de outros países, pois o escoamento da safra brasileira ocorre na entressafra americana, quando os preços atingem as maiores cotações, em conformidade com (EMBRAPA, 2022).

A produção de soja avançou significativamente no mundo nestes últimos 20 anos, com taxas de crescimento de 3,7% ao ano. Os três maiores produtores de soja mundial, juntos, respondem por mais de 80% da produção, que são Brasil, Estados Unidos (EUA) e Argentina. E destes, o país que mais apresentou taxa de crescimento da produção nessas últimas décadas é o Brasil, com aumento de 5,9% a. a., superior entre os seus concorrentes (EUA avançou 2,7% a. a. e a Argentina 1,6% a. a.), segundo (CEPEA, 2022).

Neste mesmo período o Brasil deixou o posto de segundo lugar para ser o maior

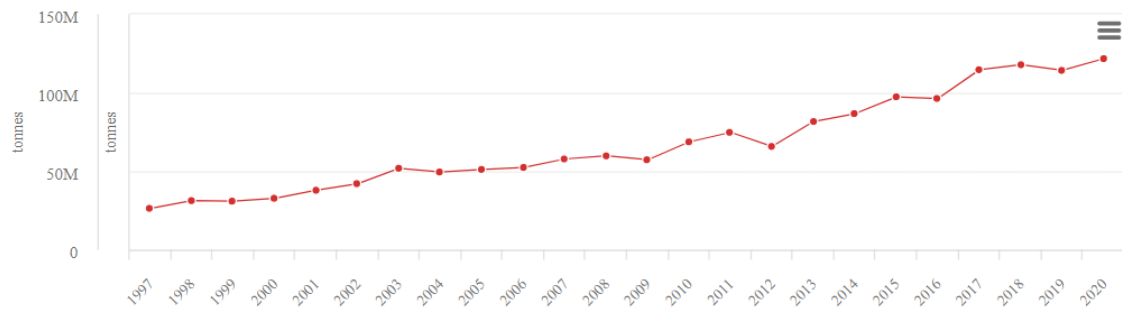
Figura 1 – Produção de Soja por continentes em 2020



Fonte: (FAOSTAT, 2022)



Figura 2 – Produção de Soja por ano no Brasil de 1994 a 2020



Fonte: ([FAOSTAT](#), 2022)

produtor de soja - na safra de 2001/2002, a produção brasileira era aproximadamente 52 milhões de toneladas, nesta mesma safra o EUA era aproximadamente 75 milhões de toneladas, agora na safra 2021/2022 o Brasil alcançou perto de 122 milhões toneladas, superior ao EUA que entorno de 120,7 milhões de toneladas, conforme dados da ([CONAB](#), 2022).

Diante do exposto deste trabalho, pretende à aplicação de modelos **ARIMA** (Autoregressivos Integrados de Médias Móveis) nos índices de preços da *commodities* soja com a finalidade de compreender o comportamento da comercialização deste produto.

## 2 Séries Temporais

Qualquer conjunto de observações ordenadas no tempo, é uma série temporal. Por exemplos:

- i)* temperaturas médias diárias de uma cidade;
- ii)* vendas mensais de uma empresa;
- iii)* valores de fechamento diários da IBOVESPA;
- iv)* preços diários de *commodities*;
- v)* valores mensais do IPCA (Índice Nacional de Preços ao Consumidor Amplo).

As séries temporais podem ser tanto discreta quanto contínua, ou seja, discreta é o intervalo entre as observações pertencentes a um conjunto discreto, e contínua é o intervalo entre as observações pertencentes a um conjunto contínuo. Observa-se que quando dizemos que uma série é discreta, estamos fazendo referência ao tempo entre as observações e não a escala da variável.

Além disso, temos dois enfoques usados na análise de séries temporais. O objetivo de ambos é construir modelos para as séries. No primeiro enfoque é feita análise no *domínio temporal* e os modelos sugeridos são *modelos paramétricos* (números de parâmetros finitos), o segundo é conduzido no *domínio de frequências* e os propostos são *modelos não-paramétricos* (MORETTIN; TOLOI, 2006). O modelo utilizado neste estudo, **ARIMA** é um modelo paramétrico.

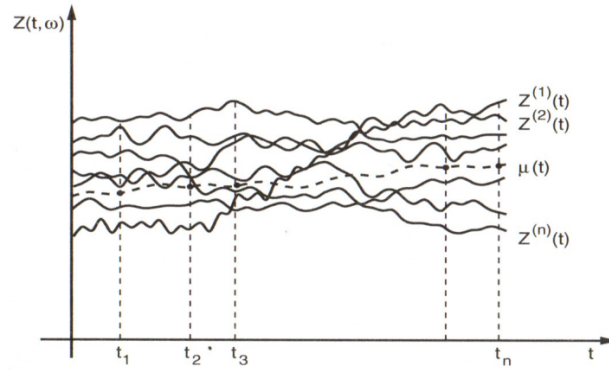
### 2.1 Processos Estocásticos

Podemos definir um processo estocástico como um conjunto cronológico de observações de um determinado fenômeno de forma que o seu comportamento pode ser descrito por uma ou mais distribuições de probabilidade. Conforme (MORETTIN; TOLOI, 2006),

**Definição 2.1.1.** Seja  $T$  um conjunto arbitrário, um processo estocástico é uma família de variáveis aleatórias  $\{Z(t), t \in T\}$ , tal que,  $\forall t \in T$ ,  $Z(t)$  é uma variável aleatória.

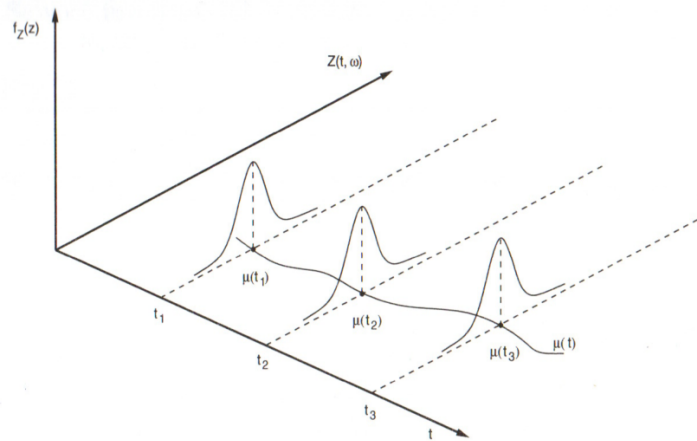
Neste caso, um processo estocástico é uma família de variáveis aleatórias, definidas num mesmo espaço de probabilidade  $(\Omega, \mathcal{A}, \mathcal{P})$ , tendo o conjunto  $T$ , normalmente tomado como conjunto dos inteiros  $\mathbb{Z} = \{0, \pm 1, \pm 2, \dots\}$  ou conjunto dos reais  $\mathbb{R}$ .

Figura 3 – Família de trajetórias de processos estocásticos.



Fonte: (MORETTIN; TOLOI, 2006)

Figura 4 – Um processo estocástico interpretado como uma família de variáveis.



Fonte: (MORETTIN; TOLOI, 2006)

A Figura 4 ilustra uma interpretação de um processo estocástico, ou seja, para  $t \in T$ ,  $Z(t)$  é uma variável aleatória definida em  $\Omega$ , e  $Z(t, \omega)$  é uma função de dois argumentos  $Z(t, \omega)$ ,  $t \in T$ ,  $\omega \in \Omega$ . Na figura, cada  $t \in T$ , temos uma v.a.  $Z(t, \omega)$ , com uma distribuição de probabilidade, é possível que a função densidade de probabilidade (fdp) no instante  $t_1$  seja distinta da fdp no instante  $t_2$ , entre os instante  $t_1$  e  $t_2$  quaisquer, mas usualmente é aquela em que a fdp de  $Z(t, \omega)$  é a mesma,  $\forall t \in T$ .

## 2.2 Modelos ARIMA

Segundo (MARTINS; WERNER, 2014) a metodologia Box-Jenkins possibilita detectar distintos modelos ARIMA e entre eles: o modelo Autoregressivo de ordem  $p$  (AR), o modelo de Médias Móveis de ordem  $q$  (MA), o modelo Autoregressivo de Médias Móveis

de ordem  $p$  e  $q$  (ARMA), o modelo Autoregressivo Integrado de Médias Móveis, de ordem  $p$ ,  $q$  contendo diferença  $d$  (ARIMA) que, apresenta como o principal caso contido nessa técnica. Segundo Ramser et al. (2015), eles apresentam conduta estacionária nos modelos AR( $p$ ) e MA( $p$ ), assim como a junção destes, o modelo ARMA( $p, q$ ).

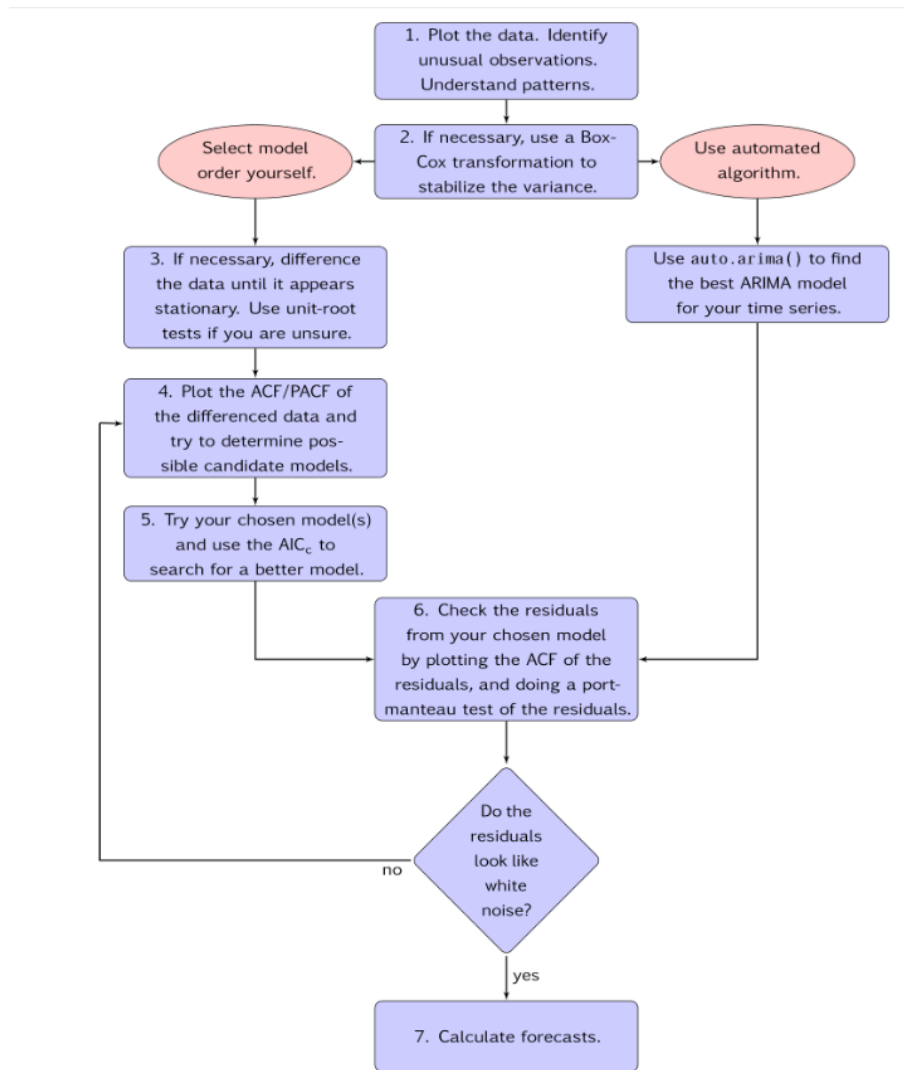
O modelo ARIMA ( $p, d, q$ ) é o efeito do uso de diferenciação na séries onde a média é variante, transformando-os estacionárias. Os modelos estacionários e os não estacionários, diferenciam-se pela quantidade de diferenças necessárias para estabilizar a série, simbolizado pela letra  $d$  (MARTINS; WERNER, 2014).

Ao ajustar um modelo ARIMA a um conjunto de dados de séries temporais, o procedimento a seguir fornece uma abordagem geral útil (HYNDMAN; ATHANASOPOULOS, 2018).

- Plote os dados e identifique quaisquer observações incomuns;
- Se necessário, transforme os dados (usando uma transformação Box-Cox) para estabilizar a variação;
- Se os dados não forem estacionários, tome as primeiras diferenças dos dados até que os dados sejam estacionários;
- Examine o ACF/PACF: É um ARIMA( $p, d, 0$ ) ou ARIMA( $0, d, q$ ) modelo apropriado?;
- Experimente o(s) modelo(s) escolhido(s) e use o AICc para procurar um modelo melhor;
- Verifique os resíduos do modelo escolhido plotando o ACF dos resíduos e fazendo um teste portmanteau dos resíduos. Se eles não se parecerem com ruído branco, tente um modelo modificado;
- Quando os resíduos parecerem ruído branco, calcule as previsões.

Este processo é usado pela função "auto.arima" do pacote "forecast" Software R. Resumido na Figura 5.

Figura 5 – Processo geral de previsão usando um modelo ARIMA.



Fonte: (HYNDMAN; ATHANASOPOULOS, 2018)

## 3 Metodologia

Para obtenção das análises estatísticas dos dados, fez-se o uso do *software* **R** (versão 4.2.1). Com posse dos dados, foi realizada uma análise inicial dos dados para identificar o comportamento da série temporal e analisar suas características quanto tendência, sazonalidade, aleatoriedade e existência de ruídos.

O índice de Preços da *commodities* Soja foi adquirido no site do CEPEA (Centro de Estudos Avançados em Economia Aplicada), com intervalo de outubro de 1997 até outubro de 2022.

## 4 Resultados e Discussões

A Tabela 1 apresenta as principais medidas descritivas obtidas a partir da série mensal condizente ao indicador de preços em reais da soja (reais por saca de 60 kg) compreendido no período de outubro de 1997 a outubro de 2022, proporcionando uma visão panorâmica do comportamento da série em estudo. Com isso é possível observar-se que ao longo de 25 anos o menor valor do preço foi de R\$ 13,87 (agosto de 1998) e o maior valor R\$ 195,85 (março de 2022), além de algumas medidas que quantificam a dispersão dos dados, tais como o desvio-padrão  $sd = 40,91$  e o coeficiente de variação  $CV = 69,77$ .

Tabela 1 – Estatística Descritiva

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
13.87	31.34	46.8	58.63	72.91	195.8

A Figura 6 apresenta a série original do indicador de preço médio mensal em reais da soja no intervalo de outubro de 1997 a outubro de 2022. Observa-se que o comportamento da série apresenta uma tendência não uniforme em seus preços.

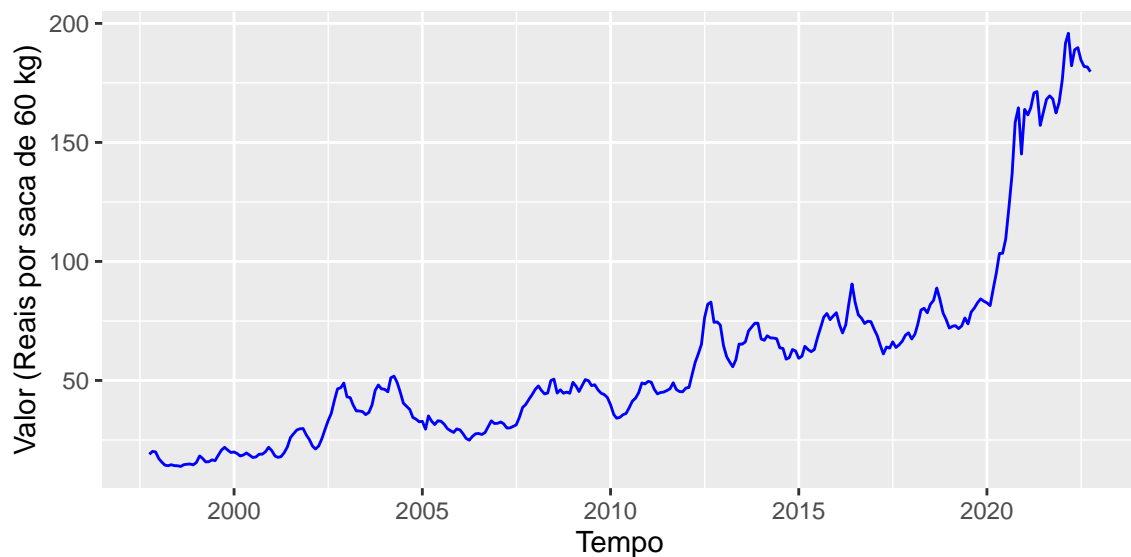


Figura 6 – Indicador de preço da soja no período de outubro de 1997 até outubro de 2022

Pode-se observar na apresentação da Figura 7 a decomposição da série em componentes: série original, parte aleatória, sazonalidade e tendência. Em vista disso, constata-se que a série oscila em torno de um valor médio do preço da soja e com um padrão recor-

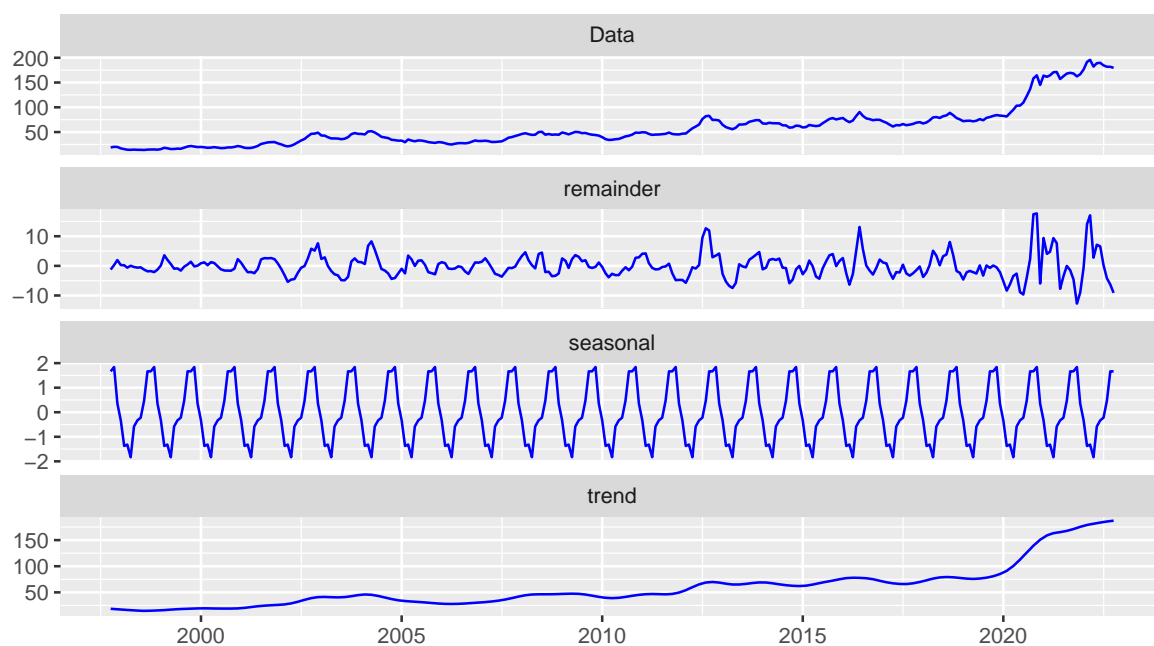


Figura 7 – Decomposição da série de preço da soja.

rente bem definido, o que deixa implícito a presença de sazonalidade aditiva e a existência de tendência nos dados.

Na Figura 10, é possível verificar o comportamento anual do preço médio mensal, composto de 301 observações, onde pode-se observar que os preços mais altos concentram nos meses de maio e junho e com mais variabilidade dos preços. Mas os preços médio em meses apresentaram comportamentos de variação similares entre si. Pode-se notar que todos os meses apresentaram no mínimo um valor atípico (outlier).



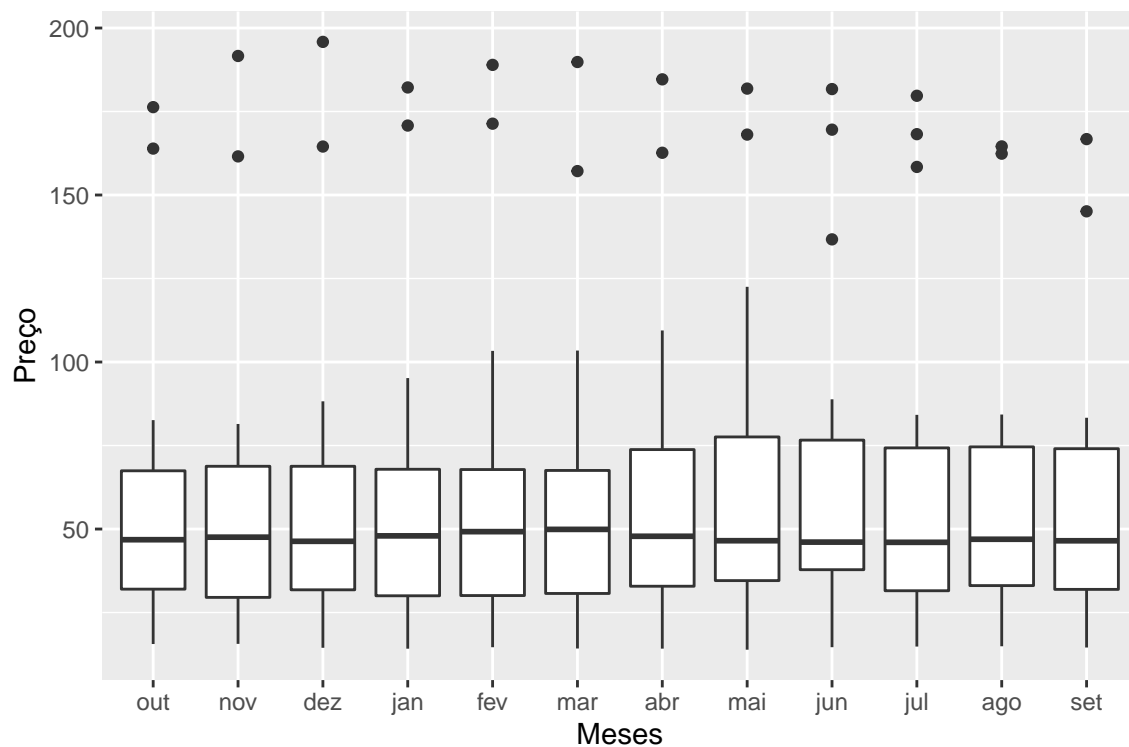


Figura 8 – Gráfico de Boxplot, da série distribuída entre os meses (período de 10/1997 a 10/2022).

As funções de autocorrelação (ACF) e autocorrelação parcial (PACF) da série, representadas na Figura 9, nota-se que ela decresce rapidamente, indicando a estacionariedade da série

Após ser aplicada a diferença, foi plotado novamente as funções para confirmar a eliminação da componente sazonal, e identificar as ordens de possíveis modelos **ARIMA** para a série. A Figura 10c, observa que ao aplicar a diferenciação, a série aparenta está estacionária na média, mas a variância é crescente ao longo do tempo. Como um dos pressupostos da teoria Box & Jenkins é que a série seja também estacionária na variância. Para tal, aplicaremos o logaritmo na série em questão. A Figura 10d nos mostra que temos a série estacionária na média e na variância.

Analisando os correlogramas anteriores, alguns coeficientes se mostraram não significativos. É claro que aplicando a diferenciação, a série torna-se estacionária. Para uma melhor verificação da função de autocorrelação e autocorrelação parcial, aplicaremos testes para verificar a estacionariedade da série, tais: o teste de *Dickey-Fuller Aumentado*, *Raiz Unitária*; teste de *Phillips-Perron*; teste de KPSS (Kwiatkowski, Philips, Schmidt e Shin).

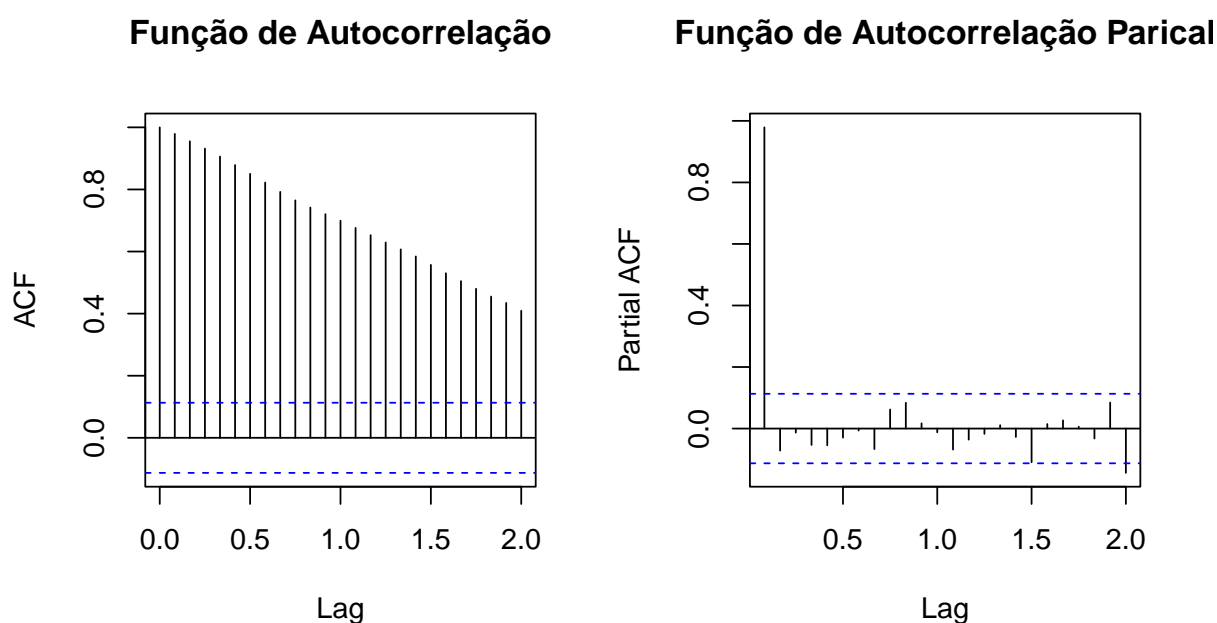


Figura 9 – Correlograma da série.

```
# Teste ADF.
adf.test(Soja.log.dif1)

##
## Augmented Dickey-Fuller Test
##
## data: Soja.log.dif1
## Dickey-Fuller = -7.6002, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary

# Teste de Phillips-Perron.
pp.test(Soja.log.dif1)

##
## Phillips-Perron Unit Root Test
##
## data: Soja.log.dif1
## Dickey-Fuller Z(alpha) = -186.81, Truncation lag parameter = 5, p-value
## = 0.01
## alternative hypothesis: stationary
```

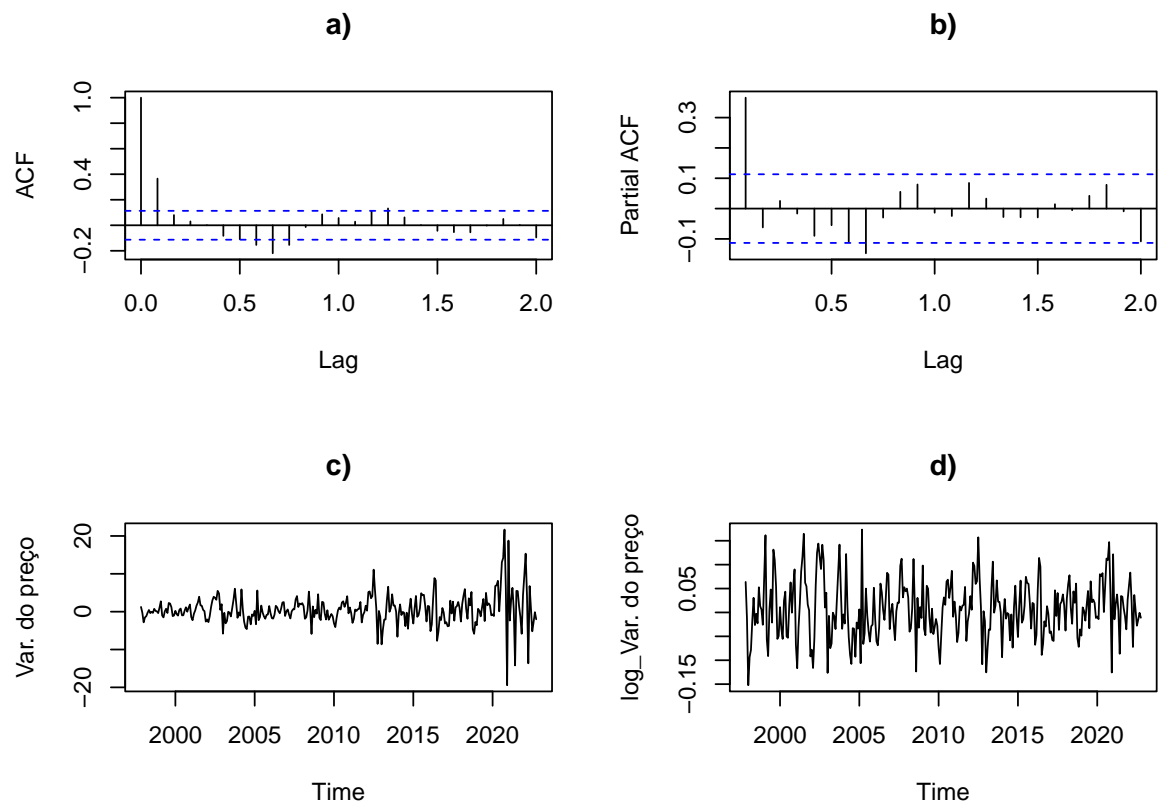


Figura 10 – Gráficos: a) Função de Autocorrelação da série logaritma diferenciada; b) Função de Autocorrelação Parcial da série logaritma diferenciada; c) Comportamento da variação da série diferenciada; e d) Logaritmo da diferenciação da série, para estabilizar a variância.

```
# Teste KPSS.
kpss.test(Soja.log.dif1, null = "Level")

##
## KPSS Test for Level Stationarity
##
## data: Soja.log.dif1
## KPSS Level = 0.050398, Truncation lag parameter = 5, p-value = 0.1
```

Anunciando nossa hipótese, a regra de decisão da rejeição ou não da hipótese nula ( $H_0$ ) é: se o *p-valor* é menor que o valor significativo, então rejeita-se  $H_0$ . Caso contrário, não rejeita-se  $H_0$ . Com isso, podemos verificar pelo teste que o *p-valor* é menor que o valor significativo de 5%. Apenas no teste de KPSS o anunciado das hipóteses são inversas, neste caso todos os testes indicam que a série é estacionária. Após a identificação das ordens de diferenciação sequencial e sazonal, tem-se um modelo com a seguinte característica

$\text{ARIMA}(p, 1, q)(P, 1, Q)$ .

$H_0$  : A série não é estacionária.

$H_1$  : A série é estacionária.

```
# Teste ADF.
adf.test(Soja.log.dif1)

##
## Augmented Dickey-Fuller Test
##
## data: Soja.log.dif1
## Dickey-Fuller = -7.6002, Lag order = 6, p-value = 0.01
## alternative hypothesis: stationary

# Teste de Phillips-Perron.
pp.test(Soja.log.dif1)

##
## Phillips-Perron Unit Root Test
##
## data: Soja.log.dif1
## Dickey-Fuller Z(alpha) = -186.81, Truncation lag parameter = 5, p-value
## = 0.01
## alternative hypothesis: stationary

# Teste KPSS.
kpss.test(Soja.log.dif1, null = "Level")

##
## KPSS Test for Level Stationarity
##
## data: Soja.log.dif1
## KPSS Level = 0.050398, Truncation lag parameter = 5, p-value = 0.1
```

Como é demasiado o serviço para selecionar valores apropriado para cada um desses três parâmetros, a função "forecast::auto.arima()", faz automaticamente.

```
# Estimativa do modelo ARIMA.
mod.arima <- auto.arima(Soja_log)
summary(mod.arima)

## Series: Soja_log
## ARIMA(1,1,0)(2,0,0)[12] with drift
##
## Coefficients:
##          ar1      sar1      sar2    drift
##          0.3645  0.0334  -0.1166  0.0076
## s.e.    0.0539  0.0585   0.0615  0.0047
##
## sigma^2 = 0.003132:  log likelihood = 440.99
## AIC=-871.97   AICc=-871.77   BIC=-853.45
##
## Training set error measures:
##                               ME          RMSE          MAE          MPE          MAPE          MASE
## Training set -4.198801e-05  0.05550114  0.04369263 -0.01376519  1.172031  0.2158349
##
##          ACF1
## Training set 0.02564065
```

O modelo resultante é um  $ARIMA(1,1,0)(2,0,0)$ , o que corresponde a uma combinação a uma série com um grau de diferenciação ( $p = 1$ ) aplicada em um modelo  $AR(1)$  e em um  $MA(0)$ . Antes de fazer a previsão, faremos uma análise de resíduos. Podemos verificar na Figura 11 que o gráfico ACF dos resíduos do modelo **ARIMA**(1,1,0)(2,0,0) mostra que todas as autocorrelações estão dentro dos limites, indicando que os resíduos estão se comportando como ruído branco. E o teste de Ljung-Box retorna um grande p-valor (maior do que o valor significativo de 5%), sugerindo também que os resíduos são ruído branco.

```
##
## Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,0)(2,0,0)[12] with drift
## Q* = 26.983, df = 21, p-value = 0.1714
##
## Model df: 3.    Total lags used: 24
```

Agora faremos a previsão da série, apresentado na Figura 12 seguinte.

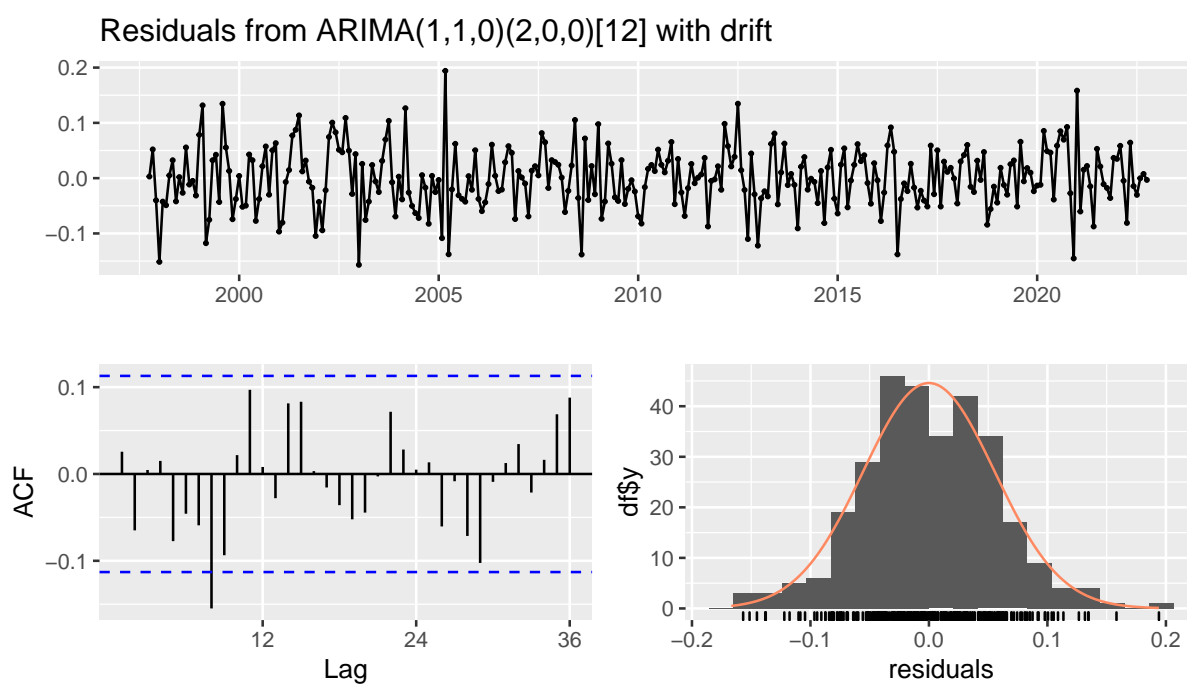


Figura 11 – Gráficos de resíduos para o modelo ARIMA(1,1,0)(2,0,0).

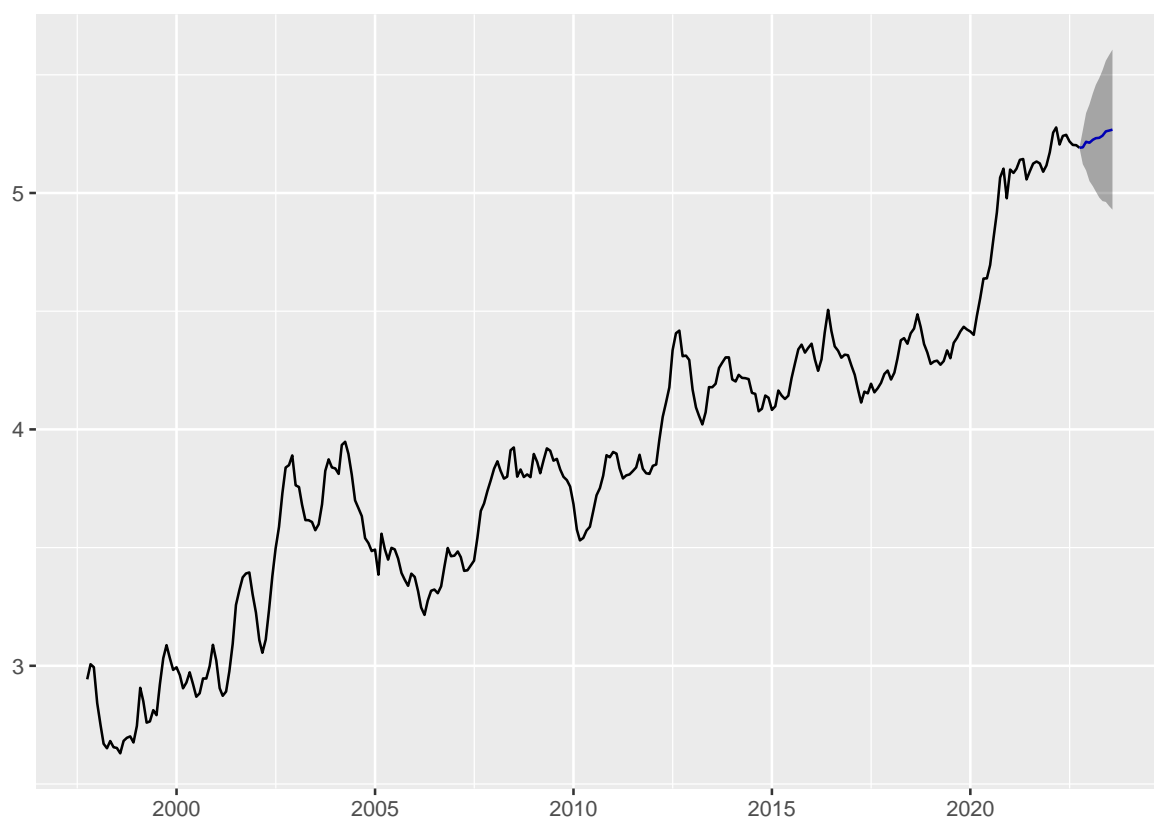


Figura 12 – Previsão para o índice de preço da soja.

## 5 Considerações Finais

O presente trabalho teve como objetivo estudar o comportamento do preço da Soja no Brasil, e testar a metodologia Box & Jenkins para previsão. Para isso foi utilizado uma função do pacote "forecast" do Software R para obter o modelo ideal para série, função está que é automatizada que reduz o número de combinações de modelos possíveis adequados. O modelo destacado foi o  $ARIMA(1,1,0)(2,0,0)$ . De posse deste modelo é possível inferir o preço futuro da saca de 60kg da soja comercializada no Brasil, possibilitando assim melhor planejamento da compra, venda e produção desse grão.

Algumas limitações podem ser revistas em estudos futuros, como: (a) relacionar outras técnicas de previsão; (b) relacionar eventos pontuais em cada ano e identificar o quanto influência no preço da soja. Este estudo pode contribuir para pesquisas acadêmicas de métodos quantitativos de previsão e para entender os valores futuros do produto para pretensão de aumentar seus lucros.

## Referências

CEPEA. 2022. Acesso em 26 de novembro de 2022. Disponível em: <<https://www.cepea.esalq.usp.br/br>>. Citado na página 7.

CONAB. 2022. Acesso em 26 de novembro de 2022. Disponível em: <<https://www.conab.gov.br/busca?searchword=soja&searchphrase=all>>. Citado na página 8.

EMBRAPA. 2022. Acesso em 26 de novembro de 2022. Disponível em: <<https://www.embrapa.br/soja>>. Citado na página 7.

FAOSTAT. 2022. Acesso em 26 de novembro de 2022. Disponível em: <[www.fao.org/faostat/en/#data/QCL/visualize](http://www.fao.org/faostat/en/#data/QCL/visualize)>. Citado 2 vezes nas páginas 7 e 8.

HYNDMAN, R. J.; ATHANASOPOULOS, G. *Forecasting: principles and practice*. [S.l.]: OTexts, 2018. Citado 2 vezes nas páginas 11 e 12.

MARTINS, V. L. M.; WERNER, L. Comparison of individual forecasts and their combinations: a study with industrial series. *Production*, SciELO Brasil, v. 24, p. 618–627, 2014. Citado 2 vezes nas páginas 10 e 11.

MORETTIN, P.; TOLOI, C. Análise de séries temporais 2ª edição revista e ampliada. *ABE-Projeto Fisher*, Editora Edgar Blücher, 2006. Citado 2 vezes nas páginas 9 e 10.



## Apêndices

# APÊNDICE A – SCRIPT R

```
library(forecast) # Ferramentas para exibir e analisar previsões de séries temporais univariadas.
library(timeSeries) # Ferramentas para séries temporais financeiras.
library(urca) # Testes de raiz unitária e de cointegração.
library(randtests) # Testes de aleatoriedade não paramétricos para sequências numéricas.
library(tseries) # Análise de séries temporais e finanças computacionais.
library(stats) # Funções para cálculos estatísticos e geração de números aleatórios.
library(readxl) # Importe arquivos Excel para R.
library(tidyverse) # Conjunto de pacotes que funcionam em harmonia para data analysis.
library(ggfortify) # Ferramentas de plotagem unificadas.
library(magrittr) # Fornece o encadeamento de comandos, pipeline (%>%).
library(kableExtra) # Gerador de tabelas.

soja <- read_xls("../dataset/soja.xls") # Importar o conjunto de dados.

soja <- ts(soja$Preço, start = c(1997,10), frequency = 12) # Criar o objeto de série temporal.

class(soja) # Retornar os valores do atributo de classe de um objeto.
as.ts(soja) # Testar se o objeto é um série temporal.

### Informações descritiva da série (resumo).
summary(soja)

### Plotagem da série original.
autoplot(soja, xlab = "Tempo", ylab = "Valor (Reais por saca de 60 kg)", ts.colour = 'blue')

### Decomposição da série e plotagem.
soja %>%
  stl(s.window = "periodic") %>%
  autoplot(ts.colour = "blue")

###
```