UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

CURSO DE GRADUAÇÃO EM ESTATÍSTICA

DOUGLAS CHAVES MOURA

Existe um Ciclo Anual na Produção de Petróleo no Brasil? Uma Análise de Dados de 2001 a 2025

FORTALEZA

2025

DOUGLAS CHAVES MOURA

Existe um Ciclo Anual na Produção de Petróleo no Brasil? Uma Análise de Dados de 2001 a 2025

Trabalho apresentado à disciplina Análise de Séries Temporais, do curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como parte dos requisitos para a avaliação da referida disciplina, no semestre 2025.1.

Professora: Drª Jeniffer Johana Duarte Sanchez

FORTALEZA

2025

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

O estudo de séries temporais constitui um pilar fundamental da análise de dados, sendo aplicado para investigar qualquer variável de interesse observada em múltiplos instantes ao longo do tempo. Seja no setor financeiro, ao avaliar os preços de um ativo, ou no setor industrial, ao monitorar a produção de uma fábrica, a análise de séries temporais é uma ferramenta essencial para a inteligência de negócios. Sua força reside na capacidade de não apenas decodificar as características e o comportamento de uma variável no passado, mas, sobretudo, de desenvolver modelos para a previsão de valores futuros, subsidiando decisões mais estratégicas e informadas.

Alinhado a essa perspectiva, o presente trabalho se dedica a analisar a dinâmica da produção mensal de petróleo (em m^3) no Brasil, compreendendo o período de janeiro de 2001 a maio de 2025. O estudo se baseia nos dados públicos disponibilizados pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) através do Portal de Dados Abertos do Governo Federal, buscando identificar padrões, tendências e sazonalidades que caracterizam a evolução de um dos setores mais estratégicos para a economia nacional.

2 Metodologia

Como parte do estudo, foi desenvolvida uma análise exploratória da série a fim de se obter um entendimento inicial de seus padrões. Para a modelagem preditiva, foi adotada a metodologia de Box-Jenkins, desenvolvida por George Box e Gwilym Jenkins na década de 1970 (BOX; JENKINS, 1970). O método Box-Jenkins consiste em um conjunto de procedimentos sistemáticos para identificar, estimar e validar modelos de séries temporais.

A metodologia foca especialmente nos modelos ARIMA (AutoRegressive Integrated Moving Average), que combinam três componentes:

- **AR (Auto-Regressive)**: Modelo autorregressivo, que utiliza a relação entre uma observação e um número de observações defasadas.
- **I (Integrated)**: Representa o número de diferenciações aplicadas para tornar a série estacionária.
- **MA (Moving Average)**: Modelo de médias móveis, que utiliza a dependência entre uma observação e o erro residual de um modelo de média móvel aplicado a observações defasadas.

O método original segue uma abordagem de modelagem iterativa em três estágios:

1 **Identificação e Seleção do Modelo**: Nesta etapa, verifica-se a estacionariedade da série e identifica-se a presença de sazonalidade. Por meio da análise gráfica das Funções de Autocorrelação (Autocorrelation Function - ACF) e Autocorrelação Parcial (Partial Autocorrelation Function - PACF), decide-se a ordem dos componentes autorregressivo (p) e de médias móveis (q) a serem incluídos no modelo.

2 **Estimação dos Parâmetros**: Utiliza-se algoritmos computacionais para estimar os coeficientes que melhor se ajustam ao modelo ARIMA selecionado. Os métodos mais comuns são a estimação por máxima verossimilhança ou por mínimos quadrados não lineares.

3 **Verificação do Modelo**: Realiza-se um diagnóstico para testar se o modelo estimado atende aos pressupostos de um processo estacionário. Em particular, os resíduos do modelo devem se comportar como um ruído branco, ou seja, ser independentes e identicamente distribuídos, com média zero e variância constante. Gráficos dos resíduos e testes estatísticos, como o de Ljung-Box, são aplicados. Se o modelo for considerado inadequado, retorna-se à primeira etapa para refinar a identificação.

4 **Previsão**: Uma vez validado, o modelo é utilizado para fazer projeções de valores futuros da série.

A base teórica deste trabalho foi fundamentada principalmente na obra “Análise de Séries Temporais” de Morettin e Toloi (MORETTIN; TOLOI, 2006). Toda a aplicação foi desenvolvida no software R (Versão 4.4.2) (R Core Team, 2024), utilizando a IDE (Integrated Development Environment) RStudio (Versão 2024.12.0) (Posit team, 2024).

3 Análise Exploratória

3.1 Descrição e Visualização dos dados

A série temporal em estudo refere-se à quantidade do volume produzido de petróleo - em metros cúbicos (m^3) - no Brasil com frequência mensal, entre janeiro de 2001 e maio de 2025, totalizando 293 observações. Trata-se, portanto, de uma série temporal univariada e discreta. Algumas estatísticas descritivas da série são apresentadas na tabela 3.1.

Tabela 3.1: Estatísticas descritivas da série mensal de produção de petróleo no Brasil (em m^3).

Variável	**Média**	**DP**	**EP**	**Mediana**	**Assimetria**	**Curtose**	**Min**	**Max**	*n*
Produção	10.927.721	3.072.831	179.516,7	10.236.311	0,39	-0,84	5.806.609	18.132.705	293

O volume de produção de petróleo no período analisado foi de aproximadamente 10.927.721 m^3 por mês, com um desvio padrão de 3.072.831 m^3, indicando variação significativa ao longo do tempo. Os volumes mensais oscilaram entre 5.806.609 m^3 e 18.132.705 m^3, com uma mediana de 10.236.311 m^3.

Considerando que a série em estudo possui características cíclicas anuais, é apresentada na tabela 3.2 as estatísticas descritivas mensalmente desagregadas do volume de petróleo produzido no Brasil em metros cúbicos (m^3).

Tabela 3.2: Estatísticas descritivas mensais da produção de petróleo no Brasil (em m^3), com base na série histórica entre jan/2001 e mai/2025.

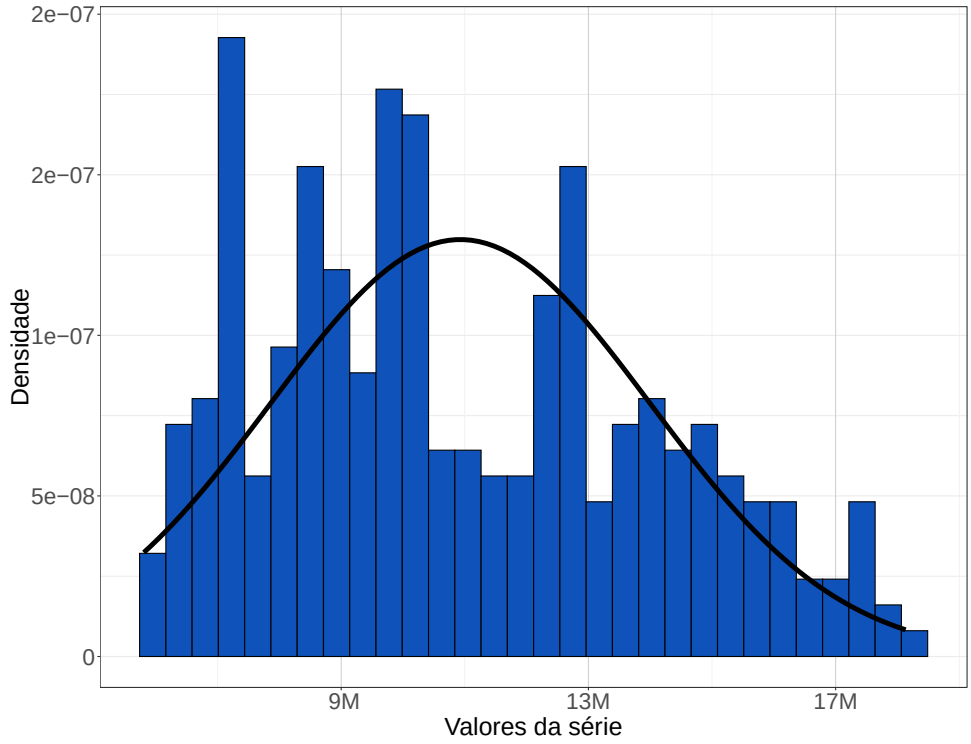
Mês	**Média**	**DP**	**EP**	**Mediana**	**Assimetria**	**Curtose**	**Min**	**Max**	*n*
Janeiro	11.204.832	3.309.201	661.840,2	10.459.863	0,34	-1,18	6.498.405	17.344.377	25
Fevereiro	10.104.573	2.918.000	583.600,0	9.302.576	0,41	-1,00	5.971.240	15.898.965	25
Março	11.004.393	3.177.167	635.433,3	10.287.594	0,44	-0,92	6.211.220	17.845.551	25
Abril	10.710.849	3.028.075	605.614,9	9.907.401	0,41	-0,96	6.052.810	17.323.715	25
Maio	11.152.062	3.191.202	638.240,3	10.227.859	0,35	-0,87	5.920.527	18.132.705	25
Junho	10.608.091	2.912.820	594.576,9	10.106.700	0,30	-1,01	6.173.469	16.272.346	24
Julho	11.076.293	3.087.005	630.132,3	10.183.902	0,33	-1,14	6.428.124	17.312.824	24
Agosto	11.173.941	3.164.227	645.895,1	10.178.062	0,31	-1,25	6.338.025	17.063.068	24
Setembro	10.854.486	3.151.938	643.386,6	9.999.890	0,45	-0,95	6.223.285	17.515.200	24
Outubro	11.116.218	3.169.541	646.979,8	10.309.414	0,28	-1,05	5.806.609	17.463.214	24
Novembro	10.793.576	3.078.555	628.407,4	10.201.005	0,39	-0,89	6.384.420	17.544.090	24
Dezembro	11.352.584	3.158.622	644.751,0	10.828.356	0,25	-1,05	6.616.203	17.669.743	24

Janeiro se destacou com uma média de produção de 11.204.832 m³ e o maior desvio padrão entre os meses (3.309.201 m^3), indicando uma variabilidade substancial na produção inicial do ano. Em maio, registrou-se a maior produção de um único mês durante todo o período analisado. Já em outubro, observa-se o menor valor de produção mensal.

No geral, a assimetria leve e a curtose negativa sugerem que, embora haja uma distribuição relativamente equilibrada dos valores mensais de produção ao longo dos anos, há uma quantidade não desprezível de valores mais extremos que puxam esse balanço.

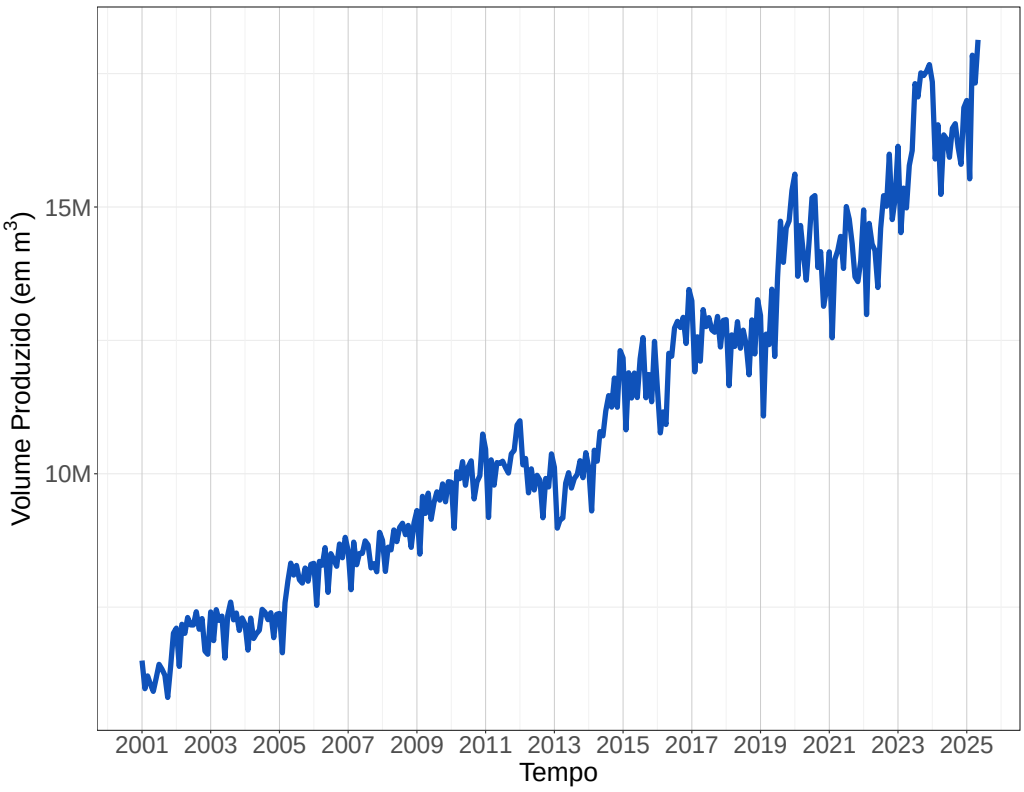
A figura 3.1 ilustra a densidade dos dados da série, destacando os padrões de distribuição.

Figura 3.1: Distribuição da densidade da produção mensal de petróleo no Brasil entre jan/2001 e mai/2025.



Prosseguindo, a figura 3.2 apresenta o gráfico da série temporal da produção de petróleo.

Figura 3.2: Série temporal da produção mensal de petróleo no Brasil (jan/2001 – mai/2025).



O gráfico revela alguns pontos interessantes. Visualmente, é nítida a ausência de estacionariedade, marcada por uma clara tendência de crescimento e pela presença de sazonalidade. A tendência mostra um aumento quase contínuo do volume produzido ao longo dos anos. As flutuações regulares sugerem a presença de um componente sazonal. Em particular, observa-se um crescimento acentuado entre 2019 e 2020, seguido de uma queda, e outro avanço expressivo entre 2021 e 2024. Esses movimentos podem estar associados a fatores externos ou choques específicos, como variações de demanda, alterações regulatórias ou eventos geopolíticos. Algumas hipóteses:

- Impactos da pandemia de COVID-19 a partir de 2020, que afetaram cadeias logísticas, demanda internacional e decisões de produção em diversos países, inclusive no Brasil.
- Aumento da produção em função da entrada em operação de novos campos do pré-sal e políticas de incentivo à exploração entre 2021 e 2024, impulsionadas por preços internacionais mais atrativos e mudanças regulatórias no setor.

As figuras 3.3 e 3.4 apresentam os gráficos de boxplot e violino, respectivamente, para visualizar a distribuição mensal dos dados de produção.

Figura 3.3: Boxplot da produção mensal de petróleo por mês.

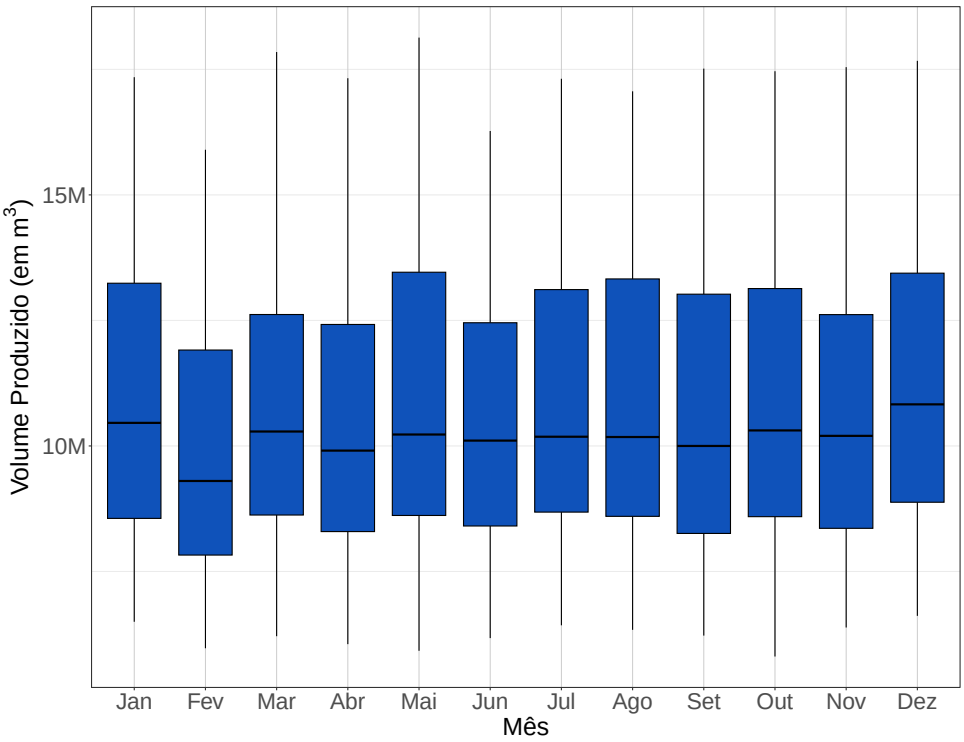
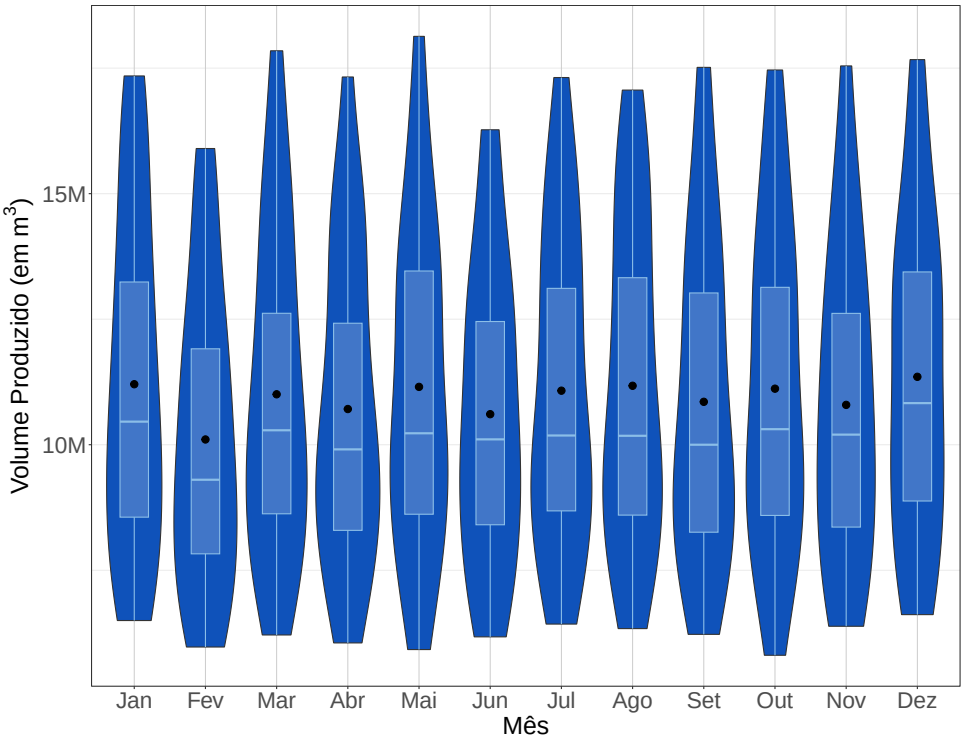


Figura 3.4: Distribuição da produção mensal de petróleo por mês — gráfico de violino.

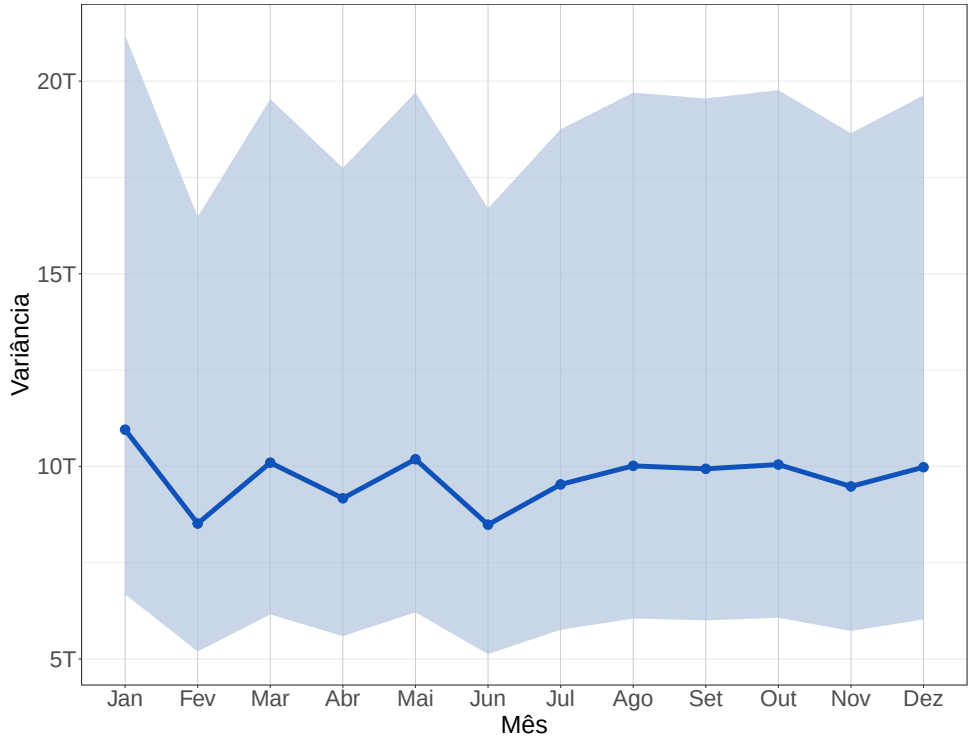


Os meses exibem medianas relativamente próximas, com exceção de fevereiro, que possui a menor mediana. Conforme mostrado nos gráficos, há uma maior concentração de dados em

valores mais baixos de produção de petróleo, enquanto valores mais elevados tendem a elevar a média. Não foram identificadas observações discrepantes significativas em nenhum dos meses.

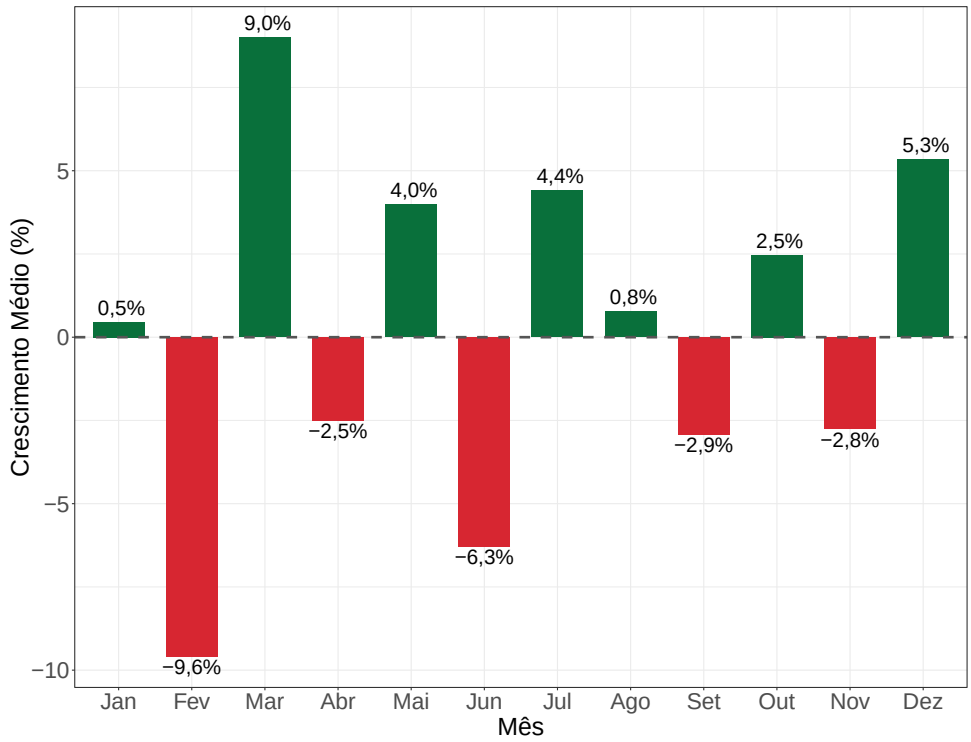
A figura 3.5 ilustra a variância mensal do volume produzido de petróleo, incluindo o intervalo de confiança de 95%. Observa-se que janeiro apresenta a maior variabilidade, sugerindo maior incerteza na produção. De fevereiro a junho, há um padrão de aumentos e quedas na variância. Nos meses de julho a dezembro, a variância demonstra um comportamento mais uniforme.

Figura 3.5: Variância da produção mensal de petróleo por mês do ano com intervalo de confiança de 95%.



O gráfico da variação percentual média, apresentado na Figura 3.6, ilustra as oscilações mensais no volume de produção em relação ao mês anterior, evidenciando padrões sazonais relevantes. Destaca-se o mês de fevereiro, que apresenta a maior queda média (-9,6%), uma retração que pode ser majoritariamente atribuída ao menor número de dias do mês. Abril, junho, setembro e novembro também registram variações médias negativas. Em contraste, um destaque positivo vai para março, que historicamente apresenta o maior aumento percentual médio (9,0%), recuperando a queda de fevereiro.

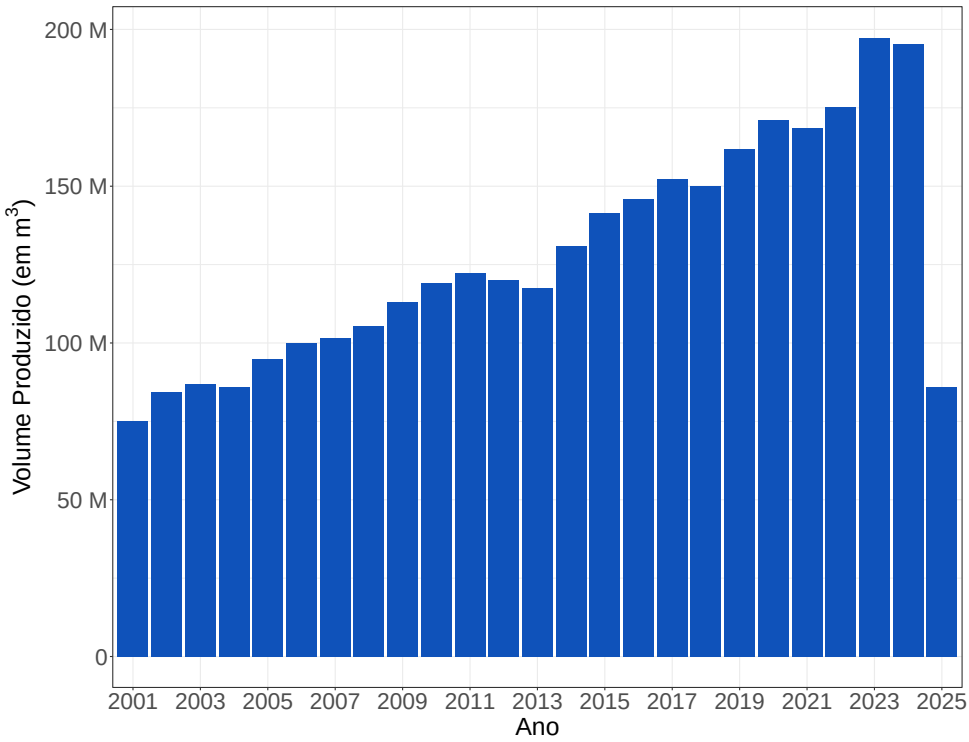
Figura 3.6: Variação percentual média da produção de petróleo em relação ao mês anterior.



A presença de um padrão sazonal se torna mais evidente na oscilação entre fevereiro e março, como observado também nos gráficos anteriores. Essa dinâmica de queda e recuperação no início do ano é uma das características mais marcantes da sazonalidade da série.

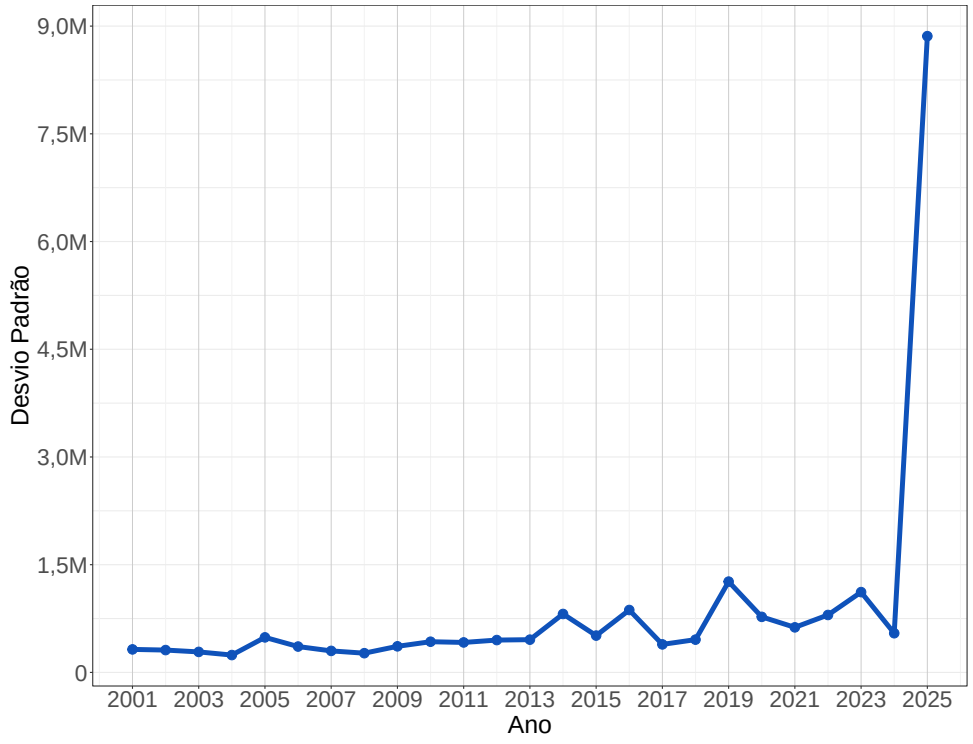
O gráfico de produção total anual (Figura 3.7) oferece uma visão macro da evolução do volume de petróleo produzido ao longo dos anos. Observa-se um crescimento quase contínuo, com os maiores picos atingidos em 2023 e 2024. O ano de 2025 apresenta um panorama parcial, pois os dados disponíveis vão apenas até maio.

Figura 3.7: Volume total anual de produção de petróleo no Brasil entre 2001 e 2025.



A Figura 3.8 mostra o desvio padrão anual, representando a volatilidade do volume de petróleo produzido dentro de cada ano.

Figura 3.8: Desvio padrão anual da produção de petróleo.

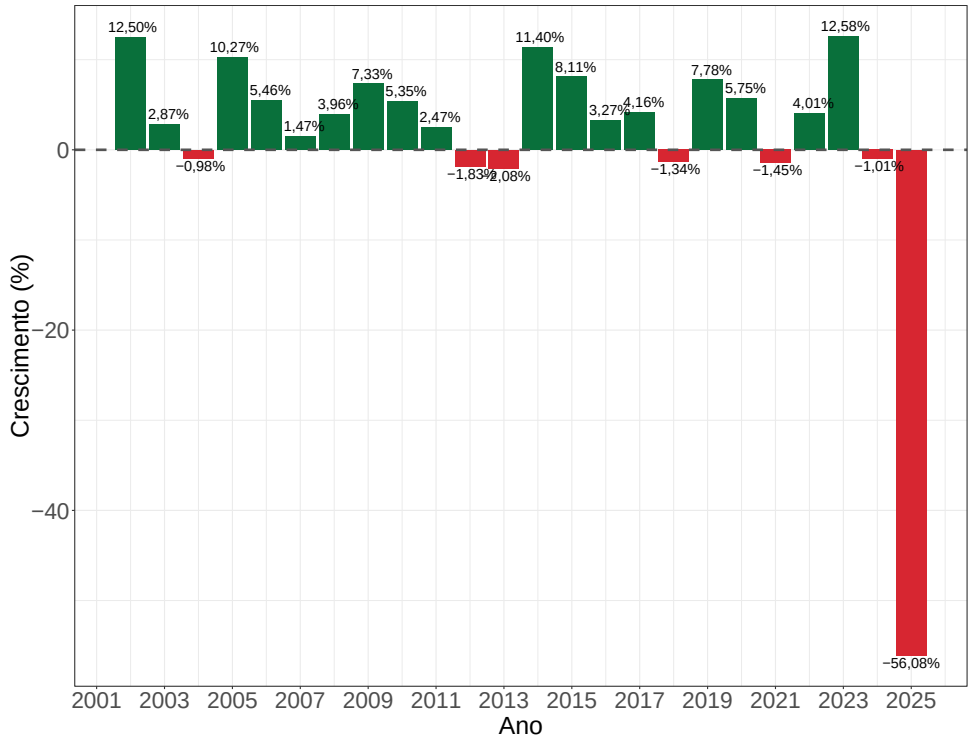


Observa-se um comportamento relativamente estável na volatilidade da produção entre

os anos de 2001 e 2013. A partir desse período, a volatilidade intra-anual começa a apresentar maiores oscilações, com aumentos e diminuições mais pronunciados. No entanto, entre 2024 e 2025, há um aumento significativo do desvio padrão (de 545.662 para 8.860.164), possivelmente devido à disponibilidade de dados apenas parciais para 2025.

Finalmente, a figura 3.9 apresenta o gráfico de crescimento anual da produção em comparação ao ano anterior, uma métrica similar à variação mês a mês, porém em escala anual.

Figura 3.9: Variação percentual anual da produção de petróleo em relação ao ano anterior.



O gráfico confirma que, de modo geral, houve um aumento do volume produzido ao longo do período, com a maioria dos anos apresentando crescimento positivo. Os poucos anos de queda foram 2004, 2012, 2013, 2018, 2021 e 2024. Dentre esses, a maior retração foi em 2013 (-2,1%). É importante destacar novamente que 2025 conta apenas com dados parciais, resultando atualmente em uma queda de crescimento de -56%.

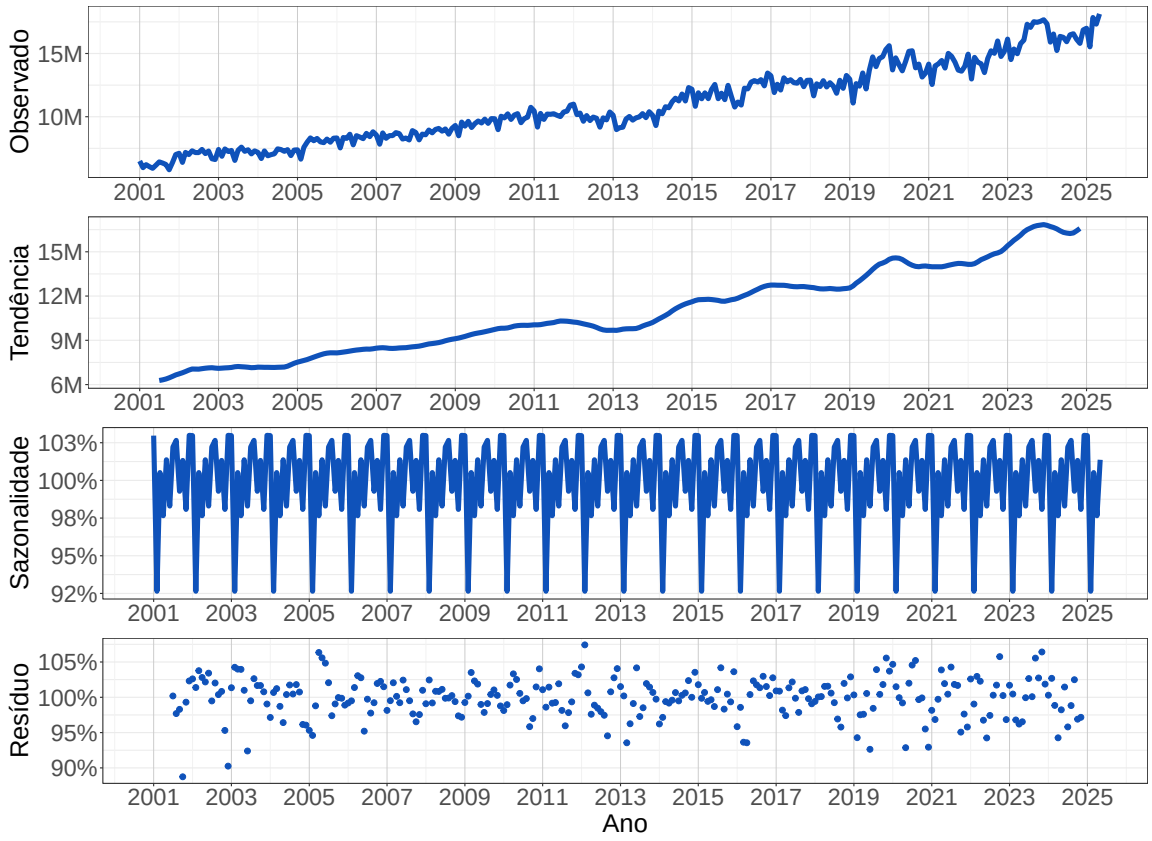
3.2 Decomposição da Série

Para analisar os componentes da série de forma isolada, realizou-se sua decomposição. Optou-se por um modelo multiplicativo, pois, como visto na análise exploratória, a magnitude das flutuações sazonais aparenta crescer com o nível da série. A estrutura do modelo de decomposição multiplicativa é dada por:

$$Y_t = T_t \times S_t \times \varepsilon_t$$

sendo Y_t o valor observado da série no tempo t; T_t o componente de Tendência; S_t o componente Sazonal; e ε_t o componente residual (aleatório), que representa o que sobra após a extração da tendência e da sazonalidade.

Figura 3.10: Decomposição multiplicativa da série de produção de petróleo: tendência, sazonalidade e resíduos.



A partir da Figura 3.10, é possível visualizar cada componente de forma clara. O segundo painel isola a tendência de longo prazo, confirmando a trajetória de crescimento já pontuada em análises anteriores. O terceiro painel revela um padrão sazonal bem definido e repetitivo ao longo dos anos. Por fim, o último painel mostra os resíduos, que, idealmente, deveriam se comportar como um ruído branco. Observa-se que eles aparentam ser aleatórios na maior parte do tempo, mas com uma dispersão um pouco maior nos extremos da série (nos anos iniciais e finais).

4 Pré-processamento e Estacionariedade da Série

A metodologia de Box-Jenkins, pilar deste estudo, exige como premissa fundamental que a série temporal a ser modelada seja estacionária. Uma série é dita estacionária quando suas propriedades estatísticas — como média, variância e autocorrelação — não variam ao longo do tempo, dependendo apenas da defasagem entre os períodos. A presença de tendência ou sazonalidade, como as visualmente identificadas no capítulo anterior, são as violações mais diretas da premissa de estacionariedade.

Neste capítulo, realizaremos a etapa de pré-processamento dos dados. O objetivo é, primeiramente, confirmar formalmente a não estacionariedade da série original através de testes de hipóteses e, em seguida, aplicar as transformações necessárias para remover a tendência e a sazonalidade, obtendo assim uma série estacionária apta para a etapa de identificação do modelo.

4.1 Testes Formais na Série Original

A análise exploratória visual forneceu indícios de que a série do volume de petróleo produzido no Brasil não é estacionária. Contudo, para validar essas ideias iniciais, é fundamental aplicar os respectivos testes de hipóteses. Esta seção, portanto, se dedica a submeter as observações visuais a um rigor estatístico, aplicando testes de hipóteses formais para verificar a presença de não estacionariedade, tendência e sazonalidade na série original.

4.1.1 Teste para Estacionariedade

Para investigar a estacionariedade, foram aplicados os testes de Dickey-Fuller Aumentado (ADF) e Kwiatkowski-Phillips-Schmidt-Shin (KPSS). Utilizar ambos é uma boa prática, pois eles partem de hipóteses nulas opostas, oferecendo uma confirmação mútua dos resultados.

O teste ADF é um teste de raiz unitária. Uma raiz unitária é uma característica de processos estocásticos que indica que os choques na série têm um efeito permanente, não se revertendo à média, o que é uma causa de não estacionariedade.

O teste foi aplicado à série original, obtendo-se o seguinte resultado:

TESTE DE ESTACIONARIEDADE - ADF

```
> # Teste de Dickey-Fuller Aumentado (ADF)
> # H0: a série possui raiz unitária (é não estacionária)
> # H1: a série  não possui raiz unitária (é estacionária)
> adf.test(serie_petroleo)

  Augmented Dickey-Fuller Test

data:  serie_petroleo
Dickey-Fuller = -3.0453, Lag order = 6, p-value = 0.1357
alternative hypothesis: stationary
```

Considerando um nível de significância usual de $\alpha = 5\%$, não há evidências estatísticas para rejeitar a hipótese nula. Portanto, o teste ADF indica que a série do volume de petróleo produzido possui uma raiz unitária, confirmando sua característica de **não estacionariedade**.

O teste KPSS, por sua vez, funciona de maneira complementar ao ADF. Ele testa se a série é estacionária em torno de uma média ou de uma tendência. A versão do teste aplicada aqui, conhecida como KPSS Level, avalia especificamente a estacionariedade em torno de um nível (média) constante.

TESTE DE ESTACIONARIEDADE - KPSS

```
> # Teste de Kwiatkowski-Phillips-Schmidt-Shin (KPSS)
> # H0: a série é estacionária em torno de uma média ou
   tendência
> # H1: a série possui raiz unitária (não estacionária)
> kpss.test(serie_petroleo)

  KPSS Test for Level Stationarity

data:  serie_petroleo
KPSS Level = 4.7847, Truncation lag parameter = 5, p-value
   = 0.01
```

Sendo o valor-p inferior ao nível de significância, rejeita-se a hipótese nula. A convergência de ambos os testes fornece uma forte evidência de que **a série original é não estacionária**.

4.1.2 Teste para Tendência

A presença de tendência é uma das principais razões para a não estacionariedade. Para confirmar formalmente a tendência crescente observada na análise visual, foram aplicados o teste não paramétrico de Mann-Kendall e um modelo de regressão linear.

O teste de Mann-Kendall é um teste não paramétrico utilizado para detectar a presença de uma tendência monotônica (consistentemente crescente ou decrescente, não necessariamente linear) em uma série temporal.

TESTE DE TENDÊNCIA - Mann-Kendall

```
> # Teste de Mann-Kendall
> # H0: não há tendência monotônica na série
> # H1: existe tendência monotônica
> MannKendall(serie_petroleo)
tau = 0.878, 2-sided pvalue =< 2.22e-16
```

A estatística de teste $\tau = 0,878$ foi positiva, o que sugere uma tendência crescente. Já o valor-p é significativamente inferior ao nível de significância $\alpha = 5\%$. Assim, rejeita-se a hipótese nula, concluindo que **há uma tendência monotônica** positiva estatisticamente significativa na série.

Para avaliar a presença de uma tendência de caráter linear, ajustou-se um modelo de regressão simples, tendo a produção de petróleo como variável resposta e o tempo como variável preditora. O teste de interesse avalia se o coeficiente angular é significativamente diferente de zero.

TESTE DE TENDÊNCIA - Regressão Linear Simples

```
> # Teste de Regressão Linear para Tendência
> # H0: coeficiente de tendência = 0 (sem tendência)
> # H1: coeficiente de tendência != 0 (com tendência)
> tempo <- 1:length(serie_petroleo)
> modelo_tendencia <- lm(serie_petroleo ~ tempo)
> summary(modelo_tendencia)

Call:
lm(formula = serie_petroleo ~ tempo)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-2344569  -402645    37009   391853  2201805

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5753985.4    87033.0   66.11   <2e-16 ***
tempo         35195.5      513.2   68.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
   ' 1

Residual standard error: 743000 on 291 degrees of freedom
Multiple R-squared:  0.9417,  Adjusted R-squared:  0.9415
F-statistic:  4704 on 1 and 291 DF,  p-value: < 2.2e-16
```

O coeficiente estimado para a variável `tempo` foi de 35.195,5, indicando um crescimento médio de aproximadamente 35,2 mil m^3 no volume de produção a cada mês. O valor-p para este coeficiente foi inferior a $\alpha = 5\%$. Portanto, rejeita-se a hipótese nula, **confirmando a presença de uma tendência linear positiva** e significativa.

4.1.3 Teste para Sazonalidade

Finalmente, para validar a impressão visual de sazonalidade, foi aplicado o teste QS (*Quadratic Spectral*), que avalia a presença de sazonalidade determinística. Para isso, a frequência da série foi definida como 12, correspondendo ao ciclo mensal-anual.

TESTE DE SAZONALIDADE – QS

```
> # Teste QS (Quadratic Spectral) para Sazonalidade
> # H0: não há sazonalidade estável na série
> # H1: existe sazonalidade estável na série
> qs(serie_petroleo, freq = frequency(serie_petroleo))
Test used:  QS

Test statistic:  258.08
P-value:  0
```

O valor-p do teste foi extremamente baixo (zero), levando a uma forte rejeição da

hipótese nula. Isso confirma formalmente o que a análise exploratória e a decomposição da série sugeriram: **a série do volume de petróleo produzido no Brasil possui um componente sazonal forte** e estatisticamente significativo.

4.2 Transformação Logarítmica

A análise exploratória (Figura 3.8) revelou que a variabilidade da série não é constante, tendendo a aumentar conforme o nível da produção cresce. Uma das premissas dos modelos ARIMA é a estabilidade da variância (homocedasticidade), e a transformação logarítmica é uma técnica clássica para estabilizar essa variância.

O objetivo desta transformação não é remover a tendência ou a sazonalidade, mas sim tornar as flutuações da série mais homogêneas ao longo do tempo.

ANÁLISE DE VARIABILIDADE - Série Original

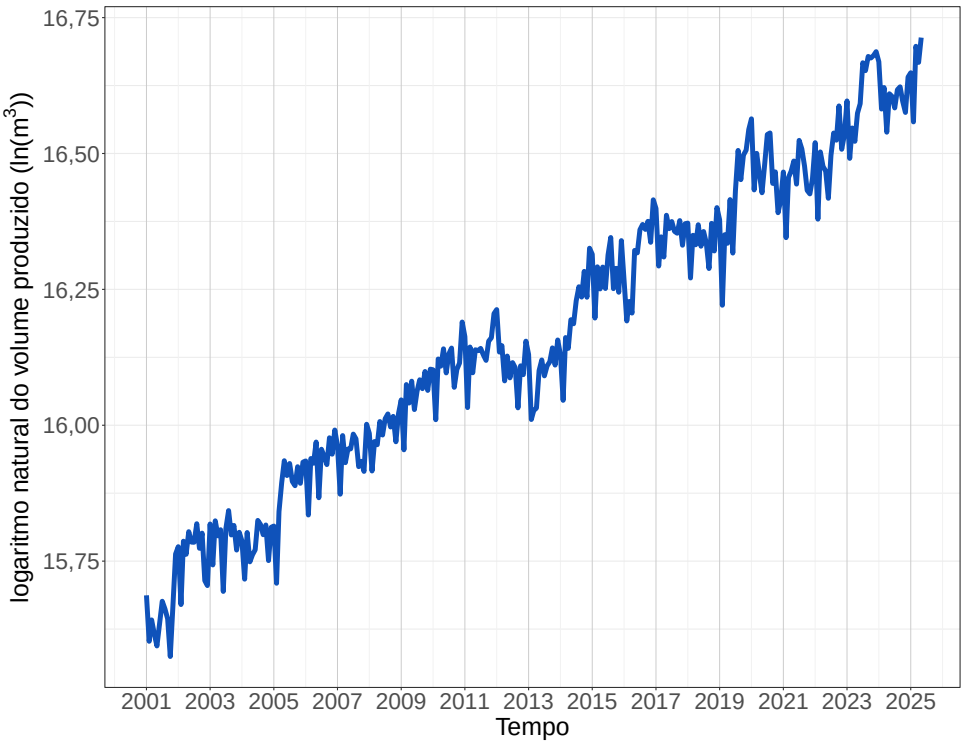
```
=== Variância da série original em escala de trilhões ===
> var(serie_petroleo)
[1] 9.442288e+12
```

ANÁLISE DE VARIABILIDADE - Série Transformada (log)

```
=== Variância da série após transformação logarítmica ===
> var(log(serie_petroleo))
[1] 0.08029983
```

Visualizando o gráfico da transformação logarítmica da série, temos:

Figura 4.1: Série temporal da produção de petróleo após transformação logarítmica.

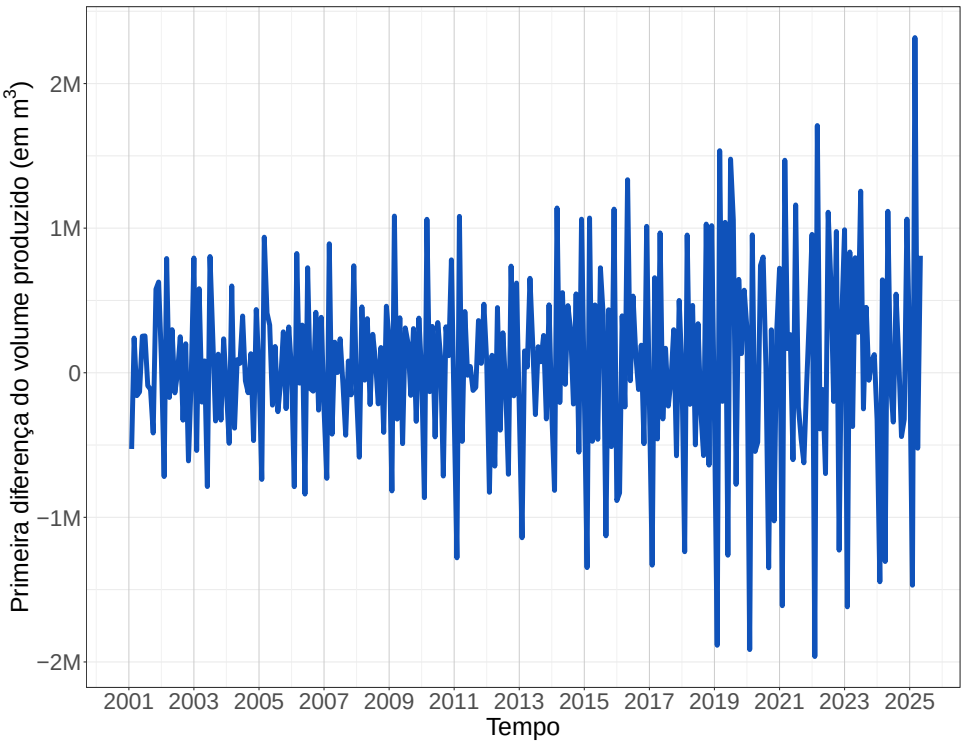


Como esperado, o valor absoluto da variância reduziu drasticamente. O gráfico da série em log (Figura 4.1) preserva a tendência e a sazonalidade. A principal mudança ocorre na interpretação: enquanto a série original é medida em m^3, a série em log reflete variações relativas. Mudanças na série logarítmica podem ser interpretadas como variações percentuais aproximadas na série original.

4.3 Diferenciação para Estacionarização da Média

Confirmada a não estacionariedade, o próximo passo é aplicar a diferenciação para remover a tendência e tornar a média da série constante. A primeira diferença, $Z'_t = Z_t - Z_{t-1}$, transforma a série de níveis em uma série de variações entre um período e outro. Com essa técnica, perde-se a primeira observação da série.

Figura 4.2: Série da produção de petróleo após uma diferenciação.



O gráfico da série diferenciada (Figura 4.2) já não exibe uma tendência clara, com os valores flutuando em torno de uma média aparentemente constante e próxima de zero. No entanto, a variabilidade ainda parece heterogênea, com picos mais acentuados nos anos finais da série histórica. Agora, validamos formalmente a eficácia dessa transformação.

TESTES DE ESTACIONARIEDADE PARA A 1ª DIFERENÇA

```
> # Teste ADF para a série com 1 diferença
> adf.test(diff(serie_petroleo))

  Augmented Dickey-Fuller Test

data:  diff(serie_petroleo)
Dickey-Fuller = -8.0478, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary


> # Teste KPSS para a série com 1 diferença
> kpss.test(diff(serie_petroleo))

  KPSS Test for Level Stationarity
```

```
data:  diff(serie_petroleo)
KPSS Level = 0.068974, Truncation lag parameter = 5, p-
   value = 0.1
```

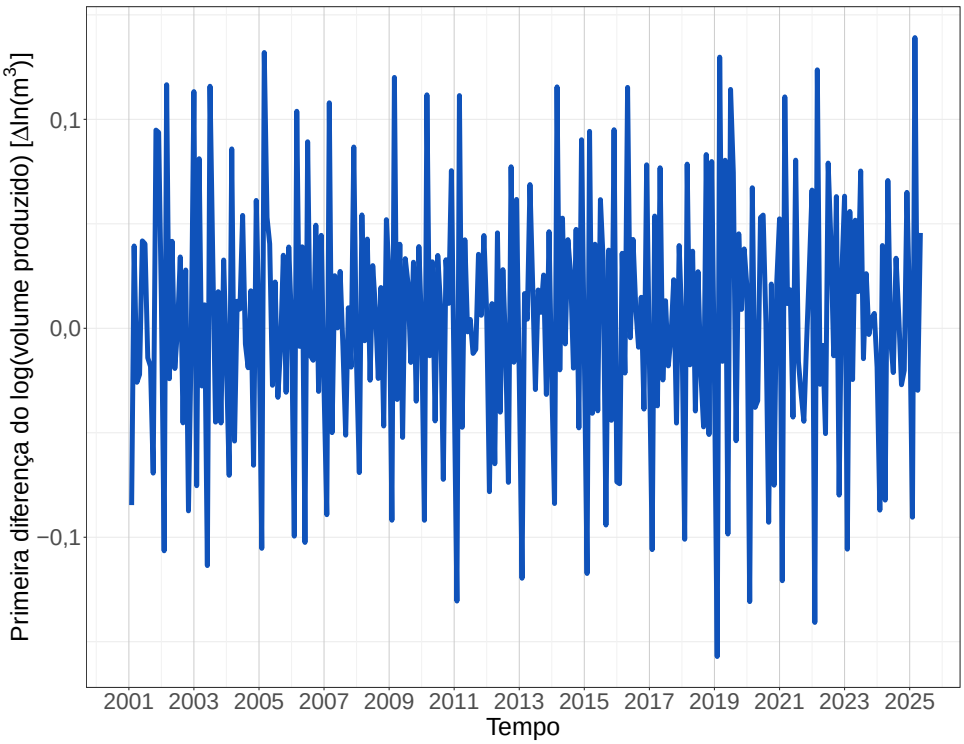
Os resultados agora são conclusivos. O teste ADF (com valor-$p < 0{,}05$) rejeita a hipótese nula, indicando ausência de raiz unitária. O teste KPSS (com valor-$p > 0{,}1$) não rejeita a hipótese nula de estacionariedade. Ambos os testes confirmam que **uma única diferenciação foi suficiente para tornar a série estacionária na média**.

Adicionalmente, o teste de tendência de Mann-Kendall na série diferenciada apresentou um valor-p de 0,48, confirmando que a tendência foi removida com sucesso. Por curiosidade, foi calculada a variância da segunda diferença, que resultou em um valor de $1,28 \times 10^{12}$, muito superior à variância da primeira diferença ($4,29 \times 10^{11}$). Este aumento é um sinal clássico de "sobrediferenciação" (*over-differencing*), reforçando que apenas uma diferenciação é necessária.

4.4 Transformação Combinada: Logaritmo e Diferença

Apesar de a primeira diferença ter removido a tendência, a variabilidade ainda não parece constante. Para resolver ambas as questões — estacionariedade na média e na variância — a abordagem mais robusta é combinar as duas técnicas: aplicar a primeira diferença na série já transformada por logaritmo.

Figura 4.3: Série da produção de petróleo após transformação logarítmica e uma diferenciação.



O resultado (Figura 4.3) é uma série que visualmente aparenta ser muito mais estável, flutuando em torno de zero e com uma variabilidade mais homogênea. A variância desta nova série é de apenas 0,00336, um valor extremamente baixo que indica uma grande estabilização. A interpretação desta série transformada é a da **taxa de crescimento mensal (aproximada) da produção de petróleo**.

5 Identificação e Estimação do Modelo

Uma prática fundamental na construção de modelos preditivos é a divisão da série histórica em conjuntos de treino e de teste. O conjunto de treino é utilizado para que o modelo "aprenda" os padrões e a estrutura de dependência temporal dos dados. Posteriormente, na fase de teste, o modelo já treinado é aplicado a uma porção de dados que ele nunca viu, permitindo uma avaliação honesta de sua capacidade de predição ao comparar os valores previstos com os valores reais.

Para este estudo, o conjunto de dados foi particionado da seguinte forma:

- **Conjunto de Treino:** Período de janeiro de 2001 a abril de 2021.
- **Conjunto de Teste:** Período de maio de 2021 a maio de 2025 (49 observações).

A escolha deste ponto de corte é estratégica. A análise exploratória mostrou que os anos mais recentes da série (a partir de 2021) apresentam não apenas um novo patamar de produção, mas também uma volatilidade distinta. Ao treinar o modelo até abril de 2021, garantimos que ele aprenda com um longo histórico e, em seguida, testamos sua robustez e capacidade de generalização em um período mais recente e desafiador.

Antes de prosseguir, é importante verificar se a série de treino, após as transformações de logaritmo e primeira diferença, é de fato estacionária.

TESTES DE ESTACIONARIEDADE - Série de Treino Transformada

```
> # H0: série possui raiz unitária (não estacionária)
> adf.test(diff(log(treino)))
Dickey-Fuller = -7.2502, Lag order = 6, p-value = 0.01

> # H0: série é estacionária
> kpss.test(diff(log(treino)))
KPSS Level = 0.017139, Truncation lag parameter = 4, p-
   value = 0.1
```

Os testes confirmam a estacionariedade da série de treino transformada (ADF com valor-p < 0,05 e KPSS com valor-p > 0,1). Com isso, podemos analisar suas funções de autocorrelação para identificar a estrutura do modelo candidato.

Figura 5.1: Função de Autocorrelação (ACF) da série de treino transformada.

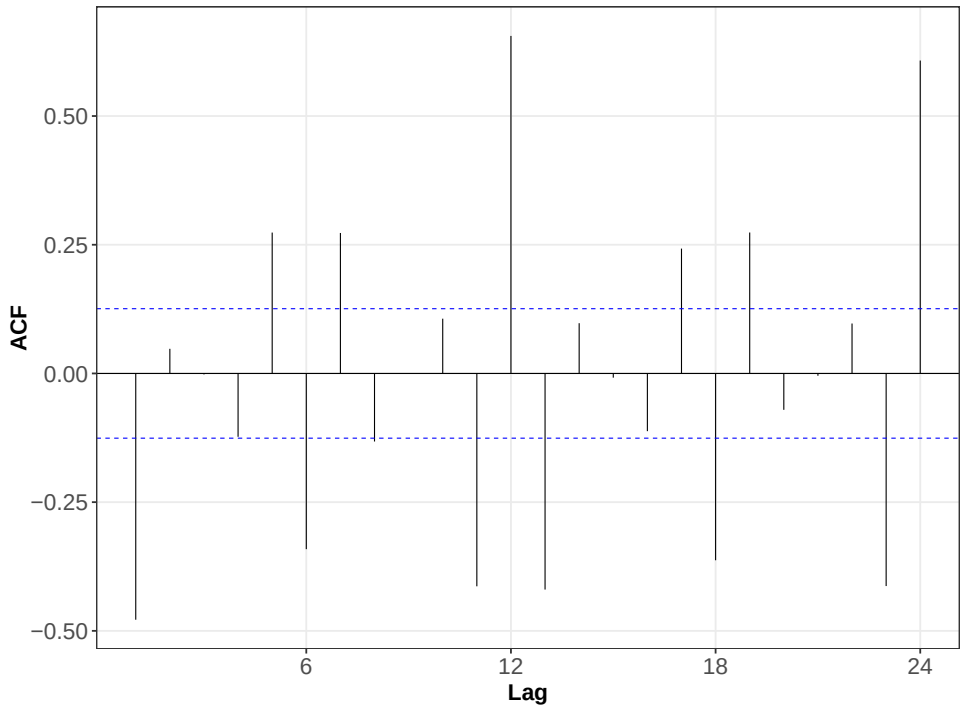
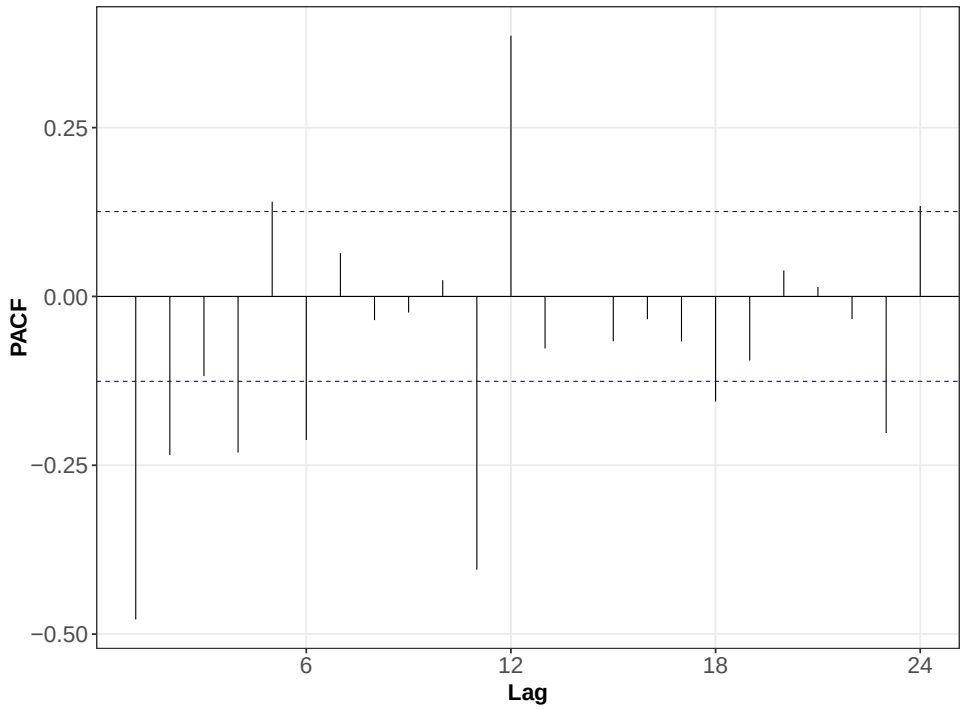


Figura 5.2: Função de Autocorrelação Parcial (PACF) da série de treino transformada.



A análise de autocorrelação da série de treino (Figuras 5.1 e 5.2) nos guiará na escolha das ordens do modelo:

- **ACF:** O gráfico mostra múltiplos picos significativos, especialmente nos lags 1, 6, 12, 18, 24, 30 e 36. O fato desses lags múltiplos do período da série (12, 24, 36, etc.) e seus vizinhos serem significativos confirma a presença de um forte componente sazonal, que não foi removido pela primeira diferença. O decaimento lento nos lags sazonais sugere a necessidade de uma diferenciação sazonal ou de componentes sazonais no modelo (SARIMA).
- **PACF:** O gráfico revela um padrão de alta significância estatística no primeiro e segundo lag, e no lag 11 e 12. Embora os lags múltiplos do período da série não apresentem o mesmo impacto observado no ACF, esses picos ainda indicam a presença de componentes sazonais. A combinação dos padrões no ACF e PACF sugere que o modelo mais adequado seria um ARIMA misto, com componentes sazonais.

A análise conjunta destes gráficos será o guia para a próxima etapa: a especificação e estimação dos modelos candidatos.

5.1 Diagnóstico dos Modelos

Para os diagnósticos, utilizaremos os modelos ARIMA Sazonais (SARIMA), uma extensão do ARIMA que inclui explicitamente os efeitos sazonais. Um modelo SARIMA$(p,d,q)(P,D,Q)_s$ é representado por:

$$\phi_p(B)\Phi_P(B^s)(1-B)^d(1-B^s)^D Z_t = \theta_q(B)\Theta_Q(B^s)a_t, \quad a_t \sim \text{RB}(0,\sigma_a^2), \tag{5.1}$$

onde (p,d,q) são as ordens não sazonais, (P,D,Q) são as ordens sazonais, s é o período da sazonalidade (neste caso, $s = 12$), e B é o operador de defasagem.

5.1.1 Modelo SARIMA(0,1,1)(0,1,1)$_{12}$

Com base na análise das autocorrelações, o primeiro modelo candidato a ser testado é um SARIMA(0,1,1)(0,1,1)$_{12}$. A escolha das ordens é justificada da seguinte maneira: a diferenciação não sazonal ($d = 1$) foi aplicada para remover a tendência, enquanto a diferenciação sazonal ($D = 1$) de ordem 12 é sugerida pelo decaimento lento da ACF nos lags sazonais. O pico significativo no lag 1 da ACF indica a necessidade de um termo de médias móveis não sazonal ($q = 1$), e os lags vizinhos do 12 na ACF sugerem a inclusão de um termo SMA(1) ($Q = 1$).

A representação do modelo SARIMA(0,1,1)(0,1,1)$_{12}$ para a série $Z_t = \log(\text{producao}_t)$ é:

$$(1-B)(1-B^{12})Z_t = (1-\theta_1 B)(1-\Theta_1 B^{12})a_t$$

Expandindo os termos, temos:

$$(1 - B - B^{12} + B^{13})Z_t = (1 - \theta_1 B - \Theta_1 B^{12} + \theta_1 \Theta_1 B^{13})a_t$$

O que pode ser reescrito como:

$$Z_t - Z_{t-1} - Z_{t-12} + Z_{t-13} = a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-12} + \theta_1 \Theta_1 a_{t-13}$$

ou,

$$Z_t = Z_{t-1} + Z_{t-12} - Z_{t-13} + a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-12} + \theta_1 \Theta_1 a_{t-13} \tag{5.2}$$

onde a_t é o ruído branco, θ_1 é o coeficiente de médias móveis não sazonal e Θ_1 é o coeficiente de médias móveis sazonal.

Após estimar o modelo, o próximo passo é realizar o diagnóstico de seus resíduos. Idealmente, os resíduos de um bom modelo devem se comportar como um ruído branco, ou seja, não devem possuir nenhuma autocorrelação remanescente. Os gráficos de autocorrelação dos resíduos são apresentados nas Figuras 5.3 e 5.4.

Figura 5.3: Função de Autocorrelação (ACF) dos resíduos do modelo SARIMA(0,1,1)(0,1,1)$_{12}$.

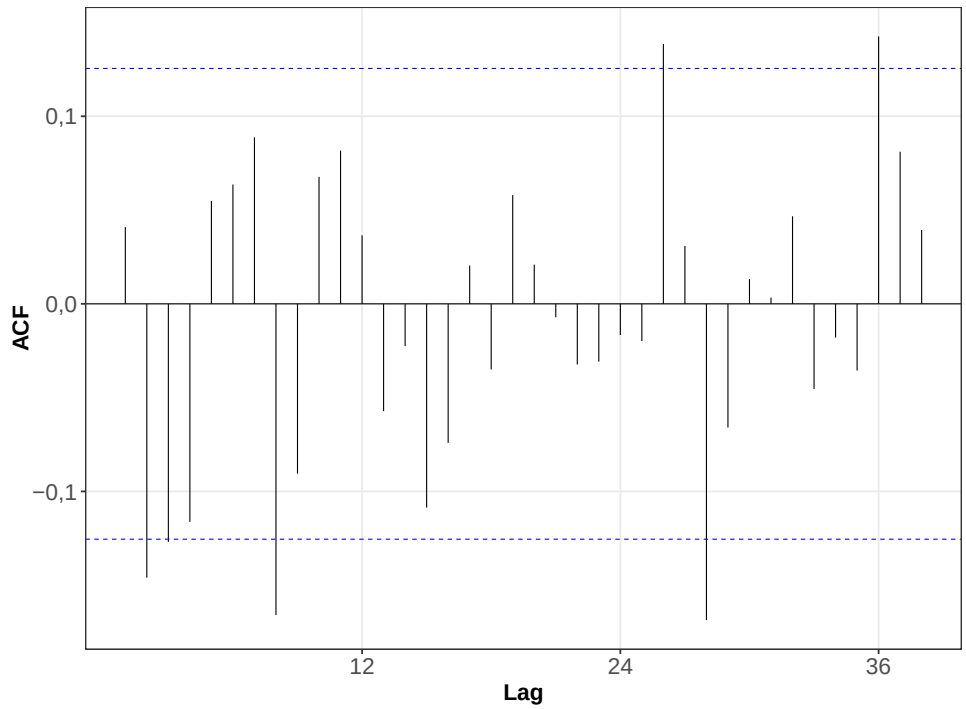
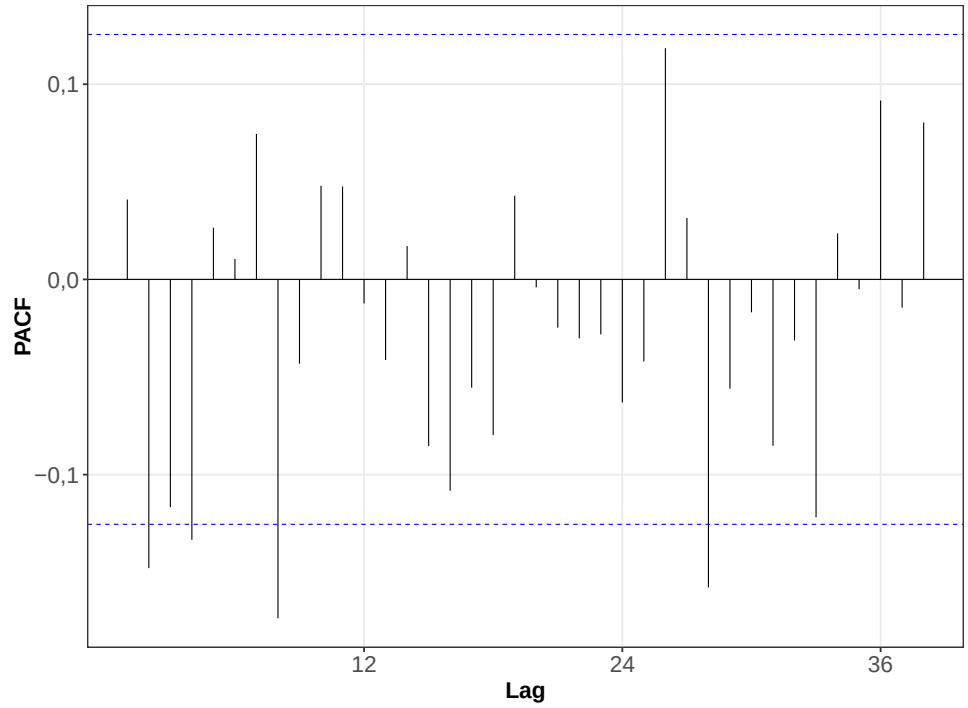


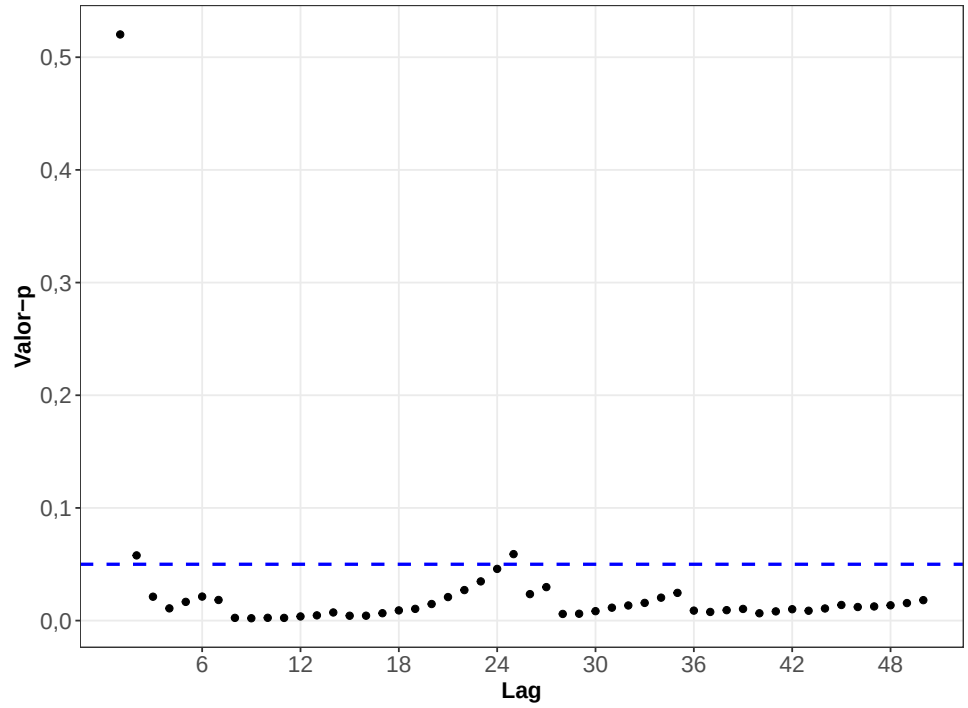
Figura 5.4: Função de Autocorrelação Parcial (PACF) dos resíduos do modelo SARIMA$(0,1,1)(0,1,1)_{12}$.



A análise dos correlogramas mostra-nos que o modelo não foi capaz de capturar toda a estrutura de dependência dos dados. O gráfico da ACF (Figura 5.3) ainda exibe autocorrelação significativa nos lags 2, 3, 8, 26, 28 e 36. A PACF (Figura 5.4) confirma essa deficiência, com picos também significativos nos lags 2, 4, 8 e 28.

Para um diagnóstico mais rigoroso, foi aplicado o teste de Ljung-Box, que avalia a hipótese nula de que os resíduos são independentes (ausência de autocorrelação) para um conjunto de lags.

Figura 5.5: Valores-p do teste de Ljung-Box para os resíduos do modelo SARIMA$(0,1,1)(0,1,1)_{12}$.



O resultado (Figura 5.5) é conclusivo: para quase todos os lags testados, o valor-p situa-se abaixo do nível de significância $\alpha = 5\%$ (linha tracejada). Rejeita-se, portanto, a hipótese de independência, indicando que o modelo SARIMA$(0,1,1)(0,1,1)_{12}$ é inadequado.

5.1.2 Modelo SARIMA$(1,1,1)(0,1,1)_{12}$

Dado que o primeiro modelo falhou em capturar a estrutura de correlação, e observando os picos remanescentes no lag 2 dos resíduos, além do pico original no lag 1 da PACF da série de treino (Figura 5.2), um refinamento natural é a adição de um termo autorregressivo não sazonal ($p = 1$). Com isso, o novo candidato é o modelo SARIMA$(1,1,1)(0,1,1)_{12}$. Sua representação é dada por:

$$\begin{aligned}(1-\phi_1 B)(1-B^{12})(1-B)Z_t &= (1-\theta_1 B)(1-\Theta_1 B^{12})a_t \\ (1-\phi_1 B - B^{12} + \phi_1 B^{13})(1-B)Z_t &= (1-\theta_1 B - \Theta_1 B^{12} + \theta_1\Theta_1 B^{13})a_t \\ Z_t - \phi_1 Z_{t-1} - Z_{t-12} + \phi_1 Z_{t-13} &= a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-12} + \theta_1\Theta_1 a_{t-13}\end{aligned}$$

ou,

$$Z_t = \phi_1 Z_{t-1} + Z_{t-12} - Z_{t-13} + a_t - \theta_1 a_{t-1} - \Theta_1 a_{t-12} + \theta_1\Theta_1 a_{t-13} \tag{5.3}$$

sendo ϕ_1 o coeficiente autorregressivo de ordem 1 não sazonal.

Repetimos o processo de diagnóstico para este novo modelo.

Figura 5.6: Função de Autocorrelação (ACF) dos resíduos do modelo SARIMA(1,1,1)(0,1,1)$_{12}$.

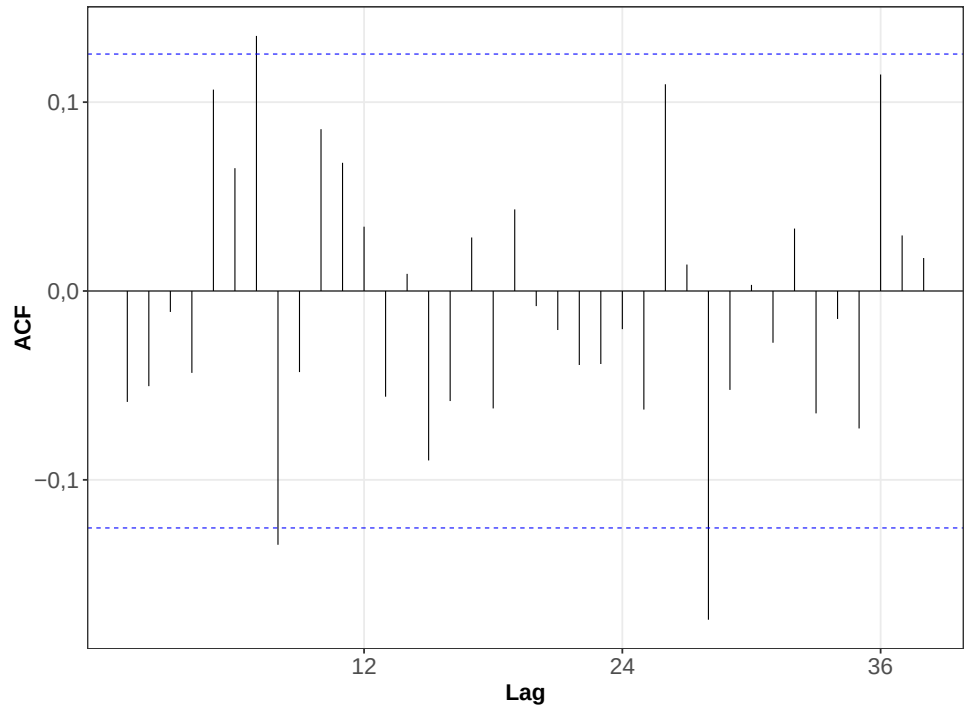
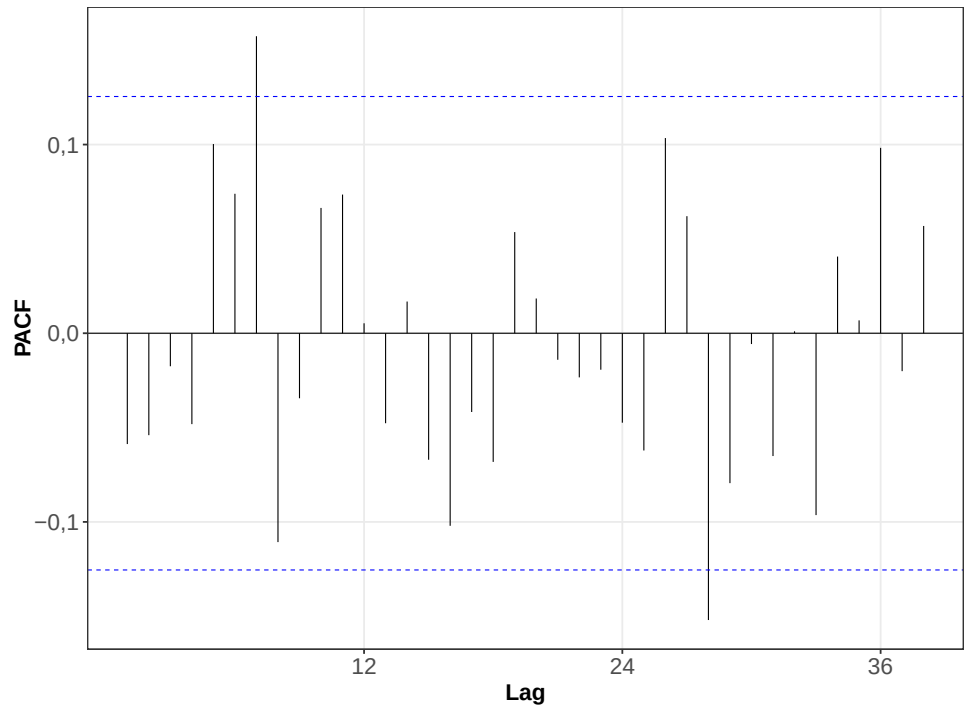


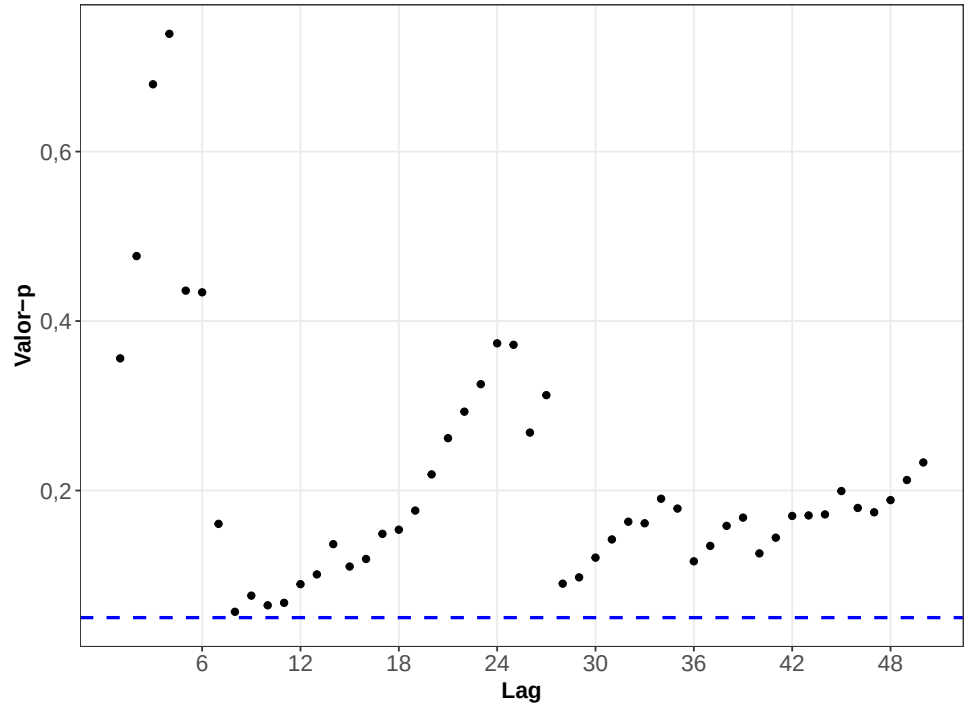
Figura 5.7: Função de Autocorrelação Parcial (PACF) dos resíduos do modelo SARIMA(1,1,1)(0,1,1)$_{12}$.



Houve uma melhora substancial. Os correlogramas dos resíduos (Figuras 5.6 e 5.7) agora mostram menos picos fora dos limites de significância. Apenas os lags 7, 8 e 28 destacam-se

como marginalmente significativos na autocorrelação, e na autocorrelação parcial os lags 7 e 28.

Figura 5.8: Valores-p do teste de Ljung-Box para os resíduos do Modelo 2.



O teste de Ljung-Box (Figura 5.8) confirma o grande avanço na qualidade do ajuste. Para todos os 50 lags testados, nenhum valor-p é inferior ao nível de significância de 5%. Portanto, não há evidências para rejeitar a hipótese de que os resíduos são independentes. Isso indica que o modelo SARIMA$(1,1,1)(0,1,1)_{12}$ é estatisticamente adequado e capta bem a estrutura de dependência da série.

5.1.3 Refinamento e Comparação de Outros Modelos

Apesar de o modelo SARIMA$(1,1,1)(0,1,1)_{12}$ ter se mostrado estatisticamente adequado, a busca por um ajuste ainda mais preciso levou a investigar outras estruturas. O objetivo é tentar capturar a autocorrelação residual remanescente, especialmente nos lags iniciais, e encontrar o modelo mais parcimonioso e com maior poder preditivo.

Ainda com base nos correlogramas da série de treino transformada (Figuras 5.1 e 5.2), são propostos dois novos modelos, aumentando a ordem dos componentes não sazonais:

- **SARIMA$(2,1,1)(0,1,1)_{12}$:** Adiciona-se um segundo termo autorregressivo ($p = 2$), motivado pelo pico significativo no lag 2 da PACF (5.4). Sua equação é:

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - B^{12})Z_t = (1 - \theta_1 B)(1 - \Theta_1 B^{12})a_t$$

- **SARIMA$(1,1,2)(0,1,1)_{12}$:** Alternativamente, adiciona-se um segundo termo de médias

móveis ($q = 2$), motivado pelo comportamento dos primeiros lags da ACF (5.3). Sua equação é:

$$(1-\phi_1 B)(1-B)(1-B^{12})Z_t = (1-\theta_1 B-\theta_2 B^2)(1-\Theta_1 B^{12})a_t$$

Ambos os modelos apresentaram resultados semelhantes ao do SARIMA(1,1,1)(0,1,1)$_{12}$, sem uma melhora conclusiva na eliminação da correlação residual nos lags 7 e 8. Isso motivou o teste de modelos com ordens ainda mais altas, como o SARIMA(3,1,2)(0,1,1)$_{12}$ e o SARIMA(2,1,3)(0,1,1)$_{12}$, que, embora tenham capturado a correlação nesses lags, ainda apresentaram picos residuais em lags mais altos (como o 28).

Essa investigação levou a uma última abordagem: refinar o modelo SARIMA(2,1,3)(0,1,1)$_{12}$ com base na significância estatística de seus próprios coeficientes. Na estimação inicial deste modelo, o coeficiente ϕ_1 (AR(1)) se mostrou não significativo. Isso sugere que ele pode ser removido do modelo sem perda de poder explicativo, resultando em um modelo mais parcimonioso. Assim, testou-se um modelo final com uma restrição:

- **SARIMA(2,1,3)(0,1,1)$_{12}$ com $\phi_1 = 0$:** Um modelo onde o primeiro coeficiente autorregressivo é fixado em zero. Sua equação é:

$$(1-\phi_2 B^2)(1-B)(1-B^{12})Z_t = (1-\theta_1 B-\theta_2 B^2-\theta_3 B^3)(1-\Theta_1 B^{12})a_t$$

Para selecionar o melhor modelo entre todos os candidatos, comparamos seus critérios de informação (AIC, AICc, BIC). Modelos com menores valores para esses critérios são preferíveis, pois indicam um melhor equilíbrio entre o ajuste aos dados e a complexidade do modelo. A Tabela 5.1 resume os resultados.

Tabela 5.1: Comparação dos critérios de informação para os modelos SARIMA candidatos.

Modelo	**AIC**	**AICc**	**BIC**
SARIMA(0,1,1)(0,1,1)$_{12}$	-890,19	-890,09	-879,87
SARIMA(1,1,1)(0,1,1)$_{12}$	**-898,42**	**-898,25**	**-884,65**
SARIMA(2,1,1)(0,1,1)$_{12}$	-897,55	-897,29	-880,34
SARIMA(1,1,2)(0,1,1)$_{12}$	-897,82	-897,55	-880,60
SARIMA(3,1,2)(0,1,1)$_{12}$	-895,12	-894,62	-871,02
SARIMA(2,1,3)(0,1,1)$_{12}$	-895,51	-895,01	-871,41
SARIMA(2,1,3)(0,1,1)$_{12}$ com $\phi_1 = 0$	-897,43	-897,06	-876,78

A seguir, são apresentados os resultados das métricas do ajuste para os diferentes modelos SARIMA.

Tabela 5.2: Erros de ajuste dos modelos SARIMA.

Modelo	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**
SARIMA(0,1,1)(0,1,1)$_{12}$	-0,0008	0,0324	0,0241	-0,0054	0,1501	0,3886	0,0409
SARIMA(1,1,1)(0,1,1)$_{12}$	-0,0029	0,0312	0,0229	-0,0184	0,1424	0,3688	-0,0587
SARIMA(2,1,1)(0,1,1)$_{12}$	-0,0027	0,0311	0,0229	-0,0172	0,1425	0,3691	-0,0089
SARIMA(1,1,2)(0,1,1)$_{12}$	-0,0027	0,0312	0,0229	-0,0169	0,1426	0,3694	0,0030
SARIMA(3,1,2)(0,1,1)$_{12}$	-0,0027	0,0312	0,0228	-0,0169	0,1419	0,3674	-0,0074
SARIMA(2,1,3)(0,1,1)$_{12}$	-0,0026	0,0312	0,0229	-0,0165	0,1423	0,3684	0,0028
SARIMA(2,1,3)(0,1,1)$_{12}$ com $\phi_1 = 0$	-0,0026	0,0312	0,0229	-0,0168	0,1421	0,3680	0,0022

Como era de se esperar, a análise apresentada na tabela 5.1 indica que o modelo SARIMA(1,1,1)(0,1,1)$_{12}$ apresenta os melhores valores para os critérios AIC, AICc e BIC, reforçando sua qualidade, que já foi observada nos diagnósticos anteriores, além de destacar sua simplicidade. Por outro lado, as métricas de erro de ajuste (5.2), como o RMSE, mostram-se extremamente semelhantes entre todos os modelos.

Entretanto, é importante destacar a alta paridade entre os modelos, com um foco especial para o modelo SARIMA(2,1,3)(0,1,1)$_{12}$ com $\phi_1 = 0$, que, além de apresentar boas métricas na análise dos critérios de informação, destacou-se por ser o mais eficaz na eliminação das autocorrelações nos lags 7 e 8. Esse desempenho torna esse modelo o escolhido.

5.1.4 Diagnóstico do Modelo SARIMA(2,1,3)(0,1,1)$_{12}$ com $\phi_1 = 0$

Com base na análise comparativa, um dos candidatos mais fortes é o modelo SARIMA(2,1,3)(0,1,1)$_{12}$ com o primeiro coeficiente autorregressivo fixado em zero ($\phi_1 = 0$). Esta seção se dedica ao seu diagnóstico detalhado.

Sua equação, com os coeficientes estimados, é:

$$(1-0,7088B^2)(1-B)(1-B^{12})Z_t = (1+0,2711B+0,8748B^2-0,1460B^3)(1+0,9117B^{12})a_t \tag{5.4}$$

A primeira etapa do diagnóstico é analisar os resíduos do modelo. As figuras 5.9 e 5.10 apresentam suas funções de autocorrelação.

Figura 5.9: Função de Autocorrelação (ACF) dos resíduos do modelo final.

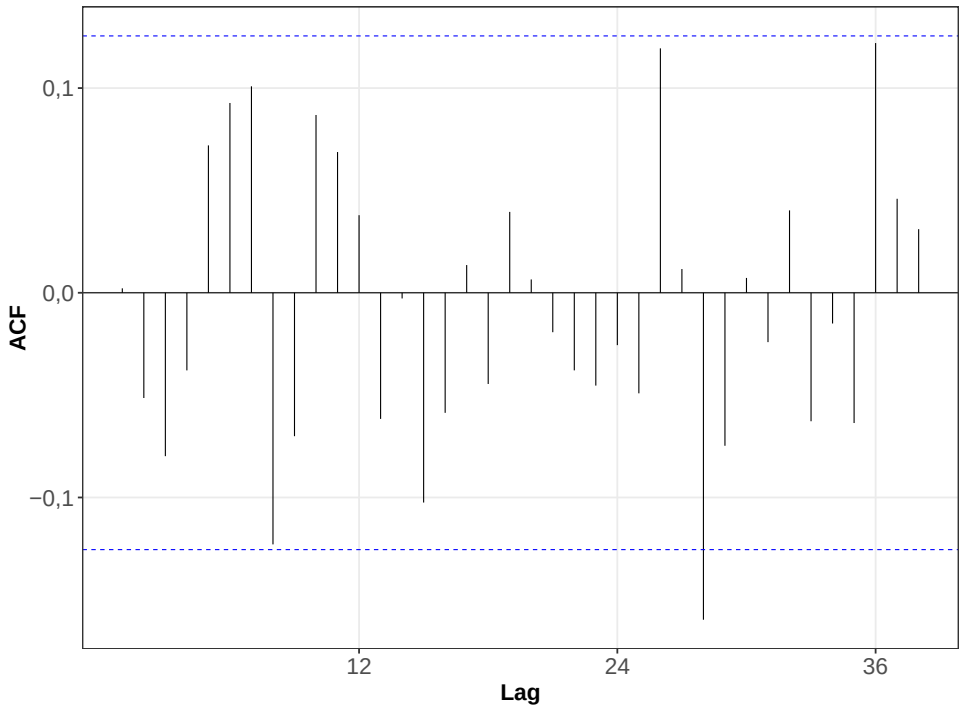
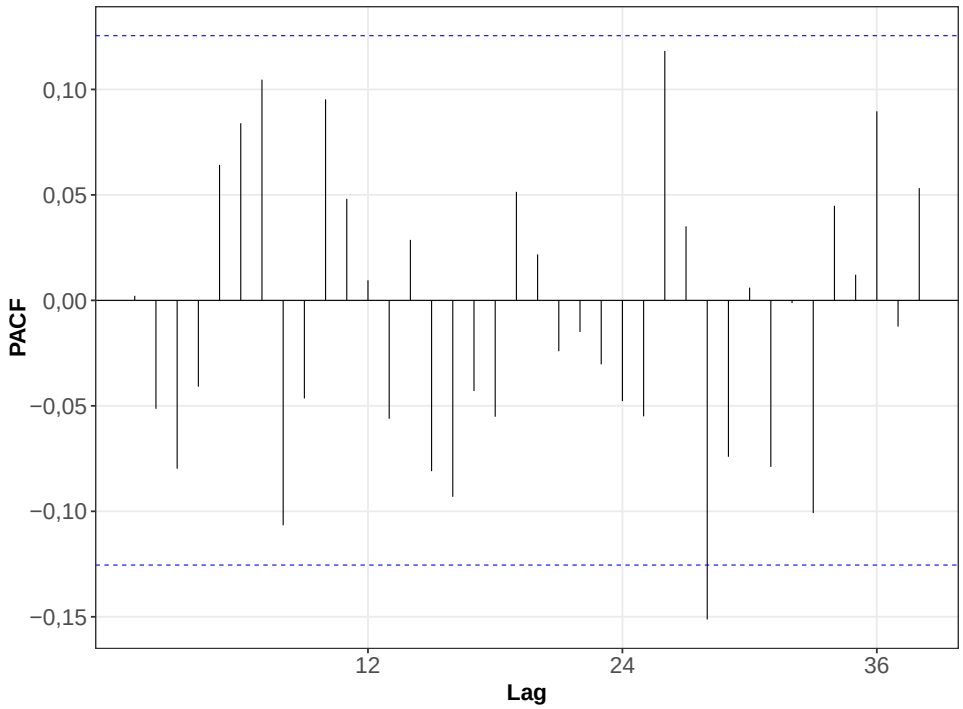


Figura 5.10: Função de Autocorrelação Parcial (PACF) dos resíduos do modelo final.

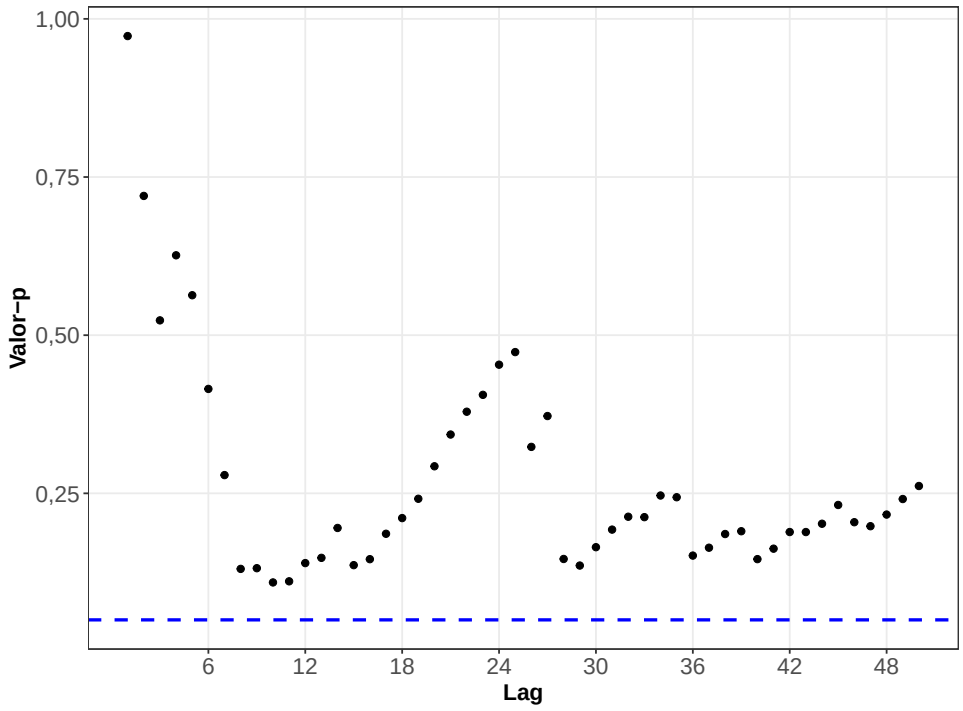


Os gráficos mostram que o modelo foi eficaz em controlar a autocorrelação que existia nos lags 7 e 8. Ainda se observa um pico marginalmente significativo no lag 28, porém, essa

autocorrelação em um lag tão alto (mais de dois anos) será desconsiderada neste relatório, pois tem pouca relevância prática para o estudo.

Para validar a independência dos resíduos de forma mais rigorosa, aplicou-se o teste de Ljung-Box.

Figura 5.11: Valores-p do teste de Ljung-Box para os resíduos do modelo final.

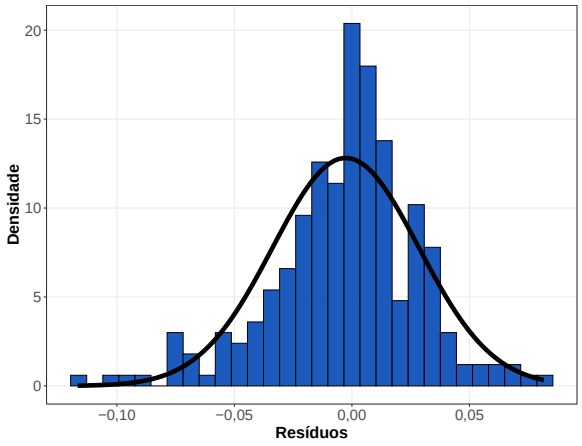


A figura 5.11 comprova a qualidade do ajuste. Para todos os lags testados, os valores-p se mantêm confortavelmente acima do nível de significância de 5%, levando à não rejeição da hipótese nula. Conclui-se, portanto, que os resíduos do modelo podem ser considerados não correlacionados.

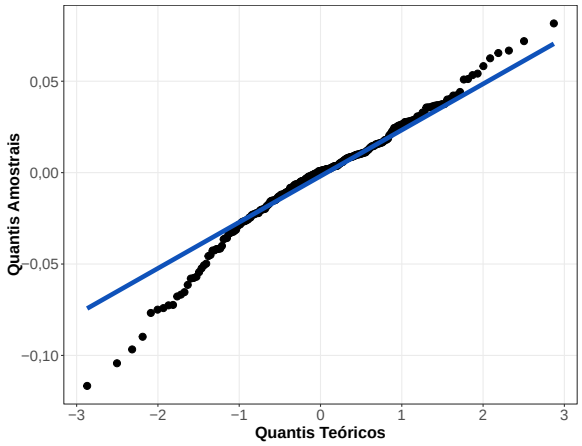
Finalmente, avalia-se a premissa de normalidade dos resíduos, visualmente através do histograma e do gráfico Q-Q e formalmente com testes de hipóteses.

Figura 5.12: Histograma com curva de densidade e gráfico Q-Q de normalidade para os resíduos do modelo final.

(a) Histograma.



(b) Gráfico Q-Q.



Visualmente, o histograma (5.12a) mostra que a distribuição dos resíduos é centrada em zero, mas com uma leve assimetria à esquerda (cauda mais longa para valores negativos). O gráfico Q-Q (5.12b) confirma essa impressão, com os pontos se afastando da linha de normalidade teórica nas caudas. Os testes formais confirmam a suspeita:

TESTE DE NORMALIDADE - Shapiro-Wilk

```
> shapiro.test(modelo$residuals)
W = 0.9724, p-value = 0.0001115
```

TESTE DE NORMALIDADE - Jarque-Bera

```
> jarque.bera.test(modelo$residuals)
X-squared = 27.7, df = 2, p-value = 9.67e-07
```

Com valores-p extremamente baixos, ambos os testes rejeitam a hipótese nula, indicando que os resíduos não seguem uma distribuição normal. É importante notar que, embora a normalidade seja desejável (principalmente para a construção de intervalos de predição precisos), sua violação não invalida o modelo para previsões pontuais, que costumam ser robustas a esse desvio.

6 Previsão e Avaliação do Desempenho

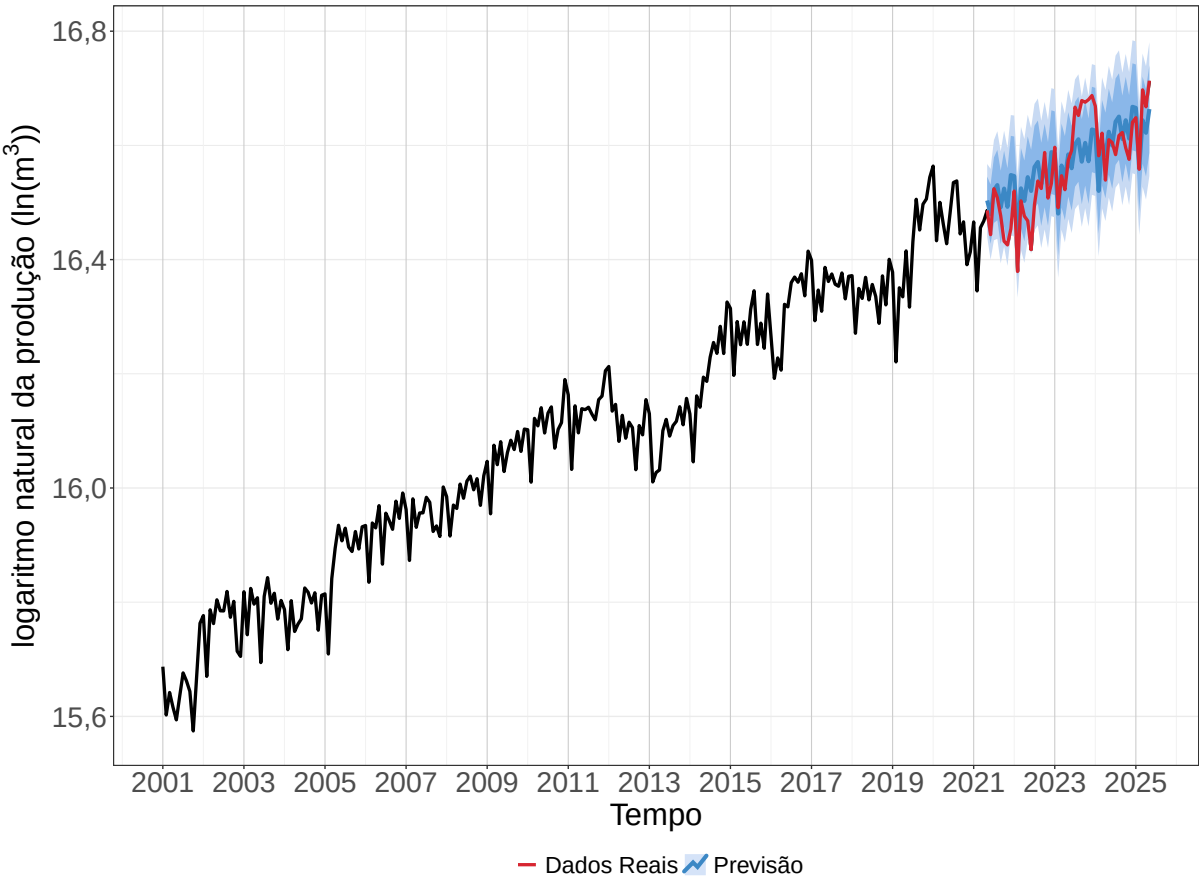
Após a etapa de identificação e diagnóstico, o passo final consiste em utilizar o modelo selecionado para prever o comportamento da série no futuro e avaliar seu real desempenho. Nesta fase, utilizamos o conjunto de teste — um período de dados que o modelo não utilizou em seu treinamento — para simular uma previsão em um cenário real.

Para este estudo, compararemos o desempenho de todos os modelos SARIMA candidatos no horizonte de previsão de maio de 2021 a maio de 2025. Adicionalmente, incluiremos um modelo de Suavização Exponencial (Holt-Winters) como um *benchmark*, ou seja, um ponto de referência para julgar a performance dos modelos mais complexos.

6.1 Análise da Previsão do Modelo Selecionado

Iniciamos com a análise visual do modelo escolhido na fase de diagnóstico, o SARIMA$(2,1,3)(0,1,1)_{12}$ com $\phi_1 = 0$. A figura 6.1 compara os valores previstos pelo modelo com os valores reais da série, incluindo o intervalo de confiança de 80% e 95% para as previsões.

Figura 6.1: Previsão do modelo SARIMA(2,1,3)(0,1,1)$_{12}$ com $\phi_1 = 0$ em comparação com os dados reais do conjunto de teste.



A visualização da figura 6.1 revela que as previsões do modelo são, de modo geral, bastante acuradas, capturando bem a tendência e os padrões sazonais da produção. Algumas observações mais detalhadas podem ser feitas:

- Entre 2021 e meados de 2022, o modelo tende a superestimar ligeiramente a produção, embora os valores reais permaneçam quase sempre dentro do intervalo de confiança de 80%.
- Nos anos de 2023 e 2024, período de máxima na produção de petróleo, o modelo por vezes subestima essa alta de produção, não conseguindo antecipar totalmente a magnitude desses aumentos. Este é o período em que a previsão mais se distancia dos dados reais.
- Para o final da série (2024 e 2025), o modelo volta a prever os valores com grande proximidade e consistência.

6.2 Comparação da Acurácia dos Modelos

Complementarmente, foi analisada a acurácia tanto do modelo selecionado para o processo de previsão quanto dos outros modelos candidatos, proporcionando uma visão mais abrangente do processo de construção do modelo final. Esta análise comparativa permite avaliar se a escolha do modelo foi adequada e identifica possíveis alternativas de desempenho similar.

Os erros de mensuração apresentados na tabela 6.1 são provenientes da comparação entre os modelos ajustados durante o treino e os valores reais do conjunto de teste. As métricas utilizadas para avaliar a performance dos modelos incluem o erro médio (ME), o erro quadrático médio (RMSE), o erro absoluto médio (MAE), o erro percentual médio (MPE), o erro percentual absoluto médio (MAPE), a autocorrelação no primeiro lag (ACF1) e a medida de Theil's U. Essas métricas fornecem uma visão mais ampla da precisão dos modelos em termos de magnitude do erro, direção do viés e variação dos resíduos.

Tabela 6.1: Erros de previsão para os modelos SARIMA.

Modelo	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**	**ACF1**	**Theil's U**
SARIMA(0,1,1)(0,1,1)$_{12}$	-0,0013	0,0489	0,0393	-0,0091	0,2371	0,7030	0,8487
SARIMA(1,1,1)(0,1,1)$_{12}$	-0,0077	0,0498	0,0409	-0,0477	0,2470	0,7070	0,8637
SARIMA(2,1,1)(0,1,1)$_{12}$	-0,0075	0,0496	0,0408	-0,0464	0,2461	0,7056	0,8618
SARIMA(1,1,2)(0,1,1)$_{12}$	-0,0074	0,0496	0,0407	-0,0456	0,2455	0,7046	0,8604
SARIMA(3,1,2)(0,1,1)$_{12}$	-0,0074	0,0496	0,0407	-0,0460	0,2457	0,7033	0,8615
SARIMA(2,1,3)(0,1,1)$_{12}$	-0,0071	0,0494	0,0405	-0,0442	0,2445	0,7013	0,8585
SARIMA(2,1,3)(0,1,1)$_{12}$ com $\phi_1 = 0$	-0,0072	0,0495	0,0406	-0,0448	0,2450	0,7019	0,8598

A análise dos resultados revela que todos os modelos apresentam desempenho muito próximo em todas as métricas de avaliação utilizadas, com exceção do primeiro modelo SARIMA(0,1,1)(0,1,1)$_{12}$, que se destaca com os menores valores de erro. Contudo, como pontuado anteriormente, este modelo não apresenta bons resultados na captura da autocorrelação da série, exibindo uma quantidade considerável de lags significativos nos gráficos de autocorrelação (figuras 5.3 e 5.4).

Considerando que o modelo SARIMA(2,1,3)(0,1,1)$_{12}$ com $\phi_1 = 0$ foi o que melhor apresentou qualidade em capturar a autocorrelação da série, teve bom desempenho nos testes de Ljung-Box (figura 5.11), bons resultados nas métricas de erro (tabela 6.1) e boa qualidade na previsão de novos valores do logaritmo da produção de petróleo, temos, portanto, um modelo conclusivo que atende satisfatoriamente aos critérios de adequação estatística e precisão preditiva.

6.3 Comparação com a Suavização Exponencial

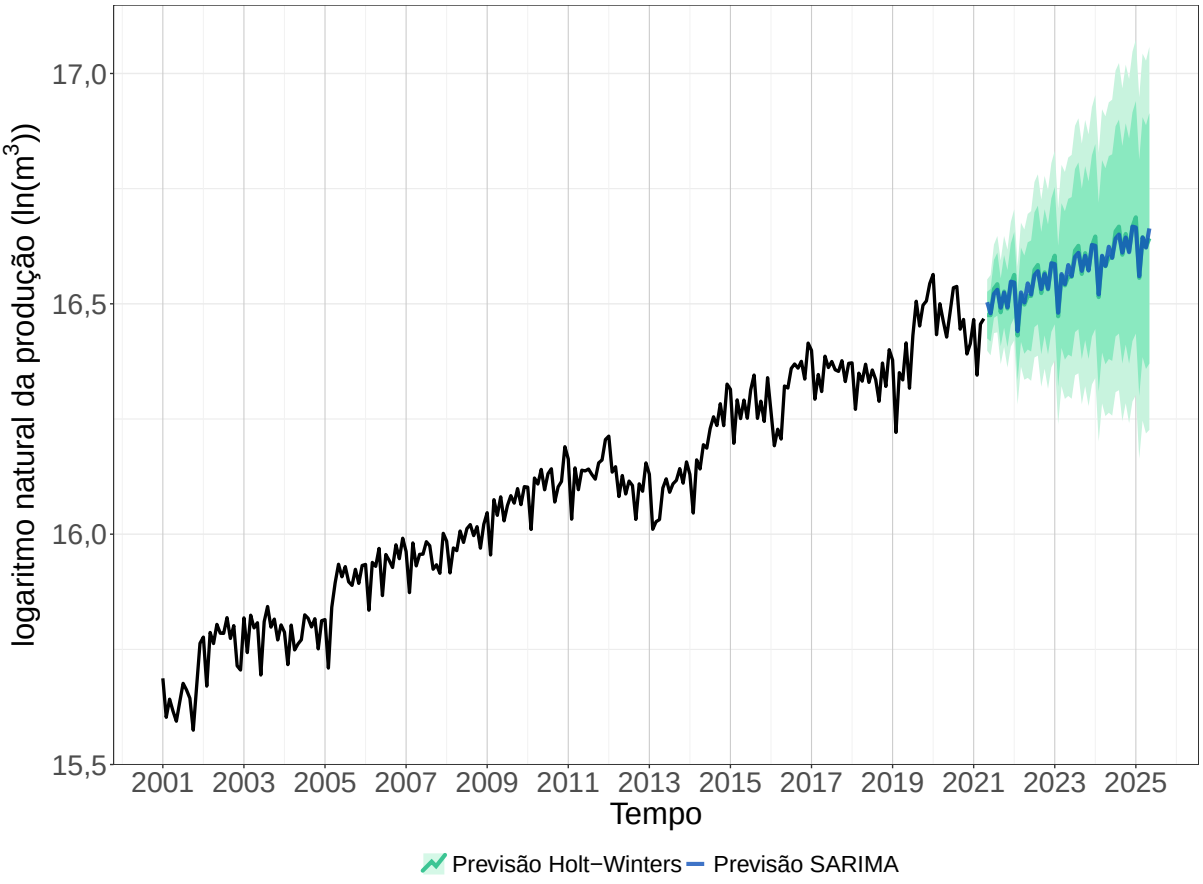
Como última parte de análise do relatório, comparamos o modelo escolhido e aprovado pelos testes da metodologia de Box-Jenkins com um modelo de suavização exponencial, buscando avaliar diferentes abordagens para a previsão da série temporal estudada.

O modelo de suavização exponencial fundamenta-se na ideia de atribuir maior peso aos valores mais recentes e menor peso aos dados mais antigos, ou seja, ele "suaviza" a série temporal de maneira exponencial. Existem alguns tipos de suavização exponencial, mas a escolhida para este comparativo é a Suavização Exponencial Tripla (ou Holt-Winters), que é uma extensão que inclui componentes de tendência e sazonalidade, sendo, portanto, adequada para a nossa série.

A versão aditiva foi escolhida, pois a série foi previamente transformada por logaritmo para estabilizar a variância.

A figura 6.2 sobrepõe as previsões geradas por ambos os métodos no conjunto de teste.

Figura 6.2: Comparativo das previsões de Holt-Winters (com seus intervalos de confianla) e o modelo SARIMA$(2,1,3)(0,1,1)_{12}$ com $\phi_1 = 0$.



Através da figura 6.2 podemos visualizar tanto a previsão de Holt-Winters quanto a previsão do modelo SARIMA$(2,1,3)(0,1,1)_{12}$ com $\phi_1 = 0$. Percebe-se que os valores de previsão

do modelo de suavização exponencial apresentam grande proximidade com os valores do modelo usando a metodologia Box-Jenkins. Em muitas previsões para os meses, os valores chegam a ser quase equivalentes, com pouca distância entre eles até mesmo nas previsões mais distintas. Um destaque importante é que o modelo de Holt-Winters gera intervalos de confiança consideravelmente maiores que os do método de Box-Jenkins, sugerindo uma maior incerteza associada às suas previsões.

A tabela 6.2 apresenta as métricas de acurácia para ambos os modelos, permitindo uma análise quantitativa de seu desempenho.

Tabela 6.2: Comparação das métricas de acurácia entre os modelos SARIMA e Holt-Winters.

Modelo	**ME**	**RMSE**	**MAE**	**MPE**	**MAPE**	**MASE**	**ACF1**	**Theil's U**
SARIMA(2,1,3)$(0,1,1)_{12}$ com $\phi_1 = 0$	-0,0072	0,0495	0,0406	-0,0448	0,2450	0,6530	0,7019	0,8598
Holt-Winters	-0,0070	0,0501	0,0410	-0,0434	0,2472	0,6591	0,6385	0,8702

Analisando as métricas de acurácia no conjunto de treinamento, observa-se que o modelo SARIMA parece ter uma leve vantagem em termos de precisão e qualidade das previsões.

De modo geral, levando em consideração que a suavização exponencial de Holt-Winters - neste caso aditiva - gera valores de previsão satisfatórios e relativamente próximos dos modelos com metodologia Box-Jenkins, porém com uma implementação consideravelmente mais simples, temos um modelo alternativo bastante promissor na parte preditiva. Entretanto, o fato de apresentar intervalos de confiança substancialmente maiores pode indicar menor precisão deste modelo.

A escolha final entre a suavização exponencial de Holt-Winters e o modelo SARIMA$(2,1,3)(0,1,1)_{12}$ com $\phi_1 = 0$ depende fundamentalmente da questão de praticidade versus maior controle no desenvolvimento da previsão, para este conjunto de dados. Enquanto o Holt-Winters oferece simplicidade de implementação, o modelo SARIMA proporciona maior rigor estatístico e melhor ajuste aos dados.

7 Apêndice A

Série Temporal da base de dados.

```
> serie_petroleo
          Jan      Feb      Mar      Apr      May      Jun
        Jul      Aug      Sep      Oct      Nov      Dec
2001  6498405  5971240  6211220  6052810  5920527  6173469
    6428124  6338025  6223285  5806609  6384420  7011827
2002  7106895  6389095  7179072  7007921  7305933  7166751
    7164729  7413679  7085702  7286273  6676713  6616203
2003  7409792  6872130  7453190  7250214  7332231  6545205
    7349096  7595173  7261756  7390239  7062968  7297703
2004  7182535  6694564  7294742  6911528  7001749  7066628
    7458958  7403277  7264852  7396430  6927263  7364453
2005  7382254  6644407  7581902  7995070  8324684  8100731
    8282310  8013086  7950711  8233697  7985490  8302393
2006  8321362  7533460  8358067  8286098  8615745  7776547
    8502763  8393094  8265597  8683793  8425352  8808735
2007  8557460  7827167  8719200  8294579  8506144  8507653
    8742543  8666584  8234483  8315685  8162621  8902509
2008  8753001  8169007  8624510  8573236  8947344  8727800
    8993784  9073102  8857868  9032539  8619957  9080024
2009  9310661  8493596  9577172  9256419  9636554  9146200
    9456357  9661504  9504934  9810383  9474178  9852167
2010  9841034  8977144 10038414  9907401 10227859  9784466
   10131670 10242388  9528102  9845848  9963874 10744711
2011 10459863  9179919 10261545  9786788 10210190 10192674
   10236133 10113736 10012434 10373449 10438137 10912001
2012 10993516 10166701 10287594  9641935 10092617  9695665
    9971773  9877331  9175222  9912636  9752828 10373350
2013 10121333  8980299  9130967  9171083  9824229 10020726
    9731279  9911445  9987346 10245379  9926112 10395815
2014 10116220  9302576 10441978 10236311 10790535 10710183
```

```
   11174551 11465498 11249119 11794528 11245741 12307870
2015 12171006 10823538 11893518 11419328 11888262 11427871
   12153759 12552588 11424773 11860458 11349421 12481025
2016 11597169 10765924 11159478 10923091 12257983 12202867
   12733059 12857465 12741490 12932062 12441672 13454317
2017 13241004 11910378 12567784 12109928 13076946 12757633
   12927061 12696405 12651851 12949556 12375461 12875354
2018 12889260 11651492 12603924 12385403 12851125 12352575
   12691799 12428694 11855957 12884028 12245480 13262950
2019 12966799 11082031 12617704 12419623 13459874 12197978
   13675415 14733726 13961986 14607873 14740356 15311227
2020 15615835 13701310 14654648 14110272 13629741 14371559
   15171754 15214597 13865864 14162777 13138310 13437135
2021 14159958 12548843 14018537 14185436 14449446 13848290
   15008887 14772716 14312681 13689776 13601431 13989220
2022 14945904 12983467 14693081 14304601 14188437 13491106
   14601126 15212749 15014034 15990952 14764266 15149381
2023 16138146 14519304 15354377 14982265 15777909 16057262
   17312824 17063068 17515200 17463214 17544090 17669743
2024 17344377 15898965 16541658 15236163 16352784 16272346
   15931278 16474641 16562424 16121037 15799691 16861902
2025 16997022 15527775 17845551 17323715 18132705
```

Referências Bibliográficas

BOX, G. E. P.; JENKINS, G. M. **Time Series Analysis: Forecasting and Control**. San Francisco: Holden-Day, 1970. Tradução recomendada: Análise de Séries Temporais: Previsão e Controle.

MORETTIN, P. A.; TOLOI, C. M. C. **Análise de séries temporais**. 2. ed. São Paulo: Blucher, 2006.

Posit team. **RStudio: Integrated Development Environment for R**. Boston, MA, 2024. RStudio version 2024.12.0+467, "Kousa Dogwood". Disponível em: <http://www.posit.co/>.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2024. R version 4.4.2 (2024-10-31). Disponível em: <https://www.R-project.org/>.