



---

# Análise de Séries Temporais

A2 - Séries Temporais

---

FGV EMAp

Ana Júlia Amaro Pereira Rocha  
Henrique Borges Carvalho  
Maria Eduarda Mesquita Magalhães  
Mariana Fernandes Rocha  
Paula Eduarda de Lima

Ciência de Dados e Inteligência Artificial  
6º Período

Rio de Janeiro, 2025

## Discussão sobre métricas e métodos de avaliação

Existem duas vertentes principais de métricas de erro para previsões: métricas absolutas e métricas percentuais. Cada grupo avalia o desempenho de forma diferente e é adequado para contextos distintos.

**Métricas absolutas** (MAE, MSE, RMSE) comparam previsões e valores reais na escala original da série. São estáveis quando os valores podem ser pequenos ou variar muito ao longo do tempo, pois não dependem de proporções. Têm interpretação direta (“erro médio em unidades da variável”).

**Métricas percentuais** (MAPE e sMAPE) medem erro relativo, comparando o erro ao tamanho do valor observado. São úteis quando a escala da série muda e quando o interesse é avaliar o erro como proporção do valor real. Porém, distorcem resultados quando há valores muito pequenos, já que o denominador torna o erro artificialmente grande.

Cada vertente oferece uma perspectiva diferente: métricas absolutas focam no desvio real; métricas percentuais focam na relevância relativa do erro.

Dentro do nosso contexto serão usadas métricas absolutas que preservam a escala original dos dados, permitindo comparar diretamente as diferenças entre valores observados e previstos. Essa escolha facilita a interpretação dos resultados, mantendo a coerência com a unidade de medida da variável analisada. Essas métricas permanecem adequadas mesmo com covariáveis, pois não dependem da escala relativa dos valores observados.

Além disso, existem diferentes formas de treinar o modelo para previsões em séries temporais. Os principais tipos são:

**One-shot:** Gera todos os passos futuros de uma vez. A entrada é a janela histórica, a saída é um vetor contendo todos os pontos de previsão. Essa método evita acúmulo de erro, porém mais difícil de treinar quando o horizonte é grande.

**Iterativo(One-Step-Ahead Rolling Forecast):** Prevê um passo por vez. Para prever o passo seguinte, usa como entrada a própria previsão anterior. Gera  $\hat{y}(t+1)$ , depois usa  $y(t+1)$ , se disponível ou  $\hat{y}(t+1)$ , para gerar  $\hat{y}(t+2)$ , e assim por diante. Uma desvantagem dessa abordagem é o erro que se acumula a cada passo.

**Lag(Cross Validation):** O histórico recente vira feature e o próximo valor vira alvo, a cada passo a janela anda também. Um ponto é que é flexível e compatível com qualquer modelo, porém é sensível a escolha dos lags. Com covariáveis, é possível usar valores no mesmo instante da variável-alvo (contemporâneos), conhecidas no presente, e também versões defasadas dessas covariáveis, quando só é conhecido o passado e não o presente.

## Discussão sobre a necessidade de transformação de variáveis

A série apresenta **heterocedasticidade**, ou seja, a variância dos valores não é constante — tende a aumentar conforme o nível da série se eleva. Essa característica é comum em séries com crescimento de natureza exponencial ou multiplicativa, e pode comprometer o desempenho de modelos que assumem variância constante.

Assim, é recomendada a aplicação de uma **transformação de variáveis**, especialmente a *transformação logarítmica* ou a **transformação de Box-Cox**. Testamos o uso das duas transformações nos dados, a segunda se saiu melhor na regressão e o desempenho de ambas nos erros dos modelos baselines não teve resultados significativos.

## Discussão sobre a decomposição entre tendência e sazonalidade

A série apresenta **tendência de longo prazo** ( $T_t$ ) e **sazonalidade periódica** ( $S_t$ ). Assim, a série pode ser representada por:

$$y_t = T_t + S_t + R_t$$

onde  $R_t$  é o componente residual ou ruído. Então, essa decomposição seria necessária para, principalmente, isolar a tendência estrutural, identificar padrões sazonais, remover o efeito da sazonalidade e analisar o componente residual.

Após a decomposição, a série ajustada sazonalmente mostrou-se muito semelhante à série original o que indica que a tendência sobrepõe a sazonalidade na série temporal em questão, há pouca ou nenhuma sazonalidade.

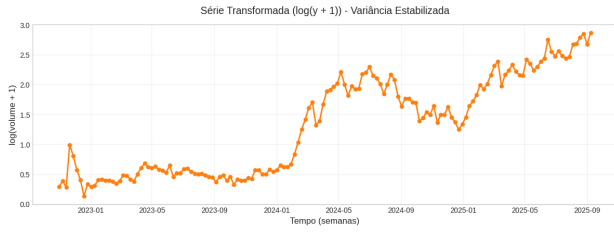


Figure 1: Série transformada



Figure 2: Série ajustada sazonalmente

## Modelos Baseline

Modelos baseline são abordagens de previsão simples que servem como um ponto de referência para avaliar o desempenho de modelos mais complexos. A sua principal função é estabelecer um limiar de performance: se um modelo sofisticado não conseguir superar um baseline simples, isso sugere que o modelo mais complexo não está capturando bem os padrões da série ou que a série é inerentemente difícil de prever. Para este trabalho, foram implementados cinco modelos baseline (Mean, Naive, Seasonal, Drift e Rolling Mean), cujos resultados foram comparados utilizando as métricas MAE, RMSE e MAPE.

## Metodologia de Validação

Para avaliar o desempenho dos modelos de forma robusta, foi empregada a técnica de *Time Series Cross-Validation* com janela expansível. Esta abordagem consiste em:

- **Janela inicial de treino:** 120 observações
- **Horizonte de previsão:** 4 pontos por fold
- **Passo de deslocamento:** 1 observação entre folds consecutivos
- **Número de folds:** 34 folds

Em cada fold, o modelo é treinado com os dados até o ponto de corte e realiza previsões *one-step-ahead rolling forecast* para o horizonte de 4 pontos. Nesta estratégia, ao prever o ponto  $h + 1$ , utilizamos todos os dados observados até o ponto  $h$  (incluindo os valores reais do teste até esse ponto), garantindo uma avaliação realista da capacidade preditiva dos modelos.

A Figura 3 ilustra o comportamento de todos os modelos baseline no último fold, permitindo visualizar como cada método se comporta em relação aos valores observados.

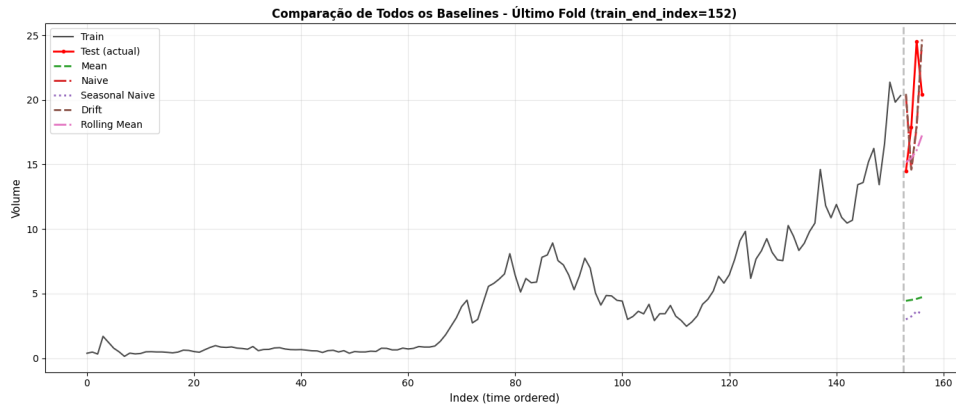


Figure 3: Ultimo fold

## Resultados

Os resultados obtidos para as métricas de erro, calculados como a média sobre os 8 folds de validação cruzada, estão consolidados na Tabela 1. O **Drift Method** apresentou o melhor desempenho geral, seguido de perto pelo Naive Method.

Modelo	MAE	RMSE	MAPE (%)
Drift Method	1.734616	2.047200	14.25
Naive Method	1.747206	2.053144	14.30
Rolling Mean (K=12)	2.480043	2.904220	18.88
Seasonal Naive	6.507206	6.658787	49.04
Mean Method	8.655486	8.793149	69.88

Table 1: Comparação de baselines

O modelo com melhor desempenho (**Drift Method**) foi selecionado para uma análise de resíduos detalhada, apresentada na próxima subseção.

Para aplicações práticas, saber a margem de erro é tão importante quanto a previsão em si. Avaliamos os intervalos de confiança (IC) de 95% gerados pelo Drift Method utilizando métricas probabilísticas robustas. A Figura 4 apresenta a previsão visual juntamente com a faixa de incerteza estimada.

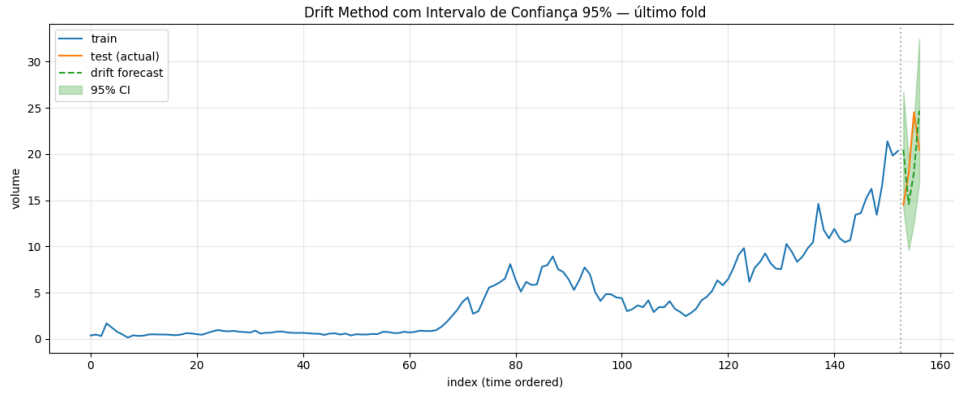


Figure 4: Previsão *one-step-ahead* do Drift Method no último fold, exibindo o Intervalo de Confiança de 95% (área sombreada). A caixa de texto destaca as métricas de calibração do intervalo.

As métricas obtidas demonstram que o estimador de incerteza do Drift Method é confiável:

- **Cobertura (Coverage):** O modelo obteve **75.00%**, um valor abaixo do nível de confiança nominal de 95%. Isso indica que o modelo apresenta *overconfidence* (excesso de confiança), gerando intervalos de incerteza mais estreitos do que a real volatilidade da série exige. O modelo subestima o risco de caudas (eventos extremos).
- **Winkler Score (22.49):** O score aumentou em relação aos testes anteriores, refletindo as penalidades pelas violações do intervalo. Embora a Largura Média ( $\approx 12.28$ ) busque um equilíbrio, o score indica que, para este horizonte de curto prazo, seria preferível um intervalo mais largo para garantir maior segurança (cobertura) nas previsões.
- **Violação (25.00%):** A taxa de violação é de 25.00%, significativamente acima dos 5% esperados teoricamente. Isso significa que, em média, uma a cada quatro previsões cai fora da margem de erro estimada. Este resultado justifica estatisticamente a necessidade de modelos mais sofisticados (como ARIMA com erros não-gaussianos ou GARCH) para capturar adequadamente a dispersão dos dados.

Esta validação da incerteza reforça a escolha do Drift Method como um benchmark sólido: ele não apenas prevê a tendência corretamente, mas também fornece uma estimativa de risco realista para o negócio.

## Análise de Resíduos do Melhor Modelo (Drift)

A análise de resíduos do melhor modelo baseline foi realizada em três etapas:

**1. Série Temporal dos Resíduos:** A série temporal dos resíduos parece apresentar algum padrão temporal, podemos ver alguns intervalos que se repetem em direção e diferem pouco em amplitude. A maioria dos resíduos concentrados em torno de zero. No entanto, a variância pareceu aumentar ligeiramente em alguns períodos, sugerindo possível heteroscedasticidade.

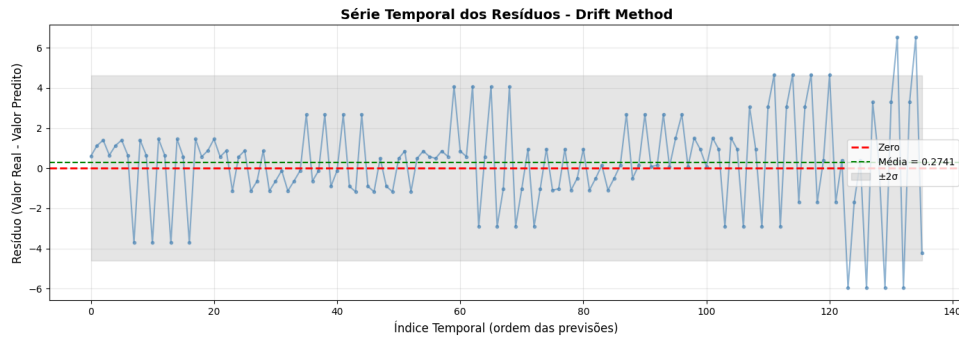


Figure 5: Série Temporal dos Resíduos

**2. Normalidade dos Resíduos:** A análise de normalidade através do histograma e gráfico Q-Q revelou desvios em relação à distribuição normal, especialmente nas caudas. Testes estatísticos confirmaram a rejeição da hipótese de normalidade, indicando que, embora o modelo capture a tendência central (média), ele pode não capturar perfeitamente os eventos extremos.

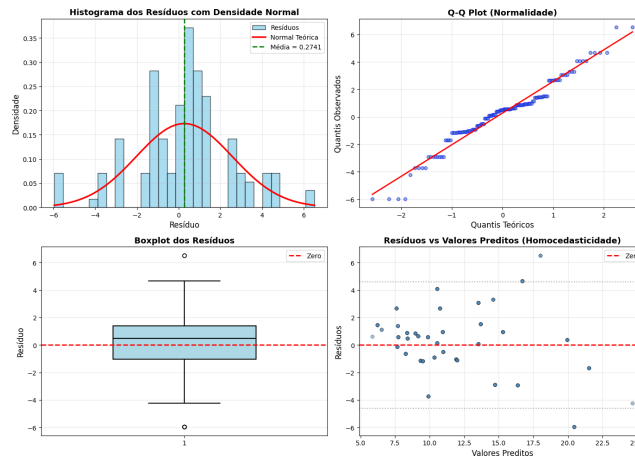


Figure 6: Normalidade dos Resíduos

**3. Autocorrelação dos Resíduos:** A análise de autocorrelação (ACF) mostrou correlação serial significativa em alguns lags iniciais. Isso era esperado para um modelo baseline simples e indica que há estrutura temporal de curto prazo remanescente que não foi capturada apenas pela tendência linear do Drift.

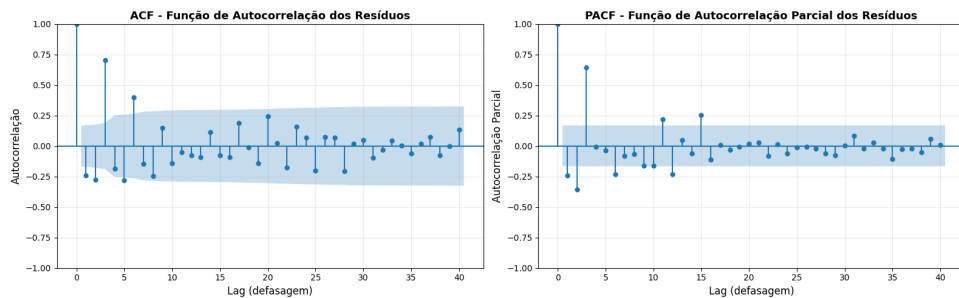


Figure 7: Autocorrelação dos Resíduos

## Transformação Box-Cox e Heteroscedasticidade

A análise inicial dos dados sugeriu heteroscedasticidade. Para mitigar este problema, aplicou-se a transformação de Box-Cox.

Após aplicar a transformação, os modelos foram reavaliados. Os resultados mostraram diferenças marginais nas métricas de erro.

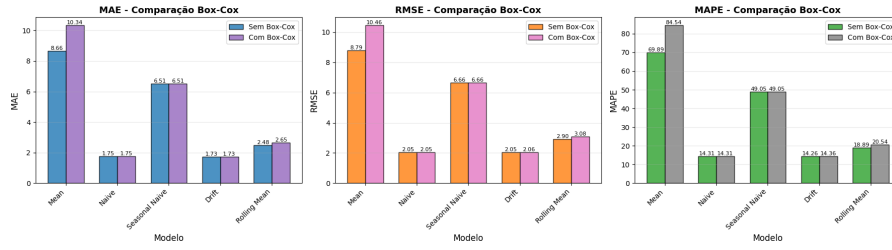


Figure 8: Comparação: Box-Cox nos Baselines

Esta pequena variação pode ser explicada por dois fatores principais:

- **Dominância da Tendência:** A série possui uma tendência linear forte que é bem capturada pelo Drift Method independentemente da escala de variância.
- **Horizonte Curto:** Com previsão *one-step-ahead*, o modelo é re-calibrado a cada passo, reduzindo o impacto negativo da heteroscedasticidade que seria mais visível em horizontes longos.

## Modelos de Regressão linear

Foi realizada uma análise completa de séries temporais utilizando regressão linear simples e múltipla. O objetivo geral foi capturar componentes de tendência, sazonalidade e autocorrelação, além de comparar o desempenho preditivo entre os modelos. Além disso, comparamos desempenhos em diferentes tipos de cross-validations e também utilizando transformações log e Box-Cox.

**Modelo Simples:** Regressão linear simples utilizando apenas a variável temporal  $t$ . O modelo captura uma tendência linear ao longo do tempo para explicar o volume.

$$\hat{y}_t = \beta_0 + \beta_1 t$$

**Modelo Múltiplo** Regressão linear múltipla com variáveis explicativas  $t$ ,  $inv$  e  $users$ . O modelo captura não só a tendência temporal, mas também a influência das variáveis externas.

$$\hat{y}_t = \beta_0 + \beta_1 t + \beta_2 inv_t + \beta_3 users_t$$

**Modelo Múltiplo + Sazonalidade** Extensão do modelo múltiplo, adicionando variáveis dummy de sazonalidade do mês. Captura padrões cíclicos periódicos além da tendência e efeitos das variáveis externas.

$$\hat{y}_t = \beta_0 + \beta_1 t + \beta_2 inv_t + \beta_3 users_t + \sum_{i=1}^{11} \gamma_i Season_i$$

## Comparação dos Modelos

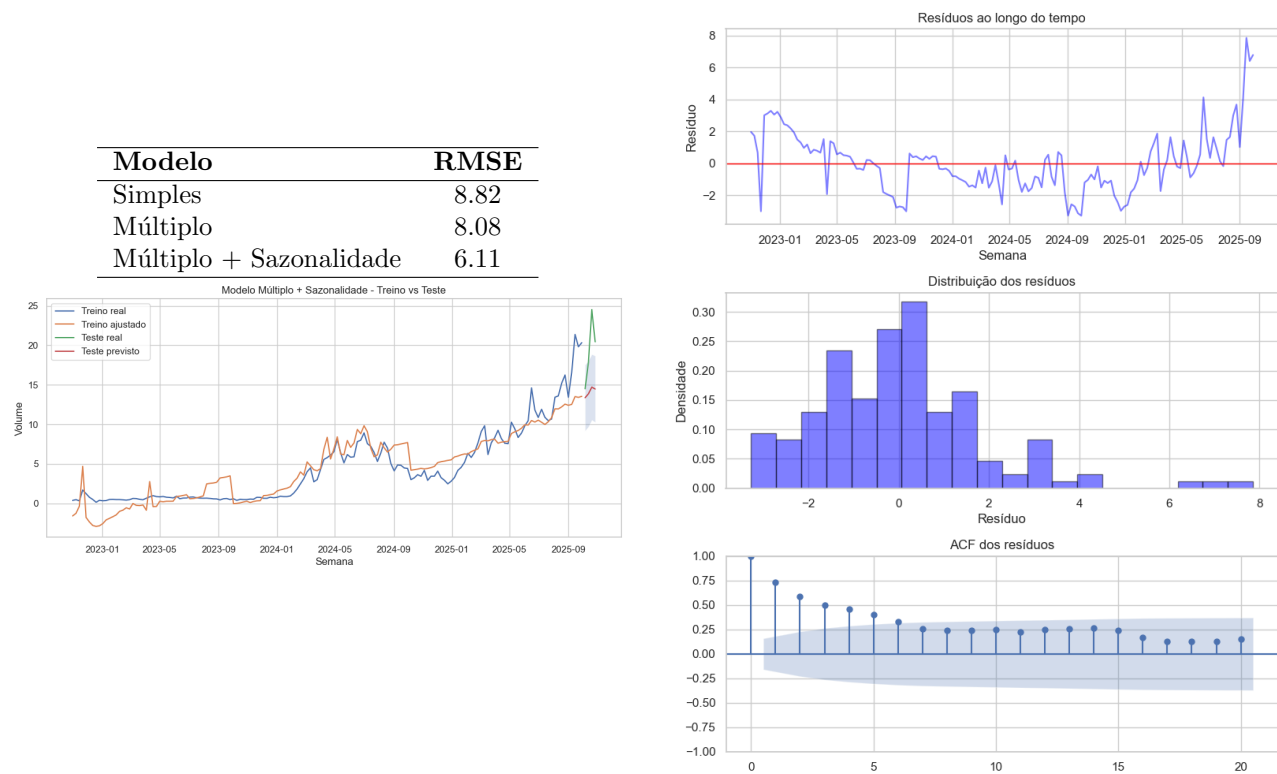


Figure 9: Comparação dos modelos de regressão: tabela de RMSE e figuras do modelo múltiplo ajustado e análise dos resíduos.

Podemos observar que os resíduos não estão completamente de acordo com as premissas de normalidade e homocedasticidade, apresentando aumento significativo no final da série. Isso indica que o modelo ainda não capturou totalmente todas as características da série temporal, como tendências locais, mudanças abruptas ou efeitos sazonais residuais.

## Transformações

Com o intuito de melhorar esse ajuste e resíduos, aplicamos transformações log e Box-cox:

**Regressão com Transformação Box-Cox** Reduz assimetria, melhora homocedasticidade e tende a produzir resíduos mais próximos de uma distribuição normal, o que aumenta a confiabilidade das inferências e intervalos de confiança.

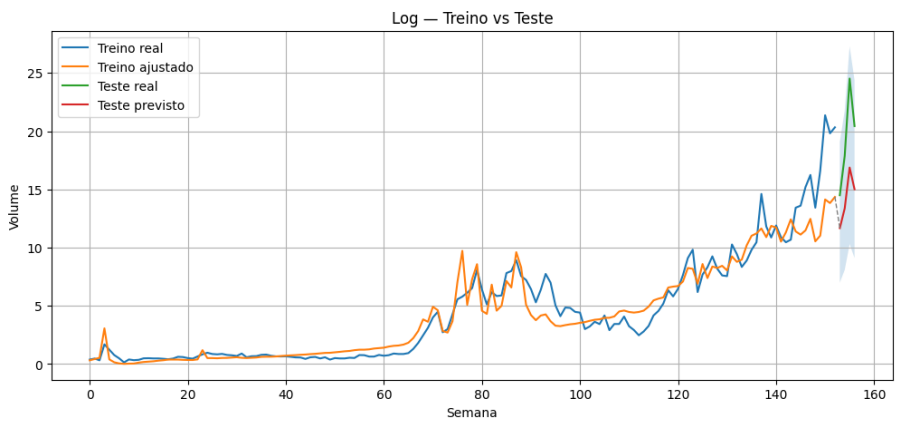
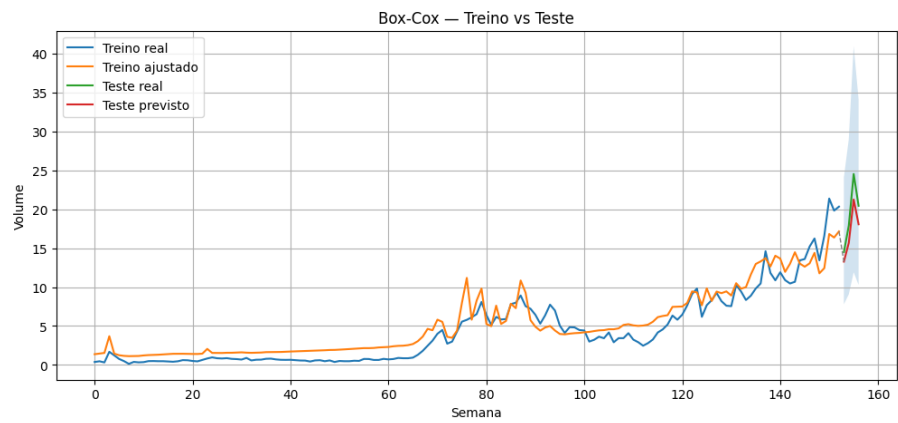
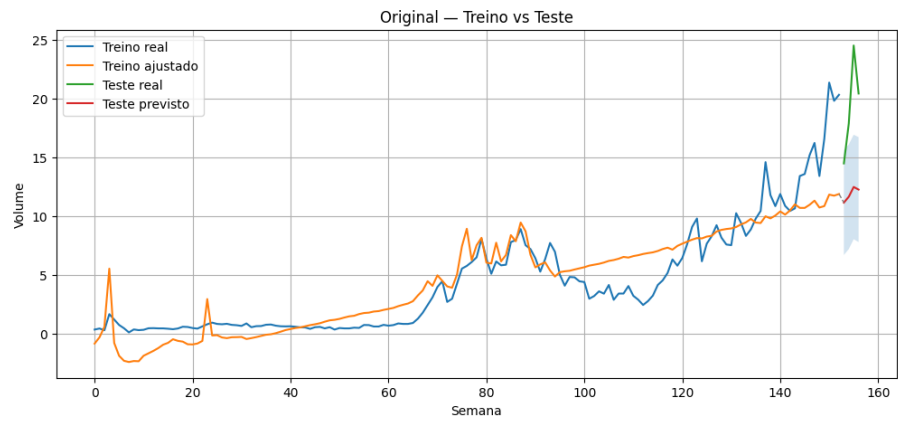
**Regressão com Transformação Logarítmica:** Similar à Box-Cox, é útil para dados positivos e assimétricos, simplifica efeitos multiplicativos em termos aditivos no modelo linear.

## Comparação Final dos Modelos

A tabela a seguir apresenta as métricas de desempenho (RMSE, MAE e MAPE) calculadas no conjunto de teste para cada abordagem:

Modelo	RMSE	MAE	MAPE (%)
Box-Cox	2.386	2.273	11.46
Log	5.389	5.101	25.60
Original	8.077	7.434	36.68

Table 2: Comparação de desempenho das transformações aplicadas na regressão.





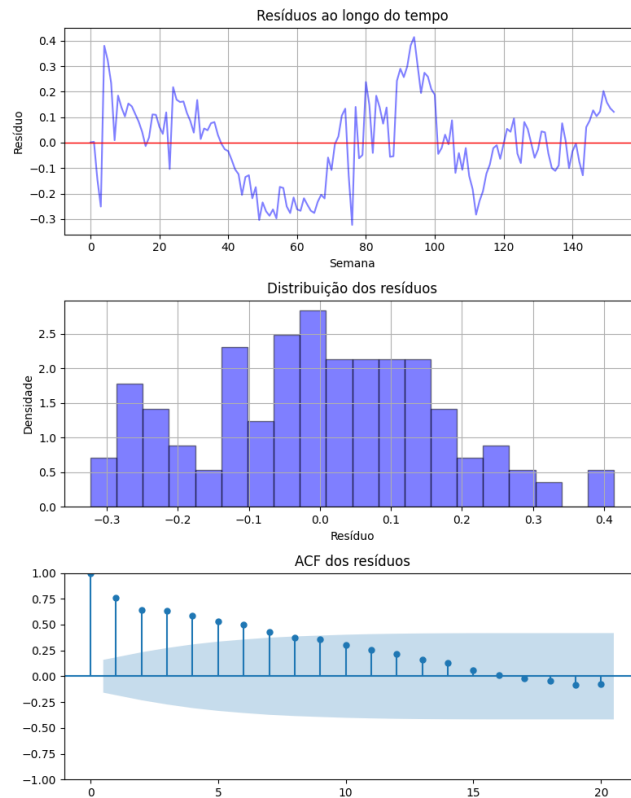


Figure 10: Análise dos resíduos do melhor modelo: com Box-Cox

A transformação Box-Cox proporcionou a melhor performance, reduzindo significativamente erros de previsão e estabilizando a variância, enquanto o modelo original apresentou o pior desempenho devido à alta assimetria e heterocedasticidade da série. O log também melhorou a performance, mas de forma menos expressiva que o Box-Cox.

Outros testes que fizemos foi a análise de diferentes tipos de cross-validation para o treinamento com dados temporais:

**Rolling Window (Janela Deslizante):** Nesse método, o modelo é treinado usando uma janela fixa de observações consecutivas e, em seguida, faz-se a previsão para o próximo ponto (ou horizonte  $h$ ). Depois, a janela "desliza" uma posição para frente e o processo se repete.

**Expanding Window (Janela Expansiva):** Aqui, o conjunto de treino cresce a cada iteração, começando com uma janela inicial e adicionando novos pontos conforme avançamos no tempo.

**TimeSeriesSplit (Cross-Validation para séries temporais):** É uma generalização das duas anteriores. Divide a série em vários "folds" temporais: Cada fold mantém a ordem temporal (não embaralha os dados). Cada iteração usa uma parte inicial como treino e a seguinte como teste.

## Modelos Autorregressivos

Foram testados diferentes modelos de séries temporais para prever o volume semanal, avaliando o desempenho no conjunto de teste ( $h = 4$  semanas).

### Modelos AR, MA e ARMA

- **AR( $p$ ) — Autoregressivo:** Assume que o valor atual da série depende linearmente dos  $p$  valores passados:  $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \epsilon_t$  onde  $\epsilon_t$  é ruído branco. Foram testados AR(1), AR(2) e AR(3).
- **MA( $q$ ) — Média Móvel:** Considera que o valor atual da série depende de uma média ponderada dos erros passados:  $y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q}$  Foram avaliados MA(1) e MA(2) para capturar choques recentes.

- **ARMA(p,q) — Combinação AR e MA:** Combina dependência linear de valores passados e efeitos de choques recentes. Testado: ARMA(1,1), combinando AR(1) e MA(1).

## Modelos SARIMA

SARIMA é uma extensão do ARIMA que incorpora componentes sazonais:  $\text{SARIMA}(p, d, q)(P, D, Q)_s$

- $p, d, q$ : ordem AR, diferença integrada e MA para a série não sazonal.
- $P, D, Q$ : ordem AR, diferença e MA para a componente sazonal.
- $s$ : período sazonal (52 semanas, neste caso).

Modelos testados:

- SARIMA(1,0,0)(1,0,0)[52]: captura tendência anual e autocorrelação simples.
- SARIMA(1,0,1)(1,0,0)[52]: combina AR e MA na série não sazonal com componente sazonal, modelando padrões mais complexos.

Modelo	RMSE	MAE	MAPE (%)
SARIMA(1,0,1)(1,0,0)[52]	3.6566	3.1336	17.3268
SARIMA(1,0,0)(1,0,0)[52]	3.6738	3.1562	17.4426
ARMA(1,1)	3.7949	3.2103	18.0513
AR(2)	3.8019	3.2230	18.1081
AR(1)	3.8166	3.2426	18.2096
AR(3)	3.9055	3.2369	18.4547
MA(2)	13.5649	11.8699	56.9861
MA(1)	14.4880	12.8108	61.8514

Table 3: Comparação dos modelos de séries temporais com base nas métricas RMSE, MAE e MAPE.

O modelo que teve menor erro foi o SARIMA(1,0,1)(1,0,0).

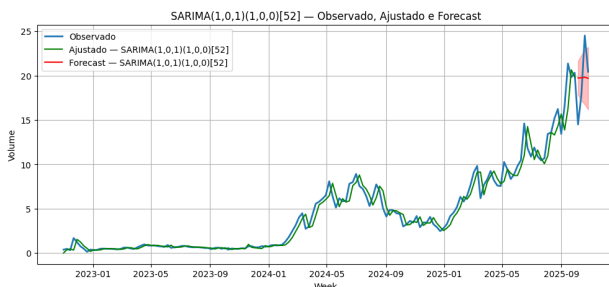


Figure 11: Ajuste do melhor modelo - SARIMA(1,0,1)(1,0,0)

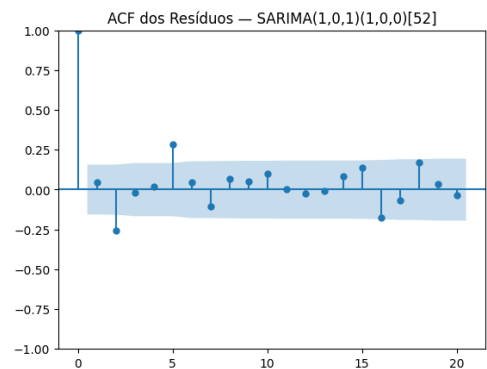


Figure 12: ACF - SARIMA(1,0,1)(1,0,0)

## Modelos de suavização exponencial

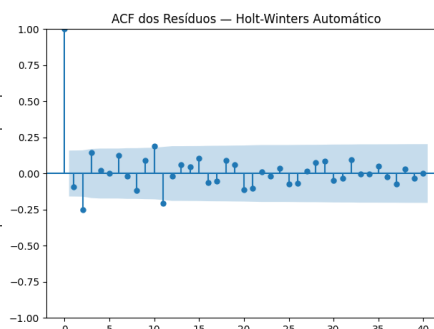
**Modelos de suavização exponencial** atualizam previsões dando mais peso às observações recentes. A versão simples ajusta apenas o nível; a dupla incorpora tendência; a tripla inclui tendência e sazonalidade.

Nos experimentos, o SES funcionou melhor com valores altos de  $\alpha$ , pois respondeu rápido às mudanças, o ajuste automático teve desempenho semelhante. No método de Holt, combinações moderadas de  $\alpha$  e  $\beta$  pioraram a estabilidade e aumentaram os erros; o ajuste automático encontrou um equilíbrio com  $\alpha$  alto e  $\beta$  baixo. No Holt-Winters, parâmetros moderados tiveram bom desempenho, enquanto valores agressivos aumentaram muito o erro; o ajuste automático teve o melhor MAE.

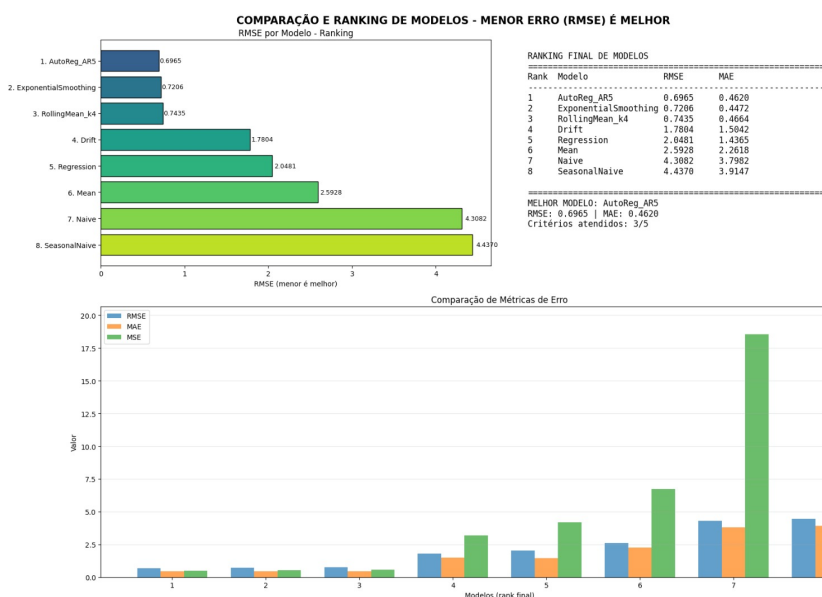
O melhor desempenho geral foi do Holt-Winters, indicando que nível, tendência e sazonalidade são componentes importantes da série. Nos resíduos, a ACF mostra padrão forte no SES, padrão menor na suavização dupla e comportamento próximo de ruído branco no Holt-Winters, reforçando que o modelo triplo capturou melhor a estrutura temporal.

Modelo	MAE
SES automático	2.593354
Holt automático	2.525490
Holt-Winters automático	2.344490

Desempenho dos modelos



## Análises de resíduos e ajuste dos modelos



A análise realizada teve como objetivo avaliar a qualidade dos resíduos dos modelos de previsão verificando se atendiam a pressupostos essenciais como média zero, normalidade, ausência de autocorrelação e homocedasticidade. Para isso, foi utilizada a classe *ResidualAnalysis*, que aplicou testes estatísticos formais e gerou múltiplos gráficos diagnósticos, permitindo uma avaliação robusta do comportamento residual em oito modelos distintos, incluindo abordagens baselines, estatísticas e de regressão.

Os critérios de avaliação envolveram testes como Anderson-Darling, Ljung-Box, ARCH e a estatística de Durbin-Watson, complementados por métricas descritivas dos resíduos. Os resultados mostraram que o modelo AutoReg\_AR5 apresentou o melhor desempenho geral, com resíduos bem comportados e excelente acurácia preditiva, seguido pelo ExponentialSmoothing, que obteve resultados satisfatórios, embora com leve indício de heterocedasticidade. Em contraste, os modelos baseline apresentaram erros substancialmente maiores, evidenciando sua inadequação para a série analisada.

A integração entre métricas de erro e qualidade dos resíduos permitiu estabelecer um ranking final dos modelos, no qual o AutoReg\_AR5 se destacou por captar eficientemente a estrutura temporal dos dados. Recomenda-se seu uso para previsões futuras, embora o ExponentialSmoothing permaneça como alternativa robusta. Sugere-se ainda considerar uma combinação ponderada entre os dois melhores modelos, de modo a aumentar a estabilidade e robustez das previsões em aplicações práticas.

## Repositório Github

Link: [Github](#)