Fundação Getúlio Vargas
Escola de Matemática Aplicada

Séries Temporais

Relatório da A2 de Séries Temporais

Leonardo Alexandre da Silva Ferreira
Guilherme Moreira Castilho

Rio de Janeiro
Dezembro / 2025

Sumário

1 Introdução

Este trabalho apresenta uma análise detalhada de séries temporais para modelar a variável `volume` utilizando técnicas fundamentais de previsão. A metodologia empregada inclui modelos *baseline*, transformação de variáveis, decomposição temporal (STL), modelos de suavização exponencial, modelos ARIMA/SARIMA/SARIMAX e regressão linear múltipla com covariáveis exógenas (`inv` e `users`).

O objetivo principal é desenvolver um modelo preditivo robusto para a variável `volume`, considerando suas características temporais, sazonalidade, tendências e a relação com as covariáveis. A abordagem metodológica visa: estabelecer *benchmarks* utilizando modelos *baseline* simples; estabilizar a variância por meio de transformações adequadas; capturar sazonalidade e tendência através de decomposição STL; desenvolver modelos preditivos utilizando técnicas avançadas (suavização exponencial, ARIMA/SARIMA/SARIMAX); incorporar covariáveis exógenas para melhorar a capacidade preditiva; validar a adequação dos modelos por meio de análise de resíduos.

2 Metodologia

Inicialmente, foram implementados modelos *baseline* para estabelecer benchmarks de desempenho: **Mean** (média simples dos dados históricos); **Naive** (último valor observado); **SNaive** (último valor da mesma estação); **Drift** (tendência linear baseada no primeiro e último valor). Adicionalmente, foram aplicados modelos de suavização exponencial (SES, Holt, Holt-Winters), modelos ARIMA, SARIMA e SARIMAX e regressão linear múltipla com covariáveis exógenas.

Para avaliação dos modelos, foram utilizadas as métricas: **MAE** (Erro Absoluto Médio), **RMSE** (Raiz do Erro Quadrático Médio), **MAPE** (Erro Percentual Absoluto Médio) e **MASE** (Erro Absoluto Escalado Médio).

A validação foi realizada utilizando uma estratégia de divisão temporal, com o último ano (52 semanas) reservado para teste e os dados anteriores para treino. Para os modelos de regressão, foi aplicada uma estratégia de previsão iterativa em blocos: inicialmente, o modelo é treinado com todas as semanas exceto as últimas 52; em seguida, são previstas as próximas 4 semanas; essas previsões são incorporadas ao conjunto de treino e o modelo é reajustado para prever as 4 semanas subsequentes; este processo é repetido iterativamente até que todas as 52 semanas sejam previstas. Esta abordagem simula condições reais de previsão, onde previsões anteriores são incorporadas ao histórico disponível para previsões futuras, oferecendo **realismo** (simula condições reais de previsão), **robustez** (evita *overfitting*) e **interpretabilidade** (resultados mais confiáveis).

Para análise dos resíduos, foram utilizados: **Gráficos de Autocorrelação (ACF)** e **Teste de Ljung-Box** para verificar a não correlação dos resíduos, e **Histograma dos Resíduos** para verificar se a distribuição dos erros é aproximadamente normal e centrada em zero.

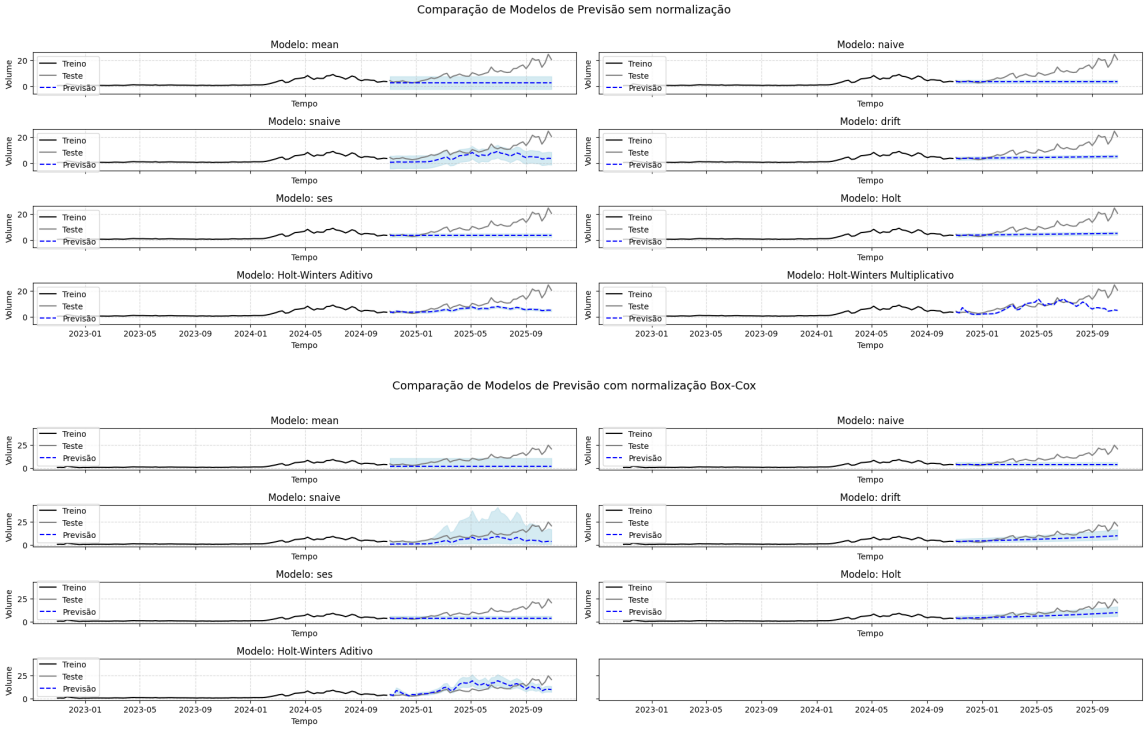
3 Transformação de Variáveis

A análise inicial das séries temporais revelou heteroscedasticidade (variância não constante ao longo do tempo) na variável `volume`, necessitando de uma transformação para estabilizar a variância. Para a variável `volume`, foi aplicada a transformação **Box-Cox** com parâmetro $\lambda = 0.0381$.

A aplicação da transformação Box-Cox resultou em melhorias expressivas nos modelos baseline e de suavização exponencial. Conforme mostrado nas métricas, modelos como *Holt-Winters Aditivo* passaram a apresentar desempenho substancialmente melhor na escala transformada, alcançando o menor MAE (0,419), RMSE (0,494) e MAPE (20,252%). Em contraste, na escala original, o melhor desempenho havia sido do *Holt-Winters Multiplicativo* (MAE 3,793; RMSE 6,029; MAPE 33,555%).

Para as covariáveis exógenas `inv` e `users`, foi aplicada a transformação **logarítmica** ($\log(x)$), visando estabilizar a variância e reduzir a assimetria dessas variáveis. A transformação logarítmica das covariáveis permite uma melhor incorporação dessas variáveis nos modelos que utilizam covariáveis

exógenas, como o **SARIMAX**, facilitando a interpretação dos coeficientes e melhorando a estabilidade numérica do modelo.



Observação: O modelo **Holt-Winters Multiplicativo** não foi incluído na comparação após a aplicação da transformação Box-Cox porque essa transformação gerou valores negativos na série transformada. Modelos com sazonalidade multiplicativa exigem que todos os valores sejam estritamente positivos. Portanto, a utilização do Holt-Winters Multiplicativo torna-se inadequada nesse contexto, inviabilizando sua estimação e, consequentemente, sua representação nos gráficos.

4 Decomposição STL

Após as transformações aplicadas às variáveis temporais, a decomposição STL foi utilizada para examinar tendência, sazonalidade e resíduos em escalas mais estáveis e adequadas à modelagem.

A variável transformada `volume` apresentou tendência suavemente crescente e sazonalidade regular de baixa amplitude. Os resíduos mostraram-se bem distribuídos ao redor do zero, indicando estabilização da variância e boa separação dos componentes.

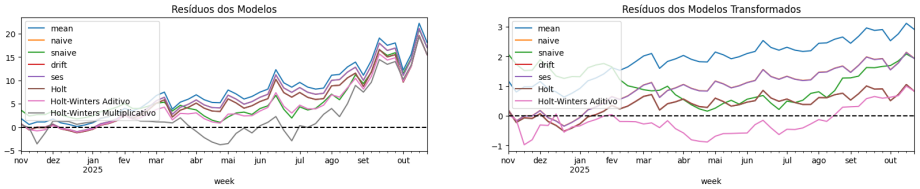
A decomposição da variável transformada `inv` revelou uma tendência levemente crescente até o final do terceiro trimestre de 2024, seguida de discreta redução. A sazonalidade manteve oscilações moderadas, enquanto os resíduos exibiram variações abruptas pontuais.

Em relação à variável transformada `users`, a tendência mostrou crescimento contínuo e sazonalidade com oscilações de baixa amplitude. Os resíduos apresentaram dispersão moderada e comportamento mais regular após a transformação logarítmica.

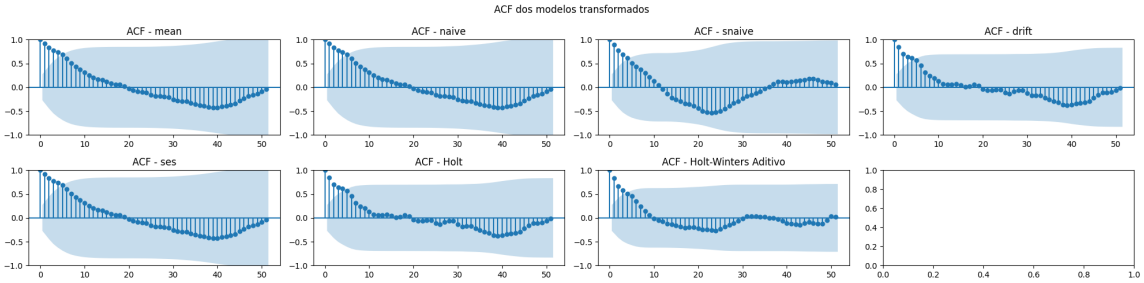
As decomposições indicam componentes mais claros e resíduos mais estáveis nas séries transformadas, reforçando a adequação das transformações para fins de modelagem e previsão.

5 Análise de Resíduos e Ajuste dos Modelos

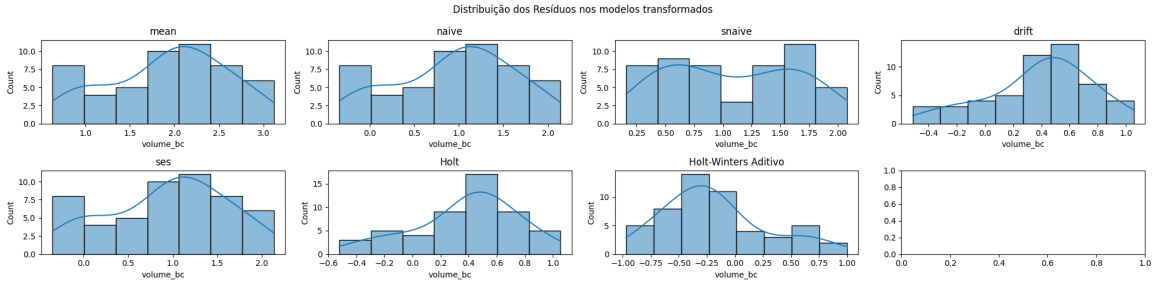
A transformação Box-Cox reduziu a variabilidade dos resíduos em todos os modelos, vemos que



Ao observar os gráficos de ACF dos modelos transformados, os que apresentam o melhor comportamento residual são: Drift, Holt e Holt-Winters Aditivo.



Ademais, seus resíduos são os que mais se aproximam de uma distribuição normal centrada em 0.



6 Modelos ARIMA, SARIMA e SARIMAX

Inicialmente, foram ajustados modelos ARIMA, SARIMA e SARIMAX utilizando configurações padrão. Observa-se que o modelo SARIMA melhora o desempenho do ARIMA ao incorporar o componente sazonal, enquanto o modelo SARIMAX, ao incluir as variáveis exógenas *inv* e *users*, apresenta desempenho superior ao SARIMA.

Em seguida, foi realizado um *grid search* para o modelo ARIMA, resultando em uma configuração que apresentou desempenho significativamente superior aos modelos baselines e aos modelos de suavização..

7 Modelos de Regressão Linear Múltipla

A regressão por Mínimos Quadrados Ordinários (OLS) foi utilizada como abordagem principal para modelagem da série.

Como as transformações das variáveis se mostraram eficientes, as regressões foram feitas sobre todas variáveis transformadas: a variável alvo (volume) com transformação Box-Cox; e as variáveis exógenas (inv, e users) com normalização log.

Para enriquecer o modelo e capturar diferentes aspectos da dinâmica da série, foram incluídas variáveis derivadas, tanto da variável alvo, quanto das variáveis exógenas, sendo elas: variaveis de calendário (ano, mês, semana do ano) que permitem calcular efeitos sazonais e padrões recorrentes associados ao calendário; defasagem (lags de 1, 2 e 52 períodos no volume e lags de 1 e 4 nas exógenas) que incorporam dependência temporal de curto e longo prazo; Diferenças a cada semana nas exógenas (”user_diff”e ”inv_diff”), capturando a variação absoluta; Interações entre as variáveis para poder avaliar o efeito conjunto das variáveis exógenas.

Para avaliar a robustez do modelo e selecionar o melhor conjunto de variáveis, foi aplicada a lógica de validação cruzada leave-one-out adaptada ao contexto temporal, de forma que não haja vazamento de informação. A vantagem é obter uma medida mais confiável do erro fora da amostra, reduzindo o risco de overfitting e permitindo comparar diferentes especificações de forma justa.

Os modelos levaram em conta um set de previsão/teste de um ano e todo o tempo que antecede esse período foi para o set de treino.

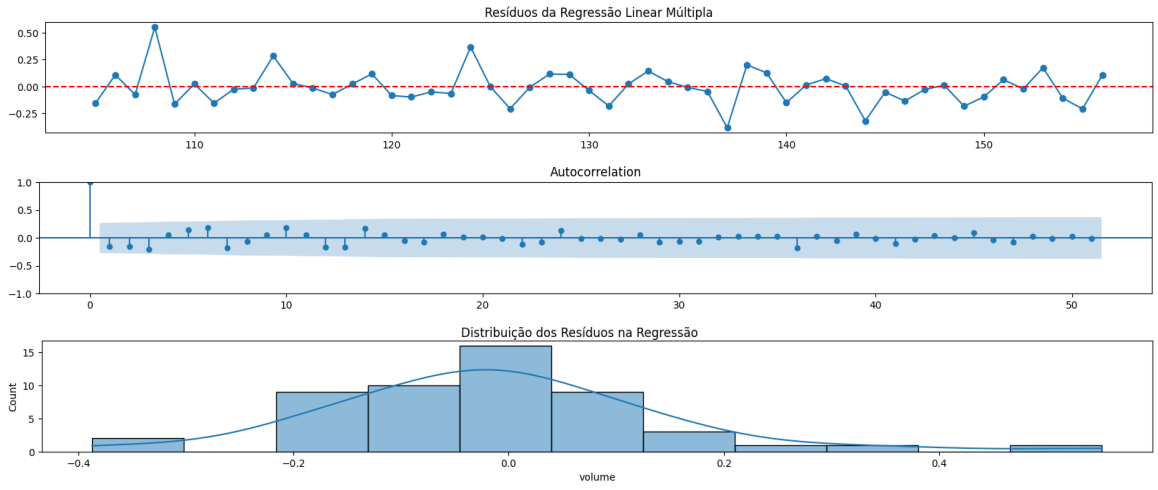
Seguindo a ideia de Leave-One-Out, foram testadas todas as combinações de features das derivações feitas. O ranqueamento foi feito pela média e desvio padrão dos erros de MAE e RMSE em cada set da Cross-Validation. Com isso, as features selecionadas foram "month", "lag1", "lag2"e "user_diff".

O desempenho apresentado pelo modelo resultante sobre a transformação box-cox da variável "volume"foi:

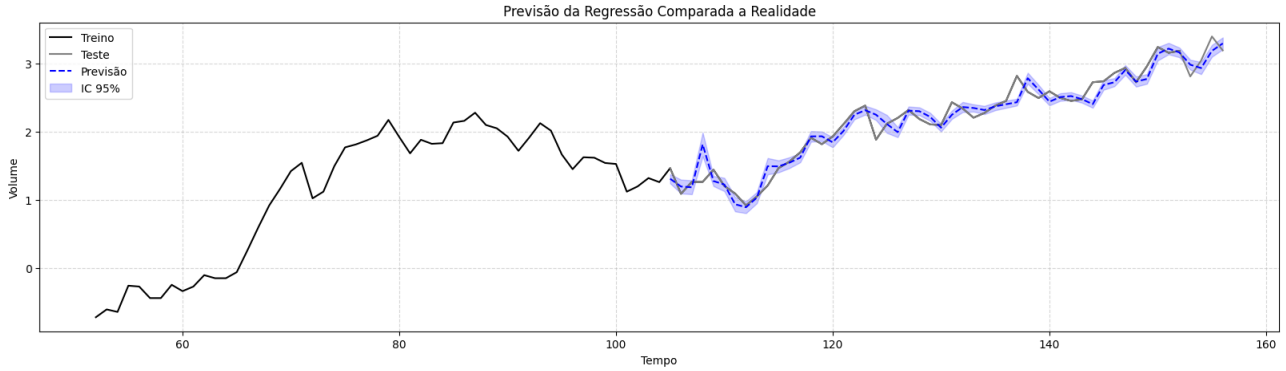
```
MAE:  0.11353090660804274
RMSE:  0.1579308911321089
                            OLS Regression Results
==============================================================================
Dep. Variable:                 volume   R-squared:                       0.979
Model:                            OLS   Adj. R-squared:                  0.978
Method:                 Least Squares   F-statistic:                     1131.
Date:                Mon, 01 Dec 2025   Prob (F-statistic):           8.30e-80
Time:                        23:06:27   Log-Likelihood:                 56.315
No. Observations:                 101   AIC:                            -102.6
Df Residuals:                      96   BIC:                            -89.55
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0818      0.042      1.969      0.052      -0.001       0.164
month         -0.0059      0.004     -1.367      0.175      -0.014       0.003
lag1           1.0045      0.078     12.923      0.000       0.850       1.159
lag2          -0.0128      0.076     -0.167      0.868      -0.164       0.139
users_diff     0.2661      0.032      8.323      0.000       0.203       0.330
==============================================================================
Omnibus:                        4.863   Durbin-Watson:                   2.190
Prob(Omnibus):                  0.088   Jarque-Bera (JB):                4.765
Skew:                          -0.318   Prob(JB):                       0.0923
Kurtosis:                       3.853   Cond. No.                         58.4
==============================================================================
```

Veja que o modelo apresenta baixos RMSE e MAE (bem menores que nas modelagens anteriores), além de altos valores de R^2.

Além disso, os resíduos do modelo se comportaram muito bem, com valores bem próximos de 0 e distribuição aproximadamente normal padrão:



E, a partir do plot de comparação entre o valor real e o valor predito, vemos o bom comportamento do modelo:



8 Discussão

Os resultados obtidos ao longo da análise evidenciam a importância das transformações aplicadas e da seleção adequada de modelos para capturar a dinâmica da série de volume. A transformação Box-Cox mostrou-se essencial para estabilizar a variância, resultando em melhorias significativas tanto nos modelos baseline quanto nos modelos de suavização. Entre esses, Drift, Holt e Holt-Winters Aditivo apresentaram os resíduos mais estáveis.

A inclusão de componentes sazonais e covariáveis também contribuiu para o ganho de desempenho: o SARIMA superou o ARIMA inicial, e o SARIMAX apresentou resultados superiores ao incorporar as variáveis exógenas *inv* e *users*. O grid search aplicado ao ARIMA resultou em uma configuração capaz de superar todos os modelos baseline e de suavização, mostrando que a modelagem autorregressiva funciona bem na série transformada.

Ainda assim, o melhor desempenho geral foi obtido pela Regressão Linear Múltipla, especialmente quando enriquecida com variáveis derivadas (calendário, defasagens e diferenças). A validação leave-one-out temporal indicou que o modelo é robusto e apresenta erros substancialmente menores que os modelos anteriores, destacando o papel central das defasagens curtas e da dinâmica das variáveis exógenas.

9 Conclusão

Este estudo mostrou que a combinação de transformação, decomposição e modelagem avançada permite construir previsões robustas para a série de volume. A transformação Box-Cox foi fundamental para estabilizar a variância e melhorar o desempenho dos modelos baseline e de suavização, resultando em resíduos mais estáveis e erros substancialmente menores. A decomposição STL revelou tendência crescente e sazonalidade bem definida, contribuindo para uma compreensão mais clara da estrutura temporal da série.

Os modelos ARIMA, SARIMA e SARIMAX apresentaram ganhos progressivos à medida que incorporavam dependência temporal, estrutura sazonal e variáveis exógenas. O *grid search* aplicado ao ARIMA evidenciou que a seleção adequada de hiperparâmetros permite superar os modelos baseline e de suavização. No entanto, o melhor desempenho geral foi obtido pela Regressão Linear Múltipla com variáveis transformadas e derivadas, avaliada via validação *leave-one-out* temporal. O modelo alcançou $R^2 = 0.979$, apresentando baixo erro preditivo e resíduos bem comportados, destacando o papel central das defasagens de curto prazo e da influência das covariáveis exógenas.