Séries Temporais dos Acessos à Página do Flamengo na Wikipédia em Português

JJ

Sumário

1 Descrição da base de dados

Os dados analisados neste trabalho foram extraídos da base WikiMedia, que registra o número de acessos diários e mensais a páginas da Wikipédia. Para este estudo, foi selecionada a página do **Clube de Regatas do Flamengo**, considerando exclusivamente acessos humanos, não automatizados e provenientes de todas as plataformas (desktop, mobile, etc.).

O período analisado abrange de julho de 2015 (2015-07) a abril de 2025 (2025-04).

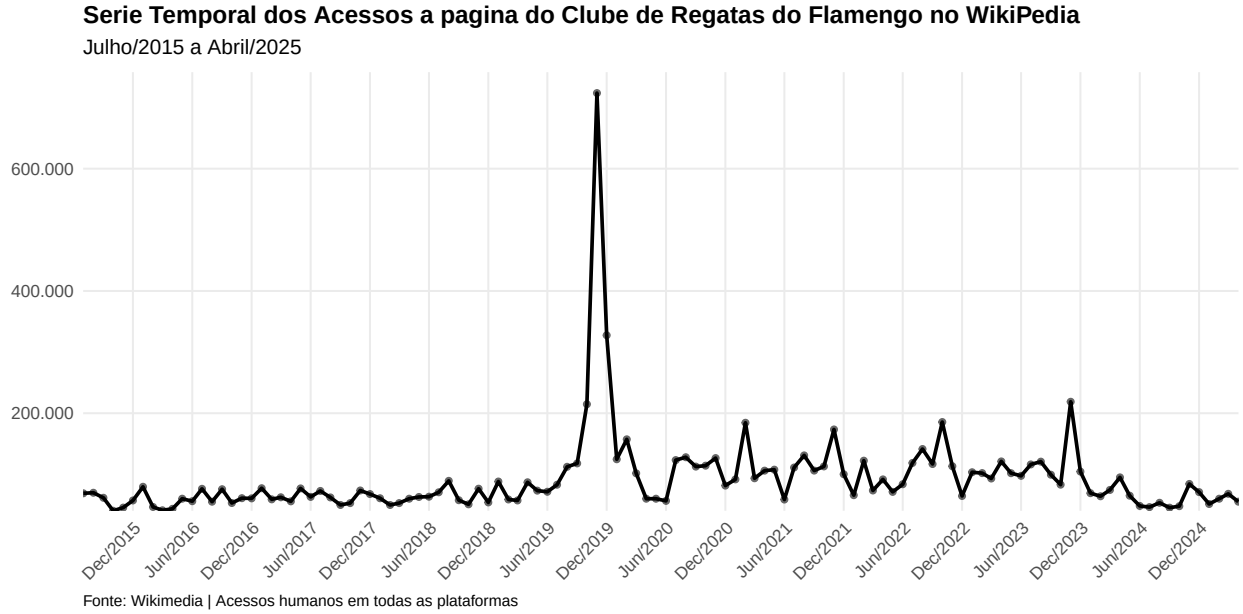
Table 1: Amostra dos dados de acessos à página do Flamengo na Wikipédia

Data	Acessos
2015-07-01	69 072
2015-08-01	69 949
2015-09-01	61 722
2015-10-01	40 512
2015-11-01	45 647
2015-12-01	57 336

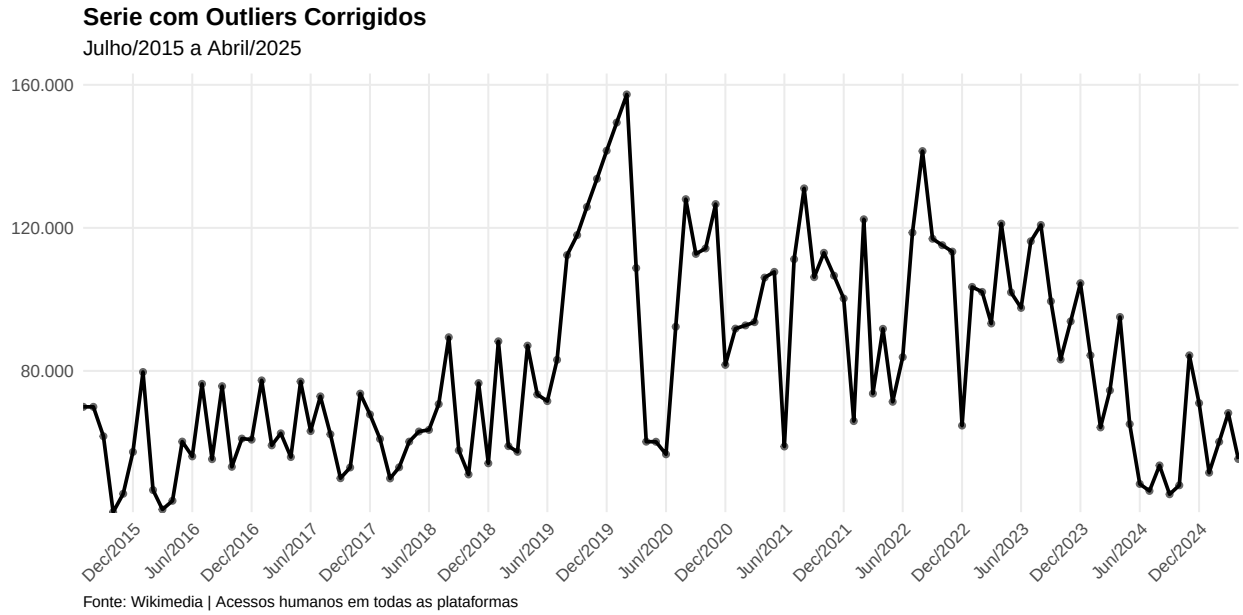
Fonte: WikiMedia; dados filtrados para acessos humanos de todas as plataformas.

2 Pré-processamento

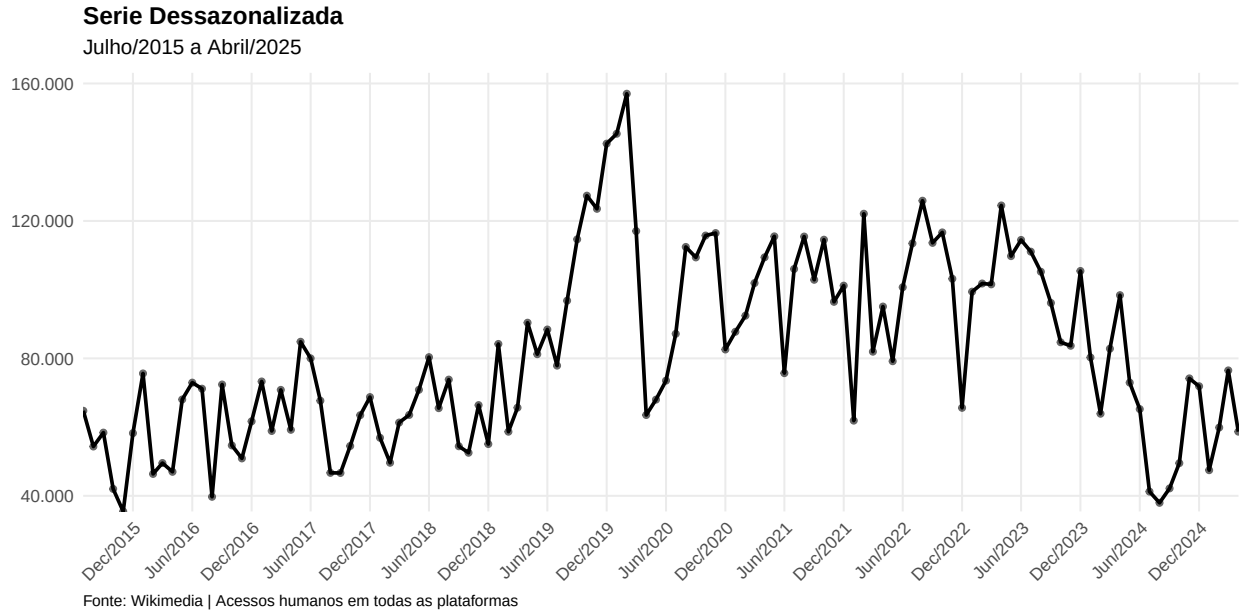
A análise visual da série indica que há sazonalidade e que há suspeita de não estacionariedade. Fora isso, também podemos observar a presença de um grande outlier.



Primeiramente, iremos nos preocupar com os outliers e, para isso, será usada a função *tso* para identificá-los e iremos retirar usando a função *na.approx* que faz, automaticamente, interpolação linear nas entradas escolhidas (se houver outliers nas extremidades, utilizaremos o valor mais próximo).



Agora, já com os outliers retirados, iremos retirar a sazonalidade usando a função *STL* que ajusta modelos de regressões locais para cada janela de um ano, extraindo o padrão médio de repetição. Assim, identificando o componente sazonal, será usada a função *seasadj* para obter a série ajustada sem a componente.



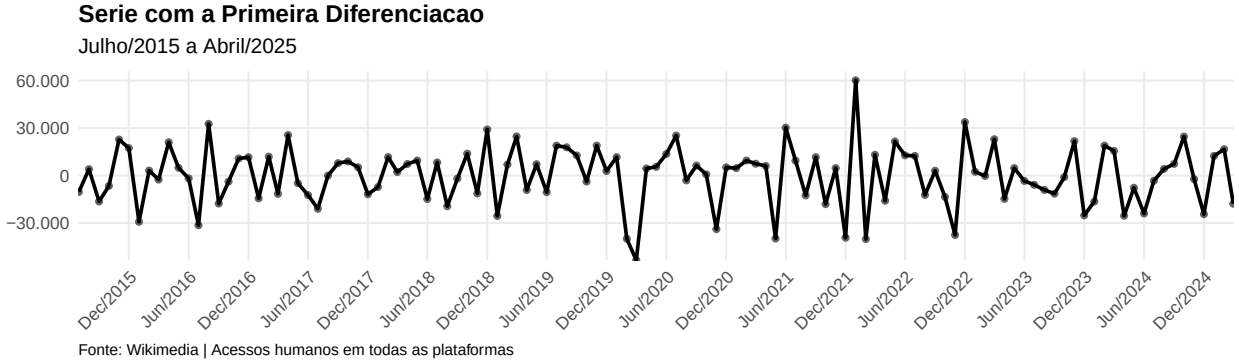
Agora, já tendo a série dessazonalizada, iremos conferir a outra suspeita inicial, que a série poderia não ser estacionária.

Para isso, faremos o teste ADF, que confere se a série é estacionária ou não. Nele, a hipótese alternativa (H1) é: "a série é estacionária".

Table 2: Resultado do Teste ADF

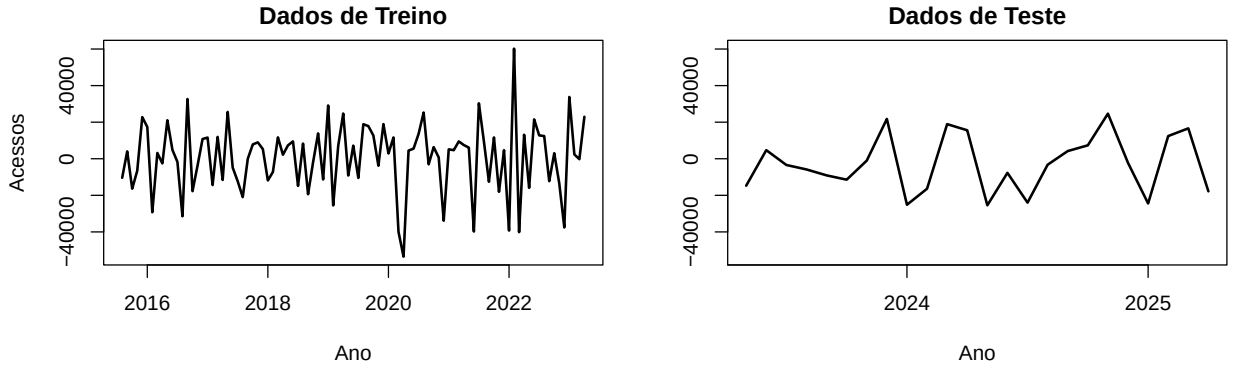
Estatistica	p_valor	Hipotese_Alternativa
-2.2015	0.4929	stationary

O valor do teste foi de -2.2015, com um p-valor de 0.4929. Esse resultado indica que a hipótese nula de não estacionariedade não pode ser rejeitada ao nível de significância de 5%. Ou seja, não há evidências para concluir que a série é estacionária. Assim, a série pode apresentar tendência ou dependência temporal de longo prazo, sendo necessária, então, a diferenciação do modelo.



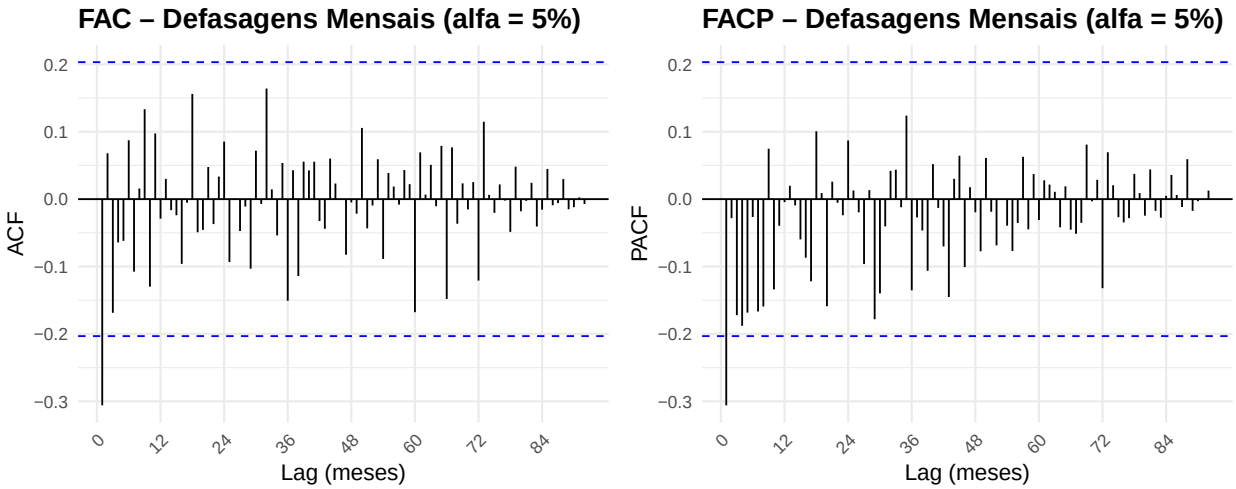
Agora a série está no formato ideal e com as características necessárias para nossas análises, identificações e estimações.

Para avaliar se os modelos conseguem prever bem os dados futuros, a série temporal foi dividida em duas partes: treino e teste. A parte de treino vai de julho de 2015 até abril de 2023, e foi usada para ajustar os modelos. Já o período de teste, de maio de 2023 a abril de 2025, foi separado para verificar se as previsões realmente funcionam em dados que o modelo ainda não viu.



3 Identificação do Modelo com Base na FAC e FACP

Para a identificação da estrutura do modelo, será considerada apenas a porção de treino da série, composta pelos dados de julho de 2015 a abril de 2023. A análise se baseia nas funções de autocorrelação (FAC) e autocorrelação parcial (FACP), aplicadas sobre a série já diferenciada. A partir da interpretação desses gráficos, será proposto um modelo inicial, que servirá como base para o resto do trabalho.



Em ambos os gráficos se pode observar truncamento no *lag 1*, o que fornece para nós a sugestão de um modelo inicial **ARIMA(1,1,1)**, ou **ARMA(1,1)** considerando que a gente já fez a diferenciação nos dados. Portanto, iremos começar a nossa análise com a ideia de que, pela *FAC* e *FACP*, temos um modelo **ARMA(1,1)**, com a seguinte equação:

$$X_t = \phi_1 X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}, \quad \varepsilon_t \sim \mathrm{N}(0, \sigma^2) \tag{1}$$

4 Sobrefixação e estimação do modelo

Na última seção, vimos que pela FAC e FACP escolhemos um modelo **ARMA(1,1)**. Nesta parte, iremos avaliar o modelo escolhido, fazer sobrefixação para tentar encontrar o melhor modelo e depois estimar os parâmetro.

Para isso, consideraremos os seguintes modelos candidatos: **ARMA(1,1)**, **ARMA(2,1)**, **ARMA(1,2)** e **ARMA(2,2)**. Em seguida, faremos a comparação entre eles, a fim de escolher o modelo final mais adequado para a série em questão.

Table 3: Coeficientes e testes-z dos modelos ARMA

(a) ARMA(1,1)

Termo	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.5457	0.0877	-6.2192	0.0000***
ma1	-1.0000	0.0277	-36.0502	0.0000***

(b) ARMA(2,1)

Termo	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.7101	0.1012	-7.0181	0.0000***
ar2	-0.2908	0.1006	-2.8897	0.0038**
ma1	-1.0000	0.0281	-35.6408	0.0000***

(c) ARMA(1,2)

Termo	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.1322	0.1056	-1.2512	0.2109
ma1	-1.9967	0.0472	-42.3056	0.0000***
ma2	0.9996	0.0471	21.2189	0.0000***

(d) ARMA(2,2)

Termo	Estimate	Std. Error	z value	Pr(>|z|)
ar1	-0.1376	0.1070	-1.2856	0.1986
ar2	-0.0294	0.1065	-0.2762	0.7824
ma1	-1.9970	0.0478	-41.7859	0.0000***
ma2	0.9999	0.0477	20.9559	0.0000***

Table 4: Comparacao de AIC e BIC entre os modelos ajustados

Modelo	AIC	BIC
ARMA(1,2)	2060.995	2071.038
ARMA(2,2)	2062.918	2075.472
ARMA(2,1)	2088.374	2098.418
ARMA(1,1)	2094.279	2101.811

Como podemos ver, em primeiro lugar, pelo teste-z, tanto **ARMA(1,1)** quanto **ARMA(2,1)** apresentam todos os seus parâmetros altamente significativos (p < 0,01). Já no **ARMA(1,2)** o termo *AR(1)* não é significativo ($p \approx 0{,}21$), e no **ARMA(2,2)** os dois coeficientes *AR* também ficam fora de significância, o que indica parâmetros RUINS nesses dois modelos.

Em segundo lugar, os critérios de informação apontam menor AIC/BIC para o **ARMA(1,2)**, seguido de **ARMA(2,2)**, mas essa vantagem de ajuste contrasta com a falta de significância do coeficiente AR. O **ARMA(2,1)**, apesar de ter AIC/BIC maiores que os do modelo **MA(2)**, tem todos os coeficientes significativos e apenas dois termos AR e um MA.

Por fim, optamos pelo **ARMA(2,1)**: ele contém apenas parâmetros que contribuem de fato para o ajuste, mantém resíduos próximos ao ruído branco e apresenta informações úteis.

Modelo Escolhido:

$$X_t = -0{,}7101\, X_{t-1} \;-\; 0{,}2908\, X_{t-2} \;+\; \varepsilon_t \;-\; 1{,}0000\, \varepsilon_{t-1}, \qquad \varepsilon_t \sim \mathrm{N}\big(0,\, 478\,526\,867\big)$$