17 e 18 de Novembro
Breves - Pa
CUMB-UFPA 2022
ISSN:1982-8691
VIII EPAMM
VIII Encontro Paraense de Modelagem Matemática
GEMM
Modelagem Matemática e Práticas Socioculturais no Contexto Amazônico

Comparação dos modelos ARIMA e LightGBM na previsão de Manchas Solares

Kalebe Carneiro Monteiro Araújo
UFPA - FAMAT
kalebe15carneiro@gmail.com

Vanilson Gomes Pereira
UFPA – FAMAT
vgomes@ufpa.br

Resumo:
O número de manchas solares são registros observados, no decorrer do tempo, da atividade do sol, possuem campos magnéticos que provocam erupções solares que causam, por exemplo, distúrbios em satélites e sistemas de telecomunicações. Neste estudo, comparamos e avaliamos três técnicas, LightGBM (Light Gradient Boosting Machine), ARIMA (Autoregressive Integrated Moving Average) e Sazonal ARIMA (SARIMA), na tarefa de previsão de um passo à frente da série temporal anual de manchas solares. Os modelos foram estabelecidos usando dados de treinamento, e os conjuntos de testes foram usados para avaliar a capacidade de previsão de cada modelo e, finalmente, avaliar a acurácia por meio das medidas estatística MAE, RMSE e MAPE. Os experimentos apresentam um melhor desempenho para o modelo SARIMA com melhor efeito de previsão em comparação com outros modelos contrastados.

Palavras-chave: Manchas solares. ARIMA. LightGBM. Modelo de Predição.

Introdução

Séries temporais são observações sobre determinado evento com suas variáveis relacionadas com o tempo. Para a previsão de uma determinada série, os dados são coletados e analisa-se suas observações anteriores para descrever a relação explicativa uma sobre as outras, que então é extrapolada para o futuro. A razão pela qual a previsão de séries temporais é tão importante é que a previsão de futuros eventos é uma entrada crítica em muitos tipos de planejamento e tomada de decisão para processos, com aplicação em áreas como economia, finanças e gestão de riscos, demografia, como também manchas solares, para prever ciclos de máxima e mínima solar, que podem causar tempestades geomagnéticas capazes de interromper satélites de comunicação em órbita e geradores de energia em solo, mas também podem causar o fenômeno conhecido como auroras boreais. Para modelar tais séries temporais, existem vários modelos de análise, nas quais utilizaremos o modelo ARIMA, SARIMA e LigthGBM.

Manchas solares são regiões mais escuras (vista por contraste), por serem mais frias do que a superfície do sol ao seu redor devido seu campo magnético ser mais intensos que inibe a convecção de plasma das camadas sub-superficiais, são contadas pelo aparecimento na parte observável do sol e sua atividade magnética está relacionada com a incidência de UVB, quantidade de Ozônio formada na atmosfera terrestre e interferência na comunicação por ondas curtas. Segundo Ezequiel Echer et al. (2003), os registros de manchas solares mostram a variabilidade cíclica regular média de 11 anos descoberto por Schwabe em 1843, datadas por Wolf que reconstruiu os dados desde 1700 à 1748 para observações anuais, de 1749 à 1817 para registros mensais e diários com alguns intervalos sem dados (de até um mês) e valores completos disponíveis desde 1818. Esses registros formam a série temporal de Manchas Solares.

Um dos modelos mais conhecido para análise e previsão de séries temporais é o modelo ARIMA, uma extensão dos modelos Auto regressivo e de Média Móvel de Box e Jenkins (2015), sendo capaz de modelar uma série não estacionária em estacionária calculando as diferenças entre observações consecutivas, procedimento conhecido como diferenciação e pode ser feito várias vezes até que se torne estacionária. Segundo Hasan, Md Rashidul, et al. (2022), ao incorporar a ordem autorregressiva (AR), diferenciação (I) e ordem de Médias Móveis (MA), o ARIMA torna-se mais flexível e robusto para prever dados de séries temporais. Além disso, ao adicionar o componente sazonal ao modelo ARIMA, temos um modelo SARIMA, uma extensão do modelo base que suporta a modelagem com componentes sazonais. Segundo Silva, Rafael (2019), O modelo ARIMA possui alto grau de complexibilidade na escolha dos parâmetros apropriados, sendo um dos desafios do modelo, devido ser necessário uma análise sobre os gráficos temporal dos dados, assim como os gráficos ACF e PACF.

Outra abordagem de previsão é o LightGBM, um modelo baseado em machine learning que utiliza lógica de aprendizado por Árvore de Decisão baseado em aumento de gradiente que fornece uma nova implementação do GBDT (Gradient Boosting Decision Tree), baseado em duas técnicas: GOSS e EFB segundo Ke, Guolin, et al. (2017). Nos seus experimentos mostra que o LightGBM acelera os processos de treinamento do GBDT convencional em até mais de 20 vezes com precisão similar.

O objetivo proposto deste estudo é fazer uma análise comparativa do desempenho de previsão dos modelos ARIMA, SARIMA e LightGBM, para previsão da série temporal anual de Manchas Solares. Por fim, avaliaremos a acurácia e a capacidade de previsões futuras dos modelos através das métricas MAE, RMSE e MAPE.

Modelo ARIMA

De acordo com Morettin e Toloi (2006) apud Silva, Rafael (2019), no procedimento ARIMA, primeiramente, modela-se a parte Autorregressiva (p) da série, onde a variável de interesse é descrita pelas observações passadas, ou seja, na defasagem de seus p valores. Descrito como:

$$y_t = c + \phi_1 \cdot y_{t-1} + \phi_2 \cdot y_{t-2} + \cdots + \phi_p \cdot y_{t-p} + \varepsilon_t \quad (1)$$

Onde c é o intercepto e εt uma sequência de variáveis aleatórias independentes igualmente distribuídas (iid) denominada resíduo.

O modelo de médias móveis é semelhante ao AR, porém o foco do processo está na defasagem dos erros, sendo seu processo MA (q):

$$y_t = \mu + \varepsilon_t + \theta_1 \cdot \varepsilon_{t-1} + \theta_2 \cdot \varepsilon_{t-2} + \cdots + \theta_q \cdot \varepsilon_{t-q} \quad (2)$$

Sendo μ a média do processo gerador da série e yt pode ser interpretado como uma média móvel ponderada do erro presente e dos seus últimos q valores.

Combinando os modelos autorregressivos (p) e de médias móveis (q), temos um modelo ARMA(p,q) que capta séries estacionárias, ou seja, séries que não possuem tendência ou sazonalidade. Para séries não estacionárias é preciso diferenciar a série usando o termo de integração (d), que suas características estatísticas, como a média e a estrutura de autocorrelação, sejam constantes ao longo do tempo. Usa-se a diferença entre seus valores sucessivos. Se mesmo após a diferenciação da equação abaixo (3), a série não tiver média constante, aplica-se a segunda diferença, ou seja Δ(Δyt). O número de vezes que uma série é diferenciada, até se estabilizar, é o valor de d do processo ARIMA(p,d,q) não sazonal, cuja equação é dada por:

$$\Delta^d y_t = c + \phi_1 \cdot \Delta^d y_{t-1} + \cdots + \phi_p \cdot \Delta^d y_{t-p} + \varepsilon_t + +\theta_1 \cdot \varepsilon_{t-1} + \cdots + \theta_q \cdot \varepsilon_{t-q} \quad (3)$$

Onde, y é a variável de interesse, c é o intercepto, Δd é a quantidade de diferenciações e ε é uma iid e os termos θ e ϕ são OS parâmetros. O intercepto c é dado pela equação:

$$C = \left(1 - \phi_1 - \cdots - \phi_p\right) \cdot \mu \quad (4)$$

Zhang, G. Peter. (2003) explica que para a construção de um modelo ARIMA é necessário seguir três etapas: identificação do modelo, estimativa de parâmetros e verificação de diagnóstico. Na etapa de identificação, a transformação de dados geralmente é necessária para tornar a série temporal estacionária, condição necessária na construção de um modelo ARIMA. Box e Jenkins (2015) propuseram utilizar a função de autocorrelação (ACF) e a função de autocorrelação parcial (PACF) dos dados amostrais como ferramentas básicas para identificar a ordem do modelo ARIMA. Uma vez que um modelo provisório é especificado, a estimativa dos parâmetros do modelo é direta. Os parâmetros são estimados de modo que uma medida geral de erros seja minimizada. A última etapa é a verificação diagnóstica da adequação do modelo, onde as informações de diagnóstico podem ajudar a sugerir modelos alternativos. Se o modelo não for adequado, um novo modelo provisório deve ser identificado, seguidos pelas mesmas etapas (de estimação de parâmetros e verificação do modelo). O melhor modelo é definido pelo menor critério de informação Akaike (AIC).

Modelo Sazonal ARIMA

O modelo Sazonal ARIMA (SARIMA) é uma extensão do modelo base. Segundo Piauhy Neto, Franklin (2021), é formado adicionando-se à equação (3) termos autorregressivos, de médias móveis e diferenciações sazonais, sendo representados por: ARIMA(p,d,q)(P,D,Q)m. Onde m é o período sazonal, D são as diferenças sazonais, para estabilizar a média, P são os termos autorregressivos sazonais, e Q são termos de médias móveis sazonais. De acordo com Hyndman e Athanasopoulos (2021) apud Piauhy Neto, Franklin (2021), para ajustar modelos SARIMA é preciso seguir as etaspas: transformar os dados para estabilizar a variância; diferenciar (d e D) para estacionarizar; determinar p, q, P e Q; ajustar o melhor modelo e testar se os resíduos são iid; e implementar o modelo para executar as previsões.

Modelo Light Gradient Boosting Model - LightGBM

LightGBM é um algorítimo de Machine Learner (Aprendizado de Máquina) que utiliza lógica de aprendizado por Árvore de Decisão baseado em aumento de gradiente. Segundo

Hasan, Md Rashidul, et al (2022), possui rapidez de processamento, além de ser eficaz em manuseio para grande conjunto de base de dados. Entretanto, não é recomendado para uso em bases de dados pequenas por ser sensível a overfitting. Em sua operação, o algoritmo utiliza duas técnicas principais: Construção de recursos exclusivos (Exclusive Feature Bundline - EBF) e amostragem de um lado baseada em gradiente (Gradient-basedd Onde-Side Sampling - GOSS). Segundo Ke, Guolin, et al. (2017) e Hasan, Md Rashidul, et al (2022), O EFB é eficaz para alavancar propriedades esparsas no histograma e traz uma aceleração para o processamento para grande escala de dados pelo fato de mesclar recursos esparsos em muito menos recurso no processo de agrupamento. Por outro lado, o GOSS permite reduzir o tamanho da amostra usando grandes amostras de gradiente e uma fração de amostras de gradiente mais baixo que multiplicadas por uma constante para dar mais peso ao conjunto de dados sub-treinado.

Procedimento metodológico

Para verificar o desempenho dos modelos ARIMA não sazonal, componente sazonal (SARIMA) e LightGBM, consideramos os dados anuais da série temporal Sunspots (SN). Inicialmente, importamos a série SN V2.0 obtida do site da SILSO (www.sidc.be/silso/datafiles). Separamos o conjuntos de dados em treinamento (1700 a 1957) e teste (1958 a 2021) para, respectivamente, estimar os modelos e avaliar a capacidade de previsão de cada modelo. A tabela 1 apresenta a descrição dos dados e na Tabela 2 mostra a estatística descritiva dos dados. Na tabela 2 mostra que série temporal mancha solar possui assimetria a direita (assimetria maior que 0) com curtose maior que 3, caracterizando uma série temporal não gausiana (curtose não é igual a 3), altamente não linear e dificil de predizer. A série temporal SN anual completa é mostrada na Figura 1.

Tabela 1 – observações da SN Anual

			Número de Observações:		
Série temporal	Descrição	Duração	Treino	Teste	Total
SN Anual	Média Anual SN	1700 – 2021	258	64	322

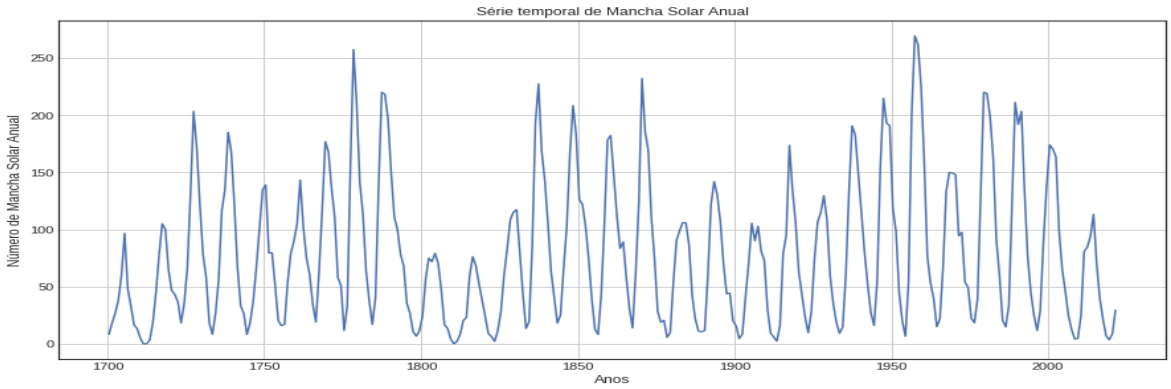
Fonte: Elaboração própria

Tabela 2 – descrição estatística da SN Anual

Série temporal	Mínimo	Máximo	Média	Desvio padrão	Assimetria	Curtose
Média Anual SN	0	269.300	78.365	62.055	0.58	4.98

Fonte: Elaboração própria

Figura 1 – Série temporal SN anual de 1700 à 2021

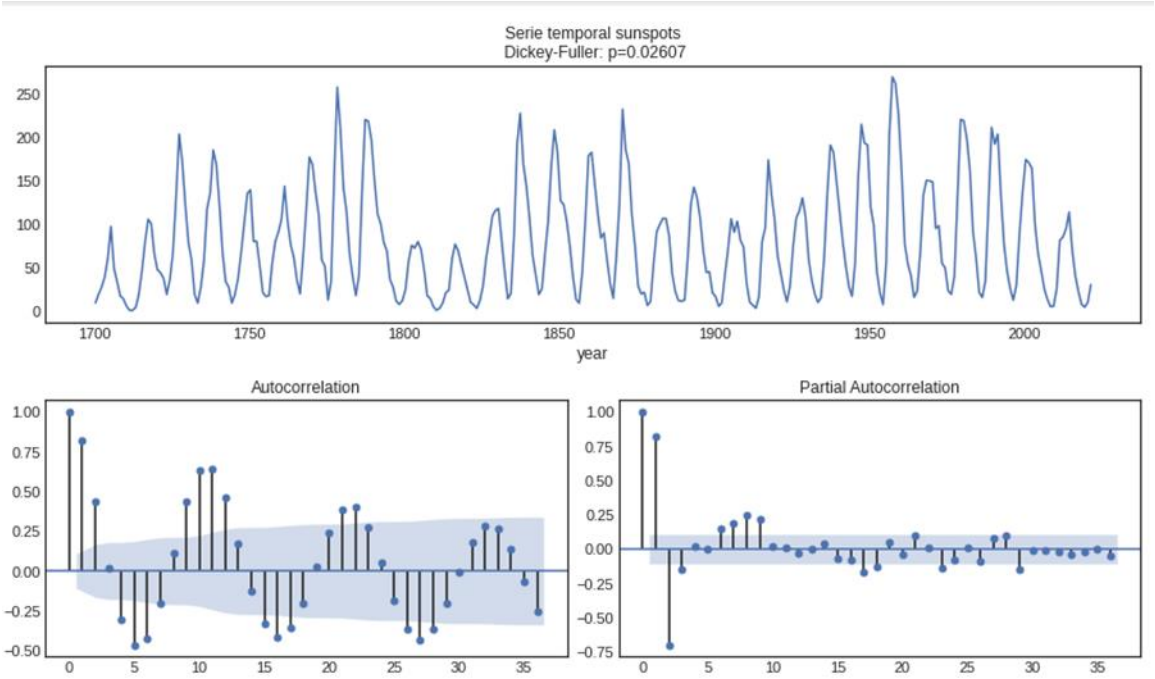


Fonte: SILSO (2022)

Procedimento Estatístico

Para determinar os padrões que melhor descrevem a série temporal de manchas solares, seguimos a abordagem Box-Jenkins para seleção dos modelos ARIMA, consistindo em 3 etapas, segundo Box e Jenkins (2015). Primeiro, a média anual de manchas solares foi plotada em relação ao tempo para detectar e corrigir a não estacionaridade da série temporal (Fig. 2) e identificar termos de média móvel e autorregressiva necessários para calcular as funções de autocorrelação (ACF) e autocorrelação parcial (PACF).

Figura 2 – Gráfico da série manchas solares e ACF – PACF



Fonte: Elaboração própria

Em seguida, modelos de ordens variadas foram ajustados e comparados através do critério de informação de Akaike (AIC) para avaliar melhorias no ajuste enquanto penalizava a complexidade do modelo. Por último, foi confirmado que a autocorrelação temporal não estava mais presente nos resíduos do modelo usando o teste de Ljung-Box. Para criar o modelo de predição foi separado 80% para treinamento e 20% para de teste.

Medidas de previsão

Para verificar a precisão das previsões de um modelo é através do ajuste dos dados, ou seja, se as diferenças entre os valores observados e os valores previstos do modelo forem pequenos e não viesados (inclinados). Para avaliar a acurácia das previsões dos modelos, utilizamos as métricas baseadas em Escalas-Dependentes, que calcula a diferença entre a série observada yi e a previsão obtida ŷi, e Erros Percentuais que calcula a diferença dos erros percentuais para cada unidade tempo. As métricas utilizadas, também são encontradas no trabalho de Prajapati, Samyak, et al. (2021), sendo elas: MAE (sigla em inglês para Mean Absolut Error) (5), que calcula o erro médio absoluto, na qual, utiliza o módulo de cada erro, evitando que os valores negativos de determinadas observações zere valores positivos, dando uma falsa impressão de que o modelo está bom; o RMSE (Root Mean Squared Error) (6), calcula a raiz quadrada do erro quadrado médio, onde cada erro é elevado ao quadrado e sua média é calculada e devido a sua raiz, o valor retorna à métrica normal; e o MAPE (Mean Absolute Percentage Error) (7), que calcula a precisão pela média do erro percentual, em forma de porcentagem. As métricas (5), (6) e (7), estão descritas nas seguintes equações:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \tag{6}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\% \tag{7}$$

Onde n é o número de observações da série temporal, y_i é i-ésimo valor da série temporal e $\hat{y}_i$ é o i-ésimo valor de previsão.

Resultados das simulações e discussões

As análises dos resultados dos modelos ARIMA, ambos não sazonal e componentes sazonais, foram selecionados com base nas funções de autocorrelação (ACF) e autocorrelação parcial (PACF) e do critério de informação de Akaike (AIC). A seleção do modelo com o menor valor de AIC foi considerado o melhor, avaliando os benefícios e as desvantagens. Um modelo de ordem $(3,0,3)$, (Modelo 1, $AIC = 2395.414$) foi selecionado e ajustado (acrescentando a componente sazonal MA com a mesma ordem) resultando em um modelo ARIMA Sazonal de ordem $(3,0,3) \times (0,0,3)_{12}$, mais parcimonioso, (Modelo 2, $AIC = 2391.013$). Na tabela 3 apresenta todos os coeficientes estatisticamente significativo (p<0.001). Todos esses parâmetros são modelados como padrão no algoritmo do método ARIMA da biblioteca sktime de Löning, Markus, et al. (2019).

Tabela 3 – Coeficientes do modelo ARIMA Sazonal

Parâmetros	Coeficientes	SE Coeficientes	Estatística t	p
AR(3)	0.8584	0.051	16.840	<0.001
MA(3)	0.1803	0.073	2.470	<0.001
MA(3) Sazonal	0.0076	0.084	0.090	<0.001

Fonte: Elaboração própria

O modelo LightGBM possui hiperparâmetros que são ajustados pelo algoritmo Grid Search (GS) realizando uma combinação de parâmetros. Alguns parâmetros mais importantes que o LightGBM utiliza é a taxa de aprendizado que foi de 0.1, com 31 arvore, profundidade da árvore 6 e o método otimizador foi GBDT. Todos esses parâmetros são configurados como padrão no algoritmo do método LightGBM, Ke, Guolin, et al. (2017).

Os resultados com as métricas estatísticas de cada modelo no conjunto de teste são apresentados na tabela 4. Pode se observar que o modelo ARIMA Sazonal (SARIMA) é superior aos demais individualmente, como pode ser observado nas medidas. O desempenho do ARIMA foi relativamente inferior, que provavelmente pode ser atribuído por não ter incluído a componente de sazonalidade. Já o modelo LightGBM obteve uma melhoria apenas na métrica MAPE com ganho de 1,3 % em relação ao modelo SARIMA. Porém as demais métricas foram inferiores. Dentre os modelos com melhor desempenho e adequação ao problema de previsão de manchas solares, podemos dizer que o SARIMA foi o que apresentou melhor precisão.

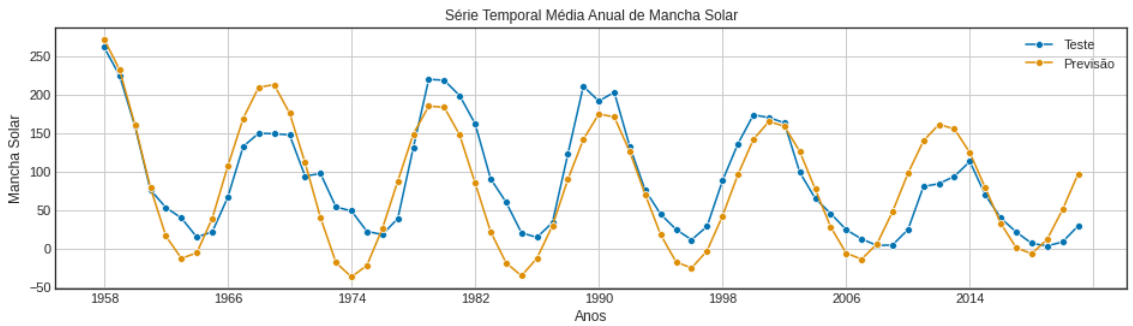
Tabela 4 – Desempenho dos modelos ARIMA e LGBM sobre o conjunto de teste

Métricas	SARIMA	ARIMA	LightBGM
MAE	**34.367**	37.546	41.200
RMSE	**41.544**	45.681	53.453
MAPE	0.965	1.218	**0.834**

Fonte: Elaboração própria

Na Fig. 3 mostra uma comparação entre a previsão e o conjunto de teste da série de média anual de manchas solares de 1958 a 2021. Observa-se que o perfil de previsão de manchas solares do modelo mais adequado (SARIMA) é capaz de corresponder aos dados reais de manchas solares. Especialmente, a flutuação pode ser capturada pelo modelo selecionado. No entanto, apresenta desvios em resposta as variações.

Figura 3 – Previsão sobre o conjunto de teste para o modelo SARIMA



Fonte: Elaboração própria

Todos os modelos experimentais são executados no ambiente de programação Python 3.6. O hardware é um Notebook com CPU Intel Core i7 e 8GB de memória.

Conclusões

Este trabalho propõe avaliar o desempenho dos modelos ARIMA, SARIMA e LightGBM, em um problema de previsão de séries temporais univariadas de manchas solares na síntese de um preditor um passo à frente. O resultado mostrou que o modelo SARIMA comparado com os demais obteve como desempenho na previsão de manchas solares. As métricas estatísticas foram avaliadas e apresentam melhor robustez e um melhor efeito de previsão em comparação com outros modelos contrastados.

Em trabalhos futuros, será considerado a influência de diferentes modelos de ajuste de parâmetros no desempenho de previsão de manchas solares. Além disso, consideraremos ainda

a possibilidade de propor um modelo híbrido e verificar a eficácia de previsão do método proposto em previsão na síntese de um preditor de longo prazo.

Referências

ECHER, Ezequiel, et al. **O número de manchas solares, índice da atividade do sol**. Revista Brasileira de Ensino de Física 25 (2003): 157-163.

BOX, George EP, et al. **Time series analysis: forecasting and control**. John Wiley & Sons, 2015.

HASAN, Md Rashidul, et al. **A Comparative Study on Forecasting of Retail Sales**. arXiv preprint arXiv:2203.06848 (2022).

SILVA, Rafael Guilherme Fernandes de Lima. **Avaliação da Precisão dos Modelos ARIMA com e sem Transformação Estabilizadora da Variância na Previsão de Séries Temporais Anuais**. (2019).

KE, Guolin, et al. **Lightgbm:** A highly efficient gradient boosting decision tree. Advances in neural information processing systems 30 (2017).

ZHANG, G. Peter. **Time series forecasting using a hybrid ARIMA and neural network model**. Neurocomputing 50 (2003): 159-175.

PIAUHY NETO, Franklin. **Métodos de seleção automática de omdelos ARIMA no software R: uma comparação dos algoritmos do pacote forecast**. (2021).

PRAJAPATI, Samyak, et al. **Comparison of traditional and hybrid time series models for forecasting COVID-19 cases**. arXiv preprint arXiv:2105.03266 (2021).

LÖNING, Markus, et al. **sktime:** A unified interface for machine learning with time series. arXiv preprint arXiv:1909.07872 (2019).

SILSO - Sunspot Index and Long-term Solar Observations. **Royal Observatory of Belgium, Brussels.** Disponível em: https://www.sidc.be/silso/home. Acesso em: 05 de setembro, 2022.