

ESTUDO DE TÉCNICAS DE PREVISÃO FUTURA DE SÉRIES TEMPORAIS DE PREÇO DE PETRÓLEO WTI UTILIZANDO DECOMPOSIÇÃO DE MODO EMPÍRICO, PROPHET, SARIMAX E REGRESSÃO LASSO

Giovani de Oliveira

USP

<https://lattes.cnpq.br/1673804055818101>

<https://orcid.org/0009-0002-2425-0955>

E-mail: giovani1996.oliveira@gmail.com

Hygor Santiago

UNICAMP - Universidade Estadual de Campinas

<http://lattes.cnpq.br/8858059011756333>

<https://orcid.org/0000-0002-4835-5498>

E-mail: hsantiagolara@gmail.com

DOI-Geral: <http://dx.doi.org/10.47538/RA-2023.V2N3>

DOI-Individual: <http://dx.doi.org/10.47538/RA-2023.V2N3-75>

RESUMO: Dados de séries temporais estão cada vez mais presentes e disponíveis para estudos e aplicações práticas à medida que a coleta, armazenamento e compartilhamento destes dados se tornam facilitados com o avanço de tecnologias como ecossistemas de big data e internet favorecem uma dispersão destas informações de maneira nunca antes vista. A proposta deste estudo é a criação de um modelo de previsão de séries temporais automático, utilizando puramente bibliotecas Python para o estudo. O método estudado retornou bons resultados, na modelagem principal (2018- 2023) vemos que o CEEMDAN retorna ótimos valores de correlação e de MAPE, EEMD retornou o melhor valor de correlação, mas foi o método EMD que chegou ao valor mais próximo do real depois de 5 anos.

PALAVRAS-CHAVE: Séries Temporais. Preço de Petróleo. Decomposição de Modo Empírico.

STUDY OF FUTURE FORECASTING TECHNIQUES OF WTI OIL PRICE TIME SERIES USING EMPIRICAL MODE DECOMPOSITION, PROPHET, SARIMAX AND LASSO REGRESSION

ABSTRACT: Time series data are increasingly present and available for studies and practical applications as the collection, storage and sharing of this data becomes easier with the advancement of technologies such as big data ecosystems and the internet, favoring a dispersion of this information. way never seen before. The purpose of this study is to create an automatic time series forecasting model, using purely Python libraries for the study. The studied method returned good results, in the main modeling (2018-2023) we see that CEEMDAN returns excellent correlation and MAPE values, EEMD returned the best correlation value, but it was the EMD method that arrived at the closest value to the real after 5 years.

KEYWORDS: Time Series. Oil Price. Empirical Mode Decomposition.

INTRODUÇÃO

Dados de séries temporais estão cada vez mais presentes e disponíveis para estudos e aplicações práticas à medida que a coleta, armazenamento e compartilhamento destes dados se tornam facilitados com o avanço de tecnologias como ecossistemas de big data e internet favorecem uma dispersão destas informações de maneira nunca antes vista.

Sensores e mecanismos estão por toda parte, resultando em quantidades sem precedentes de dados de séries temporais de alta qualidade disponíveis (Nielsen, 2019). Neste cenário, uma das aplicações que pode ser retirada dos dados é a previsão de dados futuros utilizando modelos de previsão.

Encontramos dados e análises de séries temporais em diversas áreas como: medicina (Alsallakh et al., 2021), previsão do tempo (Karevan, Suykens, 2020), economia (Sezer, Gudelek, Ozbayoglu, 2020), astronomia (Jamal, Bloom, 2020) e commodities (Zhang, Chen, Liwen, Xia, 2020).

Ao longo da história moderna, o petróleo desempenhou um papel proeminente na formação do desenvolvimento da economia mundial (Phan, Tran, Nguyen, Le, 2020). É fato que o petróleo é um fator de extrema importância na economia mundial e técnicas envolvendo aprendizado de máquina (machine learning) são muito utilizadas para tentar prever dados futuros de preço de petróleo (Na, Mikhaylov, Moiseev, 2019).

Neste contexto surge a ideia de realizar um estudo sobre utilização de técnicas de previsão de séries temporais para criar um modelo automatizado que realize previsões automáticas de um intervalo futuro de 5 anos do preço mensal do petróleo West Texas Intermediate (WTI).

REFERENCIAL TEÓRICO

Como base principal para as modelagens das séries temporais foram utilizados modelos conhecidos, o Prophet do Facebook se mostrou uma ferramenta interessante de previsão para dados de mercado de ações (Madhuri, Chinta, Kumar, 2020) e também para dados de COVID-19 (Khayyat, Laabidi, Almalki, Al-zahrani, 2021), no caso do COVID-

19 o modelo conseguiu prever bem casos com mortes, mas falhou ao tentar prever os casos de pacientes que se recuperaram.

Métodos de regressão Lasso foram utilizados para previsão de séries temporais de qualidade do ar, misturando com outras técnicas de previsão (Espinosa et al., 2021), com tentativas de previsão do dia posterior com valor de correlação chegando a 0,9 e também concentração gasosa (Song et al., 2023), mostrando que incluir uma regressão Lasso em modelos de previsão pode auxiliar a reduzir o erro final do nosso modelo.

E para complementar a análise utilizaremos o SARIMAX que já foi utilizado, por exemplo, em modelos de previsão de consumo de eletricidade (Atabay, Pagkalinawan, Pajarillo, Villanueva, Tylar, 2022), retornando bons resultados e previsões de temperatura atrelada ao clima (Elshewey et al., 2023), conseguindo modelos com correlação de 0,91.

Estudos indicam que técnicas de decomposição da série temporal, como decomposição de modo empírico (EMD), são eficientes para aumento da acurácia em modelos de previsão (Büyüksahin, Ertekin, 2019).

METODOLOGIA

A proposta deste estudo é a criação de um modelo de previsão de séries temporais automático, utilizando puramente bibliotecas Python para o estudo.

A Figura 1 ilustra a base de dados utilizada para este estudo:

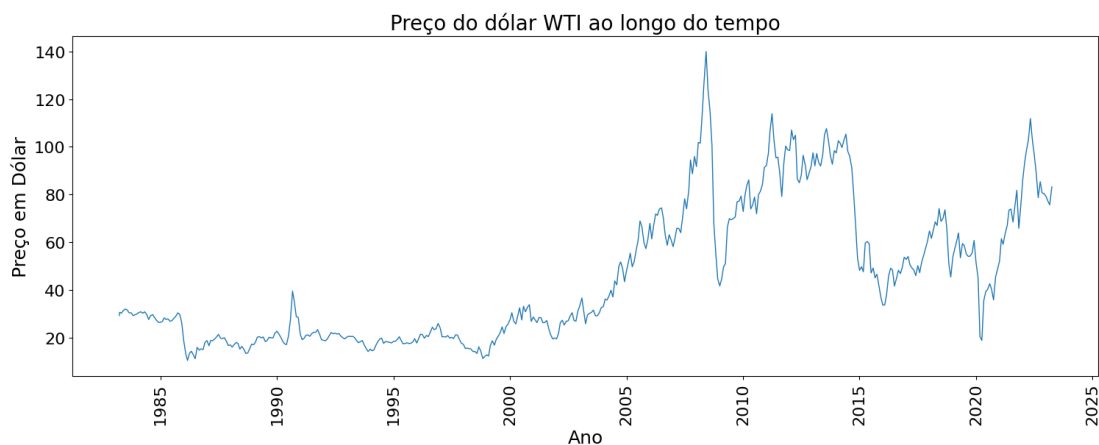


Figura 1. Preço do dólar WTI ao longo do tempo.

Com uma base de dados históricos sólida de preço mensal de petróleo WTI a ideia é realizar uma análise descritiva prévia para entendermos melhor os dados, visualização dos dados via gráfico, análise de dados faltantes e Outliers ocorrem de maneira automática nesta etapa.

Os dados datam desde março de 1983 até dados de abril de 2023, todo o treinamento dos modelos foi realizado com os dados entre 1983 a 2018 e foram separados dados de um período de 5 anos (2018 a 2023) para testes.

A proposta original é a realização da decomposição da nossa série temporal de treino, utilizando métodos de EMD, aqui utilizamos 3 métodos diferentes e ao final comparamos o resultado da utilização de cada um deles.

A Decomposição de Modo Empírico (EMD) é um procedimento iterativo que decompõe o sinal em um conjunto de componentes oscilatórios, chamados de Funções de Modo Intrínseco (IMFs), EMD será o primeiro método de decomposição utilizado neste estudo.

A Decomposição de Modo Empírico de Conjunto (EEMD) cria um conjunto de trabalhadores, cada um dos quais executa uma EMD em uma cópia do sinal de entrada com ruído adicionado. Quando todos os trabalhadores terminam seu trabalho, uma média sobre todos os trabalhadores é considerada como o resultado verdadeiro, EEMD será o segundo método de decomposição utilizado neste estudo.

O conjunto completo EMD com ruído adaptativo (CEEMDAN) realiza uma EEMD com a diferença de que a informação sobre o ruído é compartilhada entre todos os trabalhadores, CEEMDAN será o terceiro método de decomposição utilizado neste estudo.

A Figura 2 ilustra a base antes da decomposição para efeitos comparativos. As Figuras 3, 4, 5, 6, 7 e 8 ilustram como ocorre a decomposição da série temporal via EMD:

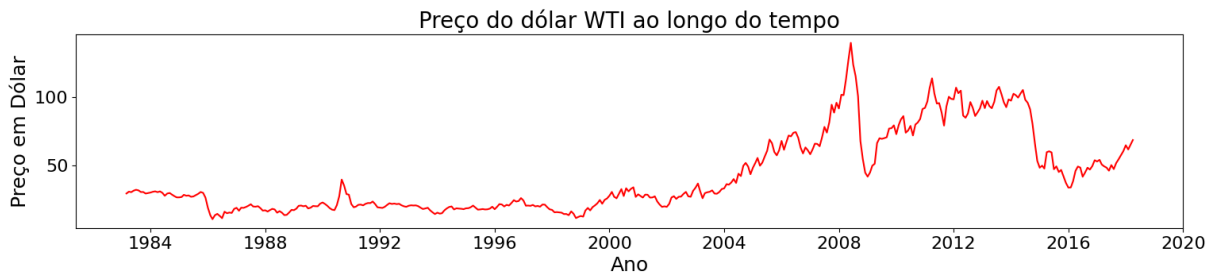


Figura 2. Preço do dólar WTI ao longo do tempo para efeito comparativo.

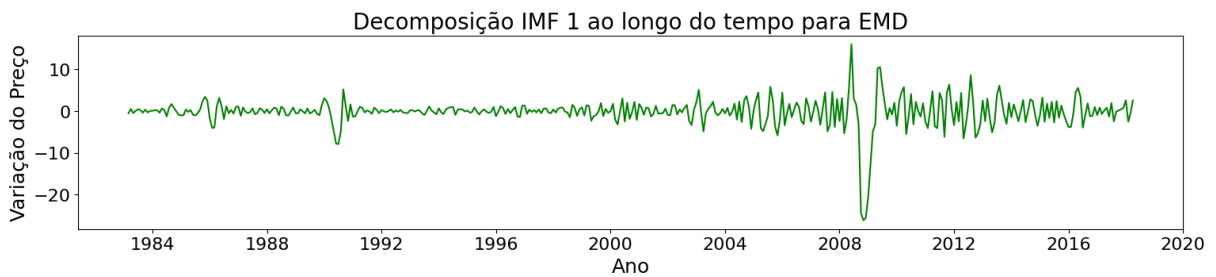


Figura 3. Decomposição IMF 1 ao longo do tempo para EMD.

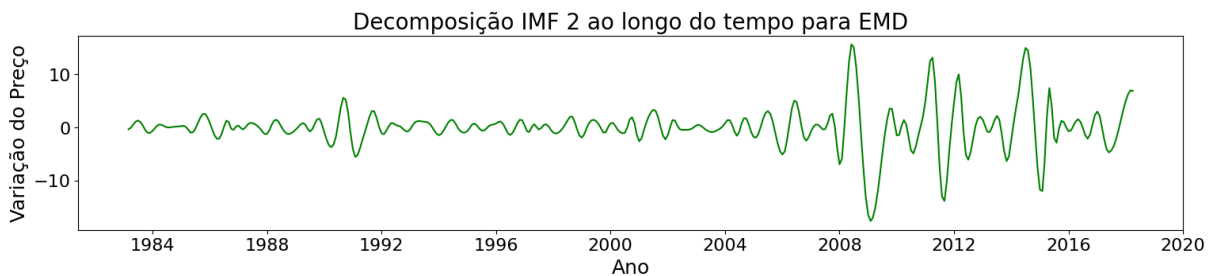


Figura 4. Decomposição IMF 2 ao longo do tempo para EMD.

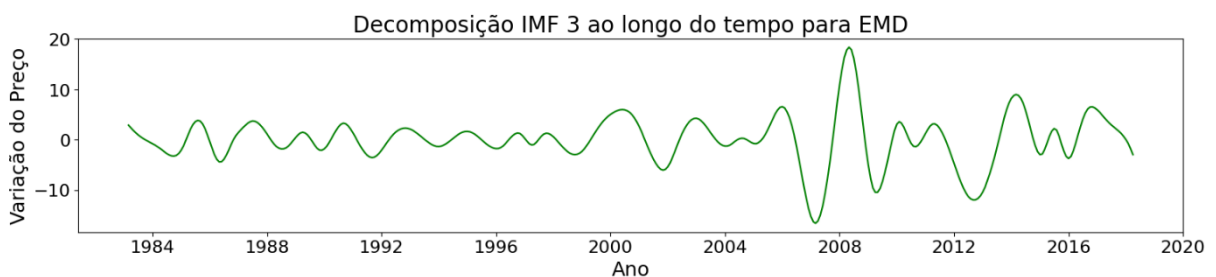


Figura 5. Decomposição IMF 3 ao longo do tempo para EMD.

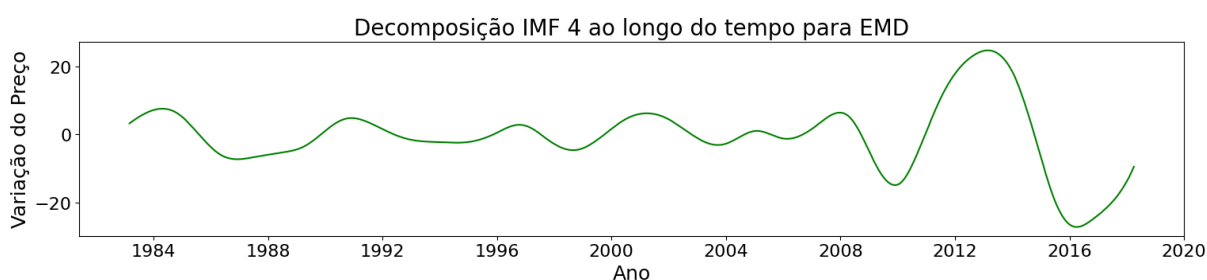


Figura 6. Decomposição IMF 4 ao longo do tempo para EMD.

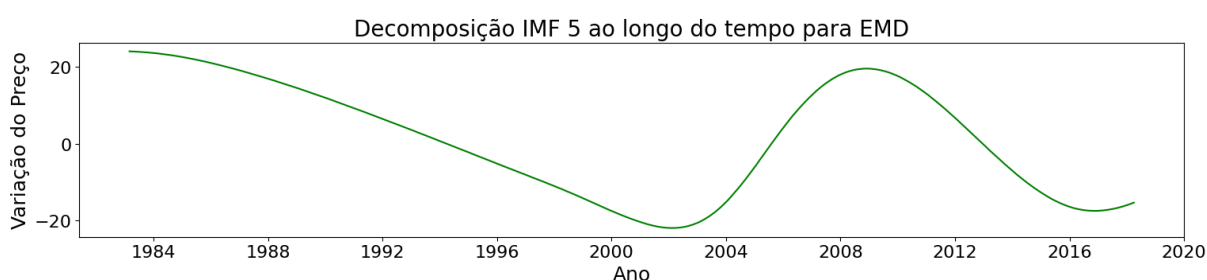


Figura 7. Decomposição IMF 5 ao longo do tempo para EMD.

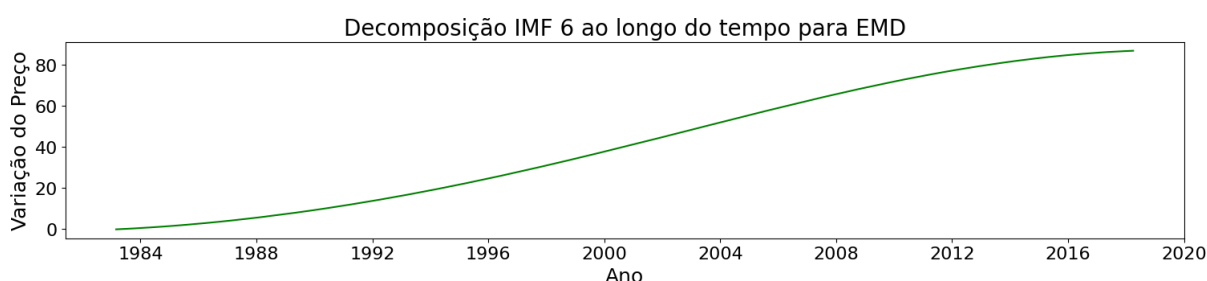


Figura 8. Decomposição IMF 6 ao longo do tempo para EMD.

As Figuras 9, 10, 11, 12, 13, 14 e 15 ilustram como ocorre a decomposição da série temporal via EEMD:

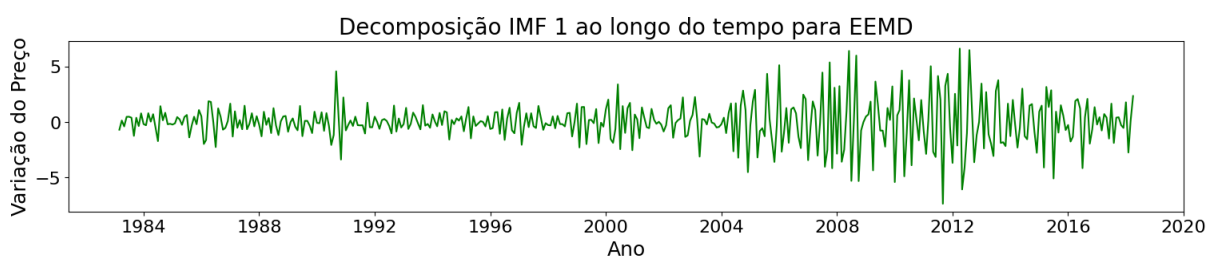


Figura 9. Decomposição IMF 1 ao longo do tempo para EEMD.

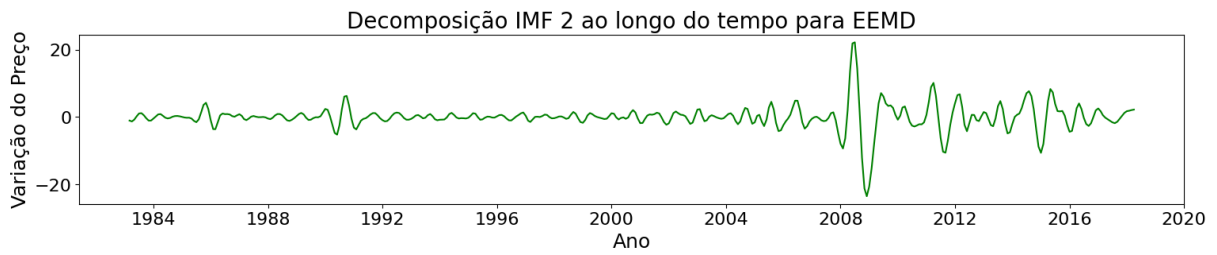


Figura 10. Decomposição IMF 2 ao longo do tempo para EEMD.

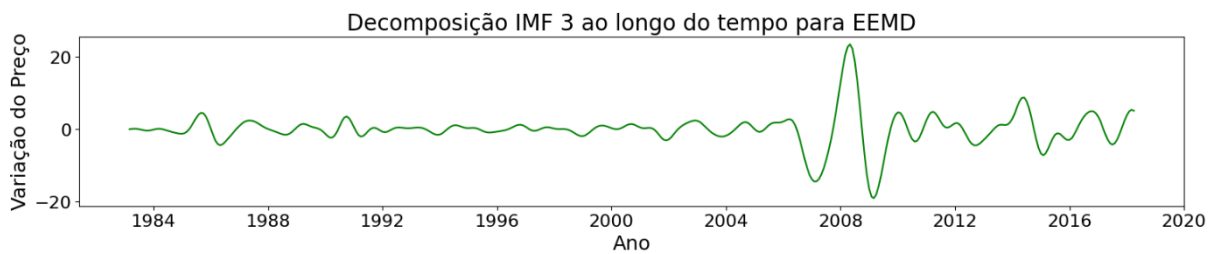


Figura 11. Decomposição IMF 3 ao longo do tempo para EEMD.

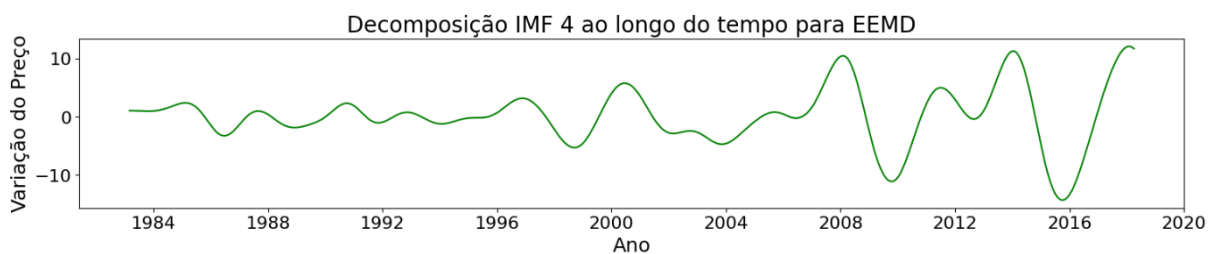


Figura 12. Decomposição IMF 4 ao longo do tempo para EEMD.

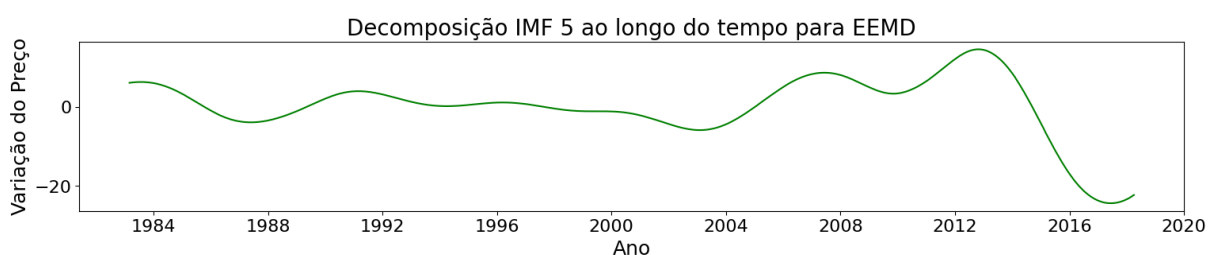


Figura 13. Decomposição IMF 5 ao longo do tempo para EEMD.

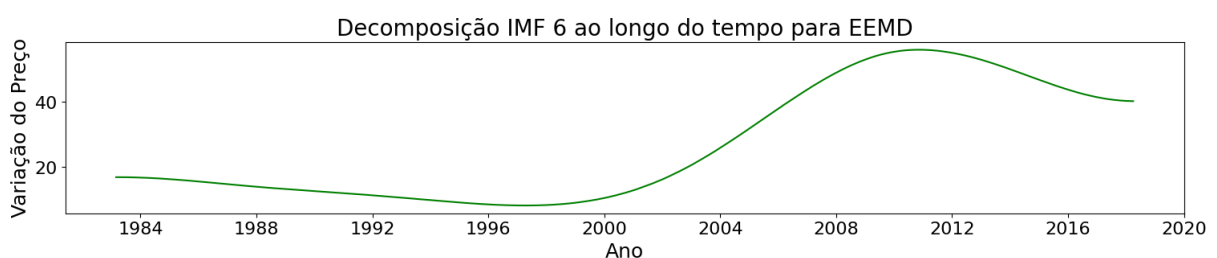


Figura 14. Decomposição IMF 6 ao longo do tempo para EEMD.

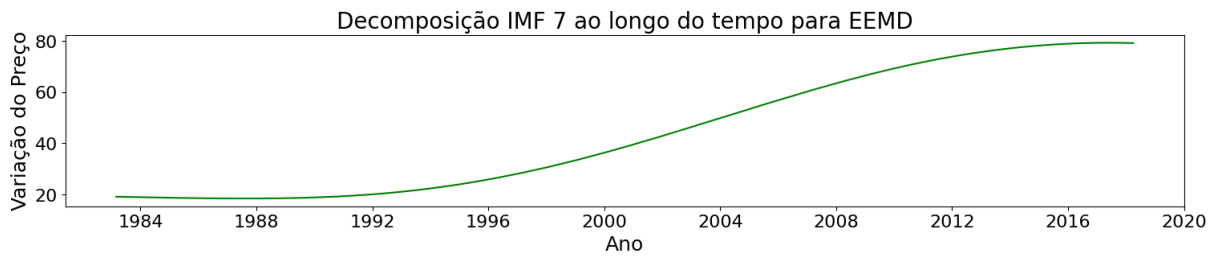


Figura 15. Decomposição IMF 7 ao longo do tempo para EEMD.

As Figuras 16, 17, 18, 19, 20 e 21 ilustram como ocorre a decomposição da série temporal via CEEMDAN:

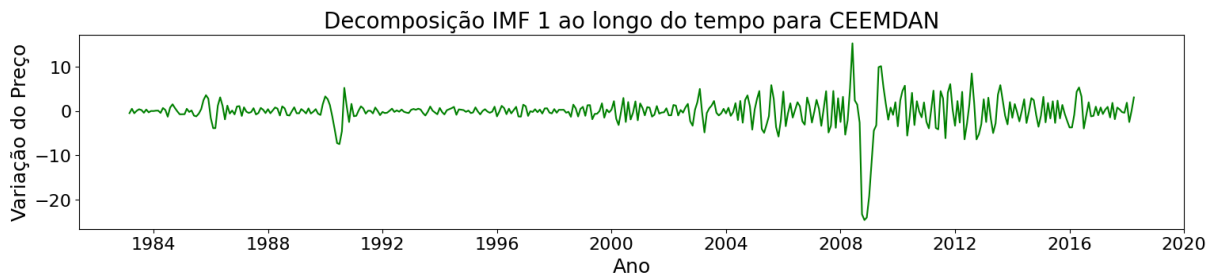


Figura 16. Decomposição IMF 1 ao longo do tempo para CEEMDAN.

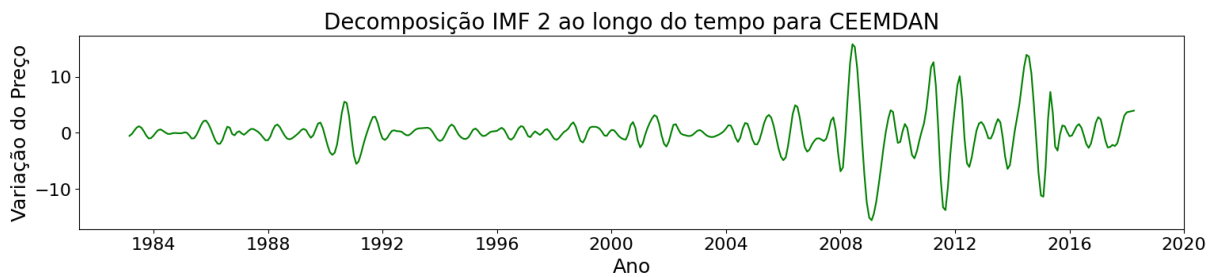


Figura 17. Decomposição IMF 2 ao longo do tempo para CEEMDAN.

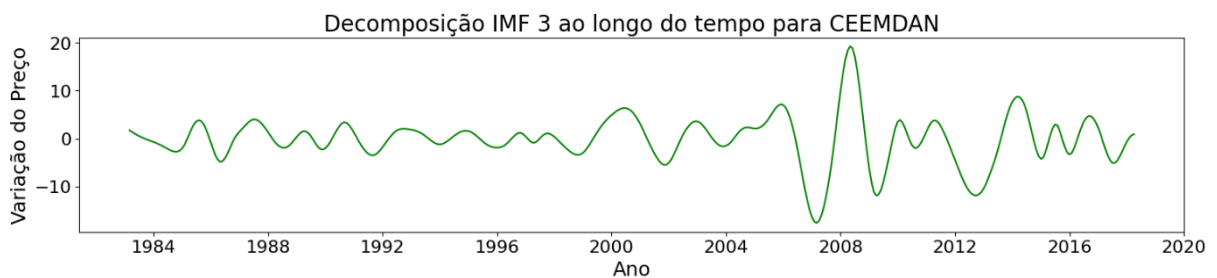


Figura 18. Decomposição IMF 3 ao longo do tempo para CEEMDAN.

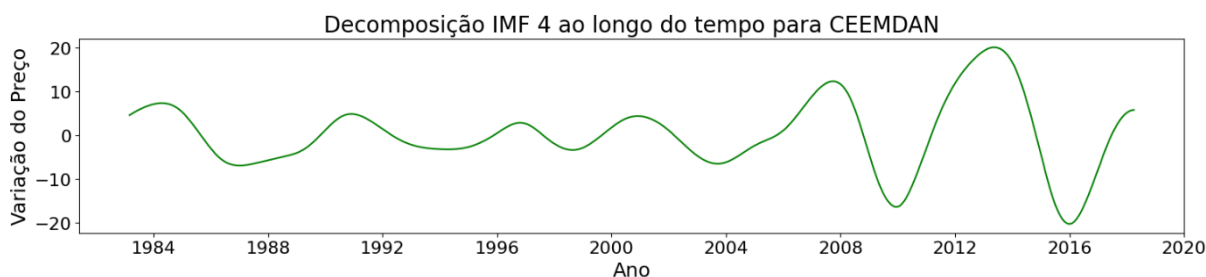


Figura 19. Decomposição IMF 4 ao longo do tempo para CEEMDAN.

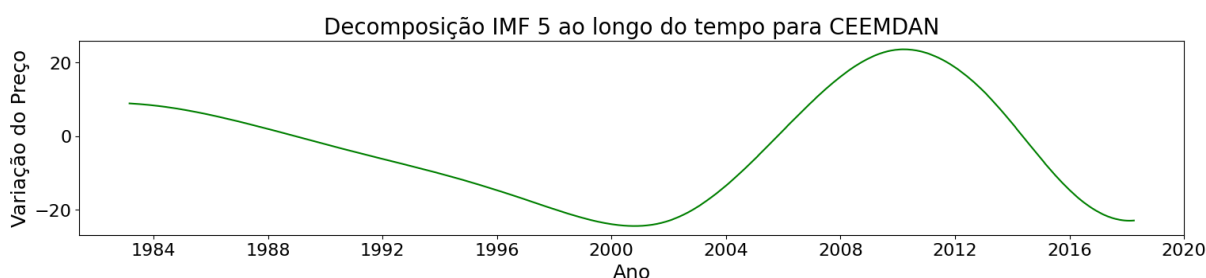


Figura 20. Decomposição IMF 5 ao longo do tempo para CEEMDAN.

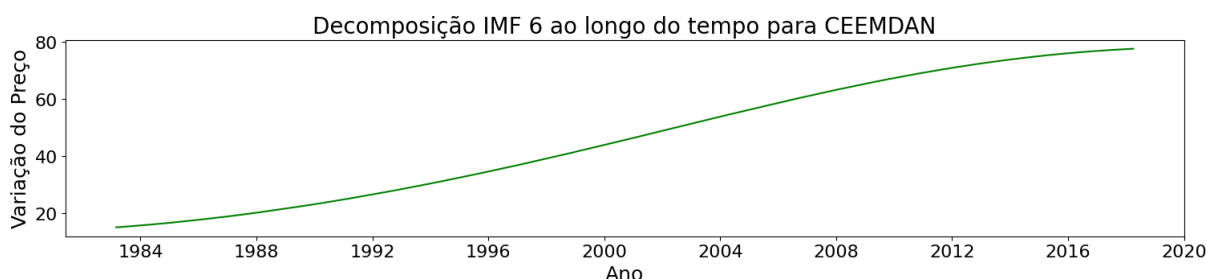


Figura 21. Decomposição IMF 6 ao longo do tempo para CEEMDAN.

Após a criação das IMFs, serão utilizados 3 tipos de modelos de previsão futura para cada IMF, ocorre a separação de cada IMF como treino e teste, seguindo a ideia de um período de 5 anos para teste. Em cada base de treino são aplicados os modelos de previsão: SARIMAX, Prophet e regressão LASSO.

Os resultados dos modelos de previsão são comparados com os dados reais de teste, retornando o erro percentual médio absoluto (MAPE).

A Tabela 1 contém os dados originados a partir da EMD, os dados destacados são os melhores em cada IMF:

Tabela 1. Comparação do MAPE gerado utilizando modelos de previsão nas IMFs geradas pelo EMD.

IMF	MAPE EMD SARIMAX	MAPE EMD Prophet	MAPE EMD Lasso
IMF 1	1,007	1,611	1,127
IMF 2	1,020	2,205	1,333
IMF 3	1,614	2,659	2,991
IMF 4	1,452	1,251	0,9701
IMF 5	0,1461	3,071	0,4323
IMF 6	0,0144	0,0447	0,0661

A Tabela 2 contém os dados originados a partir da EEMD, os dados destacados são os melhores em cada IMF:

Tabela 2. Comparação do MAPE gerado utilizando modelos de previsão nas IMFs geradas pelo EEMD.

IMF	MAPE EEMD SARIMAX	MAPE EEMD Prophet	MAPE EEMD Lasso
IMF 1	0,9785	0,9853	0,9932
IMF 2	1,188	2,446	1,555
IMF 3	1,086	1,232	1,085
IMF 4	0,9962	0,9980	0,8171
IMF 5	0,2990	2,684	2,899
IMF 6	0,0036	0,4044	0,3212
IMF 7	0,0008	0,0664	0,1023

A Tabela 3 contém os dados originados a partir da CEEMDAN, os dados destacados são os melhores em cada IMF:

Tabela 3. Comparação do MAPE gerado utilizando modelos de previsão nas IMFs geradas pelo CEEMDAN.

IMF	MAPE CEEMDAN SARIMAX	MAPE CEEMDAN Prophet	MAPE CEEMDAN Lasso
IMF 1	1,191	8,688	3,251
IMF 2	1,063	1,426	1,158
IMF 3	1,429	1,693	1,468
IMF 4	7,922	2,303	3,692
IMF 5	0,0299	8,830	1,706
IMF 6	0,0001	0,0330	0,0463

O modelo automatizado então seleciona os menores valores de MAPE entre os três disponíveis para cada IMF e com isso realiza a previsão dos dados futuros reais com base em qual dos 3 modelos de previsão apresentou o menor valor de MAPE.

Com isso será realizada a previsão de 5 anos futuros desejada. Para teste se o modelo automatizado é eficaz, foi utilizada a ideia de validação cruzada utilizando previsão de 5 anos futuros, mas para bases de treino menores do passado, e analisando valores de medida estatística tais como correlação e MAPE.

ANÁLISE DOS RESULTADOS E DISCUSSÕES

Ao se realizar o processo proposto foi obtido a Figura 22 como resultado:

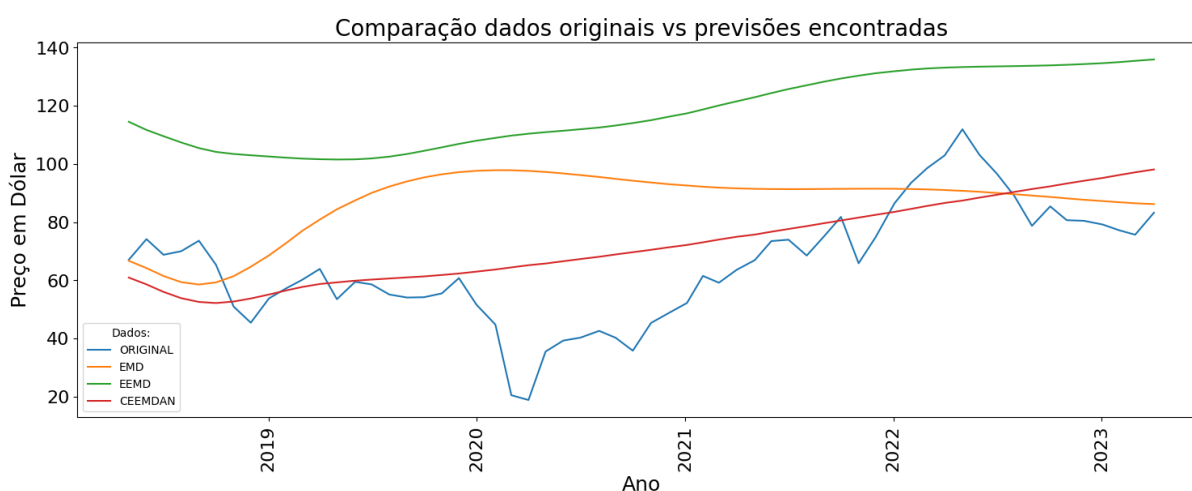


Figura 22. Comparação dados originais vs previsões encontradas.

Juntamente com a Figura 22 foram retiradas as informações de correlação, MAPE e valor final da previsão utilizado os 3 tipos de decomposição que estão na Tabela 4, os dados destacados são os melhores em cada coluna, excluindo o original:

Tabela 4. Comparação da correlação, do MAPE e do valor final gerados utilizando modelos de previsão nas IMFs de cada tipo de decomposição.

Tipo:	Correlação:	MAPE:	Valor Final:
ORIGINAL	1	0	83,24
EMD	-0,1262	0,5694	86,19
EEMD	0,706	1,001	135,9
CEEMDAN	0,6323	0,2852	98,10

É possível analisar que o as curvas de previsão acabam sendo um pouco mais suavizadas que a curva real, mas foram obtidos bons resultados, o modelo utilizando



como base a decomposição EMD conseguiu chegar a um valor muito próximo do real mesmo com um intervalo de previsão altíssimo de 5 anos.

O modelo de previsão utilizando CEEMDAN conseguiu tanto uma correlação alta em relação a curva real junto com um valor de MAPE pequeno.

E o modelo EEMD foi o que apresentou maior valor de correlação entre as 3 decomposições apresentadas.

Para verificação se este método de previsão de séries temporais é interessante será realizado validações cruzadas com 5 bases de teste do passado e serão apresentados os mesmos dados de saída que foram informados no estudo.

A Figura 23 mostra os dados do resultado da validação que possui como teste dados de 2008 à 2013:

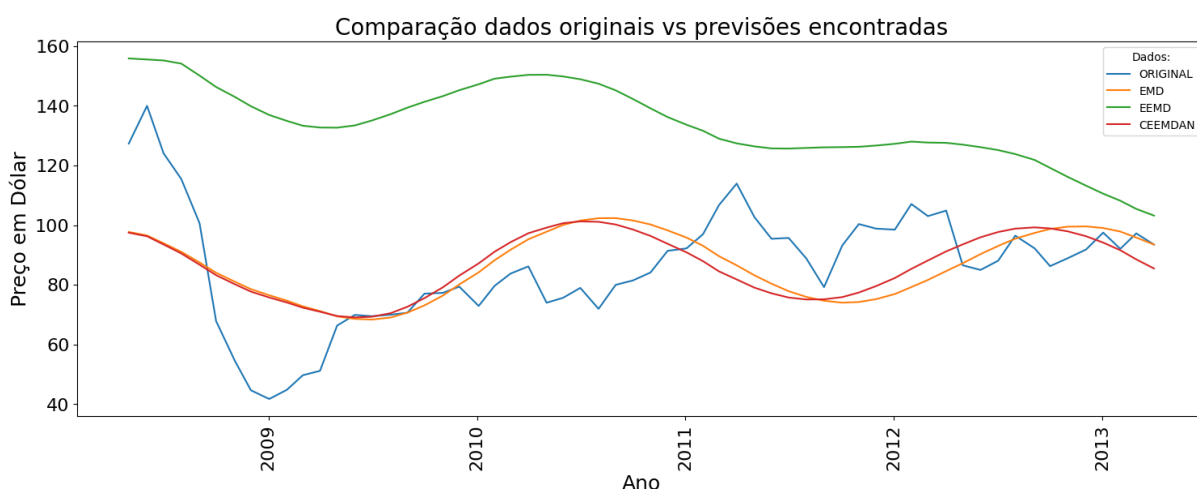


Figura 23. Comparação dados originais vs previsões encontradas, validação cruzada de 2008 à 2013.

A Tabela 5 informa correlação, MAPE e valor final da validação cruzada que possui como teste dados de 2008 à 2013:

Tabela 5. Comparação da correlação, do MAPE e do valor final gerados utilizando modelos de previsão nas IMFs de cada tipo de decomposição, base de teste de 2008 à 2013.

Tipo:	Correlação:	MAPE:	Valor Final:
ORIGINAL	1	0	93,46
EMD	0,3843	0,1866	93,48
EEMD	-0,0775	0,6593	103,2
CEEMDAN	0,3922	0,1883	85,46



A Tabela 6 informa correlação, MAPE e valor final da validação cruzada que possui como teste dados de 1994 à 1999:

Tabela 6. Comparação da correlação, do MAPE e do valor final gerados utilizando modelos de previsão nas IMFs de cada tipo de decomposição, base de teste de 1994 à 1999.

Tipo:	Correlação:	MAPE:	Valor Final:
ORIGINAL	1	0	18,66
EMD	-0,2858	0,1465	19,49
EEMD	-0,5667	1,220	48,00
CEEMDAN	-0,5326	0,6506	47,74

A Tabela 7 informa correlação, MAPE e valor final da validação cruzada que possui como teste dados de 1997 à 2002:

Tabela 7. Comparação da correlação, do MAPE e do valor final gerados utilizando modelos de previsão nas IMFs de cada tipo de decomposição, base de teste de 1997 à 2002.

Tipo:	Correlação:	MAPE:	Valor Final:
ORIGINAL	1	0	27,29
EMD	0,0788	0,2415	21,37
EEMD	0,3957	0,7090	36,10
CEEMDAN	0,1440	0,2356	21,39

A Tabela 8 informa correlação, MAPE e valor final da validação cruzada que possui como teste dados de 1993 à 1998:

Tabela 8. Comparação da correlação, do MAPE e do valor final gerados utilizando modelos de previsão nas IMFs de cada tipo de decomposição, base de teste de 1993 à 1998.

Tipo:	Correlação:	MAPE:	Valor Final:
ORIGINAL	1	0	15,39
EMD	-0,1199	0,1572	20,32
EEMD	0,0332	0,9420	36,57
CEEMDAN	-0,1978	0,1532	19,38

A Tabela 9 informa correlação, MAPE e valor final da validação cruzada que possui como teste dados de 2016 à 2021:

Tabela 9. Comparação da correlação, do MAPE e do valor final gerados utilizando modelos de previsão nas IMFs de cada tipo de decomposição, base de teste de 2016 à 2021.

Tipo:	Correlação:	MAPE:	Valor Final:
ORIGINAL	1	0	63,58
EMD	0,0240	0,6577	92,77
EEMD	0,5801	1,524	126,1
CEEMDAN	0,0564	0,5209	83,36

CONCLUSÕES/CONSIDERAÇÕES FINAIS

O método estudado retornou bons resultados, na modelagem principal (2018 – 2023) vemos que o CEEMDAN retorna ótimos valores de correlação e de MAPE, EEMD retornou o melhor valor de correlação, mas foi o método EMD que chegou ao valor mais próximo do real depois de 5 anos.

Os resultados da validação cruzada variam, mas no geral o EEMD retorna os maiores valores de correlação e EMD e CEEMDAN retornam os melhores valores de MAPE e valor final, muitas vezes até bem próximos.

Perceber uma consistência no modelo criado e obter resultados satisfatórios para previsão do preço do petróleo, que tende a variar muito e sofrer influência de diversos fatores externos, indica o quão poderoso os métodos de previsão de séries temporais são.

Tentar inserir métodos mais sofisticados como redes neurais na análise, utilizar intervalos de tempo diferentes como teste ou até mesmo modelar outros dados, de outra matéria prima, ou qualquer série temporal desejada são caminhos para seguir a partir deste estudo.

REFERÊNCIAS

ALSALLAKH, M. A., SIVAKUMARAN, S., KENNEDY, S., VASILEIOU, E., LYONS, R. A., ROBERTSON, C., SHEIKH, A. (2021). Impact of COVID-19 lockdown on the incidence and mortality of acute exacerbations of chronic obstructive pulmonary disease: national interrupted time series analyses for Scotland and Wales. *BMC Medicine* (Vol. 19, artigo 124).

ATABAY, F. V., PAGKALINAWAN, R. M., PAJARILLO, S. D., VILLANUEVA, A. R., TAYLAR, J. V. (2022). Multivariate Time Series Forecasting using ARIMAX, SARIMAX, and RNN-based Deep Learning Models on Electricity Consumption. 3rd International Informatics and Software Engineering Conference (IISEC) (pp. 1-6). Ankara, Turkey.

- BÜYÜKŞAHİN, Ü. C., ERTEKİN, S. (2019). Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition. *Neurocomputing* (Vol. 361, pp. 151-163).
- ELSHEWEY, A. M., SHAMS, M. Y., ELHADY, A. M., SHOHIEB, S. M., ABDELHAMID, A. A., IBRAHIM, A., TAREK, Z. (2023). A Novel WD-SARIMAX Model for Temperature Forecasting Using Daily Delhi Climate Dataset. *MDPI, Sustainability* (Vol. 15, Issue 1, Artigo 757).
- ESPINOSA, R., PALMA, J., JIMÉNEZ, F., KAMIŃSKA, J., SCIAVICCO, G., LUCENA-SÁNCHEZ, E. (2021). A time series forecasting based multi-criteria methodology for air quality prediction. *Applied Soft Computing* (Vol. 113, Pt. A).
- JAMAL, S., BLOOM, J. S. (2020). On Neural Architectures for Astronomical Time-series Classification with Application to Variable Stars. *The Astrophysical Journal Supplement Series* (Vol. 250, No. 2).
- KAREVAN, Z., SUYKENS, J. A. K. (2020). Transductive LSTM for time-series prediction: An application to weather forecasting. *Neural Networks* (Vol. 125, p. 1-9). Leuven, Belgium.
- KHAYYAT, M., LAABIDI, K., ALMALKI, N., AL-ZAHRANI, M. (2021). Time Series Facebook Prophet Model and Python for COVID-19 Outbreak Prediction. *Computers, Materials & Continua*.
- MADHURI, C. R., CHINTA, M., KUMAR, V. V. N. V. P. (2020). Stock Market Prediction for Time-series Forecasting using Prophet upon ARIMA. *7th International Conference on Smart Structures and Systems (ICSSS)* (pp. 1-5). Chennai, India.
- NA, J., MIKHAYLOV, A., MOISEEV, N. (2019). Oil Price Predictors: Machine Learning Approach. *International Journal of Energy Economics and Policy* (Vol. 9, Issue 5, p. 1-6).
- NIELSEN, A. (2019). *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*. Sebastopol, CA: O'Reilly.
- PHAN, D. H. B., TRAN, V. T., NGUYEN, D. T., LE, A. (2020). The importance of managerial ability on crude oil price uncertainty-firm performance relationship. *Energy Economics* (Vol. 88).
- SEZER, O. B., GUDELEK, M. U., OZBAYOGLU, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied Soft Computing* (Vol. 90). Ankara, Turkey.
- SONG, S., CHEN, J., MA, L., ZHANG, L., HE, S., DU, G., WANG, J. (2023). Research on a working face gas concentration prediction model based on LASSO-RNN time series data. *Heliyon*.
- ZHANG, D., CHEN, S., LIWEN, L., XIA, Q. (2020). Forecasting Agricultural Commodity Prices Using Model Selection Framework With Time Series Features and Forecast Horizons. *IEEE Access* (Vol. 8, p. 28197-28209).

Submissão: maio de 2023. Aceite: junho de 2023. Publicação: agosto de 2023.

