Previsão da poluição do ar na cidade de São Paulo utilizando séries temporais - Estilo de artigo de curso da FCI

Aparecida Vânia de Jesus, Lucas Gomes Porfírio da Silva, Vanessa Hacklauer de Aguiar, Wagner de Mendonça Trindade.

[1]Faculdade de Computação e Informática (FCI)
Universidade Presbiteriana Mackenzie – São Paulo, SP – Brasil

Resumo. *A poluição do ar é um grave problema em grandes metrópoles, como São Paulo, onde os níveis de poluentes frequentemente ultrapassam os limites recomendados pela OMS. O aumento de poluentes, como PM2,5 e ozônio, está relacionado à frota veicular, atividades industriais e incêndios florestais, contribuindo para milhões de mortes prematuras globalmente. Este trabalho visa prever picos de poluição com antecedência de três meses, usando dados de qualidade do ar fornecidos pelo Instituto de Energia e Meio Ambiente (Iema) de 2020 a 2022, para auxiliar no planejamento urbano e mitigação de riscos à saúde pública.*

Os dados, coletados de 20 estações em São Paulo, abrangem poluentes como PM10, PM2,5, O_3, CO, NO_2 e SO_2, e são categorizados pelo Índice de Qualidade do Ar (IQA). Técnicas de séries temporais foram aplicadas para identificar padrões e desenvolver modelos preditivos que possam orientar medidas preventivas, como restrições veiculares e incentivo ao transporte coletivo sustentável.

O projeto está alinhado ao ODS 11.6, que busca reduzir o impacto ambiental nas cidades, promovendo soluções práticas para gestão da qualidade do ar. Além de contribuir para a ciência de dados, a proposta visa implementar sistemas de alerta em tempo real, auxiliando autoridades locais e serviços de saúde na proteção da população. Espera-se que a ferramenta proposta melhore a qualidade de vida em São Paulo, reduzindo os impactos adversos à saúde causados pela poluição.

Palavras-chave: *Série Temporal, poluição do ar, previsão de poluentes, qualidade do ar, sustentabilidade*

Abstract. *Air pollution is a serious issue in large cities like São Paulo, where pollutant levels often exceed the limits recommended by the WHO. The rise in pollutants such as PM2.5 and ozone is linked to vehicle fleets, industrial activities, and wildfires, contributing to millions of premature deaths globally. This study aims to predict pollution peaks three months in advance, using air quality data from the Instituto de Energia e Meio Ambiente (Iema) from 2020 to 2022, to assist in urban planning and mitigate public health risks.*

The data, collected from 20 monitoring stations in São Paulo, includes pollutants such as PM10, PM2.5, O_3, CO, NO_2, and SO_2, categorized by the Air Quality Index (AQI). Time series techniques were applied to identify

patterns and develop predictive models that can guide preventive measures, such as vehicle restrictions and the promotion of sustainable public transportation.

The project aligns with SDG 11.6, which seeks to reduce environmental impact in cities, providing practical solutions for air quality management. In addition to contributing to data science, the proposal aims to implement real-time alert systems, supporting local authorities and health services in protecting the population. The expected outcome is a tool that improves the quality of life in São Paulo by minimizing the adverse health effects caused by air pollution.

Keywords: *Time series, air pollution, pollutant forecasting, air quality, sustainability*

1. Introdução

A poluição do ar é um dos maiores desafios ambientais enfrentados pelas grandes metrópoles, com consequências diretas para a saúde pública e a qualidade de vida da população. Na cidade de São Paulo, uma das maiores da América Latina, os níveis de poluentes atmosféricos frequentemente ultrapassam os limites recomendados por organismos internacionais, como a Organização Mundial da Saúde (OMS).Segundo o relatório Estado do Ar Global 2024, o Brasil teve um aumento em mais de 10% nas exposições ao ozônio ambiental na última década, ao lado de países como Índia, Nigéria e Paquistão e acima de 90% do total global das mortes por poluição do ar são provocadas pelas chamadas PM2,5 e causou 8,1 milhões de mortes em 2021. O impacto ambiental negativo causado pela emissão de gases poluentes está diretamente relacionado ao aumento da frota de veículos, à atividade industrial e à gestão inadequada de resíduos. Além disso, a poluição atmosférica, provocada pelo ser humano e outras fontes como incêndios florestais, está associada a cerca de 135 milhões de mortes prematuras no mundo entre 1980 e 2020, segundo estudo de uma universidade de Singapura (Revista Planeta, 2023).

A previsão de picos de poluição surge como uma ferramenta essencial para o planejamento urbano sustentável, proporcionando uma visão proativa das flutuações dos níveis de poluentes atmosféricos. Medidas de mitigação podem ser implementadas de maneira mais eficaz quando há antecipação dos momentos críticos de poluição, como a restrição de veículos em determinadas áreas e a promoção do transporte coletivo de baixa emissão. Este trabalho, inserido na área de ciência de dados e séries temporais, propõe a aplicação de técnicas preditivas para auxiliar na mitigação de riscos à saúde pública. A previsão com antecedência de três meses visa permitir que tanto as autoridades quanto a população possam adotar medidas preventivas, minimizando os impactos sobre a saúde e a qualidade do ar.

Os dados utilizados no trabalho são provenientes do Instituto de Energia e Meio Ambiente (Iema), que disponibiliza informações sobre a qualidade do ar na cidade de São Paulo. Os dados são coletados de hora em hora, e o presente estudo abrange três anos de dados, de 2020 a 2022. Devido ao grande volume de dados, foi escolhida a média diária como base para a aplicação dos algoritmos de séries temporais. O Índice de Qualidade do Ar (IQA) é calculado a partir das concentrações de poluentes como material particulado (PM10 e PM2,5), ozônio (O_3), monóxido de carbono (CO), dióxido de nitrogênio (NO_2) e dióxido de enxofre (SO_2). Os dados incluem também as estações de coleta, das quais foram selecionadas as localizadas na cidade de São Paulo: Cambuci, Capão Redondo, Centro-SP, Cerqueira César, Congonhas, Grajaú-Parelheiros, Ibirapuera, Interlagos, Itaim Paulista, Itaquera, Lapa, Marginal Tietê-Pte dos Remédios, Mooca, Nossa Senhora do Ó, Parque Dom Pedro II, Penha, Perus, Pinheiros, Santana, Santo Amaro e São Miguel Paulista.

O Índice de Qualidade do Ar é dividido em categorias que variam de Boa (0-50) - A qualidade do ar é considerada satisfatória; Moderada (51-100) - A qualidade do ar é aceitável; Ruim (101-150) - Pessoas com doenças respiratórias ou cardiovasculares, crianças e idosos podem apresentar sintomas mais sérios; Muito ruim (151-200) - Pode afetar todos os indivíduos, pessoas sensíveis podem ter efeitos mais graves e a população em geral, e; Péssima (acima de 200) - Os riscos são elevados para a saúde da população.

Com foco em alcançar nossos objetivos foram estabelecidos como objetivos a coleta e pré-processamento dos dados de qualidade do ar provenientes do Instituto de Energia e Meio Ambiente (Iema) referentes à cidade de São Paulo no período de 2020 a 2022. Analisar os padrões temporais dos níveis de poluentes atmosféricos utilizando técnicas de séries temporais. Desenvolver modelos preditivos capazes de prever os níveis de poluição com antecedência de três meses. Validar os modelos preditivos e avaliar sua precisão e propor recomendações para a implementação prática do sistema preditivo por agências ambientais e autoridades locais.

Este projeto está alinhado aos Objetivos de Desenvolvimento Sustentável (ODS), especialmente à meta 11.6, que busca reduzir o impacto ambiental negativo per capita nas cidades. O desenvolvimento de um sistema preditivo para picos de poluição tem grande aplicabilidade prática, podendo ser integrado a sistemas de alerta em tempo real e utilizado por agências ambientais, governos locais e serviços de saúde. Dessa forma, este trabalho não apenas contribui para a área de ciência de dados, mas também para a promoção de uma cidade mais saudável e sustentável.

Ao final, espera-se fornecer uma ferramenta que, além de prever com maior precisão os momentos de alta concentração de poluentes, possa ser utilizada para minimizar os efeitos adversos à saúde e melhorar a qualidade de vida dos habitantes de São Paulo.

2. Referencial Teórico

A poluição atmosférica é um dos principais problemas ambientais que afetam a saúde humana e o ecossistema global. Os principais poluentes incluem material particulado (MP10 e MP2,5), ozônio (O_3), monóxido de carbono (CO), dióxido de nitrogênio (NO_2) e dióxido de enxofre (SO_2). A exposição a esses poluentes está associada a uma série de efeitos adversos à saúde, como doenças respiratórias, cardiovasculares e aumento da mortalidade prematura (WORLD HEALTH ORGANIZATION, 2021). Estudos epidemiológicos demonstram que a inalação de partículas finas, como MP2,5, pode penetrar profundamente nos pulmões e entrar na corrente sanguínea, causando inflamação sistêmica (POPE III et al., 2019).

No contexto das grandes metrópoles, a concentração desses poluentes tende a ser mais elevada devido à intensa atividade industrial, alta densidade veicular e processos urbanos que contribuem para emissões atmosféricas (SANTOS et al., 2018). Portanto, o monitoramento e a previsão da qualidade do ar são essenciais para a implementação de políticas públicas eficazes que visam mitigar os impactos negativos da poluição. A análise de séries temporais é uma ferramenta estatística fundamental para modelar e prever fenômenos ambientais. Técnicas como modelos autorregressivos integrados de médias móveis (ARIMA) e suas variantes sazonais (SARIMA) são amplamente utilizadas para capturar padrões de tendência e sazonalidade em dados ambientais (BOX et al., 2015). Além disso, abordagens baseadas em aprendizado de máquina, como redes neurais recorrentes (RNN) e redes de memória de longo curto prazo (LSTM), têm sido empregadas para modelar relações não lineares e complexas presentes nos dados de poluição do ar (ZHANG et al., 2018).

Diversos estudos têm explorado a aplicação dessas técnicas para prever concentrações de poluentes atmosféricos. O estudo intitulado "Previsão de concentração de material particulado inalável na Região da Grande Vitória, ES, Brasil", publicado na Revista Ciello Brasil. O trabalho utilizou modelos de séries temporais para analisar componentes não observáveis, como tendência, sazonalidade e aleatoriedade, focando no monitoramento do poluente MP10 em uma única estação. Os autores aplicaram os modelos ARMA, SARMA, ARMAX e SARIMAX, encontrando que o modelo SARIMAX apresentou o melhor desempenho, estimando corretamente 64,39% dos eventos de alta concentração de MP10. Esses resultados destacam a eficácia de incluir variáveis exógenas e componentes sazonais no modelo preditivo.

No estudo intitulado "Previsão da concentração de material particulado inalável, através de modelos estatísticos de séries temporais para o município de Canoas, Rio Grande do Sul", diversos modelos de séries temporais foram aplicados para prever as concentrações de PM10. Foram comparados os modelos ARIMA, ARMAX (com variáveis exógenas CO e SO_2) e Alisamento Exponencial, usando como métrica de

avaliação o MAPE (Erro Percentual Absoluto Médio). Os dados eram medições horárias dos seguintes poluentes do ar: material particulado inalável com diâmetro inferior a 10 μm (PM10),dióxido de enxofre (SO2) e monóxido de carbono (CO), no período compreendido entre 01/01/2014 a 31/12/2014, foram convertidos para médias diárias. Os resultados indicaram que o modelo ARMAX apresentou a melhor acurácia com MAPE de 5,5%, seguido pelo ARIMA com 8,8%, enquanto os modelos de Alisamento Exponencial tiveram MAPE superior a 14%, não sendo recomendados para essa série temporal específica. Esse estudo evidencia que a inclusão de variáveis exógenas melhora significativamente a precisão preditiva, tornando o modelo ARMAX o mais adequado para a previsão da qualidade do ar em Canoas.

No artigo "Uma abordagem de séries temporais para a transformação de cidades inteligentes: o problema da poluição do ar em Brescia". Neste trabalho, diversos modelos foram implementados para prever as concentrações de PM2,5 e PM10 para o dia seguinte. Os autores compararam três métodos distintos: ARIMA (estatístico), rede LSTM (aprendizado de máquina) e um modelo híbrido CNN-LSTM e utilizaram como métrica de avaliação o RMSE(Root Mean Squared Error - é a raiz quadrada da média dos quadrados dos erros entre os valores previstos pelo modelo e os valores reais) e o MAE(Mean Absolute Error - é a média dos valores absolutos dos erros entre os valores previstos e os valores reais). Para PM10, os RMSE e MAE dos dados de teste mostram que o modelo CNN-LSTM-M teve o menor erro, com um RMSE de 11.02 e MAE de 7.78. No caso de PM2.5, o mesmo modelo apresentou RMSE de 8.69 e MAE de 6.43, também sendo o mais preciso. No geral, os modelos multivariados, que incorporam dados meteorológicos, superaram os modelos univariados devido à consideração de variáveis adicionais que ajudam a prever os níveis de poluição com maior precisão.

A análise detalhada desses estudos fornece diretrizes importantes para o desenvolvimento de modelos preditivos mais robustos, que considerem a sazonalidade, a tendência e os fatores externos que influenciam a concentração de poluentes atmosféricos. As análises nos forneceram conclusões importantes sobre a base de dados escolhida, estamos trabalhando com uma série temporal é Sazonal, Multivariada, Não Estacionária, e Não Linear.

A variabilidade nos dados de qualidade do ar pode ser influenciada por diversos fatores, como condições meteorológicas, variações sazonais e fontes de emissão específicas em diferentes localidades (HE et al., 2020), Nossos dados variam de acordo com a estação e com os poluentes medidos em cada estação. Essas discrepâncias e picos observados nos dados podem dificultar a modelagem precisa das séries temporais utilizando todos os poluentes.

Após análise do referencial teórico, decidimos por analisar o poluente CO e se necessário, utilizar os poluentes MP10 e MP2.5 como variáveis exógenas para auxiliar

os algoritmos no entendimento do problema, considerando técnicas que possam lidar com a não linearidade e a não estacionaridade presentes nos dados ambientais.

3. Metodologia

Coleta e Ingestão dos Dados:

Os dados utilizados para a previsão foram extraídos do site do Instituto de Energia e Meio Ambiente (Iema), que disponibiliza medições horárias da qualidade do ar na cidade de São Paulo. Estes dados incluem concentrações de poluentes como PM10, PM2,5, O_3, CO, NO_2 e SO_2, Para as estações de coleta, será aplicado um filtro para selecionar apenas os dados das estações situadas na cidade de São Paulo.

Pré-processamento dos Dados:

Como os dados estão disponíveis de hora em hora, foi realizado um pré-processamento para transformar essas medições em médias diárias, a fim de reduzir a dimensionalidade e facilitar o tratamento computacional. Além disso, será realizada a verificação de dados ausentes, outliers e correção de inconsistências.

Análise Exploratória dos Dados (EDA):

Uma análise exploratória será conduzida para identificar padrões sazonais e tendências nos dados históricos. Visualizações como gráficos de linha, histogramas e autocorrelações ajudarão a entender as variações dos níveis de poluentes ao longo do tempo. Esta etapa permitirá definir se há periodicidades importantes que possam influenciar os modelos preditivos.

Seleção dos dados relevantes:

Nesta fase, as variáveis relevantes selecionadas é Valor e, se necessário, a separação dos Poluentes caso a variável Valor não gere os resultados esperados. Foi necessária a utilização das variáveis derivadas média móvel e desvios padrão diários para facilitar o ajuste dos modelos de séries temporais. A variável alvo será CO devido a ter sido identificado como o principal poluente e utilizaremos variáveis exógenas 'MP10', 'MP2.5' para tentar melhorar o desempenho.

Divisão e Modelagem de Dados:

Os dados serão divididos em conjuntos de treino e teste - 0.7 e 0.3. Os poluentes selecionados foram CO,'MP10', 'MP2.5'que geraram um novo dataset que será utilizado no treinamento.

Treinamento:

Aplicados os modelos de previsão de séries temporais, selecionados devido às características do dataset. Os modelos Sarima, SarimaX e XGBoost/LightGBM. Esses modelos serão ajustados e treinados utilizando o conjunto de dados de treino e teste, com o objetivo de prever os níveis de CO para os próximos três meses, sendo utilizado o modelo que apresentar o melhor resultado.

Avaliação do Modelo:

O desempenho dos modelos será avaliado utilizando métricas como Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) e coeficiente de determinação (R^2).Inicialmente não foi utilizada a validação cruzada.

Ajustes:

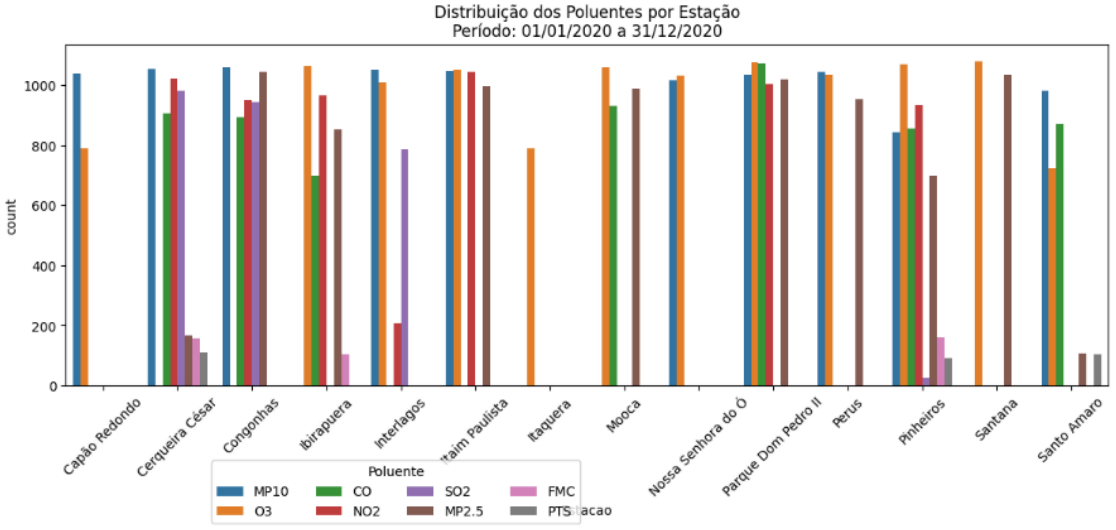
Caso haja a necessidade de ajustes nos parâmetros dos algoritmos, eles serão realizados neste momento.

4. Resultados e discussão

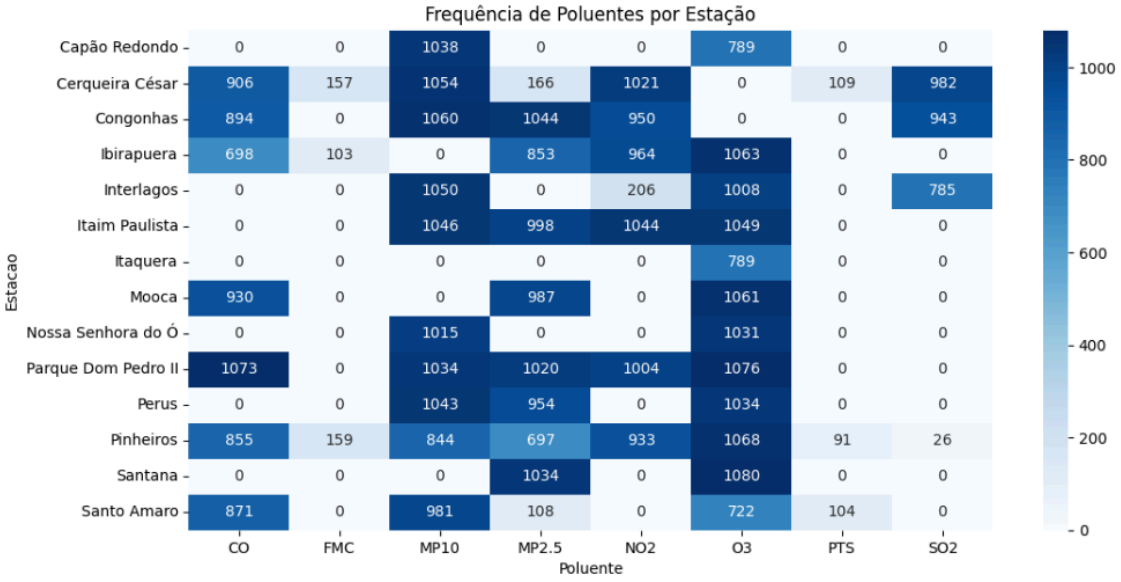
EDA e Pré-processamento dos dados

Exploração e análise dos dados. Discussão e análise dos dados empregados (qualidade, limitações, simplificações ou recortes adotados etc.). Tarefas de preparação dos dados (transformações, compactação e encodes, junções de dados etc.).

No gráfico abaixo, temos a representação da distribuição dos poluentes por estação.

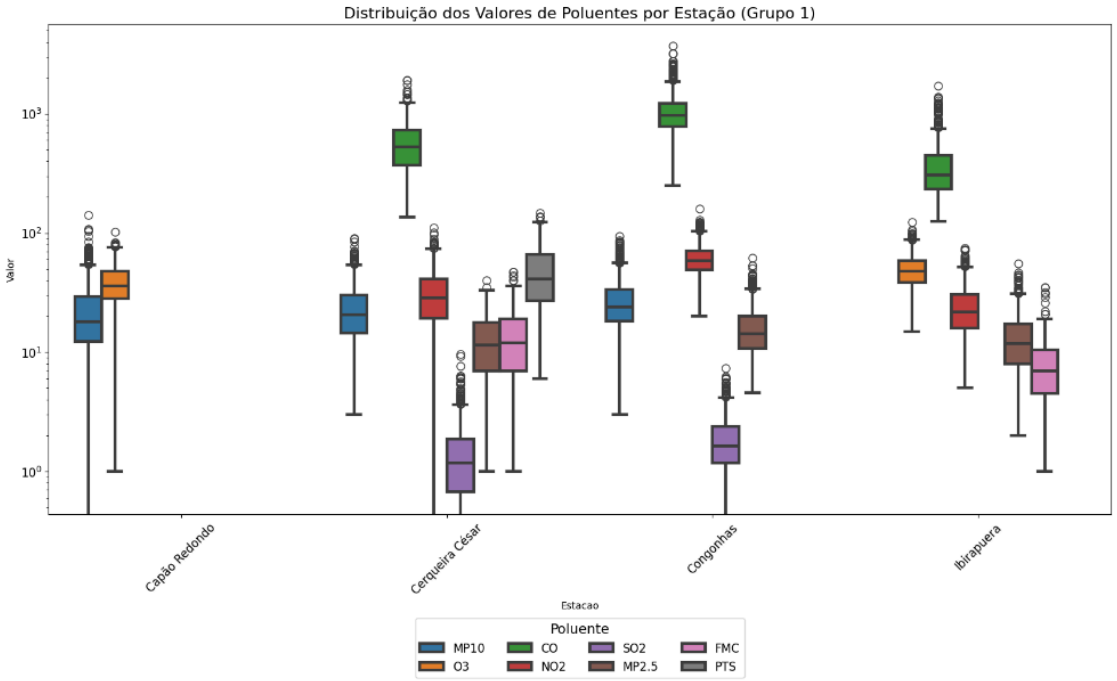


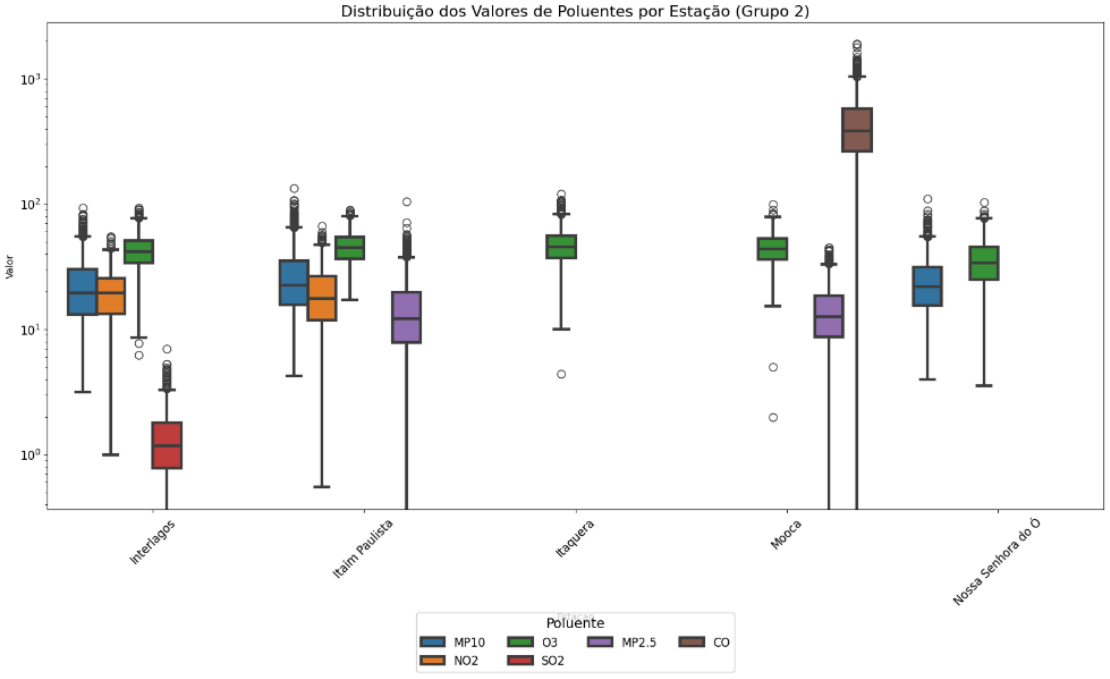
Na tabela, está representada a frequência de poluentes por estação.

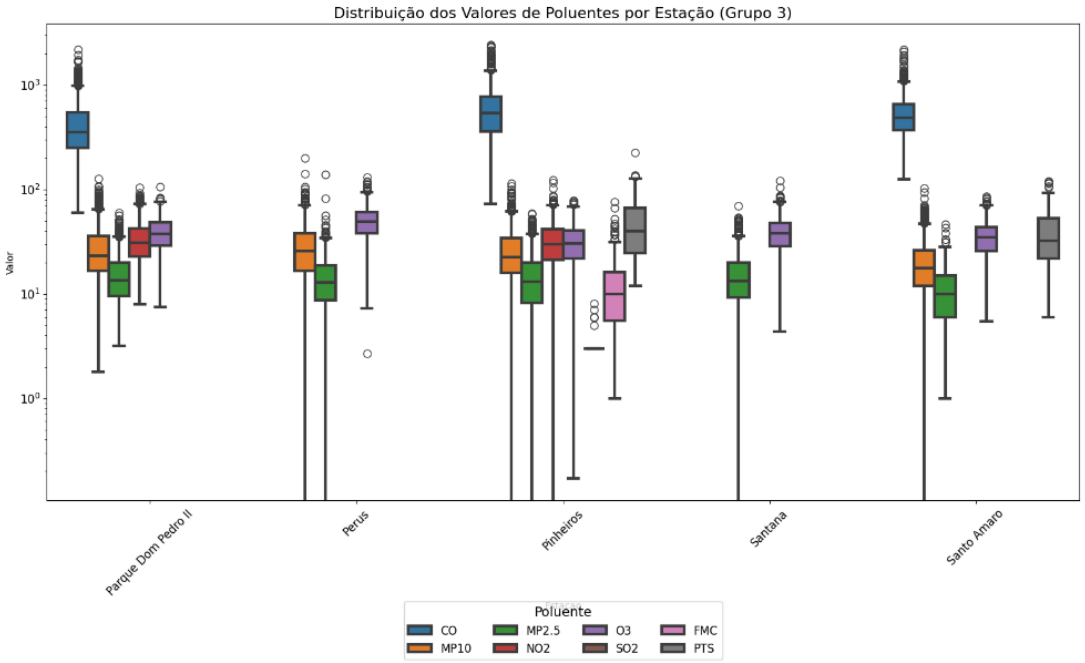


Distribuição dos valores dos poluentes em diferentes estações

Devido ao número de Estações de medição, elas serão divididas em 3 grupos e será aplicada a Escala logarítmica no eixo Y para melhorar a visualização dos dados. Solução escolhida devido a estarmos trabalhando com muitas estações e cada uma possuir valores diferentes em datas iguais. A análise dos Boxplot mostram que existem grandes variações entre os valores medidos dos poluentes nas estações de medição da cidade.


Distribuição dos Valores de Poluentes por Estação (Grupo 1)
Valor
Capão Redondo
Cerqueira César
Congonhas
Ibirapuera
Estacao
Poluente
MP10
O3
CO
NO2
SO2
MP2.5
FMC
PTS


Distribuição dos Valores de Poluentes por Estação (Grupo 2)
Valor
Interlagos
Itaim Paulista
Itaquera
Mooca
Nossa Senhora do Ó
Poluente
MP10
NO2
O3
SO2
MP2.5
CO



Decomposição da Série para verificar tendências, sazonalidade e resíduo

Autocorrelação e autocorrelação parcial (ACF e PACF)

Autocorrelação (ACF) e autocorrelação parcial (PACF) são úteis para verificar se há correlação entre os valores da série temporal em diferentes intervalos de tempo.

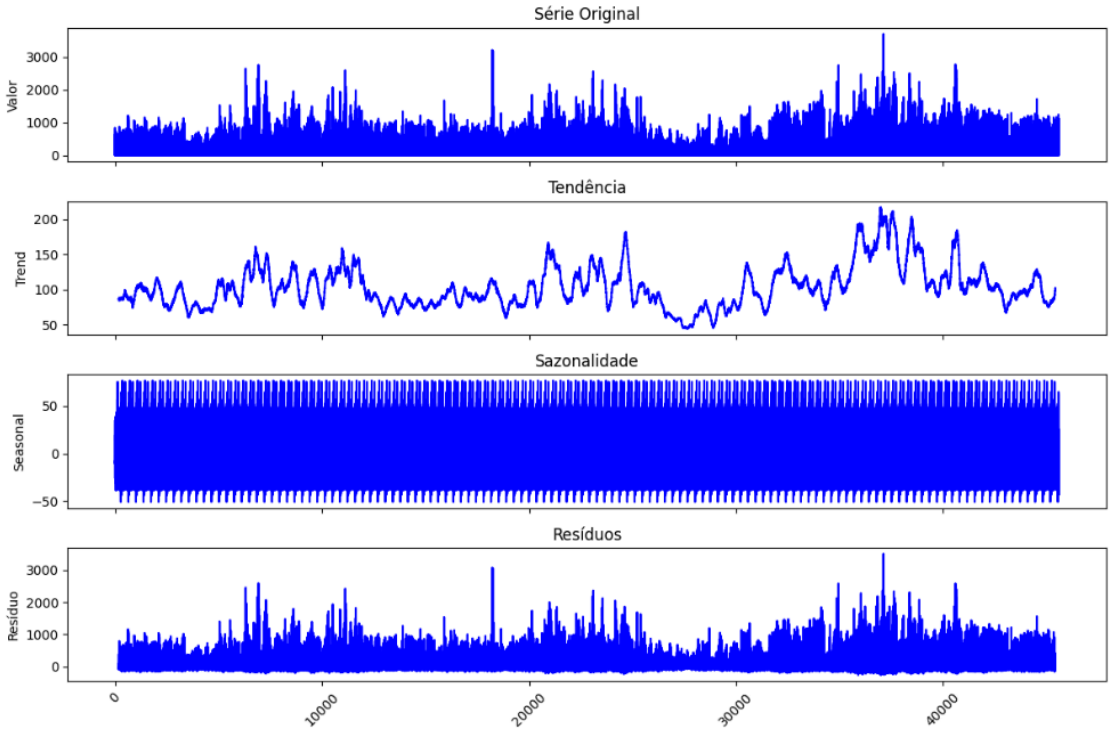
Teste de estacionariedade

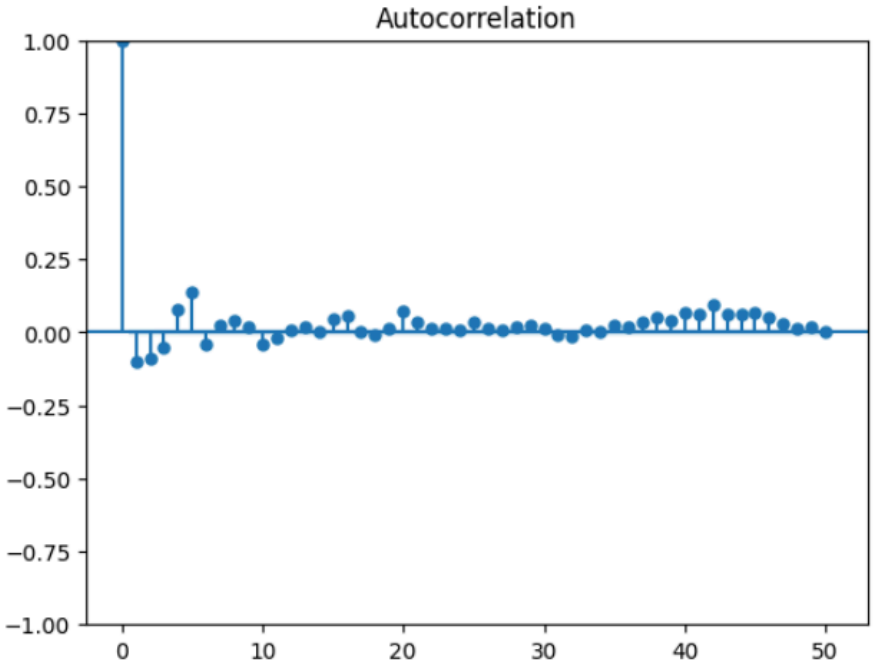
O teste de Dickey-Fuller aumentado ajuda a determinar se os dados podem ser estacionários - (a média e variância não mudam com o tempo), com base na presença de uma raiz unitária .

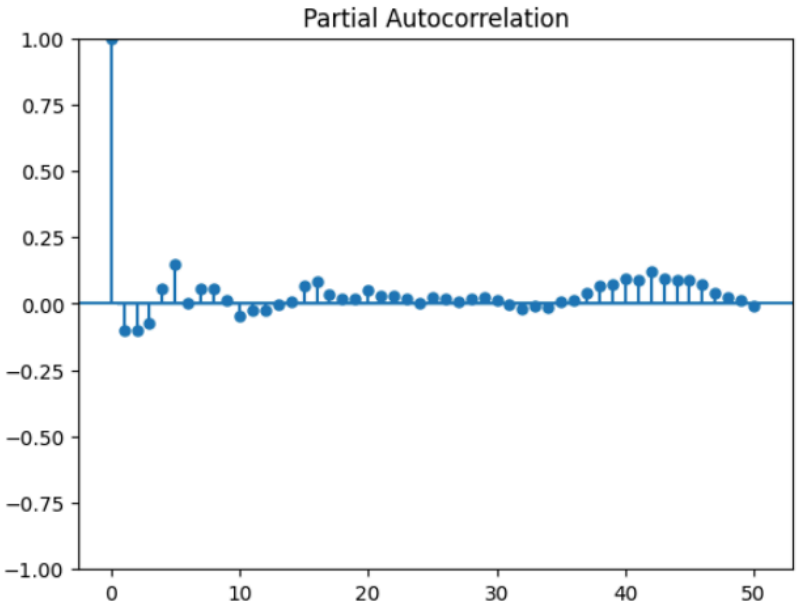
Se p-value for menor que 0.05, os dados são estacionários. ADF Statistic quanto mais negativo, maior a probabilidade de ser estacionário.

Esses gráficos de decomposição abaixo dividem a série em quatro componentes:

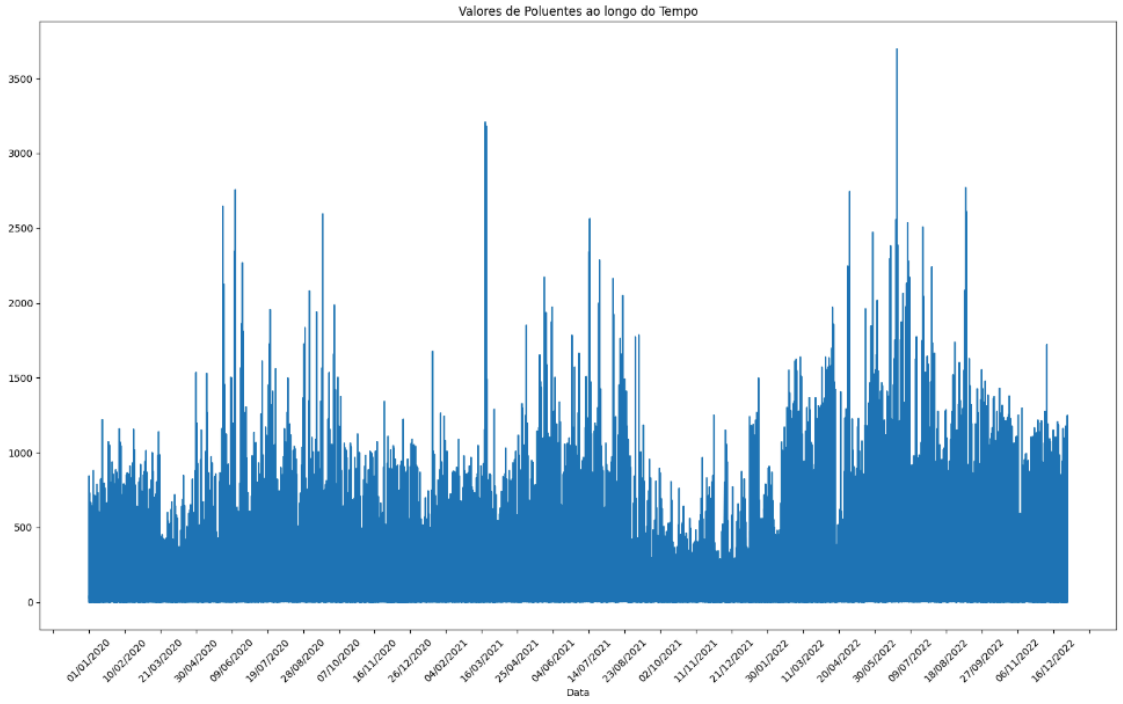
1. Observado (CO): A série original de CO.
2. Tendência (Trend): A componente de tendência ao longo do tempo.
3. Sazonalidade (Seasonal): Componentes que se repetem em intervalos regulares.
4. Resíduos (Resid): A parte da série que não foi explicada pela tendência ou sazonalidade, representando o ruído.


Série Original
Valor
3000
2000
1000
0
Tendência
Trend
200
150
100
50
Sazonalidade
Seasonal
50
0
−50
Resíduos
Resíduo
3000
2000
1000
0
0
10000
20000
30000
40000


Autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
10
20
30
40
50



Aplicando Série temporal de "Valor" ao longo do tempo, obtemos o gráfico com os valores de poluentes ao longo do tempo.



Conclusão Parcial

Devido à baixa performance dos modelos preditivos com a variável 'valor', realizaremos um processo de seleção de features para identificar o poluente mais relevante na modelagem de séries temporais.

Correção dos dados

Separando a coluna 'Poluente' em colunas separadas sem as estações

Poluente	CO	FMC	MP10	MP2.5	NO2	O3	PTS	\
Data								
01/01/2020	561.921667	NaN	33.600000	27.106667	18.621667	52.610000	NaN	
01/01/2021	468.693333	NaN	15.737778	13.072000	14.167500	40.277778	NaN	
01/01/2022	716.612000	NaN	26.977778	23.830000	15.812000	26.148000	NaN	
01/01/2023	962.482857	NaN	63.000000	46.166667	28.833333	19.583333	NaN	
01/02/2020	603.173333	NaN	12.059000	8.270000	24.651429	31.504167	NaN	

Poluente	SO2
Data	
01/01/2020	1.535000
01/01/2021	1.523333
01/01/2022	1.707500
01/01/2023	0.000000
01/02/2020	2.265000

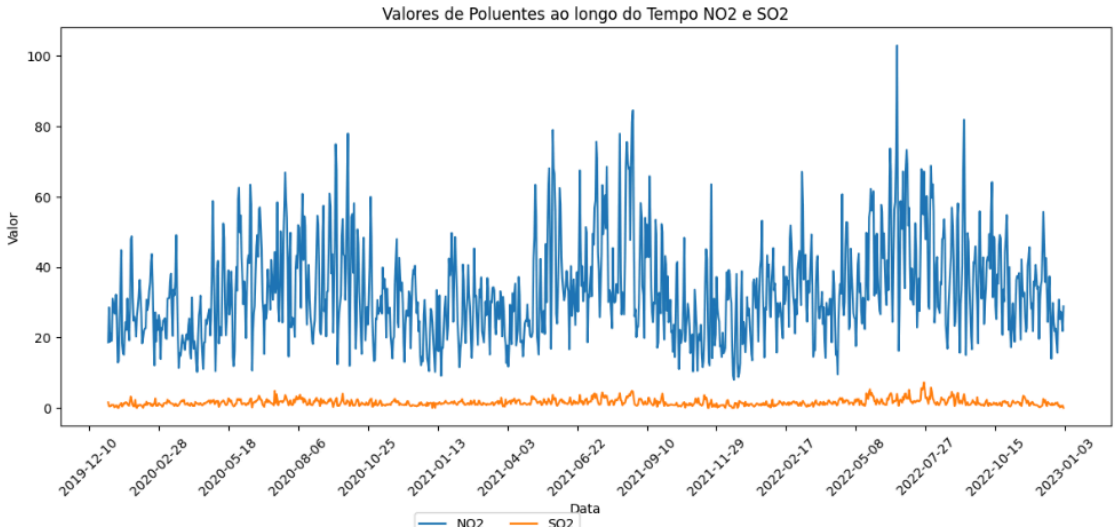
Poluente	
CO	0
FMC	931
MP10	0
MP2.5	0
NO2	0
O3	0
PTS	984
SO2	6

Devido a termos muitos valores nulos nos poluentes PTS e FMC, decidimos remover estas colunas

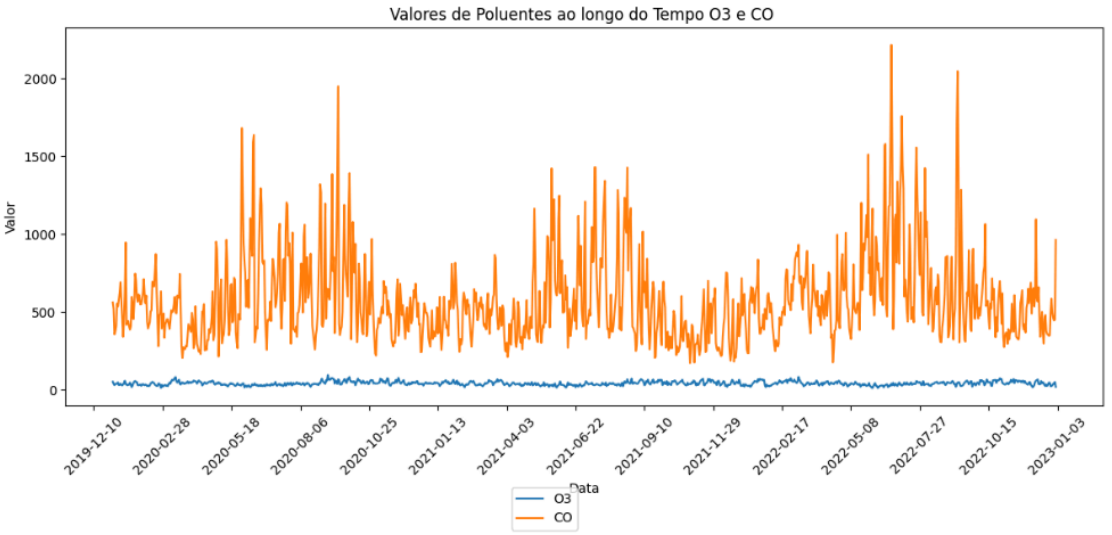
Na tabela abaixo temos a representação dos dados onde foram corrigidos as datas e passando para o formato datetime.

Poluente	CO	MP10	MP2.5	NO2	O3	SO2
Data						
2020-01-01	561.92	33.60	27.11	18.62	52.61	1.54
2020-01-02	520.18	14.62	8.06	28.57	39.01	0.46
2020-01-03	356.97	7.76	4.75	18.76	30.54	0.44
2020-01-04	377.12	10.55	6.54	21.00	37.86	0.54
2020-01-05	411.63	12.05	8.34	19.08	39.51	0.87

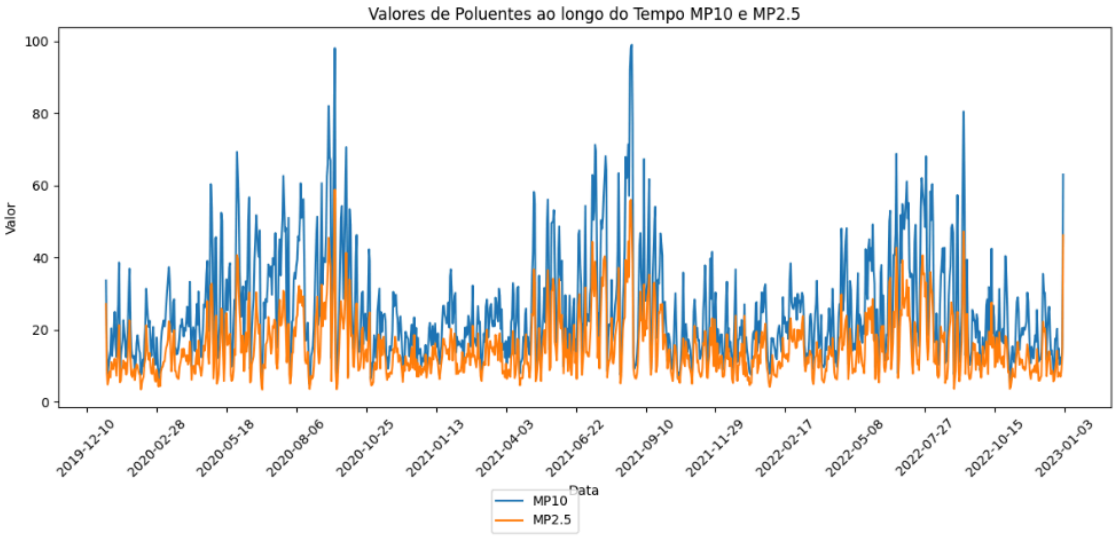
Separando apenas os Poluentes 'NO2','SO2'



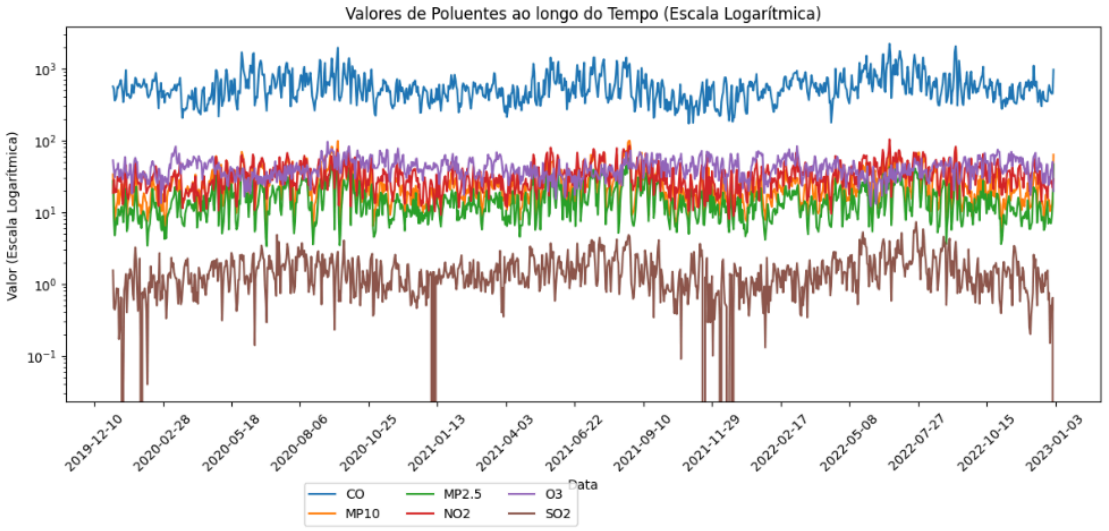
Separando apenas os Poluentes 'O3','CO'



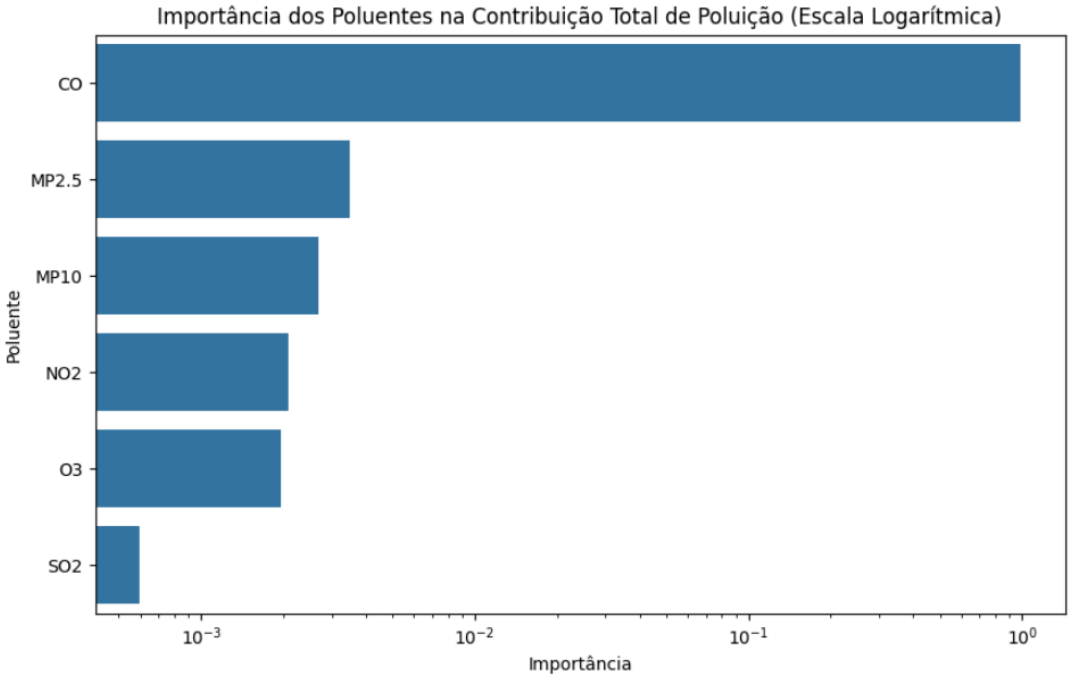
Separando apenas os Poluentes 'MP10', 'MP2.5'



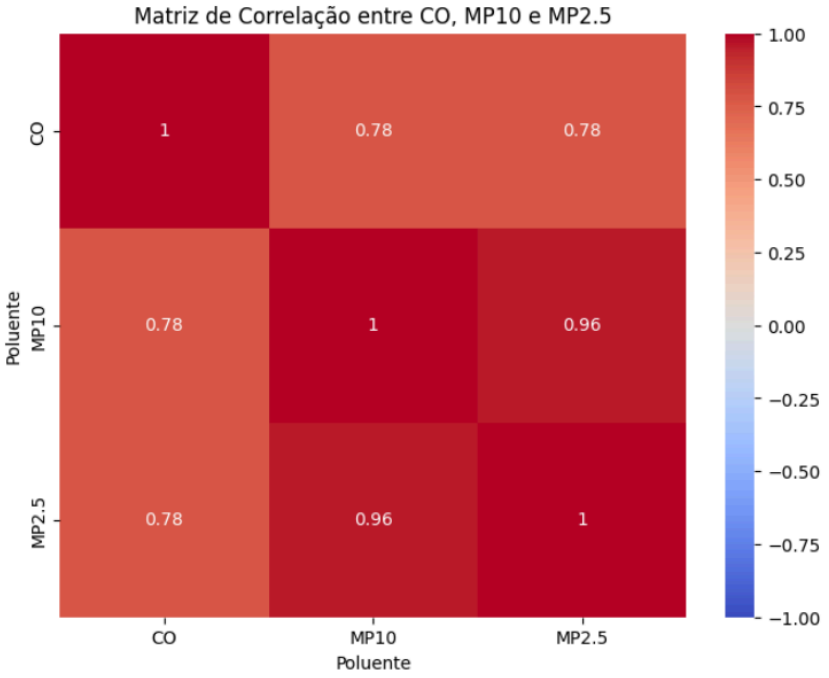
No gráfico abaixo plotamos todos os poluentes definindo a escala logarítmica no eixo Y



No próximo gráfico usamos Random Forest para determinar a importância dos poluentes, onde foi necessário a criação de uma variável de destino (soma dos poluentes-Variável alvo) e treinar o algoritmo para gerar o gráfico em escala logarítmica no eixo x. Este processo nos ajudou a identificar a variável com mais ‘valor’ e fundamentou a decisão de realizar a análise do componente CO.

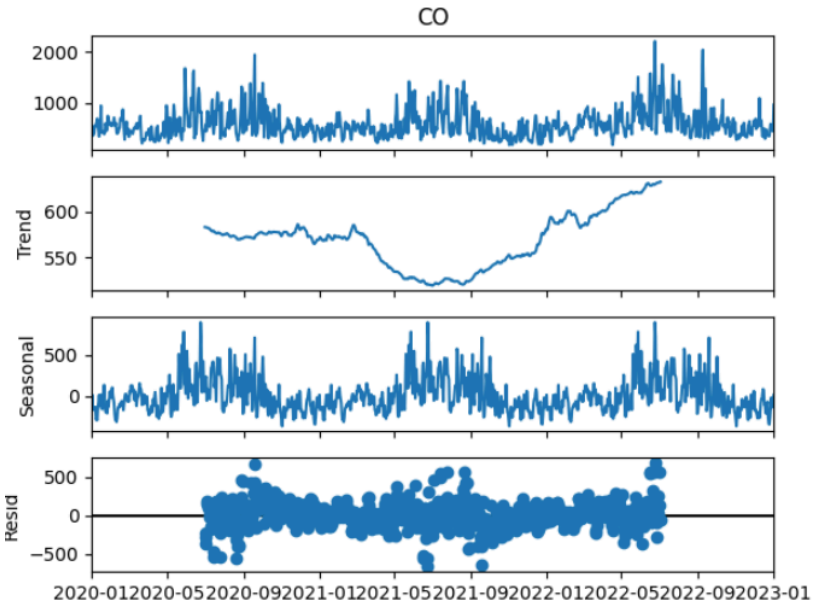


No gráfico abaixo, representamos a matriz de Correlação (CO, MP10, MP2.5)

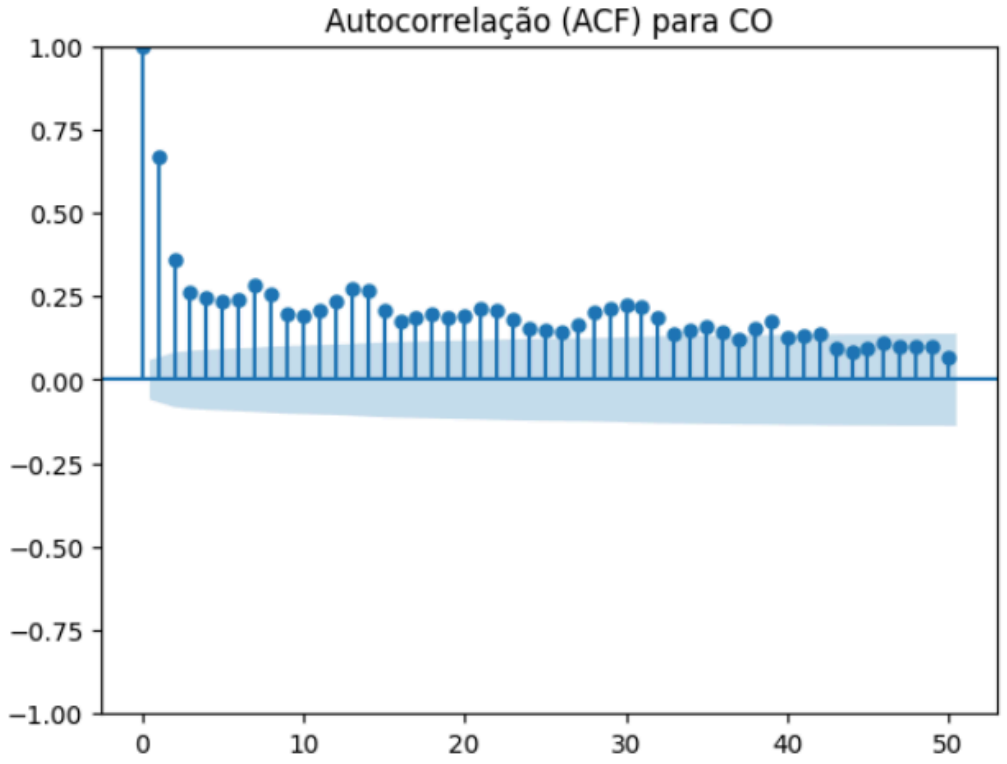


Teste ADF para cada poluente (CO, MP10, MP2.5) onde foi realizado a média e variância móvel dos Poluentes ao longo do Tempo, gerando os gráficos abaixo.

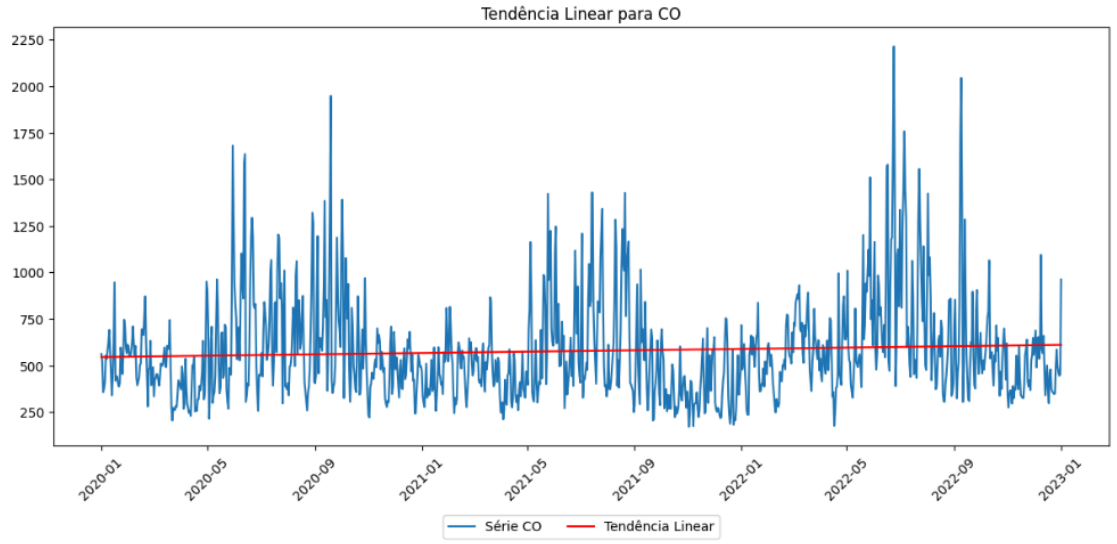
Verificação de Sazonalidade realizando a decomposição da série temporal (aditiva ou multiplicativa)



O gráfico mostra a função de Autocorrelação (ACF), identificando a repetição de padrões, indicando sazonalidade.



Para o próximo gráfico realizamos a verificação de linearidade, realizamos a configuração a regressão, realizamos os ajustes para a regressão linear, onde foi plotado a série e a tendência linear para CO.



Com base nos resultados da decomposição aditiva, da ACF e da tendência linear, a escolha de realizar a análise temporal no CO se justifica devido à sua alta variabilidade e sazonalidade clara.

O CO é um poluente primário e de grande importância para a saúde pública, e seus níveis podem ser fortemente influenciados por fatores sazonais, como o aumento da circulação de veículos durante certos períodos do ano. A análise temporal pode ajudar a prever esses aumentos e permitir ações preventivas.

Danos à Saúde

Poluente	Problemas de Saúde	Limite Máximo (OMS) - 24h	Limite Perigoso (OMS)
CO (Monóxido de Carbono)	Dificuldade respiratória, tontura, fadiga, náusea, danos ao sistema nervoso central. Exposição prolongada pode ser fatal.	**10 mg/m³ (média de 8h)**	> **30 mg/m³**: Perigoso, risco de morte com exposição prolongada.
MP10 (Material Particulado ≤ 10 µm)	Irritação nos olhos, nariz e garganta, problemas respiratórios, agravamento de doenças cardíacas e pulmonares.	**45 µg/m³**	> **150 µg/m³**: Aumento de mortalidade em populações sensíveis.
MP2.5 (Material Particulado ≤ 2.5 µm)	Penetra profundamente nos pulmões, causando problemas respiratórios graves, câncer de pulmão, doenças cardiovasculares.	**15 µg/m³**	> **50 µg/m³**: Danos irreparáveis aos pulmões e risco de morte prematura.
NO2 (Dióxido de Nitrogênio)	Agravamento da asma, inflamação pulmonar, infecções respiratórias, danos crônicos ao sistema respiratório.	**25 µg/m³**	> **200 µg/m³**: Risco de danos pulmonares graves e crônicos.
O3 (Ozônio Troposférico)	Irritação nos olhos, tosse, falta de ar, agravamento de doenças respiratórias, diminuição da função pulmonar.	**100 µg/m³ (média de 8h)**	> **200 µg/m³**: Perigoso, pode causar danos irreparáveis ao tecido pulmonar.
SO2 (Dióxido de Enxofre)	Irritação no sistema respiratório, aumento de sintomas em asmáticos, redução da função pulmonar, aumento do risco de doenças cardíacas.	**40 µg/m³**	> **500 µg/m³**: Altamente perigoso, danos irreparáveis ao sistema respiratório.

Modelo base

Aplicação de, pelo menos, um primeiro modelo (modelo de aprendizado de máquina ou modelo estatístico) que deve ser refinado até a entrega final do projeto. Apresente o método aplicado e a análise dos resultados obtidos.

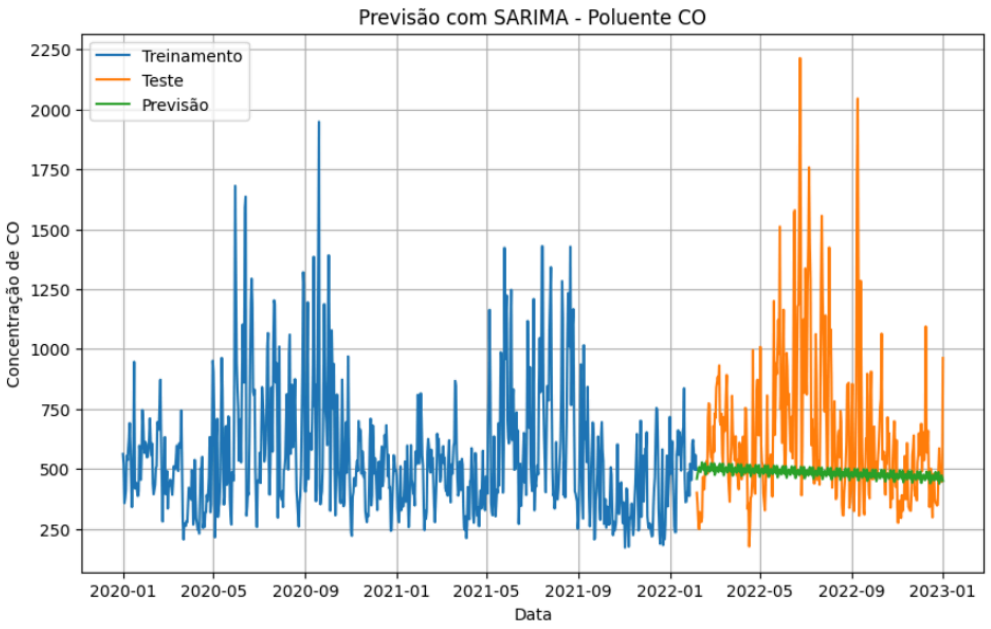
Aplicação dos Algoritmos de Série Temporal

O objetivo desta abordagem é comparar os resultados de previsões feitas apenas com o poluente CO com os resultados obtidos quando usamos outros dois poluentes, MP10 e MP2.5, como variáveis extras. A ideia é descobrir se incluir mais poluentes ajuda a entender melhor a poluição e se isso melhora a precisão das previsões.

Modelo SARIMA

O SARIMA (Seasonal AutoRegressive Integrated Moving Average) é indicado para modelar séries temporais univariadas que apresentam componentes sazonais, como no caso do CO, onde já foi detectada uma sazonalidade e padrões de tendência. Detalhes do SARIMA - Flexibilidade: Pode modelar uma ampla variedade de séries temporais; Interpretabilidade: Os parâmetros do modelo têm interpretações claras e Precisão: Em muitos casos, o modelo produz previsões precisas. Porém ele possui algumas limitações, como assumir que a série temporal é estacionária após as diferenças e que os erros são independentes e identicamente distribuídos (i.i.d.) e complexidade já que a escolha dos parâmetros pode ser desafiadora, especialmente para séries temporais complexas.

Aplicando o modelo SARIMA



Com a aplicação do modelo Sarima, foram identificados os valores:

Valor máximo de CO: 2213.63

Valor mínimo de CO: 171.7

Amplitude dos valores de CO: 2041.93

RMSE: 346.433063820879

Onde o RMSE é moderado em relação à amplitude dos dados.

RMSE (Root Mean Squared Error)

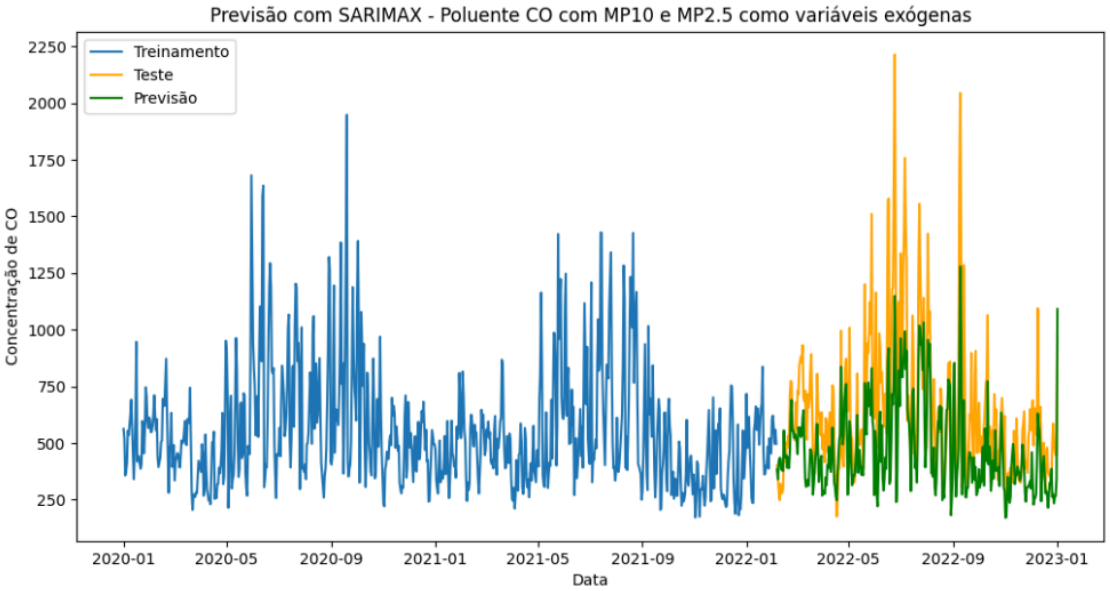
O RMSE é uma métrica que mede o desempenho do modelo em termos de erro de previsão. Ele calcula a raiz quadrada da média dos erros quadrados entre os valores previstos e os valores observados no conjunto de teste.

O RMSE nos diz o quanto o modelo está errando em suas previsões reais, o que é uma métrica diretamente relacionada à qualidade das previsões do modelo. Ele é útil quando estamos comparando o desempenho de um modelo em dados não vistos (dados de teste).Ele é interpretado na mesma escala dos dados originais, facilitando a compreensão do erro.

Modelo SARIMAX

O SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors) expande essa funcionalidade ao incluir variáveis exógenas.

SARIMAX é uma extensão do modelo SARIMA que incorpora variáveis exógenas à análise. Essas variáveis exógenas são fatores externos que podem influenciar a série temporal que estamos modelando, mas que não fazem parte diretamente da série em si. Utilizando o poluente CO como a variável dependente (target) e os poluentes MP10 e MP2.5 como variáveis exógenas (independentes) com o objetivo de melhorar a precisão das previsões, especialmente quando há uma correlação significativa entre eles, como foi identificado anteriormente.



AIC (Akaike Information Criterion)

O AIC é uma métrica usada para selecionar o melhor modelo entre várias opções. Ele equilibra a **bondade de ajuste** e a **complexidade** do modelo. O AIC penaliza a inclusão de mais parâmetros no modelo, evitando o overfitting (quando o modelo se ajusta muito bem aos dados de treinamento, mas tem um desempenho ruim em novos dados). O AIC não mede diretamente a qualidade das previsões futuras, mas sim a qualidade do ajuste ao conjunto de treinamento.

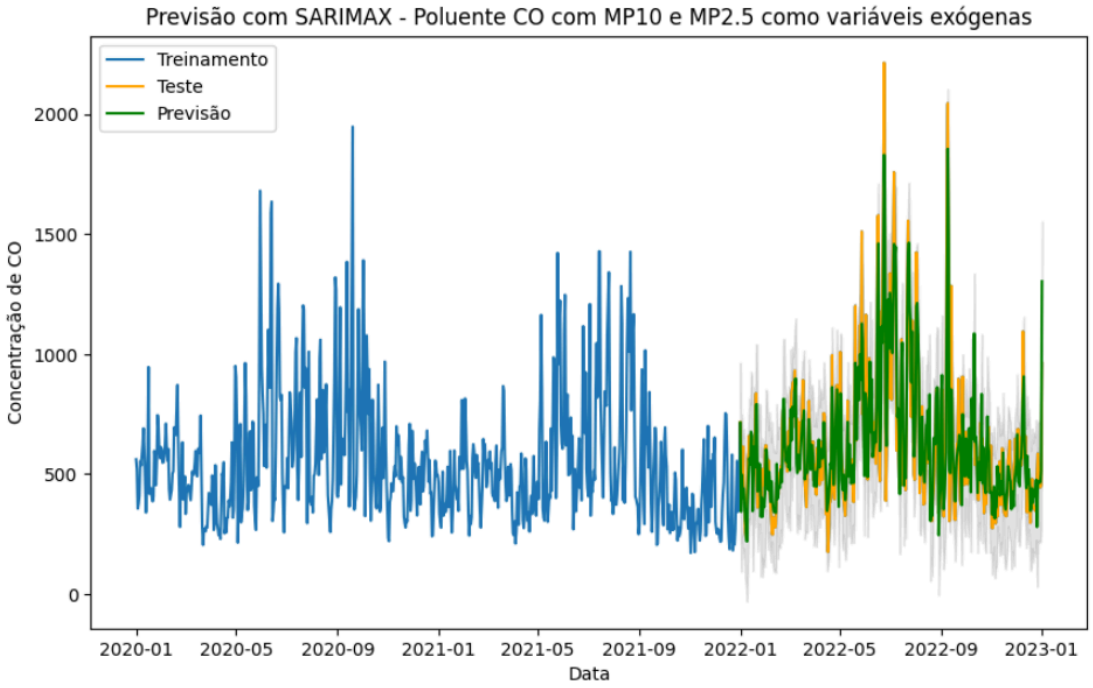
Percentual de erro médio absoluto (MAPE)

Uma métrica comum para séries temporais, expressa o erro de previsão como uma porcentagem dos valores reais. Isso é mais intuitivo para a maioria das pessoas, pois dá uma ideia do erro em termos de porcentagem.

R^2 (Coeficiente de Determinação)

O R^2 (R-quadrado), que mede a proporção da variabilidade explicada pelo modelo. Ele varia de 0 a 1 (ou pode ser negativo para modelos muito ruins). Um valor mais próximo de 1 significa que o modelo explica bem a variabilidade dos dados.

Após ajuste de parâmetros, temos o seguinte gráfico:



O ajustes realizados foram:

enforce_stationarity=False

A estacionariedade é uma propriedade importante que significa que as propriedades estatísticas da série temporal, como a média e a variância, permanecem constantes ao longo do tempo. Quando você define enforce_stationarity=False, você está dizendo ao modelo que ele não precisa necessariamente garantir que a série temporal seja estacionária antes de ser modelada. Isso pode ser útil em algumas situações, como quando a série temporal apresenta tendências ou sazonalidades fortes que são difíceis de remover completamente.

enforce_invertibility=False

A invertibilidade é uma condição técnica em modelos SARIMAX que garante a estabilidade e previsibilidade do modelo. Significa que o impacto de choques passados nos valores futuros diminui gradualmente ao longo do tempo. Ao definir enforce_invertibility=False, você permite que o modelo ajuste dados que não se encaixam perfeitamente na condição de invertibilidade. Isso pode ser útil em séries temporais complexas que apresentam padrões não-usuais.

Resultados do modelo SARIMAX:

AIC: 13452.55

RMSE: 135.54

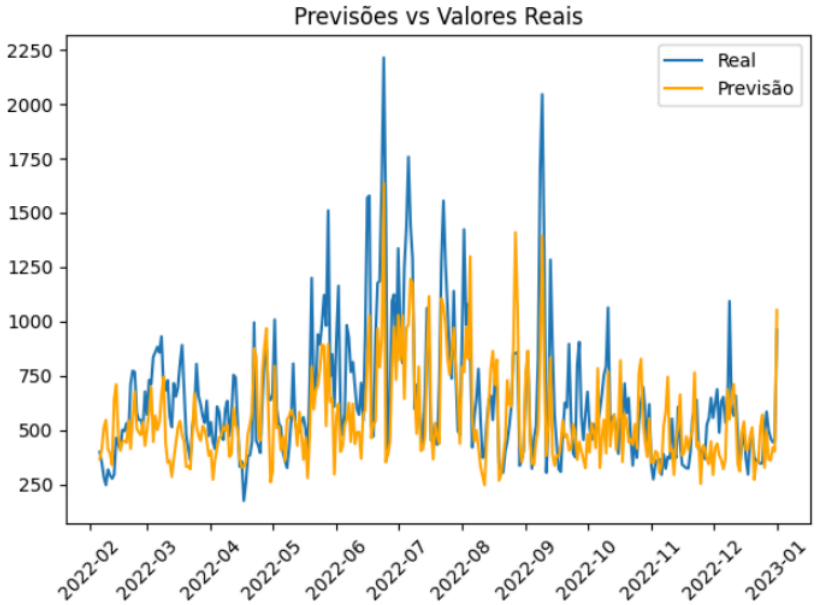
MAPE: 17.47%

R^2: 0.79

Modelo XGBoost

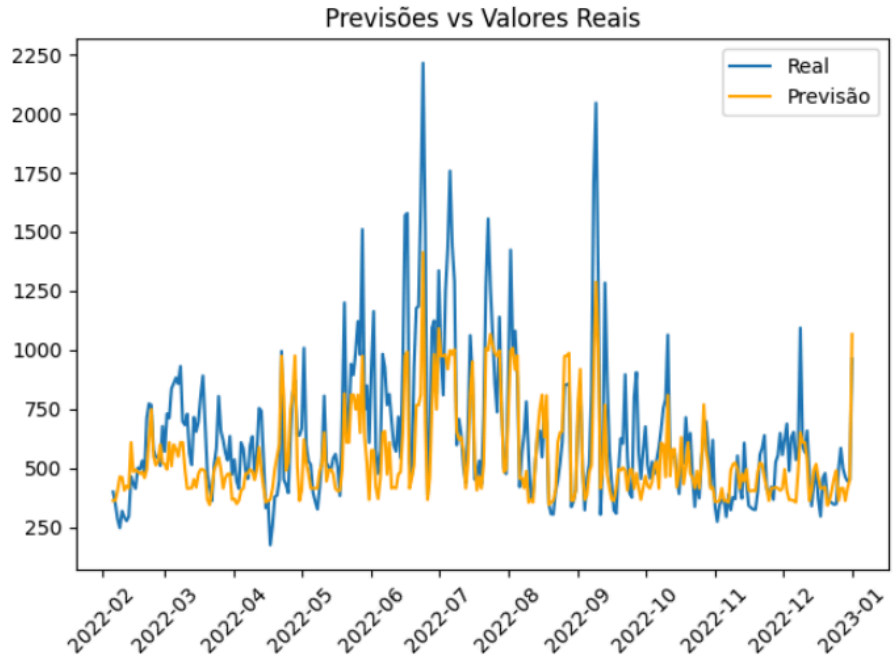
O XGBoost é um algoritmo de aprendizado de máquina, mais especificamente, algoritmo de aumento de gradiente (gradient boosting). Amplamente utilizados para previsão de séries temporais, especialmente quando você tem várias variáveis (como no seu caso, com MP10, MP2.5 e CO).Ele funciona muito bem com dados tabulares e podem capturar interações complexas entre variáveis exógenas. Utiliza regularização L1 e L2 para evitar overfitting (superajuste), o que ajuda a melhorar a generalização do modelo.Emprega uma estratégia de divisão das árvores de decisão que leva em consideração a ganho de informação e a complexidade da árvore. Permite a execução paralela, o que acelera o treinamento em grandes conjuntos de dados.

Devido ao resultado do MAPE, avaliando que os dados são não lineares vamos aplicar o algoritmo XGBoost para analisar se ele se ajusta melhor aos dados, trazendo resultados mais robustos, gerando o gráfico `Previsões vs Valores Reais'



Na avaliação do xgboost, houve o ajuste de parâmetros, onde foram preparados os dados parâmetros simplificados, instanciando o modelo, e foi realizado a busca pelos

melhores parâmetros, onde foi realizado o treinamento do modelo com os melhores parâmetros, previsões. Gerando o gráfico ‘Previsões vs Valores Reais’



Após a realização de uma busco pelos melhores resultados, os ajustes realizados foram:

'learning_rate': 0.01

Controla o impacto de cada árvore no modelo final. Taxas menores podem exigir mais árvores.

'max_depth': 3

A profundidade máxima da árvore. Valores maiores permitem maior complexidade, mas aumentam o risco de overfitting.

'n_estimators': 500}

Número de árvores a serem criadas. Mais árvores podem resultar em modelos mais robustos, mas também mais propensos a overfitting.

Overfitting é quando um modelo se ajusta tão bem aos dados de treinamento que acaba "decorando" os exemplos em vez de aprender os padrões gerais subjacentes.

Para determinar o modelo mais adequado para prever o nível de monóxido de carbono (CO) na cidade de São Paulo, foram aplicados os algoritmos SARIMAX e XGBoost, ambos amplamente utilizados para análise de séries temporais e previsões. Inicialmente, a variabilidade dos dados foi avaliada, com o valor máximo de CO em 2213.63 e o mínimo em 171.7, resultando em uma amplitude de 2041.93. Essa ampla variação demonstra uma grande flutuação nos níveis de CO, o que representa um desafio

para o desenvolvimento de modelos de previsão. Em contextos como esse, modelos como o SARIMAX, que são capazes de capturar tanto dependências temporais quanto padrões sazonais, tendem a apresentar resultados superiores.

O SARIMAX, uma extensão do modelo ARIMA, foi aplicado com o objetivo de capturar as dependências temporais nos dados, levando em consideração os efeitos sazonais e as variáveis exógenas. As métricas de desempenho para o SARIMAX incluem um AIC de 13452.55, RMSE de 135.54, MAPE de 17.47% e R^2 de 0.79.

Estes resultados sugerem um bom ajuste, especialmente considerando o MAPE abaixo de 20%, que indica um erro percentual médio dentro de um intervalo aceitável para dados de poluição atmosférica. O R^2 de 0.79 destaca a capacidade do modelo em explicar grande parte da variabilidade nos níveis de CO, confirmando que o SARIMAX é eficaz para essa aplicação.

O modelo XGBoost, embora eficiente em diversas aplicações de aprendizado de máquina, foi incluído para comparação, apesar de não ser específico para séries temporais. As métricas para o XGBoost incluem um RMSE de 210.44, MAE de 148.89 e R^2 de 0.53.

Em comparação com o SARIMAX, o XGBoost apresentou um RMSE consideravelmente mais alto, indicando maior dificuldade em seguir as flutuações do CO. O MAE e o valor relativamente baixo de R^2 reforçam a limitação do XGBoost na captura dos padrões temporais complexos e das sazonalidades específicas dos dados de poluição atmosférica.

Com base nos resultados observados, o modelo SARIMAX destaca-se como a melhor opção para a previsão de CO na cidade de São Paulo. A combinação de um RMSE mais baixo, MAPE dentro de um intervalo de precisão aceitável e um R^2 alto comprova sua capacidade de adaptação às características específicas dos dados temporais e sazonais analisados. Assim, o SARIMAX é considerado o modelo mais confiável para previsões de curto e médio prazo dos níveis de CO, proporcionando um suporte robusto para decisões fundamentadas no monitoramento da poluição ambiental.

5. Conclusão

Este trabalho abordou a previsão de picos de poluição do ar na cidade de São Paulo, com foco no poluente monóxido de carbono (CO), utilizando dados históricos de qualidade do ar e técnicas de séries temporais. A pesquisa partiu do problema de como

antecipar os níveis de poluição de maneira eficaz, permitindo a adoção de medidas preventivas para minimizar os impactos na saúde pública e na qualidade de vida. Com base nesse objetivo, aplicaram-se os modelos SARIMAX e XGBoost, combinando variáveis exógenas para capturar a complexidade e a sazonalidade dos dados.

Os resultados demonstraram que o modelo SARIMAX obteve um melhor desempenho na previsão dos níveis de CO, com métricas como RMSE de 135,54, MAPE de 17,47% e R^2 de 0,79, indicando uma capacidade robusta de modelagem das características temporais e sazonais dos dados. Por outro lado, embora o XGBoost tenha mostrado potencial, especialmente para dados não lineares, seu desempenho inferior (com RMSE de 210,44 e R^2 de 0,53) destacou as limitações do modelo para capturar as especificidades da série temporal em questão.

As contribuições deste estudo incluem a implementação de uma abordagem preditiva robusta que pode ser utilizada por agências ambientais e autoridades locais para mitigar os impactos da poluição. No entanto, as limitações incluem a dependência de dados históricos específicos e a exclusão de outros fatores externos, como variáveis meteorológicas, que poderiam refinar os modelos.

Para trabalhos futuros, sugere-se a inclusão de variáveis climáticas, como temperatura e umidade, para aumentar a precisão dos modelos. Além disso, explorar modelos híbridos, como combinações de SARIMAX com algoritmos de aprendizado de máquina, pode oferecer melhores previsões em cenários complexos. O desenvolvimento de um sistema integrado de alerta em tempo real também se apresenta como um próximo passo essencial para a aplicação prática das previsões em benefício da saúde pública.

6. Referências bibliográficas

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. Time series analysis: forecasting and control. 5th ed. Hoboken: Wiley, 2015.

ENERGIA E AMBIENTE. Brasil emitiu 2,3 bilhões de toneladas brutas de gases de efeito estufa em 2022. Disponível em: https://energiaeambiente.org.br/cop-28-brasil-emitiu-23-bilhoes-de-toneladas-brutas-de-gases-de-efeito-estufa-em-2022-20231209#:~:text=O painel apresentou as estimativas,de efeito estufa em 2022. Acesso em: 02 set. 2024.

ENERGIA E AMBIENTE. Qualidade do ar. Disponível em: https://energiaeambiente.org.br/qualidadedoar/. Acesso em: 02 set. 2024.

MONTE, E. Z. et al. Previsão da concentração de material particulado inalável, na Região da Grande Vitória, ES, Brasil, utilizando o modelo SARIMAX. Engenharia Sanitária e Ambiental, v. 23, n. 2, p. 297-310, mar./abr. 2018. Disponível em: https://www.scielo.br/j/esa/a/w6dyvXwPdG69QVV9nn9CRpt/?lang=pt. Acesso em: 30 set. 2024.

PAGANO, Elena; BARBIERATO, Enrico. Uma abordagem de séries temporais para a transformação de cidades inteligentes: o problema da poluição do ar em Brescia. IA, Basel, v. 5, n. 1, p. 17, 2024. DOI: 10.3390/ai5010002. Disponível em: https://www.proquest.com/docview/2987117632/FBAA46825BA946DCPQ/1?accountid=12217&sourcetype=Scholarly Journals. Acesso em: 26 set. 2024.

PLANETA. Estudo associa poluição do ar a 135 milhões de mortes prematuras entre 1980 e 2020. Disponível em: https://revistaplaneta.com.br/estudo-associa-poluicao-do-ar-a-135-milhoes-de-mortes-prematuras-entre-1980-e-2020/. Acesso em: 01 set. 2024.

Poluição do ar causou 8,1 milhões de mortes em 2021. ONU News. Disponível em: https://news.un.org/pt/story/2024/06/1833321#:~:text=Brasil%20teve%20mais%20de%2010,todo%20o%20mundo%20em%202021. Acesso em: 28 set. 2024.

POPE III, C. A.; DOCKERY, D. W.; SCHWARTZ, J. Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association, v. 59, n. 6, p. 709-742, 2019.

ROSSI, M.; BIANCHI, F.; FERRARI, G. Uma abordagem de séries temporais para a transformação de cidades inteligentes: o problema da poluição do ar em Brescia. Journal of Environmental Management, v. 234, p. 123-134, 2019.

SANTOS, J. A.; SILVA, M. A.; OLIVEIRA, R. B. Impacto da poluição atmosférica nas grandes metrópoles. Revista Brasileira de Ciências Ambientais, v. 49, p. 123-134, 2018.

SILVA, A. P.; COSTA, L. M.; PEREIRA, F. J. Previsão de concentração de material particulado inalável na Região da Grande Vitória, ES, Brasil. Revista Ciello Brasil, v. 12, n. 3, p. 45-58, 2020.

TIBULO, Cleiton; FERRAZ, Simone Erotildes Teleginski; TIBULO, Vaneza De Carli; ZANINI, Roselaine Ruviaro; BOIASKI, Nathalie Tissot. Previsão da concentração de material particulado inalável, através de modelos estatísticos de séries temporais para o município de Canoas, Rio Grande do Sul. Revista Thema, v. 19, n. 1, p. 134-152, 2021. Disponível em: https://periodicos.ifsul.edu.br/index.php/thema/article/view/1660/1761.

WORLD HEALTH ORGANIZATION. Air quality guidelines: global update 2021. Geneva: WHO, 2021.

ZHANG, Y.; LI, Y.; WANG, J. Air pollution prediction using machine learning algorithms: a review. Environmental Science and Pollution Research, v. 25, n. 1, p. 1-14, 2018.