Sobre o Problema de Previsão de Qualidade de Ar com Modelos de Aprendizagem de Máquina

[1]Eliseu C. de Brito, [2]Ya-Sin B. Mghazli, [2]José M. de Seixas e [1]Ricardo de A. Araújo
[1]Laboratório de Inteligência Computacional do Araripe, Instituto Federal do Sertão Pernambucano, Ouricuri, PE, Brasil.
[2]Laboratorio de Processamento de Sinais, COPPE/POLI, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil.
eliseu.cordeiro@aluno.ifsertao-pe.edu.br, yasin.barcelos@lps.ufrj.br, seixas@lps.ufrj.br, ricardo.araujo@ifsertao-pe.edu.br

***Resumo*—Neste trabalho é apresentado um estudo empírico sobre as características do fenômeno gerador de séries temporais relacionadas a qualidade do ar. Baseado neste estudo, serão investigados modelos baseados em aprendizagem de máquina para solucionar este tipo particular de problema de previsão. Por fim, uma análise comparativa será realizada com os modelos investigados conduzindo-se experimentos com um conjunto relevante de problemas de previsão de qualidade do ar relacionado a concentração de poluentes atmosféricos coletados a partir de sensores da *United States Environmental Protection Agency*.**

***Index Terms*—Qualidade do Ar, Poluentes Atmosféricos, Séries Temporais, Previsão, Aprendizagem de Máquina.**

I. Introdução

No último século, é possível observar um aumento significativo na população urbana em todo o mundo, indicando que um número considerável de pessoas que viviam em áreas rurais estão se deslocando para os centros urbanos [1]. De acordo com o relatório do *The World Bank* (WB), aproximadamente 56% da população global reside em áreas urbanas [2].Segundo a *United Nations* (UN), estima-se que esse número aumente para 70% nos próximos 30 anos [3].

Com o avanço da urbanização, o acesso à tecnologia tem viabilizado o desenvolvimento e implementação de cidades inteligentes [4], onde um número expressivo de sensores fixos ou móveis têm sido implantados para coleta e armazenamento de uma quantidade significativa de dados sobre os mais variados aspectos da vida em grandes centros urbanos [5].

Segundo a *World Health Organization* (WHO), os dados coletados revelaram que 90% das pessoas em grandes centros urbanos respiram ar com diversos tipos de poluentes, excedendo os limites definidos em diretrizes internacionais em termos de qualidade do ar [6]. Tal fato tem desecadeado uma série de problemas de saúde que podem ter consequências graves de curto a longo prazos, levando a milhares de mortes todos os anos [7]–[9].

Neste contexto, a poluição do ar decorrente de atividades humanas, industrialização e urbanização tem se tornado um fator de risco de vida em muitos países ao redor do mundo [10]. Essa questão tem despertado um crescente interesse na sociedade devido ao seu impacto negativo no bem estar e na qualidade de vida da população [11], uma vez que os diversos agentes que tornam o ar prejudicial a saúde humana têm levado a poluição atmosférica ao seu ápice [12].

Além de ser uma ameaça a saúde pública global, a falta de qualidade do ar tem impacto direto e indireto na economia, uma vez que, em relatório publicado pela *Organization for Economic Cooperation and Development* (OECD) a diminuição da qualidade do ar devido a alta concentração de poluentes atmosféricos leva a gastos que podem chegar a aproximadamente 1% do produto interno bruto mundial [13].

A qualidade do ar depende de uma série de fatores, que podem ou não contribuir para dispersão dos poluentes, dentre os quais vale destacar as condições meteorológicas, a topografia e a magnitude das concentrações dos poluentes na região [14]–[16]. Entretanto, as concentrações de monóxido de carbono (CO), dióxido de nitrogênio (NO_2), material particulado (PM_{10} e $PM_{2.5}$) e dióxido de enxofre (SO_2) representam um impacto significativo na qualidade do ar [17]–[20].

Considerando que os efeitos e sintomas da exposição a altas concentrações de poluentes não podem ser facilmente detectáveis, diversos estudos epidemiológicos têm sido apresentados na literatura sugerindo uma relação entre a exposição a poluentes e doenças respiratórias e cardiovasculares [18], [21]–[24]. Neste contexto, a previsão da concentração de poluentes no ar representa um problema de fundamental importância para sociedade, uma vez que esta permite a prevenção da exposição a altas concentrações de poluentes do ar, minimizando os efeitos adversos na saúde humana [25], [26].

Modelos estatísticos clássicos, têm sido amplamente utilizados como solução para este tipo particular de problema de previsão. O modelo mais difundido neste sentido é o modelo *Autoregressive Integrated Moving Average* (ARIMA) [27]–[30]. No entanto, este tem uma desvantagem intrínseca à sua natureza puramente linear para estimar o fenômeno gerador das séries temporais relacionadas a poluentes atmosféricos, uma vez que estas possuem algum tipo de não-linearidade [31].

Neste contexto, diversos trabalhos podem ser encontrados na literatura investigando modelos não-lineares para previsão de fenômenos temporais relacionados a concentração de poluentes no ar [32]–[38]. No entanto, a não-linearidade presente

nestes modelos não é capaz, na prática, de apresentar um ganho expressivo, em termos de desempenho preditivo, quando comparado modelo ARIMA [39], o que sugere o desenvolvimento de um estudo sobre as características do fenômeno gerador deste tipo particular de série temporal, para escolha de um modelo adequado para prevê-las.

Desta forma, este trabalho apresenta um estudo empírico sobre o fenômeno gerador de problemas de previsão de qualidade do ar, na tentativa de demonstrar que este tipo particular de série temporal possui algum tipo de não-linearidade em seu fenômeno gerador e é passível de previsão. Baseado nesta análise, são investigados modelos baseados em aprendizagem de máquina para solucionar este tipo particular de problema de previsão.

Além disso, realizamos uma análise comparativa entre os modelo investigados e modelos estatísticos clássicos utilizando um conjunto relevante de séries temporais relacionadas a concentração de poluentes atmosféricos coletados a partir de sensores da *United States Environmental Protection Agency* (US EPA). Além disso, o *mean squared error* (MSE) e o *mean absolute percentage error* (MAPE) são usados para avaliar o desempenho da previsão, que é validado estatisticamente usando o teste de Friedman.

Organizamos este trabalho da seguinte forma. Na Seção II, são apresentados os fundamentos e um estudo empírico das séries temporais investigadas . Na Seção III, é relatado a definição do método proposto. Na Seção IV, são descritos os resultados experimentais. Por fim, na Seção V, são expostas algumas conclusões e trabalhos futuros promissores.

II. Análise das Séries Temporais

De acordo com Box *et. al.* [39], uma série temporal representa uma sequência de observações sobre um determinado fenômeno ou evento que evolui com o tempo. Tais observações são pontos temporais discretos ou contínuos, sendo estes usualmente equidistantes. Portanto, uma série temporal pode ser formalmente definida por [39]:

$$\mathbf{X} = \{x_t \in \mathbb{R} \mid t = 1, 2, 3 \ldots N\}, \tag{1}$$

onde o termo t representa um índice temporal (tempo), que determina a granularidade das observações, e o termo N representa o número de observações ou pontos da série temporal.

O principal objetivo ao se aplicar uma técnica de previsão a uma dada série temporal $\mathbf{X}$ é a construção de um mapeamento que seja capaz de estimar o futuro de um fenômeno temporal [40]. Em outras palavras, o objetivo de qualquer técnica de previsão é criar um mecanismo que permita, com certa precisão, a previsão dos valores futuros da série temporal, dados por x_{t+h}, onde h representa o horizonte de previsão de h passos a frente [40].

Neste trabalho, focamos em séries temporais relacionadas a concentração dos poluentes atmosféricos monóxido de carbono (CO), dióxido de nitrogênio (NO_2), material particulado (PM_{10} e $PM_{2.5}$), e dióxido de enxofre (SO_2) doletados a partir de sensores da *United States Environmental Protection Agency* (US EPA). Todas as séries temporais são amostradas em frequência diária (2018-01-01 a 2021-12-31 para CO, 2015-01-01 a 2018-12-31 para NO_2, 2016-01-01 a 2019-12-31 para PM_{10}, 2019-01-05 a 2022-12-31 para PM_{25} e 2013-01-01 a 2016-12-31 para SO_2), conforme apresentado na Figura 1.

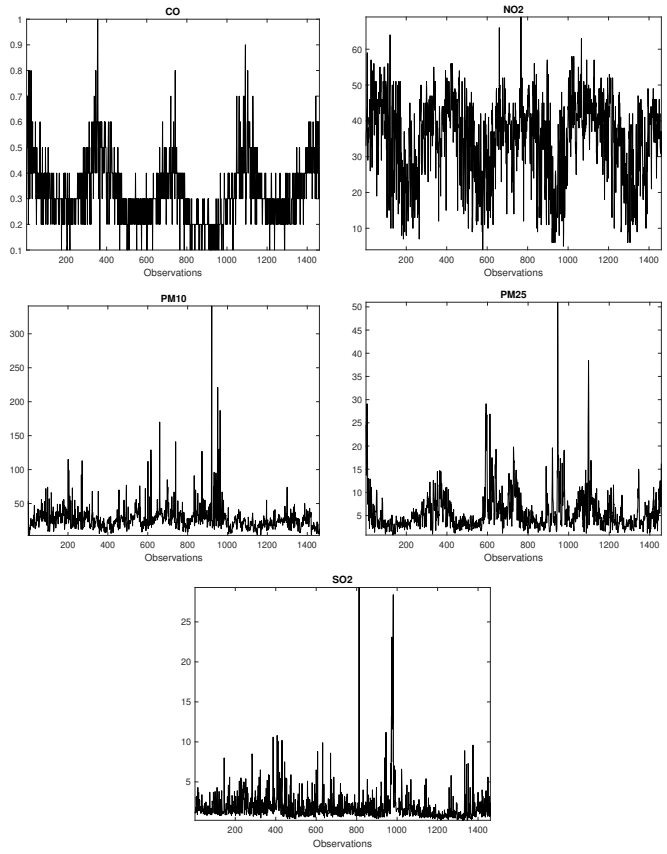

Figura 1. Gráfico das séries temporais.

Posteriormente, foram analisadas a função de autocorrelação (ACF) [39] e a função de autocorrelação parcial (PACF) [39], apresentadas nas Figuras 2 e 3, respectivamente. É possível verificar que, para as séries CO e NO2, há a presença de um decaimento lento com ciclos irregulares nas curvas da ACF. Já para as séries PM25 e SO2, é possível observar a curva da ACF com um característico decaimento hiperbólico ondular. Por fim, para a série PM10 é possível verificar padrões irregulares em sua curva da ACF. Vale mencionar que as curvas da PACF são caracterizadas por uma alta correlação em defasagens de baixa ordem, que diminui com o aumento na ordem dos retardos. Tal análise sugere a presença de algum tipo de não-linearidade nas séries temporais investigadas.

Embora sugestivo, ambas as ACF e PACF não permitem uma avaliação correta do relacionamento não-linear presente nas séries temporais investigadas. Portanto, analisamos a

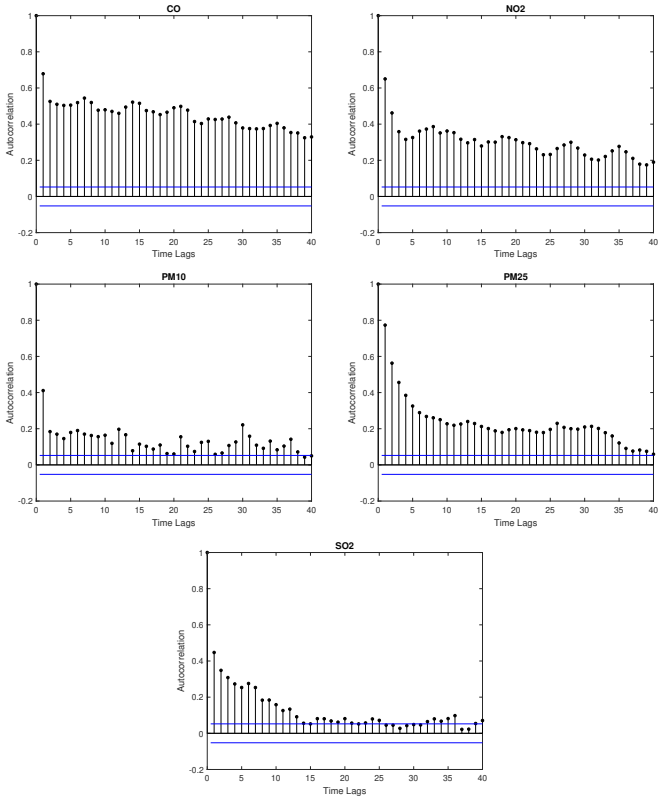

Figura 2. Função de autocorrelação das séries temporais.

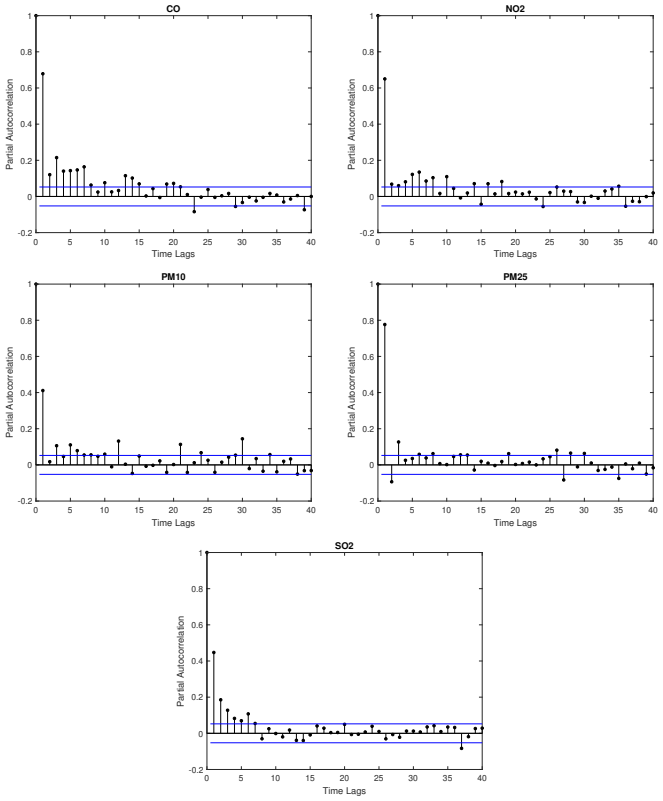

Figura 3. Função de autocorrelação parcial das séries temporais.

informação mútua média (MMI) [41], ilustrada na Figura 4, onde também é possível observar um decaimento lento com ciclos irregulares, confirmando a presença de um relacionamento não-linear.

Na tentativa de avaliar a natureza do relacionamento não-linear encontrado na análise da MMI, é investigado o parâmetro de Hurst (HP) [42], ilustrado na Figura 5. Vale mencionar, em todas as séries temporais investigadas, valor do HP está próximo de 0, associado a um comportamento anti-persistente, isto é, um processo auto-similar com dependência não-linear de longo prazo.

III. Descrição do Método Proposto

Baseado na análise de séries temporais apresentada na seção anterior, este trabalho se propõe a investigar modelos clássicos e recentemente apresentados na literatura para prever séries temporais relacionada a concentração de poluentes no ar.

A. Pré-Processamento

Neste trabalho utilizamos séries temporais com frequência diária relacionadas a concentração de poluentes atmosféricos. Os dados foram divididos em três conjuntos, de acordo com Prechelt [43]: 80% dos dados para o conjunto de treinamento, 10% para o conjunto de validação e 10% para o conjunto de teste. Todas as séries foram normalizadas no intervalo [0,1], permitindo uma análise comparativa dos resultados na mesma escala [43]. Utilizamos a estratégia de previsão *one-step-ahead*, onde o modelo é empregado para prever a série no tempo $t+1$ a partir dos retardos temporais t,$t-1$,$t-2$,...,$t-d$, sendo esta estratégia comum no contexto de problemas de previsão de qualidade do ar.

B. Configuração dos Modelos

O primeiro modelo investigado, *Autoregressive Integrated Moving Average* - **ARIMA(p,d,q)** é definido por três parâmetros, o termo autorregressivo (p), a diferenciação (d) e o termo de médias móveis (q).

O segundo modelo investigado *Linear Regression* - **LR** é um método estatístico utilizado para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. O objetivo é encontrar uma linha reta que melhor se ajuste aos dados, minimizando a soma dos quadrados das diferenças entre os valores observados e os valores previstos pela linha.

O terceiro modelo investigado *Polynomial Regression* - **PR** é uma técnica de modelagem que estende a regressão linear ao incluir termos polinomiais das variáveis independentes no modelo. Ela permite capturar relacionamentos não-lineares entre as variáveis e é útil quando os dados apresentam padrões complexos. A regressão polinomial envolve a adição de termos polinomiais de ordem superior às variáveis independentes. Isso permite que o modelo se ajuste a curvas mais flexíveis e não-lineares.

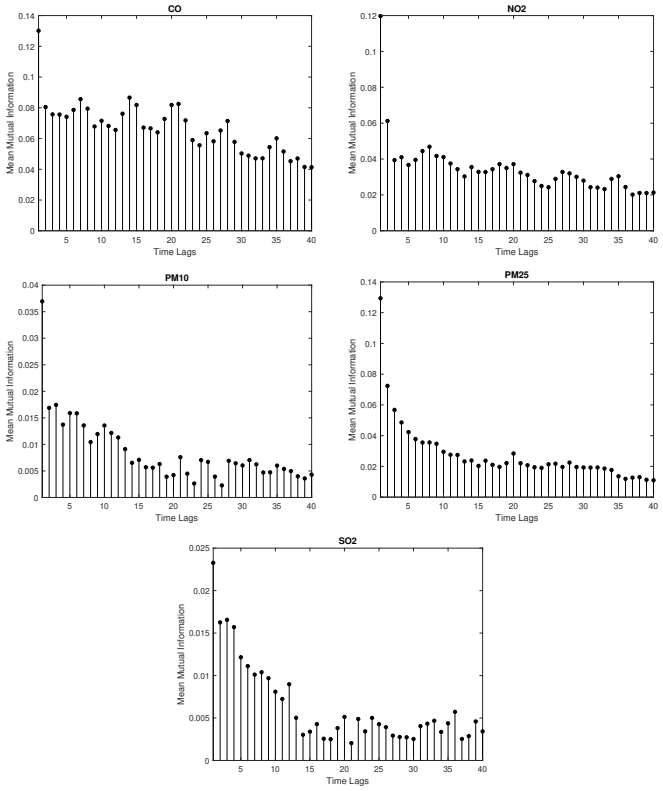

Figura 4. Informação mútua média das séries temporais.

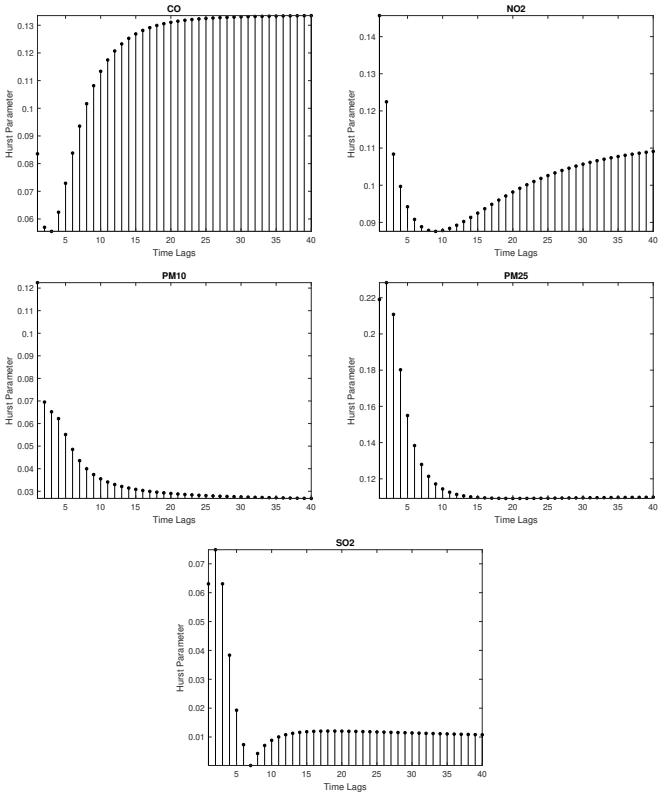

Figura 5. Parâmetro de Hurst das séries temporais.

O quarto modelo investigado, *Ridge Regression* - **RR** é um método de regressão linear regularizada que adiciona um termo de penalidade à função objetivo, a fim de lidar com o problema de multicolinearidade e reduzir a variância dos coeficientes estimados. Especificamente, o termo de penalidade é proporcional ao quadrado da norma L2 dos coeficientes, multiplicado por um parâmetro de regularização chamado de λ.

O quinto modelo investigado *Bayesian Ridge Regression* - **BRR** é um método estatístico que estende a regressão tradicional incorporando uma abordagem bayesiana. Em vez de fornecer apenas uma estimativa pontual dos coeficientes do modelo, o BRR fornece uma distribuição completa desses coeficientes, capturando a incerteza associada às estimativas. Além disso, é aplicado o termo de regularização RIDGE, que aplica uma penalidade de aprendizado L2 ao modelo.

O sexto modelo investigado, *Support Vector Regressor* - **SVR(I,KERNEL)** é definido por I retardos temporais e o KERNEL (rbf, linear e polinomial), sendo uma técnica não-paramétrica dependente das funções de *kernel*.

O sétimo modelo investigado, *Multilayer Perceptron* - **MLP(I,H,O)** é definido por uma camada de entrada (I - retardos temporais), uma camada oculta com H unidades de processamento e uma camada de saída com O unidades de processamento. Note que O=1 devido a estratégia de previsão de um-passo-adiante.

O oitavo modelo investigado, *Long-Short Term Memory* - **LSTM** é definida por uma camada de entrada (representada pelos retardos temporais), com uma camada oculta composta por unidades de processamento LSTM (responsáveis pelo mapeamento das dependências temporais da série), e uma camada de saída com uma unidade processamento clássica.

C. Definição de Hiperparâmetros

Em relação ao termos autoregressivos, de médias móveis e de diferenciação do modelo ARIMA, foi utilizada a metodologia apresentada na biblioteca *PMDARIMA*, do *Python*. Em relação aos kernels e seus respectivos hiperparâmetros do modelo SVR, foi utilizada a metodologia empregada na biblioteca *Scikitlearn* do *Python*. Para definição dos hiperparâmetros do LR, PR, RR e BRR foi utilizada a metodologia empregada na biblioteca *Scikitlearn* do *Python*.

Em relação as funções de ativação, foram investigadas a *Rectified Linear Unit* (ReLU), sigmóide e tangente hiperbólica. Para definição da quantidade de unidades de processamento, foi utilizada validação cruzada, onde foram investigados os valores 10, 50 e 100 para a primeira camada oculta.

Optou-se por utilizar o batchsize=1. Apesar desta escolha impactar em uma convergência mais lenta, no entanto esta tem um ganho na eficiência e eficácia do processo de aprendizagem, aumentando o poder de generalização do modelo. Para definição da taxa de aprendizagem, também foi empregado a validação cruzada, onde foram investigados os valores 0.1,

0.01 e 0.001. Para o algoritmo de aprendizagem, foram investigados os otimizadores sgd, adam, adamax e nadam. Foi utilizada a função de custo *Mean Squared Error* (MSE), que é comumente utilizada no processo de aprendizagem de diversos modelos de redes neurais.

Na tentativa de superar o problema de *overfitting*, foram utilizados três critérios de parada [43]: (i) épocas de treinamento 10^6, (ii) *generalization loss* (GL$> 5\%$), e (iii) *process training* (PT$< 10^{-6}$). Para definição dos retardos temporais para cada série temporal investigada, foi utilizada a metodologia apresentada na Seção II, onde foram investigados os retardos temporais utilizando a ACF, PACF, MMI e HP.

Vale mencionar que todos os modelos de redes neurais investigados foram desenvolvidos e implementados utilizando a biblioteca *keras* com o *backend Tensorflow*, do *Python*. Para cada configuração de hiperparâmetros, foram realizados trinta experimentos, onde foram calculadas a média (MEAN) e o desvio padrão (STD). Foi aplicado o teste de Friedman [44] com nível de significância $\alpha = 0.05$, pois estabelece um ranqueamento para todos os modelos.

D. Medidas de Desempenho

Diversas medidas de desempenho são encontrados na literatura. No entanto, a maior parte da literatura existente sobre previsão de séries temporais freqüentemente emprega apenas um critério de desempenho para avaliação de previsão. O critério de desempenho mais utilizado é o *mean squared error* (MSE), dado por:

$$\mathrm{MSE} = \frac{1}{N}\sum_{j=1}^{N}(\mathrm{target}_j - \mathrm{output}_j)^2, \qquad (2)$$

onde N é o número de padrões, target_j é a saída desejada para o padrão j e output_j é o valor previsto para o padrão j.

A medida MSE pode ser usada para direcionar o modelo de predição no processo de treinamento, mas não pode ser considerada sozinha como uma medida conclusiva para comparação de diferentes modelos de predição. Por esse motivo, outros critérios de desempenho devem ser considerados para permitir uma avaliação de desempenho mais robusta. Uma medida que apresenta com precisão a identificação dos desvios do modelo é o *mean absolute percentage error* (MAPE), dado por:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{\mathrm{target}_j - \mathrm{output}_j}{target_j}\right|. \qquad (3)$$

IV. Resultados Experimentais

Um resumo dos resultados para as séries temporais investigadas é apresentado na Tabela I, de acordo com as estatísticas MEAN e STD para a medida MSE. Vale ressaltar que o melhor desempenho é alcançado pelo modelo SVR (para a série CO), pelo modelo BRR (para a série NO2), pelo modelo ARIMA (para as séries PM10 e SO2) e pelo modelo LSTM (para a série PM25).

Tabela I
Desempenho no conjunto de teste para a medida MSE.

Modelo	CO	NO2	PM10	PM25	SO2
ARIMA	8.4100e-03 ±0.0000	1.3206e-02 ±0.0000	**4.8580e-04** ±0.0000	2.3565e-03 ±0.0000	**1.7043e-03** ±0.0000
BRR	6.2740e-03 ±0.0000	**1.3142e-02** ±0.0000	7.1769e-04 ±0.0000	2.1262e-03 ±0.0000	1.9521e-03 ±0.0000
LR	6.3345e-03 ±0.0000	1.3122e-02 ±0.0000	6.2081e-04 ±0.0000	2.1233e-03 ±0.0000	2.1021e-03 ±0.0000
LSTM	9.3582e-03 ±1.5198e-03	1.8178e-02 ±1.9723e-03	5.8983e-04 ±8.0594e-05	**1.7935e-03** ±2.5493e-04	1.9882e-03 ±1.6424e-04
MLP	6.6388e-03 ±8.4128e-04	1.4088e-02 ±1.3211e-03	7.2767e-04 ±5.9790e-05	2.5412e-03 ±2.2994e-04	1.9994e-03 ±7.1894e-05
PR	6.3345e-03 ±0.0000	1.3539e-02 ±0.0000	6.0369e-04 ±0.0000	2.1233e-03 ±0.0000	1.9848e-03 ±0.0000
RR	6.2818e-03 ±0.0000	1.3169e-02 ±0.0000	6.1906e-04 ±0.0000	2.2820e-03 ±0.0000	1.9558e-03 ±0.0000
SVR	**6.1483e-03** ±0.0000	1.3486e-02 ±0.0000	2.0738e-03 ±0.0000	3.1315e-03 ±0.0000	3.6173e-03 ±0.0000

O resultado obtido para o teste de Friedman considerando a medida MSE é apresentado na Tabela II, onde é possível confirmar estatisticamente os resultados apresentados na Tabela I.

Tabela II
Resultado do teste de Friedman para a medida MSE.

	χ^2=8.5333 and p-value=2.8792e-01	
Posição	Modelo	*Rank*
1	BRR	3.20
2	RR	3.60
3	ARIMA	3.80
4	LR	4.00
5	PR	4.00
6	LSTM	4.80
7	SVR	6.00
8	MLP	6.60

Um resumo dos resultados para as séries temporais investigadas é apresentado na Tabela III, de acordo com as estatísticas MEAN e STD para a medida MAPE. Vale ressaltar que o melhor desempenho é alcançado pelo modelo LR e PR (para as séries CO e PM25), pelo modelo PR (para a série NO2), pelo modelo LSTM (para a série PM10) e pelo modelo LR (para a série SO2).

O resultado obtido para o teste de Friedman considerando a medida MAPE é apresentado na Tabela IV, onde é possível confirmar estatisticamente os resultados apresentados na Tabela III.

Na Figura 6 são apresentados os resultados da previsão gerada pelos melhores modelos dentre os investigados neste trabalho, para cada poluente em particular, referentes ao conjunto de teste.

V. Conclusões

Neste artigo foi apresentado um estudo empírico sobre séries temporais relacionadas à concentração de poluentes atmosféricos, considerando a função de autocorrelação e a função de autocorrelação parcial (para análise da dependência

Tabela III
DESEMPENHO NO CONJUNTO DE TESTE PARA A MEDIDA MAPE.

Model	CO	NO2	PM10	PM25	SO2
ARIMA	0.3046 ±0.0000	0.3394 ±0.0000	0.5599 ±0.0000	0.5226 ±0.0000	1.2374 ±0.0000
BRR	0.2652 ±0.0000	0.3368 ±0.0000	0.8475 ±0.0000	0.5509 ±0.0000	1.6552 ±0.0000
LR	**0.2649** ±0.0000	0.3352 ±0.0000	0.7321 ±0.0000	**0.5507** ±0.0000	**1.2366** ±0.0000
LSTM	0.3282 ±2.8931e-02	0.4147 ±4.7001e-02	**0.5301** ±9.5620e-02	0.7083 ±0.1329	1.5695 ±0.3413
MLP	0.2744 ±8.0469e-03	0.3457 ±2.9861e-02	0.6816 ±6.2211e-02	0.5762 ±0.1150	1.2436 ±0.1486
PR	**0.2649** ±0.0000	**0.3314** ±0.0000	0.7055 ±0.0000	**0.5507** ±0.0000	1.4786 ±0.0000
RR	0.2651 ±0.0000	0.3385 ±0.0000	0.7587 ±0.0000	0.5740 ±0.0000	1.7562 ±0.0000
SVR	0.2692 ±0.0000	0.3267 ±0.0000	1.2004 ±0.0000	0.9166 ±0.0000	3.0488 ±0.0000

Tabela IV
RESULTADO DO TESTE DE FRIEDMAN PARA A MEDIDA MAPE.

	χ^2=10.2000 and p-value=1.7752e-01	
Posição	Modelo	*Rank*
1	LR	2.60
2	PR	2.80
3	ARIMA	3.60
4	BRR	5.00
5	MLP	5.00
6	RR	5.20
7	LSTM	5.80
8	SVR	6.00

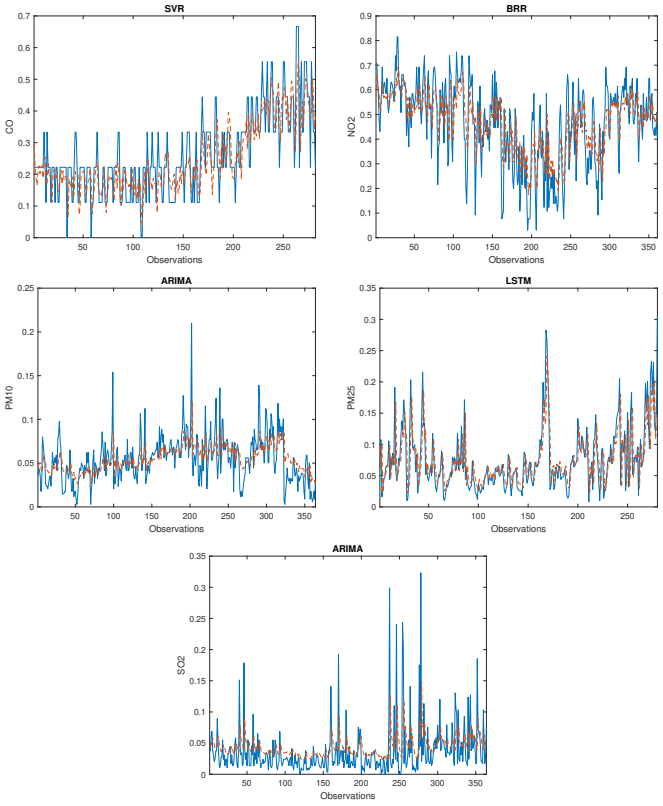

Figura 6. Previsão das séries temporais.

linear), bem como a informação mútua média e o parâmetro de Hurst (para análise da dependência não-linear).

Com base neste estudo, que demonstrou a previsibilidade deste tipo particular de série temporal, propusemos investigar modelos baseados em aprendizagem de máquina a fim de encontrar o melhor mapeamento capaz de reconstruir o fenômeno gerador destas séries.

No entanto, indo contra os resultados obtidos na análise das séries temporais, a não-linearidade presente nos modelos investigados não foi capaz, na prática, de superar em termos de desempenho preditivo, os modelos lineares investigados neste trabalho. O que leva a crer que os modelos não-lineares investigados não foram capazes de mapear a complexidade da não-linearidade presente no fenômeno gerador destas séries temporais.

Este resultado está de acordo com os resultados reportados na literatura de previsão de poluentes atmosféricos, haja visto que estes argumentam sobre a incapacidade dos modelos não-lineares capturarem a complexidade presente no fenômeno gerador destas séries, levando a um desempenho similar ou levemente inferior aos modelos lineares comumente utilizados na literatura. A avaliação das medidas MSE e MAPE corroboram com esta hipótese, que é validada estatisticamente pelo teste de Friedman, na qual nenhum modelo de redes neurais não-lineares ficou no top 3, em termos de desempenho preditivo.

Desta forma, estudos adicionais devem ser realizados como trabalho futuro na tentativa de justificar tal comportamento. Além disso, é importante confirmar os achados deste trabalho para os poluentes considerados em outras regiões, de forma a validar os resultados obtidos em contextos diferentes.

REFERÊNCIAS

[1] A. Bekkar, B. Hssina, S. Douzi, and K. Douzi. Air-pollution prediction in smart city, deep learning approach. *Journal of Big Data*, 8(161), 2021.

[2] The World Bank. Urban population (% of total population). https://data.worldbank.org/indicator/sp.urb.totl.in.zs. 2023.

[3] United Nations Department of Economic and Social Affairs. Revision of world urbanization prospects. https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html. 2018.

[4] Meric Yilmaz Salman and Halil Hasar. Review on environmental aspects in smart city concept: Water, waste, air pollution and transportation smart applications using iot techniques. *Sustainable Cities and Society*, 94:104567, 2023.

[5] P. Asha, L. Natrayan, B.T. Geetha, J. Rene Beulah, R. Sumathy, G. Varalakshmi, and S. Neelakandan. Iot enabled environmental toxicology for air pollution monitoring using ai techniques. *Environmental Research*, 205:112574, 2022.

[6] Nada Osseiran and Christian Lindmeier. 9 out of 10 people worldwide breathe polluted air, but more countries are taking action. *https://www.who.int/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action*, 2018.

[7] Mohsen Abbasi-Kangevari et. al. Effect of air pollution on disease burden, mortality, and life expectancy in north africa and the middle

east: a systematic analysis for the global burden of disease study 2019. *The Lancet Planetary Health*, 7(5):e358–e369, 2023.

[8] Isabell Katharina Rumrich, Julian Lin, Antti Korhonen, Lise Marie Frohn, Camilla Geels, Jørgen Brandt, Sirpa Hartikainen, Otto Hänninen, and Anna-Maija Tolppanen. Long-term exposure to low-level particulate air pollution and parkinson's disease diagnosis - a finnish register-based study. *Environmental Research*, 229:115944, 2023.

[9] Gan Wu, Miao Cai, Chongjian Wang, Hongtao Zou, Xiaojie Wang, Junjie Hua, and Hualiang Lin. Ambient air pollution and incidence, progression to multimorbidity and death of hypertension, diabetes, and chronic kidney disease: A national prospective cohort. *Science of The Total Environment*, 881:163406, 2023.

[10] Guangzhi Qi, Jiahang Che, and Zhibao Wang. Differential effects of urbanization on air pollution: Evidences from six air pollutants in mainland china. *Ecological Indicators*, 146:109924, 2023.

[11] Haimeng Liu, Weijia Cui, and Mi Zhang. Exploring the causal relationship between urbanization and air pollution: Evidence from china. *Sustainable Cities and Society*, 80:103783, 2022.

[12] Anil Nanda, Syed Shahzad Mustafa, Maria Castillo, and Jonathan A. Bernstein. Air pollution effects in allergies and asthma. *Immunology and Allergy Clinics of North America*, 42(4):801–815, 2022.

[13] OECD. *The Economic Consequences of Outdoor Air Pollution*. 2016.

[14] H. J. S. Fernando, M. C. Mammarella, G. Grandoni, P. Fedele, R. Di Marco, R. Dimitrova, and P. Hyde. Forecasting pm10 in metropolitan areas: Efficacy of neural networks. *Environmental Pollution*, 163:62–67, 2012.

[15] Zhongshan Yang and Jian Wang. A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction. *Environmental Research*, 158:105–117, 2017.

[16] Celeste Eusébio, Maria João Carneiro, Vitor Rodrigues, Margarita Robaina, Mara Madaleno, Carla Gama, Kevin Oliveira, and Alexandra Monteiro. Factors influencing the relevance of air quality in the attractiveness of a tourism destination: Differences between nature-based and urban destinations. *Tourism Management Perspectives*, 44:101045, 2022.

[17] S. G. Ilieva, A. Ivanov, D. Voynikova, and D. Boyadzhiev. Time series analysis and forecasting for air pollution in small urban area: an sarima and factor analysis approach. *Stochastic Environmental Research and Risk Assessment*, 28(4):1045–1060, 2014.

[18] M. Vedrenne, R. Borge, J. Lumbreras, and M. E. Rodriguez. Advancements in the design and validation of an air pollution integrated assessment model for spain. *Environmental Modelling and Software*, 57:177–191, 2014.

[19] Terhi Kangas, Sylvie Gadeyne, Wouter Lefebvre, Charlotte Vanpoucke, and Lucía Rodriguez-Loureiro. Are air quality perception and pm2.5 exposure differently associated with cardiovascular and respiratory disease mortality in brussels? findings from a census-based study. *Environmental Research*, 219:115180, 2023.

[20] Xun Deng, Bin Zou, Shenxin Li, Jian Wu, Chenjiao Yao, Minxue Shen, Jun Chen, and Sha Li. Disease specific air quality health index (aqhi) for spatiotemporal health risk assessment of multi-air pollutants. *Environmental Research*, 231:115943, 2023.

[21] V. A. Reisen, A. J. Q. Sarnaglia, N. C. Reis, C. L. Leduc, and J. M. Santos. Modeling and forecasting daily average pm10 concentrations by a seasonal long-memory model with volatility. *Environmental Modelling and Software*, 51:286–295, 2014.

[22] G. Kiesewetter, W. Schoepp, C. Heyes, and M. Amann. Modelling pm2.5 impact indicators in europe: Health effects and legal compliance. *Environmental Modelling and Software*, xxx:1 – 11, 2015.

[23] Mohammad Ali Akbarzadeh, Isa Khaheshi, Amirsina Sharifi, Negin Yousefi, Mohammadreza Naderian, Mohammad Hasan Namazi, Morteza Safi, Hossein Vakili, Habibollah Saadat, Saeed Alipour Parsa, and Negin Nickdoost. The association between exposure to air pollutants including pm10, pm2.5, ozone, carbon monoxide, sulfur dioxide, and nitrogen dioxide concentration and the relative risk of developing stemi: A case-crossover design. *Environmental Research*, 161:299–303, 2018.

[24] Arnold D. Bergstra, Bert Brunekreef, and Alex Burdorf. The influence of industry-related air pollution on birth outcomes in an industrialized area. *Environmental Pollution*, 269:115741, 2021.

[25] P. S. G. Mattos, F. Madeiro, T. A. E. Ferreira, and G. D. C. Cavalcanti. Hybrid intelligent system for air quality forecasting using phase adjustment. *Engineering Applications of Artificial Intelligence*, 32:185–191, 2014.

[26] Bihter Das, Omer Osman Dursun, and Suat Toraman. Prediction of air pollutants for air quality using deep learning methods in a metropolitan city. *Urban Climate*, 46:101291, 2022.

[27] P.J. García Nieto, F. Sánchez Lasheras, E. García-Gonzalo, and F.J. de Cos Juez. Pm10 concentration forecasting in the metropolitan area of oviedo (northern spain) using models based on svm, mlp, varma and arima: A case study. *Science of The Total Environment*, 621:753–761, 2018.

[28] Erdinc Aladag. Forecasting of particulate matter with a hybrid arima model based on wavelet transformation and seasonal adjustment. *Urban Climate*, 39:100930, 2021.

[29] Lingxiao Zhao, Zhiyang Li, and Leilei Qu. Forecasting of beijing pm2.5 with a hybrid arima model based on integrated aic and improved gs fixed-order methods and seasonal decomposition. *Heliyon*, 8(12):e12239, 2022.

[30] Jun Luo and Yaping Gong. Air pollutant prediction based on arima-woa-lstm model. *Atmospheric Pollution Research*, 14(6):101761, 2023.

[31] Licheng Zhang, Xue Tian, Yuhan Zhao, Lulu Liu, Zhiwei Li, Lixin Tao, Xiaonan Wang, Xiuhua Guo, and Yanxia Luo. Application of nonlinear land use regression models for ambient air pollutants and air quality index. *Atmospheric Pollution Research*, 12(10):101186, 2021.

[32] Xiang Li, Ling Peng, Xiaojing Yao, Shaolong Cui, Yuan Hu, Chengzeng You, and Tianhe Chi. Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, 231:997–1004, 2017.

[33] Yongan Li, Peng Jiang, Qingshan She, and Guang Lin. Research on air pollutant concentration prediction method based on self-adaptive neurofuzzy weighted extreme learning machine. *Environmental Pollution*, 241:1115–1127, 2018.

[34] Wei Sun and Chenchen Huang. A hybrid air pollutant concentration prediction model combining secondary decomposition and sequence reconstruction. *Environmental Pollution*, 266:115216, 2020.

[35] Hong Yang, Zehang Liu, and Guohui Li. A new hybrid optimization prediction model for pm2.5 concentration considering other air pollutants and meteorological conditions. *Chemosphere*, 307:135798, 2022.

[36] Bo Zhang, Yi Rong, Ruihan Yong, Dongming Qin, Maozhen Li, Guojian Zou, and Jianguo Pan. Deep learning for air pollutant concentration prediction: A review. *Atmospheric Environment*, 290:119347, 2022.

[37] Bo Zhang, Guojian Zou, Dongming Qin, Qin Ni, Hongwei Mao, and Maozhen Li. Rcl-learning: Resnet and convolutional long short-term memory-based spatiotemporal air pollutant concentration prediction model. *Expert Systems with Applications*, 207:118017, 2022.

[38] Bo Zhang, Yuan Liu, RuiHan Yong, Guojian Zou, Ru Yang, Jianguo Pan, and Maozhen Li. A spatial correlation prediction model of urban pm2.5 concentration based on deconvolution and lstm. *Neurocomputing*, 544:126280, 2023.

[39] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, New Jersey, third edition, 1994.

[40] Ricardo de A. Araújo, Adriano L.I. Oliveira, and Silvio Meira. On the problem of forecasting air pollutant concentration with morphological models. *Neurocomputing*, 265:91–104, 2017.

[41] M. B. Stojanovic, M. M. Bozic, M. M. Stankovic, and Z. P. Stajic. A methodology for training set instance selection using mutual information in time series prediction. *Neurocomputing*, 141:236–245, 2014.

[42] E. Hurst. Long term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–799, 1951.

[43] L. Prechelt. Proben1: A set of neural network benchmark problems and benchmarking rules. Technical Report 21/94, 1994.

[44] M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Statist.*, 11(1):86–92, 03 1940.