A DEEP LEARNING APPROACH WITH WAVELETS TO FORECASTING TROPOSPHERIC OZONE IN A TROPICAL METROPOLITAN REGION

Clovis Carmo Junior[1], Ingrid Winkler[1], Erick Giovani Sperandio Nascimento[1]

[1] Grupo de Pesquisa em Modelagem e Inteligência Computacional (GPMIC), SENAI CIMATEC - BA, Brazil

Abstract: Deep neural networks (DNN) have been successfully applied to develop air pollutant forecasting models. Once trained, they can learn the complex relationships and non-linearities present in atmospheric variables, delivering solutions that require less computational resources than numerical and analytical models. This study aims to develop a DNN to forecast concentrations of ozone (O_3) for the next 24hs. We tested several DNN using, as input, a multivariate time series dataset. When wavelets (WL) were integrated into the DNN's architectures, they improved the models' performance, pointing towards a consistent modeling architecture for air pollution forecasting. These deep learning models showed flexibility, strong nonlinear fitting capabilities and an ability to map nonlinear complexity from the data for air pollution forecasting.

Keywords: Air pollution; Forecast; Deep learning; Wavelets, Ozone.

UMA ABORDAGEM DE APRENDIZAGEM PROFUNDA COM WAVELETS PARA A PREVISÃO DE OZÔNIO TROPOSFÉRICO EM UMA REGIÃO METROPOLITANA TROPICAL

Resumo: Redes de aprendizado profundo (RAP) têm sido usadas com sucesso em modelos de predição de poluentes do ar. Após treinadas, elas aprendem as relações complexas e não-linearidades das variáveis atmosféricas e acham soluções que requerem menos recursos computacionais que modelos numéricos e analíticos. Este estudo visa desenvolver uma RAP para predição das concentrações de O_3 nas próximas 24hs. Testou-se várias arquiteturas de RAP, usando séries temporais multivariáveis como dados de entrada. Integrando as WL na arquitetura, temos melhores resultados, indicando consistência para modelos de predição de poluentes. Estes modelos de aprendizagem profunda demonstram flexibilidade, capacidade de ajuste e mapeamento da complexidade não-linear de dados de predição de poluentes.

Palavras-chave: Poluição do ar; Predição, Aprendizado profundo; Wavelets, Ozônio.

1. INTRODUCTION

Metropolitan areas may suffer an increase in air pollution due to the growth of urbanization, transportation, and industrial sectors [1]. O_3 is an air pollutant with a significant impact on the environment and human health [2]. At-risk populations that suffer most from the harmful effects of air pollution are children, the elderly and people with respiratory and cardiovascular diseases [3].The current Brazilian legislation for air quality follows resolution number 491/2018 [4]. It aims for the standard levels recommended by the World Health Organization (WHO). It starts from lower targets until it reaches the recommended WHO guidelines established in 2005. Even though there is an apparent effort to improve air quality in Brazil, there is still a problematic lack of data with most Brazilian cities missing air monitoring systems [5].

Several works propose different techniques. In [6], it is proposed a denoising autoencoder deep network (DAEDN) model that is based on long short-term memory (LSTM). They created a noise reduction autoencoder with an LSTM network to extract the inherent air quality characteristics of original monitoring data to implement noise reduction processing to improve accuracy for their predictions, they report improvement versus their unidirectional LSTM model. In [7], particulate matter (PM_{10}) concentration is forecasted applying a linear and a nonlinear model, which was based on a feed forward configuration of a multilayer perceptron (MLP). They evaluated both models using the full original signal and modeled the residual component after removing periodic component from the original signal. They found that the approach based on the periodic component removal presented the best results. In [8], the study uses the technique Gray Level Co-occurrence Matrix (GLCM) to feature extraction from the data. The extracted features are then optimized. So, they apply the combined Support Vector Regression (SVR) and an LSTM based deep learning model to predict their air quality indexes. In [9], the study presents a hybrid prediction model using information gain, wavelet (WL) transform, and LSTM applied to daily concentration prediction of air pollutants in Beijing-China. They used the six major air pollutants concentration to build their dataset. So, they select one each time as an output variable and the remaining five as input variables to explore the interactions among them, repeating the process to cover all six targets. They reported results with higher prediction accuracy and stability than other models used for comparison. In [10], the work features a predicting model based on discrete WL transform (DWT) and LSTM network for predict the next day nitrogen dioxide (NO_2) concentration in the city of Tianjin-China. Their results show improvement in model accuracy and generalization. Their model seems to be more suitable to perform the job than the other methods considered for comparison in their work. In [11], LSTM and DWT are applied to model the spatial variability of hourly NO_2 levels at six urban sites in Central London-UK. They train the model using only the NO_2 concentration data from the neighboring sites. As benchmark models for comparison, they use plain LSTN and MLP. At most sites, their model presents better forecasting results. In [12], the authors evaluate the support vector regression (SVR) model with and without applying WL transformation for the six major air pollutants prediction, in Isfahan metropolis, central Iran. Their results bring the WL-SVR model with improved performance versus SVR model. In [13], the viability of using recurrent and convolutional neural network models to predict ozone levels in the troposphere is evaluated. They concluded it is viable and the LSTM model has presented the best result. Both references reinforce the importance of understanding the input data and how to treat the noise from it, given the site boundary conditions and available attribute data to build the baseline dataset.

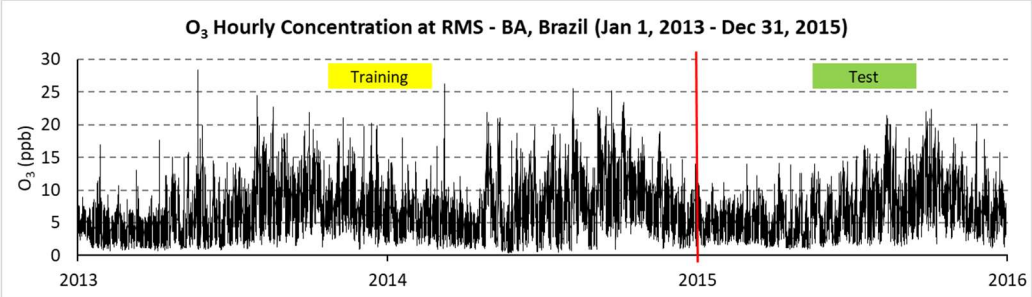
ISSN: 2357-7592

The application of deep neural network forecast models is a faster and better alternative to understanding the air quality status in Brazilian cities. As these forecast models have been proved accurate and efficient, they can help public authorities drive policies to improve air quality and issue daily air quality forecasts for the population, which can positively contribute to public health. Many air quality studies in Brazil apply numerical models and investigate urban and/or industrial areas located in the south-southeast regions[14-16]. Nevertheless, Brazil is a vast country, with the northeast region presenting both significant industrial and population growth in the last decades [5]. This study aims to develop a DNN to forecast concentrations of O_3 for the next 24hs. We will develop the models and test them on the metropolitan region of Salvador (MRS), the biggest metropolitan area in the northeast of Brazil, contributing to the increase of air quality level understanding in this region. We use as input a multivariate time series dataset composed of meteorological parameters and other air pollutants, comprising three years of measurements. To improve model performance, we apply wavelets to denoise our input data and increase our dataset dimensionality. This paper is organized as follows: Section 2 introduces the site, collected data, explains the modeling process and applied methods. Section 3 features the results and discussion. At the end, Section 4 summarizes our conclusions.

2. METHODOLOGY

The models were evaluated using a dataset from the MRS in Bahia state in the northeast of Brazil. MRS is in the tropics with reference latitude 12º 58' 16" S, longitude 38º 30' 39" W, and altitude 8m. The MRS is an urban, industrial, and coastal area, currently composed of 13 cities with a total area around 4.375 km^2. The Atlantic Forest surrounds it, and its population is over 4 million inhabitants. The region's economy is based on tourism, commerce, and several industries such as plastics, petrochemical, thermoplastic resins, fertilizers, copper metallurgy, and automotive among others [5]. The models were developed and evaluated using data from eight air quality monitoring stations, distributed in MRS downtown. The raw dataset contains the hourly average from three years of measurements of meteorological and air pollutants parameters. The measured atmospheric variables are temperature, humidity, wind speed, wind direction, and rain. The air pollutants parameters are carbon monoxide (CO), sulfur dioxide (SO_2), NO_2, PM_{10} and the target O_3. Figure 1 shows the original time series signal for O_3 concentration.

Figure 1. Original hourly data from Jan 2013 to Dec 2015



This study was divided into five steps. First, we cleaned the raw data and chose the metrics to evaluate the models' performance. Second, we took the updated dataset

with the cleaned data and applied the wavelets transform to denoise it and increase the input data dimensionality. In addition, a new dataset was built with the original data and the decomposed signals from the wavelet application, so we normalized the input parameters. Then, we worked on the modeling process building and training three different neural network architectures LSTM, RNN and MLP. Their hyperparameters were optimized to achieve good generalization, avoiding overfit and underfit. Fourth, we evaluated them with the original normalized input signal and the WL denoised signal producing the following models LSTM-WL, RNN-WL and MLP-WL to compare their performance to each other and with the plain model results. Finally, we compared and discussed the results.

To build effective models, we need to consider as inputs the atmospheric variables and all other available boundary conditions that may affect the air quality such as air pollutants concentration in the area, emissions inventory, land use, and area topography. It is also essential to capture and understand the noise in the input data to carefully select a methodology or technique to treat it, improving model forecast results without losing model fidelity. We ran a qualitative analysis in the dataset identifying and deleting samples with issues. After cleaning, we ended up with 26,226 measurement samples corresponding to more than 99.7% of the total three years original hourly data. Right after this step, we ran a quantitative analysis to understand the input data behavior and extract features to help with the selection method for the signal denoising preprocessing phase. We split our dataset of hourly data in training and test subsets. The first two years, 2013 and 2014, were used for training and the year, 2015, for tests. Along the training process to set up the hyperparameters and get a good model generalization, we used 70% of the training dataset for training and 30% for validation. Once the models are trained, we tested them applying the test dataset and used the predicted values to compare with the corresponding actual target. In the process of choosing the optimal wavelet for signal denoise we used the wavelet function objective selection method, which applies to the selected data the discrete wavelets transform using different families (48 different mother wavelet functions) [17]. The goal is to compare the original signal from each variable with the performance of approximate coefficients for each variable identifying the wavelet family with best representation for the observed data. After the decomposition, the new dataset is built with the original signal and the decomposed signals, evaluating the different wavelet family levels representation to find the one with best performance for model training, validation, and test. This work uses two performance metrics the coefficient of regression (R^2), then the normalized mean squared error (NMSE). To categorize NMSE into low, medium, or high performance, the use of standard deviation to normalize is a good option as the standard deviation based NMSE gives the variation not explained by the regression vs the overall variation in the desired target [18].

O_3 hourly concentration series is nonstationary as it is affected by other air pollutants and meteorological parameters levels variation. A wavelet transform is very suitable for dealing with nonstationary time series including air pollutant data [10]. Air pollution time series generally have short-time transient components. The wavelet transformation can effectively extract time-frequency information from that, as it is free from the stationary assumption [12]. The effectiveness of the wavelet transformation method is affected by the choice of the wavelet function and the number of decomposition levels [11]. The decomposed signal has more stable variance and fewer singular data points than the original data. The low-frequency signal can reflect the actual change trend of the signal data and random disturbing high-frequency signal [9].

So, the denoised signal can mimic the original signal information more effectively, producing a better prediction for the related data. As part of the preprocessing phase the input data is also normalized using the MinMaxScaler function, from sckit-learn library version 0.24.2, within the feature range from -1 to 1. The target variable though is kept with its original dimension not being affected by the normalization process. Regarding the computational modeling process, we adopted the computational intelligence deep learning model using wavelet decomposition, MLP, RNN, and LSTM. Various studies apply these kinds of neural networks in this field [9]-[12]. We applied the wavelet transform to decompose the eleven atmospheric variables considered in this study's original dataset. Moreover, we also applied an algorithm to verify the effectiveness of several wavelet functions for the original signal decomposition seeking for the lower root mean squared error (RMSE). Using the selected most effective wavelet for each variable in the dataset, we trained and tested the models up to the level five decomposition to find the best result for each model evaluated. During the modeling process we run a design of experiment (DOE) on each model MLP, RNN, and LSTM targeting the best combination for the models hyperparameters to optimize the models output performance. This selection is fundamental for good model performance. The basic architecture details for all models are presented in Table 1.

Table 1: Models Configuration (*number of neurons)

Model	Input Node*	1st Hidden Layer*	2nd Hidden Layer*	Dropout Layer	Output Layer*	Epochs	Batch Size
MLP	11	12	1	-	24	150	4
WL-MLP	33-55-77-99-121	12	1	-	24	150	4
RNN	11	4	-	0.4	24	600	18
WL-RNN	33-55-77-99-121	4	-	0.4	24	100	18
LSTM	11	50	-	0.35	24	300	18
WL-LSTM	33-55-77-99-121	50	-	0.35	24	300	18

The MLP is a popular type of feedforward artificial neural network (ANN) architecture composed of an input layer, an output layer and one or more hidden layers formed by neurons and their connections. Our developed MLP models contain an input layer, two hidden layers, and an output layer. The activation functions are ‘relu’ for the hidden layers and ‘linear’ for the output layer. The RNN is an ANN architecture applied for sequential data. Its key characteristic is the internal memory which recalls the information from its input throughout the next layer, making it adequate for application in sequential data problems as time series. The developed RNN in this study contains an input layer, one hidden layer, and the output layer. For the hidden layer it is used the TensorFlow Core v2.5.0 package library tf.keras.layers.SimpleRNN. The linear activation function is used for the hidden and output layers. The LSTM is also an artificial RNN architecture, but with the ability to delete, write, and read information from its memory, being able also to recall inputs over a long period. They have many applications, including predictions built out of time series, fitting our study’s purpose.

ISSN: 2357-7592

The base LSTM architecture in this study consists of one input layer, one LSTM hidden layer, one dropout layer, and one output layer. For the hidden layer, we used the TensorFlow Core v2.5.0 package library tf.keras.layers.LSTM. The linear activation function is used for the output layer. Other hyperparameters are set as follows: the optimizer is tf.keras.optimizers.SDG, from TensorFlow Core v2.5.0 package library with a learning rate of 1%.

3. RESULTS AND DISCUSSION

The top results for the performance of wavelets by parameter signal are reported in Table 2 for air pollutants and meteorological variables. All models required several DOE loops for hyperparameters optimization ending up with good generalization avoiding overfit. Table 3 brings NMSE and R^2 average comparison from O_3 24h time horizon prediction in all studied models. Figure 2 highlights the LSTM models showing the best performance with WL level 5. WL-LSTM level 5 24h average results are R^2 = 0.72 and NMSE = 1.02, the best ones shown on Table 3. The results show the potential of the wavelets function when applied to nonstationary time series data. It decomposes the signal and captures the features and behavior from parameters overtime. Additionally, it better represents them on the signal recomposition, reducing signal noise and helping the models to achieve a higher performance in their predictions. This fact is observed in all studied models, with significant percentage improvement. Therefore, the exploration of this method, together with the variety of the deep neural network alternatives in the development of air pollutants prediction models, becomes very promising.

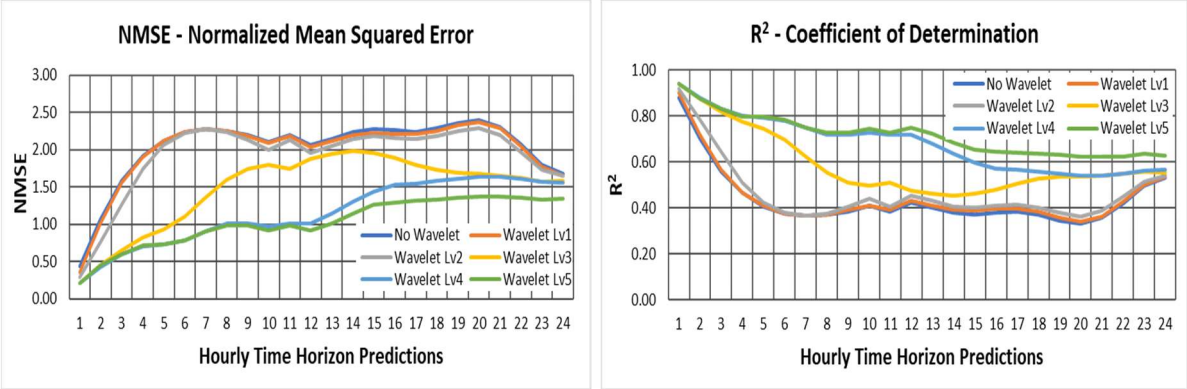
Table 2: Air pollutants and Meteorological data signals wavelet selection

Data	PM_{10}	O_3	SO_2	NO_2	CO	Rain	T	Humidity	Wind Speed	Wind Dir.
Train	db38	db28	bior2.4	db35	db38	coif1	db34	db35	db35	db37
Test	db31	db36	db2	db36	db38	db9	db33	db33	db34	sym20

Table 3: O_3 model results

Model	Configuration	NMSE	R^2
MLP	No Wavelet Application	2.00	0.44
WL-MLP	Wavelet Level 5	1.08	0.70
RNN	No Wavelet Application	2.01	0.44
WL-RNN	Wavelet Level 5	1.06	0.71
LSTM	No Wavelet Application	2.01	0.44
WL-LSTM	**Wavelet Level 5**	**1.02**	**0.72**

Figure 2: O_3 LSTM model - metrics results for prediction from 1h to 24h time horizon.



4. CONCLUSION

The deep learning approach integrated with wavelets is a worthy alternative to be further explored in the field of air pollutants prediction due to its faster results, low computational cost, and flexibility to adjust the wavelet selection for the atmospheric parameters inputs for different sites' conditions. This last characteristic is fundamental as the qualitative and quantitative properties from the input data affect model performance. It may drive the model to choose a different wavelet family, during the input signal optimization loop, to decompose the new site original signal and achieve the optimal model performance. The findings in this work are the first of its field in Brazil. We did not find any studies about deep neural networks integrated with wavelets to forecast the next 24 hours of air pollutants ground level concentration in a tropical metropolitan region from Brazil's Northeast. Our performance model can support local public authorities by providing short-term air pollutants forecasts for the next 24h. This information would allow authorities to warn at-risk populations about the expected air quality conditions, allowing them to reschedule activities in certain areas and avoid unnecessary exposure to non-recommended levels of air pollutants. This can be explored in future research for other air pollutants in this site or others that share similar boundary conditions. Future studies can also compare model performance results between the deep neural networks integrated with wavelets approach and numerical and analytical models.

5. REFERENCES

[1] PEDRUZZI, Rizzieri et al. Performance evaluation of a photochemical model using different boundary conditions over the urban and industrialized metropolitan area of Vitória, Brazil. **Environmental Science and Pollution Research**, [S.L.], v. 26, n. 16, p. 16125-16144, 10 abr. 2019. Springer Science and Business Media LLC.

[2] WHO 2013. Review of evidence on health aspects of air pollution-REVIHAAP project: final technical report. World Health Organization Regional Office for Europe. https://www.euro.who.int/en/health-topics/environment-and-health/air-quality/publications/2013/review-of-evidence-on-health-aspects-of-air-pollution-revihaap-project-final-technical-report Accessed May 3, 2021

[3] RYBARCZYK, Yves et al. Machine Learning Approaches for Outdoor Air Quality Modelling: a systematic review. **Applied Sciences**, [S.L.], v. 8, n. 12, p. 2570, 11 dez. 2018. MDPI AG.

[4] CONAMA Environmental National Committee, New national air quality standards, Res. 491/ 2018. www2.mma.gov.br/port/conama/legiabre.cfm?codlegi=740. Accessed on: 23 Apr. 2021

[5] KITAGAWA, Yasmin Kaore Lago et al. ASSESSMENT OF PRIMARY AIR POLLUTANTS IN A TROPICAL METROPOLITAN REGION BY COMBINING LOCAL AND GLOBAL EMISSIONS INVENTORIES. **Air Pollution XXVII**, [S.L.], v. 236, p. 99-110, 4 set. 2019. WIT Press.

[6] CAI, Jianxian et al. An Air Quality Prediction Model Based on a Noise Reduction Self-Coding Deep Network. **Mathematical Problems in Engineering**, [S.L.], v. 2020, p. 1-12, 15 maio 2020. Hindawi Limited.

[7] RUSSO, Ana et al. Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. Atmospheric Pollution Research, [S.L.], v. 6, n. 3, p. 540-549, maio 2015. Elsevier BV.

[8] JANARTHANAN, R. et al. A deep learning approach for prediction of air quality index in a metropolitan city. **Sustainable Cities and Society**, [S.L.], v. 67, p. 102720, abr. 2021. Elsevier BV.

[9] LIU, Bingchun et al. Air Pollutant Concentration Forecasting Using Long Short-Term Memory Based on Wavelet Transform and Information Gain: a case study of beijing. **Computational Intelligence and Neuroscience**, [S.L.], v. 2020, p. 1-12, 30 set. 2020. Hindawi Limited.

[10] LIU, Bingchun et al. A Novel Method for Regional NO2 Concentration Prediction Using Discrete Wavelet Transform and an LSTM Network. **Computational Intelligence and Neuroscience**, [S.L.], v. 2021, p. 1-14, 7 abr. 2021. Hindawi Limited.

[11] CABANEROS, Sheen McLean et al. Spatial estimation of outdoor NO2 levels in Central London using deep neural networks and a wavelet decomposition technique. **Ecological Modelling**, [S.L.], v. 424, p. 109017, maio 2020. Elsevier BV.

[12] EBRAHIMI-KHUSFI, Zohre et al. Predicting the ground-level pollutants concentrations and identifying the influencing factors using machine learning, wavelet transformation, and remote sensing techniques. **Atmospheric Pollution Research**, [S.L.], v. 12, n. 5, p. 101064, maio 2021. Elsevier BV.

[13] R. JUNIOR, Amilton S. et al. Assessing Recurrent and Convolutional Neural Networks for Tropospheric Ozone Forecasting in the Region of Vitória, Brazil. **Air Pollution XXVIII**, [S.L.], v. 244, p. 101-112, 27 jul. 2020. WIT Press.

[14] ALBUQUERQUE, Taciana Toledo de Almeida et al. WRF-SMOKE-CMAQ modeling system for air quality evaluation in São Paulo megacity with a 2008 experimental campaign data. **Environmental Science and Pollution Research**, [S.L.], v. 25, n. 36, p. 36555-36569, 29 out. 2018. Springer Science and Business Media LLC.

[15] GIDHAGEN, Lars et al. Experimental and model assessment of PM2.5 and BC emissions and concentrations in a Brazilian city – the Curitiba case study. **Atmospheric Chemistry and Physics Discussions**, [S.L.], p. 1-37, 10 dez. 2018. Copernicus GmbH.

[16] BORREGO, C. et al. Modelling the photochemical pollution over the metropolitan area of Porto Alegre, Brazil. **Atmospheric Environment**, [S.L.], v. 44, n. 3, p. 370-380, jan. 2010. Elsevier BV.

[17] Zucatelli, Pedro Junior et al. Nowcasting prediction of wind speed using computational intelligence and wavelet in Brazil, **International Journal for Computational Methods in Engineering Science and Mechanics**, 2020.

[18] OTTO, S.A. (2019, Jan.,7). How to normalize the RMSE [Blog post]. Retrieved from https://www.marinedatascience.co/blog/2019/01/07/normalizing-the-rmse/. Ass. in Apr 2021.