UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

ANA PAULA FERNANDES LUCIO MENEZES

PREDIÇÃO DE OZÔNIO TROPOSFÉRICO NO MONITORAMENTO DA QUALIDADE DO AR

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2021

ANA PAULA FERNANDES LUCIO MENEZES

PREDIÇÃO DE OZÔNIO TROPOSFÉRICO NO MONITORAMENTO DA QUALIDADE DO AR

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Yuri Kaszubowski Lopes

DOIS VIZINHOS
2021

ANA PAULA FERNANDES LUCIO MENEZES

PREDIÇÃO DE OZÔNIO TROPOSFÉRICO NO MONITORAMENTO DA QUALIDADE DO AR

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 06/novembro/2021

Yuri Kaszubowski Lopes
Doutorado
Universidade do Estado de Santa Catarina

Rafael Alves Paes de Oliveira
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

Rodolfo Adamshuk Silva
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Dois Vizinhos

DOIS VIZINHOS
2021

PREDICTION OF TROPOSPHERIC OZONE IN AIR QUALITY MONITORING

Ana Paula F L Menezes[1], Bruna H Daniel, Fernando Armani[2] ,Yuri K Lopes[3]

[1]Departamento Engenharia de Software, Universidade Tecnológica Federal do Paraná, Dois Vizinhos - PR - Brasil
[2]Departamento Ciências do Mar, Universidade Federal do Paraná, Curitiba - PR - Brasil.
[3]Departamento de Ciência da Computação, Universidade do Estado de Santa Catarina, Joinville - SC - Brasil.

amenezes@alunos.utfpr.edu.br, bhd.bruna@gmail.com, fernando.armani@ufpr.br, yuri.lopes@udesc.br

ABSTRACT

Resulting of the chemical reaction between primary pollutants, tropospheric ozone (O_3) causes extensive damage to human health, vegetation and animals due to its high oxidative power. Sensors for accurately measuring O_3 are expensive, difficult to handle, and require regular maintenance, which renders its availability at air quality stations difficult, unlike those used to measure the elements that compose O_3. In this paper, we use a forecasting time series method developed by the Facebook - Prophet model. The approach is to apply this model to real data collected during the period of 2018 to 2020 from an air quality station in Parana, Brazil, to predict O_3 from meteorological variables combined with as few sensors of primary pollutants as possible but aiming for accurate results.

Keywords: times series, Prophet, O_3, meteorological variables

INTRODUCTION

O_3 is considered a secondary pollutant responsible for several damages to human health, such as respiratory and cardiovascular diseases and other irritations (ALVES; SANTOS; COUTO, 2020). Techniques integrated with robust equipment can be used to monitor air pollution, aiding in policy development and research aimed at implementing air quality control strategies and encouraging environmental awareness (PENZA et al., 2014). Despite the importance of measuring O_3, the cost of specific sensors to accurately measure its concentration is high. In addition, you need: (i) an appropriate and safe space for installation due to its size and price, (ii) a place with a constant supply of stable electricity, and (iii) a professional specialized in handling the sensors (PANG et al., 2017; LIU et al., 2020). Therefore, not all air quality stations have the recommended equipment. In this paper, we describe a pipeline to predict ozone levels based on real data collected from common sensors found in air quality stations. The pipeline steps are: (i) characterize and analyze data unavailability; (ii) purge large chunks of missing data; (iii) fix short periods of unavailable data with imputation; and (iv) apply machine learning methods to obtain a predictive model. In this paper, we report the use of Facebook's Prophet to capture the shifts in the trend works with the presence of missing data.

MATERIAL AND METHOD

The available variables collected during the period of December/2016 to December/2020 are: Date/Time (date and time), O_3 (ozone tropospheric), CO (carbon monoxide), NO (nitric oxide), NO_2 (nitrogen dioxide), NO_x (nitrogen oxide), SO_2 (sulfur dioxide), CH4 (methane), NHMC (non methane hydrocarbon), THC (total hydrocarbons), PM10 (particulate matter), PTS (suspended

particulates), AT (air temperature), RH (relative humidity), BP (atmospheric pressure), SR (solar radiation), WS (wind speed), WD (wind direction) and Rain. The dataset contains a total of 660,347 records sampled hourly, being characterized as a times series. Due to an electricity outage between the first quarter of 2017 to mid- 2018, there was a large chunk of missing data that could not be treated. Therefore, only the data from June 2018 to December 2020 was considered, still, this period had a 40.8% missing data; and therefore, data imputation was required. In time series, continuous data are fundamental as a condition for its analysis. Therefore, multiple imputation techniques were used to overcome data unavailability. A set of relevant variables was identified based on its correlation with O_3. The identified set of variables is composed of RH, AT, SR, WS, added to ozone producers in the troposphere ($NOx = NO + NO_2$).

Prophet Forecast Model

Proposed in 2017 by two Facebook researchers, the model Prophet can predict time series with seasonal effects; working well in changing trends and the presence of outliers. However, it requires more than one year of historical data. This model can be decomposed as shown in Eq. (1), where *g(t)* is the trend, *s(t)* is seasonality, *h(t)* is holidays and *ξ(t)* is an error term (Taylor and Letham, 2017; Oo and Phyu, 2021):

$$y(t) = g(t) + s(t) + h(t) + \xi(t) \quad (1)$$

After the time series data from the sensors have been collected and the missing data imputed with the variables selected due to the correlation, the model can be trained with the historical dataset. The Prophet model expects at least two columns in the data frame: (i) ds -- with DateTime format and (ii) y -- with number format corresponding to the target will be predictable, O_3 in this case. The other variables must be included as regressors using specific class methods one by one. The variables selected due to the correlation with O_3 (RH, AT, SR, WS) were tested using the Prophet model, combining several situations with NO_x, NO e NO_2, and also with added delay in the regressors, to simulate the real tropospheric ozone production. The model was evaluated using Root Mean Squared Error (RMSE) over a subset of data not used in training.

CONCLUSIONS

The model was trained with the combinations of variables described in Table 1, where *t* is the current time step, and *t-d* is a delay of *d* time steps. The RMSE obtained indicates that the combination of RH, AT, SR, WS, NO e NO_2, and a delay of one timestep produced the best model.

Table 1 – Variables selected

Variables	RMSE
RH(t),AT(t),SR(t),WS(t)	4.82
RH(t),AT(t),SR(t),WS(t),NO(t)	4.57
RH(t),AT(t),SR(t),WS(t),NO(t),NO_2(t)	4.54
RH(t),AT(t),SR(t),WS(t),NOx(t)	4.57
RH(t-1),AT(t-1),SR(t-1),WS(t-1),NO(t-1),NO_2(t-1),RH(t),AT(t),SR(t),WS(t),NO(t),NO_2(t)	4.52

RH(t-1),AT(t-1),SR(t-1),WS(t-1),NOx(t-1),RH(t),AT(t),SR(t),WS(t),NOx(t)	4.57

The residual values graph shown in Figure 1 from the resulting RMSE = 4.52 allows us to state that this model is adjusted to historical data considering combinations of regression variables. The meteorological variables added to two more sensors are capable to produce adequate results, allowing cheaper air quality stations to be installed in places where this service is not yet available. The next step will be to compare the residuals of the Prophet model with others for time series models such as Sarimax.

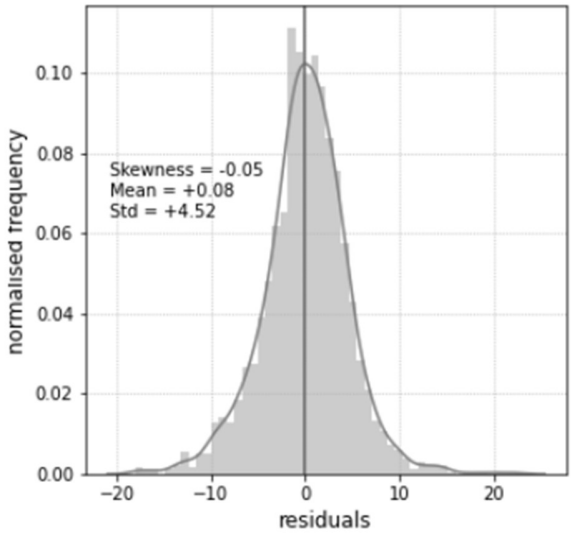

Figure 1 – Visualization of residuals distribution of model selected

REFERENCES

L. Alves, L. Santos, and E. Couto, **Distribuição das concentrações de ozônio (O_3) na área de influência do pólo industrial de Camaçari – Bahia:Prováveis impactos à saúde humana e ao meio ambiente.** Revista Brasileira de Meio Ambiente, v.8, n.1, p.113-130, 2020.

M. Penza, D. Suriano, M. G. Villani, L. Spinelle, and M. Gerboles, **Towards air quality indices in smart cities by calibrated low-cost sensors applied to networks**, in SENSORS, 2014 IEEE, 2014, pp. 2012–2017.

Meichen Liu, Karoline K. Barkjohn, Christina Norris, James J. Schauer, Junfeng Zhang, Yinping Zhang, Min Hu, and Michael Bergin. **Using low-cost sensors to monitor indoor, outdoor, and personal ozone concentrations in Beijing, China**. Environ. Sci.: Processes Impacts, 22:131–143, 2020.

Taylor SJ, Letham B. 2017. **Forecasting at scale**. PeerJ Preprints 5:e3190v2
https://doi.org/10.7287/peerj.preprints.3190v2

Xiaobing Pang, Marvin D. Shaw, Alastair C. Lewis, Lucy J. Carpenter, and Tanya Batchellier. **Electrochemical ozone sensors: A miniaturized alternative for ozone measurements in laboratory experiments and air-quality monitoring.** Sensors and Actuators B: Chemical, 240:829–837, 2017.

Z. Oo and S. Phyu, **Time Series Prediction Based on Facebook Prophet: A Case Study, Temperature Forecasting in Myintkyina**, International Journal of Applied Mathematics Electronics and Computers, vol. 8, no. 4, pp. 263-267, Dec. 2021, doi:10.18100/ijamec.816894