Forecasting Ozone and Nitrogen Oxides for Air Quality Monitoring

César Bouças, cesarboucas@dei.uc.pt
Catarina Silva, catarina@dei.uc.pt
Alberto Cardoso, alberto@dei.uc.pt
Filipe Araujo, filipius@uc.pt
Joel Arrais , jpa@dei.uc.pt
Paulo Gil, pgil@dei.uc.pt
Bernardete Ribeiro, bribeiro@dei.uc.pt

Universidade de Coimbra
CISUC - Centro de Informática e Sistemas
FCTUC-DEI - Departamento de Engenharia Informática
Portugal

Abstract

Ozone (O3) and nitrogen oxides (NOx) emissions can harm ecosystems, agriculture and public health through their direct and indirect effects on the air quality. Thus, the ability to predict future concentrations of such gases is of strategic importance, especially in the current climate changing scenario. This work presents three methods to predict O3 and NOx concentrations for the upcomming 24 hours, given a sequence of past window of the same gas concentrations as input: a moving average, a linear regression and a Long short-term memory (LSTM) network that exibited the best result, being able to forecast NOx series with an average root mean squared error (RMSE) of 115ppb and mean absolute percentage error (MAPE) of 36% with respect to the ground truth series of the test set. The presented strategy was used to empower the NanoSen-AQM air quality platform.

1 Introduction

Gas concentrations observed at a regular interval of time (step) consist in a time series that can be used to predict future observations in a process called forecasting [1]. The forecast aim is to estimate how the observations will sequence into the future. Classical models used to forecast time series include ARIMA models, decomposition models and exponential smoothing [2]. Moreover, hybrid methods demonstrated advantages combining classical models with neural networks, such as in [3] that used exponential smoothing in conjunction with a Long short-term memory (LSTM) network and reached state-of-the-art results.

In this work, three methods are used to forecast hourly averaged NOx concentrations and two methods were used to forecast hourly averaged O3 concentrations. The number of future steps predicted was set to 24 and only the gas measurements were used as input to forecast future concentrations. Making the proposed methods simple enough to enable a smooth integration in the NanoSen-AQM online platform[1] [4].

2 Proposed approach

A moving average technique and a linear regression model were used as baseline, then a LSTM model was designed to enhance the performance. We avoided using extra features and specificities of the series in order to make our methods suitable to integrate and generalize well in the online platform dynamic environment.

2.1 Moving Average and Linear Regression

Two methods were used as baseline: a simple moving average since it produces predictions with no need for training, and a ordinary least squares regression, since it counts as a machine learning solution with low computational costs allied with resonable performance.

As the input for the Linear Regression, a sequence of 72 past measurements were used to predict the upcoming 24 measurements on Devito's data. For the Badajoz data, the lenght of the input sequence was reduced to 48 due the small number of examples.

The moving average were implemented as a simple arithmetic mean of past measurements. The mean counts as a future predicted step, thus, the method is repeated until the 24 future steps are predicted. Moving the window of past measurements towards the more recent values, one step at each iteration.

[1] https://nanosenaqm.dei.uc.pt/

We experimented several window lengths to calculate the mean, and 18 was the value that led to the best balance between error metrics and visual perception of the predictions.

2.2 Long short-term memory (LSTM) network

A NOx series forecast model was designed as a neural network whereas the input x_d sequence containing 72 past measurements is first transformed by a LSTM layer with 20 neurons (units) activated by a hyperbolic tangent function. Then by another LSTM layer with 8 rectified linear units and finally by an identity layer that outputs a sequence of 24 values that corresponds to the predicted future. Figure 1 illustrates the architecture.

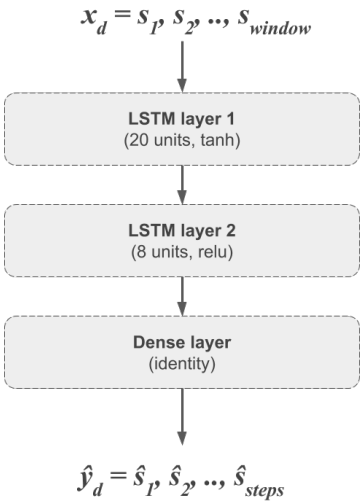

Figure 1: The neural network architecture with the number of neurons/units of each layer and its activation functions.

The network was trained through 50 epochs of backpropagation using gradient descent algorithm and mean squared error (MSE) as loss function.

The hyperparameters tuned at training were the number of hidden units and its activation functions. Whereas 20 and 8 hidden units with tanh and relu activation functions demonstrated to be sufficient to reach the best average results at training phase.

3 Experimental setup

3.1 Dataset

Series from two datasets were used to develop and test the proposed methods. As main source, averaged Nitrogen Oxides (NOx) concentrations recorded from March 2004 to February 2005 in Italy were used. This data is part of the Air Quality Data Set (Devito) [5] that is publicly available.

Ozone concentrations measured with reference sensors at Extremadura University campus (Badajoz) from September 21 to September 25 of 2017 were also used to deal with low data availability under the NanoSen-AQM data.

Method	RMSE (ppb)	MAPE (%)
LSTM	115.21	36
Linear Regression	131.34	41
Moving Average	215.86	84

Table 1: Results over Devito's NOx test set.

Method	RMSE (ppb)	MAPE (%)
Linear Regression	22.91	38
Moving Average	44.98	57

Table 2: Results over Badajoz's O3 test set.

3.2 Evaluation metrics

Given the the original ground truth series with the expected output ($y = s_1, s_2, ..., s_n$) and the series predicted by the model ($\hat{y} = \hat{s}_1, \hat{s}_2, ..., \hat{s}_n$). The RMSE measures the root average of the squares of the errors and its calculated as:

$$RMSE(y,\hat{y}) = \sqrt{MSE(y,\hat{y})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(s_i - \hat{s}_i)^2} \quad (1)$$

The MAPE is calculated as a percentage:

$$MAPE(y,\hat{y}) = \frac{100}{n}\sum_{i=1}^{n}\frac{|s_i - \hat{s}_i|}{s_i} \quad (2)$$

3.3 Preprocessing

The gas concentration time series can be defined as a sequence of values v_i as such $D = (v_i)_{i=1..|D|}$. After removing empty rows and filling missing values with the last valid observation, the original values were rescaled since machine learning models tend to behave better when feature values are in a limited range near zero:

$$s_i = \frac{v_i - min(D)}{max(D) - min(D)} \quad (3)$$

In order to train a forecast model using supervised learning, we need to derivate from D, a new set D', consisting of pairs $(x_d, y_d)_{d=0..|D'|}$ where y_d is the expected output for an input x_d.

Having defined a constant *window* that is the number of steps used as input features. And a constant *steps* that is the number of future steps to predict:

$$(x_d, y_d) = ((s_i)_{i=1..window}, (s_j)_{j=window+i+1..window+i+1+steps}) \quad (4)$$

After derivating D' we splited it into train and test sets. The first was the major portion of the examples (75%) and was used to train the machine learning based methods. While the latter was left untouched and was used only to evaluate the models at the test phase.

4 Results

For each example in the test set, the trained models were used to make a prediction as well as the moving average was calculated. After doing this process over all the set, the evaluation metrics were calculated using the set of predicted values ($\hat{y}$) and the ground truth values (y) of the test set. Obtaining the final average errors for each test set: Badajoz and Devito.

Tables 1 and 2 sumarize the obtained metrics for the Devito and Badajoz test sets respectively. In the Badajoz case, the number of examples were insuficient to train the LSTM model.

Figure 2 illustrates an example from the Devito test set and the predicted outputs for this example. Offering a visual perception of the input, expected output and predictions of each method.

The results demonstrated that all the methods should be improved, expecially the moving average, which presented much worse metrics than the others methods despite the sufficient visual perception of it's predictions.

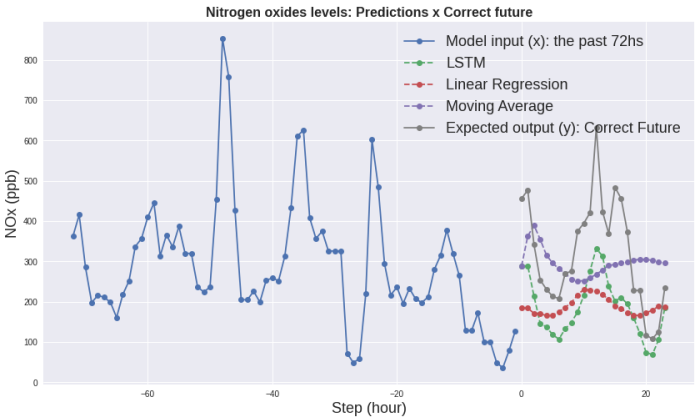

Figure 2: The predictions of the three methods for an example from the Devito's test set.

5 Conclusion and further work

The proposed methods were developed without using hand crafted features or series manipulation. Them demonstrated resonable performances, and were sucessfuly integrated into the NanoSenAQM online platform, where it is expected some generalization potential without requiring human intervention. Furthermore, the baselines showed to be attractive for it's simplicity and low memory consumption.

Results suggests that the seasonality of the series harm the performances, expecially of the moving average. Methods for automatic seasonality removal should be considered instead of classic manual removal methods, since the latter would not be suitable to be implemented as part of the online platform.

Besides removing the sasonality of the series, performance improvements can be reached by developing an exponential smoothing strategy within the LSTM such as [3]. The linear regression models might be benefited from exaustive hyperparameters search. Also, the moving average can be extended to a exponential implementation, giving greater importance to recent measurements in the inputs.

Finally, the use of informative features about temperature, humidity, wind and other factors that have impact in such gases behaviors could benefit the Linear Regression and the LSTM methods.

Acknowledgements

We acknowledge the Program Interreg-Sudoe of the European Union under grant agreement SOE2/P1/E0569 (NanoSen-AQM) and funding from the FCT, I.P., within CISUC Project UID/CEC/00326/2019 and CTS-UID/EEA/00066/2019.

References

[1] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[2] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020.

[3] Slawek Smyl. A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, 36(1):75–85, 2020.

[4] Pedro Henrique Saraiva Lucas et al. *Development of the server for the NanoSen-AQM Project*. PhD thesis, Universidade de Coimbra, 2019.

[5] Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757, 2008.

*/