HOURLY TROPOSPHERIC OZONE CONCENTRATION FORECASTING USING DEEP LEARNING

LUCAS ALVES, ERICK GIOVANI SPERANDIO NASCIMENTO & DAVIDSON MARTINS MOREIRA
Senai CIMATEC, Brazil

ABSTRACT

The purpose of this work is to build, train and evaluate a deep learning-based model to forecast tropospheric ozone levels hourly, up to twenty-four hours ahead, using data gathered from the automatic air quality monitoring system in the metropolitan region of Vitória city, Espírito Santo (ES), Brazil. Observational data of air pollutant concentrations and meteorological parameters were used as the input variables of the model once they represented the state of the atmospheric fluid in terms of its properties and chemical composition throughout the time. Several topologies of multilayer perceptron neural networks were tried and evaluated using statistics of the predictions over unseen data. The best architecture was compared with reference models and the results showed that deep learning models can be successfully applied to hourly forecasting of ozone concentrations for urban areas. Once such models are fitted to the data, the forecasting procedure has a very low computational cost, meaning that it can be used as an alternative approach in comparison with numerical modelling systems, which require much more computational power.

Keywords: air quality forecasting, ozone, neural networks, deep learning.

1 INTRODUCTION

Ozone (O_3) is a secondary pollutant in the troposphere and one of the photochemical oxidants causing air quality problems. It is formed from chemical reactions between gases emitted by natural and anthropogenic sources, such as nitrogen oxides and volatile organic compounds in the presence of solar radiation. O_3 can irritate the respiratory system, reduce lung capacity, and aggravate asthma problems [1]. Moreover, it can damage plants and affect agricultural production [2]. The World Health Organization (WHO) air quality guidelines provide thresholds for health-harmful pollution levels and the 2005 publication sets the recommended value for ozone concentration at 100 μg/m³ for a daily maximum 8-hour average [3]. Therefore, it is important to develop a powerful forecasting model that could help authorities and the population to take preventive measures and avoid imminent health risks, even before the recommended limits are reached.

2 DATA

The data used in the experiments is publicly available and was gathered from the Automatic Air Quality Monitoring Network (*Rede Automática de Monitoramento da Qualidade do Ar – RAMQAr*) owned by the State Institute of Environment and Water Resources of Espírito Santo (*Instituto Estadual de Meio Ambiente e Recursos Hídricos do Espírito Santo – IEMA-ES*). The monitoring station chosen for this study is located in Cariacica, a city in the metropolitan region of Vitória, ES, Brazil. This station measures hourly averages of the twelve atmospheric pollutant concentrations and meteorological parameters displayed on Table 1.

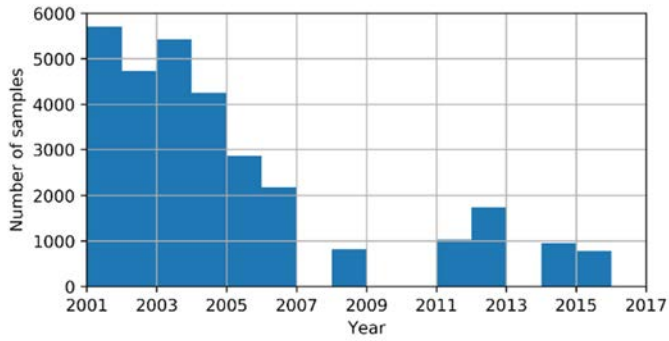
Data from the years 2001 to 2015 were collected and treated to eliminate records with invalid or missing measurements in one or more sensors. Furthermore, only valid data and hourly sequences with at least twenty-four consecutive samples were kept in order to make possible the generation of ozone concentration ground truth targets, required on supervised machine learning algorithms for the model fitting. These steps discarded most part of the

WIT Transactions on Ecology and the Environment, Vol 236, © 2019 WIT Press
www.witpress.com, ISSN 1743-3541 (on-line)
doi:10.2495/AIR190131

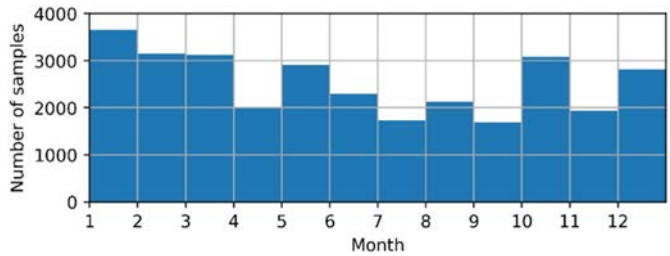
data, including all the measurements from the years 2007, 2009, 2010 and 2013, resulting in a data set with 30,492 samples and their corresponding targets. Fig. 1 shows the amount of useful data by year and month.

Table 1: Parameters measured in Cariacica's air quality monitoring station.

Parameter	Parameter characteristics	
	Type	Unit
Particulate matter below 10μm (PM_{10})	Atmospheric pollutant	(μg/m³)
Total suspended particulate matter	Atmospheric pollutant	(μg/m³)
Sulphur dioxide (SO_2)	Atmospheric pollutant	(μg/m³)
Nitrogen monoxide (NO)	Atmospheric pollutant	(μg/m³)
Nitrogen dioxide (NO_2)	Atmospheric pollutant	(μg/m³)
Nitrogen oxides (NO_x)	Atmospheric pollutant	(μg/m³)
Carbon monoxide (CO)	Atmospheric pollutant	(μg/m³)
Ozone (O_3)	Atmospheric pollutant	(μg/m³)
Temperature	Meteorological	(ºC)
Humidity	Meteorological	(%)
Scalar wind direction	Meteorological	(º)
Scalar wind speed	Meteorological	(m/s)



(a)



(b)

Figure 1: Histograms showing the distribution of useful data by time period. (a) Number of samples by year; and (b) Number of samples by month.

In the machine learning domain, one of the main objectives is to create computational models with the ability to generalize well the extracted attributes to new data. Poor generalization is often characterized by overfitting, and a common method to avoid that is to evaluate a model by splitting a data set into two. The first one is the training set, on which the model is built and optimized. The second is the test set, on which the finished model is evaluated with unseen data [4].

For this research, 22,987 data points from the years 2001 to 2005 were separated for the training and validation set, and the remaining 7,505 data points from the years 2006, 2008, 2011, 2012, 2014 and 2015 were used as test data. This procedure was made in order to guarantee that the two batches of data have samples for all months, days of a month, days of a week and hours of a day.

3 METHODS AND MODELLING

The modelling of ozone fluctuations can be made through two types of models: deterministic or stochastic. Deterministic models use several equations to represent the atmosphere behaviour and thus forecast the ozone concentrations in a limited domain. Due to the complexity of this process, developing and maintaining them are expensive tasks and demands a large amount of computational power, since it has to process many chemical and physical interactions between diverse parameters like emissions, meteorology and land cover. Stochastic models, otherwise, have a simpler implementation because they try to formulate a mathematical relationship between the input and output variables based on the detection of some patterns [5]. Once such models are fitted to the data, the predictions are made using few computational resources.

Artificial neural networks are one type of stochastic models and, in the deep learning field, there are currently many different architectures available for implementation, being essential to examine which one best fits the problem that needs to be solved. Previous works used recurrent neural networks (RNN) to predict daily maximum concentrations of tropospheric ozone in the city of Palermo, Italy [6], and in the Mexicali (Mexico)-Calexico (USA) border area [7]. In Biancofiore et al. [8], RNN models were applied to predict O_3 concentration at time $t+\Delta t$, where Δt can be 1, 3, 6, 12, 24 and 48 h. A convolutional neural network (CNN) was employed in Eslami et al. [9] to predict the hourly ozone concentration on each day using parameters from the previous day. Eight separated multilayer perceptron (MLP) networks were used in Agirre et al. [10] to forecast the values of the variables $O_3(t+k)$, being $k = 1, 2, \ldots, 8$ h, at two rural stations located in the Autonomous Community of the Basque Country (North Central Spain). An MLP predictor was built in Tamas et al. [11] using one single output to forecast O_3 concentration 24 hours ahead in Corsica, France, in order to be able to anticipate pollution peaks formation. In Coman et al. [5] two MLP models were evaluated. The "dynamic" model used a cascade of 24 multilayer perceptrons arranged so that each MLP feeds the next one, and the "static" model was a classical single MLP with 24 outputs. For both configurations, the outputs were ozone concentrations for a 24 h horizon.

The present research focus on multilayer perceptron neural network models, due to its simplicity and large application for short-term forecasts. These networks have universal function approximation capabilities, with applicability in non-linear problems and ability to generalize to unseen data, being effective for prediction purposes [12]. However, prior studies that used this type of network to forecast hourly concentrations of ozone aim attention on MLPs with one single hidden layer, which can lead to models with limited representational power. Moreover, few of them uses a single model to forecast hourly ozone concentrations for all time steps ahead in a 24 h horizon. The proposed model employs this approach, since it can take advantage of a shared internal representation for all the forecasts and obtain a

better generalization of the problem. Additionally, in the performed simulations the best results were achieved with deeper network topologies.

Multilayer perceptron networks have a flexible topology and among their main parameters are the number of layers and the number of neurons in each layer. At least three layers are required: an input layer, a hidden layer, and an output layer. The definition of the number of layers and neurons is variable, and the best composition is problem-specific [13].

Since the model objective is to predict tropospheric ozone levels hourly, up to twenty-four hours ahead, the output layer of the proposed model is composed by twenty-four neurons, one for each hour in advance. Several network designs were tested varying the number of inputs, hidden layers and nodes in each hidden layer. Besides, different nonlinear activation functions were experimented on nodes in the hidden layers, keeping the output neurons with linear activation function. In order to choose a good set of parameters for the training procedure, some optimization algorithms and values of learning rate, batch size and L2 regularization strength were tried.

The generated models were evaluated using the training data set with a 5-fold cross-validation. In this type of cross-validation, the data set is divided into parts of the same size. One part forms the validation set and the other parts form the training set. This process is repeated for each part of the data, and the combination of tests is used to make a reliable estimate of the model error [4].

The models' performances were measured based on statistics such as the mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), Pearson's correlation coefficient (r), and regression coefficient (R^2). These metrics are described in the following equations:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|O_i - F_i|, \tag{1}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(O_i - F_i)^2, \tag{2}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{O_i - F_i}{O_i}\right|, \tag{3}$$

$$r = \frac{\sum_{i=1}^{n}(O_i - \overline{O})(F_i - \overline{F})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2}\sqrt{\sum_{i=1}^{n}(F_i - \overline{F})^2}}, \tag{4}$$

$$R^2 = \frac{\left(\sum_{i=1}^{n}(O_i - \overline{O})(F_i - \overline{F})\right)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2 \sum_{i=1}^{n}(F_i - \overline{F})^2}, \tag{5}$$

where O_i is the observed value, $\overline{O}$ is the mean of all observed values, F_i is the forecasted value, $\overline{F}$ is the mean of all forecasted values and n is the number of samples.

Table 2 summarizes the topology and characteristics of the MLP neural network that achieved the best results in the metrics evaluated. The training of this network used the Feed-forward Backpropagation algorithm with Adadelta optimizer, learning rate of 1.0, batch size of 4.0, L2 regularization strength of 2.10^{-5} and mean squared error as the loss function.

3.1 Data preparation

The best obtained model has eighteen inputs, composed by six temporal variables and by the twelve parameters measured in the air quality station. The wind representation was converted

Table 2: Multilayer perceptron neural network chosen topology.

Layers	Topology		
	Neurons	Activation function	Trainable parameters
Input	18	N/A	0
1st Hidden layer	15	ReLU	285
2nd Hidden layer	185	ReLU	2,960
3rd Hidden layer	250	ReLU	46,500
4th Hidden layer	200	ReLU	50,200
5th Hidden layer	225	ReLU	45,225
6th Hidden layer	185	ReLU	41,810
Output	24	Linear	4,464

from scalar values of direction and speed to vector components U and V, using eqns (6) and (7). This transformation was made because the scalar representation could mislead the model, since direction values close to 0º or to 360º indicate that the wind is blowing in the same direction, although these values are numerically distant

$$U = speed * \sin((270 - direction)(\frac{\pi}{180})) , \quad (6)$$

$$V = speed * \cos((270 - direction)(\frac{\pi}{180})) . \quad (7)$$

As indicated by other works [10], [11], the use of periodical variables, as sine and cosine functions, representing the time cycles, lead to better results in predict ozone concentrations. Therefore, the following temporal inputs were used:

- sin (2π (60 h + m)/1440), cos (2π (60 h + m)/1440), where h = 0, 1, 2, …, 23 is the hour of the day and m = 0, 1, 2, …, 59 is the minute of the hour;
- sin (2πd/7), cos (2πd/7), where d = 0, 1, 2, …, 6 is the day of the week with 0 representing Sunday and 6 representing Saturday;
- sin (2πy/12), cos (2πy/12), where y = 1, 2, …, 12 is the month of the year.

Before model fitting, all inputs and their ground truth targets were normalized between -1 and 1. This procedure changes the data to a common scale, avoiding that one input have excessive importance in consequence of its value range [11]. Once the output of the model is obtained, the variables are de-normalized.

3.2 Reference models

To measure the efficiency of the proposed neural network, two models were used for reference. The first one is called Persistence model and is commonly used as a baseline to evaluate the performance of a forecasting model. In this predictor, the forecasts for all time steps ahead are set as the current value, which can be expressed mathematically as $y(t+\Delta t) = y(t)$, where y is the forecast target and t is time [14].

The second reference model is composed by a group of twenty-four linear regressors, each one responsible to predict a different time step of the next 24 hours of ozone concentration. The regressors are based on multiple linear regression and were fitted using the same inputs and targets as the MLP model. Regularization of L1 type was used in the models training,

technique also known as Lasso Regression [15]. Several regularization parameters, which defines the regularization strength, were evaluated using a 5-fold cross-validation over the training data set and the parameter that produced the best performance for each regressor was chosen.

3.3 Computational tools

All computational experiments were implemented using the *Python* language. Experiments with MLP neural networks were performed using the *TensorFlow* platform through its *Keras* high-level API (Application Programming Interface), and the *Scikit-learn* machine learning library was used to evaluate the linear regression model.

4 RESULTS

A comparison between the metrics of evaluated ozone forecasting models, using the test data set, is shown in Table 3. Values close to 0.0 are best for the MAE and MSE, values close to 0.0% are satisfactory for the MAPE, and values close to 1.0 are adequate for the R^2 and r. A Pearson's correlation coefficient of −1.0 implies a negative linear correlation between the forecasted and the ground truth values, and a value of 0.0 implies that there is no linear correlation between these variables [14].

Table 3: Comparison table between ozone forecasting models using performance metrics over the test data set.

Model	Performance metrics				
	MSE	MAE	r	R^2	MAPE
Persistence	295.19	12.90	0.384	0.148	113.08
Lasso	126.02	8.73	0.692	0.479	90.10
MLP	101.75	7.68	0.770	0.593	70.55

The values of the metrics introduced on Table 3 refers to the models as a whole, considering all the 24 predictions at the same time. Thus, the proposed multilayer perceptron is a very effective model, surpassing the reference models in all considered metrics. The persistence model obtained the worst performance, as expected due to its simplicity. Tables 4 and 5 presents the statistics for some forecasting horizons of the MLP model and of the Lasso linear model, respectively.

With exception of the predictions for the first hour ahead, where the Lasso model has slightly better results, the neural network outperforms the linear model in all other time horizons. This is shown on Fig. 2 using as reference the mean squared error and the Pearson's correlation coefficient (r). Besides, the prediction errors of the MLP have a more stable behaviour along the forecasted time steps.

The model accuracy is graphically shown in Figs 3 and 4, where ozone concentration forecasts are displayed for 1 hour ahead and 24 hours ahead in a period of approximately 7 days of measurements. The blue lines represent the actual ozone concentrations and the orange lines represent predictions of ozone concentrations. Fig. 3 demonstrates a period in the beginning of June 2006, which represents the end of autumn season in the south hemisphere. On the other hand, the Fig. 4 display a period in the end of December 2014 and beginning of January 2015, which represents the start of summer season.

Table 4: Multilayer perceptron performance metrics over the test data set for some forecasting horizons.

Forecast horizon	Multilayer perceptron performance metrics				
	MSE	MAE	r	R^2	MAPE
T+1	42.90	4.94	0.910	0.829	41.14
T+2	63.37	5.93	0.863	0.745	51.10
T+3	77.46	6.61	0.829	0.688	57.65
T+6	99.42	7.56	0.774	0.599	67.39
T+9	110.52	8.01	0.746	0.557	72.44
T+12	114.09	8.19	0.741	0.549	75.13
T+15	113.53	8.23	0.742	0.551	76.20
T+18	111.69	8.16	0.748	0.559	75.81
T+21	106.72	7.97	0.757	0.574	75.84
T+24	101.50	7.77	0.768	0.590	76.73

Table 5: Lasso linear model performance metrics over the test data set for some forecasting horizons.

Forecast horizon	Lasso linear model performance metrics				
	MSE	MAE	r	R^2	MAPE
T+1	34.90	4.26	0.924	0.855	34.23
T+2	74.28	6.54	0.831	0.691	61.31
T+3	103.40	7.94	0.755	0.571	79.04
T+6	141.71	9.43	0.641	0.411	98.89
T+9	138.04	9.23	0.653	0.426	93.60
T+12	125.95	8.80	0.695	0.483	88.08
T+15	137.99	9.25	0.656	0.431	97.88
T+18	147.52	9.58	0.626	0.392	102.67
T+21	134.72	9.12	0.669	0.448	98.68
T+24	105.80	7.92	0.755	0.570	78.69

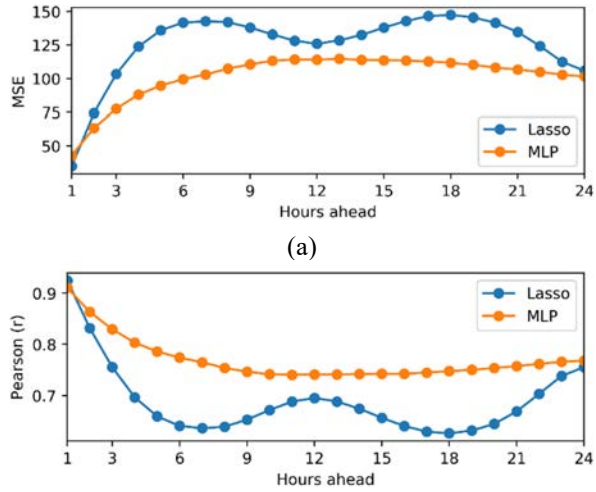


Figure 2: Comparison between MLP and Lasso predictions for each hour ahead. (a) Using mean squared error; and (b) Using Pearson's correlation coefficient.

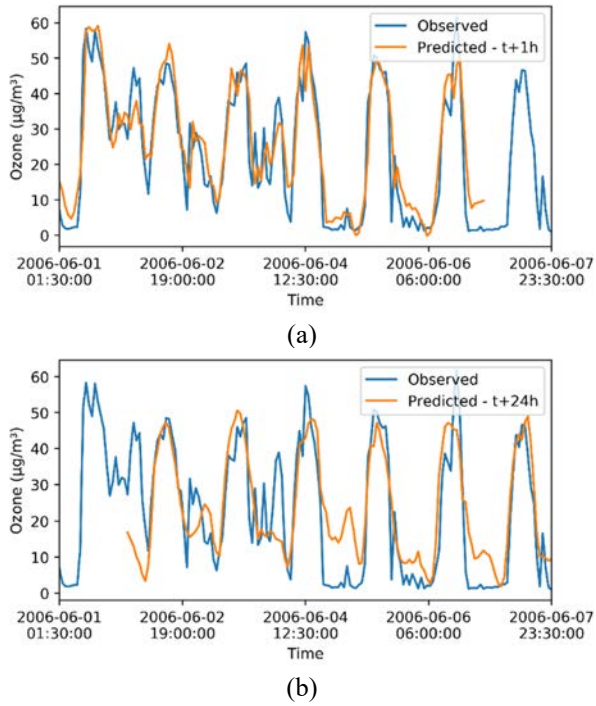
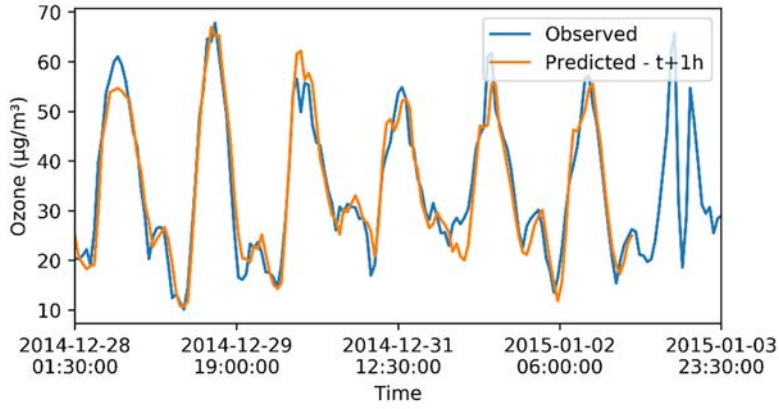
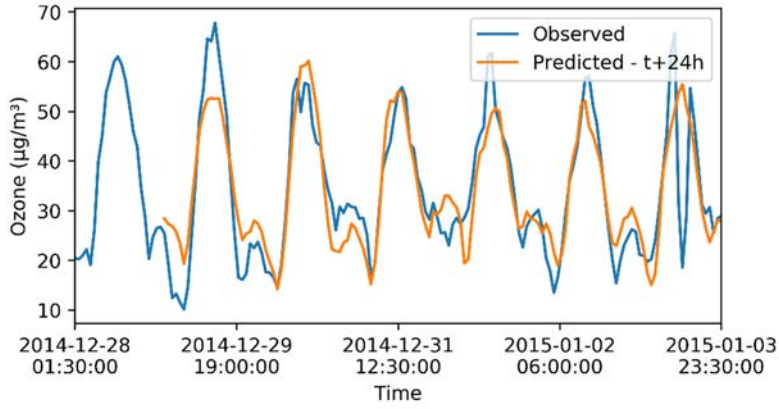


Figure 3: Multilayer perceptron predictions for a seven-day period in June 2006. (a) Predictions 1 hour ahead; and (b) Predictions 24 hours ahead.



(a)



(b)

Figure 4: Multilayer perceptron predictions for a seven-day period between December 2014 and January 2015. (a) Predictions 1 hour ahead; and (b) Predictions 24 hours ahead.

5 CONCLUSIONS

The results indicated a reasonable performance for the proposed forecasting model, which can be used by authorities and citizens to take preventive measures that avoid imminent health risks due to O_3 exposure. Moreover, it has been shown that deep learning techniques can be successfully applied to hourly forecasting of ozone concentrations in urban areas. Once such models are trained and fit to the data, the inference process, i.e. the forecasting procedure, has a very low computational cost, meaning that it can be used as an alternative approach in comparison with numerical modelling systems, which require much more computational power.

ACKNOWLEDGEMENTS

This work was partially supported by the Bahia State Research Support Foundation (Fundação de Amparo à Pesquisa do Estado da Bahia - FAPESB), in the form of a Development of Technological Innovation grant for L. Alves in the project “Implantation of

Infrastructure Research in Simulation and Computational Modelling in the State of Bahia using High Performance Processing", developed at Senai CIMATEC Supercomputing Center for Industrial Innovation.

REFERENCES

[1] Filippidou, E.C. & Koukouliata, A., Ozone effects on the respiratory system. *Progress in Health Science*, **1**, pp. 144–155, 2011.

[2] Emberson, L.D. et al., Ozone effects on crops and consideration in crop models. *European Journal of Agronomy*, **100**, pp. 19–34, 2018.

[3] World Health Organization, Occupational and Environmental Health Team, *Air Quality Guidelines for Particulate Matter, Ozone, Nitrogen Dioxide, and Sulfur Dioxide: Global Update 2005: Summary of Risk Assessment*, World Health Organization, 2006.

[4] Reitermanová, Z., Data splitting. *WDS'10 Proceedings of Contributed Papers*, Part 1, pp. 31–36, 2010.

[5] Coman, A., Ionescu, A. & Candau, Y., Hourly ozone prediction for a 24-h horizon using neural networks. *Environmental Modelling and Software,* **23**(12), pp. 1407–1421, 2008.

[6] Brunelli, U., Piazza, V., Pignato, L., Sorbello, F. & Vitabile, S., Two-days ahead prediction of daily maximum concentrations of SO_2, O_3, PM_{10}, NO_2, CO in the urban area of Palermo, Italy. *Atmospheric Environment*, **41**(14), pp. 2967–2995, 2007.

[7] Salazar-Ruiz, E., Ordieres, J.B., Vergara, E.P. & Capuz-Rizo, S.F., Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US). *Environmental Modelling & Software*, **23**(8), pp. 1056–1069, 2008.

[8] Biancofiore, F. et al., Analysis of surface ozone using a recurrent neural network. *Science of the Total Environment*, **514**, pp. 379–387, 2015.

[9] Eslami, E., Choi, Y., Lops, Y. & Sayeed, A., A real-time hourly ozone prediction system using deep convolutional neural network. arXiv preprint arXiv:1901.11079, 2019.

[10] Agirre, E., Anta, A., Barron, L.J. & Albizu, M., A neural network based model to forecast hourly ozone levels in rural areas in the Basque Country. *WIT Transactions on Ecology and the Environment*, vol. 101, WIT Press: Southampton and Boston, pp. 109–118, 2007.

[11] Tamas, W., Notton, G., Paoli, C., Voyant, C., Nivet, M.L. & Balu, A., Urban ozone concentration forecasting with artificial neural network in Corsica. *Mathematical Modelling in Civil Engineering*, **10**(1), pp. 29–37, 2014.

[12] Agirre, E., Anta, A. & Barron, L.J., Forecasting ozone levels using artificial neural networks. *Forecasting Models: Methods & Applications*, ed. J. Zhu, iConcept Press, pp. 208–218, 2010.

[13] Russell, S. & Norvig, P., *Artificial Intelligence: A Modern Approach*, 3rd ed., Pearson Education Inc., 2010.

[14] Zucatelli, P.J. et al., Short-term wind speed forecasting in Uruguay using computational intelligence. *Heliyon*, **5**(5), 2019.

[15] Tibshirani, R., Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), pp. 267–288, 1996.