**Machine Learning Models for the Prediction of the Weekly Peak of PM2.5 Concentration in Algiers, Algeria**

Sabri ghazi, Dib Ahmed and Mendjel Mehdi.

**Abstract**:

This paper aims to assess the performance of machine learning models to predict the weekly peak of PM2.5 concentration in Algiers. Two machine-learning paradigms are used:Ensembles Model, specifically XGBoost, LightGBM, CatBoost, Random Forest, and Deep learning models such as LSTM and NeuralProphet. The models are trained using dataset including three years of daily climatic parameters and measures of PM2.5 concentration. Random Forest model showed the best performances followed by LightGBM and CatBoost. NeuralProphet helped to detect the weekly and yearly seasonality of the pollutant concentration. Climatic parameters such as precipitation and temperature have a significant impact on PM2.5 concentration. Due to the weekly seasonality of the road traffic, the size of the lags values windows helps improving the performances of the models, specifically when it is a multiple of 7 days. Therefore, this confirms that the road traffic contributes the most in degrading the air quality and causing PM2.5 peaks at Algiers. With an RMSE (Root Mean Squared Error) 0.51 and R2 of 0.99, the proposed model outperformed the state of the art models, specifically those designed for prediction PM2.5 at Algiers.

**Keywords**: PM2.5, Air pollution, Ensemble Learning, Deep learning, Time series forecasting, Algiers, air pollution prediction.

**Introduction**

The degradation in air quality has emerged as a major challenge facing many cities in the world. In developing country, uncontrolled urban expansion, fossil energy based transportation and the lack of legislation to enforce air quality standards; lead to a very alarming air pollution levels. PM2.5 (Particulate Matter with aerodynamic diameter less than 2.5 micrometre) is a mixture of solid and liquid substances, mainly generated by anthropogenic activities. Combustion engine, construction, industrial process and agriculture are among the main source of PM2.5. (Bouhila et al. 2015) studied the heavy metal content of PM2.5 in Algiers, it determined that Fe and Sc are highly presents and concluded that the annual level of PM2.5 in Algiers is beyond the local and international standards. Same conclusion is confirmed in (Talbi,Kerchich 2018), in which authors analysed measured of PM1, PM2.5 and PM10 from two stations in Algiers during 2015 and 2016 in an urban and roadside. By inspecting the samples of PM2.5, the concentrations of heavy metals were determined, with Pb representing 5%. Due to its diameter and toxicity, PM2.5 can be inhaled by human been and leads to a serious public health problem (Ladji et al. 2014). Therefore, having an accurate prediction of PM2.5 peaks period can help decision-makers in managing the crisis and reduce its effects.

The remaining of this paper is organised as follow. We start by describing some related works, in which the forecasting of PM2.5 concentration is studied in different cities, with focalization about Algiers and cities with similar geographic and climatic conditions. Section 2 describes the studied region and the collected dataset. The tested models are presented in section 3, and their performances are presented in section 4. At the end, we present some conclusions and further works.

**Related works**

Many aproches have been used to predict PM concentration, we can categories them in two main categories: Linear models, Machine Learning based models and hybrid models. Also we can categories them according the used inputs, the horizon of the prediction or the studied region.

(Chellali et al. 2016) presented an MLP (Multi-Layred Perceptron) model to predict 24 hours in advance the concentration of PM10 at Algiers. The presented model was trained using two years dataset of PM10 concentration and meteorological parameters (wind speed, relative humidity and temperature), those parameters are selected using correlation with PM2.5 concentration.

(Ibrir et al. 2021) used a SVM ( Support Vector Machine) model to predict the concentration of PM of different size including PM2.5. To select the best model hyperparameters, authors used a SWARM algorithm called Dragonfly. Compared to state of the art models the presented models showed relatively convenient performances; however, the used data set is about only 4 months and it does not include the yearly seasonality pattern.

In order to avoid the sudden changes effects, ( Liou et al. 2019) used an unsupervised method to cluster anthropoginic and environmental events. As described, the uninspected event such as ranfall intensity, wind speed and road traffic have an impact on the concentration of PM2.5. The events data is collected from the error in forecast of an Adaptive Iterative Forecast model. The model used 5 min measuring rate data from Taichung ,Taiwan.

In (Gao & Li, 2021) 45 stations in Gansu, China are modelled as weighted graph with LSTM nodes each one. The weight in edge between two stations is used to include output of each node in the LSTM input of the other station. The model is able to forecast PM2.5 cocentration in every station without the need to build a model fore each station. According to the study, the model took in consideration the spaciotemporal information, and thus performed better than ensemble learning model, using the same dataset.

(Ma et al. 2020) used baysian optimazation in order to determine the hyperparamters values of a fully connected LSTM model. The model used lagged values of inputs including the weather parameters. Compared with other models using the same dataset, the model showed the best performances. However, the used data to validate the model was randomselly selected, in time series data this could lead to poorly explicative data, since it lack the time order of each observation.

An LSTM model is presented in(Zhang et al. 2020), authors used an Auto-Encoder in order to compress the feature space before passing it as input to the LSTM. The proposed model receives as input lag values of PM2.5, snow, precipitation, ambeint temperature, wind speed and direction. It was compared to classic models such as CAMx, CMAQ and other deep learning models. The proposed model showed the best performances. Authors argue that, for a long term prediction the model trained using only PM2.5 performed better then the one who includes weather paramerters, nevertheless, for small prediction horizon the models with those parameters showed better accuracy.

(Pak et al. 2020) Mutual Information estimator is used to compute the correlation between 384 station acros China, with as target the station located in Beijing. Authors argue that this help capturing the spatiotempral information.

Table XX, present a resumé of the literature review. where, weather paramters: Wind Speed, Wind Direction, Relative Humidity, Pressure, Ambian temperature and Cumulative precipitation are noted respectevelly ***WS, WD, RH, Pr, T and P***. Anthropoginic event data is noted **A.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Area and Period** | **Prediction horizon** | **Model** | **Features selection method & hyperparameters** | **Lag values** | **Inputs** | **Multi /single output** |
| (Chellali et al. 2016) | Algiers, 2015 | 24 h | MLP | Correlation | - | PM2.5,WS, RH, T | single |
| (Ibrir et al. 2021) | Algiers, 4 months | Not mentioned | SVM | Correlection, Dragonfly | - | PM2.5,WS, RH, T, P | Single |
| (Liou et al. 2019) | Taichung ,Taiwan, 2017 | 3 h | AIF | Hirarchical Clustring | - | WS, RH,T, P, A | Single |
| (Gao & Li, 2021) | Gansu, China, 2019-2020 | From 1h to 48h | G-LSTM | Adjacenty Matrix | 4 h | PM2.5, WS, WD, RH, T, P, Pr, CO,NO2, O3,SO2 PM10, PM2.5 | Both |
| (Ma et al. 2020) | Wayne ,Michigan, USA | - | Lag-FLSTM | Baysian optimization | 48 h | PM2.5, WS, WD, Press, T, CO, SO2, NO2, PM10 | Single |
| (Zhang et al. 2020) | Beijing, China | From 24h to 168h | AE-Bi-LSTM | Auto-Encoder | - | PM2.5, WS, P, Snow, T, Dewpoint | Single |
| (Pak et al. 2020) | Beijing, China, 3 years 2015-2017 |  | CNN-LSTM | Mutual Information estimator |  | CO, SO2, O3, NO2, PM2.5, PM10, T, WD, WS, |  |

**Material and methods**

**Studied Region and Period**

Algiers is located in the centre of the North of Algeria, it is a coastal city, bordered by the Mediterranean Sea on the North and surrounded by XXX. It is the economic and politic capital of Algeria. According to (ONS), in 2021 the estimated population was 8 million habitants. The city has a high economic attraction, it hosts many central administration of the country and many international corporations headquarter. Its public transportation did not expanded proportionally to its demography, therefore the transportation is mainly based on personal cars, according to public sources (ONS) Algiers’s motor fleet reached 2 million in 2019.

**Data description**

Our dataset covers 3 years of daily measures of climatic parameters and PM2.5 concentration. The measures are collected by EPA US-EMBACY station, which is located in XXXX, the climatic parameters are provided by official meteorology agency (ONM). Table XX, describes the statistic properties of the dataset.

In order to avoid sudden event effect ( Liou et al. 2019), we decided to include lagged values of the four climatic parameters. This will help our model to take in consideration the sudden changes.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **MAX\_TEMPERATUR (C°)** | **WINDSPEED\_MAX (km/h)** | **PRECIP\_TOTAL DAY (mm)** | **HUMIDITY MAX PERCENT** | **PM2.5 (µ gram/ m3)** |
| mean | 23.24 | 16.46 | 1.79 | 63.45 | 67.96 |
| Std | 5.97 | 6.751 | 4.34 | 12.77 | 14.49 |
| Min | 10.0 | 4.0 | 0.0 | 34.0 | 40.0 |
| Max | 41.0 | 44.0 | 35.0 | 94.0 | 172.0 |

**Correlation**

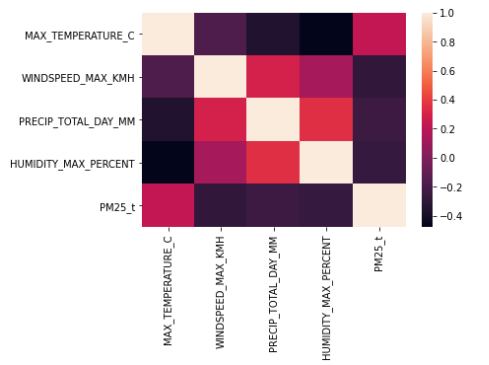
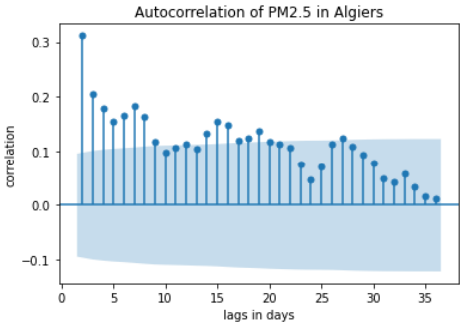


Figure –XX- Heat map correlation matrix in our dataset.

**Auto-correlation** : in figure XX, we can observe the autocorrelation, which measure the correlation betweent the lagged values and the current values. We varied the lagged values from 1 day to 35 days. As described, we can see the correlation between the value of the PM2.5 times series and the multiple of 7.



**Models**

Decision tree is a machine-learning model, which build a tree using the values of the features. First, it selects the feature that splits the training sample and build a decision node, and recursively build sub-tree. The selection of feature is done using GINI impurity metric, which calculates how well a feature split the samples.

Random Forest is an ensemble learning method; it uses decision tree models and combine their outputs. Many decision tree models are trained using random samples of the training data and random subsets of the features. XGBoost developed by Chen and Guestrin in [Chen and guestring, 2016] is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting steroid strategy. It applies the principle of boosting “weak learners” and provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way. LightGBM is a recent improvement of the gradient boosting algorithm released by Microsoft in 2017. It is prefixed by “ light”, because due to its high speed it can handle large data size and requires less memory to run. Its principal advantage over the other gradient boosting algorithms is its ability to resolve the scalability problem and the long computational time by adopting a leaf-wise tree growth strategy. LightGBM splits the tree leaf wise with the best fit whereas other boosting algorithms split the tree depth wise or level wise rather than leaf-wise. Therefore, when growing on the same leaf in Light GBM, the leaf-wise algorithm can reduce more loss than the level-wise algorithm and hence results in much better accuracy, which can rarely be achieved by any of the existing boosting algorithms.

AdaBoost algorithm, short for adaptive boosting (Freund & Schapire, 1997) is a boosting technique used as an ensemble method in machine learning. It has a solid theoretical basis and has made great success in practical applications. The algorithm adapts its strategy to the situation being used, which free its user from the difficulty of determining algorithmic parameters by re-assigning the weights to each instance, with higher weights to incorrectly classified instances.

Performances criterium:

In order to compare the performances of the tested models, we used RMSE (Root of mean Square error) and R2, as defined in **eq1** and **eq2**, respectively.

(1)

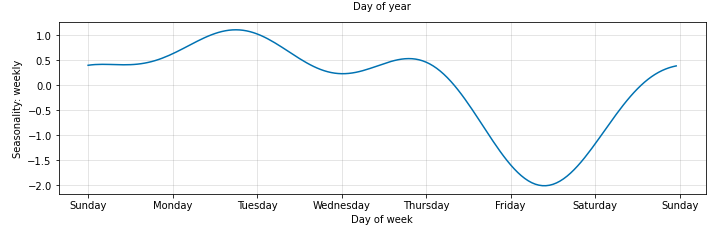
(2)

Where *yi,measured* is the *ith* measured value of a vector of *n* values, yi,predicted is the *ith* predicted value of the vector of *n* values. is the mean of the measured value.

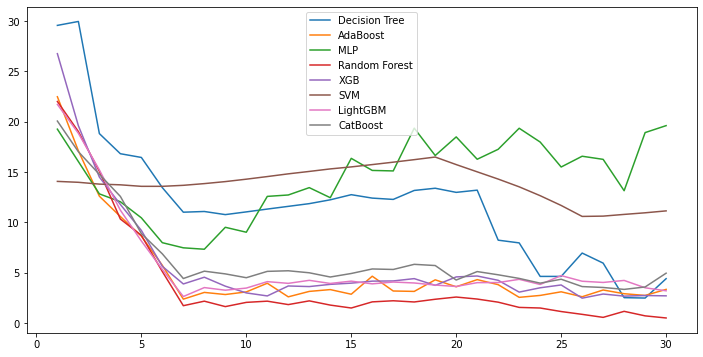
**Results and discussion**

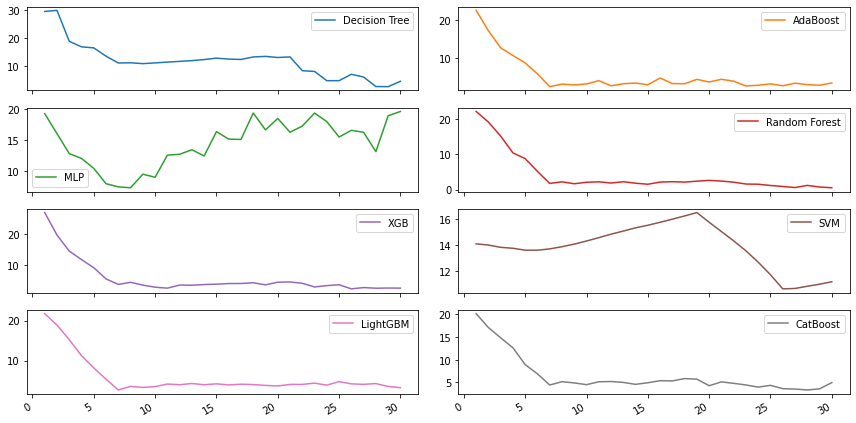
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **lags** | **Decision**  **Tree** | **AdaBoost** | **MLP** | **Random**  **Forest** | **XGB** | **SVM** | **LightGBM** | **CatBoost** |
| **1** | 29.58 | 22.49 | 19.27 | 22.02 | 26.77 | 14.08 | 21.73 | 20.075 |
| **2** | 29.97 | 17.12 | 16.05 | 19.03 | 19.61 | **13.99** | 18.86 | 17.03 |
| **3** | 18.83 | **12.61** | 12.84 | 15.08 | 14.51 | 13.81 | 15.22 | 14.78 |
| **4** | 16.83 | 10.59 | 12.06 | 10.34 | 11.79 | 13.74 | 11.22 | 12.59 |
| **5** | 16.46 | **8.64** | 10.46 | 8.72 | 9.21 | 13.59 | 8.16 | 8.90 |
| 6 | 13.46 | 5.74 | 7.99 | **5.12** | 5.64 | 13.59 | 5.33 | 6.88 |
| 7 | 11.02 | 2.38 | 7.48 | **1.73** | 3.89 | 13.69 | 2.64 | 4.43 |
| 8 | 11.09 | 3.05 | 7.34 | **2.18** | 4.56 | 13.86 | 3.53 | 5.16 |
| 9 | 10.78 | 2.84 | 9.52 | **1.64** | 3.67 | 14.06 | 3.27 | 4.89 |
| 10 | 11.05 | 3.10 | 9.03 | **2.06** | 3.00 | 14.30 | 3.48 | 4.51 |
| 11 | 11.33 | 3.96 | 12.59 | **2.18** | 2.70 | 14.56 | 4.12 | 5.14 |
| 12 | 11.60 | 2.61 | 12.73 | **1.84** | 3.69 | 14.83 | 3.95 | 5.20 |
| 13 | 11.87 | 3.15 | 13.46 | **2.21** | 3.63 | 15.07 | 4.25 | 5.00 |
| 14 | 12.25 | 3.33 | 12.46 | **1.79** | 3.85 | 15.32 | 3.94 | 4.58 |
| 15 | 12.76 | 2.87 | 16.37 | **1.50** | 3.97 | 15.52 | 4.17 | 4.93 |
| 16 | 12.42 | 4.66 | 15.17 | **2.11** | 4.17 | 15.75 | 3.90 | 5.38 |
| 17 | 12.29 | 3.19 | 15.12 | **2.22** | 4.19 | 15.99 | 4.08 | 5.33 |
| 18 | 13.18 | 3.15 | 19.37 | **2.10** | 4.42 | 16.24 | 3.987 | 5.84 |
| 19 | 13.40 | 4.30 | 16.65 | **2.37** | 3.75 | 16.50 | 3.78 | 5.72 |
| 20 | 12.99 | 3.61 | 18.50 | **2.59** | 4.59 | 15.75 | 3.64 | 4.28 |
| 21 | 13.21 | 4.32 | 16.28 | **2.39** | 4.68 | 15.03 | 4.02 | 5.12 |
| 22 | 8.24 | 3.82 | 17.27 | **2.08** | 4.253 | 14.31 | 4.03 | 4.80 |
| 23 | 7.97 | 2.56 | 19.36 | **1.56** | 3.08 | 13.54 | 4.35 | 4.44 |
| 24 | 4.64 | 2.75 | 17.99 | **1.50** | 3.51 | 12.66 | 3.84 | 3.97 |
| 25 | 4.64 | 3.11 | 15.51 | **1.15** | 3.79 | 11.69 | 4.72 | 4.36 |
| 26 | 6.96 | 2.62 | 16.58 | **0.87** | **2.48** | 10.60 | 4.16 | **3.63** |
| 27 | 5.96 | 3.29 | 16.27 | **0.57** | 2.89 | 10.63 | 4.04 | 3.54 |
| 28 | 2.53 | 2.92 | 13.155 | **1.17** | 2.68 | 10.80 | 4.24 | 3.35 |
| 29 | 2.49 | 2.75 | 18.93 | **0.72** | 2.74 | 10.96 | 3.51 | 3.59 |
| 30 | 4.415 | 3.35 | 19.62 | **0.51** | 2.71 | 11.15 | 3.21 | 4.96 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Lags | **Decision**  **Tree** | **AdaBoost** | **MLP** | **Random**  **Forest** | **XGB** | **SVM** | **LightGBM** | **CatBoost** |
| 1 | -2.86 | -1.23 | -0.64 | -1.14 | -2.16 | 0.12 | -1.08 | -0.78 |
| 2 | -2.96 | -0.30 | -0.13 | -0.60 | -0.70 | 0.13 | -0.57 | -0.28 |
| 3 | -0.56 | 0.30 | 0.27 | -0.00 | 0.07 | 0.16 | -0.02 | 0.03 |
| 4 | -0.25 | 0.50 | 0.35 | 0.52 | 0.38 | 0.16 | 0.44 | 0.30 |
| 5 | -0.19 | 0.66 | 0.51 | 0.66 | 0.62 | 0.18 | 0.70 | 0.65 |
| 6 | 0.12 | 0.85 | 0.71 | 0.88 | 0.85 | 0.18 | 0.87 | 0.79 |
| 7 | 0.46 | 0.97 | 0.75 | 0.98 | 0.93 | 0.17 | 0.96 | 0.91 |
| 8 | 0.45 | 0.95 | 0.76 | 0.97 | 0.90 | 0.15 | 0.94 | 0.88 |
| 9 | 0.49 | 0.96 | 0.60 | 0.98 | 0.94 | 0.141 | 0.95 | 0.89 |
| 10 | 0.48 | 0.96 | 0.65 | 0.98 | 0.96 | 0.12 | 0.95 | 0.91 |
| 11 | 0.45 | 0.93 | 0.33 | 0.98 | 0.97 | 0.11 | 0.93 | 0.89 |
| 12 | 0.44 | 0.97 | 0.33 | 0.98 | 0.94 | 0.090 | 0.93 | 0.89 |
| 13 | 0.42 | 0.95 | 0.26 | 0.98 | 0.95 | 0.08 | 0.93 | 0.90 |
| 14 | 0.40 | 0.95 | 0.38 | 0.98 | 0.94 | 0.06 | 0.93 | 0.40 |
| 15 | 0.37 | 0.97 | -0.04 | **0.99** | 0.93 | 0.06 | 0.93 | 0.90 |
| 16 | 0.41 | 0.92 | 0.12 | 0.98 | 0.935 | 0.05 | 0.94 | 0.89 |
| 17 | 0,44 | 0.96 | 0.15 | 0.98 | 0.93 | 0.05 | 0.94 | 0.89 |
| 18 | 0.37 | 0.962 | -0.36 | 0.98 | 0.93 | 0.04 | 0.943 | 0.88 |
| 21 | 0.23 | 0.91 | -0.16 | 0.97 | 0.90 | 0.01 | 0.93 | 0.88 |
| 22 | 0.65 | 0.93 | -0.52 | 0.98 | 0.90 | -0.04 | 0.92 | 0.88 |
| 23 | 0.60 | 0.96 | -1.34 | 0.98 | 0.94 | -0.14 | 0.88 | 0.88 |
| 24 | 0.82 | 0.94 | -1.71 | 0.98 | 0.90 | -0.34 | 0.88 | 0.87 |
| 25 | 0.70 | 0.87 | -2.29 | 0.98 | 0.80 | -0.87 | 0.69 | 0.74 |
| 26 | -1.33 | 0.67 | -12.20 | 0.96 | 0.70 | -4.39 | 0.17 | 0.37 |
| 27 | -4.78 | -0.76 | -42.01 | 0.94 | -0.34 | -17.37 | -1.65 | -1.03 |
| 28 | -0.26 | -0.68 | -33.19 | 0.73 | -0.42 | -22.03 | -2.56 | -1.23 |
| 29 | -0.63 | -1.00 | -93.46 | 0.86 | -0.98 | -30.66 | -2.24 | -2.39 |
| 30 | -7.37 | -3.83 | -164.30 | 0.87 | -2.16 | -52.42 | -3.43 | -9.56 |



**Figure : Weekly seasonality of PM2.5 in Algiers**





Conclusion

**References:**

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139.

(ONS,2018) <https://www.ons.dz/IMG/pdf/Demographie2018.pdf> ( visisted 19-07-2021)

Ladji, R., Yassaa, N., Balducci, C., & Cecinato, A. (2014). Particle size distribution of n-alkanes and polycyclic aromatic hydrocarbons (PAHS) in urban and industrial aerosol of Algiers, Algeria. *Environmental Science and Pollution Research*, *21*(3), 1819-1832.

Bouhila, Z., Mouzai, M., Azli, T., Nedjar, A., Mazouzi, C., Zergoug, Z., ... & Lounici, H. (2015). Investigation of aerosol trace element concentrations nearby Algiers for environmental monitoring using instrumental neutron activation analysis. *Atmospheric Research*, *166*, 49-59. https://doi.org/10.1016/j.atmosres.2015.06.013

Chellali, M.R., Abderrahim, H., Hamou, A., Nebatti, A., Janovec, J., 2016. Artificial  
neural network models for prediction of daily fine particulate matter concentrations  
in Algiers. Environ. Sci. Pollut. Control Ser. 14008–14017. https://doi.org/10.1007/  
s11356-016-6565-9

Talbi, A., Kerchich, Y., Kerbachi, R., & Boughedaoui, M. (2018). Assessment of annual air pollution levels with PM1, PM2. 5, PM10 and associated heavy metals in Algiers, Algeria. *Environmental Pollution*, *232*, 252-263. <https://doi.org/10.1016/j.envpol.2017.09.041>

Ibrir, A., Kerchich, Y., Hadidi, N., Merabet, H., & Hentabli, M. (2021). Prediction of the concentrations of PM1, PM2. 5, PM4, and PM10 by using the hybrid dragonfly-SVM algorithm. *Air Quality, Atmosphere & Health*, *14*(3), 313-323. <https://doi.org/10.1007/s11869-020-00936-1>

Liou, N. C., Luo, C. H., Mahajan, S., & Chen, L. J. (2019). Why Is Short-Time PM2. 5 Forecast Difficult? The Effects of Sudden Events. *IEEE Access*, *8*, 12662-12674. <http://doi.org/10.1109/ACCESS.2019.2963341>

Gao, X., & Li, W. (2021). A graph-based LSTM model for PM2. 5 forecasting. *Atmospheric Pollution Research*, 101150. <https://doi.org/10.1016/j.apr.2021.101150>

Ma, J., Ding, Y., Cheng, J. C., Jiang, F., Gan, V. J., & Xu, Z. (2020). A Lag-FLSTM deep learning network based on Bayesian Optimization for multi-sequential-variant PM2. 5 prediction. *Sustainable Cities and Society*, *60*, 102237. <https://doi.org/10.1016/j.scs.2020.102237>

Zhang, B., Zhang, H., Zhao, G., & Lian, J. (2020). Constructing a PM2. 5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. *Environmental Modelling & Software*, *124*, 104600. <https://doi.org/10.1016/j.envsoft.2019.104600>

Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., & Pak, C. (2020). Deep learning-based PM2. 5 prediction considering the spatiotemporal correlations: A case study of Beijing, China. *Science of The Total Environment*, *699*, 133561. https://doi.org/10.1016/j.scitotenv.2019.07.367