Traffic-Related Air Pollutant (TRAP) Prediction using Big Data and Machine Learning

Ismail A. Sulaimon
Big Data Technologies and Innovation Laboratory, University of Hertfordshire, Hatfield, AL10 9AB, UK
i.sulaimon@herts.ac.uk

Hafiz A. Alaka
Big Data Technologies and Innovation Laboratory, University of Hertfordshire, Hatfield, AL10 9AB, UK
h.alaka@herts.ac.uk

Abstract

The negative impact of the Increasing air pollution on the global economy, quality of life of humans and health of animals and plants has been enormous. Several works of literature, reports and news around the world have highlighted the risk posed by the ever-increasing air pollution and the threat to the lives of vulnerable groups such as children, the elderly, and people with respiratory and cardiovascular problems. The closest to home among all the air pollutants are the Traffic-Related Air Pollutants (TRAP), and they contribute the most to the risk posed to global health. This emphasises the urgency of the need for a highly accurate air pollution prediction model. Researchers have been able to achieve significant performance gain in predicting many of the pollutants except for the TRAP such as CO and NO which reported the worse prediction performance in many studies. CO and NO have been among the major pollutants of concern globally as they are linked to critical health hazards. Based on the established urgency of improving the accuracy of pollution prediction models, we collect recent data for six months and at high granularity in terms of time and location. The data is pre-processed and used to develop a Machine Learning (ML) based air pollution prediction model with high granularity and accuracy while focusing on traffic-related air pollutants CO and NO. Using the benchmarks r^2 and RMSE score, our ML models outperformed that of the studies reported in the literature for the prediction of TRAPs. This in part is due to the high data granularity we considered in terms of time and location.

1. INTRODUCTION

The United Nations (UN) reported that in 2050, the urban area is expected to be habiting about 68% of the world population (UN DESA, 2018). This alone will impact negatively on the air quality of major cities around the world if nothing is done to reduce the impact of rural-to-urban migration. The reports by the World Health Organization (WHO) elaborated how air pollution causes about seven million annual death around the world, while the air quality in above 80% of the urban areas is worse than the WHO guideline (WHO, 2014). The majority of those currently affected by this worsening situation is the vulnerable group which include children, the elderly, and people with respiratory and cardiovascular problems. Records show that in recent years, air pollution accounts for 1 out of 8 deaths globally (WHO, 2014). This highlights the importance and urgency of the need for a highly accurate air pollution prediction model.

Air pollution arises due to the increase in the proportion of pollutants in the form of particles and inhabitable gasses in the atmosphere of an area. The increase in air pollution sometimes could be due to human activities in the case of traffic, industry or home-related pollution. While other times it could be due to biological or environmental activities such as the case of pollutants like ozone (O_3), pollen, dust. Air pollutants of various nature have been linked to critical health challenges such as cardiovascular diseases, pulmonary disease, acute respiratory infection and increased risk of lung cancer (Gul & Khan, 2020). Pollutants such as Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Ozone are medically proven to irritate the airways of the lungs, increasing the symptoms of those suffering from lung diseases. Fine particles are seen to find their way deep into the lungs where they can cause inflammation and a worse condition of heart and lung diseases. Carbon Monoxide (CO) inhibits the absorption of oxygen by the blood (Department for Environment, 2021). This condition alone can lead to a shortage of oxygen supply to the heart and most likely death. Agriculturally, the yield of crops has been affected by the increasing concentration of Ozone in the atmosphere (Gul & Khan, 2020).

Researchers (such as Martínez-España et al., 2018; Tao et al., 2019; Zhang & Ding, 2017; Zhao et al., 2016) have been able to achieve significant performance gain in predicting many of the pollutants except for the traffic-related air pollutants (TRAP) such as CO and NO which reported the worse prediction performance in many studies (Rybarczyk & Zalakeviciute, 2018). CO and NO have been among the major pollutants of concern globally as they are linked to various health hazards. Worsening the situation are the particulate matters (PM) which are getting into the limelight recently. An increase in exposure to TRAPs, $PM_{2.5}$ and black carbon has been linked to the risk of asthma across childhood up to twelve years of age and a decreased cognitive function in older men (Bowatte et al., 2015; Power Melinda C. et al., 2011). The magnitude increases with age and the pattern is more prominent with $PM_{2.5}$. Other evidence shows that TRAP is also associated with eczema and hay fever (Bowatte et al., 2015). Many of these health hazards have been the driving force for governments around the world to redesign policies to reduce the impact of TRAPs on the environment. Despite government efforts, pollution is yet to reduce to a significant low. Hence, the need to be aware of where and when pollution is high. So, individuals and organizations can be well equipped to make an informed decision about pollution in their respective environments.

In the efforts to achieve an accurate pollution prediction, researchers (such as Martínez-España et al., 2018; Tao et al., 2019; Zhang & Ding, 2017; Zhao et al., 2016) have applied various methodologies on diverse sets of data. Due to the effectiveness of Machine Learning (ML) algorithms in solving many prediction problems in research and industry, researchers have been focusing largely on using various ML algorithms (Zhang & Ding, 2017). As deterministic models struggle to capture the relationships between variables that affect pollution, researchers have been pushed to considering various ML algorithms ranging from the classical ML algorithms such as Support Vector Machine (SVM), Linear Regression (LR) and the sophisticated ML algorithms like Deep-Neural Network (DNN), and Extreme Machine Learning (Iskandaryan et al., 2020; Rybarczyk & Zalakeviciute, 2018).

Diverse approaches have been used so far to achieve optimal accuracy in the prediction of pollution. Different combinations of data such as Meteorological, Traffic and Air Quality data have been used in the prediction of TRAPs, but less attention is paid to the granularity of the dataset used in terms of time and location. In the course of finding an optimally performing prediction model, researchers have trained various ML algorithms on different combinations of data (Masih, 2019). In most of the literature studied by (Iskandaryan et al., 2020; Rybarczyk & Zalakeviciute, 2018), relatively good performance was recorded for air pollutants such as O_3, PM_{10}, $PM_{2.5}$ and SO_2, while many of the models reported performed worse for TRAPs CO and NO with an average r^2 score of 0.76 and 0.78 respectively for the prediction models reported. Therefore, the situation has risen to a level of urgency due to the high risk associated with TRAPs and the low level of prediction accuracy reported so far against them.

Based on the established urgency of improving the accuracy of pollution prediction models, this research aims to study pollution prediction models reported in the literature and develop an air pollution prediction model with high granularity and accuracy while focusing on TRAPs CO and NO. The focus of our model development will be on improving the performance of the ML predictive model by increasing the granularity of the dataset used in terms of time and location. This will hopefully, broaden the focus of the research community on the need to focus on data granularity in the prediction of TRAPs.

2. LITERATURE REVIEW

Several research efforts have been contributed to predicting TRAPs using ML algorithms and different combinations of data. From the literature surveyed, the prediction of NO has received more research contribution than CO in recent years. The time-series approach implemented in (Li et al., 2017) used a three-step optimized process on a combination of datasets of traffic, meteorology, population density and elevation. To predict NO and NO_x concentration, they used an Ensemble Learning algorithm to reduce the uncertainty of the prediction. The complex approach earned a good predictive performance for locations with high density while other locations saw a less accurate prediction performance.

(Araki et al., 2018) developed a spatiotemporal model based on land use to represent relationships between pollutants and land. To estimate the concentration levels of NO_2, the study compared two algorithms - Land Use Random Forest (LURF) and Land Use Regression (LUR). Their work used a dataset recorded over four years in the Amagasaki region of Japan. The study also used datasets of the population, emission intensities, meteorology, satellite-derived NO_2, and time variables. The study was able to show the relationship between land use and pollutants, but the accuracy of the models recorded is slightly lower than the benchmark reported in related literature.

Seven Regression models are used in (Hu et al., 2017) to predict CO concentrations in Sidney, Australia. The models were trained on about ten years of data recorded through 15 stations. Comparison of the several models' performance suggested that the best of the prediction are provided by the Support Vector Regression (SVR), Decision Tree Regression (DTR), and Random Forest Regression (RFR). Their field validation also showed that SVR has the highest spatial resolution estimation and better identifies boundaries of the polluted area than the other regression models. Despite the large size of the dataset used, the performance improvement is not far from the state-of-the-art.

Many of the research contributed in literature so far developed predictive models using data with low granularity in terms of time and location. Hence, this research explores the benefit of using more granular data in training predictive models to improve the predictive performance of the TRAPs predictive model.

3. RESEARCH METHODS

As deterministic models have proved less efficient in the prediction of air pollutants, and ML shows the more promising result as reported in the literature. This research intends to apply ML in solving the accuracy challenge identified. Hence we implemented the full length of the ML pipeline, starting from Data pre-processing, Algorithm selection, Model training/validation/testing, Model evaluation. Every stage of the ML pipeline is implemented using the Python programming language (van Rossum & Drake, 1995).

3.1 Data

The availability, volume and quality of data is a major factor to consider while applying ML in solving a problem. To ensure high-quality data and selection of an optimal combination of dataset and Variable, we documented the stages of Data collection, Data merging, Variable selection. The data collected and used for the model development is to the location granularity of Latitude and Longitude, while the time granularity is to hourly.

3.1.1 Data collection

For this research, the focus is on ensuring high granularity and accuracy of our model prediction. Hence, we collected meteorological (weather) data, air quality (AQ) data and traffic data for 338 postcodes spread around the United Kingdom (UK). For this research, we used merged data for the 6 months starting from 01:00 am, 1st of December 2020, till 00:00 1st of June 2021.

3.1.1.1 Meteorological Data

The meteorological data collected is the weather data for each of the 338 UK postcodes identified for this research. These data are been sourced through the Open Weather Map API (*Weather API - OpenWeatherMap*). The API response is in a JSON object format, while the data is saved directly to the CSV file format for easy access.

3.1.1.2 Air Quality Data

The advent of the Internet-of-Things (IoT) has made the recording of Air Quality data easier. The air quality data used in this research is sourced via 338 IoT sensors stationed across the UK. The sensors record data for several pollutants (such as NO_2, PM_{10}, SO_2, FINE, O_3, PM_1, $PM_{2.5}$, TSP, CO, NO) every hour. Each of the sensors does not record readings for all the pollutants. Hence, the recordings for various pollutants were collated from several sensors. Despite the air quality data having records for other pollutants, this research focuses mainly on TRAPs such as Carbon monoxide (CO) and Nitrogen Oxide (NO).

3.1.1.3 Traffic Data

The process of getting the traffic data has been challenging as several sources does not have it readily available while many of the sources do not have the data recorded hourly nor at the level of granularity needed. Hence, for this research, we sourced traffic data through the TomTom traffic flow API (*TomTom Developer Portal | Maps APIs and SDKs for Location Applications*), analysed the data and merge them to achieve the level of granularity desired. This data is analysed and averaged to derive an hourly average of every day of the week. This is achieved by eliminating the data recorded for the period immediately after the COVID-19 lockdown was eased in the UK (between the 8^{th} of March 2021 and the 22^{nd} of March 2021). The elimination was done due to the spike in the traffic situation across the UK roads during the stated period. This ensures that the irregularity in the traffic data within this period does not distort the whole traffic data collected.

3.1.2 Data merging

Data merging is necessary to ensure we have all the data required for the ML pipeline in a single data repository. The Meteorological, Air Quality and Traffic data were merged using the location (Latitude and Longitude), Day and Time Variables. This ensures that the location granularity considered is to the level of Latitude and Longitude, and the time granularity considered is hourly.

3.1.3 Variable selection

The Meteorological, Air Quality and Traffic data collected are made up of numerous Variables, which only a few of them are of significance to the ML model development. Hence, irrelevant variables of the data were discarded while the relevant Variables are used for the model development. While the Location, Date and Time variables were used for data merging only, they were discarded after the data marging. Presented in Table 1 below are the variables selected for the model development.

Table 1: Selected variables for model development

Data	**Variable**	**Unit**	**Type**
Location	Latitude		
	Longitude		
Date	Date		
Time	Time	Hour	
Meteorological	Absolute Temperature	^{o}c	Independent
	Feels-Like	^{o}c	
	Pressure	Hg	
	Humidity	%	
	Minimum Temperature	^{o}F	
	Maximum Temperature	^{o}F	
	Wind Speed	km/h	

	Wind Degree	degree	
	Cloud	okta	
	Rainfall	mm	
Traffic	Current Travel Speed	km/h	
	Free Flow Travel Speed	km/h	
	The ratio of Current Travel Speed and Free Flow Travel Speed		
	Current Travel Time	Seconds	
	Free Flow Travel Time	Seconds	
	The ratio of Current Travel Time and Free Flow Travel Time		
	Data Confidence		
	Road Closure	0,1,2,3,4	
Air Quality	CO	$\mu g/m^3$	Dependent
	NO	$\mu g/m^3$	Dependent

The variables of the meteorological data and the traffic data are used as the independent variables, while the variables of the air quality data are used as the dependent variables.

3.2 Model development

When the data has been prepared, the next stage in the ML pipeline is model development. This involves algorithm selection, model training, validation and testing. The ML model development is done using the Scikit-Learn Python Package (Pedregosa et al., 2012).

3.2.1 Algorithm selection

Following the study by (Iskandaryan et al., 2020; Rybarczyk & Zalakeviciute, 2018), 10 well performing ML algorithms reported in the literature were selected to be trained using the merged dataset. The 10 ML algorithms are trained using their respective default parameters that are defined as the most promising by the designers of each algorithm. No hyperparameter tuning is done to enable a fair comparison of the performance of the algorithm. The algorithms considered are Extra Trees Regressor, Histogram-Based Gradient Boosting Regressor, Light Gradient Boosted Machine (LGBM) Regressor, eXtreme Gradient Boosting (XGB) Regressor, Random Forest Regressor, Bagging Regressor, Nu Support Vector Regression (NuSVR), Support Vector Regression (SVR), Gradient Boosting Regressor, K-Neighbors Regressor.

3.2.2 Training and validation

At the training stage, each of the selected ML algorithms is fitted using the merged data to create a model which can be used for TRAPs CO and NO prediction. The training is conducted by using a portion of the merged data, while the other portion is used for the model testing after it has been trained. Hence, the merged data is split randomly, as 70% is used for the training and validation. The 10-fold cross-validation was done to ensure that the performance of the trained model is accurately recorded.

3.2.3 Testing

Testing the trained model is an important stage of the model development process. This will enable researchers to evaluate the performance of the trained model. To enable accurate recording of the performance of the model, the other 30% of the merged data is used for testing. The models are used to predict the values of each of the dependent variables, the predicted values are compared with the true values. The differences are recorded for each of the true values in the dataset to enable further analysis.

4. MAIN DISCUSSION

The training, validation and testing of the model are not complete unless the result of the testing has been evaluated. This will enable deduction to be derived about the performance of the model.

4.1 Result evaluation

The prediction performance of the trained models is measured using the benchmark performance measure reported in similar literature. This will ease the comparison of performance with other similar research works. The prediction performance of each of the trained models is measured using R-Squared (r^2) and Root Mean Square Error (RMSE).

4.1.1 CO

The predictive model performance for each of the ML algorithms trained for the prediction of CO is shown in Table 2.

Table 2: Performance measure of ML Algorithms on predicting CO

Model	Adjusted R-Squared	R-Squared	RMSE
ExtraTreesRegressor	0.786	0.789	0.099
HistGradientBoostingRegressor	0.753	0.758	0.107
LGBMRegressor	0.750	0.755	0.107
XGBRegressor	0.749	0.754	0.108
RandomForestRegressor	0.740	0.745	0.110
BaggingRegressor	0.705	0.711	0.117
NuSVR	0.687	0.693	0.121
SVR	0.671	0.678	0.124
GradientBoostingRegressor	0.662	0.669	0.125
KNeighborsRegressor	0.613	0.621	0.134

The performance measure revealed that the Extra Trees Regressor recorded the best performance (r^2 = 0.789, RMSE=0.099) when compared with the other ML algorithms trained on the same data.

4.1.2 NO

The prediction of NO took a different turn as the models performed differently when compared with their respective performance for CO.

Table 3: Performance measure of ML Algorithms on predicting NO

Model	Adjusted R-Squared	R-Squared	RMSE
HistGradientBoostingRegressor	0.522	0.526	7.061
LGBMRegressor	0.511	0.514	7.148

ExtraTreesRegressor	0.507	0.511	7.171
XGBRegressor	0.485	0.489	7.329
RandomForestRegressor	0.476	0.480	7.394
BaggingRegressor	0.421	0.426	7.774
KNeighborsRegressor	0.349	0.355	8.238
GradientBoostingRegressor	0.336	0.341	8.329
NuSVR	0.163	0.169	9.349
SVR	0.132	0.139	9.521

The performance of the models measured using the r^2 and the RMSE is presented in Table 3. As seen in the performance recorded, the models performed differently from the performance recorded for CO. This reveals that some ML algorithms are more suitable for CO prediction while they are not for the prediction of NO. Here, the Hist Gradient Boosting Regressor poses the best performance ($r^2 = 0.526$, RMSE=7.061).

4.2 Training time

The training time for the models reveals how long it takes each of the ML algorithms to learn through the data presented to it. The length of the training time and the performance of the algorithm is always a concern to ML model developers.

4.2.1 CO

The training time for each of the ML algorithms trained on the CO data is shown in Figure 1.

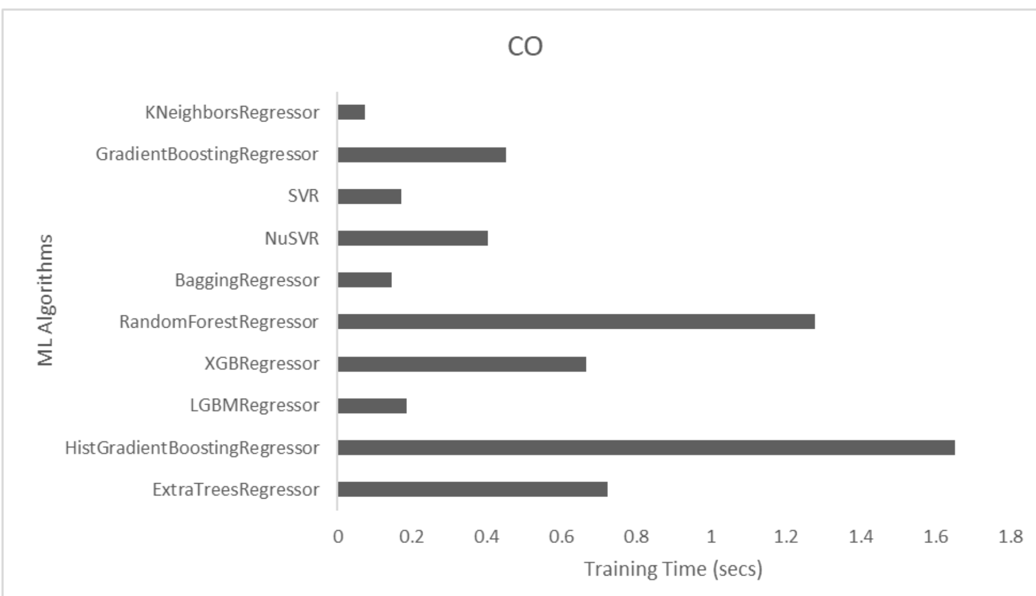

Figure 1: Training time of ML Algorithms on predicting CO

The K-Neighbors Regressor has the least training time with the worse performance, while the Extra Trees Regressor poses the best performance, but a less encouraging training time. The LGBM Regressor has shown a promising balance between the model training time and predictive performance.

4.2.2 NO

Figure 2 shows the training time for each of the ML algorithms for the prediction of NO.

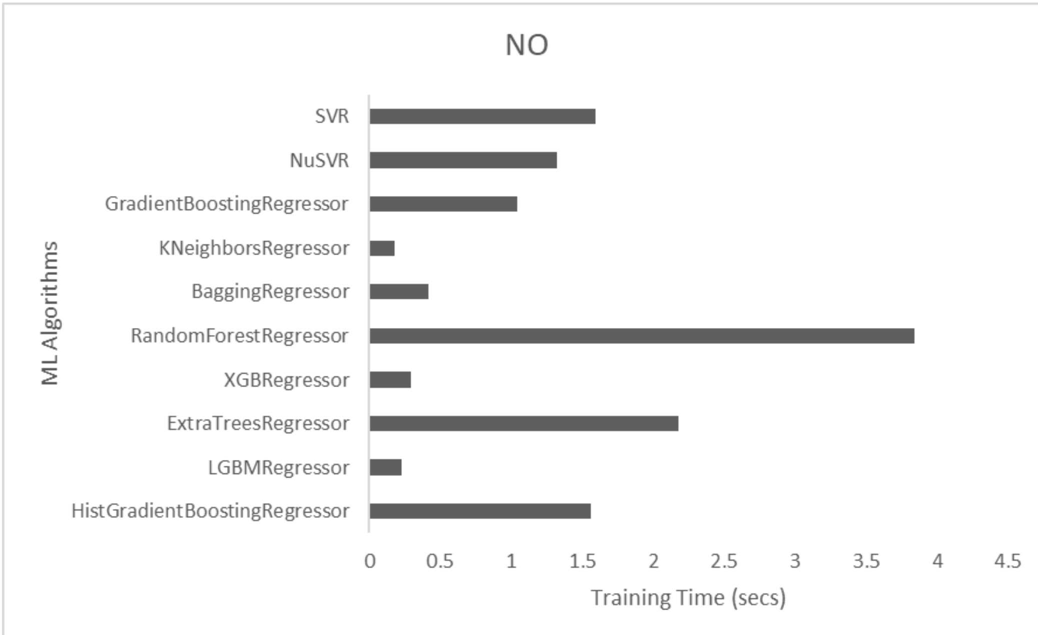

Figure 2: Training time of ML Algorithms on predicting NO

As seen in the training time for CO, the ML algorithm K-Neighbors Regressor also takes the least time to train on NO dataset while its performance is nothing close to the best performing ML algorithm in this category. While the Hist Gradient Boosting Regressor poses the best performance for the prediction of NO, the training time is high. Hence, the LGBM Regressor has a balanced trade-off between the training time and predictive performance.

5. CONCLUSIONS AND RECOMMENDATIONS

Traffic-Related Air Pollutants (TRAPs) has been linked to many severe health and economic risks globally. Despite the efforts of research, predictive models have not been able to achieve high performance in forecasting the TRAPs. This research focused on predicting TRAPs such as CO and NO with high accuracy and data granularity. The ML models are developed using high data granularity in terms of time and location. This enables better performance of the model and ensures the forecast is more localized. The performance of our ML models is evaluated using the commonly used benchmarks score r^2 and RMSE, to enable easy comparison of predictive performance. The models developed outperformed the study reported in the literature for the prediction of TRAPs. Future work is envisaged to enable the discovery of more high performing ML algorithms in the prediction of TRAPs. Another open area of research is to perform the hyper-parameter tuning of each ML algorithms in the discovery of more promising parameters for each of the algorithms considered. We also aim to use more recent data that will reflect the post-COVID-19 UK lockdown traffic situation and to ensure better predictive accuracy.

6. REFERENCES

68% of the world population projected to live in urban areas by 2050, says UN | UN DESA | United Nations Department of Economic and Social Affairs. (2018, May 16). https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html

Araki, S., Shima, M., & Yamamoto, K. (2018). Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan. Science of The Total Environment, 634, 1269–1277. https://doi.org/10.1016/j.scitotenv.2018.03.324

Bowatte, G., Lodge, C., Lowe, A. J., Erbas, B., Perret, J., Abramson, M. J., Matheson, M., & Dharmage, S. C. (2015). The influence of childhood traffic-related air pollution exposure on asthma, allergy and sensitization: A systematic review and a meta-analysis of birth cohort studies. Allergy, 70(3), 245–256. https://doi.org/10.1111/all.12561

Department for Environment, F. and R. A. (Defra) webmaster@defra gsi gov uk. (2021, March 15). Effects of air pollution- Defra, UK (UK; United Kingdom). Department for Environment, Food and Rural Affairs (Defra), Nobel House, 17 Smith Square, London SW1P 3JR helpline@defra.gsi.gov.uk. https://uk-air.defra.gov.uk/air-pollution/effects

Gul, S., & Khan, G. M. (2020). Forecasting Hazard Level of Air Pollutants Using LSTM's. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), Artificial Intelligence Applications and Innovations (pp. 143–153). Springer International Publishing. https://doi.org/10.1007/978-3-030-49186-4_13

HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation From Fixed and Mobile Sensors | IEEE Journals & Magazine | IEEE Xplore. (n.d.). Retrieved 14 June 2021, from https://ieeexplore.ieee.org/document/7892954

Hu, K., Rahman, A., Bhrugubanda, H., & Sivaraman, V. (2017). HazeEst: Machine Learning Based Metropolitan Air Pollution Estimation From Fixed and Mobile Sensors. IEEE Sensors Journal, 17(11), 3517–3525. https://doi.org/10.1109/JSEN.2017.2690975

Iskandaryan, D., Ramos, F., & Trilles, S. (2020). Air Quality Prediction in Smart Cities Using Machine Learning Technologies Based on Sensor Data: A Review. Applied Sciences, 10(7), 2401. https://doi.org/10.3390/app10072401

Li, L., Lurmann, F., Habre, R., Urman, R., Rappaport, E., Ritz, B., Chen, J.-C., Gilliland, F. D., & Wu, J. (2017). Constrained Mixed-Effect Models with Ensemble Learning for Prediction of Nitrogen Oxides Concentrations at High Spatiotemporal Resolution. Environmental Science & Technology, 51(17), 9920–9929. https://doi.org/10.1021/acs.est.7b01864

Martínez-España, R., Bueno-Crespo, A., Timon-Perez, I. M., Soto, J., Ortega, A., & Cecilia, J. M. (2018). Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain. J. Univers. Comput. Sci.

Masih, A. (2019). Machine learning algorithms in air quality modeling. Global Journal of Environmental Science and Management, 5(4). https://doi.org/10.22034/GJESM.2019.04.10

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2012). Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON, 6.

Power Melinda C., Weisskopf Marc G., Alexeeff Stacey E., Coull Brent A., Spiro Avron, & Schwartz Joel. (2011). Traffic-Related Air Pollution and Cognitive Function in a Cohort of Older Men. Environmental Health Perspectives, 119(5), 682–687. https://doi.org/10.1289/ehp.1002767

Rybarczyk, Y., & Zalakeviciute, R. (2018). Machine Learning Approaches for Outdoor Air Quality Modelling: A Systematic Review. Applied Sciences, 8(12), 2570. https://doi.org/10.3390/app8122570

Tao, Q., Liu, F., Li, Y., & Sidorov, D. (2019). Air Pollution Forecasting Using a Deep Learning Model Based on 1D Convnets and Bidirectional GRU. IEEE Access, 7, 76690–76698. https://doi.org/10.1109/ACCESS.2019.2921578

TomTom Developer Portal | Maps APIs and SDKs for Location Applications. (n.d.). TomTom Developer Portal. Retrieved 17 July 2021, from https://developer.tomtom.com/

van Rossum, G., & Drake, F. L. (n.d.). Python Reference Manual. 196.

Weather API - OpenWeatherMap. (n.d.). Retrieved 17 July 2021, from https://openweathermap.org/api

WHO | 7 million premature deaths annually linked to air pollution. (2014, March 25). WHO; World Health Organization. https://www.who.int/mediacentre/news/releases/2014/air-pollution/en/#.WqBfue47NRQ.mendeley

Zhang, J., & Ding, W. (2017). Prediction of Air Pollutants Concentration Based on an Extreme Learning Machine: The Case of Hong Kong. International Journal of Environmental Research and Public Health, 14(2), 114. https://doi.org/10.3390/ijerph14020114

Zhao, C., Heeswijk, M. van, & Karhunen, J. (2016). Air quality forecasting using neural networks. 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 1–7. https://doi.org/10.1109/SSCI.2016.7850128