Proceedings of the 7th World Congress on Civil, Structural, and Environmental Engineering (CSEE'22)
~~Lisbon, Portugal~~ Virtual Conference – April 10 – 12, 2022
Paper No. ICEPTP 172
DOI: 10.11159/iceptp22.172

NO_2 Concentration Modelling Using Meteorological and Traffic Features.

Tomasz Turek, Joanna A. Kamińska
Department of applied mathematics, Wrocław University of Environmental and Life Sciences
Grunwaldzka 53 St., 50-357, Wrocław, Poland
tomasz.turek@upwr.edu.pl; joanna.kaminska@upwr.edu.pl

Abstract –The constantly evolving urban areas also increase the concentration of emission gases. This fact has an adverse effect on habitants' health and life quality. A scientific basis, such as the development of accurate and adequate models of air pollution, is necessary to be able to influence decision-makers in such a way that real actions to improve air quality are carried out. Therefore, creating and improving models describing this phenomenon is extremely important. Both linear (Multiple Linear Regression) and non – linear methods (Random Forest) were used for modelling concentrations of pollutants in atmosphere. Based on the traffic, meteorological and pollution data from 2015 – 2020 in Wrocław, it was shown that it is possible to predict concentration of NO_2 with high accuracy: R^2 statistic reaches 88% while predicting daily average concentration using Random Forest methodology. Multiple Linear Regression provides worse fit and is biased by necessity to comply with its statistical restraints. Regardless of its costs it provides explicit interpretation of each factor in model. It was shown that there is a strong correlation between pollutants concentration and traffic volume as well as meteorological factors. Models perform significantly better when to the set of predictors concentrations of other pollutants are included. Modelling every natural phenomenon is a very challenging task. In order to excel, the further study in the area of models' optimalization, investigation relationships naturally occurring, and including new variables must be performed.

***Keywords*:** air pollution, nitrogen oxides, random forest, linear regression, traffic flow, meteorological features

1. Introduction

The matter of air pollution in a large and densely populated areas has become very important to the public health of residents. The increasing concentrations of pollutants are caused by the constantly developing urbanization of human communities. Constantly growing population of the cities directly associated with increasing usage of the internal combustion vehicles is undoubtedly a reason of increasing traffic volume which results in the continuous increment of exhaust gas emissions. In addition, in order to provide a sufficient infrastructure to live and work, the density of buildings in urban areas also increases which significantly reduces the efficiency of air corridors which used to successfully ventilate the city in the past. Wrocław as the important economical and academic center (consisting of 10 state universities) in southwest Poland is an example of such developing city. It is estimated that around 614.600 people are living in the city. The relevance of the traffic movement is proven by the literature [1] which indicates that estimated number of vehicles commuting in the city reaches 15.000. Habitants of Wrocław more often choose individual transport. In Wrocław, the 41% of residents choose to travel by car, where only 28% declares to use public transport. The remaining 24% and 6% chooses feet or bike respectively. Due to the Wrocław's water conditions it is impossible to build an underground network. These factors clearly indicate that the problem of the dense car traffic in the city is very high.

The main gases that come from the car's combustion engines are nitrogen oxides: NO_2 and NO_x. Results and insights from investigating the influence of the meteorological and traffic factors on the concentrations of the pollutants can in consequence have impact on the decisive urban managers to implement a traffic reduction in the city centres or different actions in order to reduce emissions, improve air quality, thus the quality of habitant's lives. Developed models, consisting of environmental-traffic related set of predictors can be successfully used in a location intelligence system [2] which is a base to decision support systems for local decision makers [3].

Aim of this research is to assess a possibility of estimating an NO_2 concentrations by models developed with machine learning methodology – Random Forest (RF) and compare the results with analogous Multiple Linear Regression (MLR) model. The advantage of the latter is a possibility of interpretation each of the factor on the final result. Machine learning methods, including artificial neural networks [4,5] single random trees [6] or random forest [7,8,9] also has proven to be successful in the topic of air pollution modelling despite their ambiguity and computational complexity.

2. Methods and Data Sources

The Random Forest methodology is based on the set of simple decision trees. Every single tree is created for randomly selected subset of data. The output is generated by the aggregation and averaging the individual predictions of each component tree. Due to its construction, it can be directly compared to the classical regression method as the performance can be assessed by the percentage of explained variation of the dependent variable that is predictable from the independent variables which is in line with commonly used R-squared measure (R^2). Additionally, to compare models we are going to use the root mean squared error (RMSE). We can think of this as the average difference between the predicted value for NO_2 concentration and the actual observed value. The NO_2 concentration will be estimated using hourly data and both daily, and monthly aggregated datasets where all the variables have been averaged. There will be also comparison between models built solely using the meteorological and traffic factors and those which would include other pollutants' concentrations.

2.1. Data sources

All the analyses were performed using the data sample from Wrocław, Poland which covered 5 years – from 2016 to 2020. Traffic data were provided by the Traffic and Public Transport Management Department of the Roads and City Maintenance Board in Wrocław. The data contain counts of all vehicles (cars, buses, trucks, etc.) passing through the measurement intersection in each traffic lane. Data is aggregated into the hourly intervals in order to maintain coherence with a meteorological and pollution data. Pollution concentration data are collected by the Provincial Environment Protection Inspectorate and measured at hourly intervals. The air inlet to the system is located 3m above ground level. Both traffic and air pollution data measurement stations are in located in the immediate vicinity of the Hallera and Powstańców Śląskich street intersection which is considered as one of the most important communicational canyons in the city. Meteorological data are provided by the Institute of Meteorology and Water Management (IMGW) at only one station in Wrocław, located on the outskirts of the city (9 km from the intersection in a straight line). The meteorological data set used in this research contains hourly air temperature, wind speed, and relative humidity.

3. Results

This section provides the study of the correlation between pollutants and meteorological and traffic features. There is also an insight of the air quality and traffic trends in dependence on the month (i.e., season of the year) or hour on the representative subsets of data.

3.1.1. Correlation inspection.

In order to investigate the dependencies among the variables the plot of the meteorological features along with the pollutants is presented for year 2020 (Fig.1) along with the correlation matrix for all variables (Fig. 2)

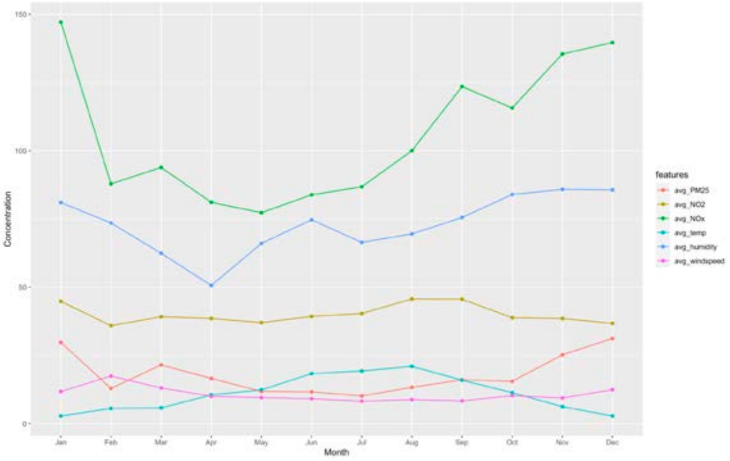

Fig. 1: Plot of the features over months in 2020.

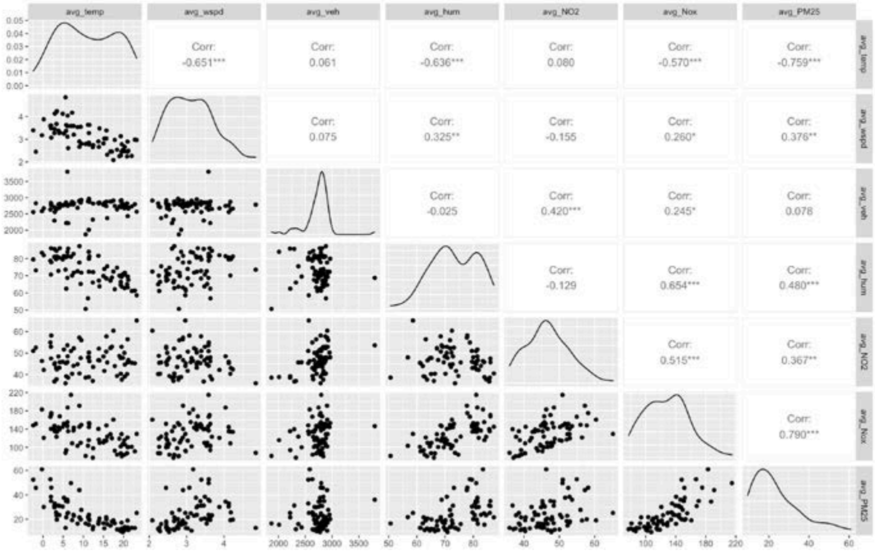

Fig. 2: Correlation matrix for all variables in the data sample.

We observe a strong correlation between pollutants which is expected as the similar conditions are required to increase their concentration. There is also clear dependence between pollutants e.g., Particulate Matter are negatively correlated with temperature. It should be mentioned, that for pollutant which is of greatest importance in this research - $NO_{2,}$ once again accordingly to expectations the highest correlation is between the numbers of vehicles, ergo traffic volume level.

3.1.2. Accuracy of the Models

This section provides information about performance of the constructed models (Table 1, Table 2). The set of all predictive variables consisted: temperature, humidity, wind speed, number of vehicles and set of other pollutants – $NO_{x,}$ $PM_{2.5}$. In the case of monthly and daily aggregates values have been averaged.

Table 1: Accuracy of the models built with meteorological and traffic features

	Meteorological and traffic features			
	Random Forest		Multiple linear regression	
	R^2	RMSE	R^2	RMSE
Hourly	44.40%	16.47	38.26%	17.36
Daily	48.60%	9.36	41.58%	9.99
Monthly	23.62%	5.33	19.49%	5.52

Table 2: Accuracy of the models built with all features.

	All features			
	Random Forest		Multiple linear regression	
	R^2	RMSE	R^2	RMSE
Hourly	87.87%	7.69	72.60%	11.56
Daily	88.15%	4.50	76.23%	6.37
Monthly	66.96%	3.50	68.89%	3.432

It can be observed that the best results are obtained for the prediction of the average daily NO_2 concentrations. As expected, including other types of pollutants significantly increased a models' goodness of fit due to their high level of correlation with target variable. RF models are generally more accurate, and do not require fulfilment set of statistical assumptions.

4. Conclusion

In this paper, influence of metrological and traffic related variables on air quality prediction has been investigated. It has been proven that there is a significant correlation between most common pollutants and weather conditions. We cannot neglect the influence and the correlation of impact of the traffic level to the pollutants' concentration. This research glimpsed at machine learning methods and demonstrated its potential in air quality models improvement. It was shown that RF model can successfully be used for air quality predictions. As future work we are going to implement different non-linear machine learning models such as support vector regression and widen our scope of variables in order to obtain more accurate results. Nevertheless, we are going to optimize RF model parameters as it has proven to be highly effective.

References

[1] M. Chalfen, J. A. Kamińska, "Identification of parameters and verification of an urban traffic flow model. Case study in Wrocław,". *ITM Web Conf.* 23, 00005, 2018.

[2] Sz. Szewrański, M. Świąder, J.K. Kazak, K. Tokarczyk-Dorociak, J. van Hoof, "Socio-Environmental Vulnerability Mapping for Environmental and Flood Resilience Assessment: The Case of Ageing and Poverty in the City of Wrocław, Poland," *Society of Evironmental Toxicology and Chemistry*, 14(5), pp. 592−597, 2018.

[3] J.K. Kazak, M. Chalfen, J.A. Kamińska, S. Szewrański, M. Świąder, "Geo-Dynamic Decision Support System for Urban Traffic Management," in *Dynamics in GIscience* I. Ivan, J. Horák, T. Inspektor, Ed. GIS OSTRAVA 2017. Lecture Notes in Geoinformation and Cartography. Springer, Cham, pp. 195−207.

[4] M.A. Elangasinghe, N. Singhal, K.N. Dirks, J.A. Salmond, "Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis," *Atmospheric Pollution Research*, 5, 4, pp. 696−708, 2014.

[5] F. Nejadkoorki, S. Baroutian, "Forecasting extreme PM10 concentrations using artificial neural network," *International Journal of Environmental Research*, 6, pp. 277−284, 2012.

[6] K.P. Singh, S. Gupta, P. Rai, “Identifying pollution sources and predicting urban air quality using ensemble learning methods,” *Atmospheric Environment*, 80, pp. 426−437, 2013.

[7] Y. Zhu, Y. Zhan, B. Wang, Z. Li, Y. Qui, K. Zhang, “Spatiotemporally mapping of the relationship between NO2 pollution and urbanization for a megacity in Southwest China during 2005–2016.” *Chemosphere*, 220, pp. 155−162, 2019.

[8] J.A. Kamińska, “The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: a case study in Wrocław,” *Journal of Environment Management*, 217C, pp. 164−174, 2018.

[9] I. Laña, J. Del Ser, A. Pedró, M. Vélez, C. Casanova-Mateo, “The role of local urban traffic and meteorological conditions in air pollution: A data-based study in Madrid, Spain,” *Atmospheric Environment*, 2016.