proceedings

MDPI

Proceedings

Lag Variables in Air Pollution Modeling Based on Traffic Flow and Meteorological Factors †

Joanna A. Kamińska [1,*], Guido Sciavicco [2], Estrella Lucena-Sánchez [2,3] and Fernando Jiménez [4]

[1] Department of Mathematics, Wroclaw University of Environmental and Life Sciences, 50-375 Wrocław, Poland
[2] Department of Mathematics and Computer Science, University of Ferrara, 44121 Ferrara, Italy; guido.sciavicco@unife.it (G.S.); estrella.lucenasanchez@unife.it (E.L.-S.)
[3] Department of Physics, Informatics and Mathematics, University of Modena e Reggio Emilia, 41121 Modena, Italy
[4] Department of Information and Communication Engineering, University of Murcia, 30100 Murcia, Spain; fernan@um.es
* Correspondence: joanna.kaminska@upwr.edu.pl
† Presented at Innovations-Sustainability-Modernity-Openness Conference (ISMO'20), Bialystok, Poland, 20–21 May 2020.

Published: 15 July 2020

Abstract: In order to refine the research on the impact of environmental factors on the concentration of pollutants in the air, in this paper, we present a mathematical model that allows the possibility of taking into account the past values of factors (explanatory variables) when modeling the current concentration of pollution. We conducted numerical analyzes based on hourly data from meteorological, traffic and air quality monitoring stations in Wrocław (Poland, Central Europe) from 2015–2017. In order to determine the optimal delay of each explanatory variable, we used a multi-objective optimization model (MO). It turned out that for the concentration of nitrogen oxides, delayed traffic flow, wind speed and sunshine duration time are more important than current ones. Then we built two random forest models: an actual model of current values of explanatory variables and a lag model with delayed variables determined by the MO method. Taking into account variables with an optimal delay (lag model) results in an increase in model accuracy for NO_2 with R^2 = 0.51 to 0.56 and for NO_x from 0.46 to 0.52. We deduced that in pollutant concentrations modeling, the possibility of greater influence of variables with delay should always be considered because it can significantly increase the accuracy of the model and indicate additional relationships or dependencies.

Keywords: air pollution; nitrogen oxides; random forest; lag variables; multi-objective optimization; traffic flow; meteorological conditions

1. Introduction

The relationship between air pollutant concentrations and environmental factors is widely studied. The quantitative and qualitative recognition of the impact of factors makes it possible to undertake actions aimed at preventing, reducing or limiting the spread of pollution. Pollution models can support urban managers in taking actions to improve air quality in the city [1–3]. The growing population of cities and increasing motorization are the reasons for the increasing number of moving vehicles and consequently, increasing exhaust gas emission. The expansion and density of city buildings reduces the phenomenon of city ventilation, which results in a decrease in the impact of low wind speeds on the evacuation of pollution. Wrocław currently has 641,600 of residents [4]. It is estimated that about 15,000 vehicles in the rush hours to below 1000 vehicles at night are moving

around the city. That means that approximately 40,000 vehicles make journeys in the city during one hour [5]. One of the main air pollutants emitted by car combustion engines is nitrogen oxides: NO_2 and $NO_x = NO + NO_2$. In the literature there exist many different air pollution concentration models, e.g., multidimensional regression models [6–8], polynomial functions [9,10], artificial neural networks [11], single random trees [12], random forest (RF) [13–15] and boosted regression trees [16,17]. These models take into account, in addition to the current values, the past values of the explanatory variables, which have been used mainly to study the impact of pollution concentration on human health and life. Lag variables are then used to take account of the exposure duration to harmful conditions [18,19].

The intensity of chemical reactions in the atmosphere depends on the duration of certain favorable conditions. Therefore, it can be assumed that the current concentration values are significantly affected not only by the current values of the explanatory variables (t), but also by previous moments ($t - 1$, $t - 2$, $t - 3$, …). Classically, this issue is described by adding to the predictor set new variables with a delay (lag variables) 1, 2, 3, … This method has two main disadvantages: first, it is not known how far back the delay variables should be created, and second, creating a set of variables for each delay significantly multiplies the number of explanatory variables, extending the time of calculations and deteriorating the quality and even the possibility of interpretation. In [20,21], it was proposed to use the multi-purpose optimization (MO) algorithm to determine the delay of each predictor that ensures maximum model fit. To be precise, a three-object optimization was developed: power (maximum 3), delay and regression coefficients for each of the variables were ultimately optimized by matching the model. To assess the influence of the variables delay, we used a random forest (RF) algorithm with lagged variables (Lag model) designated in the MO process and compared it with RF developed with original variables without delay (Actual model).

2. Data Source

We performed numerical analyzes using data from Wrocław (51.086 N, 17.012 E). Data covered the full 3 years of 2015–2017 in hourly intervals. Traffic data are provided by the Traffic and Public Transport Management Department of the Roads and City Maintenance Board in Wrocław. The data contain the number of all vehicles passing through the measurement intersection (51.08637 N, 17.01202 E) during a period of one hour. Traffic flow shows a clear, bimodal daily variability [15] with two peaks: in the morning and in the afternoon. Meteorological hourly 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 (51.10319 N, 16.89985 E; 9 km from the intersection in a straight line). One can observe clear seasonal variation in temperature, characteristic of transitional climate type subject to both oceanic and continental influences. Air pollution data are collected by the Provincial Environment Protection Inspectorate and measured at hourly intervals. The measuring station is located in the direct vicinity of the intersection with traffic measurement (30 m from the middle of intersection).

3. Results and Discussion

Using the MO, we determined the function describing the dependence of NO_2 and NO_x concentration on meteorological factors and traffic flow. We determined the delay, regression coefficient and power (maximum 3) of each variable to maximize the fit of the model to real data. Based on the 10-fold cross-validation process and on the selection of the most appropriate in terms of occurring in the atmosphere phenomena interpretation, we obtained linear functions with the delays given in Table 1. The fact of obtaining a linear function proves that the relationship is indeed linear and not of a higher degree.

Table 1. Delays [h] of variable received in multi-objective optimization process.

	Traffic Flow	Wind Speed	Air Temp.	Sunshine Duration	Relative Humidity	Air Pressure
NO_2	1	3	0	2	7	0
NO_x	1	2	0	10	0	0

For both NO_2 and NO_x, one hour of traffic flow have the major influence on actual concentration. This results from the emission and accumulation phenomenon of air pollutants.

Wind speed has an impact on the evacuation of pollution. The stronger wind speed is, the more intense evacuation and lower pollution concentration is. Due to the distance of the meteorological station (9 km in a straight line), the effect of wind speed is delayed by 2–3 h. This is a consequence of the time needed for the air masses to reach the air quality measurement station. In Wrocław, West and North-West winds prevail, therefore blowing from the meteorological station to the city center; at an average wind speed of 3.1 m s^{-1} and covering a distance of 9 km, the time this takes, taking into account the porosity of urban buildings, ranges from 2 to 3 h.

In the next step, we built two random forest models: using actual predictor values and using lagged values (predictor values with delay) for NO_2 and NO_x.

Due to the greater variation in NO_x values (coefficient of variation is equal to 46% for NO_2 and 73% for NO_x), it is more difficult to predict its values effectively. This is generally indicated by lower goodness of fit measure values than in NO_2 (Table 2). However, for both pollutants, including lag variables has improved the models fit.

Table 2. Goodness of fit coefficient for NO_2 and NO_x modeling.

	NO_2		NO_x	
	Actual	**Lag**	**Actual**	**Lag**
R^2	0.51	0.56	0.46	0.52
MADE	12.2	11.6	47.9	45.3
MAPE	0.26	0.24	0.35	0.33
RMSE	16.2	15.4	76.3	72.2
r	0.72	0.75	0.68	0.72

r—Pearson correlation coefficient, MADE-Mean Absolute Devation Error, MAPE-Mean Absolute Percentage Error, RMSE-Root Mean Square Error.

In full generality, it can be concluded that determining the optimal delay of environment variables and including such lag variables as a predictor increases the accuracy of the model. The method of determining optimal delay for each independent variable and inputting this lag variable into modeling is an absolutely general method and may be utilized in every air pollution modeling notwithstanding considered factors and type of pollution. Detailed conclusions depend on local meteorological, topographical and traffic conditions.

Author Contributions: Conceptualization, J.A.K. and G.S.; methodology, J.A.K., G.S. and E.L.-S.; software, G.S. and E.L.-S.; validation, J.A.K., G.S. and E.L.-S.; data curation, J.A.K.; writing—original draft preparation, J.A.K.; writing—review and editing, G.S. and E.L.-S.; supervision, F.J.; funding acquisition, J.A.K. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript or in the decision to publish the results.

References

1. Kazak, J.K.; Castro, D.G.; Swiader, M.; Szewranski, S. Decision support system in public transport planning for promoting urban adaptation to climate change. *IOP Conf. Ser. Mater. Sci. Eng.* **2019**, *471*, 112007, doi:10.1088/1757-899x/471/11/112007.
2. Kazak, J.; Chalfen, M.; Kamińska, J.A.; Szewrański, S.; Świąder, M. Geo-Dynamic Decision Support System for Urban Traffic Management. In *Dynamics in GIScience. GIS Ostrava 2017. Lecture Notes in Geoinformation and Cartography*; Ivan, I., Horak, J., Inspektor, T., Eds.; Springer: Cham, Switzerland, 2018; pp. 195–207, doi:10.1007/978-3-319-61297-3_14.
3. Barratt, B.; Atkinson, R.; Ross, A.H.; Beevers, S.; Kelly, F.; Mudway, L.; Wilkinson, P. Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction for a traffic management scheme. *Atmos. Environ.* **2007**, *41*, 1784–1791, doi:10.1016/j.atmosenv.2006.09.052.

4. Available online: https://wroclaw.stat.gov.pl/ (accessed on 23 January 2020).
5. Chalfen, M.; Kamińska, J.A. Identification of parameters and verification of an urban traffic flow model. A case study in Wrocław. *ITM Web Conf.* **2018**, *23*, 00005, doi:10.1051/itmconf/20182300005.
6. Ping, S.J.; Harrison, R.M. Regression modelling of hourly NO_x and NO_2 concentration in urban air in London. *Atmos. Environ*. **1997**, *31*, 4081–4094, doi:10.1016/S1352-2310(97)00282-3.
7. Aldrin, M.; Haff, I.H. Generalized additive modelling of air pollution, traffic volume and meteorology. *Atmos. Environ*. **2005**, *39*, 2145–2155, doi:10.1016/j.atmosenv.2004.12.020.
8. Zhang, Z.; Zhang, X.; Gong, D.; Quan, W.; Zhao, X.; Ma, Z.; Kim, S.-J. Evolution of Surface O_3 and $PM_{2.5}$ concentrations and their relationships with meteorological conditions over the last decade in Beijing. *Atmos. Environ.* **2015**, *108*, 67–75, doi:10.1016/j.atmosenv.2015.02.071.
9. Szyda, J.; Wierzbicki, H.; Stokłosa, A. Statistical modelling of changes in concentrations of atmospheric NO_2 and SO_2. *Pol. J. Environ. Study* **2009**, *18*, 1123–1129. doi:10.1016/j.scitotenv.2012.03.076.
10. Singh, K.P.; Gupta, S.; Kumar, A.; Shukla, S.P. Linear and nonlinear modeling approaches for urban air quality prediction. *Sci. Total Environ.* **2012**, *426*, 244–255. doi:10.1016/j.scitotenv.2012.03.076.
11. Elangasinghe, M.A.; Singhal, N.; Dirks, K.N.; Salmond, J.A. Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmos. Pollut. Res*. **2014**, *5*, 696–708, doi:10.5094/APR.2014.079.
12. Singh, K.P.; Gupta, S.; Rai, P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. *Atmos. Environ.* **2013**, *80*, 426–437, doi:10.1016/j.atmosenv.2013.08.023.
13. Araki, S.; Shima, M.; Yamamoto, K. Spatiotemporal land use random forest model for estimating metropolitan NO_2 exposure in Japan. *Sci. Total Environ*. **2019**, *634*, 1269–1277, doi:10.1016/j.scitotenv.2018.03.324.
14. Zhu, Y.; Zhan, Y.; Wang, B.; Li, Z.; Qui, Y.; Zhang, K. Spatiotemporally mapping of the relationship between NO_2 pollution and urbanization for a megacity in Southwest China during 2005–2016. *Chemosphere* **2019**, *220*, 155–162, doi:10.1016/j.chemosphere.2018.12.095.
15. Kamińska, J.A. The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław, *J. Environ. Manag*. **2018**, *217C*, 164–174, doi:10.1016/j.jenvman.2018.03.094.
16. Kamińska, J.A. Residuals in the modelling of pollution concentration depending on meteorological conditions and traffic flow, employing decision trees. *ITM Web Conf.* **2018**, *23*, 00016, doi:10.1051/itmconf/20182300016.
17. Sayegh, A.; Tate, J.A.; Ropkins, K. Understanding how roadside concentrations of NOx are influenced by the background levels, traffic density, and meteorological conditions using Boosted Regression Trees. *Atmos. Environ*. **2016**, *127*, 163–175, doi:10.1016/j.atmosenv.2015.12.024.
18. Kowalska, M.; Skrzypek, M.; Kowalski, M.; Cyrys, J.; Ewa, N.; Czech, E. The relationship between daily concentration of fine particulate matter in ambient air and exacerbation of respiratory diseases in silesian agglomeration, Poland. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1131, doi:10.3390/ijerph16071131.
19. Vanos, J.K.; Cakmak, S.; Kalkstein, L.S.; Yagouti, A. Association of weather and air pollution interactions on daily mortality in 12 Canadian cities. *Air Qual. Atmos. Health* **2015**, *8*, 307–320, doi:10.1007/s11869-014-0266-7.
20. Jiménez, F.; Kamínska, J.; Lucena-Sánchez, E.; Palma, J.; Sciavicco, G. Multi-Objective Evolutionary Optimization for Time Series Lag Regression. In Proceedings of the 6th International Conference on Time Series and Forecasting, Granada, Spain, 25–27 September 2019; pp. 373–384.
21. Brunello, A.; Kamińska, J.; Marzano, E.; Montanari, A.; Sciavicco, G.; Turek, T. Assessing the Role of Temporal Information in Modelling Short-Term Air Pollution Effects Based on Traffic and Meteorological Conditions: A Case Study in Wrocław. In *New Trends in Databases and Information Systems*; Communications in Computer and Information Science; ADBIS 2019; Welzer, T., Eder, J., Podgorelec, V., Wrembel, R., Ivanovic, M., Gamper, J., Morzy, M., Tzouramanis, T., Darmont, J., Kamišalić Latifić, A., Eds.; Springer: Cham, Switzerland, 2019; p. 1064, doi:10.1007/978-3-030-30278-8_45.