Journal of Environmental Management 217 (2018) 164–174

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wrocław

Joanna A. Kamińska

Department of Mathematics, Wroclaw University of Environmental and Life Sciences, ul. Grunwaldzka 53, 50-357 Wrocław, Poland

ARTICLE INFO

Article history:
Received 25 October 2017
Received in revised form
15 March 2018
Accepted 22 March 2018

Keywords:
Urban air pollution
Traffic flow
Meteorological conditions
Random forest
Data subsets

ABSTRACT

Random forests, an advanced data mining method, are used here to model the regression relationships between concentrations of the pollutants NO_2, NO_x and $PM_{2.5}$, and nine variables describing meteorological conditions, temporal conditions and traffic flow. The study was based on hourly values of wind speed, wind direction, temperature, air pressure and relative humidity, temporal variables, and finally traffic flow, in the two years 2015 and 2016. An air quality measurement station was selected on a main road, located a short distance (40 m) from a large intersection equipped with a traffic flow measurement system. Nine different time subsets were defined, based among other things on the climatic conditions in Wrocław. An analysis was made of the fit of models created for those subsets, and of the importance of the predictors. Both the fit and the importance of particular predictors were found to be dependent on season. The best fit was obtained for models created for the six-month warm season (April–September) and for the summer season (June–August). The most important explanatory variable in the models of concentrations of nitrogen oxides was traffic flow, while in the case of $PM_{2.5}$ the most important were meteorological conditions, in particular temperature, wind speed and wind direction. Temporal variables (except for month in the case of $PM_{2.5}$) were found to have no significant effect on the concentrations of the studied pollutants.

1. Introduction

Wrocław is Poland's fourth largest city in terms of population (638,000 inhabitants). It is situated in the south-western part of the country, at an elevation of 105–156 m above sea level, and covers an area of 293 km^2. It is the oldest city in Poland by date of granting of the town charter. Five rivers flow through the city, as well as multiple former river channels and canals, and these are crossed by more than 100 bridges and 30 footbridges. These conditions mean that the city's road system is radial in nature, and is constantly being adapted to support increasing traffic volumes. There are currently 270,000 vehicles registered in Wrocław, including 200,000 cars. The structure of the road system, in conjunction with the large number of vehicles travelling in the city, produces significant congestion, and this leads to increased exhaust emissions. According to the Provincial Environment Protection Inspectorate, in Wrocław road traffic is responsible for 56% of NO_2 emissions, 44% of CO emissions and 16% of $PM_{2.5}$ particulate emissions (Information on air quality …, 2016). The relatively low percentages (compared with other countries) of harmful emissions accounted for by road vehicle exhaust gases, particularly in the case of $PM_{2.5}$, is a result of the city's housing conditions. Wrocław's history has resulted in a situation where the city contains many century-old tenements and other houses which are still heated with solid fuels (coal and wood). It is estimated that 81% of $PM_{2.5}$, 54% of CO and 9% of NO_2 emissions in Wrocław originate from the domestic sector (Information on air quality …, 2016). Action is being taken to reduce surface emissions both from that sector and from transport. The impact of road transport on air pollution in the conurbation is beyond dispute. The unnaturally high atmospheric concentrations of the aforementioned substances have adverse consequences for human health, in particular respiratory and cardiovascular health effects (Hien et al., 2016; Adam et al., 2015; Hoek et al., 2013). Studies have also shown that air pollution may be a contributing factor to autism in children (Flores-Pajot et al., 2016) and Parkinson's disease (Pei Chen et al., 2016), and through the consequences of these may even lead to fatalities (Tang et al., 2017).

E-mail address: joanna.kaminska@upwr.edu.pl.

https://doi.org/10.1016/j.jenvman.2018.03.094
0301-4797/

Pollution models can help traffic managers to take decisions efficiently, by selecting the most adequate traffic management strategy (Barratt et al., 2007). In the literature the main input for models is traffic and meteorological data (González-Aparicio et al., 2013; Zhang and Batterman, 2013; Sayegh et al., 2016). There also exist studies based only on traffic data (Keeler, 2014) or only on meteorological data (Mlakar and Boinar, 1997). Laña et al. (2016) studied the effect of the choice of explanatory variables (traffic, meteorological, temporal) on the accuracy and fit of models; they obtained comparable results for temporal and meteorological variables and for sets also including traffic variables. However, this was not decisive for identifying the effect and significance of a traffic variable in modelling the PM_{10} concentration. In the present study, the input data include information on meteorological conditions as well as traffic and temporal data, covering the years 2015–2016. The problem of selecting an appropriate model to describe the relationships between air pollution concentration and explanatory variables becomes more and more challenging with the development of computational techniques and machine learning. These relationships are described effectively by the popular multidimensional regression models – originally linear, but now more complex, and still undergoing development. González-Aparicio et al. (2013) presented three different linear regression models – simple linear regression, linear regression with interaction terms, and linear regression with interaction terms following Sawa's Bayesian Information Criteria – to describe the dependence of PM_{10} concentration on traffic, meteorological and temporal data. Bertaccini et al. (2012) and Aldrin and Haff (2005) proposed the use of a generalised additive model to model the short-term impact of traffic and weather on air pollution. Machine learning, which is also being continuously developed, has also been used in the modelling of air pollution concentrations. Boosted regression trees are one of the classification and regression methods based on decision trees. Sayegh et al. (2016) used boosted regression trees to investigate how roadside NO_x concentrations depend on background levels, traffic density and meteorological conditions. An even more computationally complex method is that of random forests (RF), as used by Laña et al. (2016), which involves the compilation of information from multiple decision trees simultaneously. Random forests have gained momentum in the last decade by virtue of their ability to handle multidimensional classification and regression problems with excellent accuracy and low likelihood of overfitting (Breiman, 2001).

A separate issue from the choice of a modelling method is the selection of a time period for which the model parameters are to be determined or decision trees are to be constructed. At first glance it would appear best to use as long a time series as possible. However, certain relations may hold only within defined shorter periods, and become lost when large sets of continuous data are considered. The division of the calendar year into a warm season (April–September) and a cool season (October–March) makes it possible to eliminate from the modelling in the warmer period pollution produced by the domestic sector (chiefly domestic heater emissions), which in the colder period accounts for a significant part of the random error in the model. Bertaccini et al. (2012) modelled NO_2 and PM_{10} concentrations for the entire year and separately for four seasons (winter, spring, summer, autumn). Particularly in the case of temperature, two-hour wind speed and air pressure, the estimation effects for different periods differed significantly. A similar division was made by Zhang et al. (2015), who showed by means of a multiple regression method that for the winter season seven meteorological factors explained 59% of the variance in $PM_{2.5}$. In the present study, to evaluate how the quality and adequacy of the model depend on the period for which it is constructed, nine different divisions into subperiods are applied, based on a two-year measurement series. Apart from the warmer and colder subperiods and the four seasons of the year, subperiods were also defined based on associated traffic levels: working days, characterised by a bimodal distribution of traffic flow; and non-working days, when road traffic is not dependent on residents' journeys to and from work.

This paper expands on the aforementioned work by adopting a new perspective: not only does it explore the significance of individual variables for pollution levels, but it also evaluates how the length and choice of time period used in analysing the relationship between pollution and traffic, meteorological and temporal features affect the accuracy of the analysis and the significance of particular variables. The aim is to determine the impact of seasonal separation on the relationship between pollution and traffic and meteorological conditions.

The main question addressed is: How does the choice of analysed time interval affect the accuracy of the analysis and the importance of the explanatory variables used?

The paper is organised as follows. Section 2 describes the data related to traffic, pollution and meteorological conditions, and presents the theory and scheme of construction of the various models. Section 3 contains the results of modelling, together with comparisons showing how the period selected for analysis influences the quality of the model and the importance of the variables. In Section 4 the results are summed up and conclusions are drawn.

Specific models are proposed and results analysed for the pollutants NO_2, NO_x and $PM_{2.5}$, for the two years 2015–2016 and for nine separate time subsets.

2. Material and methods

2.1. Traffic

The traffic data are provided by the Traffic and Public Transport Management Department of the Roads and City Maintenance Board in Wrocław, which operates 921 video cameras distributed widely over the area of the city. Cameras manufactured by Autoscope, together with software, are used to monitor city traffic in an Intelligent Transport System (ITS). One of the pieces of information obtained is the number of vehicles passing through the measurement plane on a given traffic lane or lanes. This count includes all vehicles passing through that plane (cars, goods vehicles, public transport vehicles). A network of sensors is set up to monitor vehicular traffic at the main intersections of the city road network. A total of 68 intersections are subject to traffic measurement; marked in Fig. 1 is the one used in the present analysis: the intersection of Hallera and Powstańców Śląskich.

This extensive network makes it possible to observe the behaviour of traffic over time at multiple points throughout the city. However, having so many measuring devices also means that some of the individual time series will have a fraction of data missing or marked with an error code. These gaps in the measurement series are due to road maintenance or repair or turning of the cameras. In such cases, the missingness was handled by replacing the missing data by the average value for the time and day of the week in question, taken from the remaining data. The numbers of vehicles recorded by the camera in 15-min intervals were aggregated into hourly counts. This operation ensured that the time step size was uniform for all variables and reduced the noise produced by outliers, while maintaining the characteristics of the original distribution. The availability of meteorological and