Science of the Total Environment 651 (2019) 475–483

Contents lists available at ScienceDirect

Science of the Total Environment

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

A random forest partition model for predicting NO_2 concentrations from traffic flow and meteorological conditions

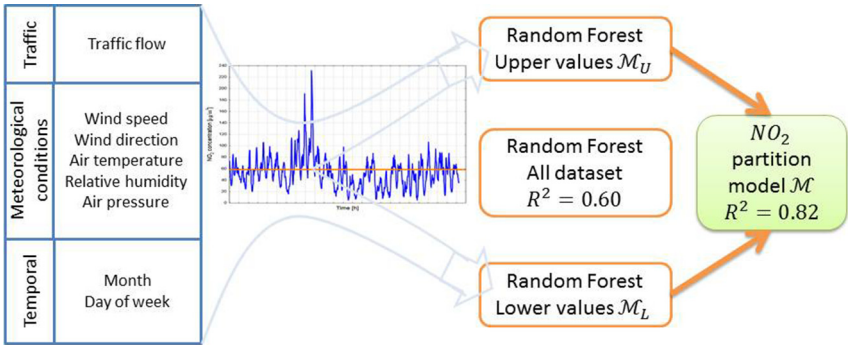
Joanna A. Kamińska

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

HIGHLIGHTS

- NO_2 pollution is caused mainly by road transport and modified by weather factors.
- A random forest describes well the dependence of NO_2 on the explanatory factors.
- A new RF-based partition model improves the description of NO_2 concentrations.
- The value of R^2 is increased from 0.60 to 0.82.
- Traffic flow has a greater impact on NO_2 in low concentration ranges.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 June 2018
Received in revised form 4 September 2018
Accepted 15 September 2018
Available online 17 September 2018

Editor: P. Kassomenos

Keywords:
Urban air pollution
Nitrogen dioxide concentration
Traffic flow
Meteorological conditions
Random forest

ABSTRACT

High concentrations of nitrogen dioxide in the air, particularly in heavily urbanised areas, have an adverse effect on many aspects of residents' health (short-term and long-term damage, unpleasant odour and other). A method is proposed for modelling atmospheric NO_2 concentrations in a conurbation, using a partition model $\mathcal{M}$ consisting of two separate models: $\mathcal{M}_L$ for lower concentration values and $\mathcal{M}_U$ for upper values. An advanced data mining technique, that of random forests, is used. This is a method based on machine learning, involving the simultaneous compilation of information from multiple random trees. Using the example of data recorded in Wrocław (Poland) in 2015–2017, an iterative method was applied to determine the boundary concentration $\tilde{y}$ for which the mean absolute deviation error for the partition model attained its lowest value. The resulting model had an R^2 value of 0.82, compared with 0.60 for a classical random forest model. The importances of the variables in the model $\mathcal{M}_L$, similarly as in the classical case, indicate that the greatest influence on NO_2 concentrations comes from traffic flow, followed by meteorological factors, in particular the wind direction and speed. In the model $\mathcal{M}_U$ the importances of the variables are significantly different: while traffic flow still has the greatest impact, the effects of temperature and relative humidity are almost as great. This confirms the justifiability of constructing separate models for low and high pollution concentrations.

1. Introduction

Air pollution in conurbations leads to serious economic, social and health problems. Unnaturally high levels of nitrogen oxides (principally NO_2) in the air have an adverse impact on human health, particularly related to the respiratory and cardiovascular systems (Hien et al., 2016; Hoek et al., 2013). Constant exposure to air pollution, originating chiefly from road vehicles, may lead to asthma, even in adults (Bowatte et al., 2018). Studies have shown that such pollution may also be a cause of autism in children (Flores-Pajot et al., 2016) and Parkinson's disease

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

https://doi.org/10.1016/j.scitotenv.2018.09.196
0048-9697/

(Pei-Chen et al., 2016), and as a consequence may even lead to death (Tang et al., 2017). According to the World Health Organization (WHO, 2012), globally 3.7 million deaths were attributable to ambient air pollution in 2012, including 480,000 in Europe. The large increase compared with the previous estimate of 1.3 million deaths in 2008 results partly from a change in the method of estimation and from the inclusion of rural areas in the analysis. Nitrogen dioxide is a major component of photochemical smog.

Nitrogen oxide pollution originates from various sources. Among the most harmful anthropogenic sources are transport emissions and domestic heating. According to the European Environment Agency, 39% of nitrogen oxides originate from road vehicles (EEA Report no. 9, 2017). In spite of a halving of emissions in this category since 1990, in 2015 the permissible annual average concentration of NO_2 (40 μg/m^3) was exceeded at a great majority (89%) of traffic stations in Europe (EEA, 2017). The dense and high buildings found in cities affect the air temperature and humidity and the wind direction, and also obstruct the evacuation of pollutants produced in the city. The accumulation of these adverse factors not only causes damage to health, but also impairs the comfort of living in a city. Attempts are made by local, national and international authorities to reduce the level of the problem to a minimum. Various scenarios of possible change are considered, together with their socioeconomic consequences. The modelling of pollutant concentrations in a longer time frame makes it possible to determine how particular types of emissions contribute to pollution and to predict the changes that will result from implementation of the considered scenarios. Pollution models can help traffic managers to take decisions efficiently, by selecting the most adequate traffic management strategy (Barratt et al., 2007; Kazak et al., 2018; Chalfen and Kamińska, 2018). The modelling of pollutant concentrations based on measurable area traits has been developed for many years. With the development of computers, the degree of complexity and accuracy of the models has increased. The still developing method of machine learning has also found applications in the modelling of air pollution levels. Catalano et al. (2016) used a neural network to model NO_2 concentrations as a function of meteorological conditions and traffic volume. Machine learning methods also include modelling based on decision trees (DT), boosted regression trees (BRT) and random forests (RF). Sayegh et al. (2016) used boosted regression trees to investigate how roadside concentrations of NO_x are influenced by background levels, traffic density, and meteorological conditions. Laña et al. (2016) modelled concentrations of nitrogen oxides, carbon monoxide and ozone using a random forest method in which information was compiled from multiple decision trees simultaneously. Generally, in an era in which machines offer huge computational power, random forests have gained momentum by virtue of their ability to handle multidimensional classification and regression problems with excellent accuracy and low likelihood of overfitting (Breiman, 2001).

When modelling pollutant concentrations as functions of the ambient conditions (without considering retrospective dependences, with a 1 h lag, for example) a fundamental problem is the low values obtained for measures of goodness of fit. Based on worldwide data from 5220 air monitors located on all continents, the method of land use regression with Lasso variable selection (Larkin et al., 2017) led to a model for NO_2 levels for which the adjusted R-square measure of fit was equal to 0.52. This means that only 52% of the variation in the pollutant concentrations was explained by the variation in the values of the independent variables. Sayegh et al. (2016) constructed 112 models for the concentration of five air pollutants based on four different sets of variables. For the largest set of input data considered by the authors, including meteorological conditions, temporal data and traffic flow, they obtained R^2 values of 0.49–0.54 for nitrogen dioxide, 0.37–0.48 for PM_{10} and 0.33–0.44 for nitrogen monoxide. In another study (Kamińska, 2018a) the three pollutant types NO_2, NO_x and $PM_{2.5}$ were modelled as functions of nine variables (including traffic volume and meteorological and temporal conditions) with a division into climatic periods for the studied area (spring, summer, autumn, winter). The R^2 values, reflecting the amounts of variation explained by the model, ranged from 0.31 to 0.58. The low values of R^2 result from the significant differences in the values of the dependent variables, the effect of unknown factors, and the difficulty of constructing a single model encompassing both typical and extreme pollutant levels. The goal of the present study is to modify the general model for the dependence of air pollutant concentrations on meteorological and temporal conditions and traffic flow, in such a way as to improve its accuracy without loss of generality. A known model described in the literature is accurate to a level of $R^2 = 0.92$ (Catalano et al., 2016); this is a prognostic model based chiefly on autocorrelation, where the predictors include the values of pollutant concentrations from the previous time point x_{t-1}. In the problem described here, a general function for any time point is sought, without reference to previous values of the dependent variable. Comparison of RF models with boosted regression trees (Kamińska, 2018b) has shown that RF models achieve a better fit to high values, but fit less well to low values. In seeking a solution to this problem, it was decided to divide the entire set of empirical data into “lower” and “upper” concentration values, and to model each of these subsets independently. The nature of the models constructed by machine learning means that it is not possible to give a function in explicit analytical form, and hence it is not possible to create a spline model (analogous to a spline function), since the assumption of continuity of the derivatives resulting from the union of functions does not hold, and it is not possible to determine uniquely the set of arguments for which the individual component functions are obtained. It is therefore proposed to describe the model constructed via this process as a partition model. Separate modelling of the impact of meteorological and temporal conditions and traffic flow in cases of low (harmless) and high (hazardous to human health) NO_2 concentrations has the aim of increasing the accuracy of fit, and thus enabling the correct interpretation of results obtained. The increasingly popular random forest method, based on the compilation of information from multiple decision trees, was applied.

2. Methodology

A random forest (RF) consists of a predetermined number of simple (binary) decision trees. Each of the component trees in an RF uses a sample subset of the available data, of predetermined size. These subsets are sampled independently for each tree, and the same instance may occur in multiple subsets (sampling with replacement). In addition, for each tree, a subset of the predictors – again of predetermined size – is selected. This means that each weak tree is trained on a different set of data and predictors. The predictive output is obtained by aggregating and averaging the individual predictions of all such compounding trees. This construction method, which blends the concepts of bagging and random feature selection, has been demonstrated to improve performance over other machine learning algorithms and linear regression models (Archer and Kimes, 2008). A random forest can describe both linear and nonlinear relationships, without any additional assumptions concerning either the independent or the dependent variables. The method is successfully used in the analysis of large datasets (data mining). In each of the models discussed in the paper, the importance of the predictive variables was determined as the sum – over all tree nodes – of the increases in the resubstitution estimate (ΔR), this value being expressed as a percentage of the maximum sum (over all variables). This means that the most important variable is assigned an importance of 100.

In spite of the wide potential for the use of the RF method, it does not provide a satisfactory description of the complex relationship between pollutant concentrations and ambient conditions. Because of the need