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

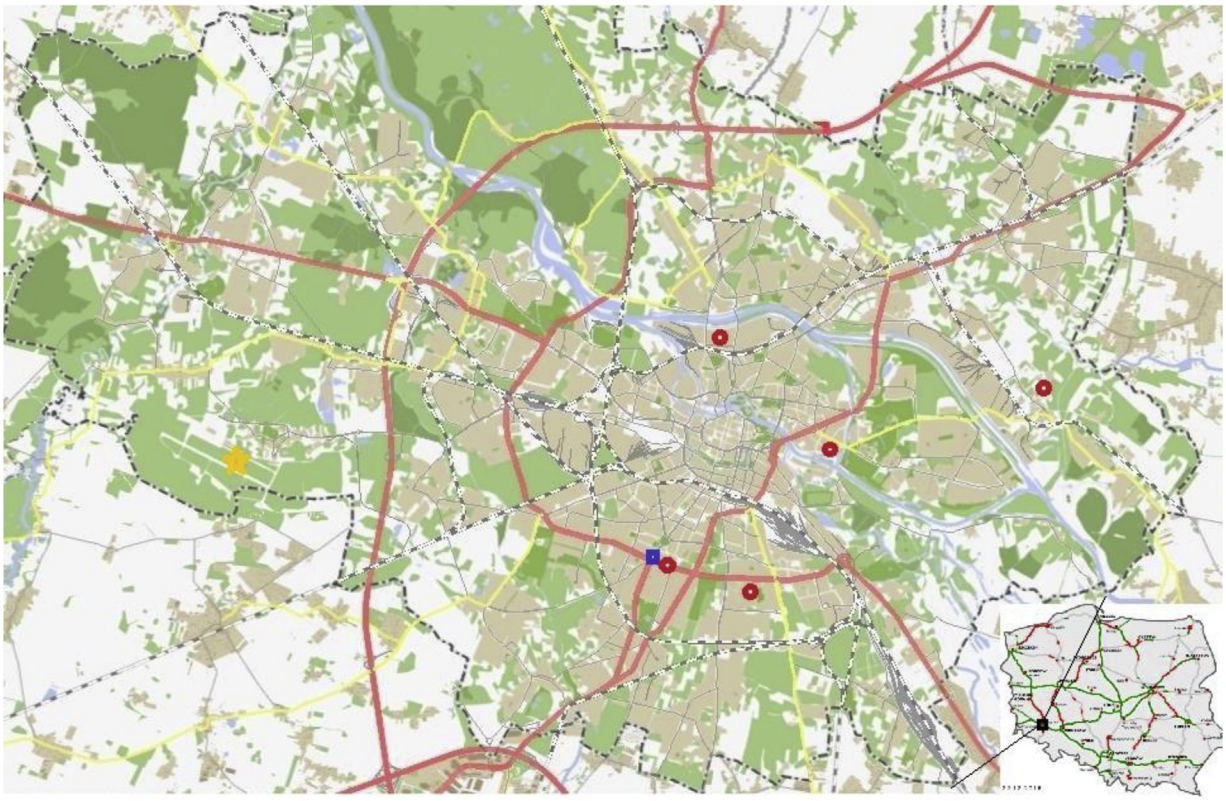

Fig. 1. Traffic, air and meteorological monitoring sites in Wrocław: the Hallera traffic counter (blue), pollution stations (red) and meteorological station (yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

pollution data further constrained the study period to the two full years from 1 January 2015 to 31 December 2016.

In the analyses described here we use hourly sums of vehicle numbers. The largest traffic volume at an intersection having an air quality measurement station in the vicinity is found close to the city centre, on Hallera street (Fig. 1). This is the place where an air quality measurement station for the transport corridor and a traffic measurement point are located closest to each other. The box plots shown in Fig. 2 represent averaged traffic variables at the Hallera intersection on three different time scales: daily, weekly and monthly.

The daily variation in traffic volume is bimodal, with peak periods in the morning between 7.00 and 8.00 and in the afternoon between 15.00 and 17.00. During night time the traffic flow is significantly lower. The extreme upper points appearing between 11.00 and 21.00 represent traffic volumes on Sunday, 21 June 2015, when a large cultural event took place in the city. This day was identified as an outlier on the daily graph. There is a clear reduction in traffic at weekends. No significant variation in volumes was recorded between successive months of the year. This results from the location of the intersection on a main road leading out of the city and on the city centre ring road.

2.2. Pollution

Pollution data are collected by the Provincial Environment Protection Inspectorate, which operates five measurement stations measuring the concentrations of different pollutants (marked in Fig. 1). In this study we focused on NO_2 and $PM_{2.5}$, which are measured at hourly intervals. In the period analysed, NO_2

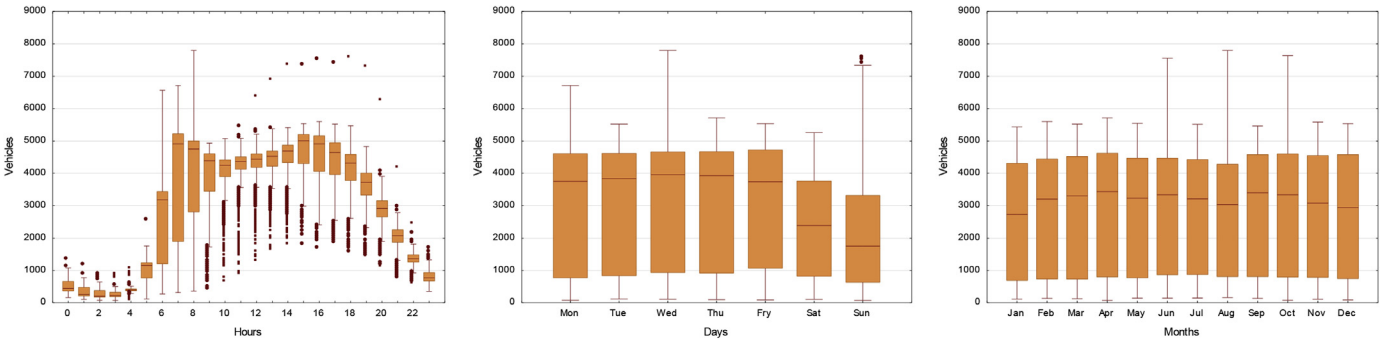

Fig. 2. Box plots of traffic volumes at the Hallera intersection for different time scales.

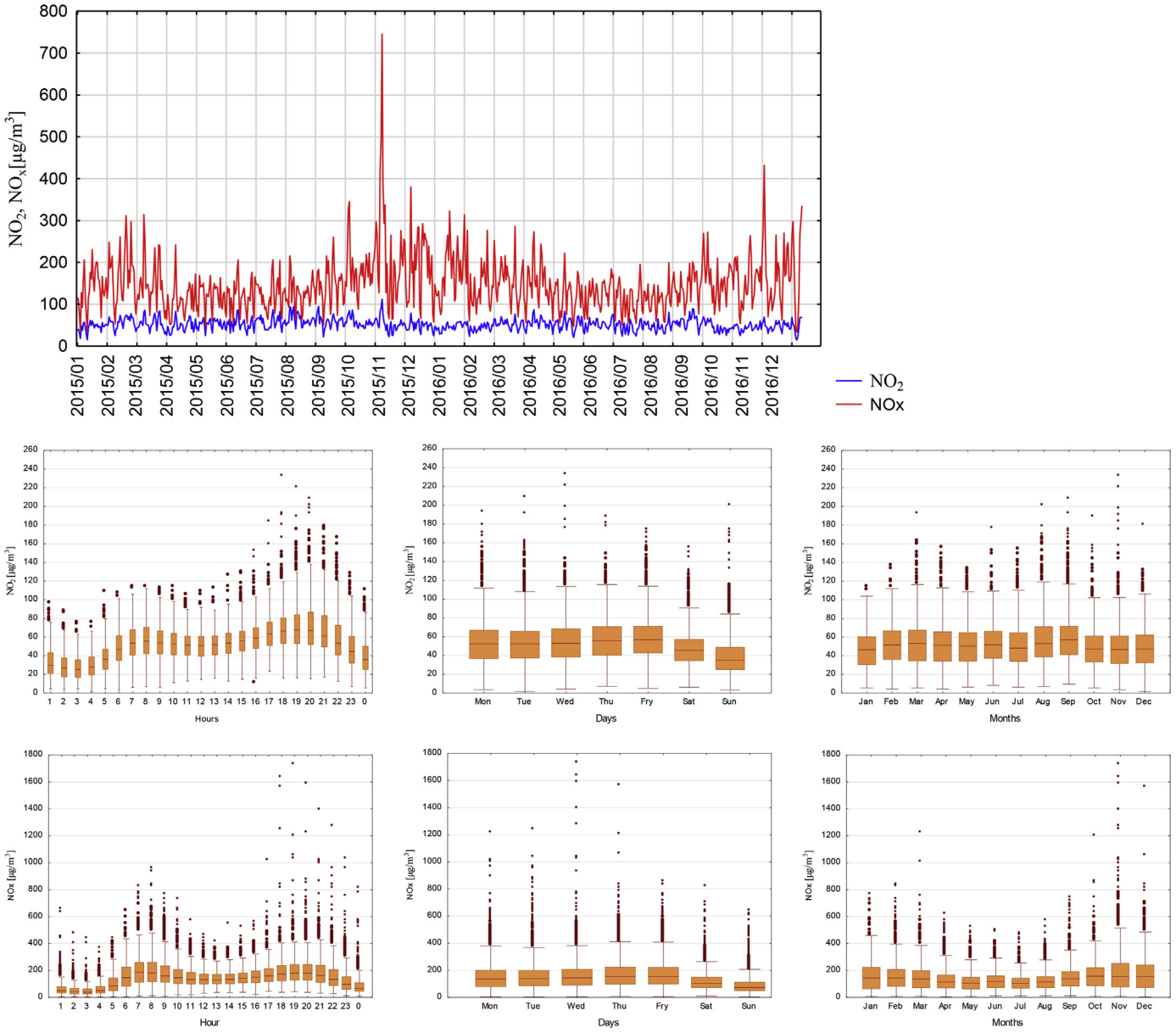

Fig. 3. Time series and box plots for NO_2 and NO_x concentrations measured at the Hallera station.

measurements were made at three stations, and $PM_{2.5}$ measurements at two stations.

To provide an example of the variation in NO_2 over time, we summarise the measurements of its concentration made at the Hallera station (Fig. 3). The permissible atmospheric concentration of nitrogen dioxide in Poland is 200 µg/m^3, and this value must not be exceeded more than 18 times in a year. The alarm level of atmospheric NO_2 is 400 µg/m^3 maintained for at least three consecutive hours (Regulation of the Minister …). In the analysed period, the permissible level was exceeded on two occasions, but the alarm level was not attained. There is marked variation in NO_2 concentration over the course of a day (Fig. 3), with higher values corresponding to peak traffic periods. The morning peak in the concentration of nitrogen oxides occurs between 7.00 and 9.00, and the afternoon peak between 18.00 and 21.00, the latter being three hours later than the traffic peak (Fig. 2). Concentrations are found to be lower at weekends, again corresponding to reduced traffic volumes. Similar daily and weekly variation is found in NO_x concentrations, although these values may be as much as seven times higher.

Polish law requires that the annual average $PM_{2.5}$ concentration be maintained at a level not exceeding 20 µg/m^3. In 2015 and 2016, for the Hallera station, located at a main city centre road intersection, the values were 30.3 and 27.5 µg/m^3. Atmospheric $PM_{2.5}$ concentrations take markedly higher values in the winter months (Fig. 4). They are not found to vary noticeably in the course of the day or over a week. The mechanism by which high $PM_{2.5}$ concentrations are generated in the winter period was described in Section 1.

2.3. Meteorological data

Meteorological data are provided by the Institute of Meteorology and Water Management (IMGW) at only one station, located on the outskirts of the city (see Fig. 1). The meteorological data set contains hourly air temperature, wind speed, wind direction,

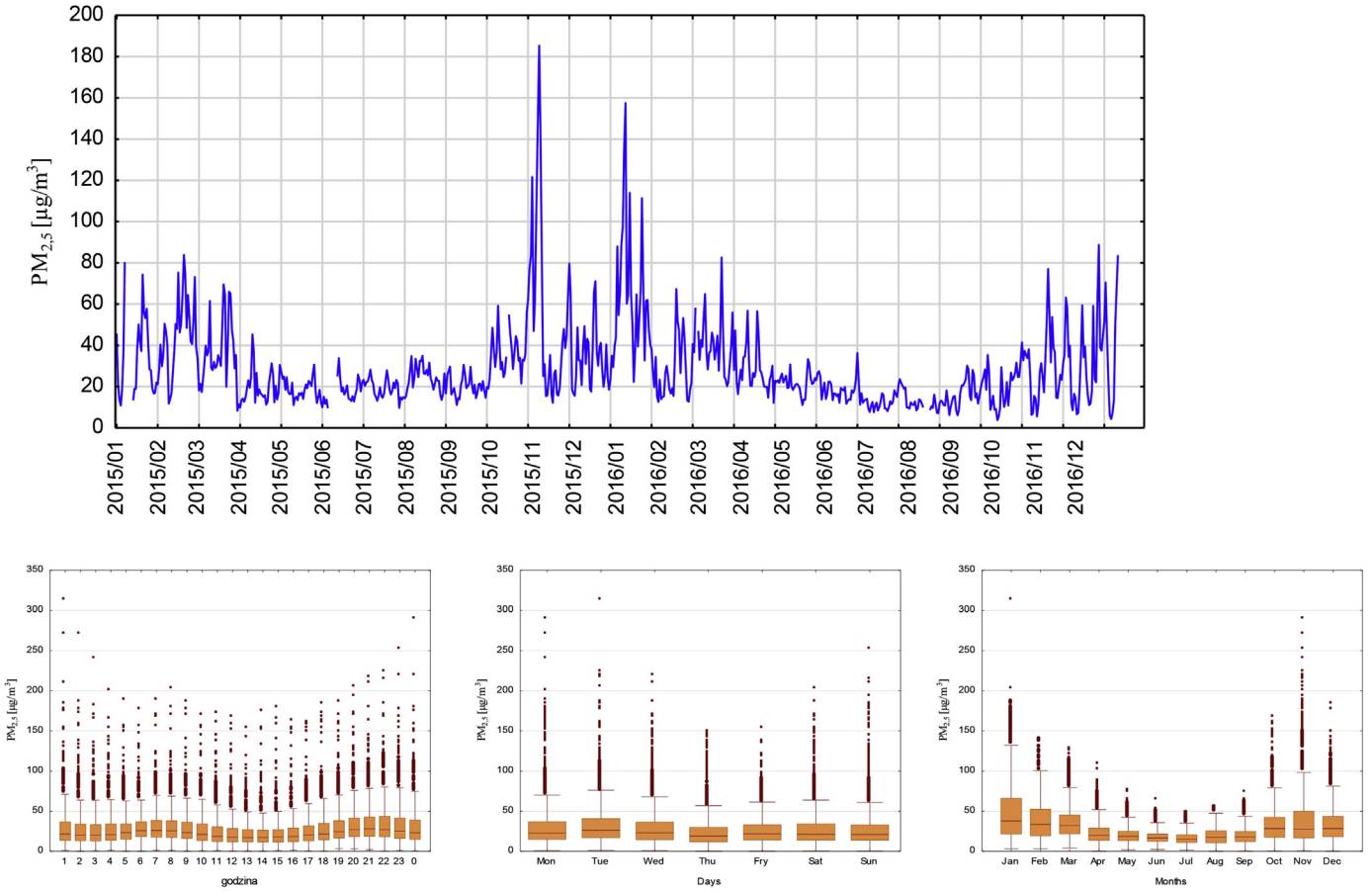

Fig. 4. Time series and box plots for $PM_{2.5}$ concentrations measured at the Hallera station.

relative humidity and atmospheric pressure.

Wrocław is situated in the temperate climatic zone of the northern hemisphere, with a transitional climate type subject to both oceanic and continental influences. Prevailing winds are westerly and south-westerly. Wrocław has climatic features typical of large industrial conurbations, in which the influence of economic activity is manifested in changes in physical features of the soil, atmospheric pollution, and the emission of artificial heat from domestic and industrial processes. The observed elevation in air temperatures (by as much as 5 °C on summer nights with fine weather) is characteristic of an urban heat island. Fig. 4 shows the time series of the collected meteorological data. There is clear seasonal variation in temperature, characteristic of temperate climates. Wrocław is one of the warmest cities in Poland, with an average annual temperature of 9.1 °C. In the winter of 2015/2016 temperatures fell to slightly below −10 °C; that winter, like the preceding ones, was regarded as warm. In each of the years studied, summer temperatures rose above 30 °C on several occasions. Wrocław does not experience strong winds: the maximum observed wind speed was 15 m/s, and the average wind speed is 3.1 m/s. Frequencies of wind directions are depicted on a wind rose (Fig. 5.).

2.4. Regression model

A random forest (RF) consists of a set number of simple decision trees. Each of the component trees forming an RF uses a sample subset of the available data. These subsets are independent, and a given instance may occur in several subsets (sampling with replacement). For each tree, predictors are selected with equal probability. Each weak tree is taught on a different sample subset. The predicted output is taken by aggregating and averaging the individual predictions of all such component 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). In each of the models discussed here, 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 (that with the highest resubstitution sum) is assigned an importance of 100. It should be noted that a different understanding of the importance of predictors is presented by Breiman (2001). The main difference is that in the method used here, ΔR values are summed for all predictors over all nodes (and trees), not only at the nodes where the variable in question participates in the division (or is a substitution variable). An advantage of this approach is that it helps to identify variables which have significant predictive power with respect to the dependent variable, but did not participate in any division (Breiman et al., 1984).

As was stated in the introduction, nine sets of cases were defined, dependent on the time for which the analysis was performed. The number of valid variable instances in each case is given in brackets:

- full two years – 1 (17598)

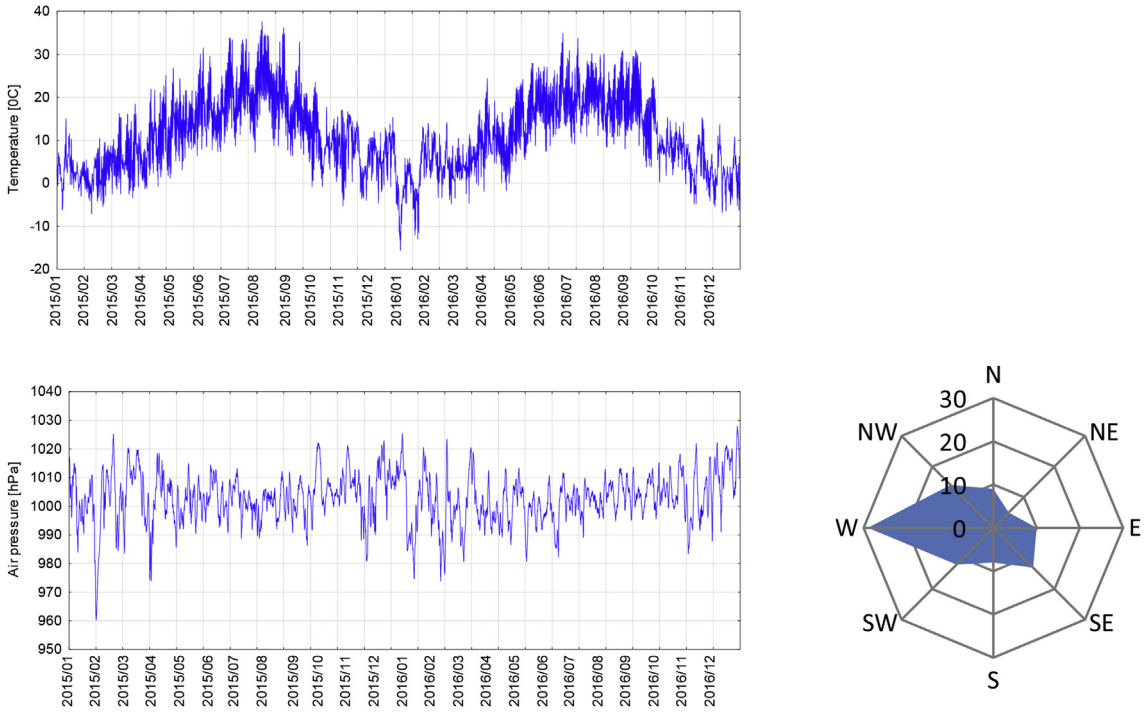

Fig. 5. Time series of hourly meteorological variables: temperature, air pressure and wind direction.

- combined warm seasons (April–September) of 2015 and 2016 – 2 (8774)
- combined cool (heating) seasons (October–March) of 2015 and 2016 – 3 (8724)
- working days – 4 (12121)
- non-working days – 5 (5477)
- spring (March–May) – 6 (4412)
- summer (June–August) – 7 (4407)
- autumn (September–November) – 8 (4357)
- winter (December–January) – 9 (4323)

Nine explanatory variables were used in the models, grouped into thematic categories:

- traffic volume;
- temporal features (day of the week, public holidays, month);
- meteorological conditions (air temperature, wind speed, wind direction, relative humidity, air pressure).

The 'public holidays' variable identifies days which are national holidays or Sundays.

The output of the model is the concentration of one of the three components of air pollution: NO_2, NO_x, $PM_{2.5}$.

Among the variables listed, four are categorical: day of the week, public holidays, month, and wind direction. Wind direction data were initially obtained in continuous numerical form, but it was not appropriate to use the wind direction in degrees as an explanatory variable, because values with a large difference may correspond to a very similar direction (for example, 1° and 360°). For this reason, wind direction was instead expressed using eight categories with 45° separations (N, NE, E, etc.).

For the construction of each tree, five explanatory variables were sampled from the set of nine variables described above. Exceptions were the random forests constructed for cases 4 and 5, for which the 'public holidays' variable was removed from the set, since its variability was too small. Along with the reduction of the number of available variables to eight, the number of predictors used was reduced to four. For the sake of statistical significance, three train–test splits were made for each period (1–9). The training set consisted of 50% of all samples, and the test set of 30%. Thus each train-split consisted of different randomly selected cases making up 50% of the dataset. It was decided that the learning process (the addition of further trees) would be stopped when for 10 cycles the error was below 5%. This condition determined the number of trees (caused the process of creating further trees to be stopped) in only two cases: for NO_x in 1 (90, 100, 80 trees) and for $PM_{2.5}$ in 1 (50, 50, 60 trees). Given that the number of variables was 9 (8), the number of predictors randomly selected for the construction of a tree was 5(4), and consequently the number of possible subsets of the variables was 126 (70), the number of trees was limited to a maximum of 200. The goodness-of-fit coefficient and feature importance were determined as the arithmetic mean of three computed values. To ensure that the models have adequate predictive power and that valid conclusions can be drawn from the results, an investigation was made of linear dependences between the input variables. For all pairs of variables, scatter plots were produced to enable visual evaluation of the existence of collinearity. Next the categorical variables were coded numerically, and values of the Pearson correlation coefficient r were determined (Table A.1). In view of the large number of data points ($N = 17332$) statistical significance was not tested because for large N the null hypothesis is practically always rejected, even if logically contradictory (Rao and Lovric, 2016). The acceptable limit for values of the correlation coefficient was established using the coefficient of determination r^2. It was assumed that if one of the variables explains not more than 10% of the variation of the other ($r^2 \leq 0.1$, $|r| \leq 0.316$), then any

dependence between them will not endanger the correctness of the model and the conclusions. There were identified two pairs of variables requiring detailed investigation: relative humidity–air temperature and relative humidity–traffic flow. The first dependence originates from the definition and method of determining relative humidity. The primary meteorological quantity is the current water vapour pressure in the air, which together with the current air temperature (determining the maximum water vapour pressure) gives the relative humidity. This is an exponential relationship, hence the correlation coefficient of −0.6 computed in the analysed temperature range (the higher the air temperature the greater the ability of the atmosphere to take up water; this implies a higher maximum pressure, and thus a lower relative pressure). A scatter plot, showing the 95% confidence interval, appears in Fig. A1. The very small width of the confidence interval confirms the absence of a significant linear relationship. The second pair of variables considered, with a Pearson correlation coefficient of −0.44, is relative humidity and traffic volume. There is no logical reason for such a dependence. On the scatter plot (Fig. A2) it is seen that almost all values lie outside the 95% confidence interval, implying that there is no significant dependence. It can thus be concluded that the correlation of relative humidity with other variables is not sufficient to generate invalid conclusions.

3. Results and discussion

Air pollution in large cities is produced to a significant extent by road vehicle emissions, with the modifying influence of meteorological conditions (Laña et al., 2016). This effect is also observed in Wrocław. The time series graphs 3 and 4 show peaks on the first days of November, indicating significantly greater concentrations of both nitrogen oxides and particulates ($PM_{2.5}$). The concentrations of the analysed pollutants increased successively from 2 November to 4 November. The traffic flow on these days did not differ from that recorded on other days. The causes of the increasing pollutant concentrations should therefore be sought in the meteorological conditions. These were days with wide temperature fluctuations with amplitudes of up to 20.6 °C (from −2.2 to 17 °C, from −5.2 to 15.4 °C, and from −3.9 to 14.2 °C), with little wind (wind speed 0–2 m/s) and with a constant pressure of 1006 hPa. On the following day the amplitude of temperature fluctuation fell to 10 °C, and the wind speed rose to 3 m/s. The reason for the poor air quality on 2–4 November 2015 was the combination of increased burning of fuels for heating purposes with unfavourable meteorological conditions, leading to the accumulation of pollutants in the lower layer of the atmosphere.

The variables discussed in Sections 2.1–2.3 were used in the construction of models as described in Section 2.4. Random forest regression models were built for each of the 27 datasets. The goodness of fit of each model was evaluated using the following coefficients of determination: R^2, MFB, MAD, MAPE. Popular information criteria such as BIC and AIC were not considered, owing to the fact that the number of variables in the model was predefined and constant. Comparison of the computed values of coefficients makes it possible to evaluate which model is best fitted to the data. The coefficient of determination R^2 is one of the fundamental measures of a model's goodness of fit. It takes values in the range ⟨0,1⟩: the closer it is to 1, the smaller are the differences between the estimated values of the dependent variable and the empirical values. Other measures of fit, independent of the mean value, include MAD (mean absolute deviation error), MAPE (mean absolute percentage error) and MFB (mean fractional bias)

Table 1
Goodness of fit.

	Equation
R^2	$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
Mean fractional bias	$MFB = \frac{1}{N}\sum_{i=1}^{N}\frac{\widehat{y}_i - y_i}{\frac{1}{2}(\widehat{y}_i + y_i)}$
Mean absolute deviation error	$MAD = \frac{1}{N}\sum_{i=1}^{N}\lvert\widehat{y}_i - y_i\rvert$
Mean absolute percentage error	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert\widehat{y}_i - y_i\rvert}{\lvert y_i\rvert}$

Where $\widehat{y}_i$ is the ith theoretical value (from the model), y_i is the ith empirical (real) value, $\bar{y}$ is the mean empirical value, and N is the sample size.

(Table 1). MFB is a measure recommended in the literature for use in the analysis of pollutant concentrations (Boylan and Russell, 2006) because it builds upon the concept of bias, which measures the tendency of a model to over- or underpredict. MAD denotes the mean absolute error, that is, the mean difference between the empirical and modelled values. MAPE is a similar measure to MAD, but it represents the mean relative error. The mathematical formulae for these coefficients of goodness of fit are given in Table 1. Their values, for each of the models described in Section 2.4, are given in Table 2.

It is found that the values of R^2 are generally low (under 0.57). The MFB values are less than 0.2, which indicates the fair performance of the models for all pollutants. The mean absolute deviation errors and mean absolute percentage errors indicate that the models are most accurate for NO_2. The values of MAPE are close to 30%, which is an acceptable relative error. In the models for NO_x the MAD and MAPE values are larger, reaching 51.9%. The models proved least accurate for $PM_{2.5}$, where the MAPE ranged from 41.2% to 78.9%, calling into question the appropriateness of modelling particulate matter concentrations in this way for periods with highly differentiated meteorological conditions (all years, autumn). In any case, the results provide a comparative insight on the predictability of pollution based on traffic and meteorological conditions. The best fits were obtained for models 7 (summer season) and 2 (warm season). This provides justification for the division of the studied period into subperiods based on atmospheric conditions. The summer season proved to be more predictable; that is, air quality is then more strongly dependent on traffic, temporal and meteorological conditions. For the other climatic subperiods (3, 6, 8, 9) the errors in the fit of the model to the empirical values are larger. Out of the analysed pollutants, the best fits were obtained for NO_2 in all periods. For the shortest series 4 (holidays), the fit of the models is equally poor as for the longest series 1, which includes the whole set of measurements for 2015–2016. It can be concluded that for long time series of measurements, the fit of the models is significantly poorer than for subsets of time series based on a defined feature (here the climatic season of the year). No significant relationship was found between the type of day (working or non-working) and the effectiveness of modelling of pollutant emissions.

Fig. 6 shows the relative importances of the variables, as proposed by the author. These are expressed in terms of share of validity (SoV) – the percentage contribution to the total sum of importance in a given model. This operation reveals the relative importance of the predictors in the model, and thus their structure and contribution to the process of explaining the variation in the dependent variable. A side-effect, however, is the masking of information as to which variable is the most important. Thus, to achieve a full picture of the importance of the predictors in the

Table 2
Values of goodness of fit coefficients for the three pollutant concentrations in the analysed time subsets [1] – [9].

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
NO_2	R^2	0.478	0.518	**0.542**	0.509	0.405	0.517	0.512	0.445	**0.569**
	MFB	0.0635	0.0732	0.0642	0.0689	0.0853	**0.0585**	0.0740	**0.0408**	0.0744
	MAD	12.75	12.63	11.28	12.49	**11.21**	12.40	12.60	12.91	**10.82**
	MAPE [%]	32.0	29.6	31.1	29.8	34.5	**29.4**	29.7	**29.4**	31.4
NO_x	R^2	0.443	**0.504**	0.481	0.453	0.394	0.473	**0.516**	0.333	0.496
	MFB	0.1270	0.1057	0.1419	0.1203	0.1387	0.1133	**0.0966**	**0.0993**	0.1593
	MAD	48.83	36.68	56.56	50.89	**32,05**	40.93	**32.32**	60.53	55.29
	MAPE [%]	46.8	**36.6**	50.1	43.7	43.9	41.1	**34.7**	43.8	51.9
PM 2.5	R^2	0.444	0.306	0.409	0.435	**0.576**	0.446	0.268	0.388	**0.521**
	MFB	0.1563	**0.1039**	0.1744	0.1449	0.1490	0.1173	**0.1068**	0.1861	0.1216
	MAD	11.08	**6.32**	15.22	11.53	10.15	8.76	**5.72**	12.87	14.65
	MAPE [%]	72.8	54.9	73.3	51.2	51.7	**41.9**	**50.9**	78.9	55.6

Bold – the best two values of each coefficient.

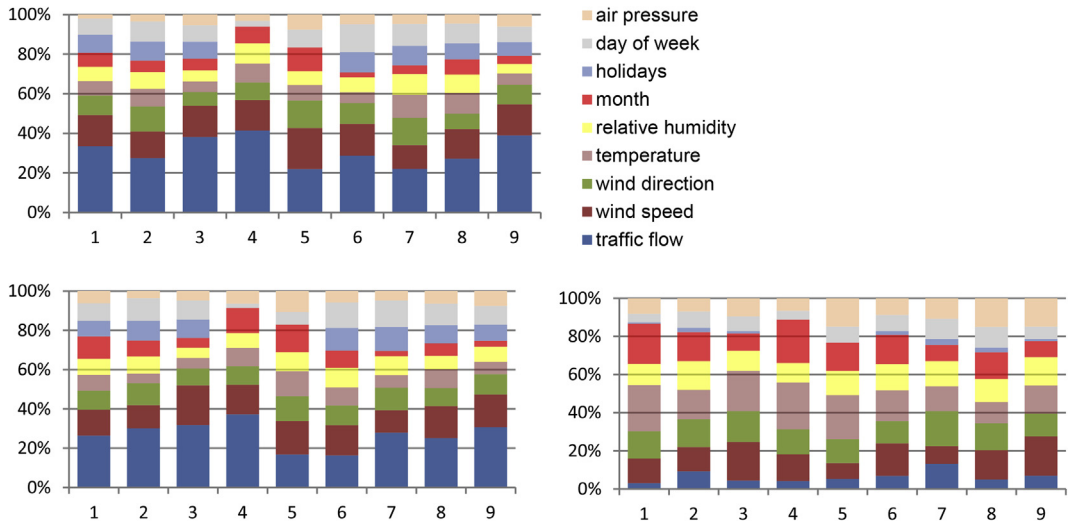

Fig. 6. Shares of validity for time subsets [1] – [11], for the pollutants NO_2 (top), NO_x (bottom left) and $PM_{2.5}$ (bottom right).

model, it is necessary to examine in parallel the importance values and the aforementioned percentage shares of validity (SoV). In models of the concentration of nitrogen oxides, the most important variable is traffic flow, except in model [5] (holidays), where the most important (with an importance of 100) was wind speed, and traffic flow had an importance of 95. The next most important variables are wind speed and direction, where the importance of wind speed is greater by between 43 (for non-working days) and 2 (for the warm season) then wind speed. The least significant variable in five models out of nine was air pressure. There is a visible difference in the structure of the importances of the variables between models [2], [7] and [3], [9]. The first two relate to warmer times of the year, respectively the warm season and the summer season. In these the variables have relatively even shares of validity: for the most important variable, traffic flow, the SoV is respectively 27% and 22%, while for the least important, month and air pressure, the values range from 9% do 13%. In the models for the cool season and the winter season the SoV values are more highly differentiated: again the most important variable is traffic flow, this time with SoVs of 38% and 39%, followed by wind speed, for which the SoV is less than half of that of traffic flow (16% in each case). The remaining variables have very similar levels of impact, between 4% and 9%. The most balanced influence of the predictors is found in the model with the smallest dataset, [5] (non-working days), where the values ranged from 6% do 22%, this being a distinctive feature of the importance of the predictors in the modelling of both NO_2 and NO_x. In that model the concentrations of nitrogen oxides are found to be affected to a greater degree by non-traffic variables, showing that those variables are of greater importance relative to road traffic in generating pollution than in the case of other measurement series. This is certainly linked to the smaller traffic flows occurring on non-working days and in winter, which lead to lower exhaust gas emissions. In the modelling of NO_x, models [4] and [5] exhibit the highest share of validity for air pressure and the lowest for day of the week out of all of the considered models. The low significance of day of the week is obvious in view of the small variation in values of that explanatory variable, particularly in model [5] (non-working days), since the great majority of non-working days (apart from national holidays) are Saturdays and Sundays. Models [4] and [5] also feature respectively the largest (38%) and smallest (17%) SoV values for traffic flow, in spite of the fact that the importance of that variable is 100 in the first of these models and 95 in the second. This results from the more balanced importance of the other variables in model [5], which increases the sum of the importances of all variables, thus lowering the SoV value.

The effect of the explanatory variables is notably different in the case of the $PM_{2.5}$ concentration. Here the greatest effect comes from temperature, followed by wind speed, wind direction and month.

The 'holidays' variable has the lowest importance. For the cooler periods [3] and [9] the SoV for the temporal variables was respectively 17% and 14%, compared with 26% and 22% for the warmer periods [2] and [7], while the importance of the 'holidays' variable is negligible (below 3%). The traffic variable attains its highest SoV in the $PM_{2.5}$ models built for the warmer periods [2] and [7] (10% and 13% respectively). This is linked to the fact that particulate emissions from the domestic sector are much lower in the warm season than at other times of year. For other periods, the particulate concentration is much less (if at all) dependent on traffic flow, and thus the importance of the latter variable is lower.

Feature importance for the predictors for four selected subperiods (seasons), determined by the method described in Section 2.4, are shown in Fig. 7. The variables were grouped into meteorological conditions, temporal conditions, and finally traffic flow. There is a similarity in the significance of the explanatory variables in the modelling of nitrogen oxide concentration. As was noted above, traffic flow has the greatest importance (for all seasons of the year), while meteorological conditions and temporal conditions are similar in importance, the former set having somewhat higher values. Exhaust gas emissions from vehicles increase the pollutant load in the air. Meteorological conditions may reinforce or weaken this effect. The flat relief of Wrocław, combined with the density of buildings, which obstruct the flow of air through the city, reinforces the effect of traffic flow on nitrogen oxide concentrations. The greatest effect of traffic flow on concentrations of nitrogen oxides (NO_2 and NO_x), compared with other factors, occurs in winter. The importance of wind speed in the modelling of nitrogen oxide concentrations does not depend on the season; its impact is approximately half that of traffic flow (feature importance from 41 to 61). In the summer period meteorological variables were found to have relatively the greatest effect on NO_2 concentration. This is linked to the process of transformation of NO and NO_2 into ozone via reactions with volatile organic compounds under the influence of the sun - chiefly in high-pressure summer weather conditions: warm sunny days with low wind (Rethinking the Ozone Problem …, 1991).

For $PM_{2.5}$ the importances of the variables are distributed very differently. The effect of meteorological conditions is the largest (being about twice as large as in the nitrogen oxide models), followed by that of temporal conditions; the least significant factor is traffic flow. Only for the summer period is the significance of traffic flow comparable to that of the meteorological variables. This results from the low level of emissions from domestic heating during this period, which means that vehicle emissions become noticeable. This finding is in accordance with analyses performed for other cities (Z. Zhang et al., 2015; Laña et al., 2016). Aldrin and Haff (2005) also obtained a significantly different distribution of importances of variables when modelling nitrogen oxide concentrations as compared with $PM_{2.5}$. The most significant variable for all of the pollutants studied by them in Oslo was the number of vehicles; for nitrogen oxides the next most important variables were wind direction and speed, whereas for $PM_{2.5}$ the next most important was a temporal variable (day number).

4. Conclusions

The division of the studied period into subperiods based on thermal conditions has been shown to be justified, and can bring to light relationships which are not visible for a dataset covering entire calendar years. For better mapping of reality by mathematical models using random forests to describe the dynamically changing concentrations of NO_2, NO_x and $PM_{2.5}$ and their relationship with the also highly variable meteorological and traffic conditions, it is essential to select an appropriate time series of data and to divide it into suitable subperiods. The division into subperiods should be based on analysis of the climatic conditions of the region in question. In the case of Wrocław there is a clear

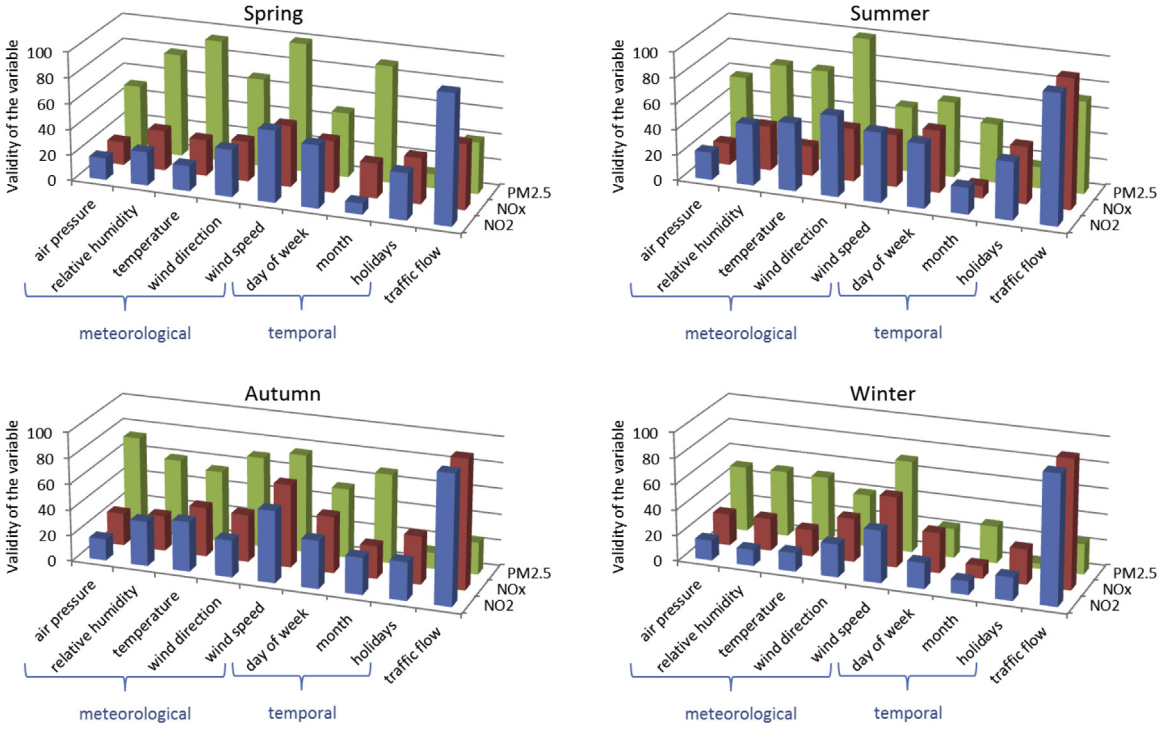

Fig. 7. Feature importance of each variable for three pollutants.

division into four seasons of the year – spring, summer, autumn and winter – where spring and autumn are transitional periods between the warm or hot summer and the cool or cold winter, which is part of the atmospherically cooler domestic heating season. The application of such a division proved to be justified. The best model fit was obtained for the warm season (April–September) and for the summer season (June–August). For other cities it may be necessary and justified to apply a different division, corresponding to the prevailing climatic conditions and possibly a different type of seasonality. In the case of London, a division into cold and warm periods was found to have no effect in an analysis of variation in $PM_{2.5}$ levels (Kassomenos et al., 2014). A division into subperiods consisting of working or non-working days was not found to have any justification. The weekend effect applies to ozone, and does not transfer to the pollutants studied in that work. The values of the goodness of fit coefficients obtained for the models in this study are comparable to those obtained by Laña et al. (2016). They show that the degree to which the models fit actual conditions is good, but not entirely satisfactory. The largest errors occur for the periods (hours, days) with very high pollutant concentrations, which the model fails to explain. This problem applies to all of the methods considered by the author and discussed in Section 1. Further research is necessary to identify peaks of pollutant concentrations and possibilities of modelling and predicting them. The dataset may be divided into "small" and "large" values, and each of the sets can be modelled separately. A problem will undoubtedly be determining the level at which the split is to be made. Another possibility is to model only the maximum values (daily, for example); it will then be important to determine the input variables so as to minimise information loss.

In the modelling of nitrogen oxide concentrations, the most important variable is traffic flow. However, for the modelling of $PM_{2.5}$ the most important are meteorological conditions: wind speed, wind direction and temperature. The presentation of the importances of variables in a model in the form of a share of validity (SoV) graph makes possible the rapid evaluation and identification of the structure of the importances of the predictors.

The conclusions drawn from this work may be put to effective use in, among others, decision support systems applied in the management of road traffic and road developments, which have recently been growing in popularity (Kazak et al., 2018; Malytska and Balabukh, 2018; Tribby et al., 2013).

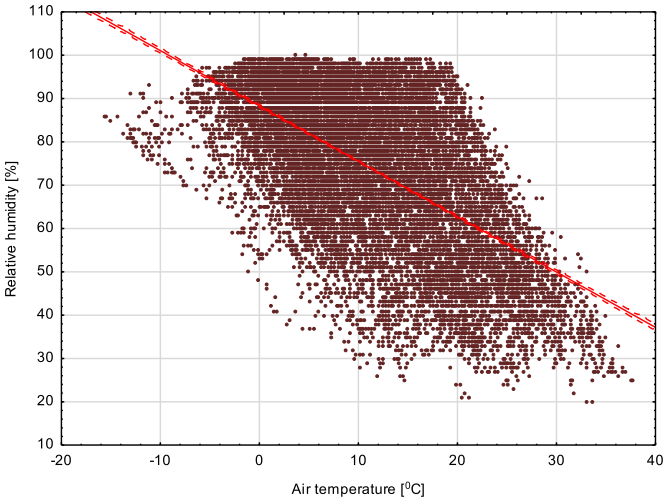

Fig. A.1. Scatter plot with linear regression and 95% confidence area.

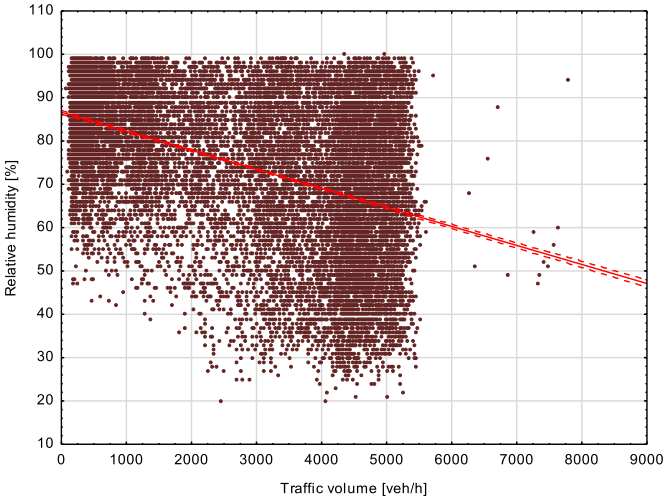

Fig. A.2. Scatter plot with linear regression and 95% confidence area.

Appendix

Table A.1
Pearson correlation coefficients - *r*.

	traffic volume	air temp.	wind speed	wind direction	relative humidity	air pressure	public holidays	month	day of the week
traffic vol.	1,00								
air temp.	0,25	1,00							
wind speed	0,20	0,01	1,00						
wind direct.	0,05	−0,02	0,15	1,00					
rel. humidity	**−0,44**	**−0,60**	−0,27	0,03	1,00				
air pressure	0,00	−0,16	−0,15	−0,01	0,05	1,00			
public holidays	−0,24	0,00	0,03	0,00	−0,01	−0,04	1,00		
Month	0,02	0,18	−0,03	−0,04	0,08	0,30	−0,02	1,00	
day of the week	−0,16	0,01	0,00	−0,02	−0,03	−0,02	0,75	−0,01	1,00

Bold: $|r| > 0.316$.

References

Adam, M., Schikowski, T., Carsin, A.E., Cai, Y., Jacquemin, B., Sanchez, M., Vierkötter, A., Marcon, A., Keidel, D., Sugiri, D., Kanani, Z.A., Nadif, R., Siroux, V., Hardy, R., Kuh, D., Rochat, T., Bridevaux, P.-O., Eeftens, M., Tsai, M.-Y., Villani, S., Phuleria, H.C., Birk, M., Cyrys, J., Cirach, M., Nazelle, A., Nieuwenhuijsen, M.J., Forsberg, B., Hoogh, K., Declerq, K., Bono, R., Piccioni, P., Quass, U., Heinrich, J., Jarvis, D., Pin, I., Beelen, R., Hoek, G., Brunekreef, B., Schindler, C., Sunyer, J., Krämer, U., Kauffmann, F., Hansell, A.L., Künzli, N., Probst-Hensch, N., 2015. Adult lung function and long-term air pollution exposure. ESCAPE: a multicentre cohort study and meta-analysis. Eur. Respir. J. 45, 38–50. https://doi.org/10.1183/09031936.00130014.

Aldrin, M., Haff, I.H., 2005. Generalized additive modelling of air pollution, traffic volume and meteorology. Atmos. Environ. 39 (11), 2145–2155. https://doi.org/10.1016/j.atmosenv.2004.12.020.

Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Comput. Stat. Data Anal. 52 (4), 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015.

Barratt, B., Atkinson, R., Ross Anderson, H., Beevers, S., Kelly, F., Mudway, L., Wilkinson, P., 2007. 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. 41 (8), 1784–1791. https://doi.org/10.1016/j.atmosenv.2006.09.052.

Bertaccini, P., Dukic, V., Ignaccolo, R., 2012. Modeling the short-term effect of traffic and meteorology on air pollution in Turin with generalized additive models. Adv. Meteorol. 2012, 1–16. https://doi.org/10.2139/ssrn.1422567.

Boylan, J.W., Russell, A.G., 2006. PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmos. Environ. 40, 4946–4959. https://doi.org/10.1016/j.atmosenv.2005.09.087.

Breiman, L., 2001. Random forests. Mach. Learn. 1, 5–32, 45. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA.

Flores-Pajot, M.-C., Ofner, M., Do, M.T., Lavigne, E., Villeneuve, P.J., 2016. Childhood autism spectrum disorders and exposure to nitrogen dioxide, and particulate matter air pollution: a review and meta-analysis. Environ. Res. 151, 763–776. https://doi.org/10.1016/j.envres.2016.07.030.

González-Aparicio, I., Hidalgo, J., Baklanov, A., Padró, A., Santa-Coloma, O., 2013. An hourly PM10 diagnosis model for the Bilbao metropolitan area using a linear regression methodology. Environ. Sci. Pollut. Res. 20 (7), 4469–4483. https://doi.org/10.1007/s11356-012-1353-7.

Hien, T.T., Linh, H.N., Luong, L.M.T., Thai, P.K., 2016. Air pollution and risk of respiratory and cardiovascular hospitalizations in the most populous city in Vietnam. Sci. Total Environ. 557–558, 322–330. https://doi.org/10.1016/j.scitotenv.2016.03.070.

Hoek, G., Krishnan, R.M., Beelen, R., Peters, A., Ostro, B., Brunekreef, B., Kaufman, J.D., 2013. Long-term air pollution exposure and cardio-respiratory mortality: a review. Environ. Health 12 (1), 43. https://doi.org/10.1186/1476-069X-12-43.

Information on air quality in Wrocław [Informacja o jakości powietrza na terenie miasta Wrocławia], 2016. Provincial Environment Protection Inspectorate [WIOS], Wrocław.

Kassomenos, P.A., Vardoulakis, S., Chaloulakou, A., Paschalidou, A.K., Grivas, G., Borge, R., Lumbreras, J., 2014. Study of PM_{10} and $PM_{2.5}$ levels in three European cites: analysis of intra and inter urban variations. Atmos. Environ. 87, 153–163. https://doi.org/10.1016/j.atmosenv.2014.01.004.

Kazak, J., Chalfen, M., Kamińska, J., Szewrański, S., Świąder, M., 2018. Geo-dynamic decision support system for urban traffic management. In: Ivan, I., Horak, J., Inspektor, T. (Eds.), Dynamics in GIScience. GIS Ostrava 2017. Lecture Notes in Geoinformation and Cartography. Springer, Cham, 195-207 https://doi.org/10.1007/978-3-319-61297-3_14.

Keeler, R.H., 2014. A Machine Learning Model of Manhattan Air Pollution at High Spatial Resolution. PhD thesis. http://hdl.handle.net/1721.1/90659.

Laña, I., Del Ser, J., Pedró, A., Vélez, M., Casanova-Mateo, C., 2016. The role of local urban traffic and meteorological conditions in air pollution: a data-based study in Madrid, Spain. Atmos. Environ. 145, 424–438. https://doi.org/10.1016/j.atmosenv.2016.09.052.

Malytska, L., Balabukh, V., 2018. Atmospheric self-cleaning coefficients as indicators of the atmospheric ability to dissipate pollutants in Ukraine. Meteorol. Hydrol. Water Manage 6 (1), 59–65. https://doi.org/10.26491/mhwm/79450.

Mlakar, P., Boinar, M., 1997. Perception neutral network-based model predicts air pollution. In: Intelligent Information Systems, pp. 345–349. https://doi.org/10.1109/IIS.1997.645288.

Pei-Chen, Lee, Liu, Li-Ling, Yu Sun, M.D., Yu-An, Chen, Chih-Ching, Liu, Chung-Yi, Li, Hwa-Lung, Yu, Beate, Ritz, 2016. Traffic-related air pollution increased the risk of Parkinson's disease in Taiwan: a nationwide study. Environ. Int. 96, 75–81. https://doi.org/10.1016/j.envint.2016.08.017.

Rao, C.R., Lovric, M.M., 2016. Testing point null hypothesis of a normal mean and the truth: 21st century perspective. J. Mod. Appl. Stat. Methods 15 (2). https://doi.org/10.22237/jmasm/1478001660. Article 3.

Regulation of the Minister of Environment of 14 August 2012 on Levels of Certain Substances in the Air [Rozporządzenie Ministra Środowiska z dnia 24 Sierpnia 2012 R. w Sprawie Poziomów niektórych Substancji w Powietrzu] (Dz.U.2012.1031).

Rethinking the Ozone Problem in Urban and Regional Air Pollution, 1991, The National Academies Press, Washington.

Sayegh, A., Tate, J.A., Ropkins, K., 2016. Understanding how roadside concentrations of NO_x are influenced by the background levels, traffic density, and meteorological conditions using boosted regression trees. Atmos. Environ. 127, 163–175. https://doi.org/10.1016/j.atmosenv.2015.12.024.

Tang, G., Zhao, P., Wang, Y., Gao, W., Cheng, M., Xin, Y., Li, X., Wang, Y., 2017. Mortality and air pollution in Beijing: the long-term relationship. Atmos. Environ. 150, 1238–1243. https://doi.org/10.1016/j.atmosenv.2016.11.045.

Tribby, C.P., Miller, H.J., Song, Y., Smith, K.R., 2013. Do air quality alerts reduce traffic? An analysis of traffic data from the Salt Lake City metropolitan area, Utah, USA. Transp. Pol. 30, 173–185. https://doi.org/10.1016/j.tranpol.2013.09.012.

Zhang, K., Batterman, S., 2013. Air pollution and health risks due to vehicle traffic. Sci. Total Environ. 450–451, 307–316. https://doi.org/10.1016/j.scitotenv.2013.01.074.

Zhang, Z., Zhang, X., Gong, D., Quan, W., Zhao, X., Ma, Z., Kim, S.-J., 2015. Evolution of surface O_3 and $PM_{2.5}$ concentrations and their relationships with meteorological conditions over the last decade in Beijing. Atmos. Environ. 108, 67–75. https://doi.org/10.1016/j.atmosenv.2015.02.071.