Residuals in the modelling of pollution concentration depending on meteorological conditions and traffic flow, employing decision trees

Joanna A. Kamińska[1,*]

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

Abstract. Two data mining methods – a random forest and boosted regression trees – were used to model values of roadside air pollution depending on meteorological conditions and traffic flow, using the example of data obtained in the city of Wrocław in the years 2015–2016. Eight explanatory variables – five continuous and three categorical – were considered in the models. A comparison was made of the quality of the fit of the models to empirical data. Commonly used goodness-of-fit measures did not imply a significant preference for either of the methods. Residual analysis was also performed; this showed boosted regression trees to be a more effective method for predicting typical values in the modelling of NO_2, NO_x and $PM_{2.5}$, while the random forest method leads to smaller errors when predicting peaks.

1 Introduction

The modelling of concentrations of air pollutants, on different scales and for different purposes, is a highly topical issue. Anthropogenic factors are the chief source of air pollution; hence there is a natural need to monitor, model and counteract such pollution, which has adverse effects on human health [1]. According to the Provincial Environment Protection Inspectorate in Wrocław, 56% of NO_2 emissions and 16% of $PM_{2.5}$ emissions are produced by road vehicles, while 81% of $PM_{2.5}$ emissions and 9% of NO_2 emissions originate from household and municipal sources [2]. Action continues to be taken to reduce surface emissions both from transport and from the municipal and household sector [3,4]. Unnaturally high concentrations of the aforementioned substances in the air chiefly affect respiratory and cardiovascular health [5–7]. Research has also shown that air pollution may be a cause of autism in children [8] and of Parkinson's disease [9], and consequently may even lead to death [10]. Pollution models can help traffic managers to take decisions efficiently, by selecting the most adequate traffic management strategy [11] or decision support system [12]. They also enable assessment of the capacity of the atmosphere for self-cleaning [13].

The main input for the models described in the literature is traffic and meteorological data [14–16]. There are also some researchers who use only traffic data [17] or only meteorological data [18]. Laña et al. [19] investigated the effect of the choice of explanatory variables (traffic, meteorological and temporal) on the correctness and fit of the model; they obtained comparable results for sets of temporal and meteorological variables and for sets expanded to include traffic data.

The input data used in the present study consist of information on meteorological and traffic conditions, as well as temporal variables, from the years 2015–2016. With the development of computational techniques and machine learning, an ever greater number of models is becoming available. The popular and still developing multidimensional regression models – originally linear, but now more complex – describe the relationships between variables in an effective manner. González-Aparicio et al. [14] used 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. Betraccini et al. [20] and Aldrin and Haff [21] proposed the use of a generalised additive model for modelling the short-term effects of traffic and weather on air pollution. Machine learning, which continues to be developed, has also been applied in the modelling of air pollution concentrations. The method of boosted regression trees (BRT) is one of the classification and regression methods based on decision trees. Sayegh et al. [16] used boosted regression trees to investigate how roadside concentrations of NO_x are influenced by the background levels, traffic density, and meteorological conditions. Even more computationally complex is the random forest (RF) method, as used in [19], where the procedure involves the compilation of information from multiple decision trees simultaneously [22].

A fundamental problem arising in modelling is the quality of the fit of the model, as measured by various goodness-of-fit coefficients. Even if the overall fit is

* Corresponding author: joanna.kaminska@upwr.edu.pl

good or very good, the model may fail to estimate the concentration peaks correctly. The problem may be approached in two ways: in terms of the short-term forecasting of pollution concentration [23–25] or in terms of the multidimensional modelling of dependences in the search for a model that will identify a set of conditions generating the actual value of pollutant concentration – in other words, that will effectively predict the concentrations based on easily available values of input variables. The first approach, important from an environmental or public service standpoint, is based on short-term forecasts, which can be highly accurate when made for one hour ahead, for example [23]. The second approach is the one relevant to the present study, where two machine learning models – a random forest and boosted regression trees – are constructed to determine the effect of meteorological, temporal and traffic flow variables on the concentrations of the atmospheric pollutants NO_2, NO_x and $PM_{2.5}$. The models are compared in terms of quality of fit to the empirical data, and reference is made to other results reported in the literature. A key part of this work is the comparison of the constructed models in terms of fit errors. The models also underwent verification using data from the year 2017.

2 Material and methods

2.1 Data

The analysis is based on hourly data obtained in the city of Wrocław (southwestern Poland) in the years 2015–2016.

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). Marked on Fig. 1 is the camera site used in the present analysis: the intersection of Hallera and Powstańców Śląskich.

Pollution data are collected by the Provincial Environment Protection Inspectorate, which operates five measurement stations measuring the concentrations of different pollutants (marked on Fig. 1). In this study we focused on NO_2, NO_x and $PM_{2.5}$, which are measured at hourly intervals. There were 17,332 data points for nitrogen oxide concentrations, and 17,003 for particulate matter.

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 dataset contains hourly air temperature, wind speed, wind direction, relative humidity and atmospheric pressure.

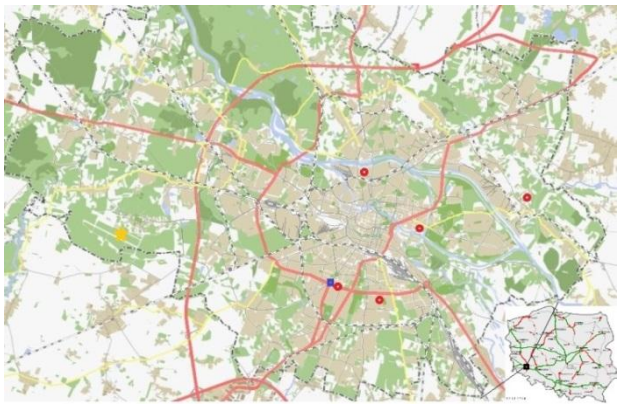

Fig. 1. Traffic, air and meteorological monitoring sites in Wrocław: the Hallera traffic counter (blue), pollution stations (red) and meteorological station (yellow).

Table 1. Descriptive statistics for continuous predictors and pollution concentration [μg/m^3] in the years 2015-2016.

Variable	$\bar{x}$	min	max	s
NO_2 concentration	51.46	1.70	231.56	23.67
NO_x concentration	147.48	3.90	1728.03	107.80
$PM_{2.5}$ concentration	28.85	0.00	311.80	23.96
Traffic flow [veh/h]	2787	76	7797	1807
Temperature [°C]	10.85	-15.70	37.70	8.38
Wind speed [m/s]	3.10	0.00	15.00	1.94
Relative humidity [%]	74.43	20.00	100.00	17.79
Air pressure [hPa]	1003	960	1028	8.54

$\bar{x}$ – mean value; s – standard deviation

2.2 Boosted regression trees

The principal idea of boosted regression trees (BRT) is the creation of a series of simple binary trees consisting of a root and two descendants (one division), where each successive tree is constructed to predict the residuals generated by the preceding trees [26,27]. At successive algorithm boosting steps, a single (best) division of the data is determined, and the deviations of the observed values from the means (the residuals in the division) are calculated. At the next division, the algorithm works on the deviations obtained as a result of the previous division. The method of stochastic gradient boosting used in the algorithm means that each subsequent tree is constructed on the basis of a random sample containing 50% of the entire dataset. Thus subsequent trees are constructed to predict the residuals in independently chosen samples. The introduction of a certain degree of randomness into the analysis is intended to prevent overtraining, and leads to models with the property of generalisation and good predictive accuracy. The described algorithm leads to a good fit between predicted and observed values even if the relationship between the predictors and the dependent variable is highly complex in nature (non-linear, for example). The use of decision trees with the C&RT method of division in an exhaustive search for single-dimensional divisions enables the quantitative evaluation of the importance of variables as a sum, over all nodes of the tree, of increases in the resubstitution estimate, and the expression of this value

as a fraction of the maximum sum (over all predictive variables).

The importance of the variables was determined by the procedure described in [28]. A key advantage of regression trees is the possibility of including qualitative variables in the set of explanatory variables.

2.3 Random forest

A random forest (RF) consists of a set number of simple decision trees. Each of the component trees in an RF uses a sample subset of the available data. These subsets are independent, and the same instance may occur in multiple subsets (sampling with replacement). For each tree, the predictors are selected with equal probability. Each weak tree is trained on a different sample subset. The predicted output is obtained by aggregating and averaging the individual predictions of all such compounding trees. This particular 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 [29]. In each of the models described, the importance of the predictive variables was determined as the sum – over all tree nodes – of increases in the resubstitution estimate (ΔR) and the expression of this value as a fraction of the maximum sum (among all variables; expressed as a percentage). 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 et al. [22]. 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.

The model included eight input variables, categorised as relating to:

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

The output of the model was the concentration of one of three air pollutants: NO_2, NO_x and $PM_{2.5}$.

Of the variables listed, three are categorical: day of the week, month, and wind direction. Wind direction data were originally 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 represent winds with a very similar direction (for example, 1º and 360º). For this reason, wind direction was instead expressed using eight categories with 45º separations (N, NE, E, etc.).

The training set consisted of 50% of all samples, and the test set of 30%. For the construction of each tree, five explanatory variables were sampled from the set of eight variables described above. It was decided that the learning process (addition of further trees) would stop when the error fell by less than 5% for 10 cycles. This condition determined the number of trees (stopped the process of creating further trees) only in the case of $PM_{2.5}$, when the process stopped at 90 trees. Given that the number of variables was 8, the number of predictors randomly selected for the construction of a tree was 5, and consequently the number of possible different subsets of the variables was 56, the number of trees was limited to a maximum of 100.

To increase the significance of high values of concentration in the model, in the construction of each of the decision trees making up the forest, *a priori* probabilities proportional to the value of the instance were used. This meant that the generated forest was more sensitive to high pollution values. This operation represents an approach to extreme values similar to that in the BRT scheme, where the method of creating subsequent binary trees from the residuals of the preceding tree takes account of those values on each occasion.

2.4 Goodness-of-fit measures

The following goodness-of-fit measures were used to evaluate the quality of the fit of each model: R^2, MFB, MADE and 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 better 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 MADE (mean absolute deviation error), MAPE (mean absolute percentage error) and MFB (mean fractional bias) (Table 2). MFB is a measure recommended in the literature for use in the analysis of pollutant concentrations [30] because it builds upon the concept of bias, which measures the tendency of a model to over- or underpredict. MADE denotes the mean absolute error, that is, the mean difference between the empirical and modelled values. MAPE is a similar measure to MADE, but it represents the mean relative error (expressed here in percent). The mathematical formulae for these coefficients of goodness of fit are given in Table 2. Their values, for each of the models considered, are given in Table 3.

Table 2. Goodness of fit measures.

R^2	$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
Mean fractional bias	$MFB = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{y}_i - y_i}{\frac{1}{2}(\hat{y}_i + y_i)}$

Mean absolute deviation error	$MADE = \frac{1}{N}\sum_{i=1}^{N}\lvert\hat{y}_i - y_i\rvert$
Mean absolute percentage error	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert\hat{y}_i - y_i\rvert}{\lvert y_i\rvert}$

where $\hat{y}_i$ is the *i*th theoretical value (from the model), y_i is the *i*th empirical (real) value, $\bar{y}$ is the mean empirical value, and *N* is the sample size.

3 Results and discussion

3.1 Goodness-of-fit measures

As mentioned at the outset, a fundamental difficulty in describing relationships between pollution concentration and explanatory variables is the low values obtained for measures of the goodness of fit of the models. Based on global data from 5220 air monitors located on all continents, a study using the method of land use regression with Lasso variable selection [31] produced a model for NO_2 with an adjusted R^2 goodness-of-fit measure equal to 0.52. Sayed et al. [16] constructed 112 models for the concentration of five airborne pollutants using four different sets of predictors. For explanatory variables covering the largest set of input data – meteorological conditions, temporal variables and traffic flow – they obtained R^2 values in the range 0.49–0.54 for nitrogen dioxide, 0.37–0.48 for PM_{10} and 0.33–0.44 for nitrogen monoxide. In [32] models were constructed for pollutant concentration in different time subsets, using the RF method. The obtained R^2 values included 0.57 for NO_2 in the winter period, 0.52 for NO_x in the summer period (June–August), and 0.58 for $PM_{2.5}$ on non-working days.

Based on the data and computational techniques described in section 2, BRT and RF models were constructed for each of the considered pollutants. The values of the goodness-of-fit measures are given in Table 3.

Table 3. Values of goodness-of-fit measures.

	Boosted regression trees			Random forest		
	NO_2	NO_x	$PM_{2.5}$	NO_2	NO_x	$PM_{2.5}$
R^2	0.43	0.40	0.36	**0.55**	**0.52**	**0.57**
MFB	**0.07**	**0.12**	**0.15**	0.14	0.23	0.28
MADE	13.3	**52.3**	12.1	**12.3**	52.3	**10.9**
MAPE [%]	**33**	**51**	68	34	60	**68**

Bold indicates the best goodness-of-fit measures for each pollution type.

The values of the goodness-of-fit measures do not deviate significantly from those reported in the literature. The R^2 value, as a measure of explanatory power, indicates that the model explains up to 57% (43%) of the variation in the dependent variable (for the RF and BRT model respectively). The MFB values indicate that the BRT model achieves a better fit for all pollutants. The mean absolute deviation error is 26% (24%) of the mean of concentrations of NO_2 (for BRT and RF respectively), 35% of the mean of concentrations of NO_x (for both models), and 42% (38%) of the mean value of $PM_{2.5}$ (for BRT and RF respectively). The MAPE values are higher than those given above because of the occurrence of empirical values close to zero (these appear in the denominator of the formula given in Table 2). The goodness-of-fit measures used indicate that values of pollutant concentration are predicted more effectively by the RF method in the case of $PM_{2.5}$ and by the BRT method in the case of NO_x. For NO_2 the measures do not indicate an unambiguously better model. Deeper analysis was therefore carried out on the values of the errors occurring in modelling using the techniques described in sections 2.2 and 2.3, and the conditions in which they were observed.

3.2 Importance of predictive variables

The importances of variables (Table 4) were determined in order to identify the variables that exert the largest influence on pollutant concentrations. The different strategies and methods used for tree construction in the RF and BRT cases led to differences in the importances of particular variables in the respective models (Table 4).

Table 4. Importance of the predictive variables, in %.

	Boosted regression trees			Random forest		
Variable	NO_2	NO_x	$PM_{2.5}$	NO_2	NO_x	$PM_{2.5}$
Traffic flow	74	75	30	100	100	22
Wind speed	100	100	76	62	76	36
Wind direct.	79	82	79	54	71	33
Temperature	45	59	100	59	88	100
Air pressure	8	48	40	40	88	36
Rel. humidity	46	58	56	57	73	56
Day of week	84	87	48	53	78	26
Month	57	60	93	59	78	91

Generally speaking, the concentrations of nitrogen oxides (NO_2 and NO_x) in the air were most strongly influenced by traffic flow, wind speed and day of the week. Although day of the week is statistically significantly uncorrelated with traffic flow ($r = –0.16$), there is a link between them, in view of the weekly variability of traffic volumes. There are clear differences between the values of importance obtained using the BRT and RF methods. According to BRT the greatest impact on nitrogen oxide concentrations comes from wind speed, which is responsible for the evacuation of pollutants; this is in agreement with the results reported in [16]. According to the RF models, however, nitrogen oxide concentrations are most affected by traffic flow, the principal source of emissions of those pollutants; this agrees with the findings of Laña et al. [19].

There is a marked difference in the importances of the variables in the case of particulate matter

concentration. According to both models, the most important factor is air temperature, which has a direct influence on heating emissions. For the same reason, the next most important factor is month. In spite of the obvious causal relationship between month and air temperature in Wrocław, the non-parametric correlation coefficient gamma shows, with statistical significance, a lack of correlation (0.14). The next most important variables are wind factors, which are responsible for the evacuation of pollutants.

3.3 Residual analysis

RF and BRT modelling do not require the distributions of the input variables to satisfy any normality assumptions. The mean values of the residuals (differences between real and modelled values) in each case did not differ statistically significantly from zero (the *t*-statistic values ranged from –1.65 to –0.22, and the p-value for the *t*-test from 0.16 to 0.84). In none of the cases considered did the distribution of residuals correspond to a normal distribution. The errors for each of the six models exhibit right-sided asymmetry and are leptokurtic – that is, they have a longer tail on the right-hand side, and there is a greater concentration of values around the mean than in the case of a normal distribution (Table 5).

Table 5. Descriptive statistics for residuals.

		BRT	RF
NO_2	Mean	-0.271	-0.54
	Standard deviation	17.90	15.70
	Skewness	1.33	1.32
	Kurtosis	5.16	5.50
NO_x	Mean	-0.256	-1.07
	Standard deviation	81.92	73.88
	Skewness	3.92	3.48
	Kurtosis	38.00	29.56
$PM_{2.5}$	Mean	-0.402	-1.17
	Standard deviation	18.93	14.15
	Skewness	2.66	1.09
	Kurtosis	15.44	8.10

The non-normality of the distribution of the residuals results from the specific nature of the phenomenon under analysis. All of the dependent variables have right-sided asymmetry (outliers and extreme values above the median). This feature is least pronounced in the case of NO_2 (where the skewness is 1.3) and is significantly stronger in the case of NO_x (skewness 3.5). For the residuals in the modelling of particulate matter, the greatest difference between the distributions for the two models is observed (Table 5). The models provide good predictions for values that are close to average, but underestimate the extreme values (Fig. 2). This phenomenon occurs to a lesser degree in RF than in BRT models. Figures 2 and 3 show underestimated values in the BRT models (points beneath the line $y=x$) for large values of concentration. In the BRT models, however, there is always visible some kind of upper bound on the modelled values: at 80 µg/m^3 for NO_2 and $PM_{2.5}$, and at 300 µg/m^3 for NO_x. This results from the way in which successive tree boosts are constructed in BRT, where in spite of the large weights of high values of residuals in the previous division, they are divided at the next step into only two subgroups (binary trees). In effect, the high values, which are not numerous, fail to be represented in the model.

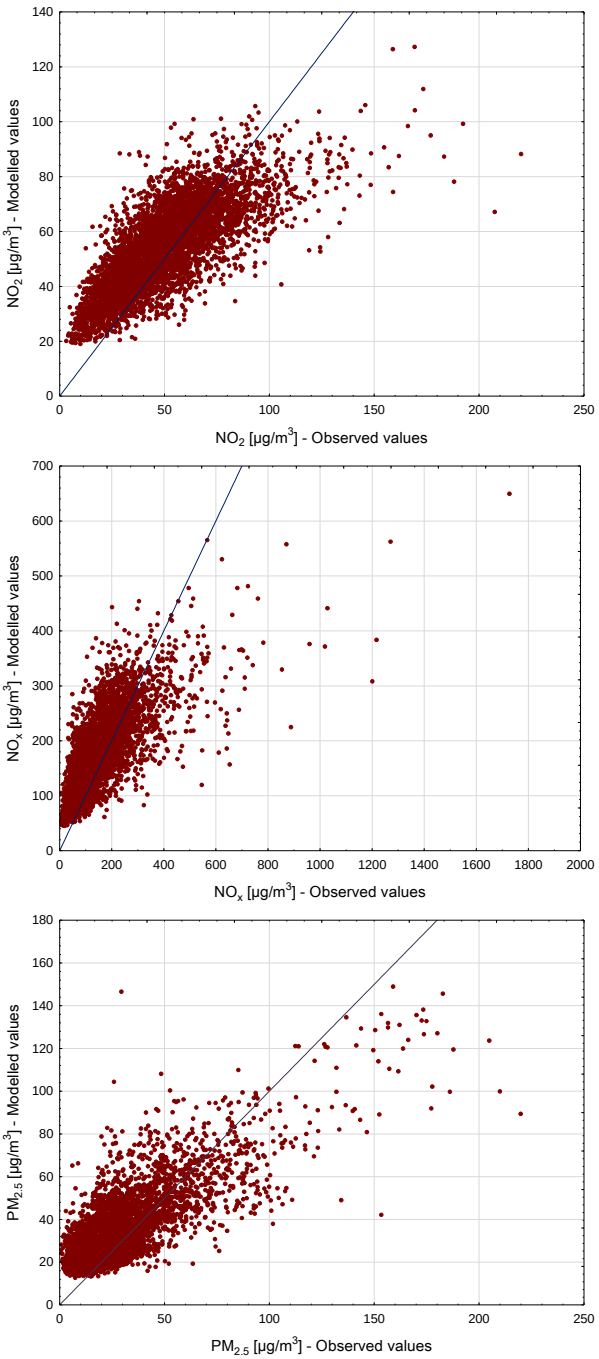

Fig. 2. Real values and values modelled by the RF method, with the line $y=x$ shown.

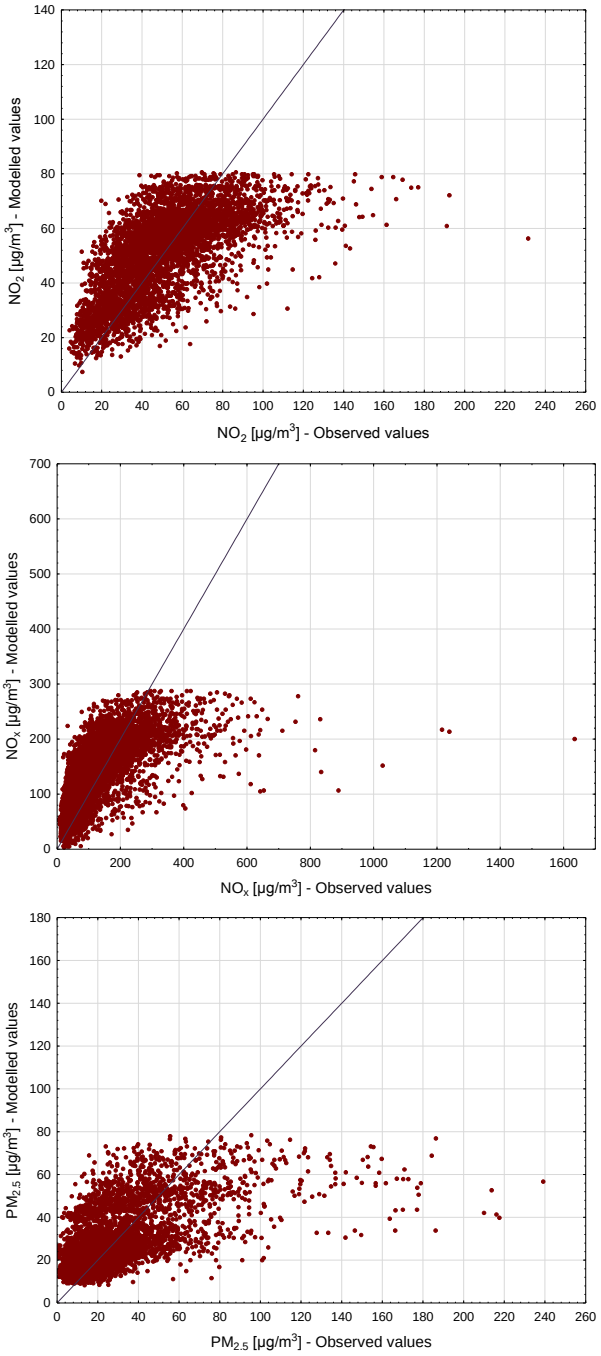

Fig. 3. Real values and values modelled by the BRT method, with the line $y=x$ shown.

3.4 Verification of models

Each of the models underwent verification using data from the year 2017, including hourly values of eight variables: traffic volume, day of the week, month, air temperature, wind speed, wind direction, relative humidity and air pressure. The values of the main statistics for the numerical variables are given in Table 6. The amplitudes of variation in pollutant concentrations were smaller in 2017 than in 2015 and 2016.

Values of concentrations of NO_2, NO_x and $PM_{2.5}$ were predicted based on the RF and BRT models constructed as described in sections 2.2, 2.3 and 3.3. Time series of measured concentrations of these three pollutants, together with the values predicted by the RF and BRT models, are shown in Fig. 4.

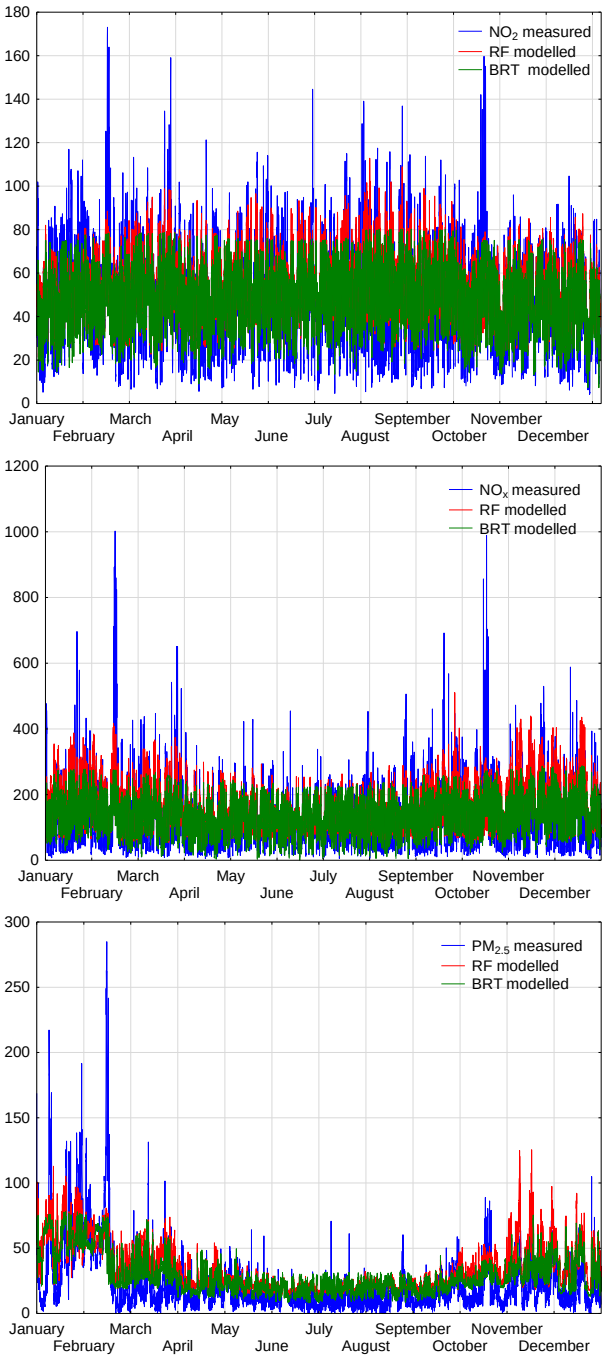

Fig. 4. Real values and values predicted by the RF and BRT models in 2017.

Table 6. Descriptive statistics for continuous predictors and pollution concentration [µg/m^3] in the year 2017.

Variable	$\bar{x}$	min	max	s
NO_2 concentration	48.41	4.30	173	21.95
NO_x concentration	132.7	5.6	1002.2	94.27
$PM_{2.5}$ concentration	23.00	0.10	285.0	25.25
Traffic flow [veh/h]	2758	0	5425	1776
Temperature [°C]	10.24	-15.20	35.0	8.30
Wind speed [m/s]	3.2	0.0	19.0	1.99
Relative humidity [%]	75.7	23.00	100.00	16.16

Air pressure [hPa]	1002	970	1026	8.67

$\bar{x}$ – mean value; s – standard deviation

The smallest relative errors between real values and the values predicted by the models (26% and 27% for BRT and RF respectively) were obtained for NO_2. Neither of the models predicted the peaks. Verification of the models with respect to the prediction of values of NO_x concentration gave worse results. MADE values and relative errors (MADE/mean value) are given in Table 7.

Table 7. Verification errors.

MADE MADE/mean [%]	BRT	RF
NO_2	12.7 24.6%	13.0 25.3%
NO_x	49.7 33.7%	57.2 38.8%
$PM_{2.5}$	13.5 46.6%	16.3 56.6%

The smallest error was obtained in the verification of the models for nitrogen dioxide concentrations. The error value is comparable to that obtained for the test data from 2015–2016. The greater variation in NO_x values led to a reduction in the accuracy of prediction of concentrations. Nonetheless, the mean absolute deviation error took lower values than in the case of the test set. The highest relative error (greater than in the model testing process) was recorded in the verification of the models for $PM_{2.5.}$ While the models predicted low values in summer, and higher values with greater variation in winter, the relative error for this pollutant was the highest (46.6% for BRT and 56.6% for RF). This is due to the overestimation of summer values and underestimation of winter values. The RF model also predicted a peak in $PM_{2.5}$ values – at the end of November and start of December – which did not occur in reality (the real values were almost two times smaller than the modelled values).

4 Conclusions

Two data mining methods – a random forest and boosted regression trees – were used to model the values of roadside air pollutant concentrations based on meteorological and traffic conditions, using the example of data obtained in the city of Wrocław in 2015–2016. In the models, account was taken of eight continuous or categorical explanatory variables. A comparison was made of the quality of the fit of the models to the empirical data. As in other cases reported in the literature, relatively low values of goodness-of-fit measures were obtained (R^2 was approximately 0.5). The effectiveness of the modelling methods was compared for each of the considered pollutants: NO_2, NO_x and $PM_{2.5}$. However, the commonly used goodness-of-fit coefficients did not provide an unambiguous answer. For the modelling of NO_x they indicated that the BRT model achieved greater predictive accuracy, while for $PM_{2.5}$ the random forest model proved superior. Residual analysis was performed for each of the models, and led to the conclusion that the effectiveness of a modelling method depends on the assumed priorities. If the prediction of typical values is most important, then better results were obtained by the BRT method. However, for the prediction of higher values (although not the peaks themselves) the RF method gives better results. The RF models, with imposition of the additional condition of the weighting of instances according to their value, predict higher concentration values, although unfortunately they also overestimate the low values. For the BRT model there was observed to be an upper bound on the predicted values of all modelled pollutants. This is a result of the way in which the method operates, in combination with the very large dataset and the small number of atypical values.

The decision trees-based approach described in this paper, fits to the concept of decision support systems. Such systems are commonly used in a variety of urban management domains, including energy planning [33], climate control [34] and water management [35]. The development of decision support systems also leads to the advancement of technologies, which are helpful in processing increasing large amounts of data [36].

References

1. World Health Organization, *Health Risks of Air Pollution in Europe (HRAPIE)* (WHO, Copenhagen, 2013)
2. *Information on air quality in Wrocław* [Informacja o jakości powietrza na terenie miasta Wrocławia] (Provincial Environment Protection Inspectorate [WIOS], Wrocław, 2016)
3. *Low-Emission Management Plan for Wrocław Municipality* [Plan Gospodarki Niskoemisyjnej dla Gminy Wrocław], p. 184 (2016)
4. J. Zwoździak DEng (ed.), *Limitations of low emissions from household coal heating in Wrocław in 2016–2020* [Ograniczenia niskiej emisji z indywidualnego ogrzewania węglowego na terenie Wrocławia w latach 2016–2020] (2017), http://bip.um.wroc.pl/artykul/643/25539/ograniczenia-niskiej-emisji-z-indywidualnego-ogrzewania-weglowego-na-terenie-wroclawia-w-latach-2016-2020
5. T.T. Hien, H.N. Linh, L.M.T. Luong, P.K. Thai, Sci. Total Environ. **Vol. 557–558**, 322-330 (2016)
6. M. Adam, T. Schikowski, A.E. Carsin, Y. Cai, B. Jacquemin, M. Sanchez, A. Vierkötter, A. Marcon, D. Keidel, D. Sugiri, Z.A. Kanani, R. Nadif, V. Siroux, R. Hardy, D. Kuh, T. Rochat, P.-O. Bridevaux, M. Eeftens, M.-Y. Tsai, S. Villani, H.Ch. Phuleria, M. Birk, J. Cyrys, M. Cirach, A. Nazelle, M.J. Nieuwenhuijsen, B. Forsberg, K. Hoogh, K. Declerq, R. Bono, P. Piccioni, U. Quass, J. Heinrich, D. Jarvis, I. Pin, R. Beelen, G. Hoek, B. Brunekreef, Ch. Schindler, J. Sunyer, U. Krämer, F. Kauffmann, A.L. Hansell, N. Künzli, N. Probst-Hensch, Eur. Resp. J. **45**, 38-50 (2015)

7. G. Hoek, R.M. Krishnan, R. Beelen, A. Peters, B. Ostro, B. Brunekreef, J.D. Kaufman, Environ Health. May **28**, 12(1), 43 (2013)
8. M.-C. Flores-Pajot, M. Ofner, M.T. Do, E. Lavigne, P.J. Villeneuve, Environ. Res. **151**, 763-776 (2016)
9. L. Pei-Chen, L. Li-Ling, S.M.D. Yu, C. Yu-An, L. Chih-Ching, L. Chung-Yi, Y. Hwa-Lung, B. Ritz, Environ. Intern. **96**, 75–81 (2016)
10. G. Tang, P. Zhao, Y. Wang, W. Gao, M. Cheng, Y. Xin, X. Li, Y. Wang, Atmos. Environ. **150**, 1238-243 (2017)
11. B. Barratt, R. Atkinson, H. Ross Anderson, S. Beevers, F. Kelly, L. Mudway, P. Wilkinson, Atmos. Environ. **41 (8)**, 1784-1791 (2007)
12. J. Kazak, M. Chalfen, J. Kamińska, S. Szewrański, M. Świąder, In: I. Ivan, J. Horak , T. Inspektor (ed.) *Dynamics in GIScience* (GIS Ostrava 2017, Lecture Notes in Geoinformation and Cartography, Springer, Cham, 195-207, 2018)
13. L. Malytska, V. Balabukh, Meteorol. Hydrol. Water Manage. **6(1)**, 59-65 (2018)
14. I. González-Aparicio, J. Hidalgo, A. Baklanov, A. Padró, O. Santa-Coloma, Environ. Sci. Pollut. Res. **20(7)**, 4469-4483 (2013)
15. K. Zhang, S. Batterman, Sci. Total Environ. **450-451**, 307-316 (2013)
16. A. Sayegh, J.A. Tate, K. Ropkins, Atmos. Environ. **127**, 163-175 (2016)
17. R.H. Keeler, *PhD thesis* (2014)
18. P. Mlakar, M. Boinar, Intelligent Information Systems, 345-349 (1997)
19. I. Laña, J. Del Ser, A. Pedró, M. Vélez, C. Casanova-Mateo, Atmos. Environ. **145**, 424-438 (2016)
20. P. Bertaccini, V. Dukic, R. Ignaccolo, Adv. Meteo. **2012**, 1-16 (2012)
21. M. Aldrin, I.H. Haff, Atmos. Environ. **39**, 11, 2145-2155 (2005)
22. L. Breiman, Machine Learning **45 1**, p. 5-32 (2001)
23. M. Catalano, F. Galatioto, M. Bell, A. Namdeo, A.S. Bergantino, Environ. Science&Policy **60**, 69-83 (2016)
24. Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, A. Baklanov, Atmos. Environ. **60**, 632-655 (2012a)
25. Y. Zhang, M. Bocquet, V. Mallet, C. Seigneur, A. Baklanov, Atmos. Environ. **60**, 656-676 (2012b)
26. J. H Friedman., Ann. Stat. **29(5)**, 1189-1232 (2001)
27. J. H. Friedman, *Stochastic gradient boosting* (Stanford University, 1999b)
28. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and regression trees* (CA: Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, 1984)
29. K.J. Archer, R.V. Kimes, Comp. Stat. Data Analysis **52(4)**, 2249-2260, (2008)
30. J.W. Boylan, A.G. Russell, Atmos. Environ. **40**, 4946-4959 (2006)
31. A. Larkin, J.A., Geddes, R.V. Martin, Q. Xiao, Y. Liu, J.D. Marshall, M. Brauer, P. Hystad., Environ. Sci. Technol., **51**, 6957-6964 (2017)
32. J. Kamińska, J. Environ. Manage. **217C**, 164-174 (2018)
33. J. Kazak, J. van Hoof, S. Szewranski, Renew and Sust. Energy Rev. **76**, 425-433 (2017)
34. J.K. Kazak, Sustainability **10**, 1083 (2018)
35. S. Szewrański, J. Chruściński, J. Kazak, M. Świąder, K. Tokarczyk-Dorociak, R. Żmuda, Water **10(4)**, 386 (2018)
36. S. Szewrański, J. Kazak, M. Sylla, M. Świąder, In: I. Ivan, A. Singleton, J. Horák, T. Inspektor (ed,) *The Rise of Big Spatial Data. Lecture Notes in Geoinformation and Cartography* (Springer, Cham, 2017)