ESTIMATION OF ATMOSPHERIC BOUNDARY LAYER VALUES IN THE CONTEXT OF THE DAILY PREDICTION OF PM10 AIR POLLUTION

PIOTR A. KOWALSKI[1,2], MACIEJ KUSY[3], MARCIN SZWAGRZYK[4] & JAN IZYDORCZYK[4]
[1]Faculty of Physics and Applied Computer Science,
AGH University of Science and Technology, Cracow, Poland
[2]Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[3]Faculty of Electrical and Computer Engineering,
Rzeszow University of Technology, Rzeszow, Poland
[4]Airly Inc. USA

ABSTRACT
Air pollution is one of the most dynamically developing problems of the contemporary world. Due to constantly present threat of air pollution, it is essential for the society to be aware of this issue and to be able to trace the individual factors influencing the existence of smog, as well as to predict the state of air quality in the following hours and days. This paper aims to determine the feasibility of cascading prediction of atmospheric boundary layer (ABL) values for several consecutive days, and then use this information to synthesize a prediction procedure for harmful smog particulate matter (PM10) for several days as well. Various prediction methods are used in the current study, among which the linear regression algorithm proves to be the most effective. Herein, the simulations concerning the investigated prediction algorithms are based on real data provided by the Airly company network of pollution measurement stations as well as ABL from the Copernicus Climate Data Store. Evaluation of the obtained results is carried out using such measures as mean squared error, mean absolute error, Pearson correlation coefficient R, and index of agreement. As a result of the simulation, ABL and then PM10 predictors are synthesized for three consecutive days. The latter is characterized by an average daily mean absolute error in the range of 8-10 μg/m3, and index of agreement 0.88-0.89 depending on the day of the prediction and the variants of the prediction algorithm selected.
Keywords: air pollution prediction (forecasting), atmospheric boundary layer, big data, data science, machine learning, particulate matters, regression task, data science.

1 INTRODUCTION

According to the WHO report [1] on the threat of air pollution, it can be seen that it is a very serious global problem. It concerns the vast majority of the population on earth, and the funds that are spent on this purpose are allocated, on one hand, to fighting and counteracting air pollution. On the other hand, they are spent on combating the consequences of this issue. It is worth emphasizing here that these effects are far-reaching and are felt most strongly in problems associated with health care in the broad sense of the term. A few years ago, WHO stated that over 50 cities in the EU were among the most polluted, 33 of these were located in Poland. An essential element in facing this problem is the environmental awareness of city and village inhabitants throughout the world. This awareness can result from various sources such as widespread environmental education in schools, generally available literature, own experience, but also the conviction whether there is a threat of air pollution in a given area. This last element is important since such knowledge can prevent exposure to polluted air. This can be achieved by continuous monitoring of the state of air quality. An example of such a solution is the Internet portal of Airly company, which provides a free map (https://airly.org/map/en/), where the current state of air pollution can be monitored (Fig. 1).

ISSN: 2398-2640 (paper format), ISSN: 2398-2659 (online), http://www.witpress.com/journals
DOI: 10.2495/EI-V5-N3-271-283

Figure 1: A map with a network of Airly's sensors.

This map makes it possible to see both global air quality indices and individual fractions of pollution that contribute to the overall index.

Depending on the measurement point of the Airly station, everyone can read online components of air pollution such as PM10, PM2.5, PM1, NO_2, O_3 and additionally get information about basic weather parameters, namely: air temperature, air humidity, pressure and wind speed. Air pollution sensors measure particulate matter using a laser method. For gaseous pollutants, a single device is able to measure two chemical compounds using an electrochemical method. Each sensor has a special light-emitting diode that changes its colour to show the current air quality according to the CAQI scale. Thanks to an individualised adjustment procedure for each measuring device, the compliance of the sensor system measurements is very high. A very strong side of this application is the possibility of daily prediction of the air quality level, with hourly resolution. The forecasting system is based on the most recent innovations in the field of neural networks and deep learning, thanks to which its verifiability often exceeds 95%. The applied idea of multiple neural networks allows the generating of air quality forecasts for practically the majority of measuring stations in the world. Therefore, it is possible to plan all activities during the nearest day, choosing an appropriate time when we will be exposed to the least amount of air pollution.

Forecasting the state of air quality is carried out using classical statistical methods, more advanced solutions of exploratory data analysis, and intelligent approaches. However, regardless of the methods used, such a forecast is a very complicated computational task [2, 3, 4], which requires several procedures starting from data preparation and ending with the development of a prediction algorithm. This issue has been the topic of many scientific research works, which used sophisticated solutions, for example, fuzzy logic based expert systems [5], neural network procedures [6], neural networks with ARMA models [7], deep neural networks [8,9], autoencoders [10] and convolutional neural networks [11,12]. Due to the very broad spectrum of the air pollution term, most modelling and prediction systems for pollution status are focused on a single factor such as PM10, PM2.5, NO_2, O_3, etc. In most cases, the models listed above include air pollution that is realized from one day to the subsequent day. A certain exception is an air quality prediction system implemented by Airly [13], which provides the possibility of independent observation of PM10, PM2.5 pollutants and CAQI

index for the next 24 h with hourly resolution. The quality and statistical properties of the proposed models largely depend on the input data adopted as well as the spatial area of application. An interesting solution may include [14] in which Computer Fluid Dynamics was applied for images of the CO_2 concentration analysis. Another approach is the investigation of Geographic Information System software to obtain detailed coordinates and dimensions of the buildings in the explored area [15]. A further type of air quality prediction is the receptor modelling method. This procedure uses the chemical and physical properties of gases and particles measured at the source of pollution to quantify its concentration and contribution [16]. A broader overview of air quality modelling and prediction systems and methods can be found in [17, 18].

In the context of both air quality exploration and prediction, the atmospheric boundary layer (ABL) plays a crucial role. The ABL is the fraction of the atmosphere that is directly related to the mass, energy and momentum flow from the earth's surface. This part of the atmosphere is responsible for the transport of air pollutants. Thus, variability is very much important in the transport, dispersion and deposition of air pollutants. The structure of the ABL is determined by the complex interactions between the stability of the atmosphere and the mechanical processes (such as wind uplift from synoptic or terrain-formed flows). These processes may operate at different altitudes and time scales, and their dominance may vary.

The main motivation of the present study is to address the issue of prediction of air pollution in the form of particulate matter PM10, in the perspective of the next few days For such a formulated task, additionally, the ABL parameter is used. It is connected with the so-called convective boundary layer heights. This phenomenon is particularly important for areas located in various basins or mountain areas with no unrestricted airflow. The main task carried out in the present work is the synthesis of cascade predictions of ABL and PM10, which generated daily averaged values for the following 3 days. It should be emphasized here that ABL values are elements of the feature vector based on which the PM10 forecast is generated.

The main novelty of the presented approach is both PM10 pollution prediction and the forecast of one of the input factors i.e. ABL. Moreover, the ABL prediction itself is hardly available in common and open forecasting systems. It is worth emphasising that both elements of prediction are based on a consistent machine learning model.

The paper is organized as follows. In Section 2, we briefly inform the reader about the data sets applied for the considered investigation. In Section 4, the investigated prediction models, as well as evaluation measures, are introduced. Next, in Section 4, all results of numerical verification of both ABL and PM10 predictions are presented and discussed. Finally, Section 5 contains concluding remarks and plans for future research.

2 DATA SETS

Heterogeneous data sources related to both air pollution and selected meteorological parameters are used in this study. The main provider of data related to air pollution and atmospheric factors such as temperature, humidity, wind force and atmospheric pressure is Airly company, which has almost 5,000 measurement points worldwide. The resolution of Airly air pollution data is dense, but due to the necessity of the physical existence of the device, this parameter cannot be clearly defined. In the case of the city of Krakow, for which the analysis is carried out, about 100 measuring stations are located within a radius of 10 km. Due to the dense network and very precise adjustment of measuring devices, individual devices achieve very high compliance with official government measuring devices (GIOŚ). As mentioned in the

introduction, the applied prediction algorithms are enriched with ABL information, which comes from the Copernicus Climate Data Store (CCDS). The ERA5 HRES atmospheric data, provided by CCDS, have a spatial resolution of 31 km and 0.28125 degrees.

All simulation data are downloaded from six Airly measuring stations located in the city of Krakow in Poland. The choice of this city as the case study is not accidental. The first reason is its location, which resembles a basin, so there is a serious problem with fast air exchange; the second argument is significant air pollution. Because the main goal of the research is the daily prediction for the next few days, all data (PM10 concentration, temperature, humidity and wind speed) take the form of daily averages of individual physical quantities. One interesting observation obtained from real data is the autocorrelation of ABL and PM10, as well as their mutual position. Based on the analysis of the above values, which, depending on time maintain quite high values (i.e. above 0.6), it can be concluded that ABL and PM10 are related to each other, and it is worth trying to use this relationship in the PM10 forecast. Moreover, based on such a high value of autocorrelation, the correct use of daily averaged data can be assumed. In addition, it is worth mentioning that similar conclusions are obtained from other sources [19]. The investigated dataset is divided into two parts. The first one – a training subset, covers the period from the beginning of November 2020 to the end of October 2021; the second one – a testing subset, includes all days of November 2021 to March 2022 interval.

3 PREDICTION MODELS

This part of the article briefly characterizes both the predictive models and the validation measures used in this research. The models' descriptions are divided into two subsections. The first subsection is concerned with prediction methods based solely on individual/intrinsic characteristics of the predicted quantity. The second one describes the prediction algorithms based on a vector of input variables containing different data features.

3.1 Self prediction models

In this section, we present the methods used in simulations that allow the prediction of the variable under study based on the knowledge of only the historical value of the quantity under consideration, i.e., no additional information (variables of a different type than the one predicted) is used [20]. They are shortly discussed below:

- The Mean Method (MM): it generates an output mean value based on training data.
- The Simple Exponentially Smoothing Method (SESM): it is characterized by its ability to model data with a clear trend that is unobserved.
- Holt-Winters Method (HTM) [21, 22]: it is more advanced than the previous algorithms, as it has several equations in its structure, including both the forecast itself and also the seasonality and three smoothing equations. The HTM has two variations that differ in the nature of the seasonal equation. The first, the additive method, is preferred when seasonal fluctuations are approximately constant in the time series; the second – the multiplicative method, is preferred when seasonal fluctuations change proportionally to the level of the series.
- Holt's Linear Trend Method (HLTM): it is based more on linear trend-related transformations.
- The Autoregressive integrated moving average (ARIMA) procedure [23,24]: it is composed of three components: the autoregressive factor, the moving average factor, and the degree of integration. The classical Autoregressive Moving Average (ARMA) models apply to stationary series, whereas the introduction of the integrated factor to the procedure

results in the possibility of applying the procedure to nonstationary series, i.e. series which have a dynamically changing mean value in relation to the standard deviation. The last model used in this research was Auto AIRMA, which does not differ from the previous model, i.e., ARIMA, but has an individualized algorithm of optimization of internal parameters. More on the above-characterized methods, in particular their mathematical models, can be found in [25].

3.2 Prediction based on heterogeneous data feature

In this part of the article, the regression algorithms that are used for time series prediction from the ML domain are characterized.

The first model considered in this group is based on the linear regression (LR) method [26]. This method, used in statistical modelling, assumes that it is possible to linearly transform the elements of the input space into the output data. In the case studied, the elements of the input space are feature vectors comprising various types of meteorological data, while the output data are prediction values of ABL or PM10. Due to the different forms of the features in the vector, this issue is described in more detail in the further part of this paper.

The random forest regression (RFR) procedure [27,28] is an example of a supervised machine learning algorithm, which applies the idea of ensemble learning to the problem of regression. The method assumes the possibility of using the results from many simultaneously created models of the regression trees. As the outcome of the synergy of individual trees, a procedure with a much higher prediction ability than each of the models considered separately can be obtained. The RFR method consists in building a large number of independent decision trees. The trees do not interact with each other. Due to the feature bagging procedure invoked in RFR training, important features in the context of prediction are selected in the majority of grown trees. Unfortunately, due to the complexity of single trees, this procedure is often used as a 'black box' model. Algorithms from this group are characterized by very good performance, scalability and accuracy. These methods generate reasonable predictions for a wide range of data, requiring a relatively small configuration. They are very popular in both research and commercial solutions.

Decision tree regression (DTR) models [29] are used in modelling and prediction approaches in statistics and machine learning very frequently. Often referred to as DTR induction, these learning processes assume the construction of a particular graph structure, which is called a tree. Going down from the root towards the leaves, the input data is divided into smaller subsets, until it is finally associated with the output data. The final result is a tree with decision nodes and leaf nodes. Decision trees in which the target variable takes continuous values are called regression trees. This method is also used to visually and openly represent decisions and the entire decision-making process. In contrast to RFR, this model is extremely readable but has a greater tendency to memorize patterns, i.e., overfitting.

3.3 Validation measures

In this study, we consider, the task of average daily prediction of ABL and PM10 contamination ($\hat{y}^{\tau}$) for the next 3 days (i.e. for $\tau = 1, 2, 3$). The base measure allowing the model to be evaluated is the mean squared error (MSE). This is expressed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i^{\tau} - y_i^{\tau})^2, \tag{1}$$

where $\hat{y}_i^{\tau}$ constitutes the i-th prediction of investigated forecast (for the τ-th day), and the output vector consists of n observed real values of the predicted variable.

In mathematics and especially in statistics, mean absolute error (MAE) is a measure of the errors between the quantities of observations expressing the same phenomenon. Thus, both values $\hat{y}^{\tau}$, y^{τ} cover the comparison of forecast data with the observed actual data, respectively. MAE is computed as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i^{\tau} - y_i^{\tau}\right|. \tag{2}$$

A completely different type of evaluation measure is the Karl Pearson correlation coefficient R. This quantity indicates the level of the linear relationship between two variables and is defined as follows:

$$R = \frac{Cov(y^{\tau},\ \hat{y}^{\tau})}{std(y^{\tau})\ std(\hat{y}^{\tau})}, \tag{3}$$

where $std(y^{\tau})$ and $std(\hat{y}^{\tau})$ denote the standard deviations of observed value and the investigated forecast, respectively, whereas $Cov(y^{\tau}, \hat{y}^{\tau})$ is the covariance coefficient. In the presented research, the square of this coefficient (R^2) is used as a validation measure.

The last evaluation measure applied in this paper is the index of agreement:

$$IA = 1 - \frac{\sum_{i=1}^{n}\left(\hat{y}_i^{\tau} - y_i^{\tau}\right)^2}{\sum_{i=1}^{n}\left(\left|\hat{y}_i^{\tau} - \overline{y}^{\tau}\right| + \left|y_i^{\tau} - \overline{y}^{\tau}\right|\right)^2}, \tag{4}$$

where value $\overline{y}^{\tau}$ denotes mean value of real data y^{τ}. The agreement value of 1 indicates a perfect match, and 0 indicates no agreement.

4 EMPIRICAL STUDY

This section presents the simulation results obtained by the algorithms described in the previous section. They concern the prediction of the ABL parameter, which is to be finally used in the task of predicting daily PM10 levels for three consecutive days.

Firstly, the prediction procedures of the ABL parameter, which are based solely on the values of a given parameter, without the use of additional data, are examined. Several interesting results are obtained; however, due to the limited volume of the article, only selective ones are shown – the most promising ones in this group. The results obtained using the MM, SESM and HLTM algorithms are characterized by a very averaging character, which comes down to a straight line. Completely different outcomes are obtained with HTM, AIRMA and Auto AIRMA. They are characterized by a positive R^2 coefficient. The best of these methods, i.e., HTM achieves the highest result equal to $R^2 = 0.51$ for the test set covering one month. The simulation results of this algorithm can be seen in Fig. 2, from which it can be seen that the prediction of the ABL for the next day takes into account the nature of the run of the test data. However, the main drawback of HTM is its lack of scalability over time. This was found when simulating this method for almost the entire 2021 year (see Fig. 3). In this case, the correlation coefficient dropped from a value of 0.51 to –0.15, which discredits this method in the context of ABL prediction.

Further research includes the task of predicting the ABL parameters using ML algorithms that rely on both past ABL data but also on other meteorological parameters such as air temperature, atmospheric pressure, humidity and wind speed.

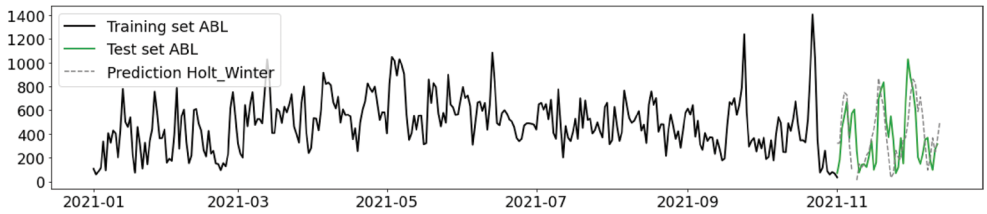

Figure 2: Results of ABL prediction for test set based on Holt-Winters method.

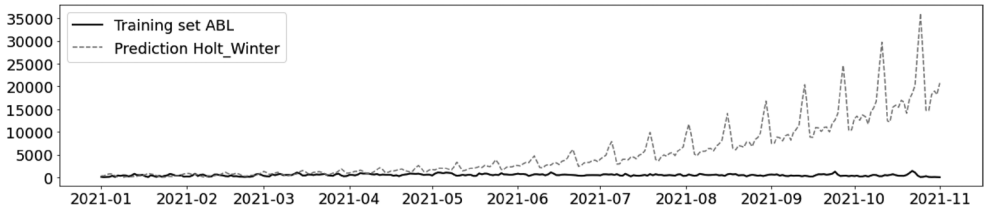

Figure 3: Results of ABL prediction for training subset based on Holt-Winters method.

Thus, the prediction model for the subsequent =1,2,3 days can be defined as a set of the following functions:

$$ABL(t+\tau) = f_{\tau}\Big(\big(ABL(t+(\tau-1),\ldots,t-1), temp(t+(\tau+1),\ldots,t-2),\\ hum(t+(\tau+1),\ldots,t-2), wind(t+(\tau+1),\ldots,t-2)\big)\Big), \tag{5}$$

where f_{τ} is the predictive model and t is the time (i.e. simulation day).

Table 1 shows the prediction results of the ABL parameter for the next three days using three methods from the ML group, that is, LR, RFR and DTR. Each simulation for every day and the prediction method are described using eight evaluation parameters. These parameters are represented by MSE, MAE, R^2 and IA. Moreover, they are divided into two groups (train and test) due to conducted computations on training and testing subsets. For each prediction day, the best results are denoted in boldface. As shown, the RFR method proves to be dominant, providing the best outcomes practically for each day for training subset while the LR model is the best choice for testing data. One needs to note that the IA for both methods for the test data exceeds 0.85.

Figure 4 shows the result of the best LR algorithm in the task of predicting the mean value of the ABL parameter for day 3 for both training and testing data. It can be seen from the figure that the prediction procedure is not averaging and takes into account variations in the amplitude of the ABL waveform.

The final stage of this research work is the synthesis of a prediction of air pollution in the form of PM10 dust for three consecutive days. In this case, similar data sets as for the ABL prediction task are considered, with the difference that the ABL prediction result, the current and historical PM10 values are included in the feature vector as additional elements. The results obtained for the PM10 prediction problem are presented in Tables 2, 3 and 4. They indicate the forecasts based on ABL predicted values for RFR, LR and DTR methods, respectively. The layout of the tables is identical to that described in Table 1. Table 2 contains the results of PM10 prediction based on the ABL predicted value obtained by the RFR procedure. In the case of training data for all days, the RFR method proves to be the best method

Table 1: Results of ABL prediction.

Prediction day	1			2			3		
Algorithm	**LR**	**RFR**	**DTR**	**LR**	**RFR**	**DTR**	**LR**	**RFR**	**DTR**
MSE (train)	19.69	**3.65**	1.07	19.15	**3.66**	10.74	19.90	**3.58**	10.61
R^2 (train)	0.61	**0.93**	0.79	0.62	**0.93**	0.79	0.63	**0.93**	0.79
MAE (train)	104.45	**43.25**	73.47	102.33	**44.12**	72.87	102.21	**43.76**	72.00
IA (train)	0.87	**0.98**	0.94	0.87	**0.97**	0.94	0.87	**0.97**	0.92
MSE (test)	**44.59**	50.70	110.74	**46.14**	51.78	111.35	**47.43**	52.99	107.12
R^2 (test)	**0.65**	0.61	0.14	**0.64**	0.60	0.14	**0.63**	0.59	0.17
MAE (test)	**156.71**	163.62	228.06	**159.79**	164.76	226.66	**162.00**	167.32	218.93
IA (test)	**0.89**	0.86	0.77	**0.89**	0.86	0.78	**0.88**	0.85	0.78

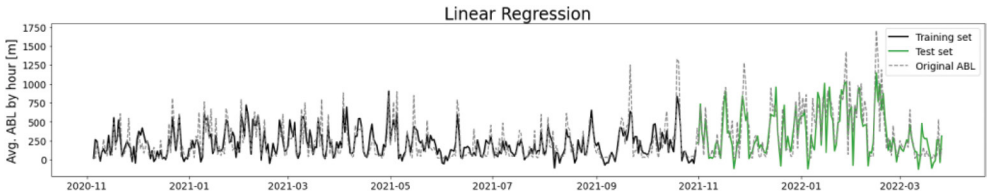

Figure 4: ABL predictions results based on LR procedure (3rd day).

for PM10 prediction. It should be emphasised here that its advantage expressed by all evaluation measures is very significant. If we analyse the outcomes on the test data, better results for the LR algorithm than for RFR methods are observed.

In the case of the analysis of PM10 prediction based on ABL predicted value with LR procedure shown in Table 3, the situation for the training as well as for testing data is the same as described above. The RFR method for PM10 prediction proved to be the most effective for all forecast days for training data. In the case of the test data, the LR method predominates.

The last of the tests pertain to PM10 prediction based on ABL predicted value with DTR procedure. The results of this numerical simulation are presented in Table 4. Again, for the training data, the RFR method wins unanimously. We also observe an overwhelming advantage of LR method over the results achieved by the RFR and DTR approaches. In the case of algorithm analysis for the test data for the first day, the best predictor is the LR algorithm, which in the case of both the MSE and the correlation measure R^2 obtains much more favourable values.

By comparing the PM10 prediction methods for individual days, it can be indicated that in each day, the LR-based prediction procedure is the most favourable for both ABL and PM10. The LR method should be indicated as the preferred prediction solution. From the above studies, it can be concluded that the RFR algorithm is prone to overfitting, as it performs practically correctly for the training data. This observation is confirmed by IA = 0.99 and very low MAE < 3.1 values. On the other hand, the DTR method has the opposite tendency. High values for all errors, often negative correlation values and frequently occurring IA < 0.7 imply very poor predictor quality.

Table 2: PM10 prediction based on ABL predicted value with RFR procedure.

Prediction day	1			2			3		
Algorithm	**LR**	**RFR**	**DTR**	**LR**	**RFR**	**DTR**	**LR**	**RFR**	**DTR**
MSE (train)	102.78	**18.21**	62.24	99.25	**18.19**	60.55	96.66	**18.08**	59.27
R^2 (train)	0.82	**0.97**	0.89	0.83	**0.97**	0.89	0.83	**0.97**	0.90
MAE (train)	7.67	**3.11**	5.71	7.53	**3.09**	5.63	7.46	**3.04**	5.52
IA (train)	0.95	**0.99**	0.97	0.95	**0.99**	0.97	0.95	**0.99**	0.97
MSE (test)	**165.26**	245.53	353.61	**156.97**	250.92	346.66	**171.99**	258.55	501.61
R^2 (test)	**0.58**	0.38	0.11	**0.61**	0.38	0.14	**0.57**	0.36	0.25
MAE (test)	**9.92**	12.83	14.08	**9.74**	12.94	13.86	**10.16**	13.26	15.91
IA (test)	**0.89**	0.79	0.73	**0.89**	0.79	0.74	**0.88**	0.78	0.68

Table 3: PM10 prediction based on ABL predicted value with LR procedure.

Prediction day	1			2			3		
Algorithm	**LR**	**RFR**	**DTR**	**LR**	**RFR**	**DTR**	**LR**	**RFR**	**DTR**
MSE (train)	103.96	**18.40**	59.36	100.48	**18.03**	58.29	97.60	**18.36**	56.46
R^2 (train)	0.82	**0.97**	0.90	0.82	**0.97**	0.90	0.83	**0.97**	0.90
MAE (train)	7.71	**3.09**	5.66	7.56	**3.05**	5.59	7.52	**3.04**	5.48
IA (train)	0.95	**0.99**	0.97	0.95	**0.99**	0.97	0.95	**0.99**	0.97
MSE (test)	**160.92**	246.22	359.12	**156.12**	257.03	370.53	**173.58**	261.52	504.97
R^2 (test)	**0.59**	0.38	0.09	**0.61**	0.36	0.08	**0.57**	0.35	0.26
MAE (test)	**9.74**	12.87	14.35	**9.70**	13.15	14.77	**10.33**	13.46	16.00
IA (test)	**0.89**	0.79	0.72	**0.89**	0.78	0.70	**0.89**	0.78	0.66

Table 4: PM10 prediction based on ABL predicted value with DTR procedure.

Prediction day	1			2			3		
Algorithm	**LR**	**RFR**	**DTR**	**LR**	**RFR**	**DTR**	**LR**	**RFR**	**DTR**
MSE (train)	103.89	**18.57**	59.63	100.27	**18.32**	58.21	97.58	**18.59**	56.83
R^2 (train)	0.82	**0.97**	0.90	0.82	**0.97**	0.90	0.83	**0.97**	0.90
MAE (train)	7.70	**3.12**	5.67	7.55	**3.09**	5.64	7.51	**3.07**	5.50
IA (train)	0.94	**0.99**	0.97	0.94	**0.99**	0.97	0.95	**0.99**	0.97
MSE (test)	**164.24**	249.00	356.51	**157.31**	260.95	352.26	**173.44**	261.46	503.31
R^2 (test)	**0.58**	0.37	0.10	**0.61**	0.35	0.13	**0.57**	0.35	0.25
MAE (test)	**9.83**	12.93	14.21	**9.79**	13.27	14.07	**10.29**	13.44	15.91
IA (test)	**0.89**	0.79	0.72	**0.89**	0.78	0.73	**0.88**	0.78	0.67

Comparison of all result options presented in Tables 2, 3 and 4 can be summarized visually. Figures 5 and 6 show the changes of PM10 prediction for consecutive days based on the ABL forecast provided by RFR and LR methods, respectively. Due to the poor outcomes obtained by the DTR method, we decide not to present additional plots showing the pollution prognosis based on this method. However, the above analyses indicate that the LR method is the best choice for ABL prediction.

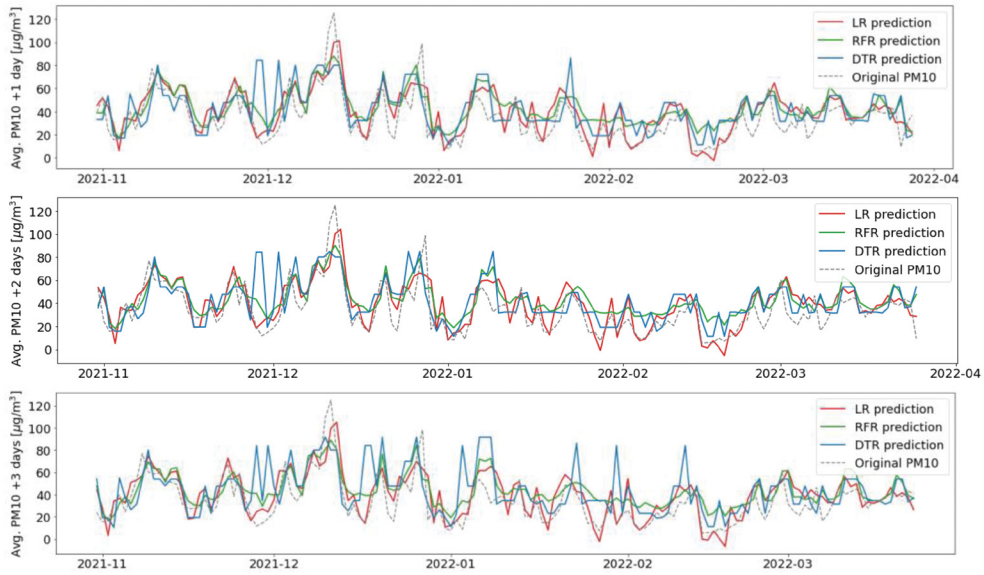

Figure 5: PM10 prediction based on ABL predicted value with RFR procedure for 1st, 2nd and 3rd day, respectively.

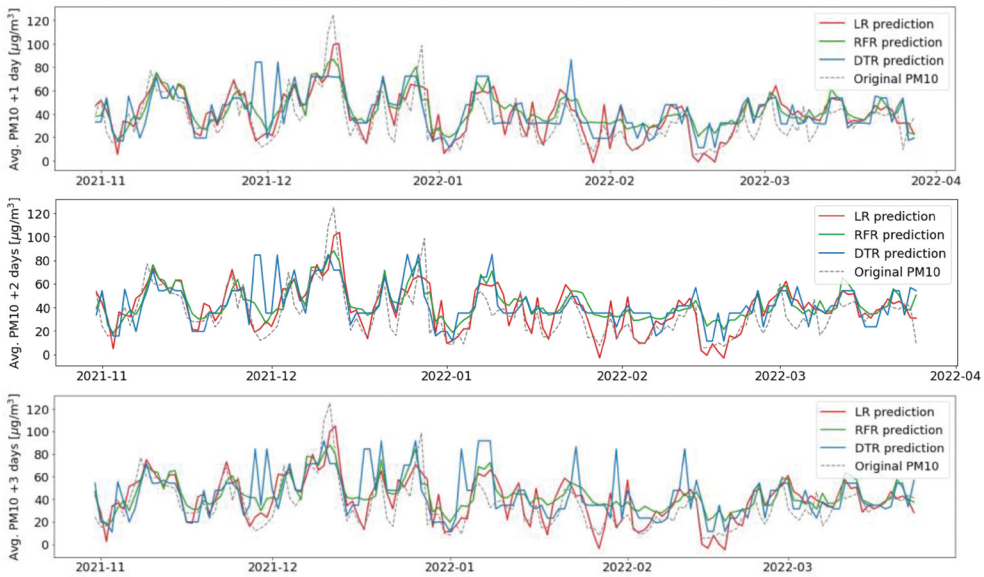

Figure 6: PM10 prediction based on ABL predicted value with RL procedure for 1st, 2nd and 3rd day, respectively.

5 CONCLUSIONS

This paper presented a study of predicting the level of PM10 of air pollution concentration as well as ABL values using several statistical and ML procedures. As a result of the analysis, the best algorithm, which turned out to be LR, was selected. The research was conducted on real data from three measurement stations of the Airly company in Kraków, Poland. The results obtained are very valuable from an application point of view. The training sample covers the whole year, which enables the algorithm to learn different behaviours and trends in the studied phenomenon. On the other hand, the test sample covers the heating period, which is characterised by very high air pollution. Despite the average correlation, the results are characterized by a very small error MAE<10 ($\mu g/m^3$) and a high agreement ($0.88 < IA < 8.9$), certifying that this identification of the smog existence predisposes the proposed algorithm to be used. The obtained results confirmed the thesis that adding real ABL values and their predictions to the feature vector would contribute to obtaining an effective PM10 air pollution prediction algorithm.

In the next research steps, further development of the proposed solution is planned, especially taking into account the issue of spatial scalability of the developed procedure and the possibility of applying the proposed algorithms to mobile devices.

ACKNOWLEDGEMENTS

This work was supported by the national centre for research and development (Project No. POIR.01.01.01-00-0425/20 NCBIR).

REFERENCES

[1] *WHO Global Air Quality Guidelines: Particulate Matter (pm2.5 and pm10), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide*, World Health Organization, 2021 (accessed 10 March 2022).

[2] Osowski, S. & Garanty, K., Forecasting of the daily meteorological pollution using wavelets and support vector machine. *Engineering Applications of Artificial Intelligence*, **20(6)**, pp. 745–755, 2007.

[3] Asghari, M. & Nematzadeh, H., Predicting air pollution in Tehran: genetic algorithm and back propagation neural network. *Journal of AI and Data Mining*, **4(1)**, pp. 49–54, 2016.

[4] Russo, A., Lind, P. G., Raischel, F., Trigo, R. & Mendes, M., Neural network forecast of daily pollution concentration using optimal meteorological data at synoptic and local scales. *Atmospheric Pollution Research*, **6(3)**, pp. 540–549, 2015.

[5] Domańska D. & Wojtylak M., Application of fuzzy time series models for forecasting pollution concentrations. *Expert Systems with Applications*, **39(9)**, pp. 7673–7679, 2012.

[6] Chakraborty K., Mehrotra K., Mohan C.K. & Ranka S., Forecasting the behavior of multivariate time series using neural networks. *Neural networks,* **5(6)**, pp. 961–970, 1992.

[7] Faruk D.O., A hybrid neural network and arima model for water quality time series prediction. *Engineering Applications of Artificial Intelligence,* **23(4)**, pp. 586–594, 2010.

[8] Perez P., Menares C. & Ramirez, C., Forecasting in the most polluted City in South America. *WIT Transaction on Ecology and the Environment*, Vol. 230, WIT Press: Southampton and Boston, pp. 199–204, 2018.

[9] Grover A., Kapoor A. & Horvitz E., A deep hybrid model for weather forecasting. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, pp.379–386, 2015.

[10] Xie, J., Wang, X., Liu Y. & Bai Y., Autoencoder-based deep belief regression network for air particulate matter concentration forecasting. *Journal of Intelligent & Fuzzy Systems*, **34(6)**, pp. 3475–3486, 2018.
[11] Liu J.N., Hu Y., You J.J. & Chan, P.W., Deep neural network based feature representation for weather forecasting. *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, Computer Engineering and Applied Computing, p. 1, 2014.
[12] Kowalski, P.A., Sapała, K. & Warchałowski, W., PM10 forecasting through applying convolution neural network techniques. *International Journal of Environmental Impacts*, **3(1)**, pp. 31–43, 2020.
[13] Kowalski, P.A., Sapała, K. & Warchałowski, W.A., Convolution neural network PM10 prediction system based on a dense measurement sensor network in Poland. *Book of abstracts of Air Quality, Pollution and Management 2019 Kuala Lumpur*, Malaysia, p. 293, 2019.
[14] Toja-Silva, F., Chen, J., Hachinger, S. & Hase, F., CFD simulation of CO_2 dispersion from urban thermal power plant: analysis of turbulent Schmidt number and comparison with Gaussian plume model and measurements. *Journal of Wind Engineering and Industrial Aerodynamics*, **169**, pp. 177–193, 2017.
[15] Chu, A.K.M., Kwok, R.C.W. & Yu, K.N., Study of pollution dispersion in urban areas using computational fluid dynamics (CFD) and geographic information system (GIS). *Environmental Modelling & Software*, **20(3)**, pp. 273–277, 2005.
[16] Lynnyk, I., Vakulenko, K. & Lezhneva, E., Analysis of the air quality in considering the impact of the atmospheric emission from the urban road traffic. *Research Methods in Modern Urban Transportation Systems and Networks*, Springer, pp. 13–27, 2021.
[17] Croitoru, C. & Nastase, I., A state of the art regarding urban air quality prediction models. *E3S Web of Conferences*, Vol. 32, EDP Sciences, p. 01010, 2018.
[18] Bai, L., Wang, J., Ma, X. & Lu, H., Air pollution forecasts: an overview. *International Journal of Environmental Research and Public Health*, **15(4)**, p. 780, 2018.
[19] Liao, Z., Sun, J., Yao, J., Liu, L., Li, H., Liu, J., ... & Fan, S., Self-organized classification of boundary layer meteorology and associated characteristics of air quality in Beijing. *Atmospheric Chemistry and Physics*, **18(9)**, pp. 6771–6783, 2018.
[20] Brown, R.G., *Statistical Forecasting for Inventory Control*, McGraw/Hill, 1959.
[21] Holt, C.E., *Forecasting Seasonals and Trends by Exponentially Weighted Averages (O.N.R. Memorandum No. 52)*, Carnegie Institute of Technology, Pittsburgh USA, 1958.
[22] Winters, P.R., Forecasting sales by exponentially weighted moving averages. *Management Science*, **6(3)**, pp. 324–342, 1960.
[23] Liu, X., Lin, Z. & Feng, Z., Short-term offshore wind speed forecast by seasonal ARIMA-A comparison against GRU and LSTM. *Energy*, **227**, p. 120492, 2021.
[24] Choudhary, A., Kumar, S., Sharma, M. & Sharma, K.P., A framework for data prediction and forecasting in WSN with auto ARIMA. *Wireless Personal Communications*, pp. 1–15, 2021.
[25] Hyndman, R.J. & Athanasopoulos, G., *Forecasting: Principles and Practice*. OTexts, 2014.
[26] Kowalski, P.A. & Warchałowski, W., The comparison of linear models for PM10 and PM2.5 forecasting. *WIT Transaction on Ecology and the Environment*, Vol. 230, WIT Press: Southampton and Boston, pp. 177–188, 2018.
[27] Xie, X., Wu, T., Zhu, M., Jiang, G., Xu, Y., Wang, X. & Pu, L., Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. *Ecological Indicators*, **120**, p. 106925, 2021.

[28] Breiman, L., Random forests. *Machine Learning*, **45(1)**, 5–32, 2001.
[29] Jumin, E., Basaruddin, F.B., Yusoff, Y.B., Latif, S.D. & Ahmed, A.N., Solar radiation prediction using boosted decision tree regression model: a case study in Malaysia. *Environmental Science and Pollution Research*, **28(21)**, pp. 26571–26583, 2021.