BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **XX**, No X

Sofia • 202X

High Performance Machine Learning Models of Large Scale Air Pollution Data in Urban Area

Snezhana G. Gocheva-Ilieva[1], Atanas V. Ivanov[1], Ioannis E. Livieris[2]

[1]Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, Plovdiv, Bulgaria
[2]Department of Mathematics, University of Patras, Patras, Greece
Emails: snow@uni-plovdiv.bg aivanov@uni-plovdiv.bg livieris@upatras.gr

Abstract: *Preserving the air quality in urban areas is crucial for the health of the population as well as for the environment. The availability of large volumes of measurement data on the concentrations of air pollutants enables their analysis and modelling to establish trends and dependencies in order to forecast and prevent future pollution. This study proposes a new approach for modelling air pollutants data using the powerful machine learning method Random Forest (RF) and Auto-Regressive Integrated Moving Average (ARIMA) methodology. Initially, a RF model of the pollutant is built and analysed in relation to the meteorological variables. This model is then corrected through subsequent modelling of its residuals using the univariate ARIMA. The approach is demonstrated for hourly data on seven air pollutants (O_3, NOx, NO, NO_2, CO, SO_2 and PM_{10}) in the town of Dimitrovgrad, Bulgaria over 9 years and 3 months. Six meteorological and three time variables are used as predictors. High-performance models are obtained explaining the data with R^2 = 90%-98%.*

Keywords: *Machine learning, Random Forest, Autoregressive integrated moving average, error correction, time series, forecasting.*

1. Introduction

Nowadays, monitoring and maintaining air quality constitute the main goals of environmental protection. Systematic air pollution with harmful aerosols in urban areas causes severe disease among the population. In the European Union provisions, standards and measures are in place to limit the concentrations of harmful pollutants. The main monitored air pollutants are nitrogen oxides (NOx) such as nitrogen monoxide (NO) and nitrogen dioxide (NO_2), ground level ozone (O_3) and particulate matter with a diameter of up to 10 micrometers (PM_{10}) [1,2]. During the last decade, a decrease of the level of pollution is noted, as evidenced by the latest reports of the European Environment Agency [3]. Bulgaria is one of the EU member states, which still has issues related to clean air in some regions [3]. There are 36 certified permanent automatic stations operating in the country in the main provincial cities and other larger urban centres. The advances of digital technology and the vigorous development of the Internet enabled the accumulation of a large volume of data, regarding the concentrations of the main air pollutants. By leveraging and analysing these data, correlated with atmospheric and meteorological observations, offers a significant potential to provide a deep insight in air pollution.

The significance of environmental studies and in particular those on air quality is topical in several aspects. It is vital for measuring to establish the trends for the concentrations of harmful air pollutants in urban areas. Their systemic exceeding causes various adverse health effects and chronic conditions such as respiratory disease, lung cancer, cardiovascular disease, and even death [4-6]. The main factors, which lead to elevated harmful aerosols in the air, include emissions released by industrial and household combustion processes, transport traffic, etc. Their impact is often combined with unfavourable meteorological and atmospheric processes. It is worth noticing that these vary between separate urban areas, taking into consideration geographical location, air quality preservation measures by local authorities, etc. Another aspect of the studies is determining the relative influence of individual factors, which cause air pollution typical for a given urban area. Here the mathematical modelling of collected empirical data comes to the rescue. A third crucial aspect in the examined issue with air pollution is the ability to make forecasts, which is fully based on mathematical and statistical approaches.

In recent years, these outlined aspects have been extensively studied using the powerful Machine Learning (ML) methods. The main characteristics of these methods include the abilities to process large scale data and extract hidden patterns and dependencies in time series data. Some of the most promising methods are Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF), Classification and Regression Tree (CART). Furthermore, the development of this type of methods and models supports more adequate practical implementation of strategies and measures to mitigate the effects of pollution in order to protect the affected population.

The objective of this study is to develop a combined RF-ARIMA methodology for improving the performance of RF method, where ARIMA states for

autoregressive integrated moving average approach. A general two-step approach is proposed, consisting of building RF regression models and the subsequent adjustment procedure of its residuals with the assistance provided by the univariate ARIMA process. Special consideration is given to error analysis and checking for lack of autocorrelation, which constitutes a significant factor for evaluating the reliability of a model [7]. The proposed approach is evaluated by building RF models and examining their performance in 7 main air pollutants in the town of Dimitrovgrad, Bulgaria, in relation to 6 meteorological time series. Hourly data for a period of 9 years and 3 months are examined.

The paper is organized as follows: Section 2 presents briefly the results of previously published studies in the area of ML and air pollution modelling. Section 3 introduces the proposed approach to improve the performance of RF models and the stages of implementation. Section 4 describes the data used and the results of their pre-processing. Section 5 contains statistical analysis, obtained RF-ARIMA models and forecasts along with a discussion on these.

2. Literature review

In the field of data modelling of concentrations of harmful air pollutants, ML methods are increasingly preferred by a large number of researchers. In [8], Durão et al. studied and forecasted O_3 levels based on hourly data by combining classification trees and multilayer perceptron (MLP) models. Biancofiore et al. [9] applied a recursive neural network, a feed-forward neural network and a multiple linear regression for modelling daily averaged PM_{10} and $PM_{2.5}$ concentrations, depending on meteorological variables. Bougoudis et al. [10] used hybrid type ML models based on feedforward NN, fuzzy logic and RF to forecast CO, NO, NO_2, SO_2 and ozone O_3 pollution levels. For particulate matter data, empirical models were built and examined using wavelet analysis and wavelet-ARIMA models by Zhang et al. [11]. Gardner and Dorling [12] utilized MLP, regression tree and linear regression models to examine the hourly surface O_3 data in relation to meteorological variables. In [13], Singh et al. utilized principal component analysis (PCA) to identify air pollution sources. In addition, the authors proposed tree ensemble models based on bagging and boosting strategies for air quality predictions. Based on their numerical experiments, they concluded that ensemble classification and regression models performed better than SVMs. Bai et al. [14] utilized a NN based on wavelet decomposition for modelling and forecasting the PM_{10}, SO_2, and NO_2, with relation to meteorological conditions. In [15], Dotse et al. combined the predictions of RF, genetic algorithm and neural network for examining daily PM_{10} exceedances. Roy et al. [16] applied and compared the forecasting ability of three methods - multivariate adaptive regression splines, RF and CART. The comparisons of various ML methods for analysis and forecasting of air pollutants are presented in the recent papers [17-22] among others.

3. Proposed RF-ARIMA approach

3.1. The framework

In this paper, we use the highly efficient ML method RF, developed by L e o B r e i m a n for the case of time series regression in combination with univariate ARIMA Box-Jenkins methodology [23, 24]. Similar error-correction procedure is used to improve the performance of CART models in [25].

Modelling will be carried out in the following basic steps.

Step 1. Pre-processing of time series. This includes: descriptive statistics, replacing the missing data, examining the presence of multi-collinearity between variables, studying for a trend, autocorrelation, etc.

Step 2. Generating RF models for each air pollutant depending on the meteorological and time variables.

Step 3. Analysis and selection of optimal RF model for each pollutant according to selected criteria.

Step 4. Checking the residuals (prediction errors) of the RF-models for presence of autocorrelation. In our case, all models have high values of the autocorrelation function (ACF) and residuals.

Step 5. Application of univariate ARIMA methodology for correction of errors by modelling the residuals and eliminating autocorrelation.

Step 6. Calculating the final RF-ARIMA models and assessment of their performance statistics.

Step 7. Application of models for forecasting pollution 24 hours ahead and comparing against holdout samples, unused in model construction.

Let us denote the current dependent variable at step 2 (time series of the pollutant) with Y, the respective RF model through RF_Y and its residuals with $res1 = Y - RF\_Y$. Thus at step 2 and 3, we get

(1) $$Y = RF\_Y + res1$$

At step 5 the dependent variable is the residual $res1$. Since in our case its values are relatively large and auto-correlate, we apply to it the univariate ARIMA method without predictors. The result is

(2) $$res1 = ARres1 + res2,$$

where $ARres1$ is the ARIMA model, and $res2$ is the respective residual.

At Step 6 for the final RF-ARIMA model $\hat{Y}$ we obtain

(3) $$\hat{Y} = RF\_Y + ARres1$$

with residuals $res2$.

3.2. Performance measures

The performance of the model in the general case is assessed using coefficient of determination R^2 and Root mean squared error (RMSE) defined by

(4) $$R^2 = 1 - \frac{\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2}{\sum_{t=1}^{n}\left(Y_t - \bar{Y}\right)^2}, \qquad RMSE = \sqrt{\frac{\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2}{n}},$$

where Y_t are the observed values of the current dependent variable Y, $\bar{Y}$ is its mean, $\hat{Y}$ are the predicted values, n is the sample volume. The respective formulas are also used for the intermediate estimates for model $ARres1$ from (2). The selected optimal models are those with the highest R^2 and minimum *RMSE*.

4. Dataset and dataset pre-processing

4.1. Investigated urban area

The proposed RF-ARIMA approach is evaluated on 7 air pollutants by a certified measurement station in the town of Dimitrovgrad, Bulgaria. The town is located in the centre of the Gornotrakiyska valley on the banks of the Maritsa River, South Bulgaria. It is the administrative centre of Dimitrovgrad Municipality, a regional commercial, industrial and transport town with a population of 34,000 inhabitants. Three European transport corridors go through Dimitrovgrad from West Europe to Turkey, Greece and the Black sea. The climate in the area is transitional temperate to Mediterranean. The average annual temperature for Dimitrovgrad is higher than that for the country of +12.6 °C. This allows intensive evaporation mainly during the summer. The distribution of precipitation is not uniform throughout the year. The geographical and climate characteristics of the region define the existing unfavourable meteorological conditions with regard to the processes in atmospheric air. The sources of harmful emissions in atmospheric air are commercial entities from the processing sector – with large industrial facilities in energy production, chemical and cement industry, the service sector – transport, commerce, services, etc., residential sector – emissions related to local residential heating.

4.2. Data description and pre-processing

The time series data used contained hourly measurements of air pollutants O_3, NOx, NO, NO_2, CO, SO_2, PM_{10} over the period from 1 January 2005 to 7 March 2014. The total sample size is $N = 80472$. The pollutants time series are considered as dependent variables. As independent variables are used 6 meteorological time series. These include wind speed, m/s (WS), wind direction, degree (WD), relative humidity, % (HUM), air temperature, $°C$ (TEMP), sun radiation, W/m^2 (GSR) and air pressure, mbar (PRESS). The studied time series show clearly expressed cyclical and seasonal behaviour. The RF does not include time variables by default, so we introduce three time variables, respectively HOUR, DAY, DAYHOURS. The descriptive statistics of available pollutant data is presented in Table 1.

Table 1 shows that the mean value of PM_{10} equals 55.280 $\mu g / m^3$ and the maximum value is over $895 \mu g / m^3$. This indicates that permanent exceedances above the allowed by European standards daily limit for PM_{10} emissions of $50 \mu g / m^3$ and yearly limit of $40 \mu g / m^3$ are available. The levels of remaining pollutants are within the permissible limits. Although their mean values are low, these pollutants form a constant background. This causes considerable health risks especially for older populations and small children.

Table 1. Descriptive Statistics of the observed pollutant time series[1]

Statistic	O_3, $\mu g / m^3$	NOx, $\mu g / m^3$	NO, $\mu g / m^3$	NO_2, $\mu g / m^3$	CO, mg / m^3	SO_2, $\mu g / m^3$	PM_{10}, $\mu g / m^3$
N valid	74932	76724	75219	76015	73793	77575	77167
N missing	5540	3748	5253	4457	6679	2897	3305
Missing, %	7%	5%	7%	6%	9%	4%	4%
Mean	50.334	15.763	7.206	19.528	0.553	33.667	55.280
Median	46.996	9.8520	1.530	15.748	0.322	13.660	38.920
Std. Deviation	35.015	23.031	21.140	16.570	0.724	61.949	56.157
Variance	1226.03	530.45	446.91	274.57	0.53	3837.66	3153.56
Skewness	3.122	5.782	7.771	2.174	3.687	6.014	3.716
Kurtosis	64.527	51.604	84.096	7.493	21.063	67.520	21.969
Minimum	0.002	0.001	0.001	0.001	0.000	0.001	0.010
Maximum	1094.2	467.6	466.4	212.0	10.4	1758.3	895.6

[1]Std. Error of Skewness is 0.09, Std. Error of Kurtosis is 0.018.

Table 1 presents large discrepancies between Mean and Median values and high values for the ratio Skewness/Std. Error of Skewness and Kurtosis/Std. Error of Kurtosis. This implies that the considered variables do not follow normal distribution. For this reason, the direct application of multivariate regression approaches is not recommended. However, ML methods, and especially RF, are not sensitive to the distribution type of the variables when building regression models. In addition, the small volume of missing data were replaced utilizing linear interpolation. The missing data for meteorological variables were processed in the same way. The resulting variables were denoted as the initial ones.

Generally, it is considered that the selection of predictors in ML methods needs to avoid mutually correlated variables. In order to study the presence of multi-collinearity between the variables, the nonparametric correlation coefficients Spearman's Rho have been calculated (see for instance [26]). For the meteorological variables, no large values of the corresponding bivariate Spearman's Rho coefficients have been found. Therefore, they can all be used as predictors.

5. Modelling results with discussion

5.1. Random Forests models

When building RF models, various values are set for the following hyper-parameters: m - number of trees in the forest, r – number of randomly selected

predictors from the pool of predictors and k - minimum cases in parent node of the tree. ML training is carried out and overfitting of the models avoidance is achieved using an out of bag (OOB) procedure for independent testing [23]. Among the numerous candidate models, the optimal ones are obtained for all pollutants with hyper-parameters $m = 200, r = 3$ and $k = 5$. RF models were built using Salford Predictive Modeler (SPM) software. Summary statistics for the selected RF models are given in Table 2.

Table 2. Summary statistics for selected RF models

RF model	R^2 OOB	R^2	RMSE OOB	RMSE	Relative variable importance[1]	DW of residuals
RF_O_3	0.836	0.958	14.557	7.861	HUM (100), TEMP (98.3), WS (64.1), DAYHOURS (58.2), DAY (49.9), PRESS (30.5), GSR (23.6), HOUR (21.8), WD (4.4)	0.612
RF_NOx	0.692	0.936	12.791	6.640	WS (100), HOUR (75.1), TEMP (54.8), DAYHOURS (48.2), DAY (41.6), GSR, (16.1), PRESS (15.9), WD (15.2), HUM (13.2)	1.101
RF_NO	0.626	0.935	12.816	6.402	WS (100), HOUR (86.3), DAYHOURS (76.6), DAY (67.2), TEMP (60.6), PRESS (22.5), GSR (19.8), HUM (19.0), WD (17.8)	1.175
RF_NO_2	0.749	0.932	8.327	4.708	WS (100), HOUR (57.4), TEMP (50.2), DAYHOURS (48.6), DAY (39.0), GSR (16.3), HUM (13.6), PRESS (12.0), WD (10.0)	0.999
RF_CO	0.805	0.958	0.330	0.168	TEMP (100), DAYHOURS (82.8), WS (71.0), DAY (70.7), HOUR (56.7), GSR (18.2), PRESS (17.9), HUM (16.6), WD (13.2)	0.996
RF_SO_2	0.551	0.899	41.943	23.696	DAYHOURS (100), DAY (91.8), HOUR (34.2), TEMP (34.1), PRESS (30.5), WS (23.2), HUM (15.0), GSR (13.9), WD (6.1)	0.871
RF_PM_{10}	0.739	0.942	30.862	16.413	TEMP (100), DAYHOURS (75.9), DAY (64.3), WS (59,5), HOUR (42.3), PRESS (32.0), HUM (32.0), GSR (15.6), WD (8.8)	0.963

[1]The first variable is considered to have weight of 100 scores, and others are relative to it.

Table 2 presents high R^2 both for the test OOBs and the RF models, with all of the latter having R^2 over 90%. The last but one column of Table 2 provides valuable information about the contribution of individual predictors in the model. For example, the pollutant PM_{10} is influenced most significantly by air temperature and wind speed, along with time variables. The latter implicitly contains information about the RF classification trees, consistent with time values (DAYHOURS, HOUR, DAY). From the relative variable importance values, the concentrations of ozone O_3 are determined by the levels of relative humidity, air temperature and wind speed. The nitrous oxides in the air NO, NOx and NO_2

depend mostly on wind speed and air temperature. Levels of CO depend basically on air temperature and wind speed and SO_2 – on air temperature and pressure.

The last column of Table 2 provides the statistics for Durbin-Watson (DW) statistics to test for presence of autocorrelation in the residuals. All DW values were significantly different from 2, therefore there exists a serial correlation in the residuals from all considered RF models. Fig. 1 shows the autocorrelation function (ACF) and partial ACF (PACF) of the residuals for the case of O_3 and PM_{10}.

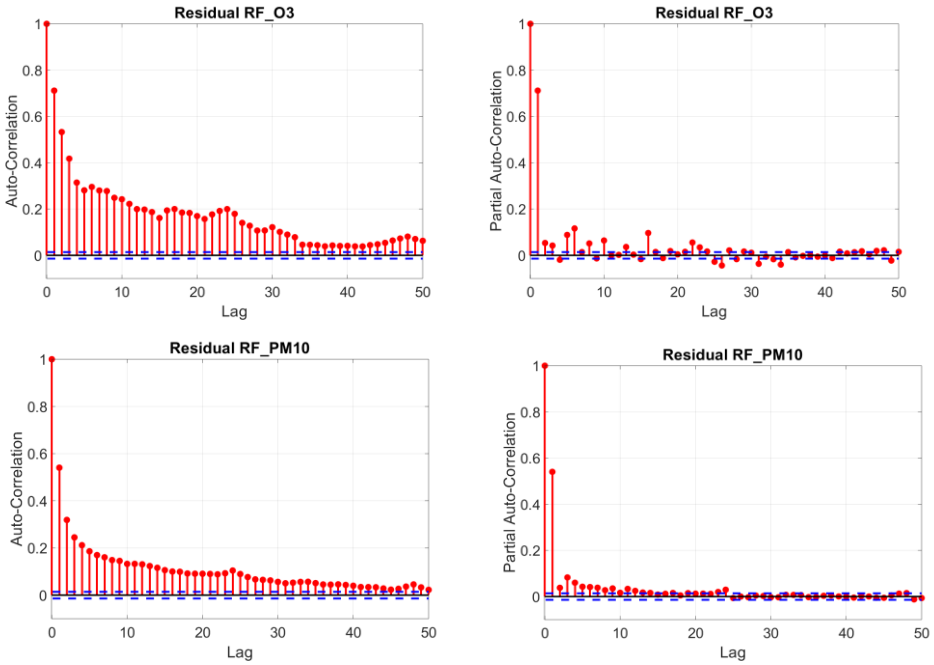

Fig. 1. ACF and PACF of the residuals of the RF models for O_3 and PM_{10}

Fig. 2 illustrates very good fit of the RF models with the time series of the pollutants. In the selected models, the outliers were underestimated. This can be explained by the relatively small number of high outliers at the sample size of $N = 80472$. Achieving too much accuracy would result in overfitting of the models.

5.2. ARIMA models of residuals of the RF models

Following the proposed modelling approach in 3.1 Step 5, we build univariate ARIMA models of the residuals. From the ACFs plots of the residuals, we conclude that the time series contain a seasonal dependence of 24 hours.

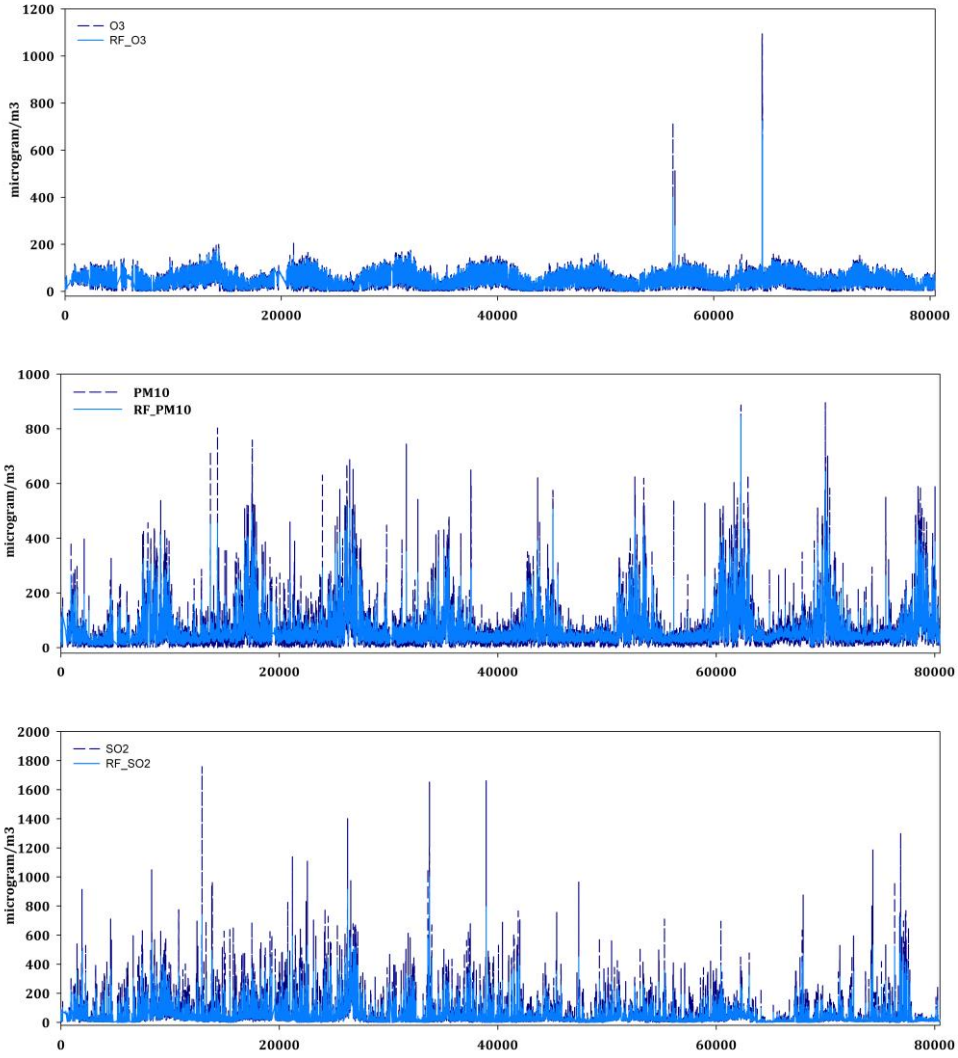

Fig. 2. Comparison of the studied data with those predicted by the RF models for O_3, SO_2 and PM_{10}

The general method is in the form ARIMA $(p,d,q)_{24}$. In addition, all PACFs have large values in lag 1, but smaller than ± 1. This suggests that each current measurement depends on the value in the previous hour, i.e. autoregressive process AR (1), without a trend, thus $p=1, d=0$. The excesses in PACFs over the confidence intervals suggest the values q of MA processes. The parameters of the obtained ARIMA $(p,d,q)_{24}$ models are given in the second column of Table 3.

Table 3. Summary statistics for ARIMA models *ARres*1 and resulting RF-ARIMA models

Pollutant	Models *ARres*1 for residuals *res*1	R^2 of $\hat{Y}$ model prediction	RMSE of $\hat{Y}$	DW of *res*2
O_3	ARIMA(1,0,16)$_{24}$	0.978	5.336	2.061

NOx	ARIMA(1,0,24)$_{24}$	0.944	5.6631	2.133
NO	ARIMA(1,0,24)$_{24}$	0.938	5.529	2.190
NO_2	ARIMA(1,0,24)$_{24}$	0.945	3.965	2.054
CO	ARIMA(1,0,24)$_{24}$	0.966	0.139	2.078
SO_2	ARIMA(1,0,5)$_{24}$	0.902	20.736	1.998
PM_{10}	ARIMA(1,0,11)$_{24}$	0.950	13.628	2.042

At the Step 6, the values of the final models RF-ARIMA, $\hat{Y}$ are calculated. Summary statistics for the built $\hat{Y}$ models are shown in Table 3. The last column shows that the DW statistics of the final residuals *res*2 are close in value to 2. This indicates that the models do not have auto-correlated residuals and are consistent. The comparisons of the RF and $\hat{Y}$ models from Tables 2 and 3, show the final models $\hat{Y}$ achieve slightly improved performance. The highest R^2 were obtained for O_3 and CO, respectively, $R^2(O_3) = 97.8\%$ and $R^2(CO) = 96.6\%$, while the smallest for SO_2, $R^2(SO_2) = 90.2\%$.

Remark. The models were generated using ordinary PCs. The calculations of one RF model with SPM were performed within 3-4 minutes. But identification and error correction with ARIMA using IBM SPSS took about 1.5 hours.

5.3. Forecasting using the obtained RF-ARIMA models

With known values of the predictors, the approach proposed in this study allows the forecasting of the dependent time series in many future moments. This will be demonstrated with the help of the holdout data sample for 24 hours ahead, not used to build the models of the 7 pollutants. The obtained forecasts were compared against the measurements. Fig. 3 shows the measured values and the forecasts for 24 hours ahead using the holdout dataset of O_3 for 7 March 2014. Very good fit is observed. It is achieved the quality of fit with $R^2 = 0.799\%$.

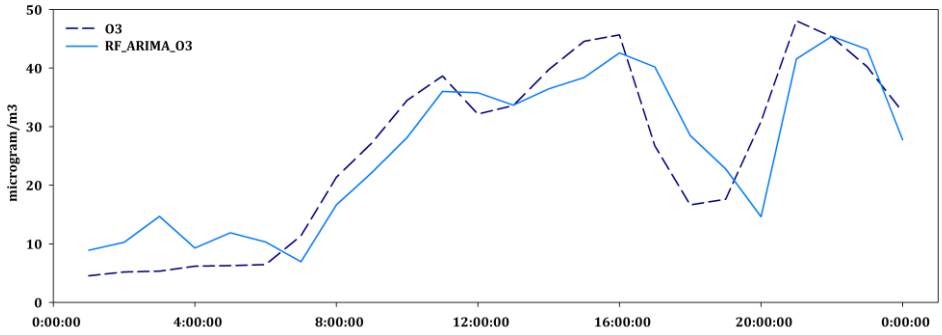

Fig. 3. Sequence pots of the measured O_3 values and the RF-ARIMA model forecasts for 24 hours ahead

The next Fig. 4 shows the forecasts of PM_{10} for 24 hours using the same holdout dataset of predictors for 7 March 2014. The behaviour of the forecast is close to the measurements, but the coincidence is inferior in quality to that of ozone. For the first 12 hours the adequacy of the model reaches $R^2 = 0.73\%$.

The obtained results should be compared with those of other authors. For example, in the intelligent system [10] for CO, NO, NO_2, SO_2 and O_3, RF regression models have been found to give the highest predictive results. In particular, the quality of fit of the obtained models achieved R^2 in the range of 87% to 94%, with the exception of SO_2 models, where R^2 varies from 66 to 78%. Authors of [16] predict O_3 with RF model with R^2 over 94%. In [20], several ML methods for predicting O_3 concentrations were compared, with the highest results of up to 93% being obtained with RF models.

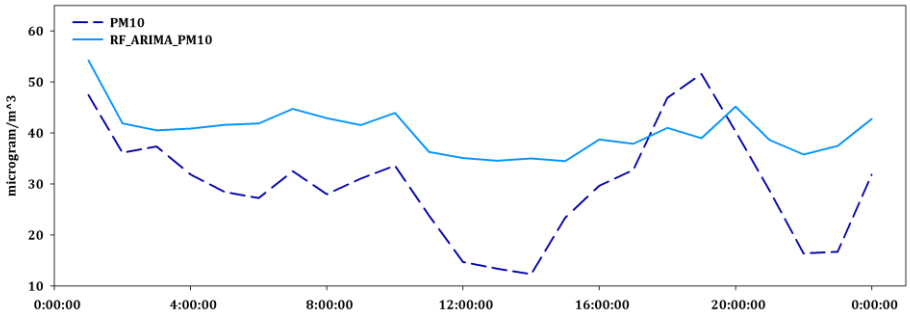

Fig. 4. Sequence plots of the measured PM_{10} values and the corresponding RF-ARIMA model forecasts for 24 hours ahead

Based on the previous analysis, we are able to conclude that the proposed approach with RF-ARIMA provides an opportunity to build high-performance models and obtain very good quality of forecasting the concentrations of air pollutants.

***Acknowledgements*:** The study is supported by the Grant No BG05M2OP001-1.001-0003, financed by the Science and Education for Smart Growth Operational Program (2014-2020), co-financed by the European Union through the European structural and Investment funds.

References

1. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe. – Official Journal of the European Union, **L 152**/1, 2008.
2. Air Quality Standards. European Commission. Environment, 2015. [Online], **http://ec.europa.eu/environment/air/quality/standards.htm**
3. Air Quality in Europe - 2019 Report. European Environment Agency. EEA Report 10, 2019 [Online], **https://www.eea.europa.eu/publications/air-quality-in-europe-2019.**
4. Brunekreef, B., S. T. Holgate. Air Pollution and Health. – The Lancet, Vol. **360,** 2002, No 9341, pp. 1233-1242.
5. Guarnieri, M., J. R. Balmes. Outdoor Air Pollution and Asthma. – The Lancet, Vol. **383**, 2014, No 9928, pp. 1581-1592.
6. Hu, W., K. Mengersen, A. McMichael, S. Tong. Temperature, Air Pollution and Total Mortality During Summers in Sydney, 1994-2004. – International Journal of Biometeorology, Vol. **52**, 2008, No 7, 689-696.

7. Livieris, I. E., S. Stavroyiannis, E. Pintelas, P. Pintelas, A Novel Validation Framework to Enhance Deep Learning Models in Time-Series Forecasting. Neural Computing and Applications, 2020. https://doi.org/10.1007/s00521-020-05169-y
8. Durão, R. M., M. T. Mendes, M. J. Pereira. Forecasting O3 Levels in Industrial Area Surroundings up to 24 h in Advance, Combining Classification Trees and MLP models. – Atmospheric Pollution Research, Vol. **7**, 2016, pp. 961-970. \
9. Biancofiore, F., M. Busilacchio, M. Verdecchia, B. Tomassetti, E. Aruffo, S. Bianco, S. Di. Tommaso, C. Colangeli, G. Rosatelli, P. Di. Carlo. Recursive Neural Network Model for Analysis and Forecast of PM10 and PM2.5. – Atmospheric Pollution Research, Vol. **8**, 2017, No 4, pp. 652-659.
10. Bougoudis, I., K. Demertzis, L. Iliadis. HISYCOL a Hybrid Computational Intelligence System for Combined Machine Learning: the case of Air Pollution Modelling in Athens. – Neural Computing and Applications, Vol. **27**, 2016, No 5, pp. 1191-1206.
11. Zhang, H., S. Zhang, P. Wang, Y. Qin, H. Wang. Forecasting of Particulate Matter Time Series Using Wavelet Analysis and Wavelet-ARMA/ARIMA Model in Taiyuan, China. – Journal of the Air & Waste Management Association, Vol. **67**, 2017, No 7, pp. 776-788.
12. Gardner, M. W., S. R. Dorling. Statistical Surface Ozone Models: An Improved Methodology to Account for Non-linear Behavior. – Atmospheric Environment, Vol. **34**, 2000, pp. 21-34.
13. Singh, K. P., S. Gupta, P. Rai. Identifying Pollution Sources and Predicting Urban Air Quality Using Ensemble Learning Methods, – Atmospheric Environment, Vol. **80**, 2013, pp. 426-437.
14. Bai, Y., Y. Li, X. Wang, J. Xie, C. Li. Air Pollutants Concentrations Forecasting Using Back Propagation Neural Network Based on Wavelet Decomposition with Meteorological Conditions. – Atmospheric Pollution Research, Vol. **7**, 2016, No 3, pp. 557-566.
15. Dotse, S.-Q., M. I. Petra, L. Dagar, L. C. De Silva. Application of Computational Intelligence Techniques to Forecast Daily PM10 Exceedances in Brunei Darussalam. – Atmospheric Pollution Research, Vol. **9**, 2018, No 2, pp. 358-368.
16. Roy, S.S., C. Pratyush, C. Barna. Predicting Ozone Layer Concentration Using Multivariate Adaptive Regression Splines, Random Forest and Classification and Regression Tree. – Advances in Intelligent Systems and Computing, Vol. **634**, 2018, pp. 140-152.
17. Liu, B., C. Shi, J. Li, Y. Li, J. Lang, R. Gu. Comparison of Different Machine Learning Methods to Forecast Air Quality Index. – Lecture Notes in Electrical Engineering, Vol. **542**, 2019, pp. 235-245.
18. Masih, A. Comparative Analysis of Tree, Meta-Learning and Function Classifiers to Predict the Atmospheric Concentration of NO_2. – Journal of Environmental Accounting and Management, Vol. **8**, 2020, No 1, pp. 31-39.
19. Masmoudi, S., H. Elghazel, D. Taieb, O. Yazar, A. Kallel. A Machine-Learning Framework for Predicting Multiple Air Pollutants' Concentrations Via Multi-Target Regression and Feature Selection. – Science of the Total Environment, Vol. **715**, 2020, 136991.
20. Martínez-España, R., A. Bueno-Crespo, I. Timón, J. Soto, A. Muñoz, J. M. Cecilia. Air-Pollution Prediction in Smart Cities through Machine Learning Methods: A Case of Study in Murcia, Spain. – Journal of Universal Computer Science, Vol. **24**, 2018, No 3, pp. 261-276.
21. Veleva, E., I. Zheleva. GARCH Models for Particulate Matter PM10 Air Pollutant in the City of Ruse, Bulgaria. – AIP Conference Proceedings, Vol. **2025**, 2018, 040016.
22. Joharestani, M. Z., C. Cao, X. Ni, B. Bashir, S. Talebiesfandarani. PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. – Atmosphere, Vol. **10**, 2019, No 7, 373.
23. Breiman, L. Random Forests. – Machine Learning, Vol. **45**, 2001, No 1, pp. 5-32.

24. Box, G. E. P., G. M. Jenkins. Time Series Analysis, Forecasting and Control. Revised Edition. San Francisco. – Holden-Day, San Francisco, 1976.
25. Gocheva-Ilieva, S.G., D. S. Voynikova, M. P. Stoimenova, A. V. Ivanov, I. P. Iliev. Regression Trees Modeling of Time Series for Air Pollution Analysis and Forecasting. – Neural Computing and Applications, Vol. **31**, 2019, No 12, pp. 9023-9039.
26. Weaver, K. F., V. Morales, S. L. Dunn, K. Godde, P. F. Weaver. Pearson's and Spearman's Correlation. – In: An Introduction to Statistical Analysis in Research, John Wiley & Sons, Inc., New Jersey, 2017, Ch. 10, pp. 435-471.