Analysis of Romanian Air Quality using Machine Learning Techniques

Andreea-Mihaela NICULAE
The Bucharest University of Economic Studies, Romania
niculaeandreea17@stud.ase.ro

Air quality monitoring has become an increasingly important subject and is one of the most important concerns of governments worldwide. Monitoring is especially important in industrial and urban areas. Due to the many forms of pollution generated mainly by fuel consumption, means of transport, coal-fired electricity generation, etc., air quality is negatively affected. As the current trend is an increase in air pollution, it is necessary to install equipment to measure air quality both in areas with a high risk of pollution and in areas where pollution is low. These types of equipment must communicate in real-time their measured values, which then can be accessed to be able to make analyzes and predictions regarding air quality in a certain geographical area, areas with a high industrialization level, or in areas with a growing population. This paper aims to investigate the application of big data and machine learning techniques to make predictions on air quality using, as a source of data, data recorded in the period 2018-2021 from measurement probes throughout Romania for PM_{10}, NO_2, O_3, and SO_2. The results of this paper's analysis show that time-series models outperform traditional models. Moreover, ANN models are successful only in classifying pollutants' AQI levels and not their actual values.

***Keywords**: Big Data, Machine Learning, Romania, Air quality, MLR, SARIMA, C5.0, Random Forest, ANN*

1 Introduction

In recent years, the world's fast development has been both beneficial and harmful to its population. While this development aims to help people live a better life, it also seems to reduce the quality of one's life, by creating problems such as global warming and air pollution. The latest one is a subject that more governments should pay attention to, as it is dangerous to the entire world's health.

According to the World Health Organization, 9 out of 10 people around the globe, live in a place where the air quality exceeds their guideline limits [1]. This means that more than 90% of the world breathes air (along with the pollutants inside it) that harms the body, exposing it to different diseases, which affect both the lungs and other organs.

With such possible side effects of air pollution, it is necessary for most governments to try and lower the damages caused by pollution, or focus on lowering the pollution itself. The latest is of growing importance nowadays, as it has more long-term benefits. In addition, the increase in data sources around the globe makes it easier to analyze air quality and provide accurate models to predict and fight pollution.

Monitors around the world provide information on overall air quality: pollutants and AQI. AQI is an index computed as the maximum of all individual pollutants values of AQI. In Europe, AQI takes values from one to six, where one means that air quality is Good, and six means that air quality is Extremely Poor. To compute AQI, the following formula is used:

$$AQI = \max(AQI_{PM_{2.5}}, AQI_{PM_{10}}, \dots, AQI_{pollutant})$$

The European Environment Agency decided on the concentration intervals of each pollutant to help compute their AQI [2]. The EEA updates these values each year. In *Table 1* are the centralized values for the 2022 pollutants AQI intervals, with the colors according to the EEA regulations.

Forecasting air quality is a complex subject that contains algorithms, techniques, and methods from numerous topics: big data, machine learning, time series, and others.
This paper focuses on some available models used in forecasting air pollution, such as MLR, ARMA, ARIMA, and SARIMA, decision trees (ID3, C4.5, and C5.0), random forest, ANN, while also presenting other popular models used by other papers for predicting air quality. In the end, the practical part of the paper contains an analysis of Romanian air quality data.

Table 1. European AQI and concentration range for each pollutant (AQI sub-index)

AQI Level	AQI	Index level (based on pollutant concentrations in µg/m^3)				
		PM_{10}	$PM_{2,5}$	NO_2	O_3	SO_2
Good	1	0-20	0-10	0-40	0-50	0-100
Fair	2	20-40	10-20	40-90	50-100	100-200
Moderate	3	40-50	20-25	90-120	100-130	200-350
Poor	4	50-100	25-50	120-230	130-240	350-500
Very Poor	5	100-150	50-75	230-340	240-380	500-750
Extremely Poor	6	150-1200	75-800	340-1000	380-800	750-1250

2. Literature Review

There are two main topics of concern regarding the analysis of air pollution: *monitoring air quality* and *forecasting air quality* [3]. Since more and more governments focus on reducing their country's pollution, more funds are available to facilitate better analyses, by opening more monitoring stations around the country and by offering relevant data sources for analysts to model.

A. Monitoring air quality

While this paper has the purpose to model air quality, it is vital to understand where the data comes from. The IoT makes it possible to monitor air pollution both using government stations (of high performance, but costly) or by using portable sensors (of decent performance, but much easier to obtain and at a low cost) [4]. These sensors monitor the most important pollutants, according to each country's needs and regulations. These are, most of the time: $PM_{2.5}$, PM_{10}, O_3, NO_2, and SO2. Using these, one can easily determine AQI and make further analyzes.

B. Air quality forecast

Forecasting air pollutants and thus, air quality itself, is a subject of increasing interest in recent years. Because air quality data presents itself mainly from sensors and monitoring stations, there is a high amount of data available to model, which makes Big Data models more favorable to use [5]. In Figure 1 are presented the most frequently used big data techniques to forecast air quality.

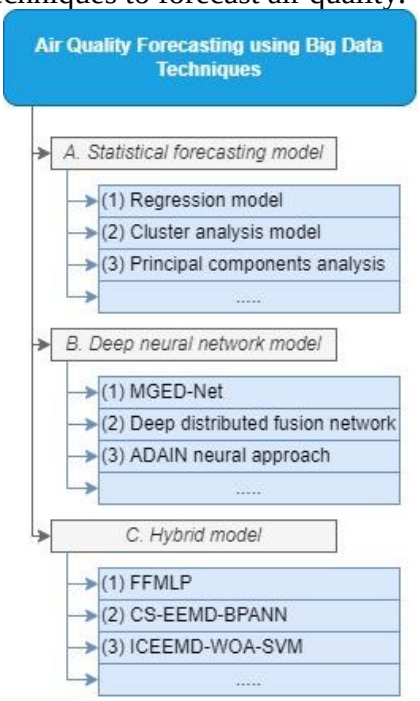

Fig. 1. Air Quality Forecasting using Big Data Techniques [5]

Most models for predicting air quality are linear regression-based [3, 6]. These models assume the linearity of air quality data; however, as some studies suggest [4, 5], pollutants do not behave or evolve linearly, thus these models are not efficient in forecasts. Other widely used linear models are time series-based [7], such as ARMA, ARIMA, and SARIMA. The same linearity problem arises for these models, but even so, most of the time these models perform better than the MLR ones.

Deep neural network models are adaptive and work better with nonlinear data [6], being used more and more in recent papers to estimate and forecast air quality. They are easier to use in model building and their results are significantly better than the MLR model. However, studies show that researchers have a hard time deciding which model is superior between DNN and ARIMA [6, 8], as further analyzes and comparisons between the two are needed.

Another category of models that work well on air quality forecasting is decision trees based on: ID3 [9], C4.5 [10], and C5.0 [11]. One paper discovered that by using the newest and improved technique, C5.0, to model air quality in New Delhi, India, the author obtained a comparatively higher accuracy than other algorithms [11] seen in the respective paper's literature review.

Most of the hybrid, complex models used to forecast air quality start from the ANN model:

- FFMLP (Feed-Forward Multi-Layer Perceptron) [12];
- ANFIS (Adaptive neuro-fuzzy inference system) [13] - which combines the fuzzy logic with the ANN model;
- NARX (Nonlinear autoregressive model with external input) [13] – which combines the logic from the ARIMA series with the ANN model;
- CS-EEMD-BPANN (Cuckoo search - Ensemble Empirical Mode Decomposition - Back-propagation artificial neural networks).

In one of these papers, the author discovered that the RMSE obtained from models based on the ANFIS and NARX methodologies is much lower than the RMSE of traditional, non-complex models [13], especially when using them on big data sets.

Apart from all these supervised models, there are also unsupervised techniques, such as clustering and principal component analysis [14]. Some researchers have used clustering and PCA, for example, to see which meteorological variables were related to the concentration of air pollutants; or to classify monitoring locations - useful for optimizing the monitoring system, by efficiently placing sensors in the network. Some researchers have also used PCA to detect errors in air quality data [14].

3. Methodology

Before presenting the algorithms used in this paper's practical part, one must know about two concepts: Big Data and Machine Learning. These two are the quintessence of proper air pollution analysis, as their concepts cover all the important aspects necessary in an analysis.

A. Big Data analysis

The term *Big Data* refers to enormous data sets that, because of their exponential growth and complexity, are hard to be efficiently processed and used by traditional DBMS. In 2017, data sets were considered Big Data if they meet the "17 Vs" [15] (initially only 3 Vs, then 4, 5, 10, 14, and lately 17): **V**olume, **V**elocity, **V**alue, **V**ariety, **V**eracity, **V**isualization, **V**alidity, **V**olatility, **V**iscosity, **V**irality, **V**enue, **V**ariability, **V**ocabulary, **V**agueness, **V**erbosity, **V**oluntariness, **V**ersatility. Air quality data sets meet many of these characteristics, thus air pollution is a Big Data set. This is a very important part for further analysis, as Big Data technologies provide better, more accurate results, easy to apply (and preferable!) to these data sets [16].

B. Machine Learning

Considered the working horse of Big Data [17], *Machine Learning* is a branch of Artificial Intelligence characterized by the basic idea that working systems can learn from the available data, identify patterns in them, and make decisions, simulating the activity of the human mind.

There are three Machine Learning techniques [17]: *supervised learning*, *unsupervised learning*, and *semi-supervised learning*. Supervised learning distinguishes itself by the fact that the user already knows the result they are trying to obtain, whereas in unsupervised learning techniques the desired result is not so clear from the data. The most widely used unsupervised learning algorithms are clustering, principal component analysis,

factor analysis, and even some kind of artificial neural networks [18]. The most used supervised learning algorithms are regressions (MLR, PLS, PCR and GLM), algorithms based on decision trees (ID3, C4.5, C5.0 and Random Forest), support vector machine and artificial neural networks [18, 19].

Machine Learning has a key element: its usage assumes the division of available data into two sets, one for training the model and one for testing the obtained model. This is very useful for obtaining efficient models, which accurately depict the studied phenomena. Moreover, by doing this, machine learning combats an occurring problem seen in modeling air quality data, *overfitting* [5].

C. Multiple Linear Regression (MLR)

The MLR algorithm used to model a linear relationship between a dependent variable, called the response, and multiple independent variables, called predictors, uses the following mathematical equation [20]:

$$y = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + \varepsilon$$

Where b_k is the regression coefficient, x_k is the predictor, y is the response, i takes values from one to k, and ε is the model's error.

As mentioned in the literature review, this model assumes that it uses linear data, which is not always the case with air pollution data sets. MLR is one of the chosen algorithms for this paper to test the hypothesis that linear models do not perform as well as other models in air quality analysis and forecasting.

D. ARMA, ARIMA, SARIMA

The time series models, ARMA (Auto-Regressive Moving Average), ARIMA (Integrated ARMA), and SARIMA (Seasonal ARIMA), have a common basis: they model a variable using both its past values and an error term. These models require the knowledge of only one pollutant evolution in time to make a forecast, making them more desirable to use for air pollution analysis.

ARMA builds a model using a variable's past values, an error term, and the error term's historical values. ARIMA builds a model starting from ARMA, but has, in addition, a backshift operator, whose role is to "send the variable in the past". SARIMA builds a model starting from ARIMA, but also takes into consideration the possible seasonality of a model as well.

The advantage of using these models is that the analysis and forecasting of each pollutant are much faster, depending only on its past values, its evolutionary trend, and the seasonality identified. The reasoning behind choosing these models is to test the hypothesis that time series models sometimes perform better than both linear and nonlinear models found in the literature review.

E. Decision Tree C5.0 algorithm

Decision Trees are instruments used both in classification problems and in predictions. Graphically presented in the form of a tree flow chart, where each node is a test decision, and each branch is a test result, a decision tree is easy to understand and use. The C5.0 algorithm is the latest, improved version of classifying data using a decision tree and it uses the concept of entropy to measure the purity of a variable: the tree's leaves will have values associated between 0 and 1, where 0 means a homogenous class and 1 means the maximum amount of disorder. Figure 2 shows an example of a decision tree constructed using the C5.0 algorithm on dummy air quality data to classify the PM_{10} AQI category using NO_2 and SO_2.

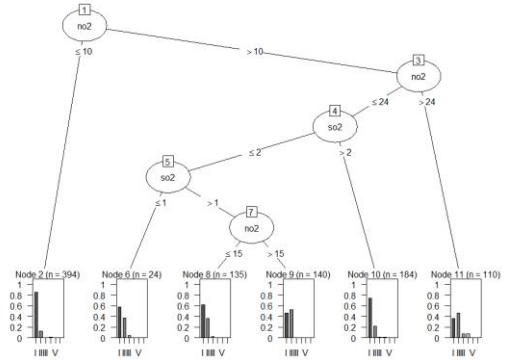

Fig. 2. Example of C5.0 Decision Tree for air quality data

This algorithm is part of the paper's used algorithms to test the hypothesis that, sometimes, this model provides the highest

forecasting accuracy among other used models.

F. Random Forest

Random Forest is a robust method that builds multiple decision trees to train, aggregating their results into one, achieving a higher model forecast accuracy. This is a powerful model because it limits the decision tree overfitting issue caused by bias and variance in data, while also increasing the forecast precision. For classification problems, the aggregated result is the majority vote, while for regression problems, the aggregated result is the computed average value.

G. Artificial Neural Network (ANN)

Artificial neural networks are complex models for solving nonlinear problems with a varied number of outputs. One of the most used ANN models is the MLP (Multilayer Perceptron). When building the MLP network, there are multiple layers of artificial neurons: one input layer (with input variables), at least one hidden layer (with computed neurons), and one output layer (used to compute the target variable). To compute the desired result, activation functions are required. Figure 3 shows the architecture of an ANN MLP with three input neurons, one hidden layer, and one output neuron.

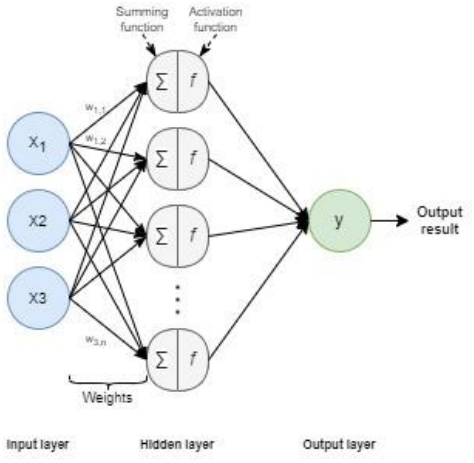

Fig. 3. Artificial Neural Network architecture

ANN is a powerful algorithm, useful in the forecast of air quality, since it does not require a thorough understanding of the dynamics between air pollution concentration levels (or AQI values, if we refer to air quality) and other explanatory variables. ANN is a part of the algorithms used in this paper to test the hypothesis that nonlinear models are superior to linear ones, such as MLR.

4. Romanian AIR Quality Analysis

A. Data source

The data used in the practical part comes from an open-source free website (https://aqicn.org/) which offers historical data of recorded air pollutants around the world. The source is not official, and the given data is subject to change. ANPM (Romanian National Environmental Protection Agency) under the World Air Quality Index Project provided Romanian air quality data.

The data set contains information on several pollutants recorded in Romania: PM_{10}, NO_2, O_3, and SO_2. $PM_{2.5}$ is missing in most of the stations in the country. The models built in this paper use only the four variables mentioned above.

For constructing the data set, the historical daily data from at least one monitoring station per county was extracted from the website. Since some stations did not record any values for O_3, they were removed from the final data set. Time-wise, the final data set contains values from January 2018 to December 2021. All missing data contain values obtained using the interpolation method. Moreover, to analyze Romanian data, the models use the computed average value of each of the four pollutants from the values of all the recorded stations.

B. Results and interpretation

Figure 4 contains the evolution of the extracted and computed Romanian air quality data: it is easy to see that the data is nonlinear. Moreover, O_3 in Romania seems to have a seasonal component in its evolution.

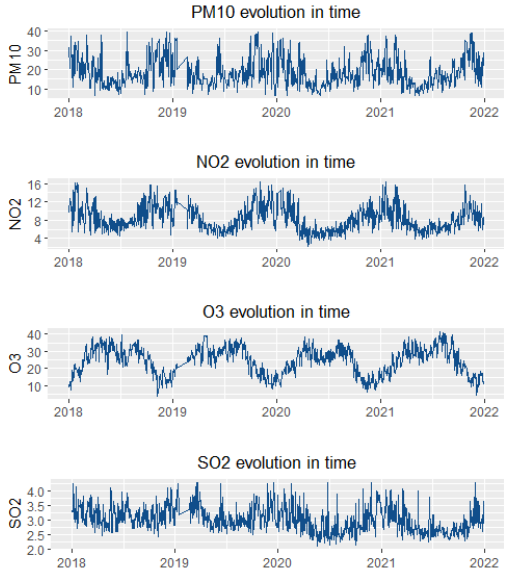

Fig. 4. Air pollutants evolution in 2018-2021

Using Table 1, the data set extends with four factorial variables, which represent the pollutants AQI category. In the available data, all the values for NO_2, O_3, and SO_2 fall within the category "Good", while the values for PM_{10} fall within two categories: "Good" and "Moderate".

All four pollutants have both continuous and factorial values. All the continuous variables were modeled using seven different models, found below: MLR, ARMA, ARIMA, SARIMA, Decision Trees (Regression), Random Forest, and ANN. PM_{10} is the only pollutant that was modeled using classification models for its interval values: Decision Trees (Classification), Random Forest, C5.0, and ANN.

For the pollutant PM_{10}, the best ARMA, ARIMA and SARIMA models were ARMA (4,1), ARIMA (4,0,1), and SARIMA (1,0,3), which can be seen in *Table 2* and Figure 5. RMSE values for all the seven models have a large range, with the lowest RMSE being of 4.405 (corresponding to the time series models) and the highest being of 7.204 (corresponding to the ANN model). ARMA (4,1) can be selected as a good model to predict PM_{10} evolution.

Table 2. RMSE values for PM_{10} models

Pollutant	Method	RMSE
PM_{10}	MLR	5.968394
	ARMA (4,1)	**4.405115**
	ARIMA (4,0,1)	**4.405115**
	SARIMA (1,0,3)	4.409546
	Classic Decision Tree	6.147519
	Random Forest	6.207245
	ANN	7.204294

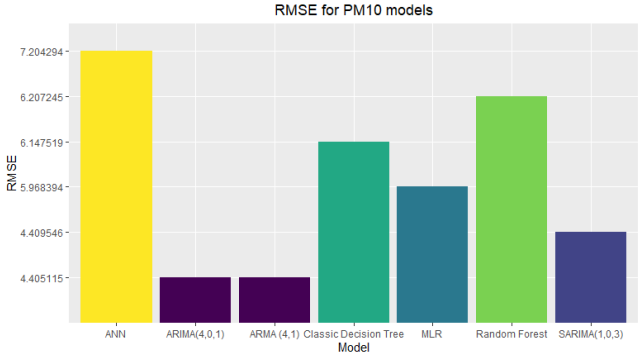

Fig. 5. RMSE for PM10 models using Romanian air quality data

For the pollutant NO_2, the best ARMA, ARIMA and SARIMA models were ARMA (0,3), ARIMA (0,1,3) and SARIMA (0,1,3), which can be seen in *Table 3* and Figure 6. RMSE values for all the seven models have a smaller range, with the lowest RMSE being of 1.523 (corresponding to the time series models) and the highest being of 2.601 (corresponding to the ANN model). ARIMA (0,1,3) can be selected as a good model to predict the pollutant NO_2 evolution.

Table 3. RMSE values for NO_2 models

Pollutant	Method	RMSE
NO_2	MLR	1.92947
	ARMA (0,3)	1.689982
	ARIMA (0,1,3)	**1.523203**
	SARIMA (0,1,3)	**1.523203**
	Classic Decision Tree	1.998766
	Random Forest	2.014126
	ANN	2.601353

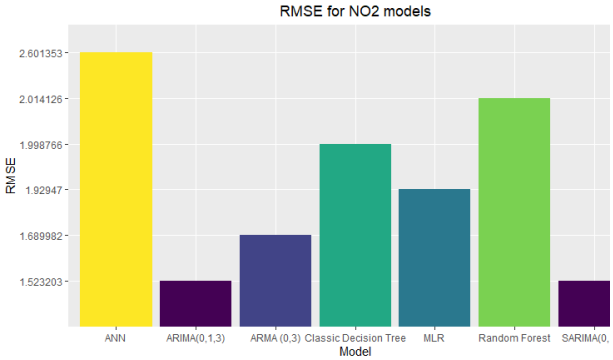

Fig. 6. RMSE for NO2 models using Romanian air quality data

For the pollutant O_3, the best ARMA, ARIMA and SARIMA models were ARMA (1,4), ARIMA (1,0,3), and SARIMA (1,0,2), which can be seen in *Table 4* and Figure 7. RMSE values for all the seven models have a higher range, with the lowest RMSE being of 2.965 (corresponding to the time series model ARMA) and the highest being of 8.444 (corresponding to the ANN model). ARMA (1,4) can be selected as a good model to predict the pollutant O_3 evolution.

Table 4. RMSE values for O_3 models

Pollutant	Method	RMSE
O_3	MLR	6.804479
	ARMA (1,4)	**2.964871**
	ARIMA (1,0,3)	2.971914
	SARIMA (1,0,2)	2.985155
	Classic Decision Tree	6.823751
	Random Forest	6.838346
	ANN	8.444393

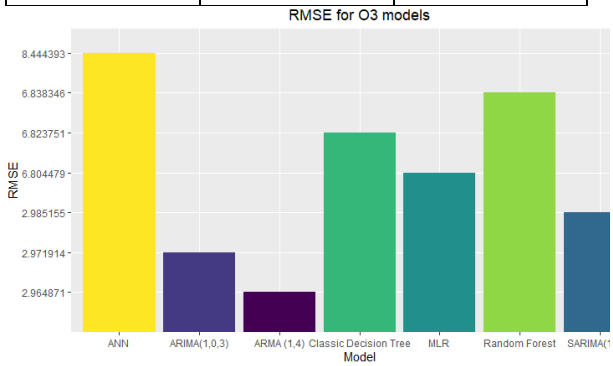

Fig. 7. RMSE for O3 models using Romanian air quality data

For the pollutant SO_2, the best ARMA, ARIMA and SARIMA models were ARMA (1,3), ARIMA (0,1,3), and SARIMA(3,1,1), which can be seen in *Table 5* and Figure 8. RMSE values for all the seven models have a very low range, with the lowest RMSE being of 0.3168 (corresponding to the time series model ARMA) and the highest being of 0.4375 (corresponding to the ANN model). ARMA(1,3) can be selected as a good model to predict the pollutant SO_2 evolution.

Table 5. RMSE values for SO_2 models

Pollutant	Method	RMSE
SO_2	MLR	0.354692
	ARMA (1,3)	**0.3168257**
	ARIMA (0,1,3)	0.3173711
	SARIMA (3,1,1)	0.317244
	Classic Decision Tree	0.3740178
	Random Forest	0.381087
	ANN	0.4375884

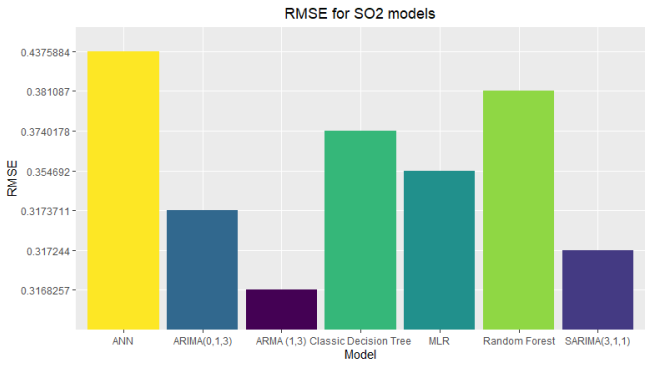

Fig. 8. RMSE for SO2 models using Romanian air quality data

For the available data, the ANN models had the highest RMSE in all four models. This means there is insufficient data to make a proper model for continuous variables using neural networks: to make better models, one would need weather information (temperature, air pressure, wind, humidity, etc.) as well.

PM_{10} is the only pollutant whose factored values can be modeled using classification models. In *Table 6*, one can see the accuracy obtained from the modelling using four different algorithms. The obtained accuracies have appropriate values, between 78.44% and 82.93%. The highest accuracy corresponds to

the ANN model, as opposed to the models constructed before, which shows that ANN is a good algorithm to predict PM_{10} interval values.

Table 6. Accuracy for PM10 models

Pollutant	Method	Accuracy
PM_{10}	Classic Decision Tree	79.64072%
	C5.0	81.13772%
	Random Forest	78.44311%
	ANN	**82.93413%**

5. Conclusions

Air pollution is a very important topic that should occupy high priority for governments around the world. Since pollution is in a continuous growth, it is mandatory to have a tool to monitor it, especially because high levels of pollution are harmful to the entire population. After monitoring air quality, it is necessary to analyze the obtained data to help with the proper-decision making to combat pollution.

Such analysis was performed in this paper, using Romania's daily air quality data ranging from 2018 to 2022. Four pollutants were used: PM_{10}, NO_2, O_3, and SO_2; $PM_{2.5}$ was missing significant amount of values, but, in the future, Romanian monitoring stations will add this important pollutant.

After analyzing multiple models, only seven regression models and four classification models qualified for this paper. All regression models had the lowest RMSE for the time series models, and the highest RMSE for the ANN model: given only the information about the four pollutants, time series model perform significantly better than other models, both linear and nonlinear. MLR models also have surprisingly lower RMSE values for the modeled pollutants, as compared to decision trees and ANN. For the classification models, however, the highest accuracy corresponds to the ANN model: factorial data behaves in an easier to model way, which makes it easier for ANN to perform better.

In conclusion, this paper covers the proposed researched topic, by presenting multiple models and computing those using Romanian data. However, this is not enough to have a very good air quality prediction, as some information is missing. To obtain the best air quality forecast, it seems to be important to have weather information in the data set. In the future, the same models should be applied for the integrated data: both air quality and weather data, to test the previously mentioned hypothesis.

Appendix

- IoT – Internet of Things – complex concept regarding a network of intelligent objects;
- $PM_{2,5}$ – fine particle with a diameter less than 2.5 μm;
- PM_{10} – fine particle with a diameter less than 10 μm;
- NO_2 – nitrogen dioxide; fatal in large quantities;
- O_3 – ozone;
- SO_2 – sulfur dioxide; toxic gas;
- AQI – Air Quality Index;
- MLR – Multiple Linear Regression;
- ARMA – Auto-Regressive Moving Average model;
- ARIMA – Auto-Regressive Integrated Moving Average model;
- SARIMA – Seasonal Auto-Regressive Integrated Moving Average model;
- ANN – Artificial Neural Network;
- RMSE – Root Mean Square Error.

References

[1] World Health Organization, "WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide," World Health Organization, Geneva, 2021.

[2] European Environment Agency, "European Air Quality Index," 2022. [Online]. Available: https://airindex.eea.europa.eu/.

[3] T. Kitchilan, M. Abeyratne and P. E. Ediriweera, "Air Quality Monitoring And Prediction Using IOT And

Machine Learning Approaches," *International Journal of Scientific and Research Publications*, vol. 12, no. 3, pp. 34-39, 2022.

[4] V. Barot and V. Kapadia, "Air Quality Monitoring Systems using IoT: A Review," *International Conference on Computational Performance Evaluation (ComPE)*, pp. 226-231, 2020.

[5] W. Huang, T. Li, J. Liu, P. Xie, S. Du and F. Teng, "An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability," *Information Fusion*, no. 75, pp. 28-40, 2021.

[6] S. M. Cabaneros, J. K. C. and B. R. Hughes, "A review of artificial neural network models for ambient air pollution prediction," *Environmental Modelling & Software*, no. 119, pp. 285-304, 2019.

[7] A. A. Adebiyi, A. O. Adewumi and a. C. K. Ayo, "Comparison of ARIMA and Artificial Neural Networks Models for Stock Price Prediction," *Journal of Applied Mathematics*, 2014.

[8] A. Alimissis, K. Philippopoulos, C. Tzanis and D. Deligiorgi, "Spatial estimation of urban air pollution with the use of artificial neural network models," *Atmospheric Environment*, 2018.

[9] G. K. Kang, J. Z. Gao, S. Chiao, S. Lu and a. G. Xie, "Air Quality Prediction: Big Data and Machine Learning Approaches," *International Journal of Environmental Science and Development*, vol. 9, no. 1, pp. 8-16, 2018.

[10] A. Y. Wang and B. T. Kong, "Air Quality Predictive Modelling Based on an Improved Decision Tree in a Weather-Smart Grid," *IEEE Access*, 2017.

[11] A. Singh, R. Kumar and N. Hasteer, "Comparative Analysis of Classification Models for Predicting Quality of Air," in *5th International Conference on Computing Communication and Automation (ICCCA)*, 2020.

[12] E. Mosadegh, K. Ashrafi, M. S. Motlagh and I. Babaeian, Modeling the Regional Effects of Climate Change on Future Urban Ozone Air Quality in Tehran, Iran, Cornell University, 2021.

[13] O. Taylan, A. S. Alkabaa, M. Alamoudi and A. Basahel, "Air Quality Modeling for Sustainable Clean Environment Using ANFIS and Machine Learning Approaches," *Atmosphere*, vol. 12, no. 713, 2021.

[14] N. S. Represa, A. Fernández-Sarría, A. Porta and J. Palomar-Vázquez, "Data Mining Paradigm in the Study of Air Quality," *Environmental Processes*, 2019.

[15] A. Panimalar, V. Shree and V. Kathrine, "The 17 V's Of Big Data," *International Research Journal of Engineering and Technology (IRJET)* , pp. 329-333, 2017.

[16] J. T. Ali, "Big data as a tool to improve air quality," 2020. [Online]. Available: https://cepei.org/en/documents/big-data-improve-air-quality/.

[17] I. E. Naqa and M. J. Murphy, "What Is Machine Learning?," in *Machine Learning in Radiation Oncology*, Springer, Cham, 2015, pp. 3-11.

[18] W. Wei, O. Ramalho, L. Malingre, S. Sivanantham, J. C. Little and C. Mandin, "Machine learning and statistical models for predicting indoor air quality," *Indoor Air*, no. 29, pp. 704-726, 2019.

[19] C. Bellinger, M. S. M. Jabbar, O. Zaïane and A. Osornio-Vargas, "A systematic review of data mining and machine learning for air pollution epidemiology," *BMC Public Health*, vol. 17, no. 907, 2017.

[20] M. Elbayoumi, N. A. Ramli, N. F. F. M. Yusof, A. S. B. Yahaya, W. A. Madhoun and A. Z. Ul-Sau,

"Multivariate methods for indoor PM10 and PM2.5 modelling in naturally ventilated schools buildings," *Atmospheric Environment*, no. 94, pp. 11-21, 2014.

NICULAE Andreea-Mihaela – student at Bucharest University of Economic Studies, attending Data Bases – Support for Business Master, Bucharest, Romania; obtained a Bachelor's Degree in Economic Cybernetics in 2020; former Erasmus+ student in Athens University of Economics and Business, attending Statistics master courses, Athens, Greece;