Journal of Environmental Engineering and Landscape Management
ISSN 1648–6897 / eISSN 1822-4199
2023 Volume 31 Issue 3: 176–185
https://doi.org/10.3846/jeelm.2023.19467

FORECASTING OF AIR POLLUTION WITH TIME SERIES AND MULTIPLE REGRESSION MODELS IN SOFIA, BULGARIA

Nikolay STOYANOV*, Antonia PANDELOVA, Tzanko GEORGIEV, Julia KALAPCHIISKA, Bozhidar DZHUDZHEV

Faculty of Automatics, Technical University of Sofia, Kliment Ohridski 8 Boulevard, 1000 Sofia, Bulgaria

Received 15 July 2022; accepted 16 May 2023

Highlights

- This study proposes an adequate analytical forecasting model of PM_{10} concentration, based on meteorological variables.
- The created time series model is based on the combination ofARIMA and Multiple Linear Regression methods.
- The adequacy requirements related to probability distribution and correlation of the residuals was established.

Abstract. Air pollution is one of the serious environmental problems. The high concentrations of particulate matter can have a serious impact over human health and ecosystems, especially in highly urbanized areas. In this regard, the present study employs a combined ARIMA-Multiple Linear Regression modelling approach for forecasting particulate matter content. The capital city of Bulgaria is used as case study. A regression analysis techniques are used to study the relationship between particulate matter concentration and basic meteorological variables – air temperature, solar radiation, wind speed, wind direction, atmospheric pressure. The adequacy of the models has been proven by examining the behavior of the residues. The synthesized time series model can be used for forecasting, monitoring and controlling the air quality conditions. All analyzes and calculations were performed with statistical software STATGRAPHICS.

Keywords: Integrated Autoregressive Moving Average (ARIMA), multiple linear regression, air pollution, PM_{10}, meteorological variables.

Introduction

In recent years, there has been a worldwide trend towards increasing the concentration of particulate matter in the air. The high levels of air pollution from various anthropogenic activities on the territory of the Republic of Bulgaria (Doncheva & Boneva, 2013) are part of this global problem (Li et al., 2021). Deterioration of air quality is most noticeable in densely populated urban areas (Stoimenova, 2016; Aarnio et al., 2016). For our country, the region of Sofia is one of the most affected by the emitted harmful substances in the air. The measured concentrations of particulate matter (PM) in the capital exceed systemically the permissible average daily norm, regulated in the Atmospheric Air Purity Act.

The air quality studying and prediction can be established by traditional and intelligent methods. The traditional methods include analytical and statistical models. The analytical (deterministic) models (Honore et al., 2008) are based on various physicochemical laws, heat-transfer and mass-transfer, etc. These are Eulerian or Lagrangian models, the main characteristics of which are high universality and wide limits of variation of conditions. Statistical models (Gocheva-Ilieva et al., 2014; Zhang et al., 2017) operate on the black box principle. They have many advantages, as high accuracy and simple computation, but are valid only in a limited range of changing conditions. Widely used in practice are combined models, representing a hybrid between deterministic and statistical models (Chaloulakou et al., 2003; Hoi et al., 2009). They combine the advantages of both types of modelling.

On the other hand, artificial intelligent-based technics are increasingly used in this field due to the many advantages, as high accuracy, efficient work with high volume of data and generalization ability. The first applications of machine learning methods are from the beginning of the 90s, as the main reason for their using is the ability to handle non-linear relationships (Roadknight et al., 1997).

*Corresponding author. E-mail: *n_stoyanov@tu-sofia.bg*

One of the mostly used method of air pollution forecasting is Artificial Neural Networks, which are based on the process in the human brain. Neural network with different structures, as Multiple Layer Perceptrone (Abderrahim et al., 2016), back propagation (Viotti et al., 2002; Kammal et al., 2006) and radial basis function (Wahid et al., 2011) have been used for prediction of wide range of pollutions and their concentration. Deep learning is a sophisticated version of neural networks and is also often used to predict of air pollution (Subbiah & Kumar, 2022). In recent years, the use of new algorithms, such as ensemble machine learning methods (Ejohwomu et al., 2022), which provide opportunities to generate more accurate and efficient forecasts, have increased.

The creation of an effective mathematical model is extremely important for the timely control of pollution sources in the areas with excessive levels of harmful substances in the air (Liping & Yaping, 2005). The statistical models are highly appropriate in many cases and can be built based on multiple linear regression (MLR), time series, Bayesian Autoregressive, etc. They provide predictive capabilities by using large arrays of numerical data.

The aim of this article is to develop a combined ARIMA-MLR model applied to air-pollution concentration of particulate matter PM_{10}. The multiple linear regression model is used for studying the relationships between the concentration of particulate matter and five basic meteorological variables. The statistical software for data processing STATGRAPHICS was used to perform the necessary analyzes and calculations.

1. Experimental area and method

1.1. Study area

Sofia is the largest city in Bulgaria and the capital of Bulgaria. It is located south of the center of the Sofia field, bordering the Stara Planina to the northeast and surrounded by the mountains Lozen, Plana, Vitosha, Lyulin to the southwest. The Sofia field is enclosed, hollow, with poor ventilation. Climatic conditions and the large number of anthropogenic sources of air pollution are the reason for the sharp increase in their concentrations in a short period of time.

Six automatic measuring stations have been installed on the territory of Sofia. The measuring station in the "Mladost" district is the only one on the territory of Sofia Municipality for which the high levels of PM_{10} pollution are formed mainly by the city traffic. "Mladost" district is located in the southeastern part of Sofia and represents about 10% of the city's territory. To the west it borders the "Darvenitsa" and "Musagenitsa" districts, to the south the Kambanite area and the Ring Road. The southeastern part of "Mladost" is about 2–3 kilometers from the "Gorublyane" district, and to the north it borders Tsarigradsko Shosse Blvd. "Druzhba" and residential area "Polygon". The investigated area is shown on Figure 1.

The mathematical experimental model was developed to reveal the relationship between PM_{10} and the five independent meteorological variables as follows: daily average air temperature T_{air}; daily average solar radiation R_{sun}; wind speed S_{wind}; wind direction D_{wind}; daily average atmospheric pressure P_{atm}. All the data of the independent and dependent variables were measured on the territory of Mladost district, Sofia for a period of one calendar year – 365 values for each variable from January to December 2017. The measurements were performed with an automatic measuring station Thermo Sharp 5030.

1.2. ARIMA-MLR method

The multiple linear regression models are based on the relationship between two or more variables to a response variable (Abdullah et al., 2017), by means of a linear equation of the following form:

$$Y_{MLR} = a_0 + \sum_{i=1}^{n} a_i X_i + \varepsilon, \qquad (1)$$

where Y_{MLR} is depended variable, X_i is the explanatory variable, a_i are regression coefficients, ε – stochastic errors due regression equation. *MLR* is an established effective method used for air pollution analysis (Ul-Saufie et al., 2011).

One of the most popular time series models for forecasting of air pollution is based by Autoregressive Integrated Moving Average (ARIMA) method. The ARIMA model includes autoregressive (AR) and moving average (MA) components, which are determine by Box-Jenkins method (Box & Jenkins, 1976). The order of a time series

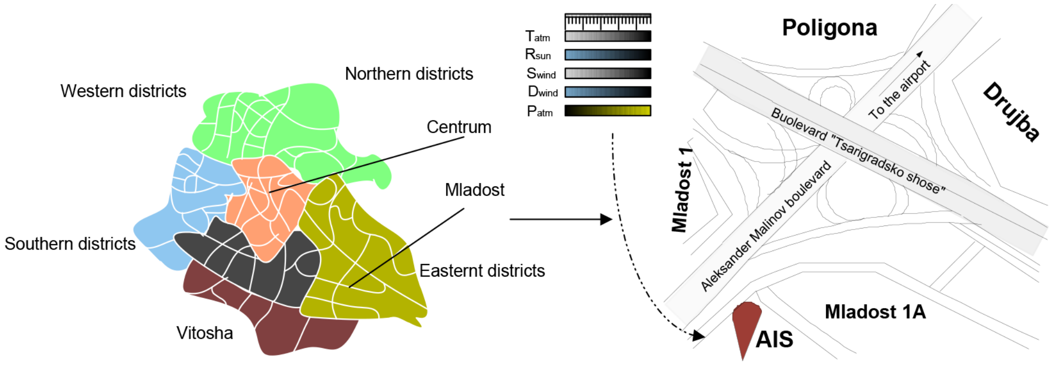


Figure 1. Investigated area – Mladost district, capital city of Sofia, Bulgaria

model is designated by following expression ARIMA (p, d, q), where p, d and q are the order of the autoregressive, the differencing and the moving-average components. The overall equation of this model can be expressed as follows (Mancini et al., 2022; Ye, 2019):

$$\Phi_P(B)(1-B)^d Y_t = Q_q(B)e_t, \tag{2}$$

where Y_t are the data at the time t, B is operator, Φ_p is the autoregressive polynomials, Q_q is the moving average polynomials, e_t is the residue referred to at the time t.

Particulate air pollution is directly related to various meteorological factors (Galindo et al., 2011). It is extremely important to carefully assess the meteorological phenomena that have the strongest impact on the process of pollutant spread in open urban areas. Combining all aspects related to climate and the spread of PM in one model is very complicated. A multi-stage calculation procedure is needed to establish the relationship between meteorological variables and particulate matter concentrations. For this reason, a combination of several mathematical methods is often used.

An interesting approach in the analysis of PM is the use of a combined method involving multiple linear regression and an ARIMA model. In this case, the purpose of the ARIMA model is to fit the residuals of the linear regression model and make a short-term forecast. The predicted residuals are superposed with values of multiple linear regression. The combined model as follows (Wei et al., 2006):

$$Y_{ARIMA-MLR} = Y_{MLR} + Y_{MLR-RES} = a_0 + \sum_{i=1}^{n} a_i X_i + \varepsilon + ARIMA(p,d,q), \tag{3}$$

where Y_{MLR} are predicted values of the multiple linear regression, $Y_{MLR\text{-}RES}$ – predicted values for the residuals of the multiple linear regression fitted for the ARIMA model.

The criteria chosen to evaluate the accuracy of the ARIMA-MLR models are RMSE (Root Mean Square Error), MAE (Mean Absolute Error), ME (Mean Error), MPE (Mean Percentage Error).

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\left(y_t - \widehat{y_t}\right)^2}{n}}; \quad MAE = \frac{1}{n}\sum_{t=1}^{n}\left|y_t - \widehat{y_t}\right|;$$

$$ME = \frac{1}{n}\sum_{t=1}^{n}(y_t - \widehat{y_t}); \quad MPE = \frac{100}{n}\sum_{t=1}^{n}\frac{y_t - \widehat{y_t}}{y_t}. \tag{4}$$

2. Results and discussion

2.1. Synthesis of a mathematical model, based on meteorological variables

The task of developing a model includes establishing of relationships between variables, distribution tests, regression analysis and estimation of residual distribution. For simplicity and convenience, the following notations of the considered variables are accepted: C – concentration of particulate matter PM_{10}; T – daily average air temperature T_{air}; R – daily average solar radiation R_{sun}; S – wind speed S_{wind}; D – wind direction D_{wind}; P – daily average atmospheric pressure P_{atm}. Data in digital form of the listed variables for a period of one calendar year were used to conduct the research, related to the creation of a mathematical model. In Table 1 are given the obtained results from the descriptive statistic of the input data.

The input data includes 2199 values and is characterized by the absence of missing data, one outlier value. Figure 2 present time sequence plot for the data of PM_{10}. Clear peaks are observed in winter, while in summer the levels are relatively lower.

From Figure 2 and the data in table, it can be concluded that the concentrations of dust particles exceed 50 μg/m^3(prescribed threshold values of PM_{10}) for certain periods of the year and represent a problem that needs special attention.

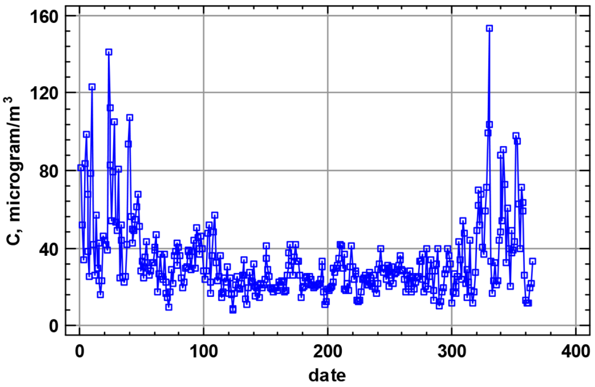


Figure 2. Time sequence plot of the observed data for concentration of PM_{10}

Table 1. Descriptive statistic of the initial data

Statistic	T, °C	R, W/m^{-2}	S, m/s^{-1}	D, degree	P, mbar	C, μg/m^{-3}
Mean	11.73	166.092	1.50786	147.992	922.723	33.671
Median	11.65	159.6	1.425	144.85	923.0	27.78
Minimum	–12.8	8.7	0.57	51.34	912.0	7.89
Maximum	27.8	379.5	3.15	261.99	934.0	153.76
Stand. Dev.	8.6435	105.597	0.419511	52.6419	3.83365	20.7459
Skewness	–0.3091	0.301092	0.982012	0.0872178	0.0479695	2.36903
Kurtosis	–0.6394	–1.24722	1.07464	–1.04098	0.503928	7.30544

2.2. Correlation analysis

The first stage in the search and synthesis of a mathematical model is the correlation analysis, which checks for the presence and direction of stochastic relationships between the variables. A study to detect the relationship between each of the six variables with the other five was performed, the result of which is presented in Figure 3.

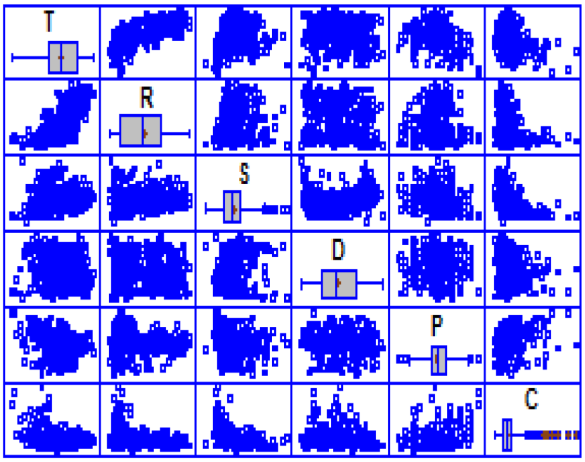


Figure 3. Correlation field and distributions of variables

The Box-and-Whisker diagram, located on the diagonal of the figure, shows a graphical interpretation of the probability distribution of the six variables and how close it is to the normal. The middle vertical line in each box represents the median of the distribution, and the so-called tails are observed on the both sides. The correlation fields below and above the diagonal represent the data between each of two variables.

The correlation of the Spearman rank for the six variables is given in Table 2. The table shows the correlations between each pair of variables. The correlation coefficients vary between –1 and +1 and show the strength of the relationship between the variables. The second number represents the number of data for each variable. The third number for the individual pairs of quantities in the table is *p*-value. It determines the statistical significance of the calculated correlations. *P*-value below 0.05 shows statistically significant non-zero correlations at a probability of 95.0%.

2.3. Investigation of the probability distributions of the variables

The study of the relationships between quantitative variables was carried out using linear regression analysis, which is one of the methods for finding dependencies, which at a later stage are used for model synthesis and forecasting. The presence of a correlation between the concentration of particulate matter and the other five variables determines the use of multiple linear regression. A needful condition for the validity of this method is the normal distribution of all variables. In the absence of sufficient coverage of the analyzed distributions with the normal one for starting the calculations with the regression model, a fixed group of other distributions are allowed – Weibull, Half Normal, Beta and others.

If the data are not described by any of the allowable distributions, an appropriate mathematical transformation is performed. The transformation process continues until an acceptable distribution is found.

The verification of distributions is performed with standardized tests. The degree of proximity of the probability distributions of the variables or of the additionally performed transformations of the data to the normal distribution is checked by the performed tests for normality – χ^2 and Shapiro-Wilk. For all other distributions the universal χ^2 is used. All tests calculate the *p*-value, from which information for the statistical significance of the results is obtained. If a *p*-value greater than 0.05 is obtained, the corresponding distribution with a 95% probability can be accepted.

After analyzing the data for each variable, it was found that the distributions of each variable differ from the normal. After many transformations and checks, suitable new variables were found, satisfying the conditions of the tests.

Table 2. Correlations of Spearman rank

	T, °C	R, W/m^{-2}	S, m/s^{-1}	D, degree	P, mbar	C, µg/m^{-3}
T, °C		0.8051	0.1912	–0.2290	–0.2854	–0.2728
p-value		0.0000	0.0003	0.0000	0.0000	0.0000
R, W/m^{-2}	0.8051		0.1251	–0.1633	–0.0859	–0.3095
p-value	0.0000		0.0171	0.0019	0.1018	0.0000
S, m/s^{-1}	0.1912	0.1251		0.2321	–0.2375	–0.4906
p-value	0.0003	0.0171		0.0000	0.0000	0.0000
D, degree	–0.2290	–0.1633	0.2321		–0.1139	–0.2058
p-value	0.0000	0.0019	0.0000		0.0301	0.0001
P, mbar	–0.2854	–0.0859	–0.2375	–0.1139		0.2240
p-value	0.0000	0.1018	0.0000	0.0301		0.0000
C, µg/m^{-3}	–0.2728	–0.3095	–0.4906	–0.2058	0.2240	
p-value	0.0000	0.0000	0.0000	0.0001	0.0000	

Below are the results and histograms for all variables.

Check of the dependent variable C. The range of values of C is from 7.89 to 153.76 μg/m^3. The variable cannot be approximated by a normal distribution. After a series of studies, an appropriate transformation of the sample was found by forming a new variable Log(Log(C)), which meets the conditions. Studies for distribution statistics were performed, as a result of which the following values were obtained: χ^2 = 48.2603 with p-value = 0.101773, Shapiro-Wilk = 0.982001 with p-value = 0.284173. The histogram of the variable Log(Log(C)) and the function are shown in Figure 4a.

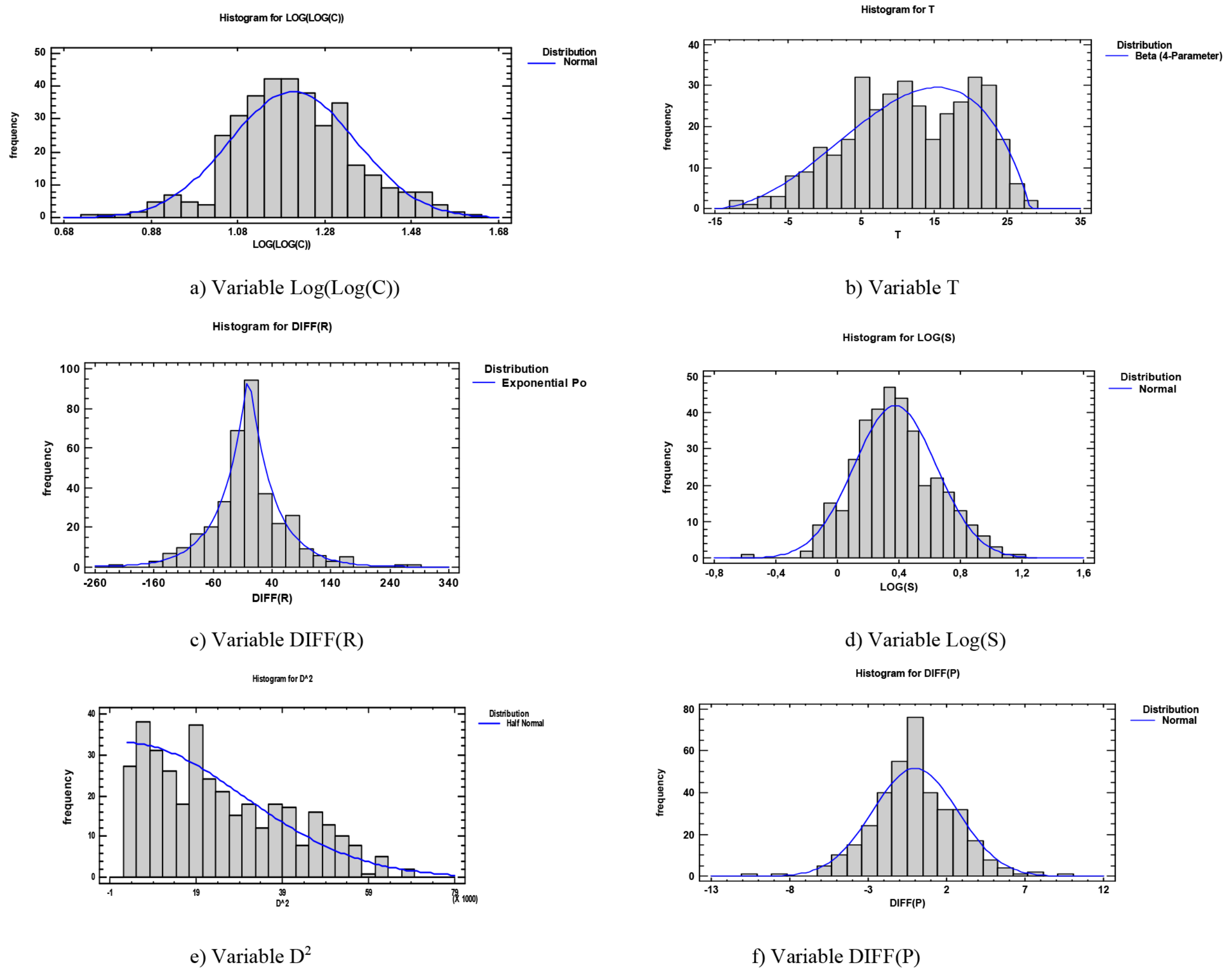
Check of the independent variable T. The values of T vary from −12.8 to +44 °C. The Beta (4-Parameter) distribution meets the set conditions. The results from χ^2 test are: χ^2 = 20.7373 with p-value = 0.145434. The histogram of T is shown in Figure 4b.

Check of the independent variable R. The probability distribution of R again differs from the normal. The data for approximation of the other admissible distributions were also checked. The results of these tests were unsatisfactory. A large number of newly created variables for possible approximation with the admissible distributions have been studied. After tests and checks, it was found that the variable DIFF(R) can be approximated by an Exponential Power distribution. The histogram and the function are shown in Figure 4c. The results obtained from the χ^2 test are: χ^2 = 47.4286 with p-value = 0.0963.

Check of the independent variable S. The range of values for S is 0.57÷3.15 m/s. From many tested transformations, a suitable one was found that fulfills the conditions for approximation with normal distribution. The new transformed variable is Log(S). The histogram of the transformed variable Log(S) and the function are presented in Figure 4d. The results of standard tests for normality are: χ^2 = 46.9452 with p-value = 0.126677, Shapiro-Wilk = 0.984525 with p-value = 0.526787.

Check of the independent variable D. The daily average values of the wind direction D measured in degrees clockwise from north are from 51.34° to 261.99°. The selected new variable is D^2, and its distribution can be approximated with the allowable Half Normal distribution. The performed universal test χ^2 over the data gives the following values: χ^2 = 46.9452 with p-value = 0.126677. The variable D^2 is presented in Figure 4e.

Check of the independent variable P. The values of P are in range 912÷934 mbar. As a result of studies with multiple transformations, a variable DIFF(P) – finite difference was found, which can be approximated with a normal distribution. The histogram of the variable



a) Variable Log(Log(C))

b) Variable T

c) Variable DIFF(R)

d) Variable Log(S)

e) Variable D^2

f) Variable DIFF(P)

Figure 4. Histograms of the meteorological variables

DIFF(P) and the function are presented in Figure 4f. The following values were obtained after performing the tests: χ^2 = 19.9232 (universal) with *p*-value = 0.0685532, Shapiro-Wilk = 0.980394 with *p*-value = 0.168024.

After the performed tests and the found distributions for all variables, a synthesis of a regression model can be started.

2.4. Sythesis of a multiple linear regression model

A multiple linear regression model has been developed to reveal the relationship between PM_{10} and the five independent variables. The functional dependence between the dependent quantity and the predictors (independent variables) is determined by the method of least squares. The method of multiple linear regression was used due to the combined influence of plurality variables on the concentration of particulate matter. The model was developed after a series of studies, as a result of which a function satisfying the requirements for adequacy was found. The equation of the synthesized multiple linear regression model describing the relationship between PM_{10} and the five meteorological variables is as follows:

$$\mathrm{Log}(C) = 1.18604\left(\frac{1}{S}\right) + 0.01429\left(\frac{1}{\exp\left(\frac{1}{T}\right)^{2}}\right) - \quad (5)$$

$$0.00037(\log(R))^{4} - 0.1698\log(D) + 0.000004P^{2}.$$

The obtained coefficient R^2 indicated that the model describes 98.68% of the data. The resulting standard error, showing the standard deviation of the residues is 0.394769, which can be used to build prediction limits for new observations. The mean absolute error is 0.301462 and represents the mean residual value. Detailed statistics of the coefficients in the obtained linear regression model is presented in Table 3. The *p*-values of all independent variables are less than 0.05, that indicating they are statistically significant at the 95% confidence interval. The graph showing the observed versus predicted values of PM_{10} is shown in Figure 5.

2.5. Investigation of the behavior of the residues

In order to evaluate the developed model, it is mandatory to study the behavior of the residues. This is done with the help of standard statistical tests. The model is considered adequate and suitable for practical use only when an analysis of the residues has been made. The compulsory requirement is that they are normally distributed. The hypothesis for normality of the distribution of the residues was tested.

In Table 4 and Table 5 the results of performed standard tests for normality of the residues are given. The high values of *p*-value > 0.05 for the χ^2 and Shapiro-Wilk tests confirm the validity of the hypothesis for normal residue distribution at 95% confidence level. The results of the two Kolmogorov-Smirnov tests are in agreement with the results of the previous two tests (χ^2, Shapiro-Wilk).

Table 3. Model coefficients assessed by MLR

Variable	Estimate	Std. error	*t*-statistic	*p*-value	Standardized coefficients Beta
1/S	1.18604	0.114625	10.3471	0.0000	0.253
$\mathrm{Log(R)}^4$	−0.000372	0.0000616	−6.02973	0.0000	−0.80
P^2	0.0000042	3.53054E-7	11.9606	0.0000	1.051
Log(D)	−0.169795	0.053489	−3.17439	0.0016	−0.241
$1/(\exp(1/T)^2$	0.0142885	0.006846	2.08712	0.0376	0.014

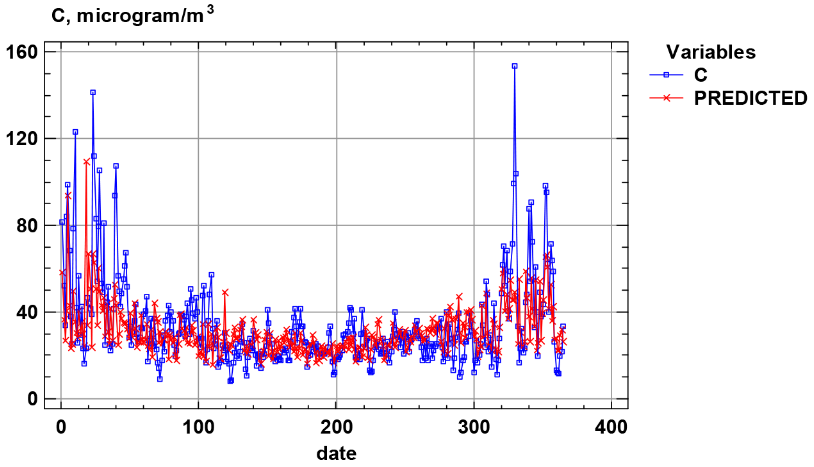


Figure 5. Plot of predicted PM_{10} against observed PM_{10} concentrations

Table 4. Residue normality test

Test	Statistic	*p*-value
Chi-Square	37.0822	0.465284
Shapiro-Wilk W	0.98665	0.733176
Skewness Z-score	1.03669	0.299878
Kurtosis Z-score	2.86051	0.004229

Table 5. Kolmogorov-Smirnov tests

Normal	Values
DPLUS	0.030511
DMINUS	0.048488
DN	0.048488
p-value	0.360509

Figure 6 shows the results of a Q-Q test (Quantile-Quantile test) of the model residues. It can be clearly seen from the figure that the residual values closely follow the line determined by the normal distribution. A slight deviation of the points is observed only in the initial section.

The performed tests give a reason for accepting the hypothesis for normal distribution of residues with a 95% level of probability.

2.6. Building of ARIMA forecasting model with regression

The aim is to build and explore ARIMA-MLR model in order to determine the relationship between the PM_{10}-values and the meteorological variables. This model depends on time and includes additionally five transfer functions. The combined model synthesis procedure in the statistical package STATGRAPHICS requires time series to be constructed for each transfer function. In order to realize a short-term forecast for PM_{10}, it is necessary to make short-term forecasts for all ARIMA models of the meteorological variables.

To achieve this goal, separate time series for each of the transformed meteorological variableswere created. The summary statistics of the ARIMA models for Tair, Rsun, Swind, Dwind and Patm are presented in Table 6.

After implementation of the short-term forecast the actual predicted values for all transformed meteorological variables were determined. These predictions after conversion to the basic variables supplemented the remaining values and formed the necessary set to find a combined ARIMA-MLR model.

Based on obtained Multiple Regression Model (5) for the analyzed 365 values of PM_{10} concentrations and the five models for transformed meteorological variables, a model of the ARIMA(p, d, q) – type is being searched to satisfy various conditions. These conditions are the smallest values of RMSE, MAE, ME, MPE, minimum of the SBIC (Schwarz Bayesian Information Criterion), normal distribution of the residues and the requirements, related to the autocorrelation and partial autocorrelation functions.

In order to obtain the best fitting many ARIMA models with different parameters were tested. Block-diagram of creating of a ARIMA-MLR model is shown in Figure 7.

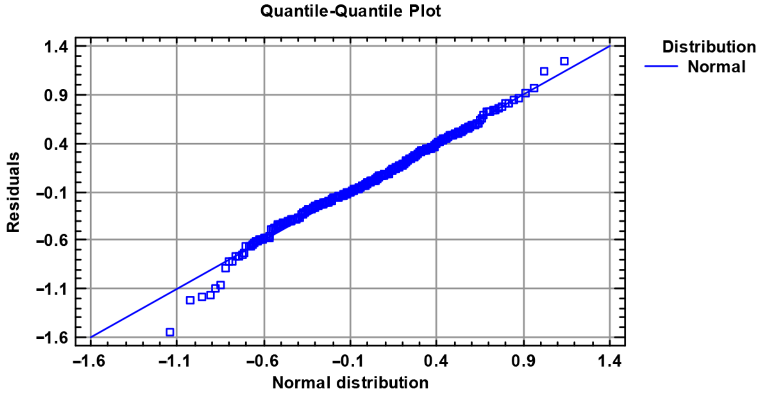


Figure 6. Curve for normality of the distribution of residues

Table 6. Statistics of ARIMA-models for the transformed meteorological variables

Meteorological variable	RMSE	MAE	ME	Ljung-Box test	Type
T	2.33094	1.78373	0.0733489	0.250659	ARIMA(4, 0, 3)
Diff(R)	52.5178	38.3049	–0.0478155	0.469194	ARIMA(6, 0, 6)
Log(S)	0.256381	0.198433	0.000397	0.694541	ARIMA(1, 0, 0)
D^2	14803.2	12006.8	29.9663	0.931213	ARIMA(3, 0, 5)
Diff(P)	2.3386	1.69874	–0.0354337	0.31978	ARIMA((1, 0, 2)

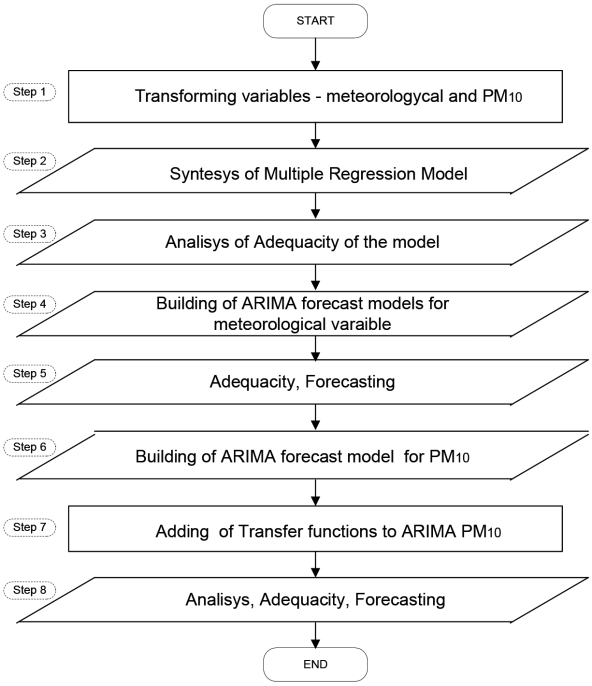

START
Step 1
Transforming variables - meteorologycal and PM10
Step 2
Syntesys of Multiple Regression Model
Step 3
Analisys of Adequacity of the model
Step 4
Building of ARIMA forecast models for meteorological varaible
Step 5
Adequacity, Forecasting
Step 6
Building of ARIMA forecast model for PM10
Step 7
Adding of Transfer functions to ARIMA PM10
Step 8
Analisys, Adequacity, Forecasting
END

Figure 7. Block-diagram of building of ARIMA-MLR model

The best adequacy was achieved for a model with following parameters – ARIMA(3, 0, 6). In this model the five meteorological variables T, R, S, D and Pwere used as transfer functions (STATGRAPHICS uses the term regressors). The general equation of the combined model $Y_{\text{ARIMA-MLR}}$ is as follows:

$$\mathrm{Log}(C) = ARIMA(3,0,6) + 1.18604\left(\frac{1}{S}\right) + 0.01429\left(\frac{1}{\exp\left(\frac{1}{T}\right)^{2}}\right) - 0.00037(\log(R))^{4} - 0.1698\log(D) + 0.000004P^{2}. \tag{6}$$

The basic statistics of the combined model are given in Table 7. The time sequence plot is shown in Figure 8a.

Table 7. The RMSE, MAE, ME and MPE values of the ARMA-MLR model

Combined model – $Y_{\text{ARIMA-MLR}}$	Statistic	Estimation
ARIMA(3, 0, 6) + Five Transfer Functions	RMSE	0.314154
	MAE	0.236558
	ME	–0.006008
	MPE	–1.12703
	SBIC	–2.26024

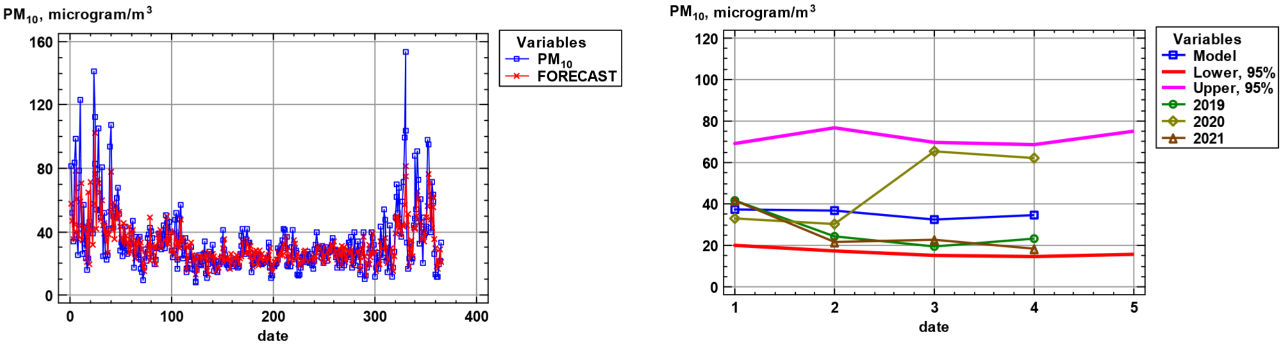
The autoregressive component AR of this model is $p = 3$, which shows that the level of pollution depends on the values for the previous 3 days. The MA component, characterizing moving average process indicate that the local stochastic changes are influences by 6 previous members of the time series. The values of RMSE, MAE, ME and MPE are low and satisfy requirements of high accuracy of the presented model.

A diagnostic procedure for the mathematical model, which includes analysis of the distribution of the residues, testing the autocorrelation (ACF) and partial autocorrelation (PACF) functions was also performed.

The results of the distribution check of the residuals are shown in Table 8. The Ljung-Box test, usually used in a ARIMA method checks of no remaining significant autocorrelation in the residuals of the model. The obtained p-value (0.5362) indicated that the autocorrelations are very small. The p-value (0.2014) of the standard Kolmogorov-Smirnov test for checking the normality of the distribution of residues is high, so it can be accepted, that they have normal distribution with a 95% confidence interval.

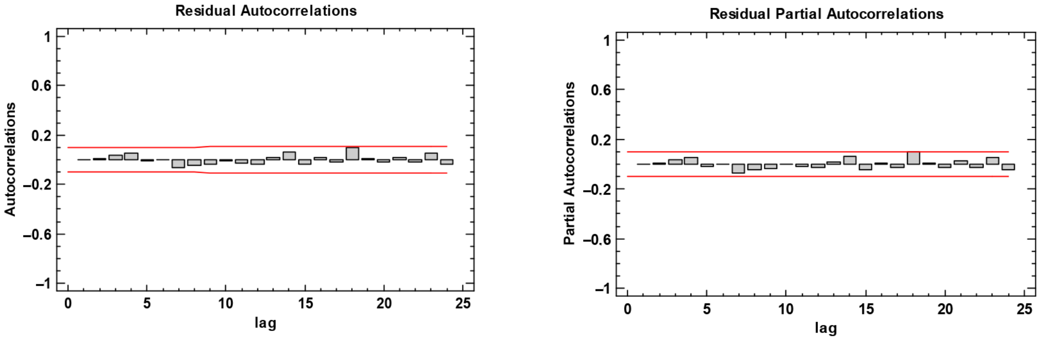
Table 8. Tests of the residuals

Tests	Test statistic	p-value
Ljung-Box	13.8593	0.5362
Kolmogorov-Smirnov	DPLUS – 0.04042; DMINUS – 0.05609 DN – 0.05609	0.2014



a) b)

Figure 8. a) Time sequence plots – ARIMA-MLR and actual values; b) Graph with 4-predicted values and actual values from 1 to 4 January 2019, 2020, 2021



a) b)

Figure 9. Graph of ACF (a) and PACF (b) of the Residuals for ARIMA (3, 0, 6)

The form of the autocorrelation ACF and partial autocorrelation PACF functions for ARIMA (3, 0, 6) model are shown in Figure 9a and Figure 9b. They show that the individual autocorrelations are very small and fall within the confidence interval. The obtained values from the various tests and the analysis of ACF and PACF show that the residues have a white noise character.

By using a developed ARIMA-MLR model a short-term forecast for 4 days and predicted values was compared with actual data for several following years was made. Figure 8b shows the result of comparison between predicted values and actual data of PM_{10} for 2019, 2020 and 2021 at 95% confidence interval.

All simulations and analyzes made with the found ARIMA-MLR model fall into required confidence intervals and demonstrate very good performance of the proposed model.

Conclusions

The forecasting of air pollution has become one of the essential areas of investigation in recent years. In the presented study a combined Auto Regressive Integrated Moving Average (ARIMA) – Multiple Linear Regression (MLR) model was applied to study and prediction of PM_{10} concentrations in Mladost district of Sofia, Bulgaria.

A multiple linear regression model was developed to reveal a causal relationship between PM_{10} and five variables (temperature, solar radiation, wind speed, wind direction, atmospheric pressure). A correlation analysis was performed for the presence of possible stochastic relationships between the particulate matter content and five meteorological variables. Using a correlation of Spearman's rank, the relationship between each of the six variables and the other five was found. Appropriate transformations of the data were determined according to the tests χ^2, Shapiro-Wilk and Kolmogorov-Smirnov for which admissible distributions with 95% probability were found. The obtained high value of the statistics shows a very strong causal relationship between the concentration of particulate matter and the independent variables. The adequacy of the developed regression model was established by analysis of the residues. It showed that their distribution can be approximated with a 95% probability to normal and constant dispersion.

The combined ARIMA-MLR model was built to assess the relationship between PM_{10} and five meteorological variables. The obtained analytical functions for combined model are used to build short-term predictions (4-days) of particulate matter content. The analysis of the residuals, as well as the conducted comparison between the predicted data from the model and actual data of PM_{10} for three following years, shows very good adequacy and predictive characteristics of the proposed model.

Acknowledgements

The author/s would like to thank the Research and Development Sector at the Technical University of Sofia for the financial support.

References

Aarnio, M. A., Kukkonen, J., Kangas, L., Kauhaniemi, M., Kousa, A., Hendriks, C., Yli-Tuomi, T., Lanki, T., Hoek, G., Brunekreef, B., Elolähde, T., & Karppinen, A. (2016). Modelling of particulate matter concentrations and source contributions in the Helsinki Metropolitan Area in 2008 and 2010. *Boreal Environment Research*, *21*(5–6), 445–460.

Abderrahim, H., Chellali, M. R., & Hamou, A. (2016). Forecasting PM_{10} in Algiers: Efficacy of multilayer perceptron networks. *Environmental Science and Pollution Research*, *23*, 1634–1641. https://doi.org/10.1007/s11356-015-5406-6

Abdullah, S., Ismail, M., & Fong, S. Y. (2017). Multiple linear regression (MLR) models for long term PM_{10} concentration forecasting during different monsoon seasons. *Journal of Sustainability Science and Management*, *12*(1), 60–69.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (Rev. ed.). Holden-Day.

Chaloulakou, A., Grivas, G., & Spyrellis, N. (2003). Neural network and multiple regression models for PM_{10} prediction in Athens: A comparative assessment. *Journal of Air & Waste Management Association*, *53*, 1183–1190. https://doi.org/10.1080/10473289.2003.10466276

Doncheva, M., & Boneva, G. (2013). Particulate matter air pollution in urban areas in Bulgaria. *Journal of Environmental Protection and Ecology*, *14*(2), 422–429.

Ejohwomu, O. A., Oshodi, O. S., Oladokun, M., Bukove, O. T., Emekwuru, N., Sotunbo, A., & Adenuga, O. (2022). Modelling and forecasting temporal $PM_{2.5}$ concentration using ensemble machine learning methods. *Buildings*, *12*(1), 46. https://doi.org/10.3390/buildings12010046

Galindo, N., Varea, M., Gil-Moltó, J., Yubero, E., & Nicolás, J. (2011). The influence of meteorology on particulate matter concentrations at an urban mediterranean location. *Water Air and Soil Pollution*, *215*, 365–372. https://doi.org/10.1007/s11270-010-0484-z

Gocheva-Ilieva, S. G., Ivanov, A. V., Voynikova, D. S., & Boyadzhiev, D. T. (2014). Time series analysis and forecasting of air pollution in a small urban area: SARIMA approach and factor analysis. *Stochastic Environmental Research and Risk Assessment*, *28*(4), 1045–1060. https://doi.org/10.1007/s00477-013-0800-4

Hoi, K. I., Yuen, K. V., & Mok, K. M. (2009). Prediction of daily averaged PM_{10} concentrations by statistical time-varying model. *Atmospheric Environment*, *43*(16), 2579–2581. https://doi.org/10.1016/j.atmosenv.2009.02.020

Honore, C., Rouıl, L., Vautard, R., Beekmann, M., Bessagnet, B., Dufour, A., Elichegaray, C., Flaud, J.-M., Malherbe, L., Meleux, F., Menut, L., Martin, D., Peuch, A., Peuch, V.-H., & Poisson, N. (2008). Predictability of European air quality: Assessment of 3 years of operational forecasts and analyses by the PREV'AIR system. *Journal of Geophysical Research*, *113*, D04301. https://doi.org/10.1029/2007JD008761

Kammal, M., Jailani, R., & Shauri, R. L. A. (2006). Prediction of ambient air quality based on neural network technique. In *4th Student Conference on Research and Development* (pp. 115–119). IEEE. https://doi.org/10.1109/SCORED.2006.4339321

Li, X., Hussain, S. A., Sobri, S., & Md Said, M. S. (2021). Overviewing the air quality models on air pollution in Sichuan Basin, China. *Chemosphere*, *271*, 129502. https://doi.org/10.1016/j.chemosphere.2020.129502

Liping, X., & Yaping, S. (2005). Modelling of traffic flow and air pollution emission with application to Hong Kong Island. *Environmental Modelling & Software*, *20*(9), 1175–1188. https://doi.org/10.1016/j.envsoft.2004.08.003

Mancini, S., Francavilla, A., Graziuso, G., & Guarnaccia, C. (2022). An application of ARIMA modelling to air pollution concentrations during covid pandemic in Italy. *Journal of Physics: Conference Series*, *2162*, 012009. https://doi.org/10.1088/1742-6596/2162/1/012009

Roadknight, C. M., Balls, G. R., Mills, G. E., & Palmer-Brown, D. (1997). Modeling complex environmental data. *IEEE Transactions Neural Network*, *8*(4), 852–862. https://doi.org/10.1109/72.595883

Stoimenova, M. P. (2016). Stochastic modeling of problematic air pollution with particulate matter in the city of Pernik, Bulgaria. *Ecologia Balkanica*, *8*(2), 33–41.

Subbiah, S. S., & Kumar, S. (2022). Deep learning based load forecasting with decomposition and feature selection technics. *Journal of Scientific & Industrial Research*, *81*, 505–517. https://doi.org/10.56042/jsir.v81i05.56794

Ul-Saufie, A. Z., Yahya, A. S., & Ramli, N. A. (2011). Improving multiple linear regression model using principal component analysis for predicting PM_{10} concentration in Seberang Prai, Pulau Pinang. *International Journal of Environmental Sciences*, *2*(2), 403.

Viotti, P., Liuti, G., & Di Genova, P. (2002). Atmospheric urban pollution: Applications of an artificial neural network (ANN) to the city of Perugia. *Ecological Modelling*, *148*(1), 27–46. https://doi.org/10.1016/S0304-3800(01)00434-3

Wahid, H., Ha, Q. P., & Duc, H. N. (2011). Computational intelligence estimation of natural background ozone level and its distribution for air quality modelling and emission control. In *Proceedings of 28th International Symposium on Automation and Robotics in Construction* (pp. 1157–1163), Seoul, Korea. https://doi.org/10.22260/ISARC2011/0212

Wei, P., Xie, S., Huang, L., Zhu, G., Tang, Y., & Zhang, Y. (2006). Prediction of $PM_{2.5}$ concentration in Guangxi region, China based on MLR-ARIMA. *Journal of Physics: Conference Series*, *2006*, 012023. https://doi.org/10.1088/1742-6596/2006/1/012023

Ye, Z. (2019). Air pollutants prediction in Shenzhen based on ARIMA and Prophet method. *E3S Web of Conferences*, *136*, 05001. https://doi.org/10.1051/e3sconf/201913605001

Zhang, H., Zhang, S., Wang, P., Qin, Y., & Wang, H. (2017). Forecasting of particulate matter time series using wavelet analysis and wavelet-ARMA/ARIMA model in Taiyuan, China. *Journal of the Air & Waste Management Association*, *67*(7), 776–788. https://doi.org/10.1080/10962247.2017.1292968