Statistical modeling of particle mater air pollutants in the city of Ruse, Bulgaria

Evelina Veleva[2] and *Ivanka* Zheleva[1,*]

[1] Department Applied Mathematics and Statistics, Ruse University, 8 Studentska Str., 7017, Ruse, BULGARIA, www.uni-ruse.bg
[2] Department Thermo engineering, Hydraulics, Ecology, Ruse University, 8 Studentska Str., 7017, Ruse, BULGARIA, www.uni-ruse.bg

Abstract This paper is devoted to examine the PM10 and PM2.5 pollution in Ruse region, Bulgaria. It is a continuation of our previous work [1] where we presented a statistical analysis and modelling of the level of PM10 air pollution in Ruse using data from one of two monitoring stations (station 2) in the city. Now in this paper we present statistical analysis of the level of PM10 pollutant on the basis of data from the another monitoring station (station 1). The measurements cover the period since 2015 up to now. The results from analysis and modelling of PM2.5 air pollutant are also presented and commented in the paper**.**

1 Introduction

Ruse is a city in the north-eastern part of Bulgaria on the Danube river coast, which is the north border between Bulgaria and Romania. The climate of the region of Ruse is temperate continental, characterized by cold winters and dry, warm summers. Spring and autumn are short. Recently it becomes clear that the mean values of the temperature in Ruse region are slightly goes up for the last 40 years and they are bigger than the mean temperature for Bulgaria [6]. Air pollution especially by particle matter (PM10 and PM2.5), which in Ruse region is going up recently, maybe affects temperature, atmospheric pressure and humidity in the region. All this in our opinion may be a reason for the increase in average temperatures nowadays for Ruse region. That is why studies on the air pollution by particle matter is needed for better understanding of these processes and for taking adequate measures to protect environment.

2 Material and methods

In the town of Ruse, Bulgaria the regular measurement of air pollution starts on 24.10.2015 [2, 3]. It is carried out by an automatic measuring station Vazrazhdane (station 1) and by a stationary station in the Zdravets - east district of the town (station 2). Since 30.09.2016 measurements are performed also by a mobile automatic station. The two stationary stations

* Corresponding author: izheleva@uni-ruse.bg

are located about 2.20 kilometers apart (the distance between them is measured by the Google Earth program) (Fig.1).

A map of the location of the stations is published on the site [3] of the Regional Inspectorate of Environment and Water in Ruse, which shows that each one covers an area with a radius of about 1,7 km.

For this study we use official data for particle matter air pollutant PM10 from stations 1 and 2 and for PM2.5 – from station 1 [3].

We use SPSS software to analyze and model data from the measurement stations in Ruse.

Fig 2 . Locations of the atmospheric air monitoring stations in the town of Ruse, Bulgaria [3]

3 Results and discussions

Station 1 monitors 8 pollutants: sulfur dioxide (SO_2), nitrogen dioxide (NO_2), nitrogen oxide (NO), carbon monoxide (CO), ground level ozone (O_3), particulate matter less than 10 micrometers in diameter (PM10), benzene (C_6H_6) and particulate matter less than 2.5 micrometers in diameter (PM2.5). Their main descriptive statistics for the period 24.10.2015 - 31.07.2017 are given in Table 1 below.

Table 1. Descriptive statistics for pollutants monitored at Station 1

Polutant	N	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
SO_2	277	5.60	22.20	8.83	2.46	2.33	6.54
NO_2	609	0.00	65.40	21.78	9.95	1.42	2.66
NO	609	0.00	211.90	9.79	15.35	6.10	58.32
CO	609	0.00	3.50	0.48	0.35	2.75	12.31
O_3	274	7.70	94.20	42.92	18.44	0.22	-0.46
PM10	584	12.00	176.00	43.10	25.75	2.03	5.02
C_6H_6	605	0.00	12,90	0.92	1.01	4.90	42.41
PM2.5	604	3.90	161,10	26.20	22.04	2.55	7.74

Above the norm are only particulate matter PM10 and PM2.5. The norm for PM10 is: average annual rate below 40 (μg/m^3) and average daily rate below 50 (μg/m^3). For year 2016 the average annual rate for PM10 is 41.17 (μg/m^3) and for the one-year period 01.08.2016 – 31.07.2017 it is 41.82 (μg/m^3) – both above the average annual norm. 146 of 584 valid observations on PM10 are above the average daily norm 50 (μg/m^3), i.e. 25% of all monitored days. The maximum value 176.00 (μg/m^3) exceeds 3.52 times the average daily norm 50 (μg/m^3).

The norm for PM2.5 is average annual rate under 25(μg/m^3). The average annual rate of PM2.5 for 2016 is 23.52, i.e. within the norm. For the one-year period 01.08.2016 – 31.07.2017 that includes the last 365 days of the sample it is 26.05, i.e. above the norm.

Four of the values of PM10 and two of the measurements of PM2.5 in august 2016 are suspiciously recorded as zero. We treated them as missing as they probably are.

As it can be seen from Figure 2, the values of PM10, measured in Station 1 (blue line), Station 2 (red line) and PM2.5 (green line) in Ruse, are highly and positively correlated. The Pearson product-moment correlation coefficients for the three variables: PM10 at Station 1 (PM10_s1), PM2.5 (measured only at Station 1, PM2_5) and PM10 at Station 2 (PM10_s2) are given in Table 2. All are statistically significant ($p < 0.0005$).

Table 2. Correlations between the three factors out of norm in Ruse

	PM10_s1	PM2_5	PM10_s2
PM10_s1	1	0.815	0.773
PM2_5	0.815	1	0.813
PM10_s2	0.773	0.813	1

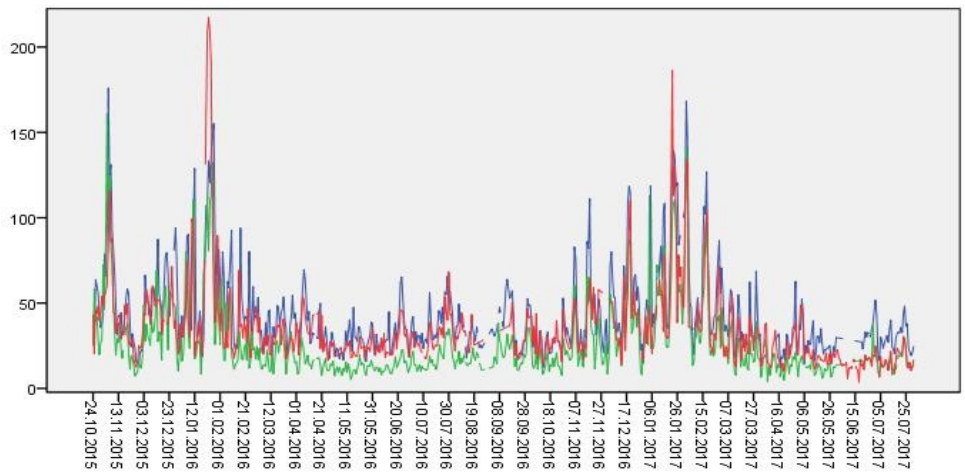

Fig. 2 Values of PM10 at Station 1 (blue line), Station 2 (red line) and PM2.5 (green line)

The distributions of PM10 at Stations 1 and PM2.5 are probably unimodal as can be seen in Figure 3.

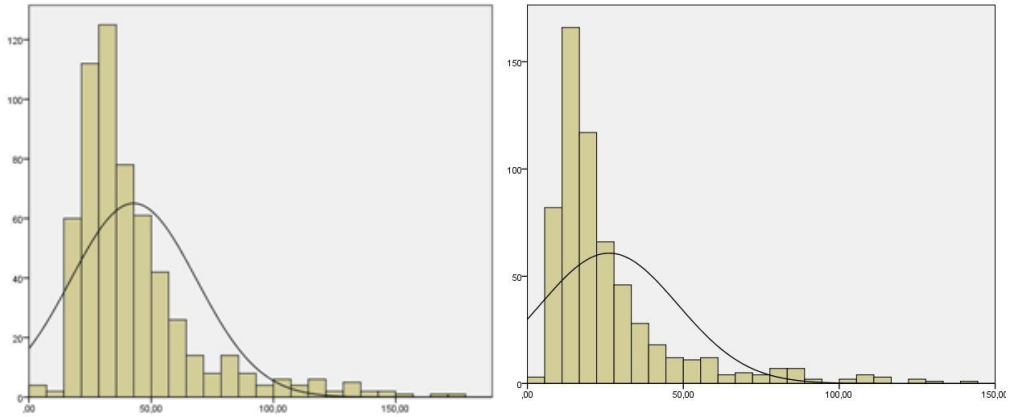

Fig. 3 Histograms with normal curve of PM10 at Station 1 (left) and PM2.5 (right)

The Kolmogorov – Smirnov and Shapiro – Wilk tests for normality are both significant ($p < 0.0005$) so we have to reject the hypotheses of normality of the distributions of PM10 and PM2.5.

3.1 Analysis and modelling of PM10 values measured at Station 1

In this section we will look in more detail at the values of the dangerous air pollutant PM10 ($\mu g/m^3$) measured in Station 1 in Ruse for the period 24.10.2015 - 31.07.2017. The study covers 647 days. Values for 584 days were measured, i.e. for 90.26% of the days, and the values for 63, i.e. for 9.74% of them are missing.

The sample quartiles are Q1=27.10; Q2=35.70 and Q3=50.75. All 41 outliers lie above 86.225, i.e. 1.5*(Q3-Q1) from the third quartile Q3; 14 of them can be classified as extreme – they exceed 121.70, i.e. 3*(Q3-Q1) from the third quartile Q3. The outliers are distributed only in winter months: 21 in January (8 extreme), 9 in February (3 extreme), 6 in November (3 extreme) and 5 in December (0 extreme). Almost all values above the norm, in particular 118 out of 146, are measured in the winter months: 32 in January, 24 in February, 10 in March, 21 in November and 31 in December. In fact, 57.14% of the valid measurements in January are above the norm. For February, November and December these percentages are 42.11%, 35% and 53.45% respectively. The monthly averages of PM10 are shown on Figure 4. Although seasonality is clearly expressed, the observation period is not long enough to examine seasonality with respect to the months of the year.

Figure 5 presents the mean values of PM10 for weekdays. The analysis of variance performed did not show statistically significant ($p = 0.94$) differences in the means over the days of the week.

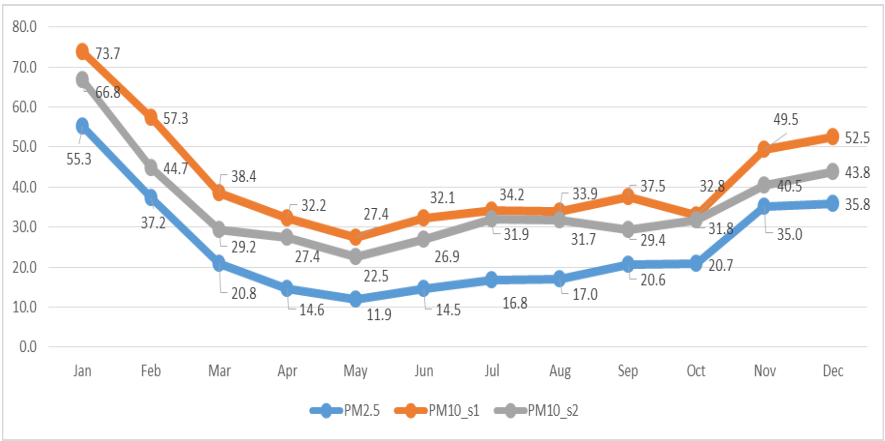

Fig. 4 Monthly averages of PM2.5 and PM10 at Stations 1 and 2

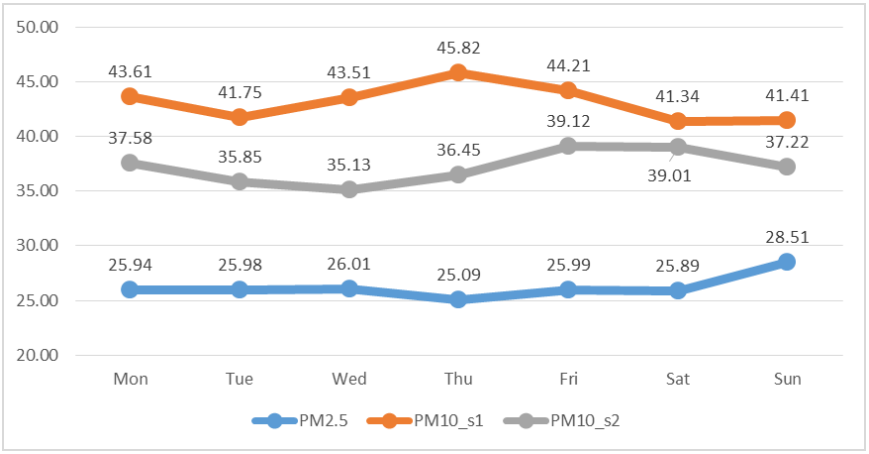

Fig. 5 Mean values of PM2.5 and PM10 at Stations 1 and 2 for weekdays

We will model the time series of values of PM10 at Station 1 using the method of ARIMA (Auto-Regressive Integrated Moving Average) as one of the better known short-term forecasting methods (see [4, 5]). We replace the 63 missing values using linear interpolation. Because in the future it is expected new extreme values to be observed, it is not appropriate to exclude the 41 outliers from the study.

The best reported model using Expert modeler mode in Forecasting procedure (Create Traditional Models) of SPSS gives the following autoregressive ARIMA(1,0,2)(1,0,1)$_7$ model:

Model	Model Fit statistics						Ljung-Box Q(18)			Number of Outliers
	Stationary R-squared	R-squared	RMSE	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.	
1	0.599	0.632	15.305	25.535	10.229	5.516	30.231	14	0.007	1

Model	Transformation	Model Parameters		Estimate	SE	t	Sig.
1	Natural Logarithm	Constant		3.612	0.055	65.597	0.000
		AR	Lag 1	0.821	0.026	31.021	0.000
		MA	Lag 2	0.152	0.046	3.338	0.001
		AR, Seasonal	Lag 1	-0.792	0.188	-4.222	0.000
		MA, Seasonal	Lag 1	-0.738	0.208	-3.553	0.000

Outliers		Estimate	SE	t	Sig.
05.01.2017	Additive	1.186	0.234	5.061	0.000

Using the Ljung – Box test for randomness of the residuals we have evidence to conclude that the fitted Model 1 is not adequate. The model determines the observation on 05.01.2017 as an additive outlier, i.e. it corresponds to a surprisingly large (in our case) or small value, but the subsequent observations are unaffected by it. We introduce in the model one intervention variable I,

$$I = \begin{cases} 0, & t \neq 05.01.2017 \\ 1, & t = 05.01.2017 \end{cases} \quad (1)$$

and run again the Expert modeler, which gives a model without outliers, but having identical characteristics with Model 1, i.e. still not adequate. Undoubtedly, the natural logarithm transformation is necessary, because the variance of the time series is obviously non-stationary (see Fig.3). Let us denote by Z_t, $Z_t = \ln(Y_t)$, the transformed series of the original data Y_t. The new series Z_t are shown on Figure 6, it has already stationary variance.

The autocorrelation (ACF) and partial autocorrelation (PACF) functions of Z_t are given on Figure 7.

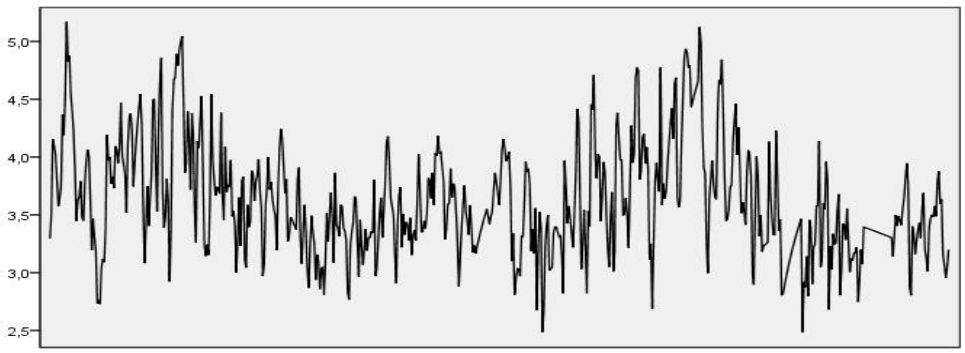

Fig. 6 Time series plot of Zt

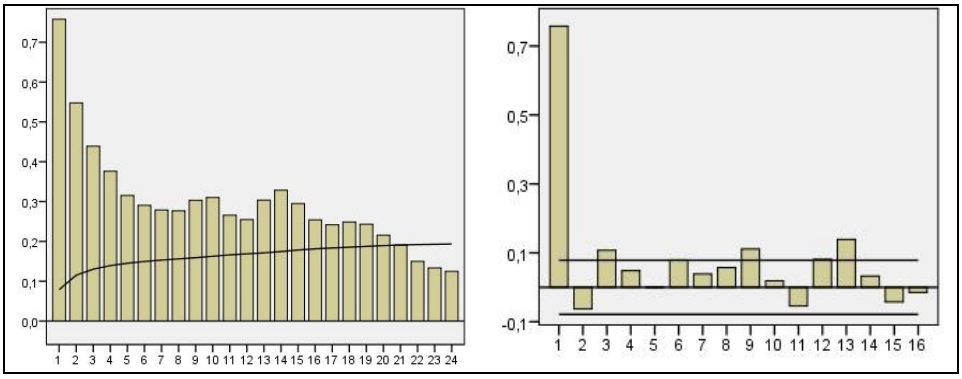

Fig.7 The ACF (left) and PACF (right) of transformed series Zt

The autoregressive part of the model, i.e. the AR component is probably 1, since the PACF "cuts off" at lag 1. In [1] we show that the best model for PM10, measured at Station 2 in the period 24.10.2015 - 25.05.2017, is the AR(1) model for the natural logarithm transformed series Z'$_t$ of the original data,

$$Z'_t = 3.544 + 0.749 Z'_{t-1} + \varepsilon'_t,$$

where ε'_t is a shock, innovation, or random error at time t. This model has 3 outliers: Innovational at 12.01.2016, Additive at 15.12.2016 and Level Shift at 07.04.2017. We used three intervention variables

$$I_1 = \begin{cases} 0, & t < 12.01.2016 \\ 1, & t \geq 12.01.2016 \end{cases}, \; I_2 = \begin{cases} 0, & t \neq 15.12.2016 \\ 1, & t = 15.12.2016 \end{cases}, \; I_3 = \begin{cases} 0, & t < 7.04.2017 \\ 1, & t \geq 7.04.2017 \end{cases}$$

to obtain a model free of outliers:

$$Z'_t = 3.867 + 0.739 Z'_{t-1} + \varepsilon'_t - 0.393 I_1 + 0.928 I_2 - 0.565 I_3$$

which explains 60% of the variation in the series Z'$_t$ (Stationary R-squared equals to 0.596).

Let us use the same model - AR(1), for the values of PM10 measured in Station 1. We obtain the model

$$Z_t = 3.613 + 0.765 Z_{t-1} + \varepsilon_t,$$

where ε_t is the error term at time t for the transformed series Z_t of PM10 at Station 1. The model is not adequate. It identifies the observation on 05.01.2017 as an Additive outlier. Analyzing the charts of the ACF and PACF of the residuals of this model (Figure 8), we see that at lags 2, 10 and 14 they are outside of the defined limits.

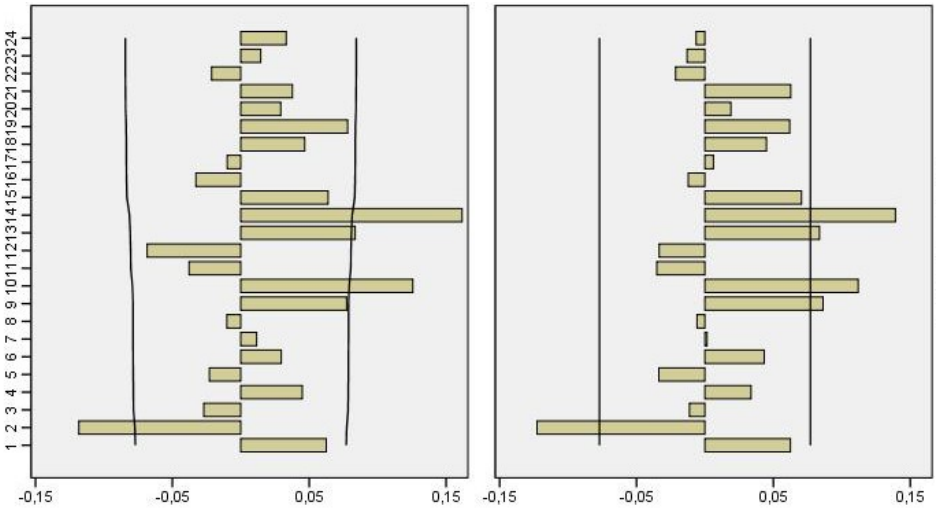

Fig.8 Residual ACF (left) and PACF (right) for the model AR(1) for Zt

The graphics of ACF on Figure 7 shows that the autocorrelations begin to decay toward zero only after the lag 13 (13=14-1). Based on these considerations, it can be assumed that a model ARIMA(1,0,14) with MA terms only at lags 2, 10 and 14 could be appropriate for describing the time series Z_t. Really, when we start Expert modeler again, but now with the option to consider only no seasonal models, it reports this model as the best one.

The obtained Model 4 has the same number of estimated parameters as Model 1, but is already adequate. It has better characteristics than Model 1 at all points. To get rid of the additive outlier on 05.01.2017 we use again the intervention variable I in (1) and get the following Model 5:

$$Z_t = 3.611 + 0.780Z_{t-1} + \varepsilon_t - 0.111\varepsilon_{t-2} + 0.107\varepsilon_{t-10} + 0.130\varepsilon_{t-14} + 1.162I\,.$$

The Model 5 has identical characteristics with Model 4, but it does not contain outliers. Stationary R-squared equals to 0.607 indicating that the model explains 61% of the variation in the transformed series Z_t of PM10 at Station 1. The forecasting for the next 10 days is given in table 3 below. The root-mean-square error for the next 7 days is 2.95 and for the next 10 days is 6.35.

Figure 9 shows the plot of the observed and fitted values of PM10 by Model 5. It can be well seen that the model gives good results and can be used for prediction of PM10 pollution in Ruse.

	Model Fit statistics						**Ljung-Box Q(18)**			
Model	Stationary R-squared	R-squared	RMSE	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.	Number of Outliers
4	0.607	0.640	15.144	25.114	10.022	5.495	18.556	14	0.183	1

Model	Transformation	Model Parameters		Estimate	SE	t	Sig.
4	Natural Logarithm	Constant		3.611	0.061	59.455	0.000
		AR	Lag 1	0.780	0.029	26.871	0.000
		MA	Lag 2	0.111	0.045	2.455	0.014
			Lag 10	-0.107	0.039	-2.740	0.006
			Lag 14	-0.130	0.039	-3.314	0.001

Outliers		Estimate	SE	t	Sig.
05.01.2017	Additive	1.162	0.235	4.950	0.000

Table 3. Forecast values for model 5

	01.08.	02.08.	03.08..	04.08.	05.08.	06.08.	07.08.	08.08.	09.08.	10.08.
Real values	31,9	37,4	36,4	38,9	35	37,2	36,7	22,8	27,1	41
Forecast	28,00	32,05	34,49	35,91	36,73	36,83	37,95	36,86	37,90	35,69
UCL	48,56	63,44	71,41	76,25	79,16	80,08	82,97	80,85	83,29	78,53
LCL	14,71	13,95	14,03	14,05	14,03	13,88	14,18	13,70	14,05	13,20

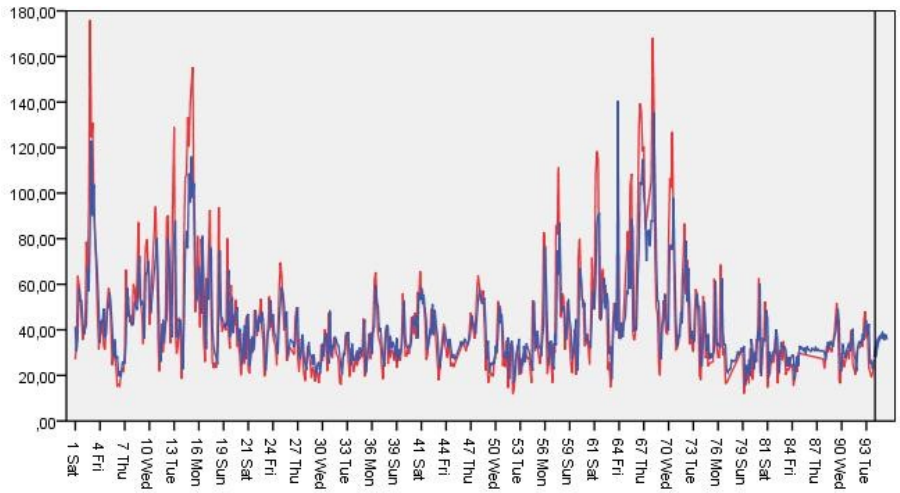

Fig. 9 Observed (red line) and fitted values (blue line) of PM10 at Station 1

3.2 Analysis and modelling of PM2.5 values

The best reported model using Expert modeler is the ARIMA(1,0,2) model:

Model	Model Fit statistics						Ljung-Box Q(18)			Number of Outliers
	Stationary R-squared	R-squared	RMSE	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.	
6	0.654	0.651	12.859	32.646	7.771	5.208	22.766	15	0.089	4

Model	Transformation	Model Parameters		Estimate	SE	t	Sig.
6	Natural Logarithm	Constant		3.027	0.137	22.123	0.000
		AR	Lag 1	0.958	0.014	63.914	0.000
		MA	Lag 1	0.241	0.042	5.752	0.000
			Lag 2	0.357	0.041	8.690	0.000

The model is adequate since the Ljung-Box statistics is not significant at alpha = 0.05. Four of the observations on PM2.5 are identified as outliers: Transient at t_1=12.01.2016 and t_2=16.11.2016, Additive at t_3=04.01.2017 and t_4=07.04.2017. The effect of a Transient (temporary) change outlier diminishes exponentially over the subsequent observations and the series returns to its normal level after several steps. We introduce four intervention variables

$$J_i = \begin{cases} 0, & t \neq t_i \\ 1, & t = t_i \end{cases}, \quad i = 1,2,3,4$$

and obtain an adequate model (Model 7) with similar characteristics, but without outliers:
$W_t = 3.037 + 0.960W_{t-1} + a_t - 0.241a_{t-1} - 0.361a_{t-2} - 1.511\delta_{1,t} + 1.450\delta_{2,t} + 1.467J_3 - 1.471J_4$,
where $W_t = \ln U_t$, U_t is the time series of values of PM2.5 at Station 1; a_t is a shock, innovation, or random error at time t and $\delta_{1,t}$, $\delta_{2,t}$ are defined by

$$\delta_{1,t} = \begin{cases} 0, & t < t_1 \\ 0.802^{(t-t_1)}, & t \geq t_1 \end{cases}, \quad \delta_{2,t} = \begin{cases} 0, & t < t_2 \\ 0.753^{(t-t_2)}, & t \geq t_2 \end{cases}. \tag{2}$$

The forecast of Model 7 for the next 10 days is given in Table 4 below. Inspecting the residual ACF and PACF of Model 7, only these of Lag 10 are beyond the defined limits. When we introduce in Model 7 an MA term of Lag 10, we obtain the following model (Model 8):

$$W_t = 3.028 + 0.925W_{t-1} + a_t - 0.188a_{t-1} - 0.323a_{t-2} + 0.122a_{t-10} - 1.498\delta'_{1,t} + 1.389\delta'_{2,t} + 1.498J_3 - 1.374J_4,$$

where $\delta'_{1,t}$ and $\delta'_{2,t}$ are the same as $\delta_{1,t}$ and $\delta_{2,t}$ in (2), but 0.802 is replaced by 0.798 and 0.753 by 0.778 respectively. The Model 8 has better characteristics than Model 7 at all points:

Model	Model Fit statistics						Ljung-Box Q(18)			Number of Outliers
	Stationary R-squared	R-squared	RMSE	MAPE	MAE	Normalized BIC	Statistics	DF	Sig.	
8	0.657	0.666	12.604	32.479	7.698	5.178	16.465	14	0.286	0

It also has a better performance in the next 10 days.

Table 4. Forecast values for models 7 and 8

	01.08.	02.08.	03.08.	04.08.	05.08.	06.08.	07.08.	08.08.	09.08.	10.08.
Real values	19,2	20,3	22,5	20,4	20,7	23,9	-	-	-	20,5
Model 7	16,20	17,15	17,48	17,81	18,12	18,42	18,71	18,98	19,25	19,51
Model 8	16,59	18,79	19,50	20,07	18,95	19,57	19,21	19,05	19,04	19,32

The root-mean-square error of the prediction for the available real values of PM2.5 (the values for the period 07 - 09 august are missing) for Model 7 is 3,555 and for Model 8 - 2,430. Figure 10 shows the plot of the observed and fitted values of PM2.5 by Model 8.

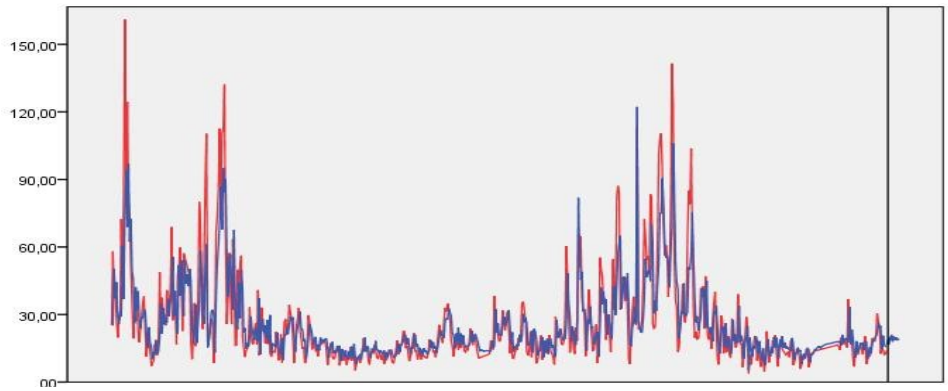

Fig. 10 Observed (red line) and fitted by Model 8 values (blue line) of PM2.5

4 Conclusions

The most dangerous air pollutants (PM10 and PM2.5) for Ruse region in Bulgaria measured in the period 24.10.2015 - 25.05.2017 by two different measuring stations were statistically analyzed in the paper. The measurements of PM10 show that during winter time the concentration of PM10 is above the norm, reaching a maximum value of 217.30 ($\mu g/m^3$) at norm 50 ($\mu g/m^3$) for the average daily rate. Next to examine better these air pollutants we create ARIMA statistical models of PM10 and PM2.5 measured by station1. The values of PM10 and PM2.5 given by the models are in a good agreement with the measurements.

The main atmospheric characteristics for Ruse region (temperature, atmospheric pressure and humidity) measured for more than 40 year period were examined in the paper [6]. For the period 1988 – 2012 the average temperature for Ruse region is higher than the mean temperature for Bulgaria. At the same time in Ruse region it is observed that water of river Danube and the air pollution increased [7, 8, 9]. Obviously the atmospheric characteristics temperature, atmospheric pressure and humidity are interrelated and maybe they are correlated with the air pollution. Especially strong is the air pollution during the autumn-winter season with particulate matter in the last decade [1]. The reasons for this are not only transportation and industry but the use of solid fuel heating. Air pollution maybe affects temperature, atmospheric pressure and humidity and this influence may cause the climate change in the Ruse region. This influence has to be studied in details further.

This paper contains results of the work on project No 2017 - FNSE – 05 and No 2017 – AIF - 3 financed by „Scientific Research" Fund of Ruse University.

References

1. I. Zheleva, E. Veleva, M. Filipova Analysis and modeling of daily air pollutants in the city of Ruse, Bulgaria Application of Mathematics in Technical and Natural Sciences AIP Conf. Proc. (2017) (in print);
2. National report on the status and protection of the environment in Bulgaria in 2012, Executive Environment Agency, Ministry of Environment and Water, available at www.eea.government.bg
3. http://www.riosv-ruse.org/kachestvo-na-atmosferniya-vazduh.html
4. A. Pankratz, Forecasting with dynamic regression models, John Wiley and Sons, New York , 1991.
5. A. Pankratz, Forecasting with Univariate Box - Jenkins Models: Concepts and Cases, John Wiley and Sons, New York , 1983.
6. I.Zheleva, M.Filipova Atmospheric Characteristics Statistic Study of Ruse Region, Bulgaria Application of Mathematics in Technical and Natural Sciences AIP Conf. Proc. 1773, 110019 (2016); doi: 10.1063/1.4965023
7. Filipova M., I. Zheleva, P. Rusev, Characteristics of PM air pollution along Bulgaria - Romania Danube region ECOLOGICA, 2013, 71, p.215-217P.
8. M. Filipova, I. Zheleva, A. Lecheva, P. Rusev, Analysis of Key Pollutant Surface Waters of the Tributaries of the Danube River in Bulgarian Section, Pliska Stud. Math. Bulgar. 24 (2015), 151-162
9. Filipova M., I. Zheleva, P. Rusev, Investigation of monitoring systems for water quality of the Danube River in the border region Romania – Bulgaria, ECOLOGICA 2015, Vol. 22 №77, p. 13-18