Seasonality of the levels of particulate matter PM10 air pollutant in the city of Ruse, Bulgaria

Cite as: AIP Conference Proceedings **2302**, 030006 (2020); https://doi.org/10.1063/5.0033628
Published Online: 03 December 2020

E. Veleva, and I. R. Georgiev

Seasonality of the Levels of Particulate Matter PM10 Air Pollutant in the City of Ruse, Bulgaria

E. Veleva[a] and I.R. Georgiev[b]

Dept of Applied Mathematics and Statistics, Ruse University, 7017 Ruse, 8 Studentska str., Bulgaria
www.uni-ruse.bg

[a] Corresponding author: eveleva@uni-ruse.bg
[b] irgeorgiev@uni-ruse.bg

Abstract. High levels of air pollutants PM10 are a problem of great importance for human health. During the months from April to September of the period 2010 - 2019 the levels in Ruse remain within the norm in 95% of the days. During the other, "cold" months of the year, only 58% of the days have values below the daily norm of 50μg/m^3. When planning their activities, it is useful for people to have forecasts for PM10 levels in the coming days. Markov chains allow such predictions to be given in a tabular form, convenient to use, without the need for calculations.

The data for the "cold" months are modelled using three Markov chains with different degrees of discretization of the original values, respectively with 12, 7 and 3 possible states. The latter, with states: {in the norm}, {slightly above the norm} and {a strong excess of the norm}, can be used without official data on the exact PM10 levels. Determining the condition of the PM10 pollutant today in this case can also be done on the basis of a personal assessment of the purity of the air over the city at the moment.

The measured levels in the period 01.01.2020 - 31.03.2020 were used as test data. They show consistency of the measured levels in 2020 with all three Markov chains considered. The obtained tabular values can be used to predict PM10 levels in the following years, in the months from October to March.

INTRODUCTION

The main air pollutant in Ruse is the particulate matter PM10, as it can be seen from our previous work ([1] – [3]). Its levels vary throughout the year, remaining moderate in the warmer months and rising, often above the average daily norm of 50 μg/m^3, during the colder months. Studies exploring the relationship of PM10 with different atmospheric characteristics for the cities of Ruse and Silistra, located close to each other along the Danube, were made in [4] and [5]. Seasonal models for the levels of this air pollutant have been estimated for other cities in Bulgaria, for example in [6] and [7]. This paper uses different approaches of time series analysis for modelling of the trend and seasonality in the data: regression models, decomposition, seasonal ARIMA models. The results are used to obtain point and interval estimates for future values of PM10 levels.

The study is based on 3652 values of the average daily levels of PM10 in the city of Ruse, measured in the period from 01.01.2010 to 31.12.2019 (Figure 1).

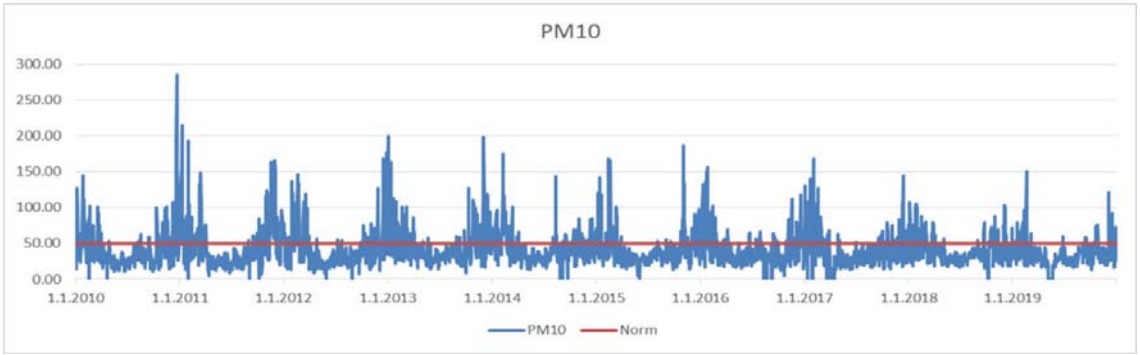

FIGURE 1. Average daily values of PM10 levels in the city of Ruse for the period 2010 – 2019

The data set contains 55 missing values (*i.e.*, 1.51% of all observations), 12 of which are consecutive days in April 2017, 15 are consecutive days in May 2019, and the remaining 28 are located at random. The minimum observed value is 1.40μg/m^3, registered on 21.4.2010, and the two largest are 285.30μg/m^3 and 261.30μg/m^3,

Application of Mathematics in Technical and Natural Sciences
AIP Conf. Proc. 2302, 030006-1–030006-8; https://doi.org/10.1063/5.0033628
Published by AIP Publishing. 978-0-7354-4036-4/$30.00

measured in two consecutive days - on 22 and 23 December 2010. Figure 2 shows the boxplots of the data for each of the years. No trend is observed in both Figures 1 and 2. There is however a strong seasonal pattern with frequency of 365 days / 12 months, that decreases in size over time.

The next Figure 3 shows the number of recorded observations above the norm for each of the years. Normally, values above 50 μg/m^3 should not be more than 35 for each year. As can be seen, this condition has been violated for each year of the period considered. Figure 3 also shows the annual changes in the mean and standard deviation of PM10 levels.

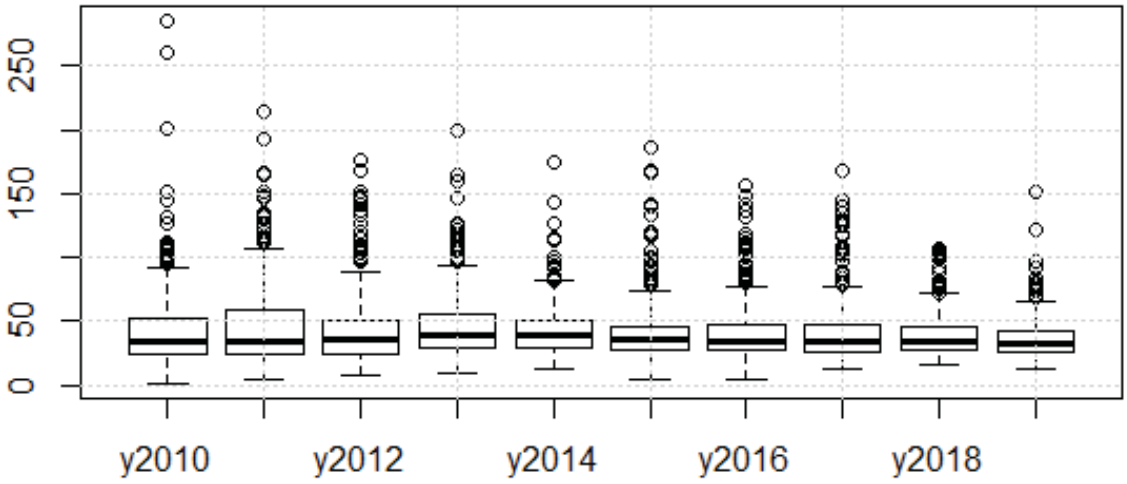

FIGURE 2. Boxplots of the values in each year

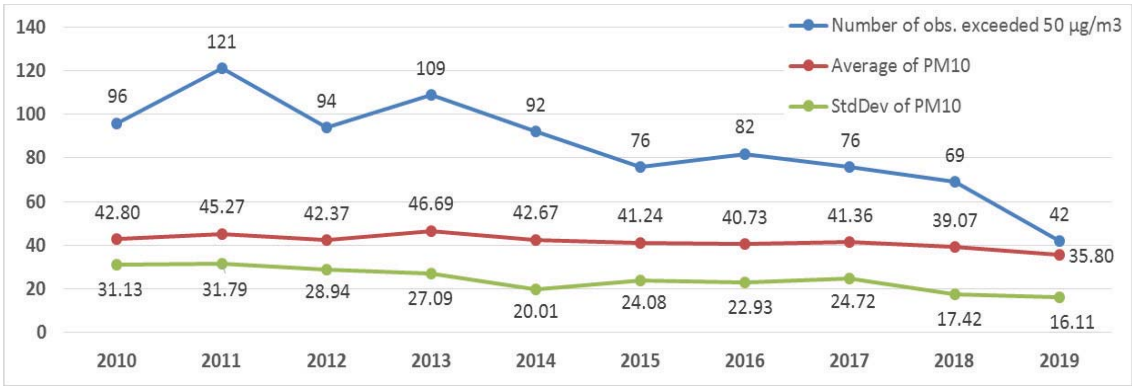

FIGURE 3. Number of observations, exceeded the norm of 50μg/m3, for each of the years

1. MODELLING THE SEASONALITY IN THE DATA

The average values of PM10 levels for the days from Monday to Sunday are respectively: 41.88, 41.29, 42.99, 42.65, 42.53, 41.11 and 40.19. The performed analysis of variance shows that the data do not lead to the rejection of the hypothesis of equality of the average levels of PM10 on different days of the week (F = 0.8280 with 6 and 3590 df, p = 0.5480). The boxplots of the samples corresponding to the different days of the week presented in Figure 4 also do not show a fundamental difference in the distributions of PM10 levels.

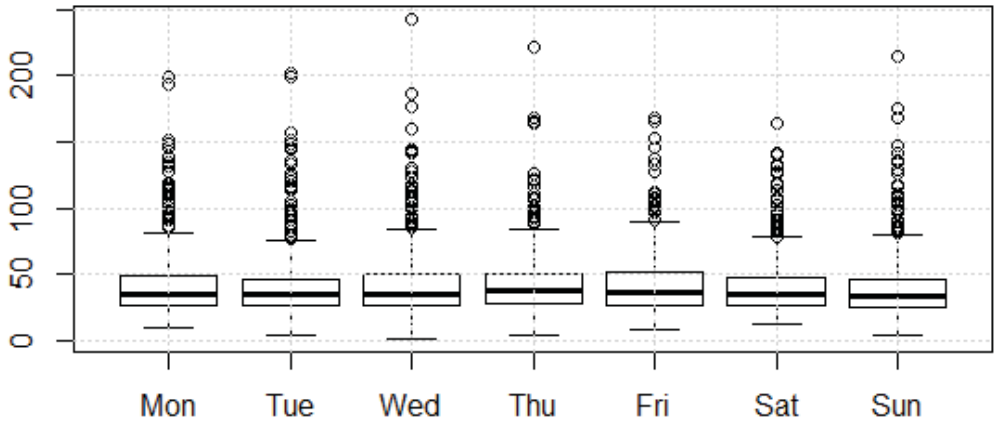

FIGURE 4. Boxplots of the observations in different weekdays

The average values of PM10 levels in the months from January to December are respectively: 62.20, 58.14, 47.24, 32.07, 26.17, 26.58, 30.39, 34.31, 35.97, 41.45, 49.24, 57.73. The performed analysis of variance clearly shows that the hypothesis of equality of the average values of PM10 levels in the different months of the year contradicts the experimental data (F= 102.0777 with 11 and 3585 df, p = 1e-202). The boxplots of measurements during the different months of the year, presented in Figure 5, show a fundamental difference in the distributions of PM10 levels.

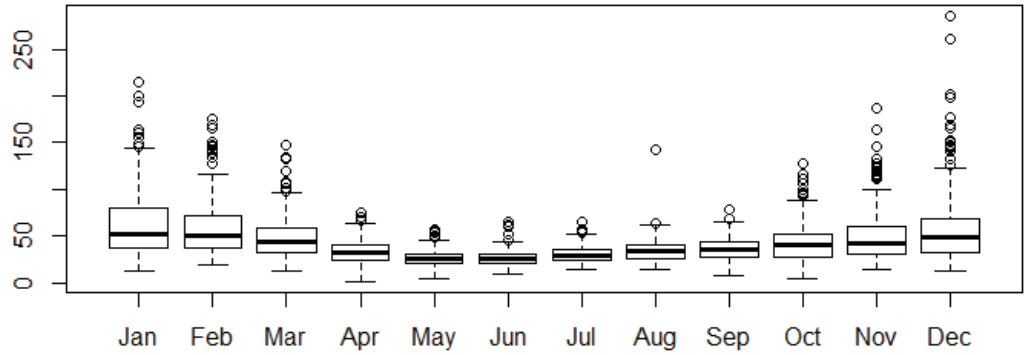

FIGURE 5. Boxplots of the observations in different months

Let us denote by Y_t the series of 120 average monthly values of PM10 levels, measured in the city of Ruse, Vazrazhdane station, in the period 01.2010 - 12.2019. In the next subsections we model the trend and seasonality (with frequency 12 months) in Y_t using different approaches of time series analysis. The five average monthly values for the period 01.2020 - 05.2020 are used as test data for the obtained models. Calculations are done by R programming language [8]. The best model of each type was selected using the Corrected Akaike's Information Criterion AICc,

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2(k+2)(k+3)}{T-k-3}\ , \qquad \mathrm{AIC} = T\log\left(\frac{SSE}{T}\right) + 2(k+2)\ ,$$

where T = 120 is the number of observations used for estimation, k is the number of predictors in the model and SSE is the sum of squared errors e_t of the predictions. Along with AICc, for the selected best model of each type, the values of the criteria AIC, BIC (Schwarz's Bayesian Information Criterion), RMSE, MAE and MAPE are also given,

$$\mathrm{BIC} = T\log\left(\frac{SSE}{T}\right) + (k+2)\log(T)\ ,$$

$$\mathrm{RMSE} = \sqrt{mean(e_t^2)}\ ,\ \mathrm{MAE} = mean\left(|e_t|\right),\ \ \mathrm{MAPE} = mean\left(|p_t|\right),\ \ p_t = 100\,e_t / Y_t\ .$$

The values of RMSE, MAE and MAPE are calculated separately for the set of 120 data (training set) and on the set of test data (test set).

1.1. Modelling by Linear Regression with Seasonal Dummy Variables

Let us consider the linear regression model for $\log(Y_t)$ with a linear trend and 11 monthly dummy variables

$$\log(Y_t) = \beta_0 + \beta_1 t + \beta_2 d_{2,t} + \beta_3 d_{3,t} + \cdots + \beta_{12} d_{12,t} + \varepsilon_t ,$$

or equivalently

$$Y_t = e^{\beta_0} (e^{\beta_1})^t (e^{\beta_2})^{d_{2,t}} (e^{\beta_3})^{d_{3,t}} \ldots (e^{\beta_{12}})^{d_{12,t}} e^{\varepsilon_t} ,$$

where $d_{i,t} = 1$ if t is in the month i of the year and 0 otherwise. The estimated model by the least squares method is

$$\begin{aligned} \textbf{Model1}: \; Y_t &= 64.9035\,(0.9989)^t\,(0.9400)^{d_{2,t}}\,(0.7616)^{d_{3,t}}\,(0.5159)^{d_{4,t}}\,(0.4221)^{d_{5,t}} \\ &\times (0.4330)^{d_{6,t}}\,(0.4956)^{d_{7,t}}\,(0.5597)^{d_{8,t}}\,(0.5915)^{d_{9,t}}\,(0.6765)^{d_{10,t}}\,(0.7863)^{d_{11,t}}\,(0.9356)^{d_{12,t}}\,e^{\varepsilon_t} \end{aligned} \tag{1}$$

where ε_t is a white noise with standard deviation 0.1710. All coefficients in (1) are significant at 1% level, except the trend, significant at 2% level, $\hat{\beta}_2$ (p=0.4204) and $\hat{\beta}_{12}$ (p=0.3872), corresponding to February and December. The mean values at these two months are close to those at January. The autocorrelation in the residuals of order up to 24 is tested with the Breusch-Godfrey test, also referred to as the LM (Lagrange Multiplier) test for serial correlation: LM test = 23.27, df = 24, p-value = 0.5039. Thus, we can conclude that the residuals are not distinguishable from a white noise series. Characteristics of the quality of Model1 are given in Table 1. Table 2 shows the point and interval forecasts for the next 5 average monthly values of PM10.

Time series plot of the average levels of PM10 for each month in the period 01.2010 – 12.2019, the fitted and forecasted by Model1 values, along with a 95% confidence intervals are given in Figure 6.

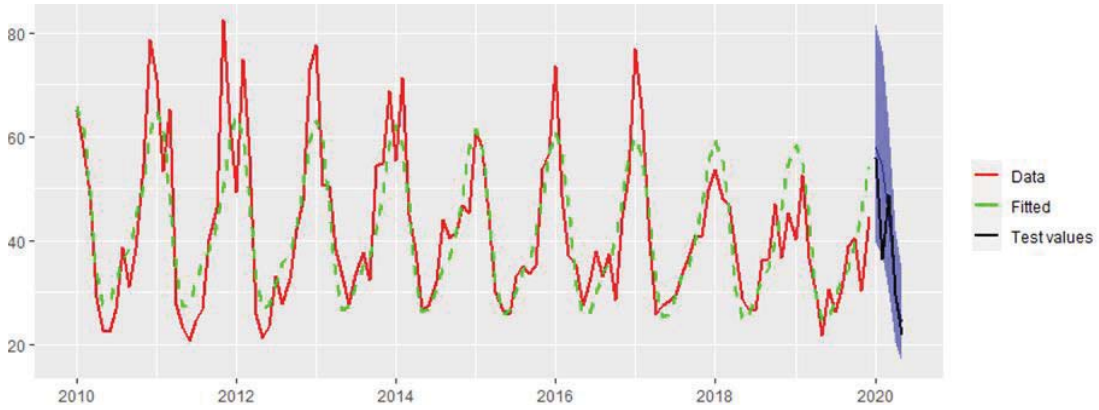

FIGURE 6. Time series plot of the monthly average levels of PM10, the fitted and forecasted by Model1 values

1.2. Modelling by STL (Seasonal and Trend Decomposition using Loess) Decomposition Method with Multiplicative Components

The monthly average levels of PM10, Y_t, are presented in the form $Y_t = T_t S_t R_t$, where T_t, S_t and R_t are the trend-cycle, seasonal and the remainder components, estimated by the STL method (see [9]) and given on Figure 7.

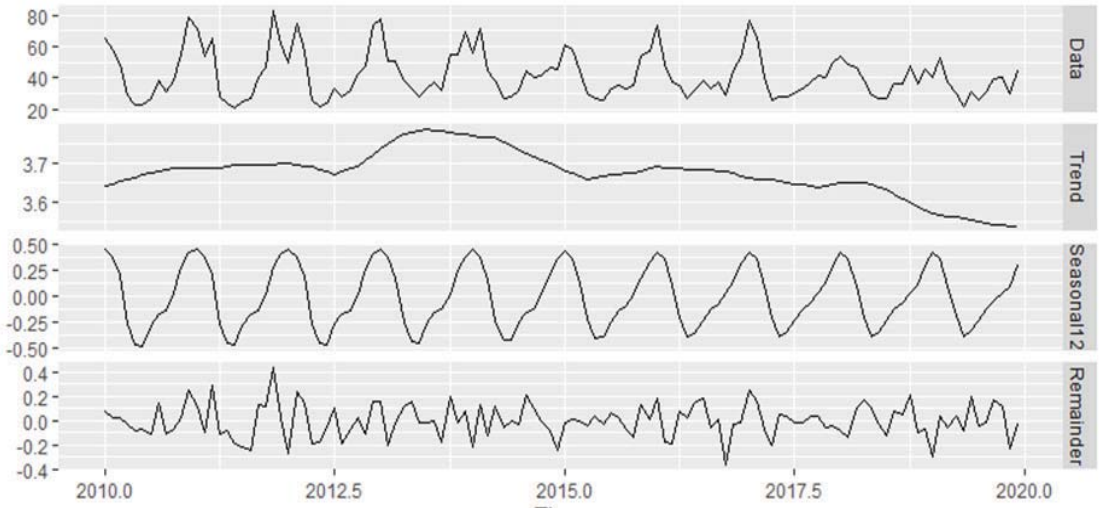

FIGURE 7. Time series plot of the data and its multiplicative components, estimated by STL method

STL is a versatile and robust method for decomposing time series. STL is an acronym for "Seasonal and Trend decomposition using Loess", while Loess is a method for estimating nonlinear relationships (see [9, 10]).

On the seasonal sub-series plot of the seasonal component of PM10 (Figure 8) we see a decreasing tendency of the levels of PM10 during the cold months and an increasing tendency during the warm months of the year.

The seasonally adjasted values are $A_t = Y_t / S_t$. Using an automated procedure in R programming language and more precisely the function *stlf*, the $Z_t = \log(A_t)$ time series is then modelled by the ARIMA(0,1,1) process $Z_t = Z_{t-1} - 0.9228\varepsilon_{t-1} + \varepsilon_t$, where ε_t is a white noise with standard deviation 0.1479.

We consider the combination of STL decomposition + ARIMA model as Model2, which produce forecasts for the monthly levels of PM10 in 2020, shown on Figure 9, along with the original, fitted data and 95% confidence intervals.

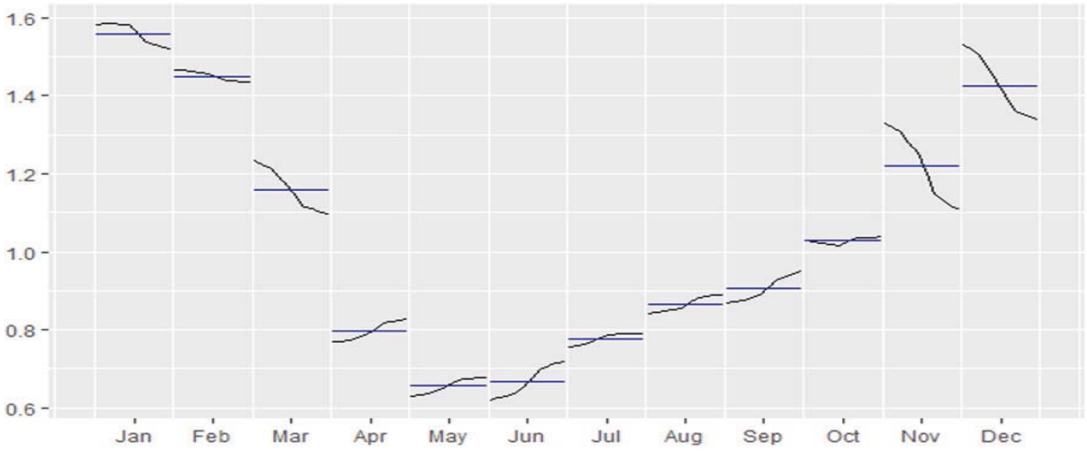

FIGURE 8. Seasonal sub-series plot of the seasonal component of PM10 levels

The autocorrelation in the residuals of order up to 24 is tested with the Ljung-Box test for serial correlation: Q* = 22.492, df = 23, p-value = 0.4908. Thus, we can conclude that the residuals are not distinguishable from a white noise series. Characteristics of the quality of Model2 are given in Table 1. Table 2 shows the predictions for the next 5 average monthly levels of PM10.

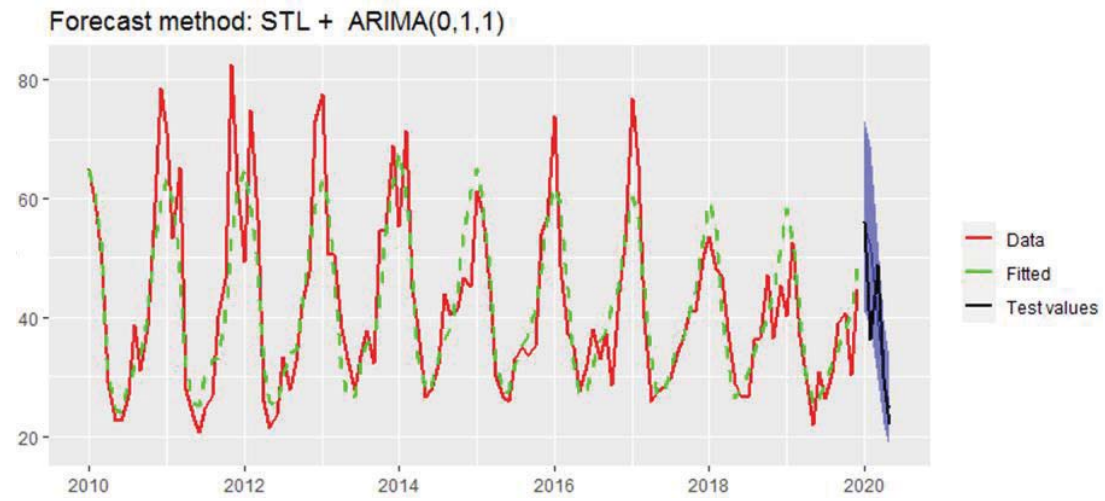

FIGURE 9. Time series plot of the monthly average levels of PM10, the fitted and forecasted by Model2 values

1.3. Modelling the Log Average Values by Seasonal ARIMA Models

We modelled the time series of $\log(Y_t)$ by seasonal ARIMA models. The autocorrelation function of $\log(Y_t)$ is not decaying and has a sinusoidal form with a period of 12. The autocorrelation (ACF) and partial autocorrelation (PACF) functions of the seasonally differenced series of $\log(Y_t)$ are shown on Fig. 10.

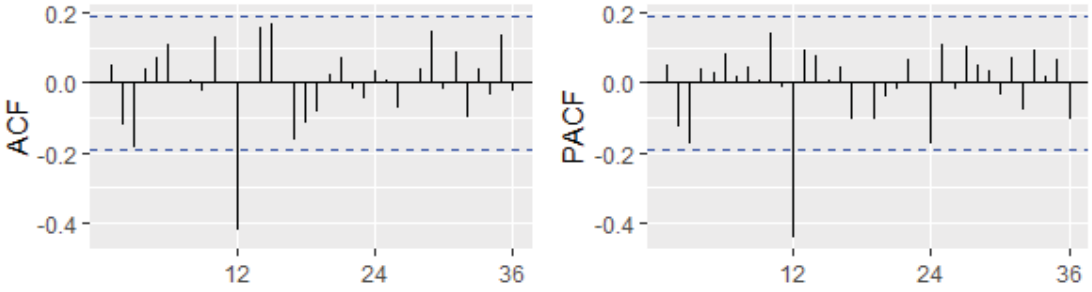

FIGURE 10. The ACF and PACF plots of the seasonally differenced series

We tried several seasonal arima models. The model with minimal AICc, estimated by the R function *Arima*, was ARIMA(0,0,0)(0,1,1)$_{12}$:

$$\log(Y_t) = \log(Y_{t-12}) - 0.0011 + \varepsilon_t - 0.5861\varepsilon_{t-12},$$

or equivalently

$$\textbf{Model3:} \quad Y_t = Y_{t-12}\, 0.9989\, e^{\varepsilon_t - 0.5861\varepsilon_{t-12}},$$

where ε_t is a white noise with standard deviation 0.1801. The autocorrelation in the residuals of order up to 24 is tested with the Ljung-Box test for serial correlation: Q^* = 23.388, df = 22, p-value = 0.3801. Thus, we can conclude that the residuals have no remaining autocorrelations. Characteristics of the quality of Model3 and predictions for the next 5 values are given in Table 1 and 2. Time series plot of the average levels of PM10, the fitted and forecasted by Model3 values are given in Figure 11.

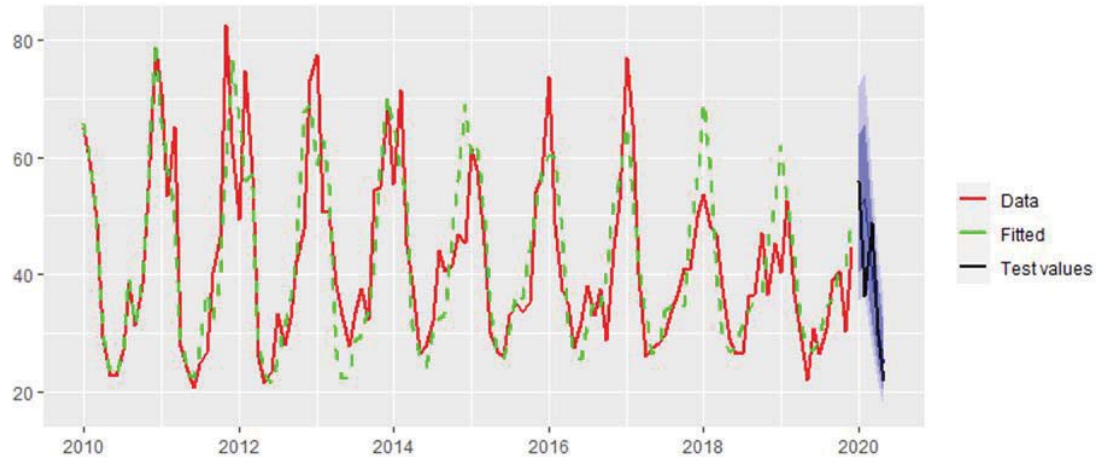

FIGURE 11. Time series plot of the monthly average levels of PM10, the fitted and forecasted by Model3 values

2. COMPARISONS BETWEEN THE THREE MODELS

The Table 1 below gives the values of one of the most commonly used criteria in practice estimating the quality of a model.

TABLE 1. Characteristics of the quality of the three models

Model	AIC	AICc	BIC	Train. RMSE	Train. MAE	Train. MAPE	Test RMSE	Test MAE	Test MAPE
Model1	-409.61	-405.61	-370.59	7.4658	5.5801	13.3947	8.5031	5.5410	15.1356
Model2	-112.30	-112.19	-106.74	7.1273	5.0963	11.7543	8.2269	5.7268	15.4338
Model3	-54.72	-54.49	-46.68	7.9553	5.7658	13.4177	8.7302	6.8906	18.1545
AV12= average of Model1 and Model2				7.2151	5.2937	12.4383	8.2734	5.5080	15.0693

On the training data set Model2 has the minimal values for the RMSE, MAE and MAPE criteria. Therefore it is the best one of all three models in describing the data. This is due to the fact that STL decomposition is effective in seasonal waves that change over time (see [9]). Model3 has the highest values for all three criteria calculated on the training and on the test data sets.

When forecasting future observations, Model2 has the smallest value for RMSE, while Model1 – for MAE and MAPE. If the mean of the values predicted by Model1 and Model2 is used for forecasting (model AV12), even lower values for MAE and MAPE and close to that of Model2 value for RMSE are obtained.

The table below gives the actual average levels of PM10 for the first 5 months of 2020 year, the lower (L) and upper (U) endpoints of the 95% confidence intervals and the mean (M) of the forecasted values of PM10 by the three models.

TABLE 2. Point and interval forecasts for the next 5 average monthly values of PM10

Model	L M U	L M U	L M U	L M U	L M U
Model1	39.75; 58.0; 81.73	37.32; 54.40; 76.75	30.21; 44.00; 62.12	20.44; 29.8; 42.03	16.71; 24.4; 34.35
Model2	40.89; 55.23; 73.0	38.50; 52.06; 68.87	29.43; 39.82; 52.72	22.13; 29.98; 39.71	18.17; 24.63; 32.66
Model3	35.61; 51.42; 72.14	36.57; 52.81; 74.10	27.75; 40.08; 56.23	21.72; 31.36; 44.00	17.15; 24.76; 34.74
Actual	55.91	36.35	48.92	29.71	21.75

In all of the three methods, the lower endpoint of the February forecast is higher than the actual observed value. The means of the lengths of the obtained 95% confident intervals for each of the three models are 30.51, 23.57 and 28.48 respectively. Model2 gives intervals with the smallest length for each of the five future values.

CONCLUSIONS

The study is based on the measured in the city of Ruse levels of the air pollutant PM10 over the ten-year period 2010 – 2019. Three different approaches have been applied to model data containing a seasonal

component. All three obtained models are multiplicative. There is a gradual slow decline in the average level of PM10, and with it the standard deviation, *i.e.*, there is a downward trend and a decrease in the amplitude of the seasonal wave in the period under review over the years. Model1 uses linear regression with seasonal dummy variables, Model2 is a combination of STL decomposition method and ARIMA(0,1,1) model and Model3 is a seasonal ARIMA$(0,0,0)(0,1,1)_{12}$. A comparison was made between them based on the criteria RMSE, MAE and MAPE. The measured levels in 2020 were used as test values to check the quality of the models to predict future levels. Model2 is the best one in presenting the collected data. For prediction, good results are obtained by Model1, Model2, as well as by the arithmetic mean of both models.

ACKNOWLEDGMENTS

This paper contains results of the work on project No 2020 - FNSE – 04, financed by Scientific Research Fund of Ruse University.

REFERENCES

1. I. Zheleva, E. Veleva, and M. Filipova, "Analysis and modeling of daily air pollutants in the city of Ruse, Bulgaria," in *AMiTaNS'17*, AIP CP1895, edited by M. Todorov (American Institute of Physics, Melville, NY, 2017), paper 030007, 10p.
2. E. Veleva and I. Zheleva, "GARCH models for particulate matter PM10 air pollutant in the city of Ruse, Bulgaria," in *AMiTaNS'18*, AIP CP2025, edited by M. Todorov (American Institute of Physics, Melville, NY, 2018), paper 040016, 9p.
3. E. Veleva and I. Zheleva (2018) Statistical modeling of particle mater air pollutants in the city of Ruse, Bulgaria, *MATEC Web of Conferences* **145**, 01010.
4. I. Tsvetanova, I. Zheleva, and M. Filipova, "Statistical study of the influence of some atmospheric characteristics upon the particulate matter (PM10) air pollutant in the city of Ruse, Bulgaria," in *AMiTaNS'18*, AIP CP2025, edited by M. Todorov (American Institute of Physics, Melville, NY, 2018), paper 110006, 8p.
5. I. Tsvetanova, I. Zheleva, and M. Filipova, "Statistical study of the influence of the atmospheric characteristics upon the particulate matter (PM10) air pollutant in the city of Silistra, Bulgaria," in *AMiTaNS'19*, AIP CP2164, edited by M. Todorov (American Institute of Physics, Melville, NY, 2019), paper 120014, 14p.
6. S. Gocheva-Ilieva, A. Ivanov, and I. Iliev, "Exploring key air pollutants and forecasting particulate matter PM10 by a two-step SARIMA approach," AIP CP2106 (American Institute of Physics, Melville, NY, 2019), paper 020004.
7. S. Gocheva-Ilieva and A. Ivanov (2019) Assaying stochastic SARIMA and generalized regularized regression for particulate matter PM10 modeling and forecasting, *International Journal of Environment and Pollution* **66**, 41-62.
8. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.
9. R.J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd edn (OTexts, Melbourne, Australia, 2018), OTexts.com/fpp2.
10. R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. J. Terpenning (1990) STL: A seasonal-trend decomposition procedure based on loess, *Journal of Official Statistics* **6**(1), 3–73, http://www.jos.nu/Articles/abstract.asp?article=613.