

Modeling Temporal Variation of Particulate Matter Concentration at Three Different Locations of Delhi

ABSTRACT

Key words: ARIMA, GARCH, particulate matter, pollution, time series, volatility

1. INTRODUCTION

The greatest environmental health concern in the modern world is exposure to polluted air [1]. Particulate matters, or PMs for short, are the air pollutants that have the worst effects on human health. Particulate matter is a term used to describe small solids or liquid droplets that enter the human body through the air we breathe and may have either anthropogenic or natural origins. Particles with a diameter of 2.5 to 10 μm (PM_{10}) can penetrate deep inside the lung, whereas $\text{PM}_{2.5}$ particles can enter the circulatory system after breaching the lung barrier. One of the most current estimates on mortality brought on by PMs suggests that, globally, fossil fuel-generated $\text{PM}_{2.5}$ is responsible for about 8.7 million premature deaths [2]. India is a major contributor as well as victim of air pollution as it is the third largest emitter of greenhouse gases [3] and the fifth largest emitter of $\text{PM}_{2.5}$ [4]. In reality, according to the World Health Organization's Air Quality Guidelines, none of the Indian cities have achieved the standard for annual $\text{PM}_{2.5}$ concentration of 5 $\mu\text{g}\text{m}^{-3}$ [5]. Air pollution is India's second-leading cause of mortality and morbidity, after malnutrition [6]. In India, 17.8% of deaths in 2019 were attributed to air pollution, with outdoor particulate matters responsible for 58.7% of those deaths [7].

Out of the fifteen most polluted cities in the world with respect to annual average $\text{PM}_{2.5}$ level, ten are situated in India and eight of them are located in the Delhi-National Capital Region (NCR) [8]. According to the 2021 World Air Quality Report, New Delhi is the fourth most polluted city in the world and the most polluted capital city. It has been well documented that every year since last few decades the average annual $\text{PM}_{2.5}$ concentration in Delhi crosses the annual National Ambient Air Quality Standard of 40 $\mu\text{g}\text{m}^{-3}$ (which is eight times higher than the WHO's standard limit) [1]. However, despite the adoption of a number of interventions by the government agencies, none of the actions proved to be effective for Delhi-NCR.

Prediction of future particulate matter levels play a crucial role in policy formulation of any country, especially those falling under low- and medium-income group. The present work explores the variation in the concentration of PM_{10} and $\text{PM}_{2.5}$ at different sites of Delhi having different pollution signatures. A proper understanding about the concentration of PM_{10} and $\text{PM}_{2.5}$ and their variation over different times of a year can be helpful for the policymakers to take proactive actions to minimize the hazard. Time series modeling approaches are used for the concentration variation study. A time series is the collection

59 of realizations of any variable over a period of time. The most important characteristic of a
60 time series is that the successive realizations are dependent. The time series analysis is
61 pioneered by Box and Jenkins' [9] Autoregressive Integrated Moving Average (ARIMA)
62 methodology. In several instances it is seen that the realizations of a time series exhibit
63 long term dependencies. This phenomenon is known as the long memory property. If any
64 time series exhibits the long memory property in its mean structure, then instead of
65 ARIMA model, Autoregressive Fractionally Integrated Moving Average (ARFIMA) model
66 [10] is used. The potential presence of long memory in a time series can be tested using
67 [11] GPH statistic. The ARIMA/ARFIMA is a linear model. To address the non-linearity of
68 a time series the Autoregressive Conditional Heteroscedastic (ARCH) [12] and
69 Generalized ARCH (GARCH) [13] models can be used as variance model along with the
70 ARIMA/ARFIMA model as mean model.

71

72 2. MATERIAL AND METHODS

73

74 2.1. ARFIMA model

75 An ARIMA model is represented as ARIMA (p, d, q) where p , d and q represents the
76 order of autoregression, integration (differencing), and moving average respectively. For
77 a linear univariate time series process $\{y_t\}$, the ARIMA process is represented as

$$78 \varphi(L)(1-L)^d y_t = \theta(L)\varepsilon_t \quad (1)$$

79 where, y_t is the actual observation and ε_t is the error term observed at time t such that
80 $\varepsilon_t \sim IID(0, \sigma^2)$. $\varphi(L)$ and $\theta(L)$ are the polynomial of lag operator L of order p and q
81 respectively. In ARIMA methodology, the order of differencing d is considered as an
82 integer. The only difference of the ARFIMA model is that here d has a fractional value.
83 For the ARFIMA models, the fractional parameter d lies between -0.5 and 0.5 [14].

84 2.2. The ARCH and GARCH model

85 A process $\{\varepsilon_t\}$ is said to have an ARCH (q) model if the conditional distribution of
86 $\{\varepsilon_t\}$ given the available information (ψ_{t-1}) up to $t-1$ time epoch can be represented as:

$$87 \varepsilon_t | \psi_{t-1} \sim N(0, h_t) \text{ and } \varepsilon_t = \sqrt{h_t} v_t \quad (2)$$

88 where, v_t is known as innovation and it is independently and identically distributed (IID)
89 with zero mean and unit variance. The distribution of innovation is data specific.

90 The conditional variance h_t for an ARCH (q) is represented as

$$91 h_t = \alpha_0 + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2, \alpha_0 > 0, \alpha_i \geq 0 \forall i \text{ and } \sum_{i=1}^q \alpha_i < 1 \quad (3)$$

92 For satisfactory model precision, a large number of parameters are needed for an ARCH
93 model. The GARCH model overcome this problem.

94 The conditional variance of a GARCH (p, q) model is defined as

$$h_t = \alpha_0 + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^p \beta_j h_{t-j}$$

$$95 \text{ provided } \alpha_0 > 0, \alpha_i \geq 0 \forall i, \beta_j \geq 0 \forall j \quad (4)$$

96 The GARCH (p, q) process is said to be weakly stationary if and only if

$$97 \sum_{i=1}^q \alpha_i + \sum_{j=1}^p \beta_j < 1 \quad (5)$$

98 2.3. Data description

99 Data for PM_{2.5} and PM₁₀ concentration ($\mu\text{g m}^{-3}$) was collected at 24 hours interval from
100 Central Pollution Control Board (CPCB) website ([https://app.cpcbcr.com/ccr/#/caaqm-](https://app.cpcbcr.com/ccr/#/caaqm-dash-board-all/caaqm-landing/caaqm-comparison-data)
101 [dash-board-all/caaqm-landing/caaqm-comparison-data](https://app.cpcbcr.com/ccr/#/caaqm-dash-board-all/caaqm-landing/caaqm-comparison-data)) for three Delhi Pollution Control
102 Committee (DPCC)-regulated air quality monitoring stations situated at Narela, Okhla
103 Phase-II and Pusa Road. These three sampling locations represent the northern (Narela),
104 central (Pusa) and southern (Okhla Phase-II) Delhi. The major source of pollution in
105 Narela and Okhla Phase-II are industrial emission and that in Pusa is vehicular emission.
106 Locations of these three weather monitoring stations have been shown in Figure 1. The
107 collected data spanned over the period of 01 July 2018 to 31 July 2022 (1492
108 datapoints). The whole data set is divided into two sets namely model building set and
109 model validation set. The last ten observations are used for the validation set and the
110 remaining portion as the model building set.

111

Mean	112.56	234.64	101.55	214.95	95.07	203.05
Median	85.35	209.33	68.06	191.86	66.20	188.77
Minimum	6.62	20.48	6.29	15.52	3.44	9.97
Maximum	689.1	717.35	601.80	884.8	570.61	726.86
S.D.	87.63	131.40	90.00	128.90	81.33	120.55
C.V. (%)	77.85	56.00	88.62	59.97	85.54	59.37
Skewness	1.55	0.72	1.75	0.93	1.68	0.73
Kurtosis	3.02	-0.04	3.41	0.8	3.29	0.34

148



Figure 2. Time plot of PM_{2.5} and PM₁₀ concentration for the selected locations

149
150
151
152
153
154
155
156
157
158

The Shapiro-Wilk test [15] is used to determine if the selected series are normally distributed. For the Shapiro-Wilk test, the null hypothesis is H_0 : The series follows normal distribution; against the alternative hypothesis H_1 : The series does not follow normal distribution. It is seen that (Table 2) none of the selected series follow normal distribution at 1% level of significance. Hence, we reject the null hypothesis. The generalized error distribution (GED) is a robust class of distribution. It is considered that all the series follow GED and also the innovations.

Table 2. Test for normality (Shapiro-Wilk test)

Site	Narela		Okhla		Pusa	
	PM _{2.5}	PM ₁₀	PM _{2.5}	PM ₁₀	PM _{2.5}	PM ₁₀
Test statistic	0.857	0.950	0.813	0.937	0.832	0.955
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

159
160
161
162
163

The stationarity of the time-series data is tested using Augmented Dickey-Fuller (ADF) test [16] and Phillips-Perron (PP) test [17]. For ADF and PP tests, the null hypothesis is H_0 : Unit root is present in the time series; against the alternative hypothesis H_1 : Unit root is not present in the time series. From the Table 3, it can be seen that for all the selected

164
165
166
167

series both the tests are significant. Hence, we reject the null hypothesis. Both the tests confirm (Table 3) that all the selected series are stationary.

Table 3. Test for stationarity of the selected series

Site	Narela		Okhla		Pusa	
Test	PM _{2.5}	PM ₁₀	PM _{2.5}	PM ₁₀	PM _{2.5}	PM ₁₀
ADF	-3.922 (0.01)	-4.429 (0.01)	-3.654 (0.03)	-3.944 (0.01)	-3.587 (0.03)	-3.829 (0.02)
PP	-9.686 (0.01)	-10.890 (0.01)	-9.468 (0.01)	-10.307 (0.01)	-8.811 (0.01)	-9.449 (0.01)

168
169
170
171
172
173
174
175
176
177

p-value is in parenthesis

The statistical dependencies among the realizations of a time series can be inspected by using the **Autocorrelation Function (ACF)** and **Partial Autocorrelation Function (PACF)** plots. The ACF plots (up to 100 lags) of the selected series are given in Figure 3. From this figure, it can be seen that the ACFs are significant for a large lags. This is known as the hyperbolic decay of ACF. This clearly indicates the presence of long memory in the mean structure. Hence, the ARFIMA model is applied to the data set, instead of the ARIMA model.

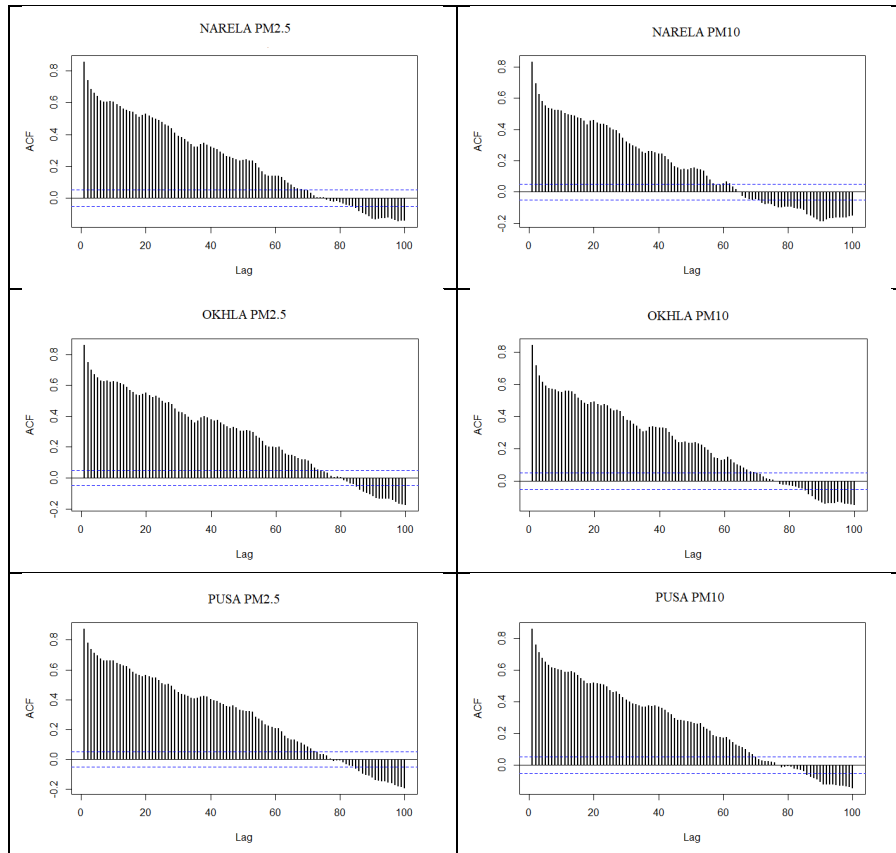


Figure 3. The ACF plots of the selected series

178
179
180
181
182
183
184
185
186
187

Various orders of autoregression and moving average of the ARFIMA (p, d, q) model are applied to each series. The residuals are obtained and are tested using the ARCH-LM test for the potential presence of conditional heteroscedasticity. **For the ARCH-LM test, the null hypothesis is H_0 : The ARCH effect is not present in the residual series; against the alternative hypothesis H_1 : The ARCH effect is present in the residual series. For each instances, the ARCH-LM test is significant. Hence, we reject the null hypothesis.** After that the GARCH(1,1) model is fitted to the residual series and the best performed order of autoregression and moving average is chosen based on minimum values of Akaike

188 Information Criterion (AIC) and Bayesian information Criterion (BIC). The estimated
 189 parameters of the selected ARFIMA (p, d, q)- GARCH(1,1) models are given in Table 4. It
 190 can be seen that almost all the estimated parameters are significant at 1% level of
 191 significance. The visualization of observed vs. fitted values is given in Figure 4. The
 192 residual series are tested for the presence variability that can be explained further. It is
 193 seen that all the residual series are white noise (WN). This proved that the selected
 194 ARFIMA (p, d, q)- GARCH(1,1) models are appropriate for forecasting the data under
 195 study. The forecast efficiency of the selected models is validated (Table 5) in the model
 196 validation set using three error functions namely RMSE, MAE and MAPE. It is seen that
 197 the values of these error functions are within permissible limit. Hence, the performances
 198 of the selected models are satisfactory.
 199
 200

Table 4. Estimate of parameters of the selected models

	Narela		Okhla		Pusa	
	PM _{2.5}	PM ₁₀	PM _{2.5}	PM ₁₀	PM _{2.5}	PM ₁₀
Model	ARFIMA (1,d,2) - GARCH (1,1)	ARFIMA (2,d,2) - GARCH (1,1)	ARFIMA (1,d,2) - GARCH (1,1)	ARFIMA (1,d,2) - GARCH (1,1)	ARFIMA (1,d,2) - GARCH (1,1)	ARFIMA (2,d,1) - GARCH (1,1)
Variable						
Mean Model						
Constant	34.770 (3.622)* **	103.945 (12.221)** *	26.059 (2.579)***	100.803 (7.831)***	34.451 (2.702)***	126.451 (4.495)***
AR(1)	0.987 (0.015)* **	1.411 (0.009)***	0.990 (0.005)***	0.983 (0.011)***	0.988 (0.005)***	1.346 (0.017)***
AR(2)		-0.419 (0.008)***				-0.356 (0.005)***
MA(1)	-0.703 (0.057)* **	-0.762 (0.045)***	-0.598 (0.013)***	-0.660 (0.032)***	-0.659 (0.018)***	-0.930 (0.052)***
MA(2)	-0.236 (0.024)* **	-0.125 (0.022)***	-0.323 (0.014)***	-0.276 (0.016)***	-0.266 (0.018)***	
d	0.459 (0.131)* **	0.147 (0.042)***	0.386 (0.029)***	0.485 (0.055)***	0.415 (0.032)***	0.391 (0.039)***
Variance Model						
Constant	12.430 (4.407)* **	46.669 (21.206)**	17.645 (4.937)***	73.193 (27.636)** *	14.947 (4.476)***	62.191 (21.912)***
α_1	0.147 (0.021)* **	0.108 (0.018)***	0.205 (0.029)***	0.161 (0.027)***	0.211 (0.033)***	0.145 (0.028)***
β_1	0.852 (0.019)* **	0.891 (0.017)***	0.794 (0.025)***	0.838 (0.023)***	0.788 (0.028)***	0.853 (0.023)***

201 ***p<0.01, **p<0.05, *p<0.10; S.E. is in parenthesis
 202

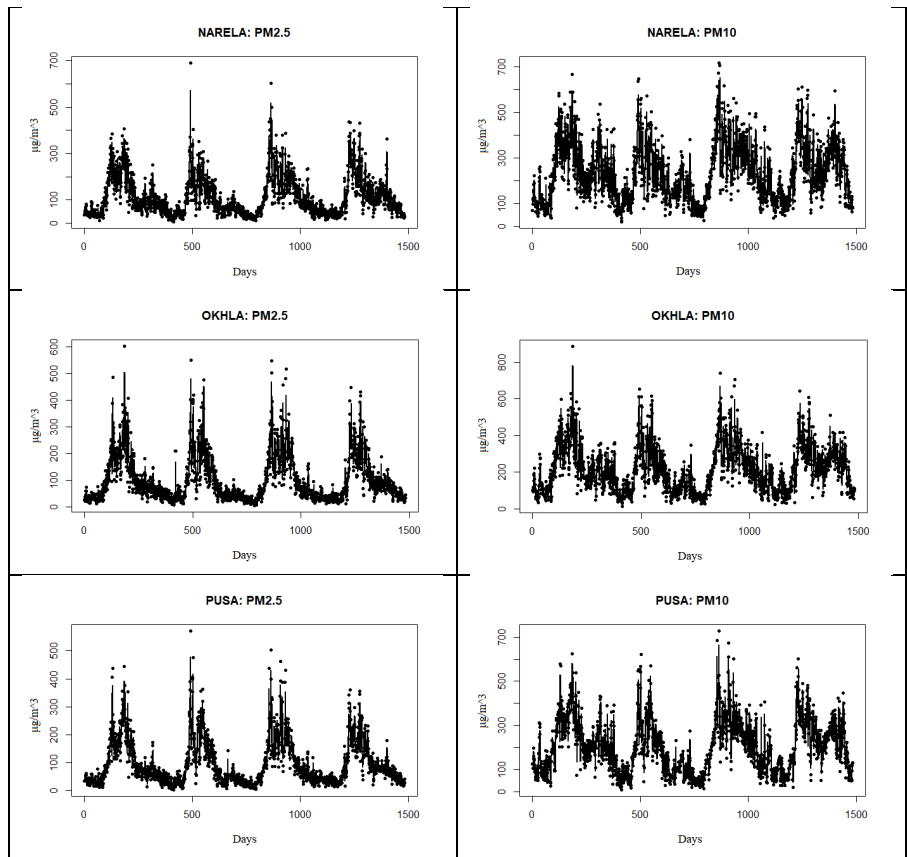


Figure 4. Observed (marker) vs. fitted (line) of the selected series

Table 5. Forecasting performance of fitted models in model validation set

Site		Model	RMSE	MAE	MAPE (%)
Narela	PM _{2.5}	ARFIMA (1,d,2) -GARCH (1,1)	15.411	14.045	8.268
	PM ₁₀	ARFIMA (2,d,2) -GARCH(1,1)	40.290	35.496	6.779
Okhla	PM _{2.5}	ARFIMA (1,d,2) -GARCH (1,1)	16.834	16.045	8.639
	PM ₁₀	ARFIMA (1,d,2) -GARCH (1,1)	41.775	39.029	6.428
Pusa	PM _{2.5}	ARFIMA (1,d,2) -GARCH (1,1)	14.677	14.128	7.928
	PM ₁₀	ARFIMA (2,d,1) -GARCH (1,1)	50.594	44.329	8.773

4. CONCLUSION

In this paper, an attempt has been made to model the variation of PM_{2.5} and PM₁₀ concentration for three selected location in Delhi. Due to the presence of long memory in mean structure, the ARFIMA model is applied as the mean model and the GARCH model as the variance model. Modeling can be helpful for understanding the fluctuation of concentrate of these two pollutants. Apart from the year-round measures already in place, the peak period of air pollution in Delhi, from late-October to early-January, demands special attention of policymakers. Few of the proactive measures that can substantially alleviate the PM-pollution are banning the burning of stubbles and firecrackers, installation of smog towers, and being more stringent about vehicular emissions.

219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274

COMPETING INTERESTS

Authors have no competing interests.

AUTHORS' CONTRIBUTIONS

Author DR and AR planned the experiment. Author DR performed the analysis and wrote the paper. Author AR interpreted the outcome and wrote the paper. Author KA, SA and V helped in data collection and paper writing.

REFERENCES

- [1]. Das M, Das A, Sarkar R, Mandal P, Saha S, Ghosh S. Exploring short term spatio-temporal pattern of PM_{2.5} and PM₁₀ and their relationship with meteorological parameters during COVID-19 in Delhi. *Urban Climate* 2021;39:100944.
- [2]. Vohra K, Vodonos A, Schwartz J, Marais EA, Sulprizio MP, Mickley LJ. Global mortality from outdoor fine particle pollution generated by fossil fuel combustion: Results from GEOS-Chem. *Environmental Research* 2021;195:110754.
- [3]. Dimitrova A, Marois G, Kiesewetter G, Samir KC, Rafaj P, Tonne C. Health impacts of fine particles under climate change mitigation, air quality control, and demographic change in India. *Environmental Research Letters* 2021;16(5):054025.
- [4]. IQAir. World's most polluted countries and regions (historical data 2018–2021). 2021a. Available in <https://www.iqair.com/world-most-polluted-countries>. (Last accessed on 14th July, 2022)
- [5]. WHO. Ambient (outdoor) air pollution. 2021. Available in [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health). (Last accessed on 14th July, 2022)
- [6]. Institute for Health Metrics and Evaluation (IHME). What risk factors drive the most death and disability combined? 2019. Available in <https://www.healthdata.org/india>. (Last accessed on 14th July, 2022)
- [7]. Pandey A, Brauer M, Cropper ML, Balakrishnan K, Mathur P, Dey S, Turkgulu B, Kumar GA, Khare M, Beig G, Gupta T. Health and economic impact of air pollution in the states of India: the Global Burden of Disease Study 2019. *The Lancet Planetary Health* 2021;5(1):25–38.
- [8]. IQAir. World's most polluted cities (historical data 2017–2021). 2021b. Available in <https://www.iqair.com/in-en/world-most-polluted-cities>. (Last accessed on 14th July, 2022)
- [9]. Box GEP, Jenkins G. Time Series Analysis, Forecasting and Control; Holden-Day: San Francisco, CA, USA, 1970.
- [10]. Granger CWJ, Joyeux R. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* 1980;4:221–238.
- [11]. Geweke J, Porter-Hudak S. The estimation and application of long-memory time series models. *Journal of Time Series Analysis* 1983;4:221–238.
- [12]. Engle RF. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* 1982; 50(4):987–1007.
- [13]. Bollerslev T. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics* 1986;31(3):307–327.
- [14]. Hosking JRM. Fractional differencing. *Biometrika* 1981;68:165–176.
- [15]. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika* 1965;52(3/4):591–611.
- [16]. Dickey D, Fuller W. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of American Statistical Association* 1979;74:427–431.
- [17]. Phillips PCB, Perron P. Testing for a unit root in time series regression. *Biometrika* 1988;75(2):335–346.