International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(3): 96-104
© 2024 Stats & Maths
https://www.mathsjournal.com
Received: 04-03-2024
Accepted: 05-04-2024

MN Megeri
Professor, Department of Statistics, Karnatak Science College Dharwad, Karnataka, India

Keerthi Astagimath
Associate Professor, Department of Statistics, Karnatak Science College Dharwad, Karnataka, India

Corresponding Author:
MN Megeri
Professor, Department of Statistics, Karnatak Science College Dharwad, Karnataka, India

An application of theoretical probability distributions to the study of ambient air pollutants in Bandra, Mumbai

MN Megeri and Keerthi Astagimath

Abstract

Rapid industrialization and population growth especially in the last decade have adversely affected urban climate, air quality and caused imbalances in the regional climate at large. There have been very few studies on the Concentration of Air Pollutants (SO_2, NO_2 and RSPM) in urban area. The distribution of air pollution parameters has an effect on the human health. In the present study, the concentration distributions of air pollution parameters were studied with respect to different seasons for one year (October 1, 2012 to September 30, 2013). The data collected for the study is from the Maharashtra Pollution Control Board (www. http://mpcb.gov.in) for Mumbai city specific to Bandra Station. Probability density functions (pdf) have been used in the analysis of the distribution of pollutant data for examining the frequency of high concentration events and also the "goodness-of-fit" of the probability density function, to the data, was evaluated, using Kolmogorov-Smirnov test. The evaluation was conducted for one year with respect to different seasons. The results of the study indicates that the best fit-distributions for air pollution parameters RSPM, SO_2 and NO_2 concentrations in Mumbai (Bandra) were extreme value, Weibull and Pearson V distributions for monsoon season. In winter and summer seasons the fit distributions were Weibull distribution for RSPM and NO_2 and Beta Distribution for SO_2. It concludes that the air pollution parameters studied (RSPM, SO_2 and NO_2) had the different statistical distributions with respect to seasons. The difference might be due to the different diffusion characteristics of individual pollutant in the air, and the interaction of diffusion characteristics and local geography, weather conditions in Mumbai (Bandra). The results can be further applied to prediction of air pollution parameters.

Keywords: Urbanization, Air pollution, RSPM, Kolmogorov-Smirnov, Weibull, Pearson V

1. Introduction

Air pollution is a universal phenomenon but, is a special case varies from area to area [9]. Rapid industrialization and population growth especially in the last decade have adversely affected urban climate, air quality and caused imbalances in the regional climate at large extent. In Mumbai, one of the largest cities in India, Sulfur dioxide (SO_2), Nitrogen oxides (NO_x) and Respirable Suspended Particulate Matter (RSPM) are the measure air pollutants and regularly monitored. During the past decades, Mumbai has undergone the most rapid development and urbanization in the history, and ambient air pollution in Mumbai has gradually changed. As shown in Figure.1.1, the trend of the air pollution (e.g. SO_2, NO_x) in Bandra, Mumbai has improved substantially from 2007 to2012, while another air pollutant factor (RSPM) became a serious public health concern in the meantime [5].

Probability distribution functions (pdf's) have been used in the analysis of air pollution data, for examining the concentration of air pollutant parameters [6-9]. Treating air pollution as stochastic process, concentration of air pollutants is usually treated as random variables with measurable statistical properties. Although "it is largely agreed that there is no a priori reason to expect that atmospheric distribution should adhere to a specific probability distribution" [10]. If certain conditions are follows, the statistical characteristics of air pollutant concentration can be described through statistical distributions. The correctly chosen distribution can help us to predict the frequency that exceeds the ambient air quality standard (AQS). For example, by making use of the direct association between emission level and some parameters (e.g., the location parameter) of the statistical distributions, Lu (2002) successfully predicted the probabilities exceeding the air quality standard and emission source reduction of PM_{10} concentration to meet the air quality standard in Taiwan [5, 8], applied some statistical distributions to determine the best statistical distribution of concentration data of major air pollutants in Shanghai, China, the results shows that the best fit distributions for PM_{10}, SO_2 and NO2 concentrations were lognormal, Pearson 5 and Extreme value distributions.

In this study, the distribution concentration of SO_2, NO_x and RSPM between the period October 1, 2012 to September 30, 2013 in Bandra, Mumbai were analyzed to simulate the frequency distribution and to estimate the parameters of the selected distributions and the main objective was to fit the seasonal concentration data to determine the optimal shape of the concentration distribution of ambient air pollutants.

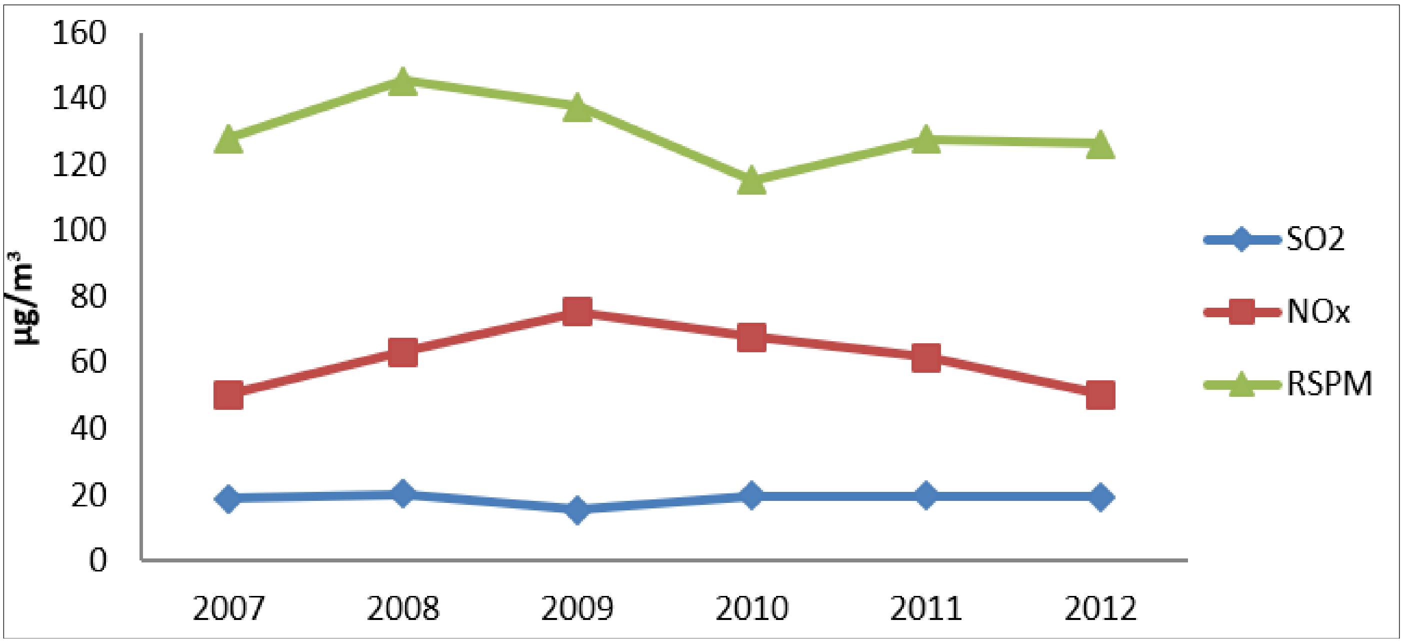


Fig 1: Annual average concentrations of SO_2, NO_x and RSPM in Bandra, Mumbai (2007-2012).

2. Methods and Materials

In the present study, the concentration distributions of air pollution parameters were studied with respect to different seasons for one year (October 1, 2012 to September 30, 2013). The data collected for the study is from the Maharashtra Pollution Control Board (*www. http://mpcb.gov.in*) for Mumbai city specific to Bandra Station. Seven probability density functions (pdf) *viz.* Beta, Extreme value, Gamma, Lognormal, Pearson type V and VI and Weibull distributions, have been used in the analysis of the distribution of pollutant data, in ordered to determine the shape of the concentration distribution and also the "goodness-of-fit" of the probability density function, to the data, was evaluated, using Kolmogorov-Smirnov test.

2.1. Statistical Distributions

The obtained ambient air pollution parameters concentrations (x) were fitted to seven selected statistical distributions, for different seasons for a period separately. The examined distributions are described as follows:

2.1.1 Beta Distribution

The pdf is given by: $f(x) = \frac{1}{B(\alpha_1,\alpha_2)} \frac{(x-a)^{\alpha_1-1}(b-x)^{\alpha_2-1}}{(b-a)^{\alpha_1+\alpha_2-1}}$, $a \le x \le b$; $\alpha_1 > 0$; $\alpha_2 > 0$; $a < b$ (1)

Where α_1 and α_2 are the scale and shape parameters of the distribution, [a, b] is the concentration range and B is the beta function and is given by

$$B(\alpha_1, \alpha_2) = \frac{\Gamma\alpha_1\Gamma\alpha_2}{\Gamma(\alpha_1+\alpha_2)} \tag{2}$$

2.1.2 Gen. Extreme Value: The pdf is given by:

$$f(x) = \begin{cases} \frac{1}{\sigma} e^{-[(1+kz)^{-1/k}(1+kz)^{-1-1/k}]} & k \neq 0 \\ \frac{1}{\sigma} e^{(-z-e^{-z})} & k = 0 \end{cases}, -\infty<x<\infty+ \text{ for } k=0 \text{ and } 1+\frac{k(x-\mu)}{\sigma}>0 \text{ for } k\neq 0 \tag{3}$$

Where k and σ (>0) are shape and scale parameters and μ is the location parameters and $z=\frac{x-\mu}{\sigma}$

2.1.3 Gamma Distribution: The pdf is given by

$$f(x) = \frac{(x-\gamma)^{\lambda-1}}{\sigma^{\lambda}\Gamma(\lambda)} e^{\left[-\left(\frac{x-\gamma}{\sigma}\right)\right]} , x \geq \gamma; \sigma>0; \lambda>0; \gamma\geq 0 \quad (4)$$

Where σ and λ are the scale and shape parameters of the distribution, γ is the location parameter and Γ is the gamma function and is given by,

$$\Gamma\lambda = \int_0^{\infty} t^{\lambda-1} e^{-t}\, dt \text{ (Eulerian integral form)} \quad (5)$$

2.1.4 Lognormal Distribution: The pdf is given by

$$f(x) = \frac{1}{\sqrt{2\pi}(x-\gamma)\sigma} e^{\left[-\left[\frac{(\ln(x-\gamma)-\mu)^2}{2\sigma^2}\right]\right]}, x>\gamma; -\infty<\mu<\infty; \sigma>0; \gamma\geq 0 \quad (6)$$

Where μ and σ are the scale and shape parameters of the distribution representing the geometric mean and standard geometric deviation respectively, while γ is the location parameter. Fitting data to a lognormal distribution evaluates the assumption that the logarithmic transformed data values follow a Gaussian distribution.

2.1.5 Pearson type V: The pdf is given by

$$f(x) = \frac{\beta^{\alpha} e^{\left[-\frac{\beta}{(x-\gamma)}\right]}}{\Gamma\alpha(x-\gamma)^{\alpha+1}} , x\geq\gamma; \alpha>0; \beta>0; \gamma\geq 0 \quad (7)$$

Where Γ is the Gamma function, α and β are the scale and shape parameters and γ is the location parameter.

2.1.6 Pearson type VI: The pdf is given by

$$f(x) = \frac{((x-\gamma)/\beta)^{\alpha_1-1}}{\beta B(\alpha_1,\alpha_2)(1+(x-\gamma)/\beta)^{\alpha_1+\alpha_1}} , x\geq\gamma; \beta>0; \alpha_1>0; \alpha_2>0; \gamma\geq 0 \quad (8)$$

Where B is the beta function α_1 and α_2 are the shape parameters, β is the scale parameter and γ is the location parameter.

2.1.7 Weibull Distribution: The pdf is given by

$$\mathbf{f(x)} = \left(\frac{\alpha}{\beta}\right)\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} e^{\left[-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right]} , x \geq \gamma; \alpha > 0; \beta > 0; \gamma \geq 0 \quad (9)$$

Where α and β are the shape and scale parameters of the distribution, and γ is the location parameter. If α=1 the Weibull distribution is identical with the Gamma distribution.

2.2. Parameter Estimation

The main purpose of the parameter estimation is to determine the specific nature of the theoretical distribution by the particular values of their parameters. The optimal values of the scale and shape parameters of the distribution were estimated using the method of maximum likelihood. The reasoning behind this method is that among the various possible values of θ the most likely value should be one that makes the probability (or probability density) of the observed **x** as high as possible [3]. This method is proposed by geneticist/statistician Sir Ronald A. Fisher around 1922.

This method helps us to calculate θ_k parameters of a k-parameter distribution in order to maximize the likelihood function L(θ).

$$L(\theta) = L(x_1, x_2,\ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta_1, \theta_2, \ldots, \theta_k) \quad (10)$$

where $(x_1, x_2,\ldots, x_n)$ are the independent observations from a random sample deriving from a population following a distribution described by a k- parameter probability density function $f(x;\theta_1, \theta_2, \ldots, \theta_k)$.

If $f(x_1; \theta_1, \theta_2, \ldots, \theta_k)$, $f(x_2;\theta_1, \theta_2, \ldots, \theta_k)$, …, $f(x_n;\theta_1, \theta_2, \ldots, \theta_k)$ are the probability functions of each of the sample values then the maximum likelihood function describes the joint probability function of the random sample.

The likelihood function being differentiable at θ_1, $\theta_{2\ldots}$ θ_k the estimation of the parameters with the maximum likelihood method is made by taking the partial derivatives of L(θ) by each parameter and solving the resulting k- equations to zero. Usually, computations are made using the logarithm of the likelihood function, and since the logarithm is also a strictly increasing function the same parameters values will maximize both the likelihood and the log-likelihood functions.

$$\frac{\partial \ln L(\theta)}{\partial \theta_k} = 0 \quad (11)$$

The method of maximum likelihood is considered advantageous for parameter estimation over the method of moments because, it usually does not yield “good” estimators. Therefore, the method of maximum likelihood is intuitively used, because we attempt to find the values of the true parameters that would have most likely produced the data that we in fact observed. For most cases of practical interest, the performance of maximum likelihood estimators is optimal for large enough data. This is one of the most versatile methods for fitting parametric statistical models to data. On the other hand, since the method of maximum likelihood requires sample processing power due to the complex numerical calculations involved, when large data sets are analyzed, computational time increases substantially. The location parameter *γ* was set to zero for the continuous distribution functions since it was desired that concentrations would exhibit behavior with physical meaning (it should be noted that accepting a certain value as a global or regional particle concentration background is rather difficult and will not be attempted). The upper bound *b* for the beta distribution was set to 300 μg/m^3 for RSPM since these concentration levels have never been reached in the Bandra station, Mumbai, on a daily basis, even in the occurrence of severe episodic conditions. The optimum scale parameters for the Weibull distribution were determined using an iterative trial and error process [1].

2.3. Goodness-of-fit

For the evaluation of the above presented probability functions was made using the Kolmogorov-Smirnov (K-S) goodness-of-fit test is used. The K-S statistic is based on the maximum difference between the hypothesized cumulative distribution function $F_0(x)$ and the empirical distribution function of the samples $S_n(x)$. Symbolically, it is given by

$$D_n = \max |F_0(x)-S_n(x)| \qquad (12)$$

The D_n value is compared with the $D_{(n,\alpha)}$ value, which is the largest difference acceptable at the α- level of significance for n- sized samples. If $D_n < D_{(n,a)}$ the hypothesis that the sample can be described by the fitted theoretical distribution is accepted at the a significance level[4].

3. Results and Discussion

3.1. Variability of measured data with respect to seasons

Table.1 to 3 summarizes the descriptive statistics of SO_2, NO_2, and RSPM concentration data in Bandra, Mumbai from 2012-20013. The daily SO_2, NO_2, and RSPM concentration variability with time is shown in Fig.2 to 5. From Table.1, we can see that the ambient air pollutant parameters had seasonal variability. Compare with the National Ambient Air Quality Standards, which was 80, 80 and 100 μg/m^3 for daily average concentration of SO_2, NO_2, and RSPM respectively, the frequency of exceedance with the standard were higher for RSPM than those for SO_2 and NO_2, suggesting that particulate air pollution has become the major environmental problem in Bandra, Mumbai.

Table 1: Descriptive Statistics with respective season for SO_2, NO_x and RSPM concentration for monsoon period in Bandra, Mumbai.

Parameter	Monsoon (μg/m^3)						
	Average	S.D	Min	P(25)	Median	P(75)	Max
SO_2	19.03	2.12	11	18	19	20	29
NO_2	48.78	13.33	21	44.75	49	53	96
RSPM	61.72	25.83	30	45	56	73	176

Source: As figure.1

In the Monsoon season, on an average, all the three parameter SO_2, NO_2 and RSPM is below the standard level of pollution but maximum value of NO_2 is above the standard value and that of RSPM approaching the standard value and it is also observed that there will be more dispersion as compare to other parameters (See Table 1).

Table 2: Descriptive Statistics with respective season for SO_2, NO_x and RSPM concentration for winter period in Bandra, Mumbai.

Parameter	Winter (μg/m^3)						
	Average	S.D	Min	P(25)	Median	P(75)	Max
SO_2	18.46	6.75	4	16	18	19	48
NO_2	50.37	10.96	21	48	52	57	66
RSPM	161.27	94.94	61	101	148	190.75	**797**

Source: *As figure.1*

The analysis of Table 3 shows that on an average value of SO_2 , NO_2 and RSPM is below the standard value but lot of variation is observed data and RSPM is attained highest value

Table 3: Descriptive Statistics with respective season for SO_2, NO_x and RSPM concentration for summer period in Bandra, Mumbai.

Parameters	Summer (μg/m^3)						
	Average	S.D	Min	P(25)	Median	P(75)	Max
SO2	18.33	2.4	13	17	18	19	25
NO_2	32.51	12.93	9	22	30	45.3	63
RSPM	118.73	44.32	32	88	109	147	**261**

Source: *As figure.1*

AS compare other two seasons *Viz* Monsoon season and Winter season, summer season value below the standard level with the exception that the maximum value of RSPM is quite above the standard level (see Table 3.3). The dispersion of data also shows there will be little variation as compare to other two seasons.

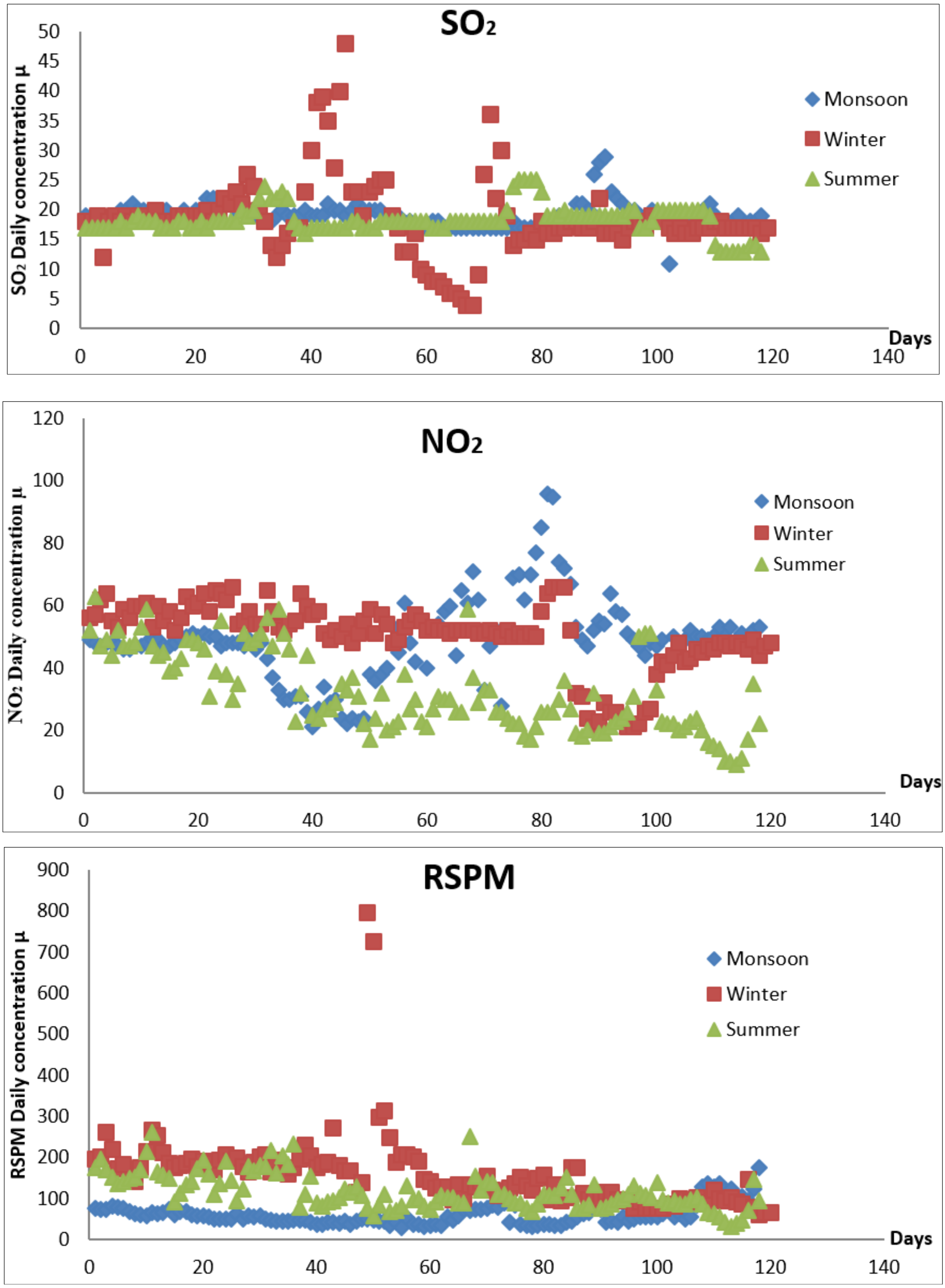


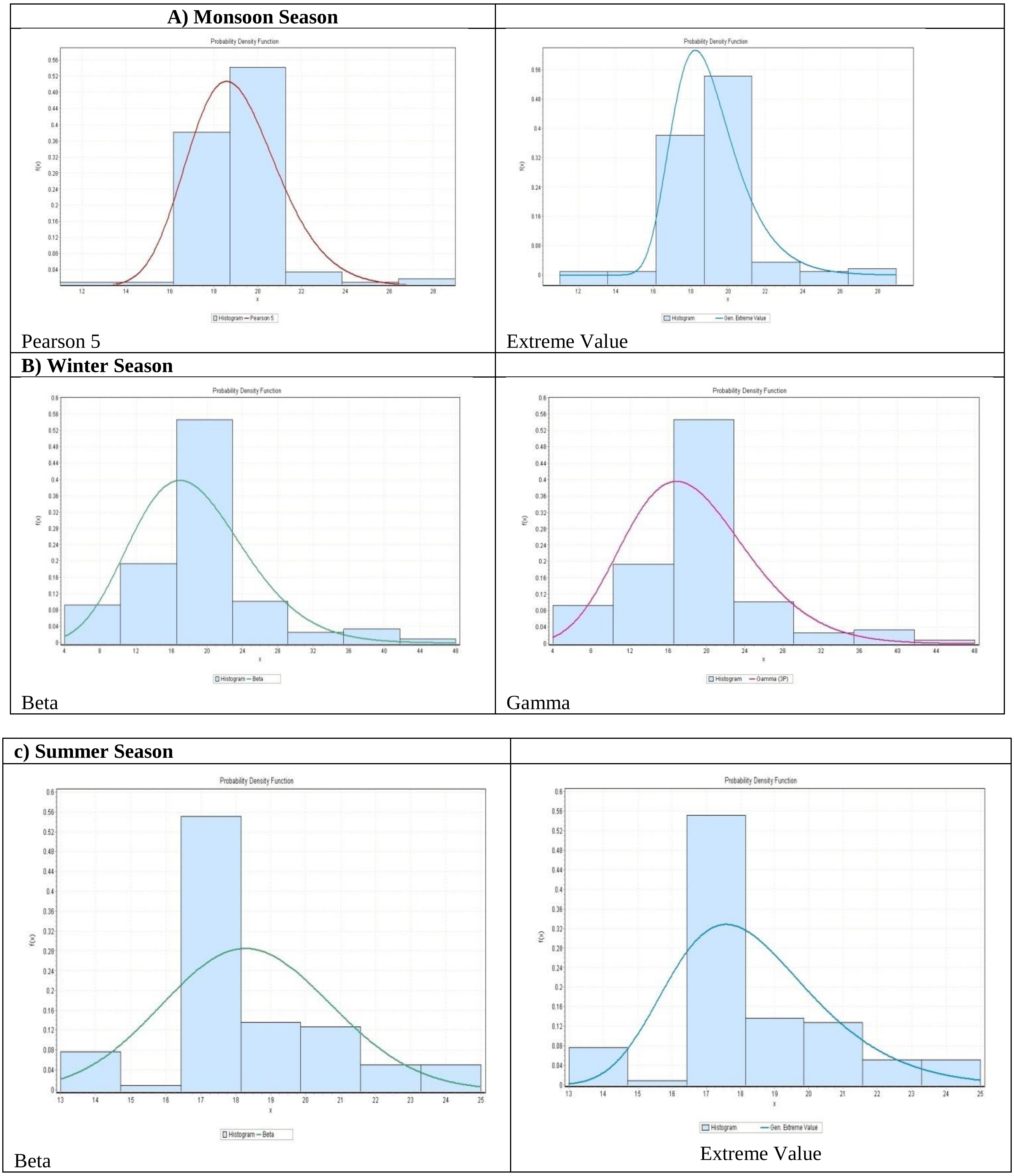
Fig 2 to 5: Daily concentration of SO_2, NO_x and RSPM for a period in Bandra, Mumbai.

Source: As figure.1

3.2. Fitting of selected distributions

In this section, we compared fitting results and parameter estimation of seven theoretical distributions namely, Beta, Extreme value, Gamma, Lognormal, Pearson type V and VI and Weibull distributions for air pollution parameters *viz*. SO_2, NO_2, and RSPM for different seasons respectively.

The results of the "goodness-of-fit" test for the distributions are shown in Table.4 to 6. Based on the Kolmogorov-Smirnov test, the results shows that, the best-fit distributions of SO_2, NO_2 and RSPM during the period monsoon were Pearson type 5, Extreme value and Weibull distributions, for winter season, the best-fit distributions were Beta for SO_2, Weibull for NO_2, RSPM and for summer season, the best-fit distributions were Beta for SO_2, Weibull for NO_2, RSPM respectively. Here the best- fit distribution referred to the one with Kolmogorov-Smirnov test value.



A) Monsoon Season

Pearson 5

Extreme Value

B) Winter Season

Beta

Gamma

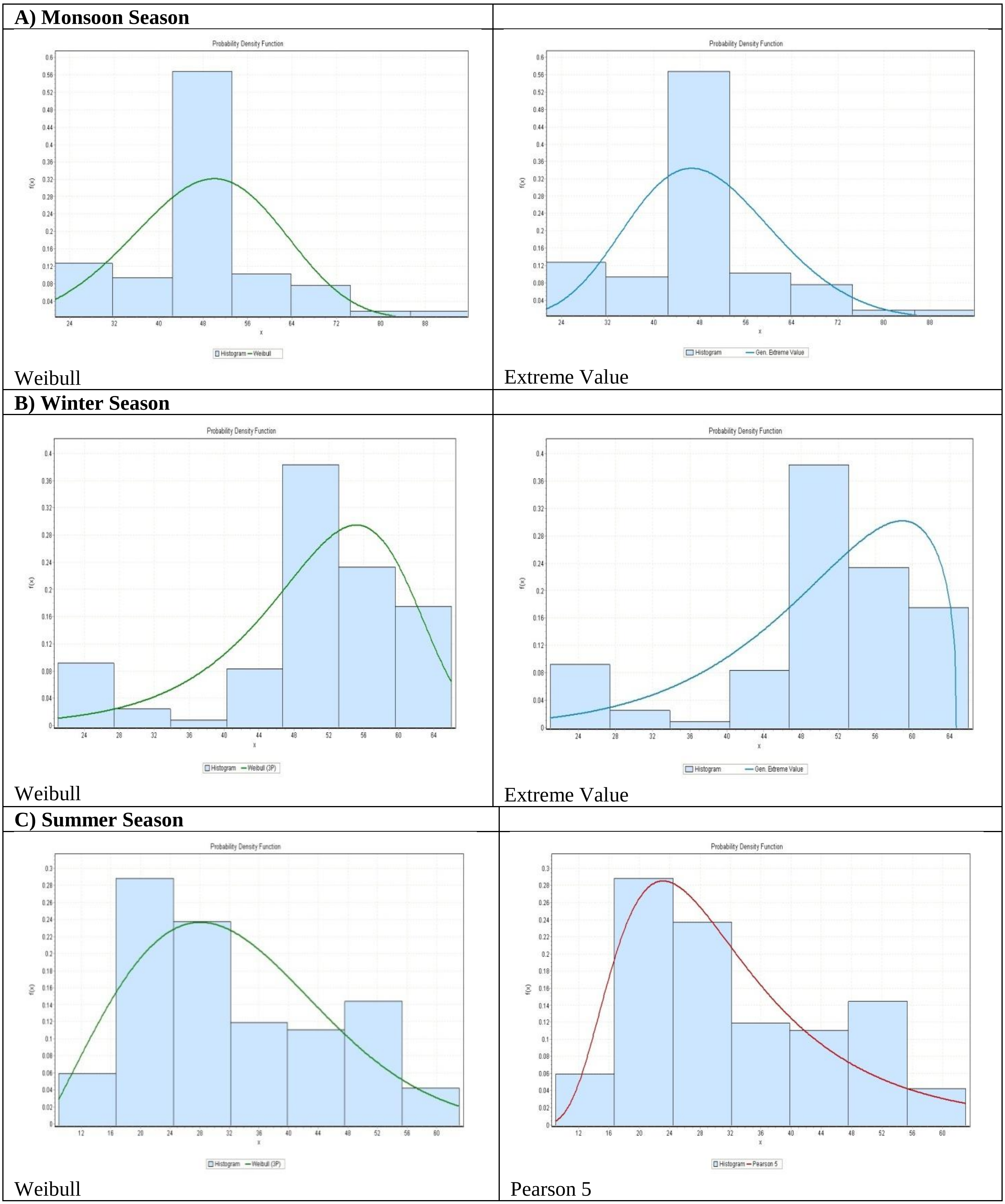
c) Summer Season

Beta

Extreme Value

Fig 6: Comparison of different statistical distributions of SO_2 concentration.

Note: X-axis is the SO_2 concentration, µg/m^3; Y-axis is the frequency value.



A) Monsoon Season

Weibull

Extreme Value

B) Winter Season

Weibull

Extreme Value

C) Summer Season

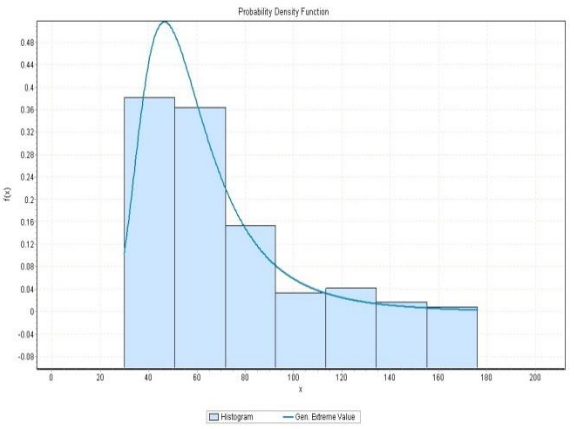
Weibull

Pearson 5

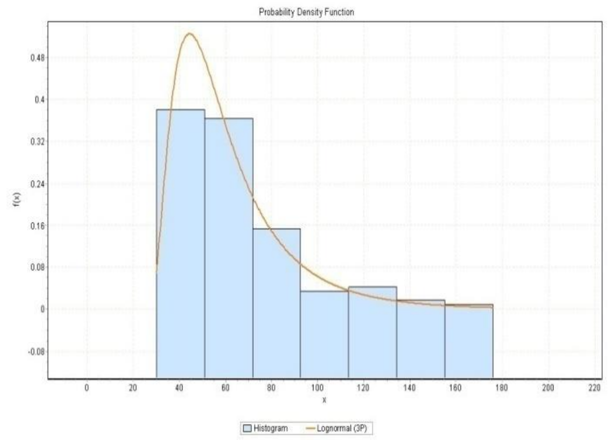
Note: X-axis is the NO_2 concentration, µg/m^3; Y-axis is the frequency value.

Fig 7: Comparison of different statistical distributions of NO_2 concentration

A) Monsoon Season

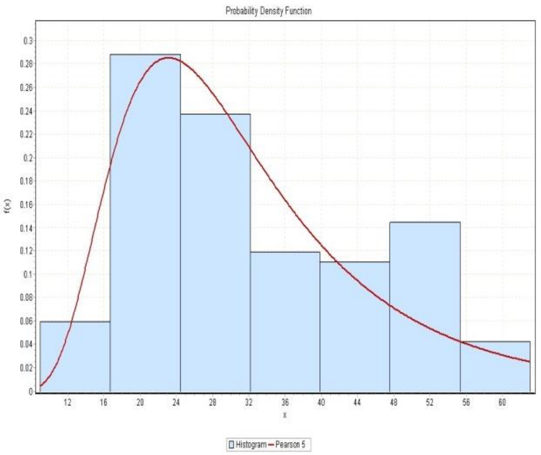


Extreme Value

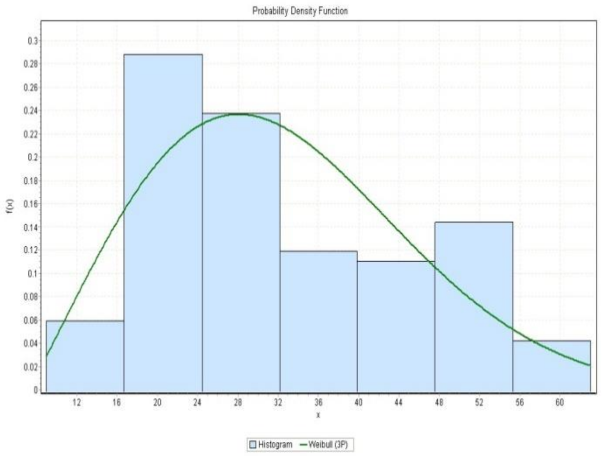


Lognormal

B) Winter Season

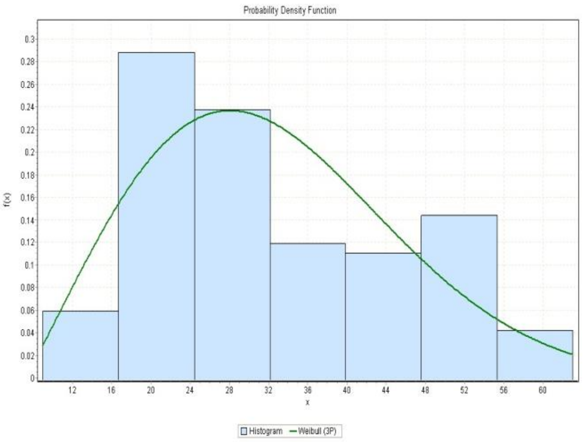


Weibull

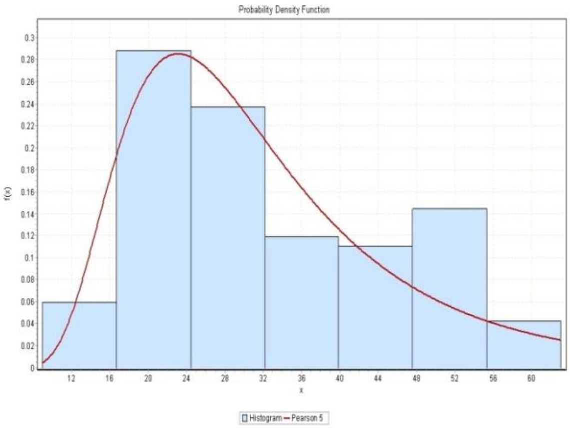


Pearson 5

C) Summer Season



Weibull



Pearson 5

Note: X-axis is the RSPM concentration, µg/m^3; Y-axis is the frequency value.

Table 4: Fitted Distribution type and "goodness-of-fit" Statistics Using the Kolmogorov-Smirnov Test for Monsoon season.

Distribution	SO$_2$	NO$_2$	RSPM
Beta	0.1636	0.1897	0.1146
Gamma	0.1685	0.1889	0.0540
Gen. Extreme value	0.1561	0.1836	0.0356*
Lognormal	0.1585	0.1901	0.3866
Pearson type 5	0.1545*	0.1907	0.0404
Pearson type 6	0.1576	0.1901	0.0399
Weibull	0.1953	0.1652*	0.0628

Note: * The best fitted distribution.

For checking the goodness of fit, Kolmogorov Smirnov test was used and results shows SO_2 best fitted probability distribution is Pearson type 5, for NO_2 Weibull and General Extreme value distribution fits well for RSPM (see Table 5). For winter season Beta distribution is best fitted for SO_2 and both NO_2 and RSPM Weibull distribution is best fitted (See Table 5).

Table 5: Fitted Distribution type and "goodness-of-fit" Statistics Using the Kolmogorov-Smirnov Test for winter season.

Distribution	SO_2	NO_2	RSPM
Beta	0.1928*	0.1923	0.0969
Gamma	0.2130	0.1850	0.1003
Gen. Extreme value	0.1989	0.1073	0.0990
Lognormal	0.1955	0.1775	0.0998
Pearson type 5	0.1962	0.1996	0.0944
Pearson type 6	0.1967	0.1748	0.1050
Weibull	0.2128	0.0966*	0.0943*

Note: * The best fitted distribution.

In summer season, again similar trend of fitting of probability distribution as we have observed in case of Winter seasons (See Table 6). In summary, Weibull probability distribution seems to be best fitted distribution all the three seasons NO_2 and RSPM.

Table 6: Fitted Distribution type and "goodness-of-fit" Statistics Using the Kolmogorov-Smirnov Test for Summer season.

Distribution	SO_2	NO_2	RSPM
Beta	0.2040*	0.0969	0.0969
Gamma	0.2156	0.1031	0.1003
Gen. Extreme value	0.2107	0.0989	0.0989
Lognormal	0.2132	0.0997	0.0996
Pearson type 5	0.2150	0.1040	0.0944
Pearson type 6	0.2136	0.1050	0.1050
Weibull	0.2207	0.0943*	0.0943*

Note: * The best fitted distribution.

4. Conclusion

In the current study, seven selected probability distribution functions were fitted to air pollutant parameters *viz*. SO_2, NO_2, and RSPM concentration data measured for a year with respective to different seasons at Bandra station, Mumbai. The study shows on an average all the parameter for three different seasons below the standard level even the maximum value of NO2 and RSPM is above the standard level. From the Kolmogorov-Smirnov goodness-of-fit test, the most appropriate probability density functions were the Weibull for NO_2 and RSPM for all the tree seasons. Beta distribution is best fitted in Monsoon and Winter seasons for SO_2 and for NO_2 Pearson type 5 distributions best fitted in summer season. The present analysis shows that the best performing statistical distributions of various air pollutants for different seasons in Bandra, Mumbai are different. And we conclude on the suitability of continuous, positively skewed distributions for describing the air pollutant parameter data in areas with increased concentration levels.

References

1. Gavriil I, Grivas G, Kassomenos P, Chaloulakou A, Spyrellis N. An application of theoretical probability distributions to the study of PM10 and PM2.5 time series in Athens, Greece. Atmospheric Environment. 2006;40(7):1212-1222. doi:10.1016/j.atmosenv.2005.09.081.
2. Georgopoulos PG, Seinfeld JH. Statistical distribution of air pollutant concentration. Environ Sci Technol. 1982;16(7):401A-416A. doi:10.1021/es00104a001.
3. Schay G. Introduction to Probability with Statistical Applications. Birkhäuser; c2007.
4. Gibbons JD, Chakraborti S. Nonparametric Statistical Inference. 4th ed. Marcel Dekker, Inc; c2003.
5. Kan H-D, Chen B-H. Statistical distributions of ambient air pollutants in Shanghai, China. Biomed Environ Sci. 2004;17(4):366-372.
6. Kao AS, Friedlander SK. Frequency distribution of PM10 chemical components and their source. Environ Sci Technol. 1995;29(1):19-28. doi:10.1021/es00001a003.
7. Larsen RI. An air quality data analysis system for interrelating effects, standards, and need source reductions. J Air Pollut Control Assoc. 1973;23(11):933-940. doi:10.1080/00022470.1973.10469882.
8. Lu H. The statistical character of PM10 concentration in Taiwan area. Atmos Environ. 2002;36(3):491-502.
9. Morel B, Yeh S, Cifuentes L. Statistical distributions for air pollution applied to the study of the particulate problem in Santiago. Atmos Environ. 1999;33(16):2575-2585. doi:10.1016/S1352-2310(98)00384-1.
10. Seinfeld J, Pandis SN. Atmospheric Chemistry and Physics. Wiley; c1998.