Particulate Matter Trends In Alwar: An Application Of Anova And Kruskal-Wallis Test

Shalini jain, Dr. V. L. Mandowara

ABSTRACT: In this paper, an attempt has been made for analysing and finding trend in the data of Air quality factor Particulate matter (PM_{10}) of Alwar city of Rajasthan State for a period of 8 years (2010 to 2017). The analysis covers 3 air quality monitoring stations in Alwar. The study area covers a substantial portion of Alwar district. First, a normality check of the distribution of data series values has been performed using Shapiro wilk test as well as graphical methods and findings suggested that for years 2015 and 2017 data is distributed non-normally and for rest of the years data is distributed normally. But due to these two years whole data series is affected. For Testing homogeneity of variances, Levene test has been used and presence of outliers has also been detected using Box plots. Second, from the class of Parametric tests Analysis of variance technique (ANOVA) has been applied for testing the homogeneity of several means year wise and as well as Kruskal Wallis test has been applied being a non parametric test for the same. As a finding, no increasing or decreasing trend is present in the monthly values of PM_{10} but there is a difference in yearly mean PM_{10} values. Both parametric test and non parametric test has given the same result that there is significant difference between the mean values of the pollutant PM_{10} year wise.

KEYWORDS: Alwar, ANOVA Test, Kruskal Wallis Test, Normality Test, PM_{10}.2010 Mathematics Subject Classification: CC:62P12

1. INTRODUCTION

According to 2017 Global Burden of Disease (GBD) report India has the second highest number of early deaths due to $PM_{2.5}$ in the world. Early deaths due to ozone pollution are the highest in India. More than a quarter of global deaths due to air pollution occur in India alone. With a large and ageing population in the grip of rising and toxic pollution, India stands at serious health risk. The analysis shows that long-term exposure to fine particulate matter has contributed to 4.2 million premature deaths and to a loss of 103 million healthy years of life globally in 2015. This is the fifth highest risk factor for deaths from heart disease and stroke, lung cancer, chronic lung disease, and respiratory infections. An additional 254,000 deaths occur due to exposure to ozone and its impact on chronic lung disease. PM2.5 was responsible for a substantially larger number of deaths than other well-known risk factors.

Early deaths in India are also the highest among the South Asian countries – it is 9 times higher than neighbouring Bangladesh and 8 times higher than Pakistan. This is an extremely serious issue – in the past, the GBD estimates had indicated that air pollution was the fifth largest killer in India (Organization). By 2030, the expected growth in many of the developing sectors (industries, residential, transportation, power generation, and construction) will result in an increase in pollution related health problems for most of the cities, mainly in cities which are emerging to metro cities. The available information on urban air pollution, their sources, and the potential of various interventions to control pollution, should help us propose a cleaner path to 2030. Fifteen of the top 20 most polluted cities in the world are located in India, according to an analysis of air quality in several cities around the world.

- *Shalini jain, Dr. V. L. Mandowara*
- *Research Scholar, Department of Mathematics & Statistics, University College of Science, Mohanlal Sukhadia University, Udaipur, Rajasthan-313001, India*
- *Professor(Rtd), Department of Mathematics & Statistics, University College of Science, Mohanlal Sukhadia University, Udaipur, Rajasthan-313001, India*

Gurugram, in **Haryana,** topped the list with an average annual particulate matter (PM $_{2.5}$) quality of 135 g/m^3 (micrograms/cubic metre), in 2018. Delhi - a frequent fixture on global pollution hotspots -was only the 11th most noxious city behind Lahore, Pakistan (10th) and Hotan, China (8th). The other cities in India that made the list of 20 were Ghaziabad, Faridabad, Bhiwadi, Noida, Patna, Lucknow, Jodhpur, Muzaffarpur, Varanasi, Moradabad, Agra, Gaya and Jind (Air Quality Index). As cities are increasing in size and population, there is a steady demand for motorized vehicles in both personal and public trans- port sectors. This puts substantial pressure on the city's infra- structure and environment, particularly since most Indian cites have mixed land use. Sharma (2016) made an attempt to investigate and ascertain air quality on major roads of Ajmer city in the form of Air Quality Index (AQI). The results reveal that gaseous pollutants such as SO_2 and NOx are within the permissible limits and particulate matter is the predominant cause of air pollution in the study area (Sharma). Deepkumar B Rathi et al.(2016) attempted a statistical analysis of ambient air in Aurangabad City (Maharastra.State) during the year 2005-2010 and concentrations of SO_2, NO_X, RSPM and SPM were studied. Results shows that SO_2, NO_X, RSPM are well below the permissible limit and SPM is above the permissible limit(B Rathi and Rathod). Mikalai Filonchyk et al.(2016) investigated hourly mass concentration of $PM_{2.5}$ and PM_{10} from June 1 to August 31, 2015 collected in the 11 largest cities of Gansu Province. This study has shown that the overall average concentrations of $PM_{2.5}$ and PM_{10} in the study area were 26 and 66 μg/m^3. In PM $_{2.5}$ episode days (when concentration was more than 75 μg/m3 for 24 hrs), the average concentrations of $PM_{2.5}$ was 2–3 times higher as compared to non-episode days. There were no observed clear differences during the weekday/weekend PM and other air pollutants (SO_2, NO_2, CO and O_3) in all the investigated cities (Filonchyk et al.). Marcelina Arkouli et al.(2010) studied temporal and horizontal distribution of the $PM_{2.5}$ and PM_{10} concentrations in Buenos Aires city .The measurements were carried out for one year from May 2006 to May 2007 .The values and temporal variations on a daily and seasonal basis were consistent with the ventilation potential of the atmosphere. The correlation between PM_{10} and $PM_{2.5}$ concentrations and frequencies of wind directions showed that the highest concentration were observed when the wind was from land wind and lowest concentrations when the wind was from fluvial wind. The ratio of $PM_{2.5}$ to PM_{10} was 0.44, which indicates the coarse particles (>2.5 μm) originated from road dust, soil re suspension and abrasion processes are the dominated fractions of PM. Results of random PM measurements showed that $PM_{2.5}$ was more homogeneously distributed over the city than PM_{10}(Arkouli et al.). Shiv Lal et al.(2018) described the air pollutions in six major cities in Rajasthan. It is stated that Oxides of Nitrogen, SO_2, ozone and ammonia are found under the prescribed limit for all locations. The $PM_{2.5}$ level is observed higher than the prescribed limit in Jodhpur, Jaipur, Pali, Ajmer and Alwar.(Lal and Kothari). T. Van Hecke (2010) described the comparison of the anova and the Kruskal-Wallis test by means of the power when violating the assumption about normally distributed populations. The permutation method is used as a simulation method to determine the power of the test. It was found that in the case of asymmetric populations the non-parametric Kruskal-Wallis test performs better than the parametric equivalent ANOVA tachnique(Hecke). Adefisoye JO et al.(2016) compared the performances of eighteen normality tests available in literature. Since a theoretical comparison is not possible, MonteCarlo simulation were done from various symmetric and asymmetric distributions for different sample sizes ranging from 10 to 1000. The performance of the test statistics are compared based on empirical Type I error rate and power of the test. The simulations results show that the Kurtosis Test is the most powerful for symmetric data and Shapiro Wilk test is the most powerful for asymmetric data(JO et al.).

2. CONCEPTS AND METHODS

2.1 Geographical and physical features of Study Area:

The Alwar District is situated in the north-east of Rajasthan state between 27°34' and 28°4' north latitudes and 76°7' and 77°13' east longitudes. The total area of district is 8380 Sq.Kms. It's greatest length from south to north is about 137 Kms. and greatest breadth from east to west is about 110 Kms. It is surrounded by the Aravali hills from all sides which protect it from the sandy and hot winds coming from the Thar Desert. Recently, Alwar city has been declared to be a part of the national capital region (NCR) because of which there has been a considerable development in the district. (http://alwar.rajasthan.gov.in/content/raj/alwar/en/home.html)

2.2 The data set:

The data used in this study consists of daily observations of PM_{10} for the period of 2010-2017. The maximum daily values were obtained by extracting the maximum pollutant levels of a dataset of hourly concentrations through a 24 hour period. Air pollutant concentrations are expressed in microgram per cubic meter (μg/m^3).

2.3 Air monitoring infrastructure :

The daily measurements of air quality parameter (daily average concentrations of pollutant) PM_{10} used in the present study was measured by CPCB, a regulatory monitoring agency in Delhi. This network consists of three air quality monitoring stations in Alwar city of Rajasthan state (mentioned in Table 1) which have been operated by Central pollution control board (CPCB), Ministry of Environment and Forest, providing hourly values for a number of Air Quality (AQ) parameters.

Table 1 : Air quality monitoring stations

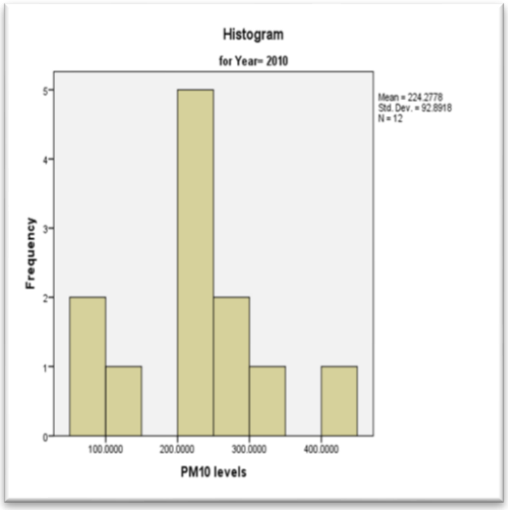
Sr.No.	Station Code	Station Name	Category
1	373	Gaurav Solvex	Industrial
2	372	Regional office	Residential/Commercial
3	219	RICCO Pump House, MIA	Industrial

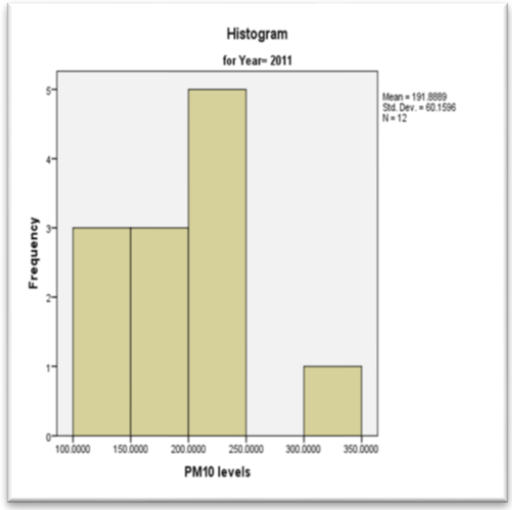
3. STATISTICAL ANALYSIS

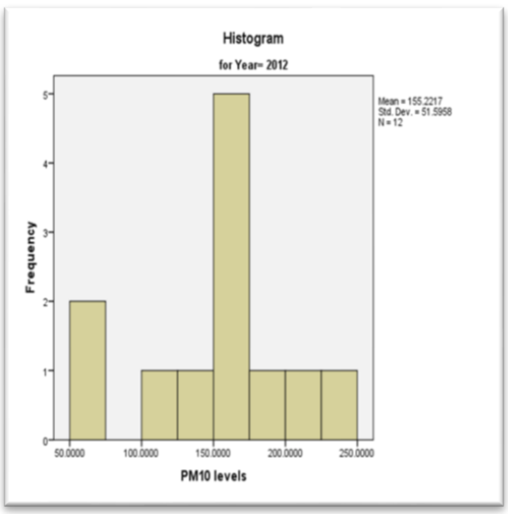
3.1 Normality Check:

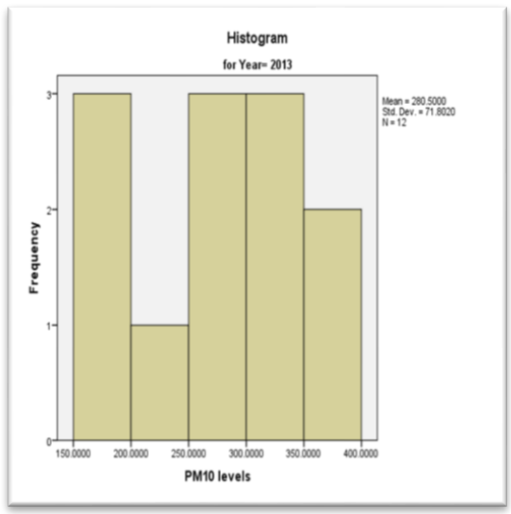
In statistics it is conventional to assume that the observations are normal. The entire statistical framework is grounded on this assumption and if this assumption is violated the inference breaks down. For this reason it is essential to check or test this assumption before any statistical analysis of data. The most commonly used graphical methods for normality check are Histograms and Q-Q plots. Keya Rani Das et al.(2016) provided a brief review of commonly used tests for normality using both graphical and analytical tests (Rani Das).

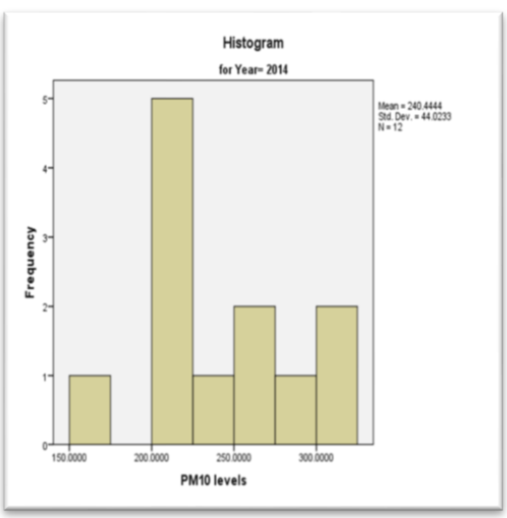
3.1.1 Histograms

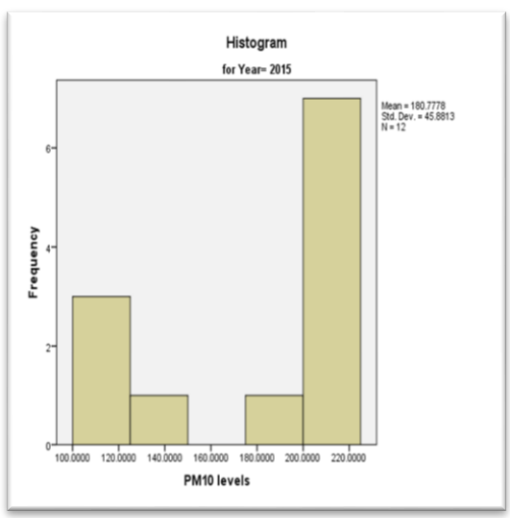

Histogram
for Year= 2010
Mean = 224.2778
Std. Dev. = 92.8918
N = 12
Frequency
PM10 levels


Histogram
for Year= 2011
Mean = 191.8889
Std. Dev. = 60.1596
N = 12
Frequency
PM10 levels


Histogram
for Year= 2012
Mean = 155.2217
Std. Dev. = 51.5958
N = 12
Frequency
PM10 levels


Histogram
for Year= 2013
Mean = 280.5000
Std. Dev. = 71.8020
N = 12
Frequency
PM10 levels


Histogram
for Year= 2014
Mean = 240.4444
Std. Dev. = 44.0233
N = 12
Frequency
PM10 levels


Histogram
for Year= 2015
Mean = 180.7778
Std. Dev. = 45.8813
N = 12
Frequency
PM10 levels

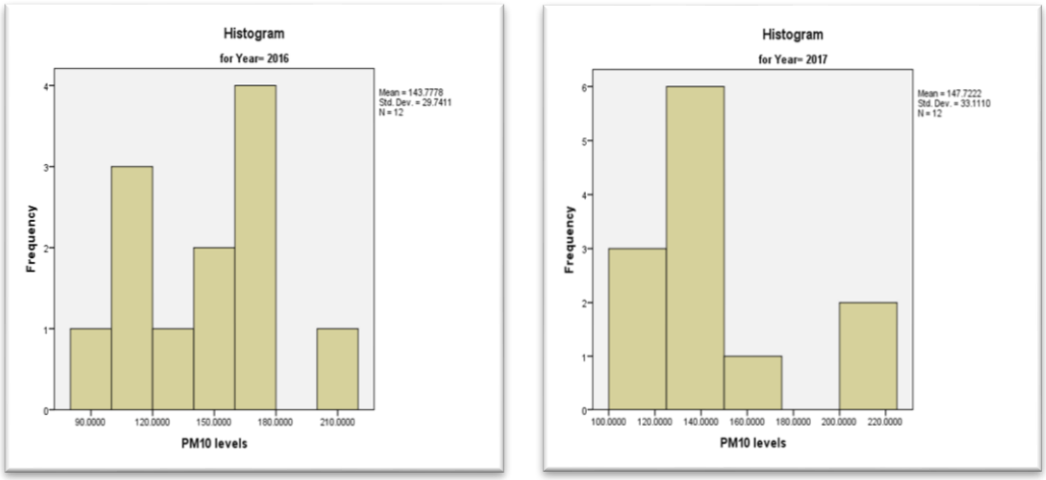


Fig 2: Histograms for years 2010-2017 PM_{10} values

By observing Histograms for different years we see that for all the years, data is not forming a Bell Shaped curve and we found that PM10 values are non-normally distributed.

3.1.2 Q-Q Plots

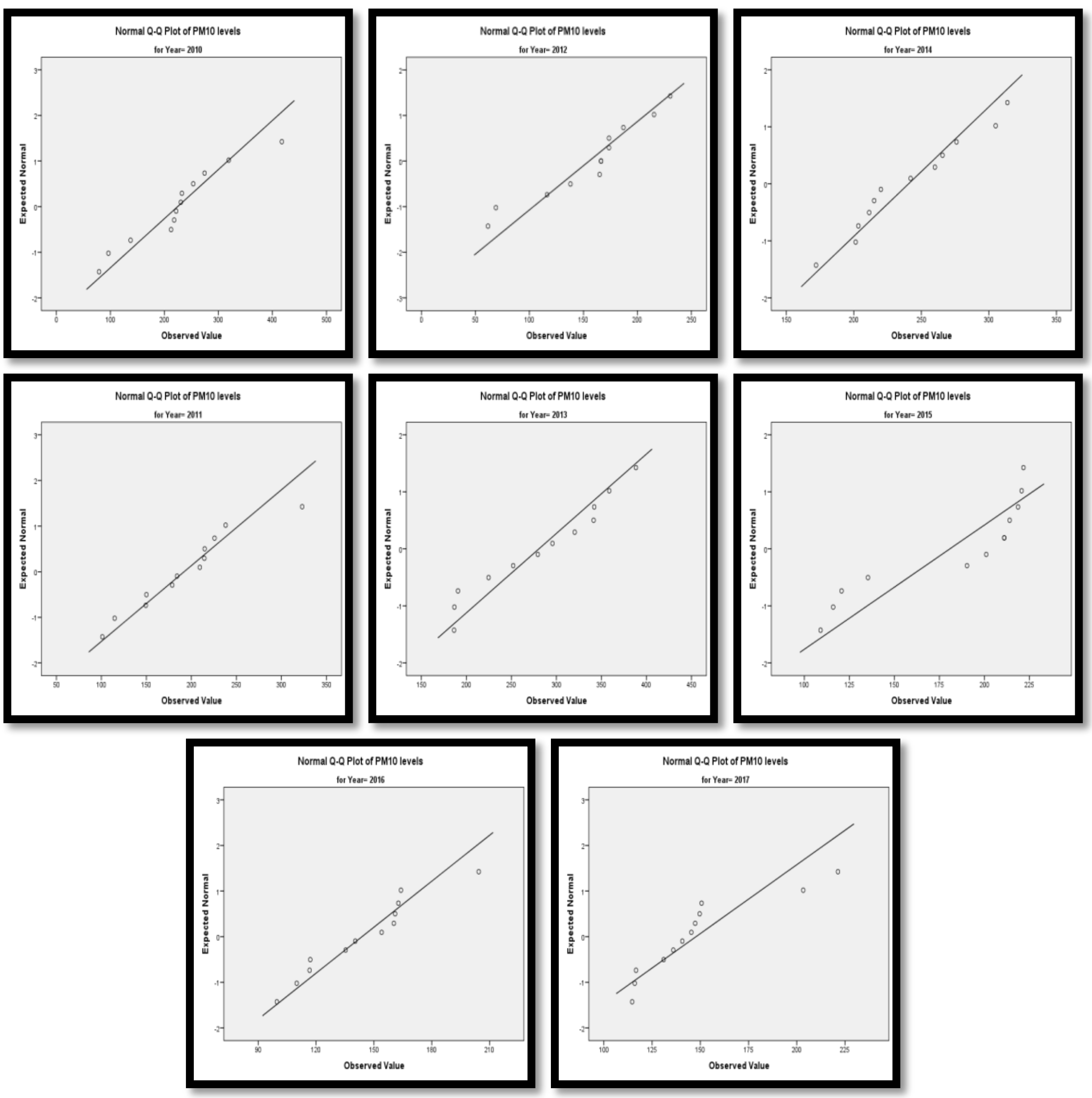


Fig 3 : Q-Q Plots for years 2010-2017

By observing Q-Q Plots we found that for the years 2013, 2015, 2016 and 2017 data is distributed non-normally and for remaining years it may follow normal distribution.

3.1.3 Shapiro-Wilk Test

The Shapiro-Wilk test is one of the most popular tests for normality assumption diagnostics which has good properties of power and it based on correlation within given observations and associated normal scores. The Shapiro-Wilk test statistic is derived by Shapiro and Wilk (1965). The form of the test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where $x_{(i)}$ is the i-th order statistics and α_i is the i-th expected value of normalized order statistics. For independently and identically distributed observations, the values of α_i can be obtained from the table presented by Shapiro and Wilk (1965) for sample sizes up to 50 (S.S. Shapiro; M.B. Wilk).

Table 2: Shapiro Wilk Test

	Year	Statistic	df	Sig.
PM10 levels	2010	.945	12	.566
	2011	.953	12	.682
	2012	.919	12	.277
	2013	.922	12	.300
	2014	.953	12	.680
	2015	.778	12	.005
	2016	.942	12	.531
	2017	.822	12	.017

It is observed from Table 2 that PM_{10} levels for years 2015 and 2017 are significant and we conclude that data for these years are non-normally distributed, otherwise for remaining years data may be distributed normally.

3.2 Test for presence of outliers :

For testing the presence of outliers Box Plots technique is most commonly used. We have drawn Box plots for different years with respect to PM_{10} values.

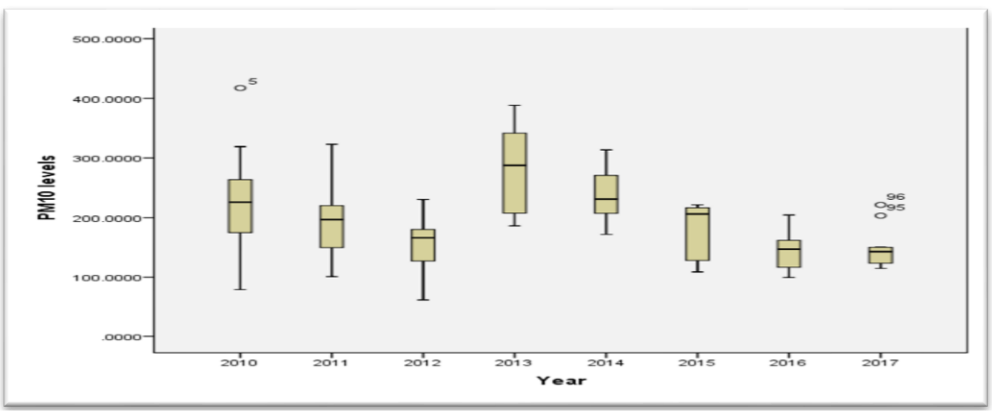


Fig 4 : Box Plots for years 2010-2017

By observing Box Plots for years 2010-2017 with respect to PM_{10} levels we found that a few outliers are present in the data in years 2010 and 2017.

3.3 Test for homogeneity of variances :

For testing homogeneity of variances, Levene statistic has been applied in Table 3 and the Sig. value (<.05) shows that there is significant difference between the variances year-wise.

Table 3 : Test of Homogeneity of Variances

Levene Statistic	df1	df2	Sig.
2.307	7	88	.033

4. RESULTS AND DISCUSSIONS

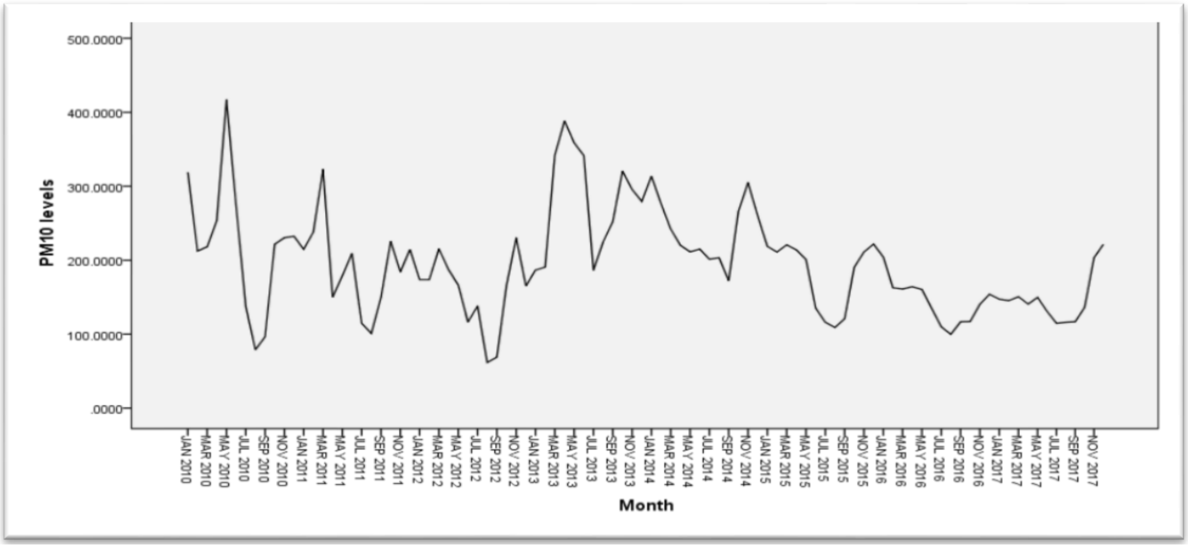


Fig 5 : Graphical presentation of PM10 values year 2010-2017

It is observed from Fig 5 that no increasing or decreasing trend is found in the monthly values of PM_{10}. Difference in mean values has been observed. We further test if this difference is statistically significant. As a finding we come to know that all the three assumptions that is, (i) Data is normally distributed, (ii) Homogeneity of variances and (iii) No outliers are present in the data, are violated, which are required to be fulfilled for application of ANOVA technique. So, we try to perform non parametric test for testing the significance of the difference between yearly mean PM_{10} levels using Kruskal -Wallis test.

4.1 Kruskal Wallis test :

The Kruskal-Wallis test is a non-parametric test, which means that it does not assume that the data come from a distribution that can be completely described by two parameters, mean and standard deviation (the way a normal distribution can). Like most non-parametric tests, it is performed on ranked data, so we convert the measurement observations to their ranks in the overall data set: the smallest value gets a rank of 1(John H. McDonald)

The formula for Kruskal-Wallis test is :
where

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{} - 3(n+1)$$

R_i is different ranks, for each of the different groups
H = Kruskal-Wallis Test statistic
N = total number of observations in all samples
The Kruskal-Wallis test statistic is approximately a chi-square distribution, with k-1 degrees of freedom where n_i should be greater than 5 (StatisticsSolutions).

Table 4 : Kruskal-Wallis Test

Ranks

Year	N	Mean Rank
2010	12	60.17
2011	12	48.29
2012	12	35.13
2013	12	77.17
2014	12	69.29
2015	12	44.21
2016	12	25.92
2017	12	27.83
Total	96	

Test Statistics[a,b]

	PM10 levels
Chi-Square	39.043
df	7
Asymp. Sig.	.000

a. Kruskal Wallis Test
b. Grouping Variable: Year

By application of Kruskal –Wallis test we found that Sig. value come out to be .000 which is <.05. And, we reject the null hypothesis of no difference and conclude that there is significant difference between year-wise mean values of PM_{10}. Now, for testing the same, we make an attempt to apply ANOVA technique, being a parametric test.

4.2 ANOVA Test:

Table 5 : ANOVA Test

PM10 levels	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	202597.793	7	28942.542	8.875	.000
Within Groups	286987.050	88	3261.216		
Total	489584.843	95			

From Table 5 it is observed that there is significant difference between yearly mean values of PM_{10} since Sig. value has come out to be .000(<.05). For year-wise multiple comparisons Tukey HSD test has been applied.

Table 6 : Multiple Comparisons
Dependent Variable: PM10 levels
Tukey HSD

(I) Year	(J) Year	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
2010	2011	32.3888889	23.3138603	.860	-40.002441	104.780219
	2012	69.0561111	23.3138603	.073	-3.335219	141.447441
	2013	-56.2222222	23.3138603	.249	-128.613552	16.169108
	2014	-16.1666667	23.3138603	.997	-88.557997	56.224663
	2015	43.5000000	23.3138603	.578	-28.891330	115.891330
	2016	80.5000000*	23.3138603	.019	8.108670	152.891330
	2017	76.5555555*	23.3138603	.030	4.164225	148.946886
2011	2010	-32.3888889	23.3138603	.860	-104.780219	40.002441
	2012	36.6672222	23.3138603	.765	-35.724108	109.058552
	2013	-88.6111111*	23.3138603	.006	-161.002441	-16.219781
	2014	-48.5555556	23.3138603	.434	-120.946886	23.835775
	2015	11.1111111	23.3138603	1.000	-61.280219	83.502441
	2016	48.1111111	23.3138603	.446	-24.280219	120.502441
	2017	44.1666667	23.3138603	.558	-28.224663	116.557997
2012	2010	-69.0561111	23.3138603	.073	-141.447441	3.335219
	2011	-36.6672222	23.3138603	.765	-109.058552	35.724108
	2013	-125.2783333*	23.3138603	.000	-197.669663	-52.887003
	2014	-85.2227778*	23.3138603	.010	-157.614108	-12.831448
	2015	-25.5561111	23.3138603	.956	-97.947441	46.835219
	2016	11.4438889	23.3138603	1.000	-60.947441	83.835219
	2017	7.4994444	23.3138603	1.000	-64.891886	79.890774
2013	2010	56.2222222	23.3138603	.249	-16.169108	128.613552
	2011	88.6111111*	23.3138603	.006	16.219781	161.002441
	2012	125.2783333*	23.3138603	.000	52.887003	197.669663
	2014	40.0555556	23.3138603	.676	-32.335775	112.446886
	2015	99.7222222*	23.3138603	.001	27.330892	172.113552
	2016	136.7222222*	23.3138603	.000	64.330892	209.113552
	2017	132.7777778*	23.3138603	.000	60.386448	205.169108
2014	2010	16.1666667	23.3138603	.997	-56.224663	88.557997
	2011	48.5555556	23.3138603	.434	-23.835775	120.946886
	2012	85.2227778*	23.3138603	.010	12.831448	157.614108
	2013	-40.0555556	23.3138603	.676	-112.446886	32.335775
	2015	59.6666667	23.3138603	.185	-12.724663	132.057997
	2016	96.6666667*	23.3138603	.002	24.275337	169.057997
	2017	92.7222222*	23.3138603	.003	20.330892	165.113552
2015	2010	-43.5000000	23.3138603	.578	-115.891330	28.891330
	2011	-11.1111111	23.3138603	1.000	-83.502441	61.280219
	2012	25.5561111	23.3138603	.956	-46.835219	97.947441
	2013	-99.7222222*	23.3138603	.001	-172.113552	-27.330892
	2014	-59.6666667	23.3138603	.185	-132.057997	12.724663

	2016	37.0000000	23.3138603	.757	-35.391330	109.391330
	2017	33.0555556	23.3138603	.847	-39.335775	105.446886
2016	2010	-80.5000000*	23.3138603	.019	-152.891330	-8.108670
	2011	-48.1111111	23.3138603	.446	-120.502441	24.280219
	2012	-11.4438889	23.3138603	1.000	-83.835219	60.947441
	2013	-136.7222222*	23.3138603	.000	-209.113552	-64.330892
	2014	-96.6666667*	23.3138603	.002	-169.057997	-24.275337
	2015	-37.0000000	23.3138603	.757	-109.391330	35.391330
	2017	-3.9444445	23.3138603	1.000	-76.335775	68.446886
2017	2010	-76.5555555*	23.3138603	.030	-148.946886	-4.164225
	2011	-44.1666667	23.3138603	.558	-116.557997	28.224663
	2012	-7.4994444	23.3138603	1.000	-79.890774	64.891886
	2013	-132.7777778*	23.3138603	.000	-205.169108	-60.386448
	2014	-92.7222222*	23.3138603	.003	-165.113552	-20.330892
	2015	-33.0555556	23.3138603	.847	-105.446886	39.335775
	2016	3.9444445	23.3138603	1.000	-68.446886	76.335775

**. The mean difference is significant at the 0.05 level.*

Table 7 : Homogeneous subsets
PM10 levels
Tukey HSD[a]

Year	N	Subset for alpha = 0.05			
		1	2	3	4
2016	12	143.777778			
2017	12	147.722222			
2012	12	155.221667	155.221667		
2015	12	180.777778	180.777778	180.777778	
2011	12	191.888889	191.888889	191.888889	
2010	12		224.277778	224.277778	224.277778
2014	12			240.444444	240.444444
2013	12				280.500000
Sig.		.446	.073	.185	.249

Means for groups in homogeneous subsets are displayed.
Uses Harmonic Mean Sample Size = 12.000.

Table 7 represents the homogeneous subsets of the years which shares the same Sig. value.

5. CONCLUSION

By analysing the monthly PM_{10} values for years 2010-2017, it is found that there is no increasing or decreasing trend in PM10 but difference in mean values is present. And further, by using both the methods for testing the significance of the difference between mean PM_{10} values for respective years we found that there is significant difference in the mean level of the pollutant PM_{10}.For this purpose, both the techniques, parametric test (ANOVA) and non parametric test (Kruskal-Wallis) has given the same result. So, we can say that ANOVA technique is not much sensitive for normality assumption and it can give same results as non-parametric tests with more Power of the test.

6. ACKNOWLEDGEMENT:

We would like to thank Central Pollution Control Board (CPCB), New Delhi for providing Air Pollution data.

REFERENCES :

[1]. Air Quality Index. World Air Quality Index. 2018, http://aqicn.org/city/mexico/mexico/camarones/.

[2]. Arkouli, Marcelina, et al. "Distribution and Temporal Behavior of Particulate Matter over the Urban Area of Buenos Aires." Atmospheric Pollution Research, vol. 1, no. 1, Elsevier, 2010, pp. 1–8, doi:10.5094/APR.2010.001.

[3]. B Rathi, Deepkumar, and Shantilal D. Rathod. "Statistical Analysis of Ambient Air Quality in Aurangabad City, Maharashtra, India." Modern Chemistry & Applications, vol. 4, no. 3, OMICS Publishing Group, Oct. 2016, doi:10.4172/2329-6798.1000184.

[4]. Filonchyk, Mikalai, et al. "A Study of PM2.5 and PM10 Concentrations in the Atmosphere of Large Cities in Gansu Province, China, in Summer Period." Journal of Earth System Science, vol. 125, no. 6, 2016, pp. 1175–87, doi:10.1007/s12040-016-0722-x.

[5]. Hecke, T. Van. “Power Study of Anova versus Kruskal-Wallis Test.” Journal of Statistics and Management Systems, vol. 15, no. 2–3, May 2010, pp. 241–47, doi:10.1080/09720510.2012.10701623.

[6]. JO, Adefisoye, et al. “Performances of Several Univariate Tests of Normality: An Empirical Study.” Journal of Biometrics & Biostatistics, vol. 07, no. 04, OMICS Publishing Group, Jan. 2016, doi:10.4172/2155-6180.1000322.

[7]. John H. McDonald. Kruskal–Wallis Test - Handbook of Biological Statistics. 3rd ed., Sparky House Publishing, Baltimore, Maryland., 2014.

[8]. Lal, Shiv, and Shulbha Kothari. Comparative Studies of Air-Pollution of Major Cities of Rajasthan , India. Vol. 6890, no. 7, 2018, pp. 506–08.

[9]. Organization, World Health. Global Burden of Disease Study 2017. 2017, pp. 1–7.

[10]. Rani Das, Keya. “A Brief Review of Tests for Normality.” American Journal of Theoretical and Applied Statistics, vol. 5, no. 1, 2016, p. 5, doi:10.11648/j.ajtas.20160501.12.

[11]. S.S. Shapiro; M.B. Wilk, Biometrika. An Analysis of Variance Test for Normality (Complete Samp1es) T. Vol. 52, no. 3, 1965, pp. 591–611.

[12]. Sharma, Dr Rashmi. “Study of Air Quality Status of Some Cities of Rajasthan.” Journal of Pharmacy and Biological Sciences, vol. 11, no. 1, 2016, pp. 36–39, doi:10.9790/3008-11113639.

[13]. StatisticsSolutions. “Kruskal-Wallis Test - Statistics Solutions.” Statistics Solutions, 2017, pp. 2–3, https://www.statisticssolutions.com/kruskal-wallis-test/.