Identifying Optimal Statistical Distributions for Air Pollution Data of Agra

Rajat Kumar Pachauri[1]*, Prof. Vineeta Singh[2], Shivam Dixit[3]

[1]Research Scholar, Department of Statistics, Institute of Social Sciences, Dr. Bhimrao Ambedkar University, Agra, Uttar Pradesh, India, Email: rajatpachauri2601@gmail.com
[2]Professor& Head, Department of Statistics, Institute of Social Sciences, Dr. Bhimrao Ambedkar University, Agra, Uttar Pradesh, India, Email: vineetasanjaisingh@gmail.com
[3]Research Scholar, Amity School of Applied Sciences, Amity University Rajasthan, Jaipur, India, Email: dixit.shivam2007@gmail.com
*Corresponding Author

Received: 18.07.2024 Revised: 12.08.2024 Accepted: 03.09.2024

ABSTRACT

High levels of air pollution pose significant risks to human health and contribute to environmental instability. Addressing these challenges requires effective tools, such as statistical air pollution modelling, which can forecast the return time of future high-level pollution episodes. This technique is also instrumental in supporting government agencies in enhancing air quality management, especially when air quality data is accurately analysed.In air pollution modelling, statistical distributions like Gamma, lognormal, and Weibull are more commonly used than other distribution models. The primary objective of this observational study is to identify the most suitable distributions for predicting air pollution levels in the city of Agra. Our findings indicate that the lognormal distribution best fits SO_2 and NO_2 levels, while the Weibull distribution is more appropriate for modelling PM_{10} and the Air Quality Index of Agra.

Keywords: Air pollution, Statistical distribution, Air pollutant prediction.

INTRODUCTION

Ambient air quality is rapidly becoming one of the most critical global issues. As the population grows, so does the demand for industrial production to meet human needs, including motor vehicles and power generation. Consequently, air pollution has escalated into a worldwide concern [1]. Research shows that air pollution has severe long-term and short-term effects on human health [2], contributing to approximately 800,000 deaths globally each year [3].

The statistics are even more alarming; according to Kim et al. [4], exposure to PM_{10} not only impacts health but is also linked to an estimated 2.1 million deaths annually. This underscores the profound impact air pollution can have, limiting and impairing human activities. These findings highlight that air pollution is not only harmful but can also be a silent killer.

In recent decades, air pollution has emerged as a major environmental issue, particularly in urban areas. As urban air quality deteriorates, the risks of heart disease, stroke, lung cancer, and respiratory disorders like asthma increase [5]. Short-term exposure to high pollutant levels raises the likelihood of premature death from respiratory and cardiovascular conditions. Moreover, prolonged exposure exacerbates the risk of high blood pressure, various cancers, and increases the mortality rate among tuberculosis patients [3].

Air pollution is classified into two types: primary and secondary pollutants. Primary pollutants are emitted directly from sources and remain unchanged [6], whereas secondary pollutants are formed through oxidation or other chemical reactions following the emission of primary pollutants [7]. In urban areas, industrial and vehicular activities are the primary sources of air pollution [8].

In India, the Central Pollution Control Board (CPCB) is responsible for overseeing air quality standards. State Pollution Control Boards (SPCB), Pollution Control Committees (PCC), and other reputable agencies monitor air pollutants. The CPCB collaborates with these authorities to ensure consistent and uniform air quality data, offering both technical and financial support.

The National Air Quality Monitoring Project (NAMP), managed by the CPCB, is a nationwide initiative for monitoring ambient air quality. The project utilizes both manual and continuous Ambient Air Quality Monitoring Stations across the country. Currently, the network comprises 1,257 monitoring stations. Manual monitoring is conducted at 883 stations across 378 cities/towns in 28 states and 7 union territories, while continuous monitoring takes place at 374 stations in 190 cities/towns across 27 states and 4 union territories [9]. Under NAMP, four air pollutants are regularly monitored at all locations: sulphur dioxide (SO_2), nitrogen oxides (NO_2), respirable suspended particulate matter (RSPM/PM_{10}), and fine particulate matter ($PM_{2.5}$). Meteorological data such as wind speed and direction, relative humidity (RH), and temperature are also recorded alongside air quality.

The National Air Quality Index (AQI) serves as a guideline for assessing air pollution levels in India, as outlined in Table 1. These standards help determine whether the air quality is safe or harmful. Over the past decade, the average AQI levels in many parts of India have consistently been above the recommended limits. Therefore, it is crucial to use statistical distributions to predict air pollution in India, ensuring that air quality levels remain within the boundaries set by the National Ambient Air Quality Standards [10].

Statistical modelling thus becomes an essential tool for predicting the recurrence of high-level air pollution events. This data and knowledge can assist government agencies in formulating air pollution regulations and promoting alternative energy sources to reduce emissions [11]. Additionally, it helps related government entities provide early warnings and prepare the public for potential high pollution events.

Table 1: National Air Quality Index [10]

AQI	**AQI category**
0-50	Good
51-100	Satisfactory
101-200	Moderately polluted
201-300	Poor
301-400	Very Poor
401-500	Severe

Statistical Distribution

A variety of statistical distributions can be applied to model air pollution data. Georgopoulos and Seinfeld [12] examined and reported on the statistical distribution of pollutant concentrations in air quality. Among the most commonly used distributions for fitting air pollution data are the lognormal, Weibull, and gamma distributions. However, other alternative distributions can also be employed to fit air pollutant distributions in different contexts.

In his research, Jim [13] evaluated the Weibull, Lognormal, and Pearson V distributions for the daily average PM_{10} concentrations across three locations in Taiwan: Hsin-chu, Sha-lu, and Jian-jim. He concluded that the lognormal distribution was the most appropriate for this purpose. Similarly, a study conducted in central Taiwan by Lu [14] found that when PM_{10} concentrations are high, the parent distribution deviates, making it difficult to reliably predict using the Weibull, Gamma, and Lognormal distributions. This study also identified the exponential and extreme value distributions as effective alternatives for fitting actual PM_{10} data. However, Sedek [15], analysing data from 1998 to 2002 in Kuala Lumpur, determined that the Lognormal distribution was the best-fitting model for predicting PM_{10} concentrations.

In Malaysia, Noor et al. [16] modelled PM_{10} concentrations in the industrial area of Nilai using Lognormal, Weibull, and Gamma distributions. Their findings indicated that the Lognormal distribution provided the best fit for data in 2006, while the Weibull distribution was more suitable in 2007. Hamid [17] also conducted research in the industrialized Nilai area, fitting the distribution with a Lognormal model.

A study in China by Shi et al. [18] employed statistical distributions to forecast PM_{10} concentrations and manage air quality in five Chinese cities. This study used statistical distributions like Lognormal, Weibull, and Gamma to determine the frequency of daily average PM_{10} concentrations and the necessary reduction in particulate matter emissions to meet Air Quality Standards (AQS). The Lognormal distribution was found to be a good fit for the data. In another study, Oguntunde et al. [19] identified the Gamma distribution as the best-fitting model for forecasting future air pollution concentrations in Nigeria.

Maciejewska et al. [20] used several statistical distributions to determine the return duration of extreme concentrations in Warsaw, Poland. The results showed that the exponential distribution satisfactorily described the high concentrations during winter months, while the Lognormal distribution was the best model for middle-range values.

Berthe et al. [21] found that the Gumbel Weibull distribution was effective for modelling and predicting rainfall in Mali, with their test results suggesting it was efficient at forecasting reported rainfall. Predicting air pollution concentrations is crucial for future planning, as air pollution negatively impacts both human health and environmental stability. In a separate study on statistical distributions, the Weibull distribution was used to forecast ozone concentrations in the coastal areas of Port Dickson, Selangor, and Klang [22].

Additionally, Al-Dhurafi et al. [23] examined the probability distribution model for air pollution, using Exponential, Weibull, Gamma, and Lognormal distributions to model the air pollution index (API). The study concluded that the Gamma distribution was the most closely fitted to the majority of air pollution data.

In the Alacati region of Izmir, Turkey, Ozay et al. [24] conducted studies using two-parameter Weibull and Rayleigh distributions for statistical analysis. The Weibull distribution was found to best fit the characteristics of the results, while other distributions were less suitable. Another study by Younes and Hassan [25] predicted hourly PM_{10} concentrations using Weibull, Lognormal, and Gamma distributions, further demonstrating the utility of these models in air pollution analysis.

Table 2: Application of Statistical modeling for prediction

Author(s)	**Pollutants**	**Time period**	**Prediction Area**	**The best distribution**
Lu, 2002 [13]	PM_{10}	1995 to 1999	Hsin-Chu, Sha-Lu, and Gian-Jim, Taiwan	Lognormal
Lu, 2004 [14]	PM_{10}	1994 to 1999	Central Taiwan	Exponential and extreme value
Sedek et al., 2006 [15]	PM_{10}	1998 to 2002	Kuala Lumpur,Malaysia	Lognormal
Noor et al., 2011 [16]	PM_{10}	2006 to 2007	Nilai,Malaysia	Lognormal
Hamid et al.,2013 [17]	PM_{10}	2003 to 2009	Nilai, Negeri Sembilan, Malaysia	Lognormal
Xi et al.,2013 [18]	PM_{10}	2004 to 2008	5 representative cities of China	Lognormal
Oguntunde et al., 2014 [19]	CO	August 2004 to August 2010	Lagos State, Nigeria	Gamma
Maciejewska et al.,2015 [20]	Black Carbon	July 2012 to June 2014	Warsow,Poland	Lognormal
Berthe et al.,2015 [21]	Rainfall	1949 to 2006	Mali	Gumbel Weibull
Nasir et al., 2016 [22]	Ozone Concentration	2012	Port Klang, Selangor and Port Dickson, Negeri Sembilan, Malaysia	Weibull
Al-Dhurafi et al., 2016 [23]	CO, SO_2 , NO_2, PM_{10}	2007 to 2011	KualaLumpur, Malaysia	Gamma
Ozayet al.,2016 [24]	Wind Speed	September 2008 to March 2014	Alacati region, Turkey	Weibull
Yunus and Hasan, 2017 [25]	PM_{10}	2014 to 2015	Seberang Perai and Petalin Malaysia	Lognormal

Methodology

In this study, we collected monthly air pollution data from UPPCB monitoring stations in Agra, spanning from January 2017 to December 2023. We extracted average concentrations of air pollutants (NO_2, SO_2, and PM_{10}) and the Air Quality Index (AQI) from all these monitoring stations. This data was then fitted to Weibull, Gamma, and Lognormal distributions.

Using R software, we generated histograms and QQ plots for each pollutant within each distribution. To assess the goodness of fit, we analysed the data using these distributions and calculated the log-likelihood value, a key indicator of how well a regression model fits the dataset. A higher log-likelihood value indicates a better fit between the model and the data, with values ranging from negative to positive infinity. By comparing the log-likelihood values across different models and distributions, we identified which distribution is best suited for each pollutant.

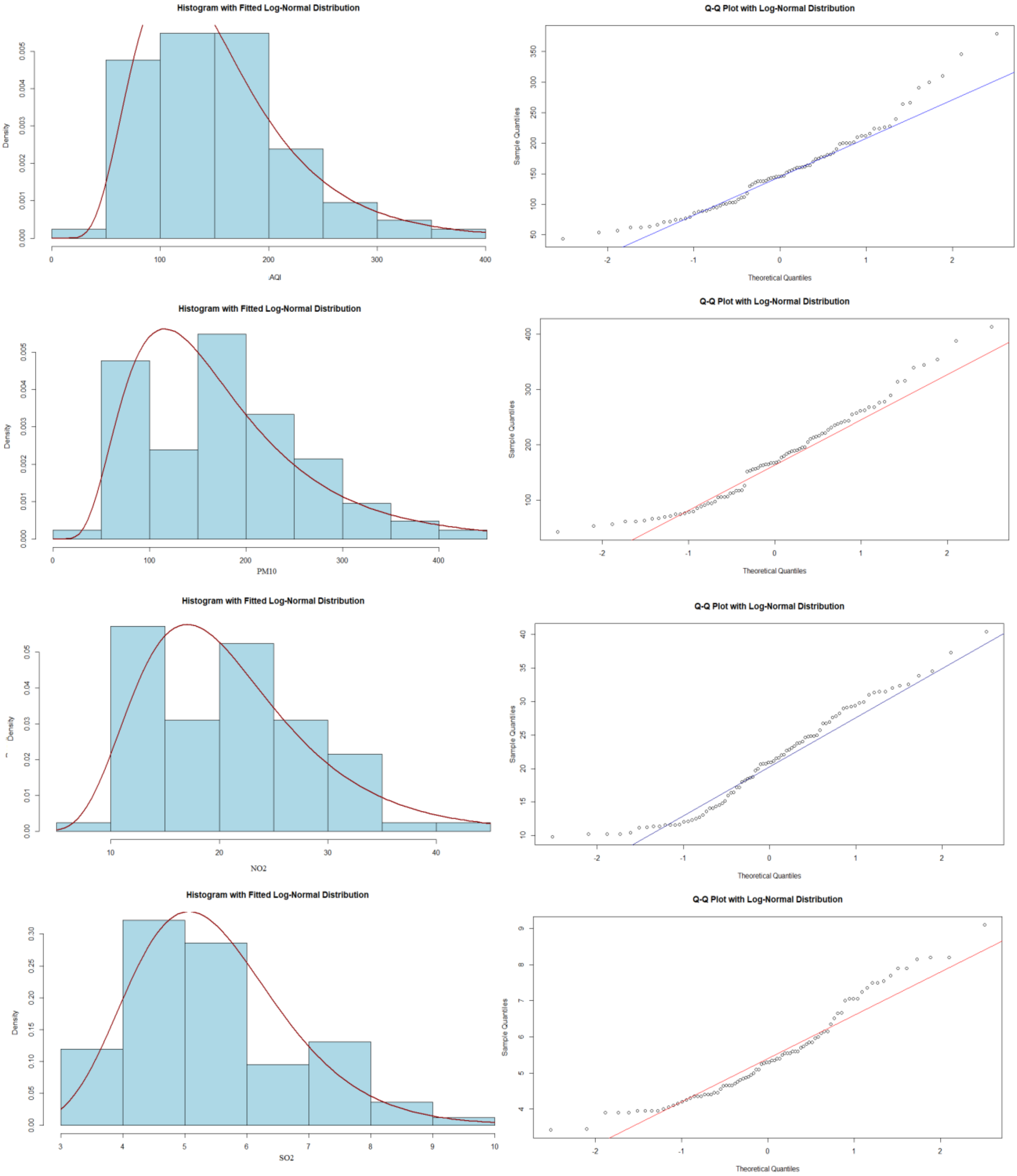

Figure: 1 Lognormal Distribution on Air Pollutants Data

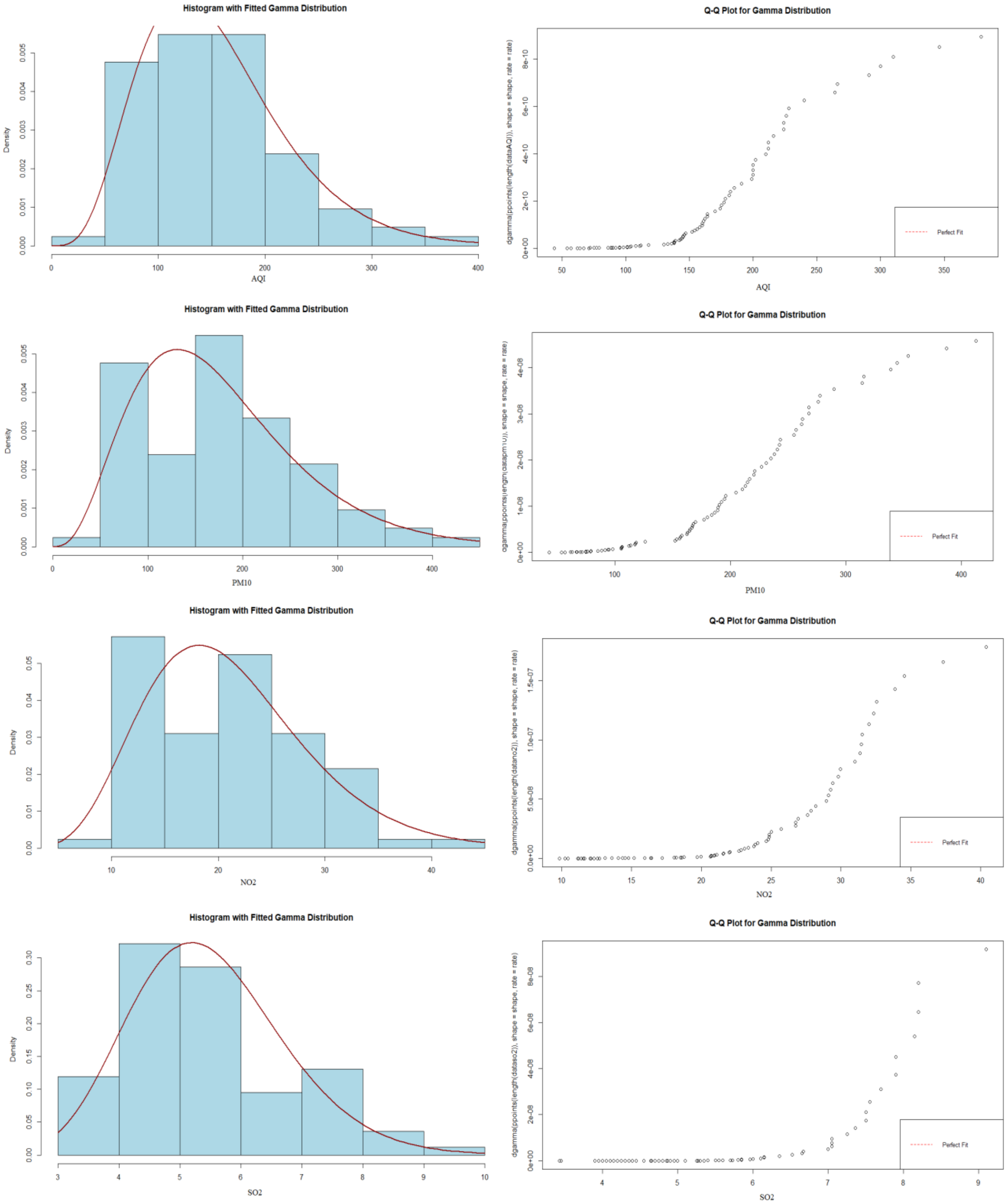

Figure: 2 Gamma Distribution on Air Pollutants Data

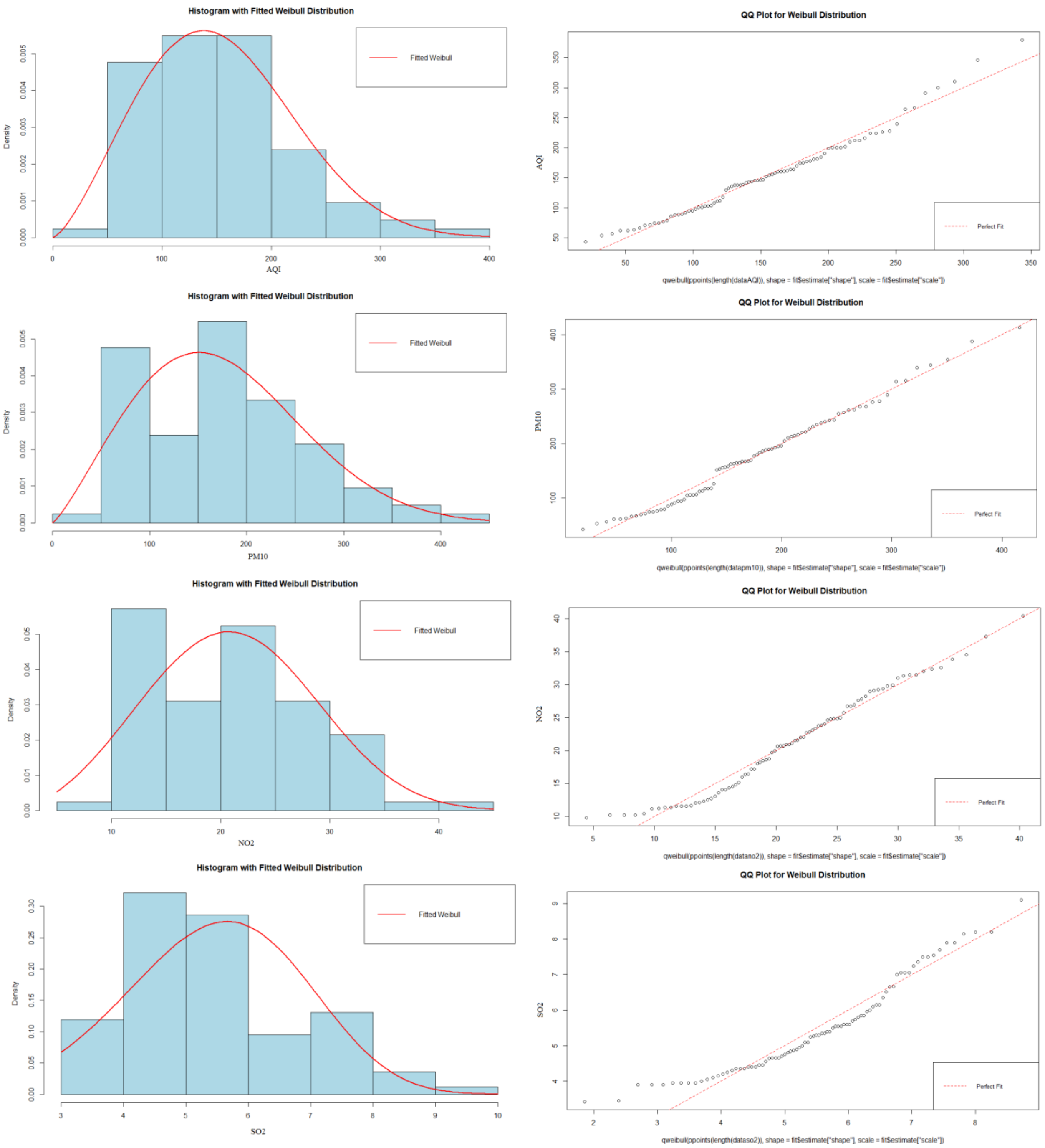

Figure: 3 Weibull Distribution on Air Pollutants Data

Table 3: Statistical Distribution on Air pollutants

		Pollutants Value			
Distribution	**Variable**	**AQI**	**PM_{10}**	**NO_2**	**SO_2**
Lognormal Distribution	**Meanlog**	4.9310958	5.035034	2.9777141	1.6742085
	sdlog	0.4589698	0.532369	0.3782971	0.2290027
	Loglik	467.9861	**489.1785**	**287.6645**	136.0065
Gamma Distribution	**Estimate Shape**	5.12400149	3.9870912	7.43839	18.86078
	Estimate rate	0.03344334	0.02276709	0.3535317	3.442273
	Loglik	467.5028	487.5363	286.9194	137.2094

Weibull Distribution	**Estimate Shape (alpha)**	2.386671	2.207347	3.059027	4.373963
	Estimate rate (beta)	173.1532	198.407439	23.607839	5.999728
	Loglik	**469.9724**	487.7763	287.4534	**143.8565**

RESULT & CONCLUSION

One effective method for estimating the future return period of high air pollution events is the modelling of statistical distributions of air pollution data. This approach not only aids in predicting such events but also enables relevant government agencies to implement preventive measures, particularly for vulnerable populations like the elderly, children, and individuals with asthma. Additionally, these tools are invaluable in the policy-making process, helping to determine whether proposed development areas require environmental assessments to mitigate health risks.

In the research discussed in this study, the Gamma, Lognormal, and Weibull distributions are the most frequently used to model air pollution data. However, other distributions, such as Gumbel, Nakagami, and Frechet, may also be suitable for representing the dispersion of air pollutants. The optimal distribution often depends on the specific characteristics of the region being studied, and findings may vary across different areas.

The objective of this observational study is to identify the distributions most commonly used to predict air pollution concentrations, thereby enhancing the accuracy of future air quality forecasts. We collected air pollution data from UPPCB monitoring stations in Agra, covering the period from January 2017 to December 2023. By fitting the data to Weibull, Gamma, and Lognormal distributions, we found that for Agra city, the lognormal distribution best fits SO_2 and NO_2, while the Weibull distribution is the most suitable for modelling PM_{10} and the Air Quality Index (AQI).

REFERENCES

[1] Zinordin.N.S.,Ramli.N.A.,Sulaiman.M.,Awang.N.R.(2014). A review of the effect of traffic, road characteristic and meteorological conditions on ozone precursors from vehicle emission. International journal of engineering research&technology. Vol,3 issue 11.

[2] Demuzere, M., van Lipziq, N. P. M. (2010). A new method to estimate air-quality level using a synoptic- regression approach. Part 1: Present-day O3 and PM10 analysis. Atmospheric Research 44, 1341-1355.

[3] Khaniabadi.Y.O.,Goudarzi.G.,Daryanoosh.S.M.,Borgini.A.,Tittarelli.A.,Marco.D.A.(2016). Exposure to PM10, NO2, and NO3 and impact on human health. Environ Sci Pollut Res. DOI 10.1007/s11356-016-8083-6.

[4] Kim, K.H., Kabir, E., & Kabir, S. (2015). A Review on the Human Health Effect of Airbone Particulate Matter. Environmental International 136-142.

[5] Yahaya, A.S., Ramli, N.A. & Hamid, H.A.(2007). Review of Fitting Distribution on Air Pollution Modelling. Prosiding Simposium Kebangsaan Sains Matematik ke-XV, Malaysia.

[6] Chen, R., Kan, H., Chen, B., Huang, W., Bai, Z., Song, G., Pan, G. (2012). Association of particulate air pollution with daily mortality the China air pollution and health effects study. Am. J. Epidemiol. 175 (11), 1173-1181.

[7] Mabahwi, N.A., Leh, O, L, H and Omar, D. (2014). Human Health and Wellbeing: Human health effect of air pollution. Proceia Social and Behavioral Science 153, 221-229.

[8] Standers, L.H. (2000). Regulatory Aspects of Air Pollution Control in the United States. Air & Waste Management Association, 8-21.

[9] https://cpcb.nic.in/namp-data/

[10] https://cpcb.nic.in/uploads/National_Ambient_Air_Quality_Standards

[11] Li, M., and Zhang, L. (2014). Haze in China: Current and future challenges. Environmental pollution, Volume 189, 85-86.

[12] Georgopoulus,P.G, and Seinfeld,J.H (1982). Statistical Distribution of air quality concentrations. Environment Science and Technolo-gy,16,401A-416A.

[13] Lu,H.C. (2002). The statistical characters of PM10 concentration in Taiwan area. Atmospheric Environment, 36,p.491-502.

[14] Lu,H.C. (2004). Estimating the emission Source reduction of PM10 in central Taiwan. Journal of chemosphere, 54(7),p.805-814.

[15] Sedek.J.N.M.,Ramli.N.A.,Yahaya.A.S.(2006). Air quality predictions using log normal distribution functions of particulate matter in Kuala Lumpur. Malaysian journal of Environmental Manage-ment 7:33-41.

[16] Noor.N.M.,Tan.C.Y.,Abdullah.M.M.A.,Ramli.N.A., Yaha-ya.A.S.(2011). Modeling of PM10 concentration for industrialized area in Malaysia : A case study in Nilai. IPCBEE vol.12 IACSIT press,Singapore.

[17] Hamid.H.A.,Yahaya.A.S.,Ramli.N.A.,Ul-saufie.A.Z.(2013). Finding the best statistical distribution model in PM10 concentration modeling by using log normal distribution. Journal of Applied Sciences, 12(2): 294-300.

[18] Xi.W.,Jie.C.R.,Heng.C.B.,Dong.K.H.(2013). Application of statistical distribution of PM10 concentration in air quality management in 5 representative cities of China. Biomed Environ Sci, 26(8): 638-646.

[19] Oguntunde.P.E.,Odetunmibi.O.A.,Adejumo.A.O.(2014). A study of probability models in monitoring environmental pollution in Nigeria. Journal Probability and Statistics, Article ID 864965,6 pages.

[20] Maciejewska.K.,Rezlar.K.J.,Reizer.M.,Klejnowski.K.(2015). Modelling of black carbon statistical distribution and return periods of extreme concentrations. Environmental Modelling & software 74: 212-226.

[21] Berthe.K.A.,Abdramane.B.,Reichenbach.S (2015). Gumbel Weibull distribution function for Sahel precipitation modeling and predicting: Case of Mali. ISSN 1996-0786. Vol. (5),pp. 405-412.

[22] Nasir.M.Y.,Ghazali.N.A.,Moktar.M.I.Z.,Suhaimi.N.(2016). Fitting statistical distribution function on ozone concentration data at coastal areas. Malaysian journal of analytical sciences, Vol 20 No3:551-559.

[23] AlDhurafi.N.A.,Razali.A.M.,Masseran.N.,Zamzuri.Z.H.(2016). The probability distribution model of air pollution index and its dominants in Kuala Lumpur. American Institute of Physics,Engineering Science and Technology. Vol,10,No12, 1560-1574, DOI, 10.1063/1.4966829

[24] Ozay.C.,Celiktas.M.S.(2016). Statistical analysis of wind speed using two-parameter Weibull distribution in Alacati region. Energy Conversion and Management 121: 49-54.

[25] Yunus.R.M.,Hasan.M.M (2017). Predicting hourly PM10 concentration in Seberang Perai and Petaling Jaya using log normal linear model. Proceeding of IASTEM International conference. ISBN: 978-93-86083-34-0.

[26] Jaffar.M.I.,Hamid.H.A.,Yunus.R.,RaffeeA.F.(2018). Fitting Statistical Distribution on Air Pollution: an Overview. International Journal of Engineering & Technology,7 (3.23) (2018) 40-44, DOI: 10.14419/ijet.v7i3.23.17256