Statistical Distribution Models for Airborne VOCs (Xylene, Toluene and Benzene) in Visakhapatnam using Burr XII 3P, Log-Logistic 3P and Dagum-I 3P Distributions

Saripalli Arun Kumar[1], Akiri Sridhar[1*], Sarode Rekha[2], Siripurapu Adilakshmi[3] and Ramanaiah M.[4]
1. Department of Mathematics, GSS, GITAM (Deemed to be University), Visakhapatnam-530045, INDIA
2. Department of Mathematics, Madanapalle Institute of Technology and Science, Chittoor-517325, INDIA
3. Department of Basic Science and Humanities, Vignan's Institute of Information Technology, Visakhapatnam-530049, INDIA
4. Department of Chemistry, Aditya Institute of Technology and Management, Tekkali-532201, INDIA
*sakiri@gitam.edu

Abstract

This study investigates the application of advanced statistical models for airborne Volatile organic compounds (VOCs)-specifically xylene, toluene and benzene-using air pollution data collected from the city Visakhapatnam from the year 2018 to 2022. To capture the non-normal distributional behavior of VOC concentrations, three flexible probability distributions were employed: Burr XII 3P, Log-Logistic 3P and Dagum-I 3P. Parameter estimation was performed via maximum likelihood estimation (MLE) while model validation was achieved through P–P and Q–Q plots. In order to identify the most suitable distribution for modeling the air pollutant data, three distinct goodness of fit test statistics and five model selection criteria were applied.

The results demonstrate that the Dagum-I 3P distribution best fits xylene and benzene, while the Log-Logistic 3P model is optimal for the moderately skewed toluene concentrations. Cross-validation confirms these findings, highlighting the reliability of tailored distributional models for VOCs. The proposed distributions provide a robust framework for predicting air quality information and conducting accuracy assessments with all calculations and visual indications carried out with R-software. This work underscores the potential of pollutant-specific modeling for improved air quality assessment and management strategies, contributing to environmental health planning in urban-industrial areas.

Keywords: Air Pollution, Environmental Data Analysis, VOC Modeling, Probability Distributions, Maximum Likelihood Estimation, Cross-Validation.

Introduction

Air pollution indicates immediate danger towards environment and health of public concern world-wide, particularly in industrialized urban areas. Among the array of harmful pollutants, volatile organic compounds like xylene, toluene and benzene are of particular concern due to their prevalence in both urban and industrial environments. These VOCs emitted from an assortment of sources, including vehicular emissions, industrial processes and the use of solvents, are known to contribute to both short- and long-term health issues such as respiratory ailments, neurological disorders and cancer, making them a significant focus for environmental monitoring and regulatory control. The city of Visakhapatnam, one of the India's major industrial hubs, has witnessed increasing VOC levels in recent years. This escalation is largely attributed to the city's rapid growth and industrialization.

Traffic density and complex environmental conditions are influenced by meteorology and topography. Understanding VOC concentration patterns and accurately modeling their statistical distribution is essential for effective air quality management in urban-industrial regions. Effective modeling can provide insights into pollutant behavior, can inform regulatory standards and aid in designing interventions to mitigate pollution's impact on public health and the environment.

Previous studies have utilized various probability distributions including normal, log-normal and other standard models, to represent pollutant concentrations in the atmosphere. However, such distributions often fall short in capturing the non-normal, highly skewed nature typical of VOC concentration. VOC distributions are commonly characterized by extreme values, skewness and heavy tails, necessitating the use of more flexible statistical models. Advanced distributions like the Burr XII 3P, Log-Logistic 3P and Dagum-I 3P offer enhanced flexibility, enabling a more accurate representation of pollutant concentration variability and distributional behavior. These models have been successfully applied in the fields such as income inequality, environmental toxicology and epidemiology but are relatively unexplored in air quality modeling for VOCs in urban-industrial environments.

Research by Gavriil et al[6] analysed eight probability density functions for PM_{10} and $PM_{2.5}$ and concluded that the Pearson type VI, inverse Gaussian and lognormal distributions provide the best fits, particularly for high concentration percentiles[6]. Ahmat et al[1] found that the three-parameter Generalized Extreme Value (GEV) distribution is highly effective in predicting extreme PM_{10} concentrations in Malaysia, demonstrating strong accuracy in forecasting

exceedances[7]. Hein et al[7] created a database of benzene, toluene and xylene measurements to develop statistical models predicting exposure levels based on various workplace determinants.

The study addresses limitations in historical data and modeling, providing parameter estimates for specific operations that are useful for community-based studies[14]. Shen et al[14] studied the embryotoxicity of benzene, toluene, xylene and formaldehyde on murine embryonic stem cells via airborne exposure, finding formaldehyde to be the most toxic with significant IC (50) values. The study supports this model as effective for predicting the embryotoxicity of volatile organic compounds[2]. Atari and Luginaah[2] developed land use regression models to predict BTEX concentrations in Sarnia, Ontario, with industrial areas, dwelling counts and highways accounting for most variability (R^2: 0.78–0.81). The study highlights the importance of modeling BTEX to assess its health impacts alongside nitrogen oxides and particulate matter[1].

Nurmatov et al[12] reviewed 8,455 studies on VOCs and their effects on asthma and allergies, selecting 53 relevant manuscripts. They found inconsistent evidence and significant bias, particularly regarding aromatics and formaldehyde and called for more rigorous research on VOC exposure's impact[12]. Thupeng[15] utilized the three-parameter Burr type XII distribution to model maximum nitrogen dioxide levels at the Gaborone fire brigade in winter 2014. Comparing it to the Dagum and log-logistic models, the Burr type XII showed the best fit, highlighting its effectiveness for modeling extreme ambient air pollution levels[15]. Bang et al[3] found that large industrial complexes in Ulsan contributed about 40% of annual BTX (Benzene, Toluene, Xylene) levels in nearby urban areas, with higher concentrations in summer. Contributions decreased with distance from the sites, offering valuable data for environmental epidemiology[3].

Rashnuodi et al[13] identified significant seasonal differences in BTEX exposure among petrochemical workers in western Iran, with strong correlations to biological indices. The study emphasized the necessity for strategies to reduce hazardous levels, especially for benzene and toluene. Kamani et al[8] found that BTEX concentrations in indoor environments in Zahedan, Iran, exceeded EPA limits. Carcinogenic risks from benzene and ethylbenzene were significant and toluene's hazard quotient was above one, indicating potential health effects. Chaudhary et al[5] introduced the New Extended Kumaraswamy Exponential Distribution to analyze air quality data in Kathmandu, revealing poor conditions with higher pollutant levels in winter. The model demonstrated flexibility for forecasting and reliability analyses, validated through various statistical tests.

This study evaluates the best suitability of the Burr XII 3P, Log-Logistic 3P and Dagum-I 3P probability distributions for modeling the concentrations of airborne xylene, toluene and benzene. The analysis utilizes air quality data from Visakhapatnam, collected over the period from 2018 to 2022. Model parameters are estimated via Maximum Likelihood Estimation (MLE). The adequacy of the fitted models is assessed through graphical diagnostic tools including P–P and Q–Q plots, while the fit quality is evaluated using statistical tests such as the Anderson–Darling, Kolmogorov–Smirnov and Cramér–von Mises tests.

The Akaike Information Criterion (AIC), Corrected Akaike Information Criterion (CAIC), Bayesian Information Criterion (BIC), Hannan–Quinn Information Criterion (HQIC) and Approximate Bayesian Information Criterion (ABIC) were applied to assess model performance and select the most appropriate distribution for each VOC. Furthermore, to ensure the robustness and generalizability of the models, K-fold cross-validation is employed. By applying these advanced distributional models to VOC concentration, this study contributes to the field of environmental data analysis, offering a methodological approach in understanding pollutant dynamics in regions experiencing rapid industrial growth. This tailored modeling approach aims to enhance air quality management strategies, contributing to informed decision-making for environmental health and regulatory practices in urban-industrial regions.

Material and Methods

Material: A dataset of daily average air quality measurements, along with the application of Burr XII 3P, Log-Logistic 3P and Dagum-I 3P statistical distributions is described to model VOC trends accurately and to assess air quality dynamics in an industrial setting.

Real Data Description: This study aims at data analysis; it was conducted by using daily average ambient air quality datasets collected from January 2018 to December 2022 by the Andhra Pradesh Pollution Control Board at the Continuous Ambient Air Quality monitoring station in Visakhapatnam (GVMC). The datasets include measurements of key airborne volatile organic compounds (VOCs), specifically xylene, toluene and benzene, alongside other pollutants such as Carbon Monoxide (CO), Ozone (O_3), Nitric Oxide (NO), Nitrogen Dioxide (NO_2), Nitrogen Oxides (NOx), Ammonia (NH_3), Sulfur Dioxide (SO_2), $PM_{2.5}$ and PM_{10}. Meteorological parameters including temperature (AT), relative humidity (RH), wind speed (WS), wind direction (WD), solar radiation (SR), barometric pressure (BP) and rainfall (RF), were also recorded.

The data collection provides daily records, allowing for a detailed assessment of air quality trends and meteorological influences over a five-year period. To ensure data integrity, preliminary processing involved handling missing values through linear interpolation and addressing outliers, defined as values exceeding three standard deviations from the mean, using Winsorization. This comprehensive dataset provides a robust foundation for advanced statistical

distributions (Burr XII 3P, Log-Logistic 3P and Dagum-I 3P) inmodeling, enabling a thorough analysis of VOC trends and variability in an industrial urban environment. It facilitates detailed evaluations to better understand airborne VOC levels in Visakhapatnam.

Continuous Probability Distributions in Air Pollution: Statistical probability distributions are essential for modeling air pollution data, offering a systematic approach to capture variability and trends in environmental quality. Given the observed non-normality and positive skewness in VOC concentration data, three advanced probability distributions were selected for fitting: Burr XII 3P**,** Log-Logistic 3P and Dagum-I 3P. Each distribution were chosen for its flexibility in capturing skewed, heavy-tailed behavior often found in environmental datasets.

Burr-XII 3P distribution: It is known effective for modeling skewed or heavy-tailed data, making it suitable for environmental data where extreme values may occur. The probability density function (PDF) is as:

$$f(x;\tau,\omega,\sigma) = \frac{\tau \cdot \omega \cdot (x-\sigma)^{\tau-1}}{\left[1+\left(\frac{x-\sigma}{\omega}\right)^{\tau}\right]^{\tau+1}}$$

and its cumulative distribution function (CDF) is:

$$F(x;\tau,\omega,\sigma) = 1 - \left[1+\left(\frac{x-\sigma}{\omega}\right)^{\tau}\right]^{-\tau}$$

where $x > \sigma$, $\tau > 0$ (Shape), $\omega > 0$ (Scale) and $\sigma > 0$ (Location).

Dagum-I 3P distribution: This distribution is versatile for capturing heavy tails and variable skewness, making it a strong candidate for pollutant data that exhibit significant kurtosis. Its PDF and CDF are:

$$f(x;\gamma,p,\phi) = \frac{\gamma \cdot p \cdot x^{\gamma-1}}{\phi\left[1+\left(\frac{x}{\phi}\right)^{\gamma}\right]^{p+1}} \quad \text{and} \quad F(x;\gamma,p,\phi) = \left[1+\left(\frac{x}{\phi}\right)^{\gamma}\right]^{-p},$$

where $x > 0$, $\gamma > 0$(Shape), $p > 0$ (Shape) and $\phi > 0$ (Scale).

Log-Logistic 3P distribution: Commonly used in survival and reliability analysis, this distribution is effective for moderately skewed data and has been applied to VOC data with moderate tail behavior. Its functions are defined as:

$$f(x;\mu,\nu,\gamma) = \frac{\mu \cdot \nu \cdot (x-\gamma)^{\mu-1}}{[\nu+(x-\gamma)^{\mu}]^{\mu+1}} \text{ and } F(x;\mu,\nu,\gamma) = \frac{1}{1+\left(\frac{\nu}{x-\gamma}\right)^{\mu}}$$

where $x > \gamma$, $\mu > 0$(Shape), $\nu > 0$ (Scale) and $\gamma > 0$ (Location).

Utilizing these three-parameter distributions enhances the predictive accuracy and reliability of our analysis, allowing for a better understanding of pollution dynamics and aiding in informing policy decisions for air quality management.

Parameter Estimation: Parameter estimation for each selected distribution, Burr XII 3P, Log-Logistic 3P and Dagum-I 3P was conducted using the Maximum Likelihood Estimation (MLE) method. MLE is a robust statistical approach that maximizes the likelihood function, yielding parameter values that best explain the observed data. For this study, MLE was implemented in R program to ensure accuracy and efficiency across the large VOC dataset[9]. The likelihood functions for each distribution were optimized by differentiating the log-likelihood to each parameter in solving these equations computationally to obtain optimal parameter values.

This approach ensured that each distribution's unique shape, scale and location parameters fitted specifically to the concentration patterns of xylene, toluene and benzene, thus capturing their distinct distributional characteristics. The use of MLE allowed for reliable parameter estimates that align with the skewed, non-normal nature of the air pollutant, enhancing model accuracy and predictive performance.

Model Validation Techniques: To assess the accuracy and appropriateness of each fitted model, a combination of graphical and statistical validation techniques was applied.

Probability-Probability (P–P) Plot: P–P plot facilitates a comparison within the empirical cumulative distribution function (CDF) of the observed data and the theoretical CDF derived from the fitted model. This methodology is effective for evaluating the degree of the model which aligns to the cumulative distribution of the actual data. Precisely, for an ordered sample $x_{(1)}, x_{(2)}, \ldots\ldots, x_{(n)}$, P–P plot involves plotting the empirical probabilities $E(i) = \frac{i}{n+1}$ against the model's theoretical probabilities $F(x_{(i)})$. A linear alignment along the 45-degree line indicates that the theoretical distribution closely approximates the observed data, thereby suggesting a strong model fit.

Quantile-Quantile (Q–Q) Plot: Q–Q plot serves as an additional graphical method for model validation that compares the quantiles of the observed dataset with those of the fitted distribution. In this plot, observed quantiles $x_{(i)}$ are plotted against theoretical quantiles $F^{-1}(E(i))$ derived from the fitted model. A near-linear relationship along the 45-degree line indicates that the model distribution aligns with the empirical distribution.

This technique is especially effective for identifying discrepancies in the tails, highlighting deviations in extreme values. Collectively, these graphical methods provide an intuitive framework for validating models by directly assessing the correspondence between observed data and theoretical models.

Goodness of Fit Test Statistics

The Anderson–Darling (AD), Kolmogorov–Smirnov (KS) and Cramér–von Mises (CVM) tests were conducted to evaluate model fit quantitatively. These tests measure how closely the fitted model matches the perceived data.

Kolmogorov–Smirnov (KS) Test: KS test is a non-parametric statistical method that quantifies the maximum distance between the empirical distribution function (EDF) of the observed data and the theoretical cumulative distribution function (CDF) derived from the fitted model. For an ordered dataset x_1, x_2,……,x_n, this test evaluates supremum of the absolute differences between EDF and theoretical CDF, thereby providing a measure of GOF. The test is computed as:

$$K_n = \max\left(K_n^{+}, K_n^{-}\right)$$

where $K_n^{+} = max_{1\le i\le n} \mid F_n(x_i) - F_0(x_i) \mid$ and $K_n^{-} = max_{1\le i\le n} \mid F_0(x_i) - F_n(x_i)$.

Here, $F_n(x)$ is EDF of the sample and $F_0(x)$ is theoretical CDF. If KS test value is below the critical value or if p-value exceeds the selected significance level, the null hypothesis says that the data follows fitted distribution-cannot be rejected, indicating a good fit.

Anderson–Darling (A^2) Test: The A-D test modifies KS test by placing more weight on the tails of distribution, making it particularly sensitive to discrepancies in extreme values. For a continuous distribution with ordered sample $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, the test statistic A^2 is calculated as:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}((2.i-1)[\ln(F(x_{(i)})) + \ln(1 - F(x_{(n+1-i)}))])$$

A lower value of $A^2$with a high p-value indicates a well-fitting model. A–D test is effective for identifying discrepancies at the distribution tails, a critical aspect in the analysis of pollutant concentrations where extreme values frequently exert significant influence.

Cramér–von Mises (W) Test: The C-vM test measures the cumulative squared differences between EDF and theoretical CDF across the entire range of data, providing a balanced assessment of the fit across all distribution points. The test statistic W for a sample of ordered data $x_{(1)}$, $x_{(2)}$, ……,$x_{(n)}$, is as:

$$W = \frac{1}{12.n} + \sum_{i=1}^{n}\left(\frac{2.i-1}{2.n} - F\left(x_{(i)}\right)\right)^2$$

This test offers a comprehensive measure of fit quality and is sensitive to deviations in both the central and tail regions of the distribution. A higher p-value in C-vM test suggests that the observed data aligns well with the fitted model.

Model Selection Criteria: Model evaluation is essential in statistical analysis for identifying the appropriate distribution of volatile organic compound (VOC) concentrations. It was based on a range of information criteria to ensure optimal model choice by balancing fit quality and model complexity.

Akaike Information Criterion (AIC): AIC is a theoretic measure that evaluates model quality by balancing the trade-off between fit accuracy and model complexity. It is defined as:

$$\mathrm{AIC} = -2.\ln(\mathrm{Likelihood}) + 2.\widehat{p_e}$$

where $\widehat{p_e}$is the number of estimated parameters in each model. Lower values of AIC indicate superior model performance, as this criterion imposes a penalty on models with extraneous parameters. By balancing fit quality with the number of parameters, AIC facilitates the identification of the most parsimonious model that adequately explains data while avoiding overfitting.

Corrected Akaike Information Criterion (CAIC): CAIC extends AIC by adding a correction for sample size, particularly useful in cases with smaller datasets. CAIC is expressed as:

$$\mathrm{CAIC} = -2.\ln(\mathrm{Likelihood}) + 2.\widehat{p_e}.\frac{n}{n - \widehat{p_e} - 1}$$

CAIC imposes a more rigorous penalty for the inclusion of additional parameters compared to AIC. This characteristic diminishes the probability of selecting overly complex models, particularly in contexts with limited sample sizes.

Bayesian Information Criterion (BIC): BIC, proposed by Schwarz, also penalizes model complexity but with a stronger focus on the sample size. This is calculated as:

$$\mathrm{BIC} = -2.\ln(\mathrm{Likelihood}) + \widehat{p_e}.\ln(n)$$

BIC employs a logarithmic penalty for the number of estimated parameters, scaled by the sample size, which biases it more strongly against complex models than AIC. Lower BIC values favor models that achieve high likelihood with fewer parameters, making BIC particularly advantageous for selecting parsimonious models in large datasets.

Hannan–Quinn Information Criterion (HQIC): HQIC serves as an alternative to AIC and BIC by implementing a penalty that increases at a rate slower than the logarithmic factor applied in BIC. This distinctive characteristic allows HQIC to balance model complexity and fit in a manner that may be more favorable in certain analytical contexts. It is estimated as:

$$\mathrm{HQIC} = -2.\ln(\mathrm{Likelihood}) + 2.\widehat{p_e}.\ln\left(\ln(n)\right)$$

HQIC is beneficial in contexts prioritizing fit quality and moderate model complexity, effectively balancing the lower penalty of the AIC with the stronger penalty of BIC.

Approximate Bayesian Information Criterion (ABIC): ABIC is a modified version of BIC designed to improve model selection for small sample sizes by adjusting complexity penalty to avoid favoring overly complex models. It is found as:

$$\text{ABIC} = -2.\ln(\text{Likelihood}) + \widehat{p_e}.\ln(n) + \frac{2.\widehat{p_e}^2}{n}$$

Lower ABIC values indicate better models, offering a more balanced trade-off between fit and complexity in small sample scenarios.

Log-Likelihood Criteria (LL): LL is a measure to assess the fit of statistical models to data. It quantifies how better a model explains the observed outcomes, with higher values indicating a better fit. It is calculated as follows:

$$L(\theta) = \sum_{i=1}^{n} \log(f(x_{(i)}, \theta))$$

where n is the number of observations, $x_{(i)}$ represents each observed data point, $f(x_{(i)}, \theta)$ is PDF of chosen model evaluated at $x_{(i)}$ parameterized by θ.

Cross-Validation: The K-fold cross-validation was employed for the robustness and predictive performance of the statistical models selected for VOC concentrations. In this study, a 10-fold cross-validation approach was chosen to assess the generalizability of each fitted distribution model, dividing the dataset into ten subsets. For each iteration, one subset was used as the validation set, while the leftover nine subsets were used for training. This iterative process was repeated ten times, calculating the average performance metrics to ensure that the models were evaluated across different data segments. Each model's performance was assessed using key selection criteria including AIC, BIC, CAIC, ABIC and HQIC. These criteria provide a balanced assessment, penalizing overly complex models while rewarding those that effectively fit the data. Lower average values across these metrics indicate a better fit and model reliability. The cross-validation approach enhances model generalizability by minimizing the risk of overfitting and ensuring that the models can reliably predict VOC concentrations across different time periods. This method provides a robust evaluation framework, reinforcing the findings from the goodness of fit test statistics and supporting the choice of pollutant-specific models for accurate air quality assessments.

Data Visualization Enhancements: To better interpret the temporal trends and external influences on VOC concentrations, figures 1 and 2 were enhanced with time series visualizations of xylene, toluene and benzene levels. Shaded regions indicate seasonal periods (monsoon, winter, summer), helping to visualize potential seasonal effects on VOC levels, particularly increases during colder months when pollutants may be trapped near the surface. Vertical lines mark significant external events, such as the COVID-19 lockdown in early 2020, allowing comparisons of pollutant levels before, during and after these events. A 30-day rolling average was calculated and overlaid on each plot to smooth short-term fluctuations and clarify long-term trends. Notable peaks and dips in VOC levels were annotated with concentration values or brief explanations of potential causes, adding context to extreme values. These enhancements provide a detailed view of VOC concentration trends, improving the assessment of model fit and pollutant dynamics in response to seasonal and event-driven factors.

Results

In air pollutant modeling, the proposed advanced statistical distributions provide distinct advantages that align with the data characteristics. Analyzing xylene, toluene and benzene reveals valuable insights into their concentration distributions and statistical properties. Table 1 presents descriptive statistics for VOC concentrations in Visakhapatnam from 2018 to 2022. Xylene has a mean concentration of 2.73µg/m³ and a standard deviation of 4.74µg/m³, indicating significant variability. Its median concentration is 1.9µg/m³, lower than the mean, with a skewness of 10.38 and kurtosis of 133.13, reflecting a right-skewed distribution with substantial outliers. The range is 0.1 to 80.4µg/m³, showing sporadic spikes.

Toluene has a mean of 10.23µg/m³ and a sd of 7.46µg/m³, with a median of 8.8 µg/m³, skewness of 2.95 and kurtosis of 15.11, indicating moderate skewness and a wide range of 78.3µg/m³. Benzene, with a mean of 3.47 µg/m³ and sd of 1.9µg/m³, has a median of 3.2µg/m³, suggesting a more symmetric distribution. Its skewness of 1.09 and kurtosis of 4.67 indicate slight right skewness, with a narrower range of 0.1 to 20.9µg/m³. These patterns emphasize the need for robust statistical modeling to accurately capture data characteristics and extreme concentrations.

To address these needs, advanced statistical distributions such as Burr XII 3P, Log-Logistic 3P and Dagum-I 3P were evaluated to enhance predictive accuracy and inform regulatory decisions in air quality management[4].

The enhanced time series plots in figures 1 and 2 illustrate seasonal patterns and responses to external events such as the COVID-19 lockdown[11]. Seasonal indicators allow for clearer comparison of VOC concentration levels across monsoon, winter and summer, while annotations reveal potential links between extreme values and known events. Figure 1 presents individual plots where xylene (blue line) shows low-level fluctuations with sporadic peaks, particularly during 2020-2021, suggesting episodic increases in emissions.

Table 1
Descriptive Statistics for Xylene, Toluene and Benzene concentrations in Visakhapatnam

Statistic	Air Pollutant		
	Xylene	**Toluene**	**Benzene**
Sample Size (n)	1,627	1627	1627
Minimum (µg/m³)	0.1	0.1	0.1
Maximum (µg/m³)	80.4	78.4	20.9
1st Quartile	1.1	5.9	2.1
Median (µg/m³)	1.9	8.8	3.2
Mean (µg/m³)	2.73	10.23	3.471
3rd Quartile	3.2	12.85	4.7
Range (µg/m³)	80.3	78.3	20.8
Standard Error of Mean (SE Mean)	0.12	0.18	0.047
Lower 95% CI for Mean	2.5	9.87	3.38
Upper 95% CI for Mean	2.96	10.59	3.56
Variance	22.47	55.67	3.6
Standard Deviation (µg/m³)	4.74	7.46	1.9
Skewness	10.38	2.95	1.09
Kurtosis	133.13	15.11	4.67
Trimmed Mean (10%) (µg/m³)	2.14	9.3	3.35
Median Absolute Deviation (MAD) (µg/m³)	1.48	4.89	1.93

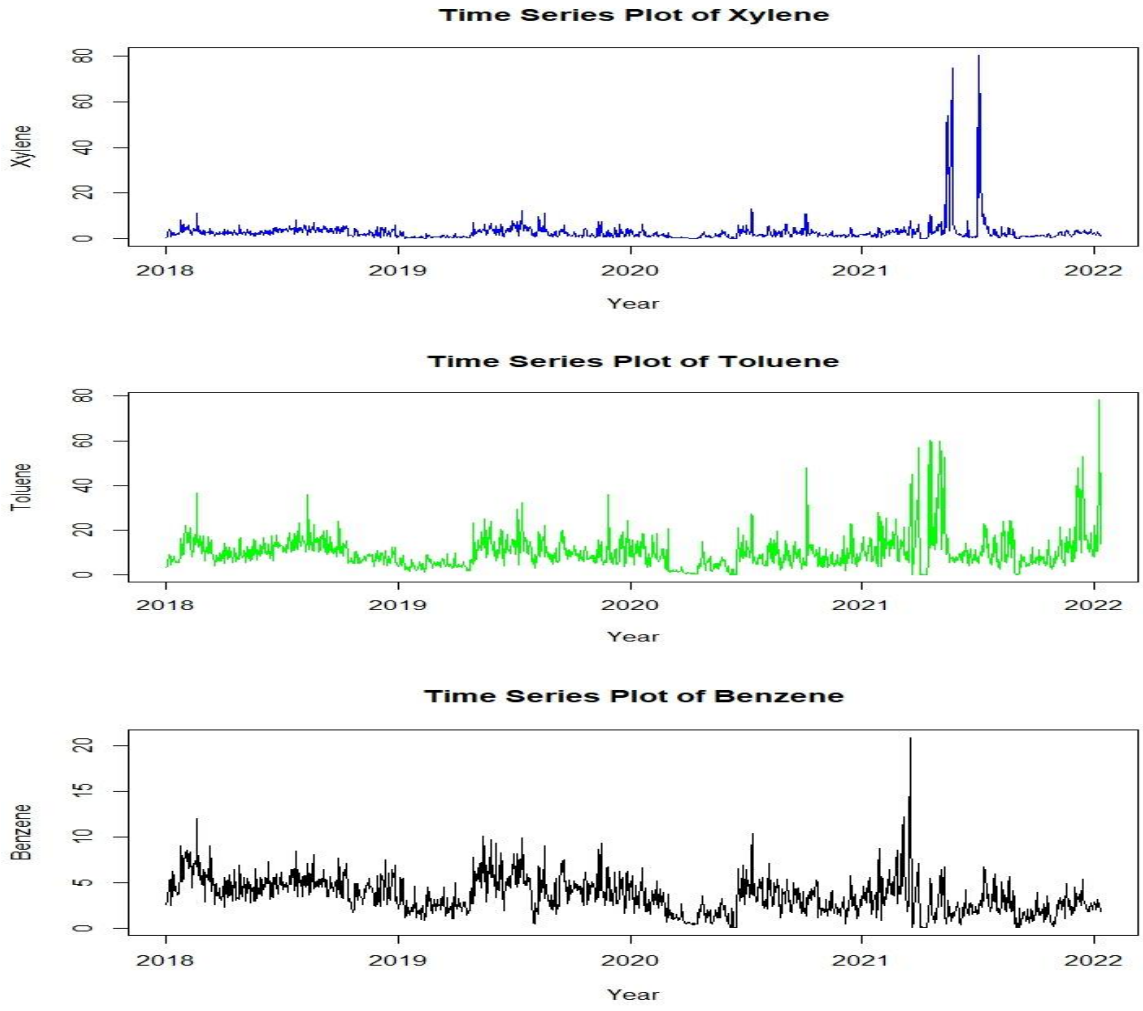

Figure 1: Time series plot of Xylene, Toluene and Benzene concentrations in Visakhapatnam (2018–2022), with seasonal indicators and annotations for significant events (e.g., COVID-19 lockdown) to highlight potential external influences on VOC levels.

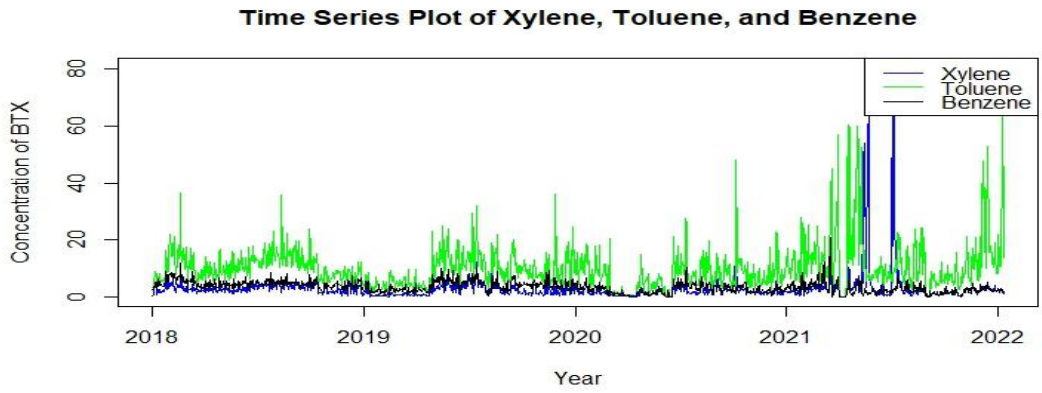

Figure 2: Combined time series plot of Xylene, Toluene and Benzene concentrations, with rolling averages and comparative shading, illustrating pollutant-specific trends and seasonal variations.

Table 2
Parameter Estimates and Standard Errors for Distribution Fits

Distribution	Parameter	Xylene		Toluene		Benzene	
		Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
Burr XII 3P	shape1	1.8127	0.0564	1.9799	0.0599	1.9761	0.0522
	shape2	0.3443	0.0257	0.1821	0.0188	0.0351	0.0139
	scale	1.4746	0.0449	6.9306	0.1589	2.6907	0.0495
Log-Logistic 3P	shape	0.4433	0.0164	0.2598	0.0141	0.1771	0.0178
	scale	0.6927	0.0326	2.4892	0.0488	1.7645	0.0966
	threshold	-0.12	0.0457	-3.271	0.5489	-2.6217	0.549
Dagum-I 3P	shape1.a	2.6033	0.1166	3.8459	0.1601	5.8476	0.3535
	scale	2.6381	0.1359	12.9034	0.3722	5.209	0.1253
	shape2.p	0.5868	0.0483	0.414	0.0275	0.2591	0.0217

Toluene (green line) exhibits higher variability and more frequent fluctuations, especially from 2019 onward, implying dynamic changes in its sources or atmospheric behavior. Benzene (black line) remains relatively stable, with intermittent spikes, especially around 2020, which may reflect specific events or seasonal factors affecting its concentration. These individual plots reveal unique temporal behaviors and emission patterns for each VOC. Figure 2 combines these time series, allowing direct comparison and emphasizing toluene's pronounced fluctuations throughout the period, contrasted by xylene's intermittent peaks and benzene's relative stability with occasional variations. Together, these enhancements underscore trends aligned with the selected distribution models, highlighting the pollutants' seasonal behavior and variability.

Parameter Estimation: Table 2 presents parameter estimates and standard errors for each distribution. They provide insights into their suitability for modeling VOC concentrations with MLE. The results for xylene, toluene and benzenedemonstrate the superior performance of the Dagum-I 3P, Log-Logistic 3P distributions compared to the Burr-XII 3P model. For xylene, Burr XII 3P distribution shows shape parameters (shape1=1.8127, shape2=0.3443) indicating moderate skewness, with a scale parameter of 1.4746. Low standard errors suggest reliable estimates. Log-Logistic 3P model presents a lower shape parameter (0.4433) and scale (0.6927), with a threshold of -0.12 indicating a leftward shift. Dagum-I 3P distribution, with higher shape 1 (2.6033) and shape 2 (0.5868), points to pronounced tail behavior, emphasizing heavy-tailed characteristics. For toluene, Burr XII 3P model's scale parameter (6.9306) captures broader data range, while Log-Logistic 3P fit, with a threshold of -3.27, indicates a significant leftward shift. Dagum-I 3P distribution maintains strong parameter estimates, suggesting a good fit. Benzene's Burr XII 3P estimates (shape1=1.9761, scale=2.6907) imply skewness towards higher values, whereas Log-Logistic 3P distribution features a lower shape parameter (0.1771) and a threshold of -2.62, suitable for describing the data's distribution.

Parameter estimates show significant variability in scale and shape, highlighting the necessity for multiple distribution models to accurately characterize pollutants. This variability is essential for choosing suitable models to improve predictive accuracy and inform air quality management strategies.

Confidence Intervals: Table 3 displays the confidence intervals for the estimated parameters of each model revealing key insights into each model's suitability and precision in capturing VOC characteristics for xylene, toluene and benzene. These intervals indicate plausible

values for the parameters, highlighting their statistical significance and estimation precision.

The results in table 3 indicate that the Burr XII 3P distribution shows narrow intervals for the shape1 parameter across pollutants, particularly for xylene, suggesting stable estimates that effectively capture distributional shape and skewness. However, benzene's shape2 parameter exhibits broader intervals, indicating greater uncertainty in modeling its tail behavior, which is essential for pollutants with extreme values.

For the Log-Logistic 3P distribution, narrow intervals for the shape parameter suggest consistent model fit, although the threshold parameter exhibits wider intervals for toluene and benzene, reflecting uncertainty in lower bound estimates. This variability suggests that the Log-Logistic 3P model is effective for central values but may be less reliable for extreme concentrations. The Dagum-I 3P distribution shows significant variability in Benzene's shape 1.a parameter, indicating higher uncertainty in modeling its heavy tail, while xylene and toluene have relatively narrow intervals for the scale parameter, demonstrating stable modeling capability. Overall, the Burr XII 3P and Dagum-I 3P distributions provide robust fits for xylene, while the Log-Logistic 3P distribution aligns best with toluene's moderate skewness. For benzene, the Dagum-I 3P model captures heavy-tailed behavior with some variability, underscoring the need for pollutant-specific models to address unique distributional profiles effectively.

Goodness-of-Fit tests: The goodness-of-fit for each distribution model fitted to air pollutant data for xylene, toluene and benzene, shown in table 4 is evaluated. The goal is to determine which distribution best describes the observed data using the KS, CVM and AD test statistics, along with their respective p-values. In this analysis, the Burr XII 3P distribution provides a reasonable fit for xylene, with a low KS statistic (0.0348) and significant p-value (0.0385). The CVM and AD statistics (0.2920 and 2.5047) further support its suitability. Conversely, the Log-Logistic 3P distribution exhibits a poor fit, evidenced by higher KS, CVM and AD values and low p-values (p=0.0076 for KS), suggesting a significant deviation from the observed data.

The Dagum-I 3P distribution has lower test statistics (KS=0.0303) and higher p-values (p=0.09998 for KS), indicating a good fit, more as strong as the Burr XII 3P. For toluene, the Burr XII 3P distribution again demonstrates a good fit, with highly significant p-values (p=0.00107 for KS) and high AD statistic (6.8843). The Log-Logistic 3P distribution performs best in this case, with the lowest KS, CVM and AD statistics (0.0149, 0.0616 and 0.8934) and high p-values, indicating an excellent fit.

Table 3
Confidence Intervals for each Model Parameters and Air Pollutant

Distribution	Parameter	Xylene		Toluene		Benzene	
		2.50%	97.50%	2.50%	97.50%	2.50%	97.50%
Burr XII 3P	shape1	1.7022	1.9233	1.8625	2.0972	1.8738	2.0784
	shape2	0.294	0.3946	0.1452	0.219	0.0079	0.0624
	scale	1.3867	1.5626	6.6192	7.242	2.5937	2.7876
Log-Logistic 3P	shape	0.411	0.4755	0.2321	0.2875	0.1423	0.2119
	scale	0.6289	0.7565	2.3935	2.5849	1.5753	1.9538
	thres	-0.2096	-0.0304	-4.3467	-2.1952	-3.6976	-1.5457
Dagum-I 3P	shape1.a	2.3748	2.8318	3.5322	4.1597	5.1547	6.5405
	scale	2.3717	2.9044	12.174	13.6328	4.9634	5.4546
	shape2.p	0.492	0.6815	0.3601	0.4678	0.2165	0.3017

Table 4
Goodness-of-Fit Statistics for each Distribution of Xylene, Toluene and Benzene

Air Pollutant	Distribution	KS (D)	CVM (W^2)	AD (A^2)	p-Value (KS)	p-Value (CVM)	p-Value (AD)
Xylene	Burr XII 3P	0.0348	0.2920	2.5047	0.0385	0.1424	0.04926
	Log-Logistic 3P	0.0414	0.4754	3.7767	0.0076	0.04602	0.01122
	Dagum-I 3P	0.0303	0.2711	2.0263	0.09998	0.1635	0.08884
Toluene	Burr XII 3P	0.0481	0.7885	6.8843	0.00107	0.00778	0.00038
	Log-Logistic 3P	0.0149	0.0616	0.8934	0.8629	0.8041	0.4183
	Dagum-I 3P	0.0417	0.4735	4.0282	0.00699	0.04652	0.00845
Benzene	Burr XII 3P	0.0349	0.3169	2.8308	0.03787	0.1254	0.03338
	Log-Logistic 3P	0.0521	0.7885	4.7209	0.00029	0.00779	0.00390
	Dagum-I 3P	0.0313	0.3119	2.7869	0.0077	0.04809	0.03088

The Dagum-I 3P distribution also shows a significant fit, though not as strong as the Log-Logistic 3P, as reflected by moderate KS and AD values (0.0417 and 4.0282).

For benzene, both the Burr XII 3P and Dagum-I 3P distributions provide significant fits, with similar KS and AD statistics (Burr XII 3P: KS=0.0349, AD=2.8308; Dagum-I 3P: KS=0.0313, AD=2.7869), although the Burr-XII has slightly more significant p-values. The Log-Logistic distribution performs poorly, with high KS, CVM and AD statistics (AD=4.7209) and low p-values, suggesting that it is not a suitable model for benzene. Overall, the BurrXII 3P consistently performs well for xylene and benzene where it provides significant and reliable fits. The Log-Logistic 3P distribution performs best for toluene but struggles to fit xylene and benzene accurately. The Dagum-I 3P distribution offers decent performance for xylene and benzene, highlighting the importance of selecting the most appropriate model for each pollutant in air quality assessment.

Graphical Assessment of Model Fits: Figures 3, 4 and 5 compare empirical and theoretical distribution functions for xylene, toluene and benzene, modeled using the Burr XII 3P, Log-Logistic 3P and Dagum-I 3P distributions fitted via maximum likelihood estimation (MLE). Each figure displays the PDF and CDF alongside the fitted theoretical versions, allowing for visual assessment of model accuracy. Q-Q and P-P plots are also included to evaluate model fit by comparing observed and theoretical quantiles and cumulative probabilities.

In these plots, closer alignment along the 45-degree line indicates a better fit, showing that the empirical data closely follows the theoretical distribution. This analysis of PDFs, CDFs, Q-Q and P-P plots collectively identifies the most suitable model for each pollutant, offering critical insights into the statistical behavior of these volatile organic compounds and enhancing predictive accuracy in air quality assessments.

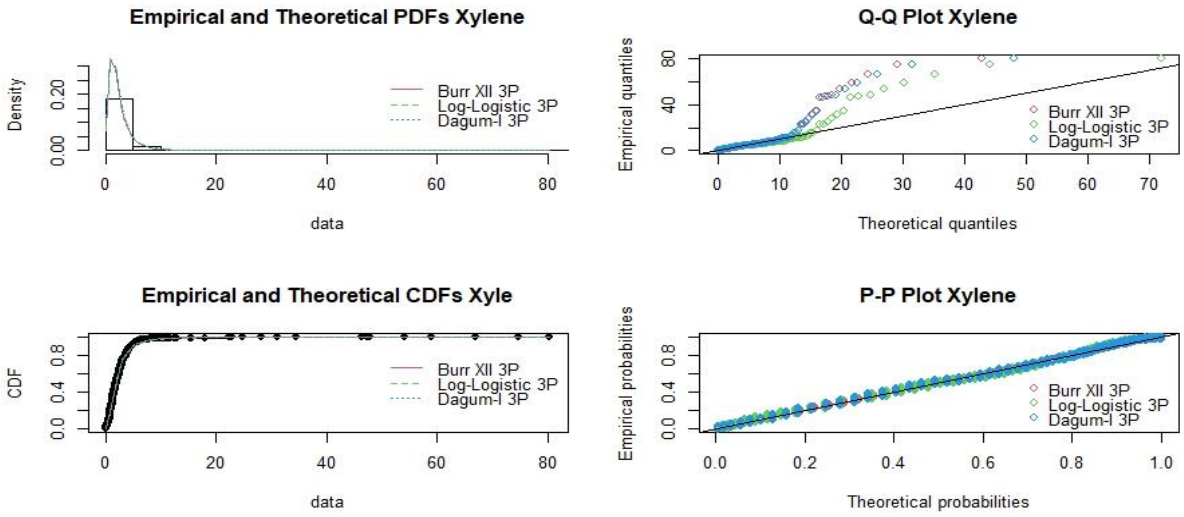

Figure 3: Empirical and Theoretical PDFs, CDFs, Q-Q and P-P Plots for Xylene using Burr XII 3P, Log-Logistic 3P and Dagum-I 3P Distributions

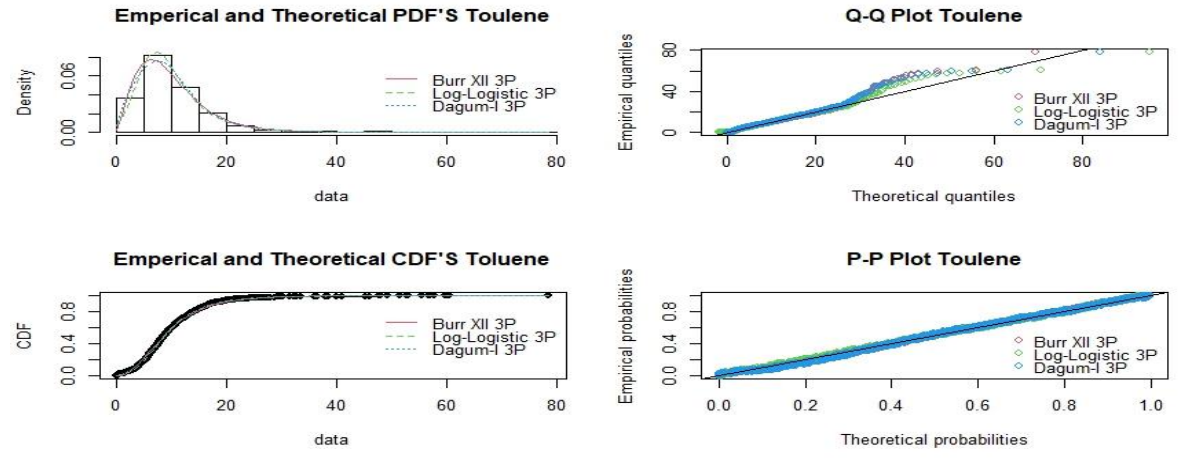

Figure 4: Empirical and Theoretical PDFs, CDFs, Q-Q and P-P Plots for Toluene using Burr XII 3P, Log-Logistic 3P and Dagum-I 3P Distributions

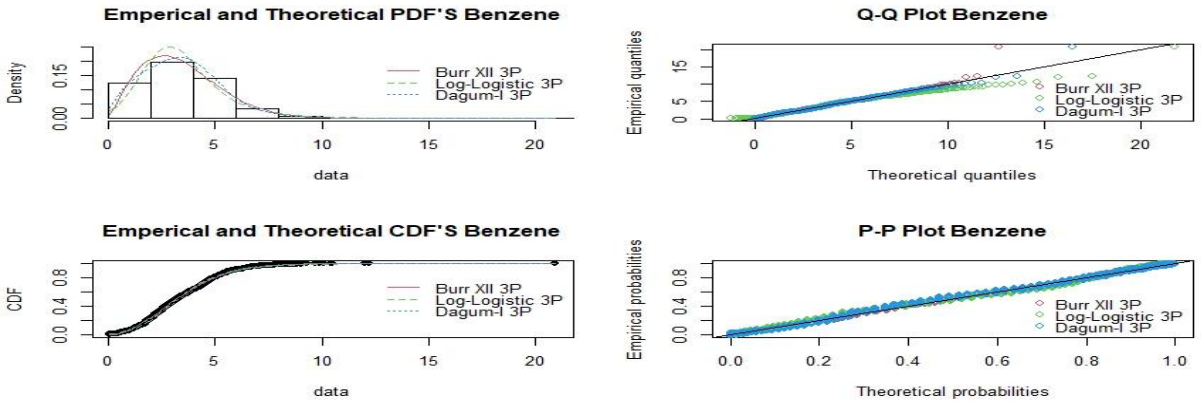

Figure 5: Empirical and Theoretical PDFs, CDFs, Q-Q and P-P Plots for Benzene using Burr XII 3P, Log-Logistic 3P and Dagum-I 3P Distributions

Table 5
Performance Metrics for each Model of Air Pollutant Concentrations

Distribution	Pollutant	LL	AIC	BIC	CAIC	HQIC	ABIC
Burr XII 3P	Xylene	-3018.396	6042.791	6058.975	6061.98	6048.8	6058.98
	Toluene	-5149.865	10305.73	10321.91	10324.91	10311.74	10321.91
	Benzene	-3252.617	6511.233	6527.416	6530.42	6517.24	6527.42
Log-Logistic 3P	Xylene	-3026.334	6058.668	6074.852	6077.85	6064.67	6074.85
	Toluene	-5112.609	10231.22	10247.4	10250.4	10237.22	10247.41
	Benzene	-3273.205	6552.411	6568.594	6571.59	6558.41	6568.60
Dagum-I 3P	Xylene	-3012.997	6031.994	6048.177	6051.18	6037.99	6048.18
	Toluene	-5117.143	10240.29	10256.47	10259.47	10246.29	10256.47
	Benzene	-3240.747	6487.494	6503.677	6506.68	6493.45	6503.68

In figure 3, the Dagum-I 3P distribution provides the best overall fit for xylene, closely matching the empirical data in both the PDF and CDF plots and showing minimal deviations in the Q-Q and P-P plots. The Burr XII 3P distribution performs reasonably well but with slight deviations at the extremes, while the Log-Logistic 3P distribution struggles with higher concentrations, as reflected in the Q-Q and P-P plots. For toluene (Figure 4), the Log-Logistic 3P distribution shows the best fit, particularly in the Q-Q and P-P plots, with the Burr XII 3P and Dagum-I 3P distributions performing moderately but with noticeable deviations at the tails.

As shown in figure 5, both the Burr XII 3P and Dagum-I 3P distributions fit Benzene well, while the Log-Logistic 3P distribution shows larger deviations, particularly at higher concentrations, indicating a weaker fit. Overall, the Dagum-I 3P distribution consistently offers the best fit for xylene and benzene, while the Log-Logistic 3P distribution performs best for toluene. The Burr XII 3P distribution performs adequately across all pollutants but shows slight limitations at extreme values, as seen in the comparative plots.

Model Evaluation using Performance Metrics: This study utilized the information criteria AIC, BIC, CAIC, HQIC and ABIC along with log-likelihood to evaluate the fit of three probability distributions-Burr XII 3P, Log-Logistic 3P and Dagum-I 3P-in modeling xylene, toluene and benzene concentrations. These metrics provide a quantitative basis for selecting the distribution that best fits each pollutant's concentration data, as lower values indicate a better fit to the observed data. Table 5 summarizes the performance metrics, offering insights into the suitability of each distribution for the air pollutant data.

For xylene, the Dagum-I 3P distribution achieved the lowest values across all criteria (AIC = 6031.994, BIC = 6048.177, CAIC = 6051.18, HQIC = 6037.99, ABIC = 6048.18), suggesting it is the most suitable model for capturing xylene's variability. These metrics indicate that the Dagum-I 3P distribution has a more efficient balance between model complexity and data fit for xylene, making it preferable to Burr XII 3P and Log-Logistic 3P. In the case of toluene, the Log-Logistic 3P distribution outperformed the others, reflected in its lower values for AIC (10231.22), BIC (10247.4), CAIC (10250.4), HQIC (10237.22) and ABIC (10247.41). This suggests that the Log-Logistic 3P distribution better represents the underlying structure of toluene concentrations, likely due to its flexibility in accommodating the specific data patterns of toluene.

Thus, the Log-Logistic 3P distribution is considered the best fit for modelling toluene variability. For benzene, the Dagum-I 3P distribution again emerged as the superior model, with the lowest AIC (6487.494), BIC (6503.677),

CAIC (6506.68), HQIC (6493.45) and ABIC (6503.68) values among the three distributions. This pattern demonstrates the Dagum-I 3P distribution's effectiveness in capturing benzene's distributional characteristics, likely due to its heavy-tailed nature, which matches the properties of benzene data. These findings emphasize the effectiveness of the Dagum-I 3P model for xylene and benzene and Log-Logistic 3P for toluene, highlighting the importance of model selection based on empirical criteria. This approach enhances understanding of pollutant dynamics and supports data-driven decisions in air quality management for Visakhapatnam.

Model Validation using 10-Fold Cross-Validation: Table 6 details the average information criteria obtained via 10-fold cross-validation for Burr XII 3P, Log-Logistic 3P and Dagum-I 3P models applied to the airborne volatile organic compounds (VOCs) xylene, toluene and benzene concentrations. The criteria include AIC, BIC, CAIC, HQIC, ABIC and log-likelihood, which indicate model performance, with lower values suggesting better fit.

For xylene, the Dagum-I 3P model demonstrated the best performance, with the lowest AIC (5428.99), BIC (5444.86), CAIC (5430.99), HQIC (5434.91) and ABIC (5429.99), alongside the highest log-likelihood (-2711.50). These metrics suggest that Dagum-I 3P effectively captures the underlying distribution of xylene concentrations, reflecting its balance of accuracy and complexity. The Burr XII 3P model performed reasonably well for xylene, as indicated by its second-lowest AIC (5438.67) and BIC (5454.54) values, though it trailed the Dagum-I 3P model. Conversely, the Log-Logistic 3P model had the highest AIC (5453.04) and other criteria values for xylene, indicating a relatively weaker fit compared to Dagum-I 3P.

In the case of toluene, the Log-Logistic 3P model provided the best fit, as evidenced by its lowest AIC (9208.48), BIC (9224.34), CAIC (9210.48), HQIC (9214.40) and ABIC (9209.48) values. These results suggest that the Log-Logistic 3P model effectively captures the unique distributional characteristics of toluene, providing a more accurate fit than the other distributions. The model's log-likelihood of -4601.24, although not the highest among the three, further supports its robustness for toluene. The Dagum-I 3P model showed comparable performance with an AIC of 9216.53 and a log-likelihood of -4605.27, whereas the Burr XII 3P model had significantly higher AIC and BIC values (9275.37 and 9291.23 respectively), marking it as the least suitable model for toluene.

For Benzene, Dagum-I 3P again emerged as the best-fitting model, achieving the lowest AIC (5839.09), BIC (5854.95), CAIC (5841.09), HQIC (5845.00) and ABIC (5840.09) values, alongside the highest log-likelihood (-2916.54). This distribution's performance for benzene indicates a strong alignment with the dataset, particularly in capturing the heavy-tailed characteristics often seen in benzene. The Burr XII 3P model was followed with slightly higher values, whereas the Log-Logistic 3P model had the highest AIC (5897.54), indicating a less optimal fit for Benzene concentrations.

In summary, the Dagum-I 3P distribution consistently emerged as the most effective model for both xylene and benzene, while the Log-Logistic 3P model demonstrated a strong fit for Toluene. The Burr XII 3P model generally performed the least well across all pollutants. These findings highlight the necessity of selecting pollutant-specific models that accurately capture the distributional nuances of each VOC, thereby enhancing our understanding of air quality dynamics. This model-specific approach is essential for accurate environmental monitoring and the development of data-driven strategies to manage VOC concentrations in Visakhapatnam, ultimately supporting informed policy decisions aimed at protecting public health.

Comparisons via Performance Criteria and Cross-Validation: Model performance metrics and cross-validation results to evaluate the robustness of the Burr XII 3P, Log-Logistic 3P and Dagum-I 3P distributions for modeling VOC (xylene, toluene and benzene) concentrations were compared. Multiple performance metrics included AIC, BIC, CAIC, HQIC and ABIC, along with log-likelihood to assess model fit quality for each distribution.

Table 6
Average Information Criteria for Model Distributions of Xylene, Toluene and Benzene using 10-Fold Cross-Validation

Distribution	Airborne	AIC	BIC	CAIC	HQIC	ABIC	LL
Burr XII 3P	Xylene	5438.67	5454.54	5440.67	5444.59	5439.67	-2716.34
	Toluene	9275.37	9291.23	9277.37	9281.29	9276.37	-4634.68
	Benzene	5863.53	5879.4	5865.53	5869.45	5864.53	-2928.77
Log-Logistic 3P	Xylene	5453.04	5468.91	5455.04	5458.96	5454.04	-2723.52
	Toluene	9208.48	9224.34	9210.48	9214.4	9209.48	-4601.24
	Benzene	5897.54	5913.4	5899.54	5903.46	5898.54	-2945.77
Dagum-I 3P	Xylene	5428.99	5444.86	5430.99	5434.91	5429.99	-2711.5
	Toluene	9216.53	9232.4	9218.53	9222.45	9217.53	-4605.27
	Benzene	5839.09	5854.95	5841.09	5845	5840.09	-2916.54

These metrics provide a multidimensional view of model performance, helping to identify the best fit for each VOC based on a balance between accuracy and complexity.

For xylene and benzene, the Dagum-I 3P distribution consistently achieved the lowest AIC, BIC, CAIC, HQIC and ABIC values, indicating it as the most suitable model. For toluene, the Log-Logistic 3P distribution demonstrated the best fit, with consistently lower criteria values, underscoring its strength in capturing the variability of toluene concentrations. The Burr XII 3P model, in contrast, showed relatively higher values across metrics, suggesting a limited ability to model these VOCs accurately.

These findings appreciate applying 10-fold cross-validation, further validating model reliability through averaged performance metrics across multiple data segments. This method reduces the potential for overfitting and strengthens the evidence of model robustness. Cross-validation results confirmed that Dagum-I 3P maintained superior performance for xylene and benzene, while the Log-Logistic 3P model remained optimal for toluene, reinforcing its effectiveness. The Burr XII 3P model again ranked lower, confirming its limitations in accurately representing VOC data patterns.

Dagum-I 3P model provides the best fit for xylene and benzene, achieving the lowest average criteria scores. For toluene, the Log-Logistic 3P model consistently performs best, reflecting its capability to capture the compound's variability. Across pollutants, the Burr XII 3P model ranks lower, underscoring its limitations in accurately representing these VOCs. This combined approach ensures that the selected models not only align well with observed VOC patterns but are also robust and reliable for use in environmental monitoring and predictive analysis. The results support targeted, data-driven decision-making for air quality management in Visakhapatnam, facilitating more effective mitigation strategies to address VOC pollution.

Discussion

This study provides a comparative analysis of statistical distribution models for predicting VOC concentrations, focusing on model effectiveness for different pollutants in Visakhapatnam and identifying the most suitable model for each compound. Our findings highlight the importance of pollutant-specific models that align with the unique distributional characteristics of each VOC. By applying Burr XII 3P, Log-Logistic 3P and Dagum-I 3P distributions, we observed distinct patterns that reveal both model suitability and the nuanced behavior of VOC concentrations in an urban-industrial setting. The Dagum-I 3P distribution proved most effective in capturing the heavy-tailed distributions of xylene and benzene, as indicated by its lowest KS, AD and CvM test values, as well as AIC, BIC, CAIC, HQIC and ABIC scores, alongside high p-values indicating a strong fit. The heavy tails observed in xylene and benzene distributions suggest episodic concentration peaks, possibly driven by intermittent emissions or specific environmental conditions in which the Dagum-I 3P model accommodates well. This distribution's ability to handle high skewness and kurtosis aligns with the concentration patterns of these VOCs, reflecting the impact of high-emission events common in industrial areas. Its robustness was further validated through 10-fold cross-validation, confirming its effectiveness in modeling heavy-tailed data. Conversely, toluene's concentration profile was best captured by the Log-Logistic 3P distribution, which achieved consistently lower goodness-of-fit and information criteria values than other models. The moderate skewness and comparatively lower kurtosis of toluene concentrations may account for this outcome, as the Log-Logistic 3P model is well-suited for data with moderate tail behavior.

This fit may reflect a more consistent concentration pattern, likely to be influenced by stable emission sources such as vehicular emissions and other diffuse sources typical in urban environments, differing from the more episodic emissions observed in xylene and benzene. The Burr XII 3P distribution, though flexible, consistently ranked lower than the Dagum-I and Log-Logistic models in terms of goodness-of-fit and selection criteria across all three VOCs. Its comparatively poorer performance highlights limitations in accurately representing the unique tail behaviors and skewness of these pollutants, underscoring the need for pollutant-specific modeling strategies instead of a one-size-fits-all approach.

These findings emphasize the practical importance of selecting models tailored to each pollutant, especially in industrial areas like Visakhapatnam, where VOC emissions vary by type and are affected by complex environmental factors. Tailoring models to each pollutant enhances accuracy in forecasting VOC levels, which is crucial for regulatory planning and public health assessment. Additionally, time series visualizations enabled us to observe seasonal trends and responses to external events such as the COVID-19 lockdown. Elevated VOC concentrations during colder months suggest that seasonal inversion effects contribute to pollutant accumulation, while declines during the lockdown highlight the role of human activity in VOC levels. These observations contextualize the extreme values in the data, validating the selected distribution models' capacity to account for variability influenced by both seasonal and anthropogenic factors.

This study underscores the utility of performance metrics and cross-validation techniques in environmental modeling, advocating for their broader use in air quality assessments. While the current analysis focuses on xylene, toluene and benzene, future research could extend this approach to other VOCs or pollutants with similar complex distributional properties. Additional studies could also examine interactions between VOC concentrations and meteorological factors to further enhance predictive accuracy in urban-industrial areas.

In summary, our pollutant-specific modeling approach, validated through rigorous goodness-of-fit testing and cross-validation, provides a robust framework for accurately capturing VOC concentration distributions. These findings are critical for environmental policy-making, enabling data-driven strategies for air quality management tailored to the distinct pollution profiles of urban-industrial contexts like Visakhapatnam.

Conclusion

This research presents a comprehensive modeling approach for VOC (Xylene, Toluene and Benzene) concentrations in Visakhapatnam, showing that advanced statistical distributions like Dagum-I 3P and Log-Logistic 3P effectively capture the distributional characteristics of xylene, benzene and toluene. By applying a suite of goodness-of-fit tests, performance metrics and cross-validation, we established that the Dagum-I 3P distribution is most suitable for the highly skewed and heavy-tailed distributions of xylene and benzene, while the Log-Logistic 3P distribution best represents the moderately skewed toluene concentrations. These findings emphasize the importance of selecting pollutant-specific models, as each compound's unique distributional profile affects model accuracy and reliability.

This tailored approach to VOC modeling offers a valuable tool for environmental monitoring, supporting more precise forecasts of VOC levels and risk assessment in regions with similar industrial profiles. Our methodology provides a framework that can guide air quality management, helping policymakers to prioritize regulatory efforts and refine interventions for specific pollutants. Future work could extend this modeling approach to additional VOCs or regions, further enhancing data-driven environmental health strategies across urban-industrial contexts.

References

1. Ahmat H., Yahaya A.S. and Ramli N.A., Prediction of PM_{10} extreme concentrations in urban monitoring stations in Selangor, Malaysia using three parameters extreme value distributions (EVD), *Jurnal Teknologi*, **77(32)**, 37-46 **(2015)**

2. Atari D.O. and Luginaah I.N., Assessing the distribution of volatile organic compounds using land use regression in Sarnia, "Chemical Valley," Ontario, Canada, *Environmental Health*, **8**, 1-14 **(2009)**

3. Bang J.H., Oh I., Kim S., You S., Kim Y., Kwon H.J. and Kim G.B., Modeling the effects of pollutant emissions from large industrial complexes on benzene, toluene and xylene concentrations in urban areas, *Environmental Health and Toxicology*, **32(4)**, 1-12 **(2017)**

4. Biçer C., Bakouch H.S., Biçer H.D., Alomair G., Hussain T. and Almohisen A., Unit Maxwell-Boltzmann Distribution and Its Application to Concentrations Pollutant Data, *Axioms*, **13(4)**, 226 **(2024)**

5. Chaudhary A.K., Telee L.B.S., Karki M. and Kumar V., Statistical analysis of air quality dataset of Kathmandu, Nepal, with a new extended Kumaraswamy exponential distribution, *Environmental Science and Pollution Research*, **31(14)**, 21073-21088 **(2024)**

6. Gavriil I., Grivas G., Kassomenos P., Chaloulakou A. and Spyrellis N., An application of theoretical probability distributions to the study of PM_{10} and $PM^{2.5}$ time series in Athens, Greece, *Global NEST Journal*, **8(3)**, 241-251 **(2006)**

7. Hein M.J., Waters M.A., van Wijngaarden E., Deddens J.A. and Stewart P.A., Issues when modeling benzene, toluene and xylene exposures using a literature database, *Journal of Occupational and Environmental Hygiene*, **5(1)**, 36-47 **(2007)**

8. Kamani H., Baniasadi M., Abdipour H., Mohammadi L., Rayegannakhost S., Moein H. and Azari A., Health risk assessment of BTEX compounds (benzene, toluene, ethylbenzene and xylene) in different indoor air using Monte Carlo simulation in Zahedan City, Iran, *Heliyon*, **9(9)**, 1-11 **(2023)**

9. Kazi Z., Filip S. and Kazi L., Predicting $PM_{2.5}$, PM_{10}, SO_2, NO_2, NO and CO air pollutant values with linear regression in R language, *Applied Sciences*, **13(6)**, 3617 **(2023)**

10. Natarajan S.K., Shanmurthy P., Arockiam D., Balusamy B. and Selvarajan S., Optimized machine learning model for air quality index prediction in major cities in India, *Scientific Reports*, **14(1)**, 6795 **(2024)**

11. Naz F., Mccann C., Fahim M., Cao T.V., Hunter R., Viet N.T. and Duong T.Q., Comparative analysis of deep learning and statistical models for air pollutants prediction in urban areas, *IEEE Access*, **11**, 64016-64025 **(2023)**

12. Nurmatov U.B., Tagiyeva N., Semple S., Devereux G. and Sheikh A., Volatile organic compounds and risk of asthma and allergy: a systematic review, *European Respiratory Review*, **24(135)**, 92-101 **(2015)**

13. Rashnuodi P., Dehaghi B.F., Rangkooy H.A., Amiri A. and Mohi Poor S., Evaluation of airborne exposure to volatile organic compounds of benzene, toluene, xylene and ethylbenzene and its relationship to biological contact index in the workers of a petrochemical plant in the west of Iran, *Environmental Monitoring and Assessment*, **193**, 1-10 **(2021)**

14. Shen S., Yuan L. and Zeng S., An effort to test the embryotoxicity of benzene, toluene, xylene and formaldehyde to murine embryonic stem cells using airborne exposure technique, *Inhalation Toxicology*, **21(12)**, 973-978 **(2009)**

15. Thupeng W.M., Use of the three-parameter Burr XII distribution for modeling ambient daily maximum nitrogen dioxide concentrations in the Gaborone fire brigade, *American Scientific Research Journal for Engineering, Technology and Sciences (ASRJETS)*, **26(2)**, 18-32 **(2016)**.

(Received 08th November 2024, accepted 04th December 2024)