

MIXTURE DISTRIBUTION ANALYSIS OF DAILY AVERAGE PM_{2.5} CONCENTRATIONS IN BANGKOK, THAILAND

Tanatip Hanpayak and Tidarut Areerak*

Received: June 11, 2024; Revised: October 17, 2024; Accepted: October 30, 2024

Abstract

Air pollution, specifically PM_{2.5}, poses significant health risks and global environmental challenges. Bangkok, the capital city of Thailand, also experiences severe levels of PM_{2.5}. The objective of this study is to determine the optimal probability distribution for PM_{2.5} concentration in Bangkok. Daily average PM_{2.5} concentrations from January 1, 2018, to December 31, 2022, were analyzed at 10 monitoring sites in the Bangkok area. The concentration patterns of PM_{2.5} were characterized using several statistical distributions, including lognormal, gamma, and Weibull distributions, as well as 2-mixture and 3-mixture distributions. Akaike and Bayesian information criteria were utilized to identify the optimal statistical distribution fit. Results indicate that the lognormal-lognormal distribution is most suitable for the PM_{2.5} concentration data from stations situated along roadways. However, stations in residential areas exhibit variability in statistical distributions, likely due to distinct factors at each location. Nevertheless, it is observed that each station in Bangkok is matched to a distribution with lognormal components.

Keywords: Fine particulate matter (PM_{2.5}); Information criteria; Mixture distribution; Statistical distribution

Introduction

Air pollution negatively influences human health and reduces life expectancy, while also significantly impacting society and the economy. The detrimental effects extend beyond health concerns, encompassing severe economic consequences. The World Health Organization identifies particulate matter as the primary air pollutant, emphasizing its substantial impact. Particulate matter, prevalent in both urban and non-urban areas, is recognized as a complex

mixture of components with varied chemical and physical properties. Fine particulate matter (PM_{2.5}), characterized by particles with a diameter equal to or less than 2.5 micrometers, is a representative example (World Health Organization, 2021). PM_{2.5} comprises varying proportions of water and key components such as sulfates, acids, nitrates, elemental carbon, organic carbon, and metals, with the composition differing based on their sources

*School of Mathematical Sciences and Geoinformatics, Institute of Science, Suranaree University of Technology, Thailand.
E-mail: tanatip.b6120009@gmail.com; E-mail: tidarut@g.sut.ac.th*

* Corresponding author

DOI: <https://doi.org/10.55766/sujst-2024-05-e05095>

Suranaree J. Sci. Technol. 31(5):030232(1-10)

(Balmes and Eisner, 2016). Between 1999 and 2014, South and Southeast Asia faced significant challenges, including a 38% increase in premature mortality associated with particulate matter. This period recorded a total of 1,447,000 premature deaths caused by PM_{2.5}, highlighting the significant threat that air pollution poses to public health (Shi *et al.*, 2018).

Air pollution, particularly PM_{2.5}, presents significant health risks and environmental challenges globally, especially in urban areas. Thailand consistently ranked 23rd worldwide in terms of PM_{2.5} levels, with estimated average concentrations reaching 26.4 µg/m³ in 2018. Although these levels declined from 2019 to 2022, they increased again in 2023 (IQAir AirVisual, 2018, 2019, 2020, 2021, 2022, 2023). Despite periodic fluctuations, there is a discernible trend suggesting that PM_{2.5} levels will rise again over the next five years. This underscores Thailand's ongoing struggle with air pollution and highlights the need for sustained efforts to mitigate its impact on public health and the environment. The Pollution Control Department (2022) identified Saraburi, Samut Songkhram, Bangkok, Phitsanulok, and Nong Khai as the top five provinces in Thailand with significant air quality issues, exceeding air quality standards for 97, 77, 68, 57, and 54 days per year, respectively. Therefore, studying PM_{2.5} is crucial for both the general public and the government. It allows the public to be aware of and protect themselves from the effects of PM_{2.5}, and it assists government agencies in formulating policies and preventing issues arising from PM_{2.5}.

Understanding the statistical distribution of data through probability distributions is fundamental to data analysis and summarization. Numerous studies have applied statistical distributions to various types of data, including financial data, insurance data, medical data, epidemic outbreaks, disaster-related data, and particulate matter studies. In the context of particulate matter, Mishra *et al.* (2021) examined the distribution of PM_{2.5} in five countries: Brazil, China, France, India, and the United States during their respective lockdown periods in 2020 and the corresponding period in 2019. They utilized gamma, lognormal, and Weibull distributions in their analysis and found that the gamma distribution provided the best fit for PM_{2.5} concentrations across these countries. The use of these three distributions has been a focal point in numerous research studies exploring particulate matter distributions, including those by Lu (2004), Mijić *et al.* (2009), and Wang *et al.* (2013).

Given that the composition of PM_{2.5} includes several major components, such as sulfates, acids,

nitrites, elemental carbon, organic carbon, and trace metals (Balmes and Eisner, 2016), it is crucial to study the statistical distribution of PM_{2.5} using a mixture distribution to account for this complexity. Taylor *et al.* (1986) examined the distribution of air pollution particles, including ozone, carbon monoxide, sulfur dioxide, oxides of nitrogen, nitrogen dioxide, and nitrogen, which are major components of PM_{2.5}. They found that these particles generally follow lognormal, gamma, and Weibull distributions. Therefore, this research employs mixture distributions to analyze the statistical distribution of PM_{2.5} concentration, incorporating these three distributions as components.

The mixture model is a flexible method for modeling multimodal data, particularly when datasets include multiple subgroups or subpopulations. This approach has seen diverse applications across numerous fields. Bernardi (2013) demonstrated that linear combinations of multivariate skew normal mixtures can be expressed as finite mixtures of univariate skew normal distributions, providing analytical formulas for various risk measures and advancing the use of skew normal distributions in risk assessment. Korolev *et al.* (2015) utilized mixtures of normal distributions with time-varying parameters to analyze trends in climate data, specifically examining turbulent heat fluxes derived from data collected by the National Centers for Environmental Prediction and National Center for Atmospheric Research (NCEP-NCAR) during 1948 to 2008 in the Atlantic. This approach captures both long-term trends and short-term variability. Malá (2016) investigated unemployment duration in the Czech Republic using a mixture of lognormal distributions to analyze interval-censored data from a Labor Force Survey, with a focus on gender-based differences in unemployment duration. Maximum likelihood estimates in the R programming language were used for the analysis. Punzo *et al.* (2018) proposed a mixture of contaminated gamma distributions for modeling skewed and multimodal insurance and economic data, estimating model parameters via maximum likelihood and evaluating Value-at-Risk. Their results highlighted the model's flexibility, showing superior performance over traditional benchmarks. Channoufi *et al.* (2018) addressed image denoising challenges by introducing an unsupervised algorithm based on finite mixtures of bounded generalized Gaussian distributions (BGGMD), achieving more effective noise reduction compared to standard Gaussian methods. Aquilanti *et al.* (2021) introduced a parameter estimation method for finite mixture models using Mean Field Games theory, which identifies critical points of the log-likelihood for

Table 1. Probability density function

Distribution	Notation	Probability density function
Gamma distribution	$X \sim \text{gamma}(a, b)$	$f(x; a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-\frac{x}{b}}$
Lognormal distribution	$X \sim \text{lognormal}(\mu, \sigma)$	$f(x; \mu, \sigma) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$
Weibull distribution	$X \sim \text{Weibull}(a, b)$	$f(x; a, b) = \frac{a}{b} \left(\frac{x}{b}\right)^{a-1} e^{-\left(\frac{x}{b}\right)^a}$

Bernoulli mixtures and extends this approach to categorical distributions. This technique has shown efficacy in cluster analysis applications. Tajuddin *et al.* (2022) proposed mixture models for count data with multimodal distributions, developing ten mixture distributions using Poisson (P), Poisson Lindley (PL), Negative Binomial (NB), and Negative Binomial Lindley (NBL) components. They found the P-NB, PL-NB, and NB-NB models are the most suitable for describing data from subpopulations with different means and dispersions. Each of these studies underscores the finite mixture model's adaptability, facilitating nuanced analyses across fields by accounting for heterogeneous data characteristics.

In air quality analysis, especially regarding PM_{2.5} particles, understanding their distribution requires accounting for the complex mixture of chemical compounds they contain. Given this complexity, employing a mixture distribution model that is suitable for data from diverse population groups provides an effective approach for analyzing PM_{2.5} concentration distributions. From previous research, the chemical components of PM_{2.5} follow lognormal, gamma, and Weibull distributions (Taylor *et al.*, 1986). Therefore, this research utilizes a mixture distribution incorporating lognormal, gamma, and Weibull components to gain more precise insights into the distribution patterns of PM_{2.5} concentration.

This research aims to analyze the daily average PM_{2.5} concentration data collected from air quality monitoring stations in Bangkok. The study distinguishes between stations located near roads and those situated in urban areas, evaluating which mixture of statistical distributions best fits the data by comparing Akaike and Bayesian Information Criteria values. The maximum likelihood method is employed to estimate the parameters.

Materials and Methods

Statistical Distribution

Statistical distributions describe the probabilities of potential outcomes within a dataset or experiment, providing insight into data variability

and patterns. Various statistical distributions have a crucial role in practical applications across various fields, including sciences, engineering, economics, and health. In this study, gamma, lognormal, and Weibull distributions are employed to analyze the distribution of PM_{2.5} concentration. Table 1 presents the probability density functions of random variables associated with these distributions, where a and b denote the shape and scale parameters, respectively, while the parameters μ and σ represent the mean and standard deviation of $Y = \ln X$, respectively (Krishnamoorthy, 2006).

Mixture Distribution

Occasionally, a distribution arises from a combination of multiple component distributions. Each component distribution represents a distinct subpopulation or data source, whereas the mixture distribution represents the overall distribution of the combined data. A mixture distribution is commonly employed in situations where the data comes from a heterogeneous population. For a finite mixture consisting of k subpopulations, the probability density function is defined by Equation (1):

$$f_X(x) = \sum_{i=1}^k p_i f_i(x), \quad (1)$$

where f_i and p_i denote the probability density function and the weight of the i th subpopulation, respectively. The weights p_i satisfy conditions $p_i \geq 0$ and $\sum_{i=1}^k p_i = 1$ (Kvam and Vidakovic, 2007).

Let $X_1, \dots, X_k$ be random variables from the k components and let X be a random variable from the mixture distribution. The expected value of the mixture distribution can be calculated by Equation (2):

$$\mu = E(X) = \sum_{i=1}^k p_i \mu_i, \quad (2)$$

where $\mu_i = E(X_i)$ denotes the expected value of the i th component distribution.

The variance of a mixture distribution can be calculated by Equation (3):

$$\text{Var}(X) = \sum_{i=1}^k p_i (\sigma_i^2 + \mu_i^2) - \mu^2, \quad (3)$$

where $\sigma_i^2 = \text{Var}(X_i)$ denotes the variance of the i th component distribution.

The n th moment of a mixture distribution can be calculated by Equation (4):

$$E(X^n) = \sum_{i=1}^k p_i E(X_i^n), \quad (4)$$

where $E(X_i^n)$ denotes the n th moment of the i th component distribution.

Mixture distributions provide a flexible approach to modeling complex data, particularly when datasets include multiple subgroups or patterns. They are highly adaptable to biological and physical phenomena, allowing for the identification of distinct subpopulations within natural and biological contexts. By distinguishing between component distributions, mixture models can uncover underlying patterns in the data. Overall, mixture distributions are powerful tools for applications that require flexibility in handling diverse data sources, enabling comprehensive analysis and enhanced pattern recognition across various fields (Everitt and Hand, 1981).

This research employs gamma, lognormal, and Weibull distributions as components of non-mixture, 2-mixture, and 3-mixture distributions in order to analyze the distribution of the dataset.

Maximum Likelihood Method

Estimation of the parameters in its probability distribution function is necessary for studying a statistical distribution. The maximum likelihood method is the most common approach for estimating these values. Consider a random sample $X_1, X_2, \dots, X_n$ taken from a random variable X , where $X_1, X_2, \dots, X_n$ are independent and identically distributed random variables. The likelihood function is defined as the product of the individual probability density values as shown in Equation (5):

$$L(\theta; x) = \prod_{i=1}^n f(x_i; \theta), \quad (5)$$

where $f(x; \theta)$ is the probability density function of X , $x = (x_1, x_2, \dots, x_n)$, and θ is an unknown parameter from a sample space Ω . The point estimator of θ is denoted by $\hat{\theta}$, and $\hat{\theta}$ is called the maximum likelihood estimator (MLE) of θ , if it satisfies Equation (6):

$$\hat{\theta} = \arg \max L(\theta; x). \quad (6)$$

This implies that $\hat{\theta}$ maximizes the function $L(\theta; x)$. In practice, the natural logarithm of the likelihood function is often times more convenient to work with; it is denoted by $l(\theta; x)$ and defined by Equation (7):

$$l(\theta; x) = \ln L(\theta; x) = \sum_{i=1}^n \ln f(x_i; \theta). \quad (7)$$

Since the logarithm function is a one-to-one function, using $l(\theta; x)$ instead of $L(\theta; x)$ does not discard information. Therefore, the estimator $\hat{\theta}$ maximizes the function $L(\theta; x)$ as well as $l(\theta; x)$. The maximum likelihood estimator $\hat{\theta}$ can be determined by solving Equation (8):

$$\frac{\partial}{\partial \theta} l(\theta; x) = \frac{\partial}{\partial \theta} \ln L(\theta; x) = 0. \quad (8)$$

In the case where this equation is nonlinear, numerical methods can be employed to find the solution (Hogg *et al.*, 2019).

This research employs the maximum likelihood method to estimate the parameters in mixture distributions. The random variables in each component are assumed to be independent. This independence assumption simplifies the maximum likelihood method process, improving the consistency of results across analyses (McLachlan and Peel, 2000). However, determining the appropriate number of components in a mixture distribution remains a significant limitation, as selecting an excessive number of components can overcomplicate parameter estimation and reduce model interpretability (McLachlan *et al.*, 2019).

Information Criteria

Optimizing the selection of an appropriate model for a given dataset involves balancing model accuracy and complexity. This study employed two widely used information criteria, namely the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), to determine the most suitable distribution for fitting the dataset, where l represents the maximum value of the loglikelihood function for the model, k represents the number of estimated parameters within the model, and n represents the sample size, as shown in Table 2 (Ding *et al.*, 2018).

Table 2. The definition of AIC and BIC

Name	Definition
Akaike information criterion	AIC = $-2l + 2k$
Bayesian information criterion	BIC = $-2l + k(\ln(n))$

A lower AIC or BIC value indicates a better goodness of fit. Models with lower AIC and BIC values are preferred. The BIC tends to favor simpler models compared to AIC, given its stronger penalty on model complexity. Consequently, AIC and BIC values are computed for each model, and the one with the lowest criterion value is selected.

Data and Methodology

This study examines a dataset containing daily average PM_{2.5} concentrations, which is provided by the Division of Air Quality and Noise Management Bureau, Pollution Control Department, Ministry of Natural Resources and Environment of Thailand. The dataset spans from January 1st, 2018, to December 31st, 2022. Analysis is conducted

using RStudio version 4.3.2 with the dataset organized into yearly intervals. The study concentrates on the 10 monitoring stations located in Bangkok, as outlined in Table 3 and depicted in Figure 1.

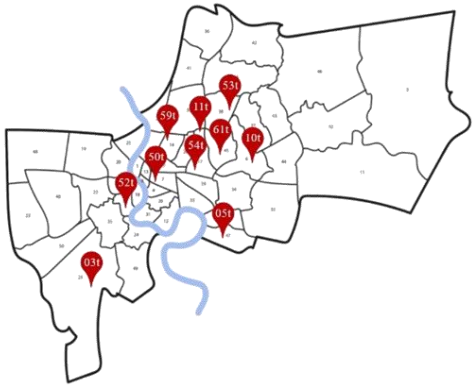


Figure 1. Geographical distribution of air quality monitoring stations in Bangkok

Table 3. Inventory of air quality monitoring stations in Bangkok

Code	Station name
05t	Thai Meteorological Department
10t	Khlong Chan Community Housing
11t	Huai Khwang Community Housing Authority
59t	Public Relations Department
61t	Bodindecha School
03t	Along Highway No. 3902
50t	Chulalongkorn Hospital
52t	Thonburi Sub-Electricity Authority
53t	Chokchai Metropolitan Police Station
54t	Din Daeng Community Housing

Table 4. Descriptive statistics of the PM_{2.5} concentration data

Station	Year	Descriptive Statistics of the PM _{2.5} Concentration Data					
		Mean	Median	SD	Variance	Min	Max
05t	2018	22.67	18	14.65	214.65	4	99
	2019	21.47	17	13.13	172.28	4	79
	2020	19.43	14	13.80	190.53	4	100
	2021	19.81	15	14.52	210.91	4	88
	2022	23.09	19	10.81	116.85	9	69
10t	2018	30.29	27	12.60	158.75	8	65
	2019	21.39	18	12.70	161.35	4	84
	2020	19.87	16	11.48	131.85	6	74
	2021	20.02	16	12.59	158.57	5	70
	2022	19.00	17	9.25	85.53	6	56
11t	2018	30.26	28	11.05	122.09	12	64
	2019	23.50	20	12.25	150.01	7	81
	2020	20.77	16	12.59	158.54	6	72
	2021	20.92	16	13.38	178.89	6	79
	2022	20.42	18	10.14	102.87	7	70
59t	2018	20.41	17	12.53	157.10	6	97
	2019	20.92	18	11.81	139.44	5	70
	2020	19.13	15	12.49	156.05	3	70
	2021	18.29	14	13.06	170.48	4	79
	2022	16.40	14	9.53	90.77	4	75
61t	2018	25.34	21	14.80	219.09	8	90
	2019	25.84	23	13.89	192.80	7	94
	2020	21.79	18	12.42	154.19	7	78
	2021	20.71	17	9.09	82.72	11	63
	2022	19.79	17	7.71	59.49	11	55
03t	2018	41.39	39	15.90	252.96	14	97
	2019	35.90	33	14.10	198.76	15	104
	2020	29.89	24	14.64	214.27	11	96
	2021	30.33	24	16.51	272.64	12	131
	2022	28.43	25	11.03	121.72	13	100
50t	2018	28.10	24	13.46	181.07	11	87
	2019	27.38	24	12.98	168.42	10	92
	2020	24.96	20	13.16	173.23	8	82
	2021	24.75	21	13.39	179.40	9	89
	2022	25.62	22	10.45	109.17	12	73
52t	2018	28.88	22	18.30	334.73	8	105
	2019	24.67	20	14.93	222.90	6	84
	2020	22.81	18	13.75	189.10	5	101
	2021	24.84	20	14.06	197.76	9	101
	2022	21.52	18	11.16	124.45	7	85
53t	2018	26.09	22	13.86	192.13	7	81
	2019	26.06	23	13.71	188.02	5	89
	2020	22.49	17	14.83	219.84	4	77
	2021	24.30	20	15.76	248.52	5	85
	2022	21.51	18	11.95	142.70	6	70
54t	2018	35.32	33	13.50	182.18	13	96
	2019	33.40	30	13.43	180.40	13	91
	2020	32.11	26	15.25	232.54	13	112
	2021	32.59	27	15.31	234.55	13	106
	2022	30.94	29	10.50	110.18	13	79

The PM_{2.5} concentration data for each station and year underwent analysis, involving the fitting of the data to various distribution types. This study categorizes distributions into three groups: non-mixture, 2-mixture, and 3-mixture. The non-mixture category comprises gamma (GM), lognormal (LN), and Weibull (W) distributions, while the 2-mixture category includes gamma-gamma (GM-GM), lognormal-lognormal (LN-LN), Weibull-Weibull (W-W), gamma-lognormal (GM-LN), gamma-Weibull (GM-W), and lognormal-Weibull (LN-W) distributions. The 3-mixture category includes gamma-gamma-gamma (GM-GM-GM), lognormal-lognormal-lognormal (LN-LN-LN), Weibull-Weibull-Weibull (W-W-W), gamma-gamma-lognormal (GM-GM-LN), gamma-gamma-Weibull (GM-GM-W), gamma-lognormal-lognormal (GM-LN-LN), gamma-lognormal-Weibull (GM-LN-W), gamma-Weibull-Weibull (GM-W-W), lognormal-lognormal-Weibull (LN-LN-W), and lognormal-Weibull-Weibull (LN-W-W) distributions. Utilizing the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), this research calculates values for each distribution to determine the most appropriate statistical distribution.

Parameter estimation is performed using the maximum likelihood method.

Results and Discussion

The descriptive statistics of PM_{2.5} concentration data for each station from 2018 to 2022, including the mean, median, standard deviation, variance, minimum, and maximum values are presented in Table 4.

The analysis of the probability density function of the PM_{2.5} concentration data in 2018 reveals multiple peaks and asymmetry, as shown in Figure 2. This indicates that the empirical probability density function of the PM_{2.5} concentration at each station is a combination of various distributions. These distributions correspond to the main components of PM_{2.5}, such as ozone, carbon monoxide, sulfur dioxide, oxides of nitrogen, nitrogen dioxides, and nitrogen oxide. Taylor *et al.* (1986) investigated the statistical distribution of these components and found that they fit the gamma, lognormal, and Weibull distributions. Consequently, this analysis focuses

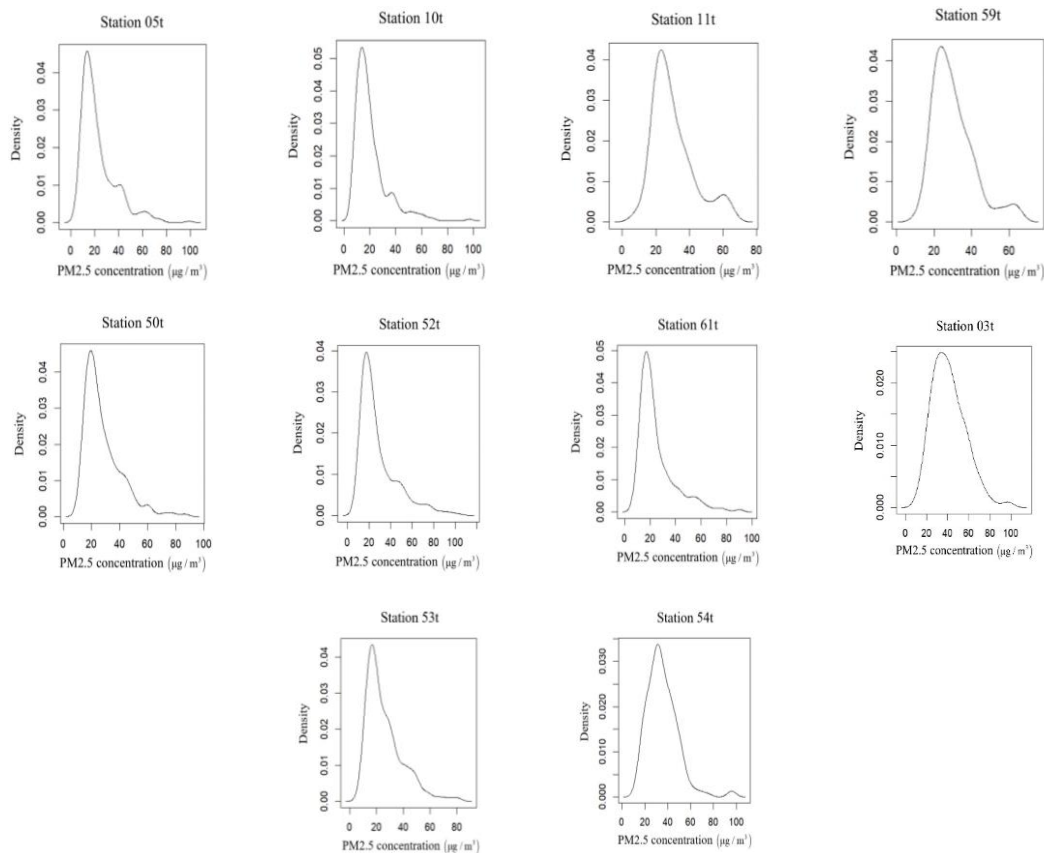


Figure 2. Empirical probability density function of the PM_{2.5} concentration data in 2018

Table 5. AIC and BIC values at station 05t

Distributions	2018		2019		2020		2021		2022	
	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
GM	2,688.09	2,680.40	2,775.67	2,767.88	2,766.59	2,758.79	2,804.86	2,797.06	2,665.66	2,657.87
LN	2,656.54	2,648.85	2,754.84	2,747.04	2,720.54	2,712.73	2,771.73	2,763.93	2,639.92	2,632.13
W	2,725.73	2,718.04	2,809.10	2,801.31	2,805.49	2,797.68	2,832.29	2,824.49	2,716.72	2,708.93
GM-GM	2,655.94	2,636.71	2,764.06	2,744.57	2,702.03	2,682.51	2,768.52	2,749.02	2,617.62	2,598.14
LN-LN	2,657.21	2,637.97	2,763.92	2,744.43	2,702.02	2,682.51	2,765.27	2,745.77	2,616.49	2,597.02
W-W	2,666.06	2,646.83	2,771.46	2,751.97	2,716.50	2,696.99	2,787.59	2,768.09	2,640.64	2,621.17
GM-LN	2,655.84	2,636.60	2,763.73	2,744.24	2,701.96	2,682.45	2,766.76	2,747.26	2,617.45	2,597.98
GM-W	2,660.65	2,641.42	2,764.33	2,744.85	2,707.82	2,688.30	2,778.78	2,759.28	2,632.45	2,612.98
LN-W	2,661.13	2,641.90	2,763.57	2,744.08	2,705.28	2,685.77	2,768.53	2,749.03	2,621.19	2,601.72
GM-GM-GM	2,671.23	2,640.46	2,780.62	2,749.44	2,718.67	2,687.45	2,782.60	2,751.40	2,633.24	2,602.09
LN-LN-LN	2,672.84	2,642.07	2,776.57	2,745.39	2,718.36	2,687.14	2,782.44	2,751.24	2,632.05	2,600.89
W-W-W	2,673.78	2,643.01	2,781.60	2,750.42	2,724.67	2,693.44	2,748.50	2,717.30	2,641.77	2,610.61
GM-GM-LN	2,668.60	2,637.82	2,776.51	2,745.33	2,718.83	2,687.61	2,782.07	2,750.87	2,634.03	2,602.88
GM-GM-W	2,673.01	2,642.24	2,781.02	2,749.84	2,720.48	2,689.25	2,782.66	2,751.46	2,633.80	2,602.65
GM-LN-LN	2,467.10	2,436.33	2,781.91	2,750.73	2,719.22	2,688.00	2,781.98	2,750.78	2,633.20	2,602.05
GM-LN-W	2,671.39	2,640.62	2,781.50	2,750.33	2,718.97	2,687.74	2,781.85	2,750.65	2,638.40	2,607.25
GM-W-W	2,673.20	2,642.43	2,777.02	2,745.84	2,723.12	2,691.90	2,784.67	2,753.47	2,635.78	2,604.62
LN-LN-W	2,671.39	2,640.62	2,781.86	2,750.68	2,721.96	2,690.74	2,777.14	2,749.84	2,632.27	2,601.12
LN-W-W	2,673.07	2,642.30	2,777.97	2,746.79	2,720.78	2,689.56	2,783.46	2,752.27	2,633.25	2,602.10

Table 6. Ranking score at station 05t

Distributions	2018		2019		2020		2021		2022		Average Ranking Score	
	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
GM	18	18	8	18	18	18	18	18	18	18	16.00	18.00
LN	4	17	1	10	13	17	6	16	15	17	7.80	15.40
W	19	19	19	19	19	19	19	19	19	19	19.00	19.00
GM-GM	3	3	5	4	3	3	4	4	3	3	3.60	3.40
LN-LN	2	2	3	2	1	1	3	3	2	2	2.80	2.60
W-W	6	9	6	5	5	10	8	15	7	15	10.80	16.40
GM-LN	5	5	4	3	2	2	2	2	1	1	2.20	2.00
GM-W	7	10	2	1	4	4	5	5	4	6	6.40	10.80
LN-W	8	16	7	17	6	16	17	17	16	16	4.40	5.20
GM-GM-GM	10	6	13	11	8	6	13	11	9	8	10.60	8.40
LN-LN-LN	9	4	9	6	9	7	11	9	12	11	9.40	7.40
W-W-W	14	12	14	12	12	11	14	12	11	10	13.60	11.80
GM-GM-LN	1	1	18	16	11	9	10	8	8	7	10.00	7.40
GM-GM-W	11	7	15	13	10	8	9	7	14	13	13.00	11.40
GM-LN-LN	16	14	11	8	16	14	16	14	13	12	9.60	8.20
GM-LN-W	13	11	10	7	7	5	12	10	5	4	11.80	9.60
GM-W-W	12	8	17	15	15	13	7	6	6	5	14.40	12.40
LN-LN-W	15	13	12	9	14	12	15	13	10	9	11.40	9.40
LN-W-W	17	15	16	14	17	15	1	1	17	14	13.20	11.20

on non-mixture, 2-mixture, and 3-mixture distributions, utilizing gamma, lognormal, and Weibull distributions as component distributions.

Tables 5 and 6 present the results for station 05t. For each year, AIC and BIC values are calculated and displayed in Table 5. The lowest values for AIC and BIC are highlighted in bold. After that, AIC and BIC values are ranked from lowest to highest. Table 6 shows the rank and the average ranking score of each distribution. The lowest values of the average ranking score are highlighted in bold. The GM-LN distribution is identified as the best fit for the dataset from station 05t, having the lowest average ranking score.

The parameters of the PM_{2.5} concentration data at station 05t from 2018 to 2022 are shown in Table 7. These are based on a two-mixture gamma-lognormal (GM-LN) distribution model. For the lognormal component, the estimated parameters

include weight, mean, and standard deviation, while the gamma component parameters consist of weight, shape, and scale.

Figure 3 illustrates the comparison between the histogram of PM_{2.5} concentration data from station 05t and the probability density function of the GM-LN distribution for each year from 2018 to 2022. In the figure, the red line represents the lognormal component, the green line represents the gamma component, and the black line represents the mixture distribution that is a combination of lognormal and gamma components.

The remaining stations are analyzed using the same procedure as station 05t, and the summary of the outcomes is presented in Table 8.

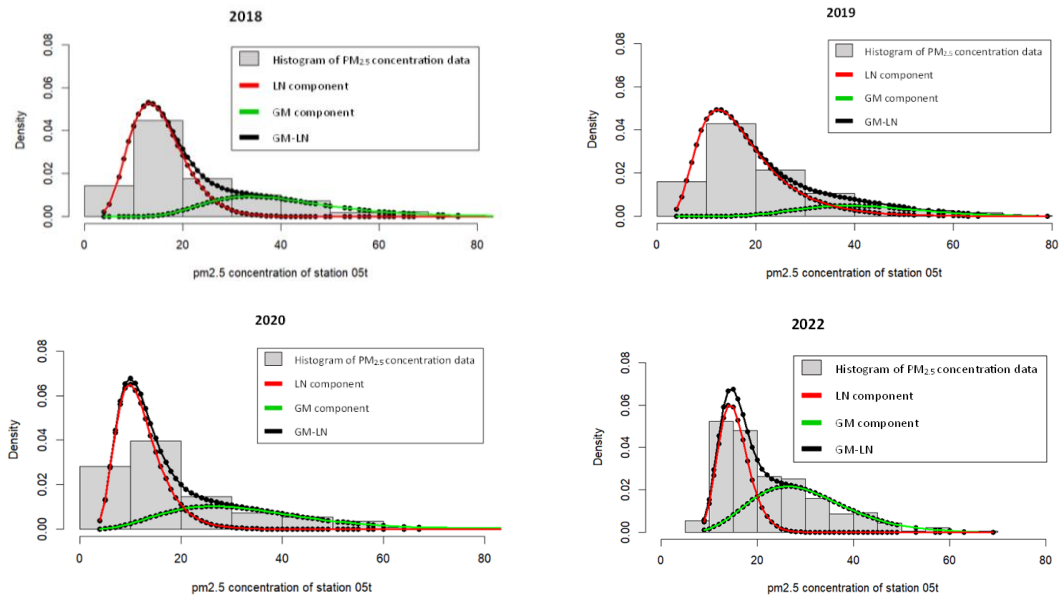
Table 8 provides a comparison of stations divided into two groups: roadside and residential areas. The analysis reveals that roadside stations generally align with the LN-LN distribution or

Table 7. Estimating parameter of PM_{2.5} data at station 05t

Parameters of 2-Mixture Gamma-Lognormal Distribution at 05t						
Year	Lognormal Distribution Component			Gamma Distribution Component		
	Weight	Mean	Standard Deviation	Weight	Shape Parameter	Scale Parameter
2018	0.293431	3.637887	0.346460	0.706570	7.312229	0.477158
2019	0.848704	2.757514	0.490280	0.151296	11.375843	0.268938
2020	0.650439	2.427644	0.376890	0.349561	5.001804	0.151818
2021	0.589536	3.174172	0.516572	0.410464	9.354529	1.029747
2022	0.465260	2.707983	0.209705	0.534740	8.275363	0.277318

Table 8. Statistical distributions selected based on AIC and BIC

Stations (residential areas)	Best distribution		Stations (along roadsides)	Best distribution	
	BIC	AIC		BIC	AIC
05t	GM-LN	GM-LN	03t	GM-LN	GM-LN-W
10t	LN-LN	GM-GM-LN	50t	LN-LN	LN-LN
11t	GM-LN	GM-LN	52t	LN-LN	LN-LN
59t	LN-LN	LN-LN-LN	53t	GM-LN	GM-LN
61t	LN-LN	GM-GM-GM	54t	LN-LN	LN-LN

**Figure 3. Histogram of PM_{2.5} data from station 05t and the probability density function of the GM-LN distribution**

distributions incorporating LN as a component, regardless of whether the AIC or BIC criterion is applied. In contrast, residential stations exhibit variability in their corresponding distributions when evaluated using the AIC criterion. However, under the BIC criterion, most residential stations also tend to align with the LN-LN distribution or distributions that include LN as a component. It is noteworthy that the BIC criterion often selects less complex models compared to the AIC criterion. The differences in the appropriate distributions for each station are likely influenced by environmental factors or the presence of nearby industries, which affect the distribution patterns.

Conclusions

This study focuses on the daily average concentrations of PM_{2.5} from January 1, 2018 to December 31, 2022, specifically examining data collected from ten monitoring stations in Bangkok. Statistical distribution functions, including non-mixture distribution and mixture distributions, are employed to characterize the statistical distribution patterns of PM_{2.5}, with the objective to identify the most representative distributions using the Akaike information criterion and the Bayesian information criterion. The analysis reveals that the daily average concentrations of PM_{2.5} at each station correspond

to different distributions. However, all stations align with distributions that include lognormal components.

Furthermore, government agencies or relevant authorities can implement policies to reduce PM_{2.5} levels by utilizing data from the distribution patterns detected at each monitoring station. This data helps us understand the specific components of PM_{2.5} in each area and allows for a clearer identification of their sources. Based on this information, authorities may consider measures to control the emission of major PM_{2.5} components in communities near monitoring stations, which could effectively reduce PM_{2.5} concentrations in the air.

Acknowledgement

I would like to express my deepest gratitude to Assoc. Prof. Dr. Eckart Schulz for his invaluable assistance in proofreading this work. I am also sincerely grateful to the School of Mathematical Science and Geoinformatics, Institute of Science, Suranaree University of Technology, Thailand, for their support in facilitating this research.

References

- Aquilanti, L., Cacace, S., Camilli, F., and Maio, R.D. (2021). A mean field games model for finite mixtures of Bernoulli and categorical distributions. *Journal of Dynamics and Games*, 8(1):35-59. <https://doi.org/10.3934/jdg.2020033>
- Balmes, J.R. and Eisner, M.D. (2016). Indoor and Outdoor Air Pollution. In: Murray and Nadel's Textbook of Respiratory Medicine. 6th ed., Elsevier, p. 1331-1342. <https://doi.org/10.1016/B978-1-4557-3383-5.00074-9>
- Bernardi, M. (2013). Risk measures for skew normal mixtures. *Statistics and Probability Letters*, 83(8):1819-1824. <https://doi.org/10.1016/j.spl.2013.04.016>
- Channoufi, I., Bourouis, S., Bouguila, N., and Hamrouni, K. (2018). A flexible statistical model for image denoising. In: 15th International Conference, ICIAR 2018, Póvoa de Varzim, Portugal, June 27-29, 2018, p. 30-38. https://doi.org/10.1007/978-3-319-93000-8_4
- Ding, J., Tarokh, V., and Yang, Y. (2018). Model selection techniques: an overview. In: *IEEE Signal Processing Magazine*, 35(6):16-34. <https://doi.org/10.1109/MSP.2018.2867638>
- Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*. 1st ed. Chapman & Hall, 143p. <https://doi.org/10.1007/978-94-009-5897-5>
- Hogg, R.V., McKean, J.W., and Craig, A.T. (2019). *Introduction to Mathematical Statistics*. 8th ed. Pearson Education, 746p.
- IQAir AirVisual. (2018). 2018 World Air Quality Report: Region & City PM2.5 Ranking. 22p. Available from: <https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2018-en.pdf>
- IQAir AirVisual. (2019). 2019 World Air Quality Report: Region & City PM2.5 Ranking. 33p. Available from: <https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2019-en.pdf>
- IQAir AirVisual. (2020). 2020 World Air Quality Report: Region & City PM2.5 Ranking. 41p. Available from: <https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2020-en.pdf>
- IQAir AirVisual. (2021). 2021 World Air Quality Report: Region & City PM2.5 Ranking. 43p. Available from: <https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2021-en.pdf>
- IQAir AirVisual. (2022). 2022 World Air Quality Report: Region & City PM2.5 Ranking. 47p. Available from: <https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2022-en.pdf>
- IQAir AirVisual. (2023). 2023 World Air Quality Report: Region & City PM2.5 Ranking. 45p. Available from: <https://www.iqair.com/world-most-polluted-cities/world-air-quality-report-2023-en.pdf>
- Korolev, V., Gorshenin, A., Gulev, S., and Belyaev, K. (2015). Statistical modeling of air-sea turbulent heat fluxes by finite mixtures of Gaussian distributions. In: Dudin, A., Nazarov, A., and Yakupov, R. (eds). *Information Technologies and Mathematical Modelling - Queueing Theory and Applications. ITMM 2015. Communications in Computer and Information Science*, Springer, 564:152-162. https://doi.org/10.1007/978-3-319-25861-4_13
- Krishnamoorthy, K. (2006). *Handbook of Statistical Distributions with Applications*. Taylor and Francis Group, NY, 344p. <https://doi.org/10.1201/9781420011371>
- Kvam, P.H. and Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering*. John Wiley & Sons, 420p. <https://doi.org/10.1002/9780470168707>
- Lu, H.C. (2004). Estimating the emission source reduction of PM10 in central Taiwan. *Chemosphere*, 54(7):805-814. <https://doi.org/10.1016/j.chemosphere.2003.10.012>
- Malá, I. (2016). On the residual duration of unemployment in the Czech Republic based on finite mixture of lognormal distributions. In *APLIMAT 2016 - 15th Conference on Applied Mathematics 2016*, p. 791-798.
- McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons, 419p. <https://doi.org/10.1002/0471721182>
- McLachlan, G.J., Lee, S.X., and Rathnayake, S.I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, 2019(6):355-378. <https://doi.org/10.1146/annurev-statistics-031017-100325>
- Mijić, Z., Tasić, M., Rajšić, S., and Novaković, V. (2009). The statistical characters of PM₁₀ in Belgrade area. *Atmospheric Research*, 92(4):420-426. <https://doi.org/10.1016/j.atmosres.2009.01.002>
- Mishra, G., Ghosh, K., Dwivedi, A.K., Kumar, M., Kumar, S., Chintalapati, S., and Tripathi, S.N. (2021). An application of probability density function for the analysis of PM_{2.5} concentration during the COVID-19 lockdown period. *Science of the Total Environment*, 782:146681. <https://doi.org/10.1016/j.scitotenv.2021.146681>
- Pollution Control Department. (2022). *Air and Noise Pollution Situation and Management in Thailand, 2022*. 66p. Available from: Ministry of Natural Resources and Environment. (In Thai)
- Punzo, A., Mazza, A., and Maruotti, A. (2018). Fitting insurance and economic data with outliers: A flexible approach based on finite mixtures of contaminated gamma distributions. *Journal of Applied Statistics*, 45(14):2563-2584. <https://doi.org/10.1080/02664763.2018.1428288>
- Shi, Y., Matsunaga, T., Yamaguchi, Y., Zhao, A., Li, Z., and Gu, X. (2018). Long-term trends and spatial patterns of PM_{2.5}-induced premature mortality in South and Southeast Asia from 1999 to 2014. *Science of the Total Environment*, 631-632:1504-1514. <https://doi.org/10.1016/j.scitotenv.2018.03.146>
- Tajuddin, R.R.M., Ismail, N., and Ibrahim, K. (2020). Several two-component mixture distributions for count data. *Communications in Statistics: Simulation and*

- Computation, 51(7):3760-3771. <https://doi.org/10.1080/03610918.2020.1722834>
- Taylor, J.A., Jakeman, A.J., and Simpson, R.W. (1986). Modeling distributions of air pollutant concentrations -I. Identification of statistical models. *Atmospheric Environment*, 20(9):1781-1789. [https://doi.org/10.1016/0004-6981\(86\)90127-7](https://doi.org/10.1016/0004-6981(86)90127-7)
- Wang, X., Chen, R.J., Chen, B.H., and Kan, H.D. (2013). Application of statistical distribution of PM10 concentration in air quality management in 5 representative cities of China. *Biomedical and Environmental Sciences*, 26(8):638-646. <https://doi.org/10.3967/0895-3988.2013.08.002>
- World Health Organization (WHO). (2021). WHO global air quality guidelines. Particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon dioxide. World Health Organization, 273p.