The Statistical Distributions of $PM_{2.5}$ in Rayong and Chonburi Provinces, Thailand

Vanida Pongsakchat[1,*] and Pattaraporn Kidpholjaroen[2]

[1,2] Department of Mathematic, Faculty of Science, Burapha University
Chonburi, Thailand

[*]Corresponding author's email: vanida [AT] buu.ac.th

ABSTRACT—*The fine particulate matter ($PM_{2.5}$) concentrations is one of the most important issues that are often discussed since it has a greater impact on human health. Statistical distribution modeling plays an important role in predicting $PM_{2.5}$ concentrations. This research aims to find the optimum statistical distribution model of $PM_{2.5}$ in Rayong Province and Chonburi Province. The daily average data from 2014 – 2019 for Rayong and from 2015 – 2019 for Chonburi were using. Five statistical distributions were compared. A proper statistical distribution that represents $PM_{2.5}$ concentrations has been chosen based on three criteria include Anderson-Darling statistic and RMSE. The results show that Pearson type VI distribution performs better compared to other distributions for $PM_{2.5}$ concentrations in Rayong. For Chonburi, the proper statistical distribution is Log normal distribution.*

Keywords— $PM_{2.5}$, Statistical distribution, Chonburi Province, Rayong Province

1. INTRODUCTION

The fine particulate matter ($PM_{2.5}$) refers to tiny particles or droplets in the air with diameter of less than 2.5 microns. $PM_{2.5}$ is an air pollutant that is a concern for people's health when levels in air are high. Exposure to $PM_{2.5}$ can affect lung function and worsen medical conditions such as asthma and heart disease.

The presence and continuous increasing levels of $PM_{2.5}$ has been a worsening issue in Thailand over the past few years. The main sources of $PM_{2.5}$ in Thailand are automobile exhaust, burning of biological material, secondary dust generated from the combination of automobile exhaust and burning of fossil fuels in factories and electrical generator plants [6].

In order to appropriately manage the problems of $PM_{2.5}$, it is important to evaluate the situation in individual areas since the amount of $PM_{2.5}$ differ wildly depending on geographic locations environment. The statistical distribution model of $PM_{2.5}$ concentrations is an important quantity used to describe and discuss air pollutant diffusion. When the statistical distribution of air pollutant is correctly chosen, it can be used to predict the mean concentration.

Many types of statistical distributions have been used to fit particulate matter concentrations including lognormal [1,2,4,5,7,9,10], gamma [4,5,7,9,10], Weibull [4,6,7,9,10], Pearson type V [4,5,10] and Pearson type VI [4] distributions. The lognormal distribution form is more wildly used to represent the type of air pollutant concentration distribution. However, the proper distribution may differ in some locations.

In this study, five statistical distribution models were fit to the daily $PM_{2.5}$ concentration data of Laem Chabang station, Chonburi province and Rayong Provincial Agriculture Office station, Rayong province and hence find the optimum distribution to represent the daily $PM_{2.5}$ concentrations in both stations.

2. METHODOLOGY

2.1 Study Areas and Data

The study considers daily $PM_{2.5}$ concentration data (in µg/m^3) from two case study stations in Chonburi and Rayong provinces which are in the east of Thailand. Laem Chabang station is located in Laem Chabang, Chonburi. The station Rayong Provincial Agriculture Office is located in Mueang Rayong District, Rayong. The studied data (from Jan 2014 to Dec 2019 for Leam Chabang station and from May 2015 to Dec 2019 for Rayong Provincial Agriculture Office station) provided by the Division of Air Quality Data, Air Quality and Noise Management Bureau, Pollution Control Department.

These stations are in the industrial cities. The air quality of both cities have been affected by industrial activities and traffic densities. Pollutants emitted from these sources into the atmosphere led to increasing level of air pollution index (API), as measured by $PM_{2.5}$ concentrations.

Table 1 shows summary statistics of the daily $PM_{2.5}$ concentration of the studied stations. On the average, daily $PM_{2.5}$ concentrations of Leam Chabang station (21.002 µg/m^3) is a little higher than Rayong Provincial Agriculture Office station (20.468 µg/m^3). The highest level of $PM_{2.5}$ concentrations for Laem Chabang station and Rayong Provincial Agriculture Office station were recorded as 95.833 µg/m^3 and 103.059 µg/m^3, respectively.

Table 1: Summary statistics of daily $PM_{2.5}$ concentration data

	Laem Chabang station	Rayong Provincial Agriculture Office station
Average	21.002	20.468
S.D.	12.887	13.887
Min.	3.778	3.750
Max.	95.833	103.059

Each dataset was separate into two datasets, the first dataset was used to obtain the optimum statistical distribution and the second dataset was used to validate the selected distribution. For Rayong Provincial Agriculture Office station, the first dataset was the data from Jan 2015 to Dec 2018 and the second dataset was the data from Jan 2019 to Dec 2019. While Leam Chabang station, the first dataset was the data from May 2015 to Dec 2018 and the second dataset was the data from Jan 2019 to Dec 2019.

2.2 Statistical Distributions

To determine the optimum statistical distribution for daily average $PM_{2.5}$ concentration of the studied stations, five statistical distributions are considered. These statistical distributions are lognormal, gamma, Weibull, Pearson type V and Pearson Type VI distributions. The maximum likelihood estimation (MLE) is used to estimate the parameter values of these distributions [4,6,8,9]. The probability density function of these distributions are as follow:

- Lognormal

$$f(x) = \frac{1}{x\sigma(2\pi)^{1/2}} \exp\left(\frac{-(\log x - \mu)^2}{2\sigma^2}\right)$$

where $x \geq 0$, σ is the shape parameter and μ is the scale parameter [3].

- Gamma

$$f(x) = \frac{x^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)} \exp(-x/\beta)$$

where Γ is the gamma function, $x \geq 0$, α is the shape parameter and β is the scale parameter [3].

- Weibull

$$f(x)=\left(\frac{\beta x^{\beta-1}}{\eta^{\beta}}\right)\exp\left[-(x/\eta)^{\beta}\right]$$

where $x \geq 0$, β is the shape parameter and η is the scale parameter [3].

- Pearson type V

$$f(x)=\frac{\exp\left(-\beta/(\mathrm{x}-\gamma)\right)}{\beta\Gamma(\alpha)\left((\mathrm{x}-\gamma)/\beta\right)^{\alpha+1}}$$

where $\gamma < x < +\infty$, α is the shape parameter, β is the scale parameter and γ is the location parameter[4].

- Pearson type VI

$$f(x)=\frac{\left[(x-\gamma)/\beta\right]^{\alpha_1-1}}{\beta\mathrm{B}(\alpha_1,\alpha_2)(1+(\mathrm{x}-\gamma)/\beta)^{\alpha_1+\alpha_2}}$$

where B is the beta function, $\gamma \leq x < +\infty$, α_1 and α_2 are the shape parameters, β is the scale parameter and γ is the location parameter [4].

2.3 Assessment of Goodness of Fit

To evaluate the goodness of fit of the statistical distributions to the daily PM2.5 concentration data, the Kolmogorov-Smirnov (KS) [4,7,8,9,10] and Anderson-Darling (AD) [4,8,10] tests were used.

- Kolmogorov-Smirnov test

The KS statistic is defined as the maximum difference between the sample cumulative distribution function ($S(x)$) and the examined statistical function ($F(x)$). Therefore, the KS statistic is given by:

$$D=\max\left|F(x)-S(x)\right|$$

- Anderson-Darling test

The AD statistic assesses whether the sample comes from the specified distribution. The formula for the AD statistic, A^2 is

$$A^2=-n-S \ ,$$

where $S=\sum_{i=1}^{n}\frac{2i-1}{n}\left[\ln\left(F(x_i)\right)+\ln\left(1-F(\mathrm{x}_{n+1-i})\right)\right]$ and $x_1,\ldots,x_n$ are the sample values sorted in order of magnitude.

In this study $p-$value is used in hypothesis testing. The null hypothesis (H_0: the data follow the specified distribution) is rejected at the chosen significance level α if $p-$value<α. While comparing the different distributions, the distribution with higher $p-$value is likely to better fit regardless of the level of significance.

2.4 Performance Indicator

The root mean square error (RMSE) was used to evaluate whether the specified distribution generate good predicted values of the validation dataset [1,2,4,7,8]. The RMSE is defined as

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(P_i-O_i\right)^2} \ ,$$

Where P_i is the predicted data point and O_i is observed data point. For good prediction, the RMSE value must approach zero. Therefore, a smaller RMSE value means that the specified distribution is more appropriate.

3. RESULTS AND DISCUSSION

Each selected statistical distributions was fitted to the first daily $PM_{2.5}$ concentrations dataset of the studied stations and the goodness of fit statistics were afterword calculated.

Table 1 shows the goodness of fit statistics and the $p-$values . Considering the $p-$values of KS and AD test, the lognormal, Pearson type V and Pearson type VI distributions fit the daily $PM_{2.5}$ concentration of Leam Chabang station at significant level $\alpha = 0.05$. For Rayong Provincial Agriculture Office station, the statistical distributions that fit daily $PM_{2.5}$ concentrations were the Pearson type V and Pearson type VI distributions.

Table 1: Goodness of fit tests for the studied distributions

Distribution	Leam Chabang station		Rayong Provincial Agriculture Office station	
	KS ($p-$value)	AD ($p-$value)	KS ($p-$value)	AD ($p-$value)
Lognormal	0.0345 (0.0985)	1.9586 (0.0968)	0.05411 (0.0001)	7.8925 (0.0001)
Gamma	0.0697 (0.0000)	11.3180 (0.0000)	0.0870 (0.0000)	23.8100 (0.0000)
Weibull	0.0860 (0.0000)	22.4390 (0.0000)	0.0911 (0.0000)	33.8800 (0.0000)
Pearson type V	0.0201 (0.6896)	0.5155 (0.7311)	0.0313 (0.0741)	2.8789 (0.0316)
Pearson type VI	0.0203 (0.6721)	0.5787 (0.6683)	0.0321 (0.0621)	2.8408 (0.0330)

Figure 1 and Figure 2 illustrate the histograms of the daily $PM_{2.5}$ concentrations of each stations. It can be noticed that the lognormal, Pearson type V and Pearson type VI distributions fit well with the data histogram of the Leam Chabang station. Likewise, for the Rayong Provincial Agriculture Office station, the distributions that fit the data histogram were the Pearson type V and Pearson type VI distributions. The graphical results corresponded to the goodness of fit results.

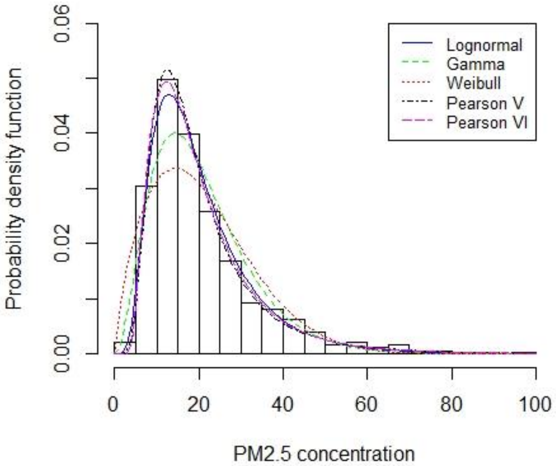

Figure 1: Comparison of probability density functions and $PM_{2.5}$ histogram of Leam Chabang station.

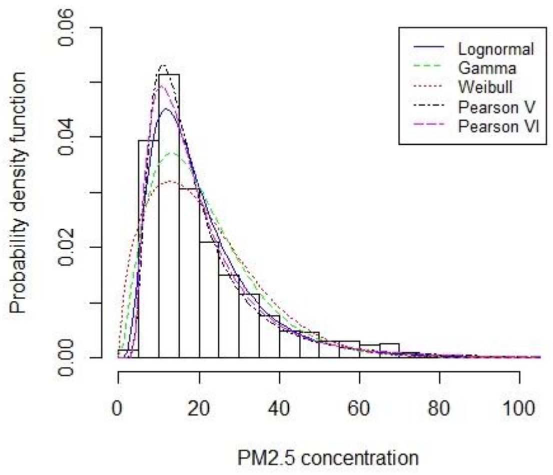

Figure 2: Comparison of probability density function and $PM_{2.5}$ histogram of Rayong Provincial Agriculture Office station.

Table 2 illustrates the root mean square error of the fitted distributions which calculated using the second dataset of each station. From the table the distribution that represent the best one to predict the daily $PM_{2.5}$ concentrations of the Leam Chabang station was the lognormal distribution since it gave the lowest RMSE. For the Rayong Provincial Agriculture Office station, the best distribution was the Pearson type VI distribution.

Table 2: Root mean square error of the fitted distributions.

Distribution	Leam Chabang station	Rayong Provincial Agriculture Office station
Lognormal	1.9244	-
Pearson type V	3.1019	9.6379
Pearson type VI	2.3485	7.3842

4. CONCLUSION

Five statistical distributions of $PM_{2.5}$ concentrations were investigated in two stations which locate in the cities in the east of Thailand. The distributions were fitted to the actual data and the parameters of each distribution were estimated by the maximum likelihood estimation. The fit distributions were determined using the two goodness of fit tests and the best fit distribution was selected using the root mean square error.

The results show that the best daily $PM_{2.5}$ concentration distribution in Leam Chabang station was the lognormal distribution which was most frequently used to represent air pollutant concentrations [1,2,4,7,9,10] and is consistent with the results of the number of previous researches [1,4,7]. For the Rayong Provincial Agriculture Office station, the best distribution was the Pearson type VI distribution which also consistent with the results of some previous researches [4]. The results of this study provide useful information on air quality status in the east of Thailand and can be used for air quality management.

5. REFERENCES

[1] Abdul A. H., Yahaya, A. S., Ramli, N. A. and Ul-Saufie, A. Z., "Finding the Best Statistical Distribution Model in PM_{10} Concentration Modeling by using Lognormal Distribution", Journal of Applied Sciences, vol. 13, no. 12, pp.294-300, 2013.

[2] El-Shanshoury, Gh. I., "Fitting the Probability Distribution Functions to Model Particulate Matter Concentrations", Arab Journal of Nuclear Sciences and Applications, vol.50, no. 2, pp. 108-122, 2017.

[3] Forbes, C., Evans, M., Hastings, N. and Peacock, B., Statistical Distributions, 4th ed., Wiley, New York, 2011.

[4] 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", GlobalNEST International Journal. 8. 241-251. 2006.

[5] Kan, H. D. and Chen, B. H., "Statistical Distributions of Ambient Air Pollutants in Shanghai, China", Biomedical and Environmental Sciences, vol. 17, pp. 366-372, 2004.

[6] Limpaseni, W. "The Harmful Effects of $PM_{2.5}$", Thailand Engineering Journal, vol. 71 no. 1, pp. 9-17, 2018. (In Thai)

[7] Lu, H. C., Fang, G. C. and Wu Y. S., "Estimating the Frequency Distributions of Particulate Matter and Their Metal Elements in a Temple", Journal of the Air & Waste Management Association, vol. 56, no. 7, pp. 1033-1040, 2006.

[8] Pobocikova, I., Sedliackova, Z. and Michalkova, M., "Application of Four Probability Distributions for Wind Speed Modeling", Procedia Engineering, vol. 192, pp. 713-718, 2017.

[9] Rossita M. Y. and Masud M. H., "Predicting Hourly PM_{10} Concentration in Seberang Perai and Petaling Jaya Using Log-Normal Linear Model", International Journal of Management and Applied Science (IJMAS), vol. 3, no. 3, pp. 103-108, 2017.

[10] Wang, X., Chen, R. J., Chen, B. H. and Kan, H. D., "Application of Statistical Distribution of PM_{10} Concentration in Air Quality Management in 5 Representative Cities of China", vol. 26, no. 8, pp. 638-646, 2013.