iJET EXPLORE International Journal of Engineering & Technology
CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS
ISSN 2049-3444

International Journal of Engineering and Technology Volume 2 No. 4, April, 2012

# Performance of Parameter Estimator for the Two-Parameter and Three-Parameter Gamma Distribution in $PM_{10}$ Data Modelling

**Hazrul Abdul Hamid[a,b] , Ahmad Shukri Yahaya[b], Nor Azam Ramli[b] , Ahmad Zia Ul-Saufie [b,c]**

[a]Penang Matriculation College, Kepala Batas, Pulau Pinang, Malaysia
[b]School of Civil Engineering, Universiti Sains Malaysia, Pulau Pinang, Malaysia
[c]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, Malaysia

## ABSTRACT

Data of $PM_{10}$ in selected cities in Malaysia were used to study the performance of parameter estimator of the two-parameter and three-parameter gamma distribution in modeling and predicting $PM_{10}$ data. This study have considered the method of moment and method of simple explicit to estimate the parameters for the two-parameter and three-parameter gamma distribution respectively. Five performance indicators are used to determine the most appropriate probability distribution for this set of data. Results show that the three-parameter gamma distribution give a better prediction compared to the two-parameter gamma distribution for Nilai, Negeri Sembilan site while the two-parameter gamma distribution predicts the $PM_{10}$ better compared to the three-parameter gamma distribution for Seberang Perai, Penang site.

**Keywords:** *Particulate Matter ($PM_{10}$), Air Pollution Modelling, Gamma Distribution*

## 1. INTRODUCTION

The impact of air pollution is noticeable, especially for human being where it can cause several significant effects including carcinogenic effects. Such health problems include cardiac arrhythmias, reducing lung function, asthma, chronic bronchitis and increasing respiratory symptoms, such as sinusitis, sore throat, dry and wet cough, and hay fever [15]. There are possible short-term and long term health effects of exposure to air pollution. In a short-term, high level of air pollution leads to an acute condition. [16] has detailed the air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulphur dioxide. Table 1 below shows the World Health Organization guidelines for particulate matter while table 2 shows the Malaysia ambient air quality guideline.

**Table 1: World Health Organization guidelines for particulate matter**

| | |
|---|---|
| $PM_{2.5}$ | 10 µg/m$^3$ annual mean |
| | 25 µg/m$^3$ 24-hour mean |
| $PM_{10}$ | 20 µg/m$^3$ annual mean |
| | 50 µg/m$^3$ 24-hour mean |

**Table 2: Malaysia ambient air quality guideline for $PM_{10}$**

| | |
|---|---|
| $PM_{10}$ | 50 µg/m$^3$ annual mean |
| | 150 µg/m$^3$ 24-hour mean |

Distribution describes the frequency or probability of possible events. There are many distributions that can be use to fit the air pollution data. [4] present methodologies and limitations in describing air quality through statistical distributions of pollutant concentration. The useful probability density functions in representing atmospheric concentrations include two-parameter distributions, three-parameter distributions and four-parameter distributions. Probability density function has been applied successfully in many physical phenomena such as wind speed, rainfall, river discharge, and air quality [5]. The probability density function of concentration in an atmospheric plume is an important quantity used to describe and discuss environmental diffusion [17]. The important step in fitting data to probability distribution is parameter estimation and the understanding of the behaviour of air pollution statistically would allow predictions to be made accurately.

The best distribution to represent the performance of high pollutants concentration is gamma distribution [10]. Gamma distribution is also commonly used in reliability. This distribution is an important distribution for nongaussian statistical modelling. Gamma distribution is positively skewed and it is often used as a model of life spans testing and many other related fields [13]. Gamma distribution is commonly used in environmental engineering to represent air pollutant concentration data especially for high density air pollutant data. Gamma distribution is also applied widely in several fields. Numerous authors have suggested that the gamma distribution is an appropriate distribution for rainfall amounts [8]. [14] also show that the gamma distribution is very useful and this distribution provide a more convenient description of the grain size distribution. Even the research also show that the gamma distribution and lognormal distribution are not much different in presenting grain size, but the log plots of the gamma function show the experimentally observed asymmetry in the grain size distributions. Another success application of gamma distribution is to represent rainfall amount. [7] use the gamma distribution to represent monthly rainfall in Africa. The gamma distribution is well suited to these applications not only because it is fairly familiar to African scientists, but because of it capability of representing a variety of distribution shapes. Not limited to environmental application, [12] used the gamma distribution in application of sleep apnea, which is sleep disorder breathing.

Several methods can be used to determine the parameters of gamma distribution. The most commonly method is method of moment and method of maximum likelihood. These methods can be used to estimate the parameters of the two-parameter gamma distribution and also the three-parameter gamma distribution. However, this research will focus on parameter estimation by using method of moment for two-parameter distribution and method of simple explicit which also use the concept of moment estimator for three-parameter distribution.

## 2. METHODOLOGY

### 2.1 Area of Study

Two areas involved in this study are Nilai in Negeri Sembilan and Seberang Perai in Pulau Pinang. Negeri Sembilan state is situated on the western coast of peninsular Malaysia. The total area of this state is 6686 $km^2$, latitude 2° 45′ 0″ N and longitude 102° 15′ 0″ E.

Pulau Pinang state is situated on the northwest coast of peninsula Malaysia. The area of this state is about 1048 $km^2$, latitudes $5^o$ 24'N and longitude $100^o$ 14'E. Pulau Pinang State consists of two parts, Pulau Pinang Island and mainland Seberang Perai. The island has an area of 285 $km^2$ and is connected to Seberang Perai by ferry and by the 13.5 km long Pulau Pinang Bridge.

Nilai in Negeri Sembilan and Seberang Perai in Pulau Pinang are rapidly growing town especially in industries along with economic sector. Since both of this area can be categorised as industrial area, air pollution can not be avoided. This was proven when several unhealthy days were recorded [2] and that is the reason why these two areas were selected in this study.

### 2.2 Air Quality Data

Annual hourly observations for $PM_{10}$ in Nilai, Negeri Sembilan and Seberang Prai, Pulau Pinang from January 2003 to Disember 2005 were selected for predicting $PM_{10}$ concentration level. Table 1 and Table 2 show the characteristic for $PM_{10}$ data for the monitoring site in Nilai Negeri Sembilan and Seberang Perai, Pulau Pinang respectively.

**Table 1: Characteristic for the $PM_{10}$ data in $\mu g/m^3$ for the monitoring site in Nilai, Negeri Sembilan**

| Year | 2003 | 2004 | 2005 |
|---|---|---|---|
| Mean | 54.88 | 59.75 | 63.25 |
| Standard deviation | 27.01 | 29.99 | 44.21 |
| Median | 50 | 54 | 54 |
| Mode | 33 | 36 | 40 |
| Minimum | 5 | 5 | 6 |
| Maximum | 270 | 406 | 542 |
| Coefficient of Skewness | 1.09 | 1.34 | 3.59 |
| Coefficient of Kurtosis | 2.25 | 4.95 | 21.72 |

**Table 2: Characteristic for the $PM_{10}$ data in $\mu g/m^3$ for the monitoring site in Seberang Perai, Pulau Pinang**

| Year | 2003 | 2004 | 2005 |
|---|---|---|---|
| Mean | 80.13 | 92.31 | 78.39 |
| Standard deviation | 35.62 | 38.02 | 49.08 |
| Median | 76 | 81 | 65 |
| Mode | 71 | 76 | 61 |
| Minimum | 5 | 5 | 5 |
| Maximum | 362 | 421 | 425 |
| Coefficient of Skewness | 0.713 | 0.734 | 1.136 |
| Coefficient of Kurtosis | 1.312 | 1.719 | 1.621 |

### 2.3 Gamma Distribution

For all types of two parameter parent distributions, λ is the shape parameter that determines the form of the distributions and σ is the scale parameter that determines the skewers of the distribution. For the three-parameter distribution, λ represent the shape parameter, σ represent the scale parameter and δ represent location parameter. The important step in fitting data to distribution is parameter estimating. To estimate the parameters, several methods can be used such as method of maximum likelihood, probability plot, method of moment, method of least squares, and method of percentiles. Probability density function for the two-parameter gamma distribution is as follow [3]:

$$f(x)=\left(\frac{1}{\sigma\Gamma(\lambda)}\right)\left(\frac{x}{\sigma}\right)^{\lambda-1} exp\left(-\frac{x}{\sigma}\right) \quad (1)$$

where $x \geq 0$. λ represents a shape parameter and σ represents a scale parameter for annual measurement of particular sites.

Probability density function for the three-parameter gamma distribution are as follow [4] :

$$f(x)=\frac{1}{\sigma\Gamma(\lambda)}\left(\frac{x-\delta}{\sigma}\right)^{\lambda-1} exp\left(-\frac{x-\delta}{\sigma}\right) \quad (2)$$

where $x \geq \delta$ ; $\sigma > 0, \lambda > 0$. λ represents a shape parameter, σ represents a scale parameter and δ represents a location parameter.

### 2.4 Parameter Estimator

#### 2.4.1 Method of Moment

There are several methods can be used to determine the parameter of the gamma distribution but the commonly used is method of moment and method of maximum likelihood. However, this study only focus on method of moment as this method is most commonly used. Method of moment also gives better fits compared to maximum likelihood estimator [11]. There is no restriction to estimate the parameter for the two-parameter gamma distribution by using this method. But, for the three-parameter distribution, the third central moment must be positive for valid solutions. This is the only restriction on this approach and that is the reason why method of moment for the three-parameter gamma distribution is not considered in this research. Parameter estimation was done by using method of moment for the two-parameter distribution and method of simple explicit for the-three parameter distribution.

For method of moments of the two-parameter distribution (MoM), estimation of the parameters can be obtained using the mean and standard deviation. The equations to obtain σ and λ for this method is as follows:

$$\lambda = \frac{1}{(\text{coefficient of variation})^2} \quad (3)$$

Where coefficient of variation $= \frac{s}{x}$, $s$ is a sample standard deviation and $\bar{x}$ is a sample mean. σ is the solution of :

$$\sigma\lambda = \bar{x} \quad (4)$$

Hence, the values of σ and λ can also be calculated by the following equations:

$$\sigma = \frac{s^2}{\bar{x}} \quad (5)$$

$$\lambda = \left(\frac{\bar{x}}{s}\right)^2 \quad (6)$$

[6] proposed a new modified moment estimation of parameter of the gamma distribution (MME) and shown the proposed estimator is more efficient compared to maximum likelihood estimator. The equations to obtain σ and λ for this method are as follows:

$$\lambda = \frac{\bar{x}^2}{s^2} - \frac{1}{n} \quad (7)$$

$$\sigma = \frac{\bar{x}}{\lambda} \quad (8)$$

### B. Simple Explicit Estimator

For the three-parameter gamma distribution, the following equations were used to estimate the parameters [1]:

$$\delta = 2x_{(1)} - 0.5\left(x_{(2)} + x_{(3)}\right) \quad (9)$$

$$\lambda = \left| \frac{-0.4 m_1(\delta) m_{0.4}(\delta)}{m_1(\delta) m_{0.4}(\delta) - m_{1.4}(\delta)} \right| \quad (10)$$

$$\mu = \frac{\bar{x} - \delta}{\lambda} \quad (11)$$

This method is known as Simple Explicit Estimator 1(SE-1). [1] not only proposed the above estimator. There are another estimator was proposed in their paper known as Simple Explicit Estimator 2 (SE-2). Equation below shows the second set of estimator.

$$\delta = 1.23x_1 - 0.23x_6 - 0.22(y_5 - y_2) - 0.21(y_4 - y_3) \quad (12)$$

$$\lambda = \left| \frac{-0.5 m_1(z_1) m_{0.5}(z_1) + 0.1 m_{0.1}(z_2) m_{1.6}(z_2) - 0.6 m_{1.1}(z_2) m_{0.6}(z_2)}{m_1(z_1) m_{0.5}(z_1) - m_{1.5}(z_1) + m_{1.1}(z_2) m_{0.6}(z_2) - m_{0.1}(z_2) m_{1.6}(z_2)} \right| \quad (13)$$

$$\mu = \frac{\bar{x} - \delta}{\lambda} \quad (14)$$

where

$$m_r = \frac{1}{n} \sum_{i=1}^{n} \left( x_i - \bar{x} \right)^r \quad (15)$$

## 2.5 Performance Indicators

Five performance indicators are used to determine the best estimator. Two of the performance indicator are error measures and another three are accuracy measures. Below are the formula and explanation of root mean square error, normalized absolute error, coefficient of determination, index of agreement and prediction accuracy [9]. The equation to obtained the normalized absolute error (NAE) is:

$$NAE = \frac{\sum_{i=1}^{n} Abs(P_i - O_i)}{\sum_{i=1}^{n} O_i} \quad (16)$$

*NAE* value closer to zero indicates better distribution.

The root mean square error (RMSE) is given by:

$$RMSE = \sqrt{\left( \frac{1}{N-1} \right) \sum_{i=1}^{N} (P_i - O_i)^2} \quad (17)$$

For a good model, the RMSE should approach zero. Therefore, a smaller RMSE means the model is more appropriate.

The index of agreement (IA) is given by:

$$IA = 1 - \left[ \frac{\sum_{i=1}^{N} (P - O_i)^2}{\sum_{i=1}^{N} \left( |P_i - \bar{O}| + |O_i - \bar{O}| \right)^2} \right] \quad (18)$$

where $0 \le IA \le 1$. When IA is closer to 1, then the model is appropriate to simulate the experimental data.

The coefficient of determination ($R^2$) is given by:

$$R^2 = \left( \frac{\sum_{i=1}^{N} (P_i - \bar{P})(O_i - \bar{O})}{N . S_{pred} . S_{obs}} \right)^2 \quad (19)$$

The prediction accuracy (PA) is given by:

$$PA = \frac{\sum_{i=1}^{N} (P_i - \bar{O})^2}{\sum_{i=1}^{N} (O_i - \bar{O})^2} \quad (20)$$

where $0 \le PA \le 1$. When PA is closer to 1, then the model is appropriate to simulate the experimental data.
where;

$N$ = Total number of annual measurements of a particular station

$P_i$ = Predicted values of one set annual monitoring record

$O_i$ = Observed values of one set annual monitoring record

$\bar{P}$ = Mean of the predicted values of one set annual monitoring record

$\bar{O}$ = Mean of the observed values of one set annual monitoring record

$S_{pred}$ = Standard deviation of the predicted values of one set annual monitoring record

$S_{obs}$ = Standard deviation of the observed values of one set annual

## 3. RESULTS AND DISCUSSIONS

This section will discuss the result of parameter estimation in predicting $PM_{10}$ data. Table 3 gives the results of the performance indicator (PI) for the Nilai, Negeri Sembilan site. Simple explicit estimator 1 (SE-1) gives the better estimation for 2003 and 2004. However, method of moment (MoM) was the best estimator for 2005. Table 4 gives the result of performance indicator for Seberang Perai, Pulau Pinang site. Method of moment 1 (MoM) was the best estimator for 2003 while for 2004 and 2005, modified moment estimator (MME) gives the better estimation. The distribution of data may cause the different of best estimator in the same site but in different year.

Table 5 summarized the final parameter for gamma distribution by using the best method of parameter estimation. For Nilai, Negeri Sembilan, three-parameter gamma distribution was selected to represent the observed data for 2003 and 2004 while two-parameter gamma distribution was selected to represent the data for 2005. Meanwhile for Seberang Perai, Pulau Pinang, two-parameter gamma distribution was selected to represent the data for 2003, 2004 and 2005.

**Table 3: Performance Indicator for Nilai, Negeri Sembilan site**

| Year | PI | 2-parameter (MoM) | 2-parameter (MME) | 3-parameter (SE-1) | 3-parameter (SE-2) | Best estimator |
|---|---|---|---|---|---|---|
| 2003 | NAE | 0.007973 | 0.007974 | 0.008338 | 0.007820 | Simple Explicit 1 |
| | PA | 0.997904 | 0.99704 | 0.998200 | 0.998187 | |
| | $R^2$ | 0.995584 | 0.995584 | 0.996176 | 0.996150 | |
| | RMSE | 1.776750 | 1.776626 | 1.637176 | 1.649795 | |
| | IA | 0.998902 | 0.998902 | 0.999071 | 0.999055 | |
| 2004 | NAE | 0.013978 | 0.013981 | 0.011785 | 0.011589 | Simple Explicit 1 |
| | PA | 0.993676 | 0.993676 | 0.994184 | 0.994150 | |
| | $R^2$ | 0.987166 | 0.987167 | 0.988178 | 0.988110 | |
| | RMSE | 3.376513 | 3.376458 | 3.242206 | 3.254711 | |
| | IA | 0.996778 | 0.996778 | 0.997028 | 0.997001 | |
| 2005 | NAE | 0.137535 | 0.137539 | 0.098351 | 0.098279 | Modified Moment Estimator |
| | PA | 0.940841 | 0.940844 | 0.940007 | 0.940062 | |
| | $R^2$ | 0.884980 | 0.884986 | 0.883412 | 0.883515 | |
| | RMSE | 15.098261 | 15.097983 | 15.279159 | 15.272762 | |
| | IA | 0.969443 | 0.969444 | 0.965348 | 0.965379 | |

**Table 4: Performance Indicator for Seberang Perai, Pulau Pinang site**

| Year | PI | 2-parameter (MoM) | 2-parameter (MME) | 3-parameter (SE-1) | 3-parameter (SE-2) | Best estimator |
|---|---|---|---|---|---|---|
| 2003 | NAE | 0.019933 | 0.019935 | 0.029491 | 0.029491 | Method of Moment |
| | PA | 0.996977 | 0.996976 | 0.995899 | 0.995899 | |
| | $R^2$ | 0.993756 | 0.993735 | 0.991589 | 0.991589 | |
| | RMSE | 2.791761 | 2.791762 | 3.342838 | 3.342838 | |
| | IA | 0.998442 | 0.998442 | 0.997839 | 0.997839 | |
| 2004 | NAE | 0.019811 | 0.019811 | 0.028781 | 0.028781 | Modified Moment Estimator |
| | PA | 0.995313 | 0.995313 | 0.994635 | 0.994635 | |
| | $R^2$ | 0.990423 | 0.990423 | 0.989074 | 0.989074 | |
| | RMSE | 3.695583 | 3.695564 | 4.099064 | 4.099064 | |
| | IA | 0.997610 | 0.997610 | 0.997164 | 0.997164 | |
| 2005 | NAE | 0.033924 | 0.033922 | 0.031857 | 0.031857 | Modified Moment Estimator |
| | PA | 0.997393 | 0.997393 | 0.997180 | 0.997180 | |
| | $R^2$ | 0.994566 | 0.994566 | 0.994140 | 0.994140 | |
| | RMSE | 3.603066 | 3.602807 | 3.714303 | 3.714303 | |
| | IA | 0.998630 | 0.998630 | 0.998575 | 0.998575 | |

**Table 5: Parameter for the gamma distribution using the best method**

| | Nilai, Negeri Sembilan | | | Seberang Perai, Pulau Pinang | | |
|---|---|---|---|---|---|---|
| Parameter | 2003 | 2004 | 2005 | 2003 | 2004 | 2005 |
| Scale | 14.7475 | 16.4249 | 30.9026 | 15.8351 | 15.7078 | 30.7271 |
| Shape | 3.3828 | 3.3336 | 2.0468 | 5.0605 | 5.8767 | 2.5515 |
| Location | 5.0000 | 5.0000 | - | - | - | - |
| Best Method | Simple Explicit 1 | Simple Explicit 1 | Method of Moment | Method of Moment | Modified Moment Estimator | Modified Moment Estimator |
| Distribution | three-parameter gamma distribution | three-parameter gamma distribution | two-parameter gamma distribution | two-parameter gamma distribution | two-parameter gamma distribution | two-parameter gamma distribution |

## 4. CONCLUSION

Parameter estimation is one of the important steps in fitting distribution allow predictions to be made accurately. The aim of this study was to compare several parameter estimators and to find the most suitable estimator and distribution for predicting $PM_{10}$ concentration in Nilai, Negeri Sembilan and Seberang Perai, Pulau Pinang. The quality and reliability of the developed models were evaluated via five performance indicators (NAE, RMSE, PA, IA and $R^2$) and the result of this study show that simple explicit 1 (SE-1) is the most suitable method to estimate the parameter for Nilai site in 2003 and 2004 by using the three-parameter gamma distribution. However, method of moment 1 (MoM) was best fitted to Nilai site for 2005 and Seberang Jaya site for 2003. Modified moment estimator (MME) is the best method to estimate the parameters to Seberang Jaya site for 2004 and 2005.

## REFERENCES


[1] Balakrishnan, N. and Wang, J. (2000), Simple efficient estimation for the three-parameter gamma distribution. *Journal of Statistical Planning and Inference*, 85, 115 – 126.

[2] Department of Environment, Malaysia (2007). Malaysia Environmental Quality Report 2007. Kuala Lumpur: *Department of Environment, Ministry of Sciences, Technology and the Environment*, Malaysia

[3] Evans, M., Hastings, N. and Peacock., B. (2000). Statistical Distributions. New York: John Willey.

[4] Georgopoulus, P.G. and Seinfeld, J.H. (1982) Statistical Distributions of Air Quality Concentrations. Environmental Science and Technology, 16, p. 401A – 416A

[5] Harikrishna, M. and Arun C., (2003). Stocastic Analysis for Vehicular Emissions on Urban Roads – A case study of Chennai. Proceedings of the Third International Conference on Environmental and Health, 15-17 December 2003, Chennai, India.

[6] Hwang, T.Y. and Huang, P.H. (2002). On New Moment Estimation of Parameters of the Gamma distribution Using its Characterization. Annals of The Institute of Statistical Mathematics, 54, 840 – 847.

[7] Husak, G.J., Joel, M., and Chris, F. (2007). Use of gamma distribution to represent monthly rainfall in Africa for drought monitoring applications. International Journal of Climatology, 27, 935 – 944.

[8] Irish, J. L. (1994). It look like a gamma distribution to me : the distribution of daily rainfall amounts. Proceedings of the Water Down Under 1994 Conference. 21-25 November 1994. National Conference Publication – Institution of Engineers, Australia.

[9] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38, 2895-2907.

[10] Lu, H.C. (2003) Comparison of statistical characteristic of air pollutants in Taiwan by frequency distribution. Journal of the Air & Waste Management Association, 53(5), 608 – 616.


 

[11] Md Yusof, N.F. (2009). Modelling and predicting of PM10 concentration during high particulate events in Malaysia. PhD thesis. Universiti Sains Malaysia, Malaysia.

[12] Nur Zakiah, M.S., Abdul Aziz J., Syed Hassan, A.A. (2008). A comparison of Weibull and gamma distribution in application of sleep apnea. Asian Jurnal of Mathematics & Statistics, 1, 132 – 138.

[13] Pang, W.K., Hou, S.H., Yu, B.W.T., Li, W.K. (2004). A simulation based approach to the parameter estimation for the three-parameter gamma distribution. European Journal of Operational Research, 155, 675 – 682.

[14] Vaz, M.F. and Fortes, M.A. (1988). Grain size distribution : The lognormal and the gamma distribution functions. Scripta metallurgica, 22, 35 – 40.

[15] World Health Organization (1998) Report of the Bioregional Workshop on Health Impacts of Haze Related Air Pollution. Manila : WHO.

[16] World Health Organization (2006) Air Quality Guidelines : Global Update 2005 Particulate matter, ozone, nitrogen dioxide and sulphur dioxide. German : WHO

[17] Yee, E. and Chan, R. (1997) A Simple Model for the Probability Density Function of Concentration Fluctuations in Atmospheric Plumes. Atmospheric Environment, 31(7), 991 – 1002.