Modeling the distribution of duration time for unhealthy air pollution events

N Masseran¹, M A M Safari² and S I Hussain³

¹,³Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
²Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Email: ¹kamalmsn@ukm.edu.my, ²aslam.safari@upm.edu.my

Abstract. The information about how long a severe unhealthy air pollution event will last is crucial for the purpose of planning a possible measure to mitigate its risk. Thus, analyzing the distribution of duration data on the past occurrences of air pollution events is important. This study analyzes the hourly data of air pollution index (API) in Klang City, Malaysia from 1997 to 2018. Air pollution duration data are determined from the period when API > 100, preceded and followed by periods when API < 100. In this study, four types of statistical distributions, namely, Exponential, Gamma, Lognormal, and Weibull are proposed as practical models. Goodness-of-fit measures are compared for each distribution to determine the best fitted one to describe the observed data. Results indicate that the Lognormal distribution provides the best fitted model among others.

1. Introduction

Air pollution is an important issue that must be addressed around the world, particularly in urban areas. To deal with this issue, elaborating the behaviors of air pollution events on the basis of observed data is informative. In parallel with that, a statistical analysis is an important tool to extract various kinds of information from the observed air pollution data. One approach is by describing its probability distribution. Statistical distribution can provide knowledge on the probability behavior of the magnitude of an air pollution event. For example, an extreme air pollution event can be modeled and described using the generalized extreme value distribution or generalized Pareto distribution [1, 2, 3, 4]. Al-Dhurafi et al. [5] showed that the distribution of general air pollution index data can be modeled using several unimodal distributions, such as Gamma, Lognormal, and Weibull. The distribution of air pollution index data, which correspond to their compositional sub-indices, can be modeled and described using various forms of mixture distribution [6]. Leiva et al. [7] suggested that the skewed Sinh–normal distribution can be used to model air pollution data. Bartoletti and Loperfido [8] found that a skew–normal distribution can provide good fit to the data PM₁₀ pollutant, particularly in estimating the probabilities of its high concentration values. Kan and Chen [9] found that Lognormal, Pearson-V, and extreme value distributions are suitable to represent the distributions of the daily data of pollutant variables, such as SO₂, NO₂, and PM₁₀ in Shanghai, China. Nadarajah [10] proposed a truncated inverted beta distribution as an alternative model for modeling air pollution data, particularly the ozone level pollutant variable.
The information about probability distribution is important, as it represents a building block for developing advance and complex statistical techniques for analyzing air pollution behaviors. For example, Sak et al. [11] proposed the copula model for air pollution risk assessment based on a generalized hyperbolic distribution, which represents a marginal model for the data of PM$_{2.5}$ in several cities in China. Masseran [12] used standard normal, standardized Student-t, and generalized residual distribution to represent the residuals in a mixing model of autoregressive integrated moving average and generalized autoregressive conditional heteroscedastic to forecast the PM$_{10}$ pollutant data in Kuala Lumpur, Malaysia. AL-Dhurafi et al. [13] employed the generalized Pareto distribution to represent unhealthy air pollution data as input for developing a hierarchical model, which combines air pollution data from several locations in Selangor, Malaysia. Masseran and Hussain [14] used the generalized Pareto model to represent the tail distribution of air pollution data as an input for dynamic copula modeling among five pollutant variables. Zhang et al. [15] employed a Gamma distribution as a building block for the Hidden Markov model for the purpose of an ozone level prediction. Masseran and Mohd Safari [16] used the generalized extreme value distribution as a building block for developing intensity–duration–frequency curves to evaluate the risks of extreme air pollution events in Klang, Malaysia.

Most studies discuss the probability distribution of air pollution data in terms of their magnitude or real value. In this research, we look at a different perspective by attempting to describe the probability distribution of air pollution events in terms of the characteristic defined as a duration size. The duration size of air pollution events refers to a state in which air pollution events indicate unhealthy conditions over an extended period. Thus, a large duration size implies prolonged air pollution events [17]. The probability distribution for this kind of data can be used as a building block for developing an advance statistical model, particularly for the purpose of air pollution risk assessment and analysis.

2. Study area and data
This study uses Klang City, which is located at Peninsular Malaysia, as a case study. Klang has a dense population and is one of the large cities in Malaysia with a land area of approximately 573 km$^2$. Figure 1 shows the location of Klang City in the map of Peninsular Malaysia [18].

![Figure 1. (a) Map of Peninsular Malaysia, the red dot indicates the location of Klang; (b) Map of Klang.](image-url)
Klang is involved in many industrial activities and plays an important role for economic interests in Malaysia. The most important one is the import and export activities, which operate in Port Klang. Klang has also been recognized as the 13th busiest trans-shipment port and the 16th busiest container port in the world [19]. However, the rapid development of Klang with respect to urban commercial and industrial areas in recent decades has elevated its risk of atmospheric pollution [20]. Given the importance of its industrial activities, analyzing the behaviors of air pollution duration events in Klang on the basis of previous data is important. The data used this study comprise the hourly API data in Klang for the period of January 1, 1997 to December 31, 2018. The API values are evaluated based on the breakpoints of 50, 100, 200, 300, 400 and 500. These breakpoints correspond to air quality status represented as; Healthy [0–50], Moderate [50–100), Unhealthy [100–200), Very Unhealthy [200–300), and Hazardous (above 300) [21]. The data has a small percentage of missing values at random points. Thus, to estimate these missing values, the method of single imputation based on the average of the last known and next known observations is used. This method is easy to implement and can provide good results for missing data with random behavior [22]. Next, based on the hourly API data, the duration size data for air pollution events are determined using the following equation

\[ D_i = \sum_{j=1}^{N} I_j \left( API_i \right), \quad \text{for} \quad i = 1, 2, 3, \ldots, n, \]  

where \( N \) is the total number of observations and \( I_j \left( API_i \right) \) is an indicator function determined as follows

\[ I_j \left( API_i \right) = \begin{cases} 1, & \text{if} \quad API_i > 100, \\ 0, & \text{if} \quad API_i \leq 100. \end{cases} \]  

3. Methodologies
Several statistical models have been selected to describe the distribution of duration time for unhealthy air pollution events in Klang. These selected distributions are popular and have been used to model the duration time data for flood or drought events in the hydrological analysis [23, 24]. We believe that these selected distributions can also be useful in modeling the distribution of duration time for unhealthy air pollution events.

3.1. Exponential distribution (EX)
EX is a simple statistical model that only has one parameter. Nevertheless, EX has been used as a basis model for many applications in diverse areas (see [25] for example). The probability density function (PDF) for EX is given as:

\[ f(x) = \frac{1}{\theta} \exp \left( -\frac{x}{\theta} \right), \]  

where \( \theta \) is a scale parameter. Its cumulative density function (CDF) can be written as

\[ F(x) = 1 - \exp \left( -\frac{x}{\theta} \right). \]
The maximum likelihood estimator (MLE) for the parameter of EX can be obtained using the
following equation

$$\hat{\theta} = \bar{x}. \quad (5)$$

3.2. Gamma distribution (GA)

GA is a popular statistical model in various fields, including theoretical and applied research. In terms
of environmental modeling, the application of GA can be found in the research on wind speed behaviors
[26], hydrological analysis [27], air pollution analysis [15, 28], and many more. The PDF for GA is
given as follows:

$$f(x) = \frac{1}{\Gamma(\alpha)} \beta^{\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right), \quad (6)$$

where $\alpha$ is a shape parameter and $\beta$ is a scale parameter. The CDF of GA can be written as

$$F(x) = \frac{\gamma(\alpha, \frac{x}{\beta})}{\Gamma(\alpha)}, \quad (7)$$

where $\gamma(\cdot)$ is lower incomplete Gamma function. The estimated parameters for GA based on MLE can
be obtained using the following equations

$$\hat{\beta} = \frac{\bar{x}}{\alpha}, \quad (8)$$

and

$$\ln(\hat{\alpha}) - \psi(\hat{\alpha}) = \ln \left(\frac{1}{n} \sum_{i=1}^{n} x_i\right) - \frac{1}{n} \sum_{i=1}^{n} \ln x_i. \quad (9)$$

3.3. Lognormal distribution (LN)

LN is a statistical distribution of logarithmic values derived from related normal distributions. LG is also
a popular model with ubiquitous applications [29, 30]. The PDF for LN is given as follows:

$$f(x) = \frac{1}{x\sigma \sqrt{2\pi}} \exp \left[\frac{-(\ln(x) - \mu)^2}{2\sigma^2}\right], \quad (10)$$

where $\mu$ is a location parameter and $\sigma$ is a shape parameter. The CDF for LN can be written as

$$F(x) = \frac{1}{2} \left[1 - \frac{1}{2} \text{erf} \left(\frac{\ln(x) - \mu}{\sigma \sqrt{2}}\right)\right], \quad (11)$$
where \( \text{erf}(\cdot) \) is complementary error function. The MLEs for LN are given as

\[
\hat{\mu} = \frac{\sum_{i=1}^{n} \ln x_i}{n},
\]

and

\[
\hat{\sigma} = \frac{\sum_{i=1}^{n} (\ln x_i - \hat{\mu})^2}{n}.
\]

### 3.4. Weibull distribution (WE)

The WE distribution is also a popular statistical model in many fields of applied research involving various kinds of environmental data [27, 31]. The PDF for WE is given as

\[
f(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^\alpha\right],
\]

where \( \alpha \) is a shape parameter and \( \beta \) is a scale parameter. The CDF for WE is given as

\[
F(x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^\alpha\right].
\]

Next, the estimated parameter for WE based on MLE can be computed using the following equations

\[
\hat{\alpha} = \left[\left(\sum_{i=1}^{n} x_i^\alpha \ln x_i\right)\left(\sum_{i=1}^{n} x_i^\alpha\right)^{-1} - n \sum_{i=1}^{n} \ln x_i\right]^{-1},
\]

and

\[
\hat{\beta} = \left[\frac{1}{n} \sum_{i=1}^{n} x_i^\alpha\right]^{\frac{1}{\alpha}}.
\]

### 3.5. Goodness-of-fit measurement

The performance of each will be evaluate using three different goodness-of-fit measures known as Akaike’s Information Criterion (AIC), Kolmogorov-Smirnov statistic (K-S statistic), and \( R^2 \) coefficient. The AIC formula is given as follow

\[
\text{AIC} = -2\log(L) + 2k,
\]

where \( k \) is the number of the parameter and \( L \) is the likelihood function on each fitted model respectively [32, 33]. While, the formula for K-S statistic is given as follow

\[
D_n = \sup_{x_i} | F_n(x_i) - F(x_i) |,
\]
where $\sup x$ is the supremum of the set of distances between empirical CDF $F_n$ and theoretical CDF $F(x)$ on each fitted model respectively. The lower the values determine by AIC and K-S statistic provide an indication of a better fitted model to the observed data [34]. Apart from that, the formula for $R^2$ coefficient given as follow

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{F}_i - \bar{F})^2}{\sum_{i=1}^{n} (\hat{F}_i - \bar{F})^2 + \sum_{i=1}^{n} (F_i - \bar{F})^2}.$$  \hspace{1cm} (20)

The higher the value of $R^2$ coefficient indicates a better fitted model to the observed data [34].

4. Results

Before conducting a detailed analysis regarding the modeling of the distribution of duration size for unhealthy air pollution events, exploring the descriptive statistics on the data can be informative. Table 1 shows the descriptive statistics for the duration data of unhealthy air pollution events. As presented in table 1, the mean duration size of unhealthy air pollution events is approximately 21.393 hours with a standard deviation equal to 35.318 hours. It implies that the variability of the duration size among the air pollution events is very high corresponds to its mean. Meanwhile, the median is found to be only three hours, which is evidently lower than its mean. The range between minimum value (one hour) and the maximum value is quite large (224 hours), indicating the presence of some points as extreme duration sizes. Likewise, the skewness and kurtosis values are unequal to zero, suggesting that the duration data of unhealthy air pollution events in Klang do not follow a normal distribution. Thus, this study attempts to determine the statistical model, which can suitably describe all these properties of the duration data of pollution events.

| Variable | Mean | Median | Standard Deviation | Min. Value | Max. Value | Skewness | Kurtosis |
|----------|------|--------|--------------------|------------|------------|----------|----------|
| Duration | 21.393 | 3      | 35.318             | 1          | 224        | 2.757    | 9.125    |

Table 2 provides the results of the estimated parameters for each model using the MLE approach. On the basis of the estimated parameters, Figure 2 illustrates the graphical representation of the density plot and the CDF plot on each fitted statistical model to the data duration time for unhealthy air pollution events in Klang. PDF shows that all the fitted statistical models can represent the properties of the empirical data. First, the distribution of empirical data is skewed to the right. Thus, instead of the mean, the median should be used as a robust estimate for the center measure. Second, the existence of long-tail properties is represented by some extreme data points. This behavior provides a similar argument with the results found in the descriptive statistics in table 1. Likewise, based on PDF and CDF plots, we believe that the Gamma, Lognormal, and Weibull can be provide good fitted models to describe the distribution of the duration size data for unhealthy air pollution events.

| Variable | Exponential | Gamma | Lognormal | Weibull |
|----------|-------------|-------|-----------|--------|
| Duration | 21.413      | 0.498 | 42.918    | 1.786  |

Table 2. Results of the estimated parameters for each model
Figure 2. PDF and CDF plots on each fitted statistical model for the duration data of unhealthy air pollution events in Klang.

However, deciding which model is the best approximated one to the empirical data on the basis of the graphical representations of probability densities is rough. Thus, a further evaluation must be conducted using several goodness-of-fit (GOF) measures, such as Akaike’s information criterion (AIC), Kolmogorov–Smirnov (KS) statistic, and $R^2$ coefficient. A low value of AIC indicates less information loss on a particular model fitted. A low value of KS statistic indicates a high level of similarity between the empirical CDF and the CDF of a fitted model. A high value of $R^2$ coefficient implies that the fitted model covers a high degree of variation on the empirical data. Table 3 presents the results of the GOF evaluation for all fitted distribution models. As shown in table 3, all the GOF measures find that Lognormal is the best model to provide an approximation to the distribution of duration size data for unhealthy air pollution events in Klang. Considering this information, we suggest that the Lognormal distribution can be used as a marginal model for a further analysis involving the risk of the occurrence of hazardous air pollution events corresponding to their duration size.

Table 3. Results of the GOF evaluation on each fitted model

| Fitted Distribution | AIC       | KS-statistic | $R^2$ coefficient |
|---------------------|-----------|--------------|------------------|
| Exponential         | 1862.884  | 0.375        | 0.3152           |
| Gamma               | 1764.419  | 0.213        | 0.1673           |
| **Lognormal**       | **1707.880** | **0.197**    | **0.9347**       |
| Weibull             | 1745.006  | 0.204        | 0.1399           |

5. Conclusion
The duration sizes for air pollution events indicate the behaviors of consecutive periods of unhealthy air pollution conditions over time. Thus, a large duration size implies a prolonged air pollution event. To provide a preliminary analysis regarding this topic, a case study is conducted using the data of Klang, Malaysia for the period of January 1, 1997–December 31, 2018. Based on the empirical data, this study investigates a potential probability model that can be used to represent such data. The suitable probability model can be used as a building block for developing an advance statistical model for the purpose of air pollution risk assessment and analysis. Given the duration data for unhealthy air pollution...
events, the properties of unimodal distributions, and distributions skewed to the right, four well-known statistical models that satisfy these properties, namely, Exponential, Gamma, Lognormal, and Weibull are fitted to represent the data. The method of maximum likelihood is used as a parameter estimation, whereas AIC, KS statistic, and $R^2$ coefficient are employed to evaluate the GOF of each fitted model. Overall, the results reveal that the Lognormal distribution provides the best fitted model to the duration size data for unhealthy air pollution events in Klang. In future studies, the Lognormal distribution can potentially be used as a marginal model for developing a complex statistical model for the risk assessment of the occurrence of hazardous air pollution events corresponding to their duration size.

Acknowledgement
The author is indebted Malaysian Department of Environment for providing air pollution data. This research would not be possible without the sponsorship from the Universiti Kebangsaan Malaysia (grant number FRGS/1/2014/SG04/UKM/03/1 and DIP-2018-038)

References
[1] Küchenhoff H and Thamerus M 1996 Extreme value analysis of Munich air pollution data Environmental and Ecological Statistics 3 pp 127–141
[2] Masseran N, Razali A M, Ibrahim K and Latif M T 2016 Modeling air quality in main cities of Peninsular Malaysia by using a generalized Pareto model Environmental Monitoring and Assessment 188 (1) p 65
[3] Martins L D, Wikauts C F H, Capucim M N, de Almeida D S, da Costa S C, Albuquerque T, Carvalho V S B, de Freitas E D, de Fátima Andrade M and Martins J A 2017 Extreme value analysis of air pollution data and their comparison between two large urban regions of South America Weather and Climate Extremes 18 pp 44–54
[4] Masseran N and Safari M A M 2020 Risk assessment of extreme air pollution based on partial duration series: IDF approach Stochastic Environmental Research and Risk Assessment 34 (3–4) pp 545–559
[5] Al-Dhurafi N A, Razali A M, Masseran N and Zamzuri ZH 2016 The probability distribution model of air pollution index and its dominants in Kuala Lumpur AIP Conference Proceedings, 1784 050010
[6] Al-Dhurafi N A, Masseran N, Zamzuri Z H and Safari M A M 2018 Modeling the Air Pollution Index based on its structure and descriptive status Air Quality, Atmosphere and Health 11 (2) pp 171–179
[7] Leiva V, Vilca F, Balakrishnan N and Sanhueza A 2010 A Skewed Sinh-Normal distribution and its properties and application to air pollution data Communications in Statistics – Theory and Methods 39 (3) pp 426–443
[8] Bartoletti S and Loperfido N 2010 Modelling air pollution data by the skew-normal distribution Stochastic Environmental Research and Risk Assessment 24 pp 513–517
[9] Kan H-D and Chen B-H 2004 Statistical distributions of ambient air pollutants in Shanghai, China Biomedical and Environmental Sciences 17 pp 366–372
[10] Nadarajah S 2008 A truncated inverted beta distribution with application to air pollution data Stochastic Environmental Research and Risk Assessment 22 pp 285–289
[11] Sak H, Yang G, Li B and Li W 2017 A copula-based model for air pollution portfolio risk and its efficient simulation Stochastic Environmental Research and Risk Assessment 31 pp 2607–2616
[12] Masseran N 2017 Modeling fluctuation of PM$_{10}$ data with existence of volatility effect Environmental Engineering Science 34 (11) pp 816–827
[13] AL-Dhurafi N A, Masseran N and Zamzuri Z H 2020 Hierarchical-Generalized Pareto model for estimation of unhealthy air pollution index Environmental Modeling and Assessment 25 (4) pp 555–564
[14] Masseran N and Hussain S I 2020 Copula modelling on the dynamic dependence structure of multiple air pollutant variables Mathematics 8 (11) p 1910
[15] Zhang H, Zhang W, Palazoglu A and Sun W 2012 Prediction of ozone levels using a Hidden Markov Model (HMM) with Gamma distribution Atmospheric Environment 62 pp 64–73
[16] Masseran N and Mohd Safari M A 2020 Intensity–duration–frequency approach for risk assessment of air pollution events Journal of Environmental Management 264 110429
[17] Masseran N 2021 Power-law behaviors of the duration size of unhealthy air pollution events Stochastic Environmental Research and Risk Assessment (Article in press)
[18] Google 2019 Source : https://maps.googleapis.com/maps/api/geocode/json?address=Klang%2CSelangor&key=xxx
[19] Al-Dhurafi N A, Masseran N, Zamzuri Z H and Razali A M 2018 Modeling unhealthy air pollution index using a Peaks-Over-Threshold method Environmental Engineering Science 35 (2) pp 101–110.
[20] Azmi S Z, Latif M T, Ismail A S, Juneng L and Jemain A A 2010 Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia Air Quality, Atmosphere & Health 3 pp 53–64.
[21] Alyousifi Y, Masseran N and Ibrahim K 2018 Modeling the stochastic dependence of air pollution index data Stochastic Environmental Research and Risk Assessment 32(6) pp 1603–1611
[22] Masseran N, Razali A M, Ibrahim K, Zaharim A and Sopian K 2013 Application of the single imputation method to estimate missing wind speed data in Malaysia Research Journal of Applied Sciences, Engineering and Technology 6 pp 1780–1784
[23] Lee T, Modarres R and Ouarda T B M J 2013 Data-based analysis of bivariate copula tail dependence for drought duration and severity Hydrological Process 27 pp 1454–1463
[24] Janga Reddy M and Ganguli P 2012 Application of copulas for derivation of drought severity–duration–frequency curves Hydrological Process 26 pp 1672–1685
[25] Masseran N, Razali A M, Ibrahim K, Wan Zin W Z and Zaharim A 2012 On spatial analysis of wind energy potential in Malaysia WSEAS Transactions on Mathematics 6 (11) pp 467–477
[26] Masseran N 2015 Evaluating wind power density models and their statistical properties Energy 84 pp 533–541
[27] Zhang L and Singh V P 2019 Copulas and their applications in water resources engineering (Cambridge, United Kingdom: Cambridge University Press)
[28] Jiang P, Li C, Li R and Yang H 2019 An innovative hybrid air pollution early-warning system based on pollutants forecasting and Extenics evaluation Knowledge-Based Systems 164 pp 174–192
[29] Chastko K and Adams M 2019 Improving long-term air pollution estimates with incomplete data: A method-fusion approach MethodsX 6 pp 1489–1495
[30] Reiminger N, Jurado X, Vazquez J, Wemmer C, Blond N, Wertel J and Dufresnea M 2020 Methodologies to assess mean annual air pollution concentration combining numerical results and wind roses Sustainable Cities and Society 59 102221
[31] Masseran N, Razali A M and Ibrahim K 2012 An analysis of wind power density derived from several wind speed density functions: The regional assessment on wind power in Malaysia Renewable and Sustainable Energy Reviews 16 pp 6476–6487
[32] Hirotugu A 1974 A new look at the statistical model identification IEEE Transactions on Automatic Control 19 pp 716–723
[33] Burnham K P and Anderson D R 2002 Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed (New York, United States: Springer-Verlang)
[34] Masseran N 2018 Integrated approach for the determination of an accurate wind-speed distribution model Energy Conversion and Management 173 pp 56–64