FITTING MALAYSIAN WIND SPEED VIA LOG-NORMAL DISTRIBUTION

Mohamed N.S., Kamisan N.A.B., Zubairi Y.Z., Hussin A.G.
Centre for Foundation Science Study Universiti Malaya, Universiti Malaya, 50603 Kuala Lumpur, Malaysia.

e-mail: nsabirin@um.edu.my, yzulina@um.edu.my, nurarinab@yahoo.com, , ghapor@um.edu.my

Abstract: Lognormal is one of the distributions in the family of the extreme value distribution type. The lognormal distribution is a single tailed probability distribution of any random variables whose logarithm is normally distributed. As may be concluded by the name, the lognormal distribution has certain similarities to the normal distribution and describes many naturally occurring populations; for example, lognormal distribution has been used in many disciplines such as economics, life sciences and meteorology. Thus paper evaluates the performance of lognormal as a model to describe the Malaysian annual maximum wind speed for four locations (Alor Setar, Langkawi, Melaka and Senai). Based on rigorous tests of root mean square error, mean absolute error and mean absolute percentage error and density plots, the results suggest that the lognormal is an adequate model in describing the maximal wind speed, thus contributing towards a better understanding of extreme wind behavior via robust statistical model.
Keyword: Lognormal distribution, maximum wind speed

1. Introduction

In statistics, lognormal distribution is the single – tailed probability distribution of any random variables whose logarithm is normally distributed. As may be concluded by the name, the lognormal distribution has certain similarities to the normal distribution. In other words, random variable is lognormally distributed if the logarithm of the random variable is normally distributed. Because of this, there are many mathematical similarities between the two distributions.

Within the area of reliability analysis, lognormal has been applied to the time to failure of equipment used and maintenance efficiency (Barabady, 2007) and coefficient of friction of wear and tear (Steele, 2007). Within the area of life sciences, the lognormal distribution was used to describe the nucleation and growth processes (Bergmann, 2008) and model defragmentation habitat in ecological studies (Heide, 2008).

Within the area of meteorology, the lognormal is used to estimate spatial and temporal statistical properties of local daily mean wind speed under global climate change (Garcia, 1997). There are also researchers that used lognormal distribution to describe the turbulence intensity distribution in optimizes the wind turbines building cost (Hansen, 2004).

In this paper, the lognormal distribution is proposed to model the Malaysian wind data. Based on the distribution of the plots of histogram of the maximum wind speed which exhibit a long tail distribution and skewness to the right, the lognormal seems to be a suitable candidate to describe the data. Based on several performance tests, the suitability of lognormal is discussed. This was also supported by some other researcher that also

used lognormal distribution to model their pressure data since the tail of the distribution is higher than the normal distribution (Gurley, 1997).

2. Formulation of Lognormal Distribution

A variable X is lognormally distributed if $Y=ln(X)$ is normally distributed with "*ln*" denoting the natural logarithm. The general formula for the probability density function of the lognormal distribution is:

$$f(x;\mu,\sigma)=\frac{e^{-((\ln(x)-\mu)^2/2\sigma^2}}{x\sigma\sqrt{2\pi}},x\geq 0;\sigma>0 \tag{1}$$

where σ is the shape parameter and μ is the location parameter.

The cumulative distribution function of lognormal distribution is:

$$F(x;\mu,\sigma)=\int_0^x \frac{1}{\sqrt{2\pi}\sigma x}\exp\left\{-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right\}dx \tag{2}$$

3. Estimation of parameters

Maximum likelihood estimation (MLE) are often used to estimate the unknown parameters μ and σ. Maximum likelihood parameter estimation will determine the parameters that maximize the probability (likelihood) of the sample data. From a statistical point of view, the method of maximum likelihood is considered to be more robust (with some exceptions) and yields estimators with good statistical properties. In other words, MLE methods are versatile and applicable to most models and to different types of data. In addition, they provide efficient methods for quantifying uncertainty through confidence bounds. Although the methodology for maximum likelihood estimation is simple, the implementation is mathematically intense. The MLE method has very desirable properties. By letting $x_1,x_2,...,x_n$ be a random sample size n drawn at random and from the pdf, $f(x;\mu,\sigma)$ unknown parameters, the likelihood function is as follows:

$$L=\prod_{i=1}^{n} f(x_i;\mu,\sigma) \tag{3}$$

Where μ and σ are unknown parameters. The likelihood function of lognormal can be expressed by

$$L(x_1,...,,x_n,\mu,\sigma)=\frac{1}{2\pi^{\frac{n}{2}}\sigma^n}\prod_{i=1}^{n}\frac{1}{x_i}e^{\frac{-\sum_{i=1}^{n}(\ln x_i-\mu)^2}{2\sigma^2}} \tag{4}$$

To maximize the μ and σ, the equation (4) is logged with the log-likelihood function as below.

$$\ln L(x_1,...,x_n,\mu,\sigma) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma + \ln\sum_{i=1}^{n}\frac{1}{x_i} - \sum_{i=1}^{n}\frac{(\ln x_i - \mu)^2}{2\sigma^2} \quad (5)$$

By differentiating equation (5) partially with respect to σ and μ, the equations (6) and (7) are obtained as below.

$$\frac{\partial \ln L}{\partial \mu} = \sum_{i=1}^{n} -\frac{\ln x_i - \mu}{\sigma^2} \quad (6)$$

$$\frac{\partial \ln L}{\partial \sigma} = \frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(\ln x_i - \mu)^2}{\sigma^3} \quad (7)$$

By letting (6) and (7) equal to zero the parameters estimate are obtained namely $\hat{\mu}$ and $\hat{\sigma}$ given as

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\ln x_i}{n} \quad (8)$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\ln x_i - \left[\frac{\sum_{i=1}^{n}\ln x_i}{n}\right]\right]^2 \quad (9)$$

4. Performance indicator

To asses the suitability of the proposed model several statistics test can be used. In this study, the performance measure used are root mean squared (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE). RMSE measures the average mismatch between each data point and the model. RMSE is one of the best measurements to communicate the results to others. In general small values obtained from the performance indicator suggest suitability of the proposed model. In comparison of the three measures, the RMSE is more sensitive to outlying observations as they contribute huge values in the square deviations measure. Nevertheless, RMSE is frequently used to measure the differences between values predicted by a model or an estimator and the observed values. These individual differences are also called residuals. The following are the formulation of the performance indicators.

The formula for RMSE is given by

$$\sqrt{\frac{\sum_{i=1}^{n}(x_{p,i}-x_{o,i})^2}{n}} \tag{10}$$

where $x_{p,i}$ is the predicted value and $x_{o,i}$ is the observed value.

MAE is a weighted average of the absolute errors, with the relative frequencies as the weight factors. The MAE gives slightly smaller value than RMSE and it is an easier statistic to understand than the RMSE. It is also recommended by *Matsuura et al(2005)* as it has an unambiguous measure of average error magnitude.

The formula for MAE is given by

$$\frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{11}$$

where f_i is the prediction and y_i is the true value.

MAPE is used to measure the accuracy in a fitted time series value in statistics, specifically trending. MAPE is also often useful for purposes of reporting because it is expressed in generic percentage terms. The formula for MAPE is given by

$$\frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{12}$$

where A_t is the actual value and F_t is the predicted value. To complement the model checking procedures, pdf and cdf plot are also used.

5. Source of Data

The source of data is from the Malaysian Meteorological Department which record maximum wind speed in four locations in year 2005. They are Alor Setar which is located in the northern part of peninsular Malaysia, Senai located in the southern part, whilst Melaka in the mid-west part of the peninsular. Maximum wind speed was also recorded in Langkawi, an island off coast of the northern part of Malaysia.

Figure 1: Map of four locations in Malaysia

Table 1: Descriptive statistics of maximum wind speed of four locations

Location	Mean (m/s)	Min (m/s)	Max (m/s)	n	$\hat{\sigma}$
Alor Setar	8.4028	2.9	16.2	357	2.3250
Senai	8.2104	3.6	16.3	364	2.0745
Langkawi	8.5523	3.8	23.8	363	2.7347
Melaka	8.6975	3.7	16.0	365	2.3138

Referring to Table 1, it can be seen that the mean of the wind speed for all of the locations is approximately 8.4 m/s and the standard deviation is small (approximately 2.4) for all of the locations which implies that the variability is small. The spread of the data is the smallest in Senai and the largest spread is observed in Langkawi.

5. Result and discussion

5.1 Parameters estimates of lognormal distribution

For each of the four locations, lognormal distribution was fitted and the parameters estimation of $\hat{\mu}$ and $\hat{\sigma}$ are obtained and given in Table 2.

Table 2: Parameter estimates of lognormal distribution

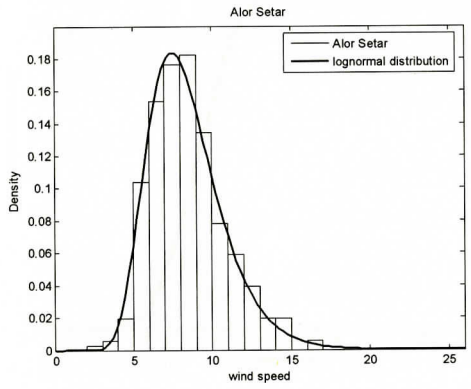
Location	$\hat{\mu}$	$\hat{\sigma}$

Alor Setar	2.0905	0.2788
Senai	2.0745	0.2487
Langkawi	2.0969	0.3150
Melaka	2.1287	0.2625

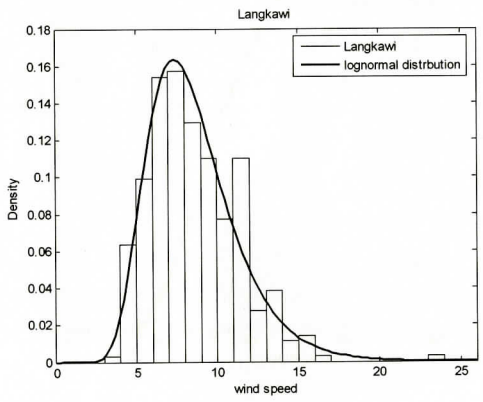
From Table 2, the parameters estimates are approximately the same in all four locations with mean about 2.0977m/s and standard deviation 0.2763. In comparison, with the sample statistics of mean and standard deviation in Table 1, there seems to be a general agreement on the estimates of measure of the central location and measurement of the spread parameter and taking he logged values, there seems to be a general agreement on the estimates of the measure of central location and measurement of the spread parameter.

5.2 The probability density function (PDF)

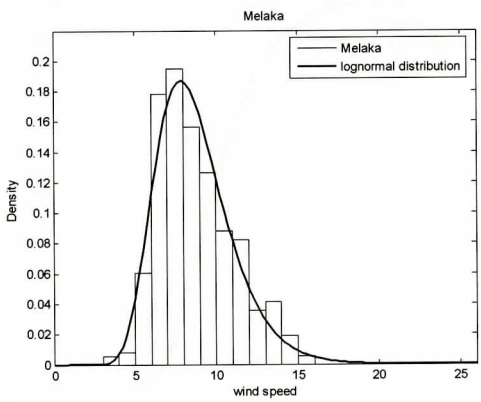
The probability density function (pdf) is a function that represents a probability distribution in terms of integrals. Informally, a probability density function can be seen as a "smoothed out" version of a histogram.



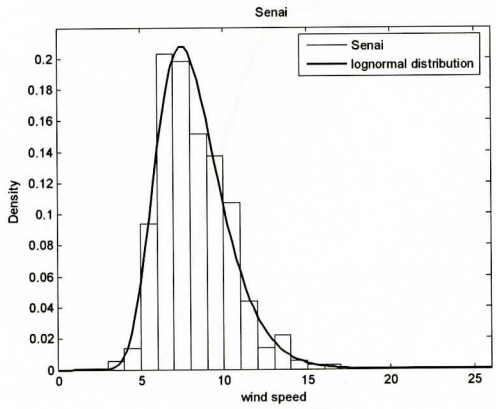
2a) Alor Setar



2b) Langkawi



2c) Melaka



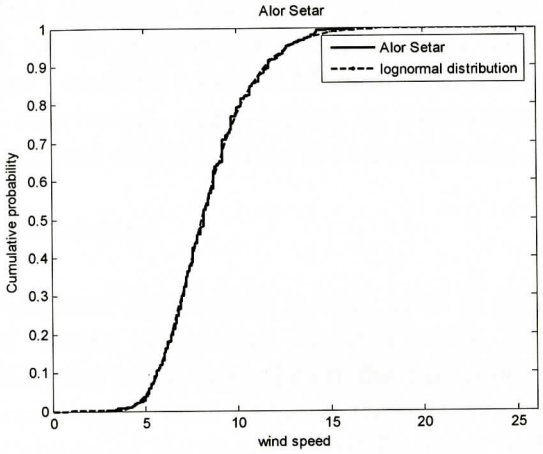
2d) Senai

Figure 2: Pdf plots of four locations

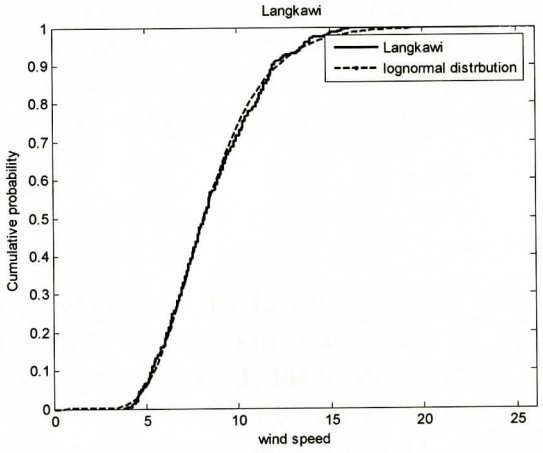
From the superimposed plots of pdf and the histogram of the observed values in Figure 1, the lognormal seems to be a sufficient model in all the locations. By comparing the pdf plots at four locations, Langkawi shows some disagreement. This is expected as the distribution of Langkawi has a wider spread as compared to others. The graph is skewed to the right and so is the pdf plot. The peak of the pdf plot is also on the highest bin of the histogram. In short, the shape of the histogram and the shape of the pdf plot are identical.

5.3 The cumulative distribution function (CDF)

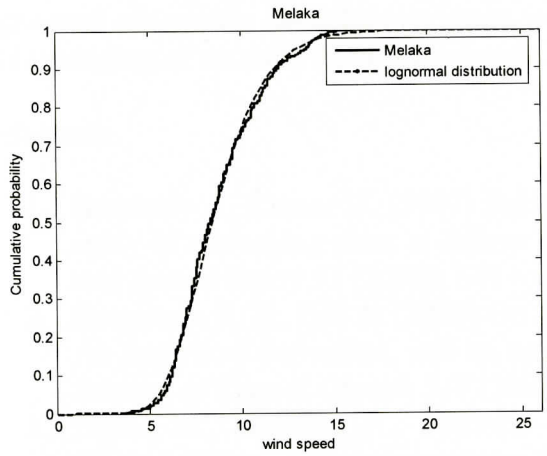
Other than pdf plot, we also could use cdf plot to check on goodness of fit of the data. The cumulative density function (cdf) plot is useful for examining the distribution of a sample of data. A theoretical cdf can be overlaid on the same plot with the empirical plot to compare the empirical distribution of the sample with the theoretical distribution. Below is the cdf plot for each of the locations.



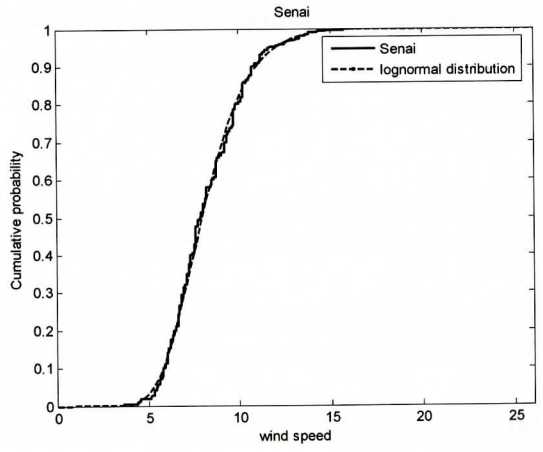
3a) Alor Setar



3b) Langkawi



3c) Melaka



3d) Senai

Figure 3: Cdf plot for four locations

The plots show that the observed data plot and the predicted data plot are very close on each other. The comparison of the empirical cdf plot (observed data) with the fitted cdf plot of lognormal distribution (predicted data) suggests lognormal distribution adequately fit the wind speed data. To further assess the goodness of fit of the model, statistical measure of RMSE, MAE and MAPE are evaluated.

Table 3: Performance indicator

Location	RMSE	MAE	MAPE
Alor Setar	3.0632	2.3604	0.2901
Senai	2.9914	2.4092	0.3002
Langkawi	3.7441	2.9419	0.3617
Melaka	3.1853	2.5864	0.3123

From the result shown in Table 3, the performance indicators show small values for each of the locations. Consistent with the earlier plots of pdf, Langkawi has the largest value for all the performance indicators in comparison to other locations whilst Alor Setar shows smallest value on MAE and MAPE and Senai shows the smallest value for RMSE. In conclusion, small values of performance indicator obtained suggest a good fit of lognormal distribution to the recorded data.

6. Conclusion

Lognormal distribution is often used in describing data due to the flexibility of the scale and shape parameters. In this study, the lognormal distribution was fitted to the Malaysian wind data. From the pdf plots and cdf plots obtained, there seems to be a general agreement between the observed and estimated values. To further assess the suitability of the model, RMSE, MAE and MAPE performance indicates used. Small values for each of the locations indicate that lognormal distribution is a good fit for Malaysian wind speed data. The significant finding suggest that two parameters lognormal adequately describe maximum wind speed; thus contributing towards a better understanding of extreme wind behavior via parametric statistical model.

Reference

1. Barabady J., Kumar U. (2007) Reliability analysis of mining equipment: A case study of a crushing plant at Jajarm Bauxite Mine in Iran. *Reliability Engineering and System Safety 93* (2008): 647 – 653
2. Bergmann R. B., Bill A. (2008) On the origin of loghrithmic-normal distributions: An analytical derivation, and its application to nucleation and growth processes. *Journal of Crystal Growth 310* (2008): 3135 – 3138
3. Evans M., Hastings N., Peacock B. (1993) *Statistical Distributions*, Second Edition, Wiley – Interscience, New York

4. Garcia A., Torres J. L., Prieto E., De Francisco A. (1997) Fitting wind speed distributions: A case study, *Solar Energy* (1998) Vol 62: 139 – 144
5. Steele C. (2007) Use of the lognormal distribution for the coefficients of friction and wear. *Reliability Engineering and System Safety 93* (2008): 1574 – 1576
6. Van Der Heide C. M., Van Den Bergh J. C. J. M., Van Ierland E. C., Nunes (2008) Economic valuation of habitat defragmentation: A study of the Veluwe P. A. L. D., the Netherlands. *Ecological Economics* (2008): 1 – 12
7. Walpole R. E., Myers R. H., Myers S. L., Ye K. (2002) *Probability & Statistics for engineers and scientists*, Seventh Edition, Pearson Education International, London
8. Kurt S. Hansen, Gunner Chr Larsen (2004) Design turbulence intensity. *IEA Annex XVII, Database on Wind Characteristics*, Riso National Laboratory, Roskilde (June 2004)
9. Gurley K. and Kareem A. (1997) Analysis Interpretation Modeling and Simulation of Unsteady Wind and Pressure Data. *Journal of Wind Engineering and Industrial Aerodynamics*, Vol.69-71, pp.657-669