Best Fitted Distribution For Meteorological Data In Kuala Krai

Siti Mariam Norrulashikin [1*], Fadhilah Yusof [2], Siti Rohani Mohd Nor[3] & Nur Arina Bazilah Kamisan[4]

[1, 2, 3, 4]*Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor Darul Takzim*

**Corresponding author: sitimariam@utm.my*

https://doi.org/10.22452/josma.vol3no1.2

Abstract

Modeling meteorological variables is a vital aspect of climate change studies. Awareness of the frequency and magnitude of climate change is a critical concern for mitigating the risks associated with climate change. Probability distribution models are valuable tools for a frequency study of climate variables since it measures how the probability distribution able to fit well in the data series. Monthly meteorological data including average temperature, wind speed, and rainfall were analyzed in order to determine the most suited probability distribution model for Kuala Krai district. The probability distributions that were used in the analysis were Beta, Burr, Gamma, Lognormal, and Weibull distributions. To estimate the parameters for each distribution, the maximum likelihood estimate (MLE) was employed. Goodness-of-fit tests such as the Kolmogorov-Smirnov, and Anderson-Darling tests were conducted to assess the best suited model, and the test's reliability. Results from statistical studies indicate that Burr distributions better characterize the meteorological data of our research. The graph of probability density function, cumulative distribution function as well as Q-Q plot are presented.

Keywords: Burr, Meteorology, Goodness-of-fit, Maximum Likelihood Estimation, Probability distribution

1. Introduction

Climate phenomena are observed to be multifaceted. The complexity of their multifaceted existence has only begun to be understood. It should be argued that even though chances of extreme weather events such as hurricanes, heat waves, floods, and storms are higher in certain areas, these occurrences can occur naturally all over the world and are not man-made. Many of the extreme weather events can be negatively blamed on natural phenomena as opposed to man-made. In this context, the normal decadal

or multi-decadal fluctuations in the environment provide the surroundings for anthropogenic climate changes (Seneviratne et al., 2012). Since this phenomenon is outlined, humans were able to understand more about the processes that happen in the broader picture of existence. In comparison, the climate is often generally defined as the average weather in a specific region, influenced by rainfall patterns, temperature, humidity, wind and seasons. Climate trends affect the natural ecosystems that form in the region.

Moreover, economic activities and human societies are highly dependent on these trends. The current atmosphere is no longer what it was before, and previous statistics are not alone as a credible predictor for forecasting the future with other determinants or variables (Ehrlen & Morris, 2015). Unexpectedly fast changes arise as a result of negative or harmful effects occurring within natural environments. As far as these are concerned, there are many processes within communities that are primarily affected by climate patterns; climate affects where and how humans, plants and animals live and communicate with each other, especially in terms of food production, water availability and use, and health risks. For example, changes in the normal timing of precipitation and temperature in the atmosphere could affect the regular timing of plants and fruits and the insect hatching process and the time needed to flow as far as possible. Other than this, the well-contained pollination of plants, the reproduction of food for migrating birds, the spawning of fish, drinking water and irrigation, etc, may also be affected.

Under normal conditions, some short-term seasonal climate change is viewed as normal. A long-term occurrence may, however, suggest a climate that changes. These effects of climate change are the product of global warming. Global warming is the rise in global temperature, while changes in the external environment entail a narrower sense of climate change. Temperature and weather are primarily influenced the precipitation characteristics (Fu, Gao, Liang, & Liu, 2021). As a result, related changes in each of these aspects may have contributed to a change in precipitation. The study and modelling of the multivariable meteorological data by using the statistical distributions could provide a valuable contribution to understanding the characteristics of global warming. Knowing how climate changes can impact us is important for mitigating the risks of climate change. Probability models are useful for the study of climatological data because it measure how well the probability distribution function is able in capturing the data for each variable and each region (Perkins, Pitman, Holbrook, & McAneney, 2007). However, in a certain study it is still a matter of debate to select a suitable model. In this analysis, five probability distributions, namely Beta, Burr, Gamma, Lognormal, and Weibull (Jamaludin & Jemain, 2007; Yusof & Mean, 2012; Zaharim, Najid, Razali, & Sopian, 2009), were selected. These methods are widely used on meteorological data to simplify the data for distribution analysis. According to Zaharim et al. (2009), the distribution of Burr is the best distribution that matches well with Malaysia's wind speed data.

2. Materials and Methods

The following subsections present a summary of the data used and a brief overview of the methodology applied in this paper.

2.1 Study area and data collection

The data used in this study was collected from Kuala Krai Station, located in the center of Kelantan State, northeast of the Peninsular Region of Malaysia. Kuala Krai Station is located at an approximate latitude of 5° N and an approximate longitude of 102° E. The land is hilly and it is a region of tropical rainforest, once in great abundance. The Kuala Krai region is the meeting point of the two major rivers that form the Kelantan River at Kelantan. It flows near the capital city of Kelantan, Kota Bharu, to the South China Sea estuary. Kuala Krai experiences intense monsoons during which their average annual temperature reaches 26.8°C. Meanwhile, the average annual rainfall is 2713mm. The Malaysian Meteorological Department collected daily meteorological data consisting 24 hours of mean temperature (°C), maximum wind speed (m/s) and precipitation (mm) data from 1985 to 2009 (Norrulashikin et al., 2015).

2.2 Continuous distribution

If random variables are continuous, the probabilities of a specific value can no longer be calculated because their corresponding probabilities are all nil. Therefore, when dealing with continuous random variables, the key concern is the possibility that random variables will take values within certain intervals. The probability of the case is afforded as follows:

$$P(a \le X \le b) = \int_a^b f(x)dx$$

For continuous random variables, in order to calculate the probability, we need to integrate the probability function called the probability density function (PDF). There are a variety of standard probability functions, but the most common one is related to the standard normal random variable (Andren, 2007). The probability density function of continuous random variable X is related to its cumulative distribution function (CDF). It is denoted in the same way as a discrete random variable. However, we have to integrate from minus infinity to the selected value for the continuous random variable, i.e.:

$$F(c) = P(X \le c) = \int_{-\infty}^{c} f(X)dX$$

The five distributions used in this study is summarized in Table 1.

Table 1 Distribution functions

Distribution	PDF	CDF
Beta	$f(x)=\frac{1}{B(\alpha_1,\alpha_2)}\frac{(x-a)^{\alpha_1-1}(b-x)^{\alpha_2-1}}{(b-a)^{\alpha_1+\alpha_2-1}}$, with, -continuous shape parameter, $\alpha_1,\alpha_2>0$ -continuous boundary parameter, $a<b$	$F(x)=I_z(\alpha_1,\alpha_2)$ where, $z\equiv\frac{x-a}{b-a}$ $a\le x\le b$
Burr	$f(x)=\frac{\alpha k\left(\frac{x}{\beta}\right)^{\alpha-1}}{\beta\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{k+1}}$, with, -continuous shape parameter, $k,\alpha>0$ -continuous scale parameter, $\beta>0$ -continuous location parameter ($\gamma\equiv 0$ for the three-parameter Burr distribution)	$F(x)=1-\left(1+\left(\frac{x}{\beta}\right)^{\alpha}\right)^{-k}$ where, $\gamma\le x\le+\infty$
Gamma	$f(x)=\frac{x^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\exp(-x/\beta)$ with, -continuous shape parameter, $\alpha>0$ -continuous scale parameter, $\beta>0$ -continuous location parameter ($\gamma\equiv 0$ yields the two-parameter Gamma distribution)	$F(x)=\frac{\Gamma_{x/\beta}(\alpha)}{\Gamma(\alpha)}$ where, Γ_z - Incomplete Gamma function Γ - Gamma function $\gamma\le x\le+\infty$
Lognormal	$f(x)=\frac{\exp\left(-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right)}{x\sigma\sqrt{2\pi}}$ with -continuous parameter, $\sigma>0,\ \mu$ -continuous location parameter ($\gamma\equiv 0$ yields the two-parameter Lognormal distribution)	$F(x)=\Phi\left(\frac{\ln x-\mu}{\sigma}\right)$ where, Φ - Laplace integral $\gamma<x<+\infty$
Weibull	$f(x)=\left(\frac{\alpha}{\beta}\right)\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$ with -continuous shape parameter, $\alpha>0$ -continuous scale parameter, $\beta>0$ -continuous location parameter ($\gamma\equiv 0$ yields the two-parameter Weibull distribution)	$F(x)=1-\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$ where, $\gamma\le x\le+\infty$

2.3 Parameter estimation method

In order to obtain the probability distribution that fit the data well, there are parameter estimation method to be implemented in the formulation of probability distribution to ensure that the data fit the distribution as close as possible (Hussin & Yusof, 2020). The most accurate and commonly used parameter estimation is the Maximum Likelihood Estimation which also known as MLE method. MLE functioned by lowering the mean square error that associates with model parameter estimates. For a given function of $f(x)$, MLE is defined as the value of x that maximizes the likelihood of $f(x)$ or the logarithm of the likelihood of $f(x)$. This process will decrease the situation of an important number of unlikely outcomes for x to occur.

2.4 Goodness-of-fit test

The goodness-of-fit test is performed on statistical model to assess the degree to which the data set can be explained by that model. The measurement used for the goodness-of-fit test generally summarizes the difference between the observed values and the predicted values of the statistical model. Two goodness-of-fit measures like Kolmogorov-Smirnov (KS), and Anderson-Darling (AD) are used for this analysis at 0.05 level of significance. The null hypotheses and the alternative ones are:

- H_0: The data follows the distribution specified.
- H_A: The data does not follow the distribution specified.

i. Kolmogorov-Smirnov (KS)

The Kolmogorov-Smirnov statistic*, D* is based on the greatest vertical disparity between theoretical and empirical cumulative distribution functions.

$$D = \max_{1 \le i \le n} \left(F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F(x_i) \right)$$

When the test statistic is greater than the critical value, the hypothesis of the null hypothesis is rejected at the 5% significance level.

ii. Anderson-Darling (AD)

The Anderson-Darling procedure discusses the disparity in cumulative distribution functions between the fit and expected cumulative distribution function. Using this measure, more weight is given to the tails than to the Kolmogorov-Smirnov test. Anderson-Darling statistic is given by,

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1).\left[\ln F(X_i) + \ln(1 - F(X_{n-i+1})) \right]$$

The statistical hypothesis is rejected at 5 percent significance level if the critical value obtained from the statistical table is lower than the test statistic, A^2.

3. Results and Discussion

The results of goodness-of-fit statistics summary and the ranking measures are shown in Table 2 and Table 3, respectively. The values in Table 2 represents the test statistics value for each goodness-of-fit test. While the ranking measures in Table 3 is identified from the values in Table 2. The smallest value of statistical test will be in rank 1 and the highest will be ranked 5. Based on Kolmogorov-Smirnov, and Anderson-Darling measures at 5% significance level, the Burr distribution best fits the patterns of temperature, wind speed, and rainfall data in Kuala Krai. As a result, the most prevalent meteorological variations in Kuala Krai are found in the Burr distribution. Meanwhile, Lognormal distributions for temperature and wind speed variables ranked last in contrast to other goodness-of-fit measures, and Beta distributions for rainfall variable did the least well.

Table 2: Goodness-of-fit test statistics summary for all variables

	Kolmogorov-Smirnov			Anderson-Darling		
Distributions	T	WS	R	T	WS	R
Beta	0.043	0.048	0.105	0.691	0.509	8.788
Burr	0.031	0.032	0.050	0.416	0.364	0.937
Gamma	0.054	0.047	0.071	1.148	0.647	3.158
Lognormal	0.056	0.058	0.087	1.263	1.006	5.219
Weibull	0.040	0.045	0.077	1.357	1.779	2.303

Note: T represents temperature, WS represents wind speed, and R represents rainfall

Table 3: Goodness-of-fit test ranking for continuous distribution of meteorological variables

	Kolmogorov-Smirnov			Anderson-Darling		
Distributions	T	WS	R	T	WS	R
Beta	3	4	5	2	2	5
Burr	1	1	1	1	1	1
Gamma	4	3	2	3	3	3
Lognormal	5	5	4	4	4	4
Weibull	2	2	3	5	5	2

Note: The ranking is in the order from 1 to 5; 1 is the best ranking and 5 is the worst ranking. T represents temperature, WS represents wind speed, and R represents rainfall.

Table 4: Fitting parameters of meteorological variables depend on various distributions.

Distributions	Temperature	Wind Speed	Rainfall
Beta	α_1 =210.2600	α_1 =3.9828	α_1 =0.9651
	α_2 =87.1670	α_2 =5.2250	α_2 =4.2680
	a =1.2281	a =4.4435	a =0.6000
	b =36.5740	b =11.9710	b =1173.9000
Burr	k =2.0590	k =2.4350	k =2.9843
	α =41.0870	α =9.1328	α =1.5994
	β =26.8680	β =8.7178	β =380.0900
Gamma	α =791.0100	α =43.8270	α =1.3524
	β =0.0331	β =0.1757	β =155.2200
Lognormal	σ =0.0356	σ =0.1537	σ =0.9621
	μ =3.2658	μ =2.0299	μ =4.9959
Weibull	α =35.0620	α =8.0566	α =1.2885
	β =26.6200	β =8.1612	β =228.4400

The fitting parameters that required for the perfect fit with various distributions are documented in Table 4. Different values may be assigned to the continuous shape parameters ($\alpha_1, \alpha_2, k, \alpha$), scale parameters ($\beta, \sigma$), and location parameters (γ, μ), depending on the types of data distribution. Figure 1 illustrates the probability density functions (PDF) of Burr and Lognormal distributions, which model the meteorological data well. Burr and Lognormal were used in this illustration since Burr represents the highest ranking and Lognormal represents the lowest ranking in the list of distributions used on this study. The PDF displays graphically different wind speed data properties including parameters of shape, scale, and location. Figure 2 demonstrates the contrast between Burr and Lognormal distribution in the cumulative distribution functions (CDF). The CDF graph is useful to precisely assess how well the data observed might match the distributions. In comparison to Lognormal distribution, the Burr distribution is clearly seen to fit well with the wind speed data. Previous research found that the distribution of Burr is the best distribution that matches well with Malaysia's wind speed data (Zaharim et al., 2009). Figure 3 shows the Quantile-Quantile (Q-Q) plot for the Burr and Lognormal distribution of wind speed data. The Q-Q plot is used to determine how close a particular distribution is to the observed data. In a nutshell, by using Q-Q plot, the testing distribution is the best model if the points lie on an approximately straight line.

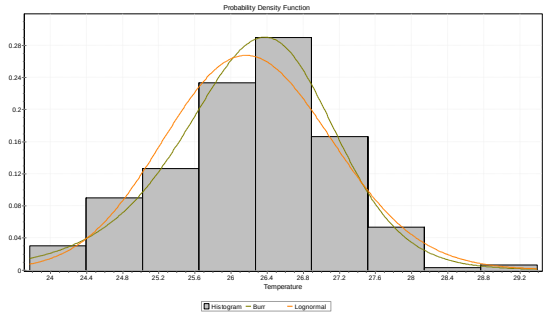

Figure 1.1: Burr and Lognormal distributions fit the temperature data.

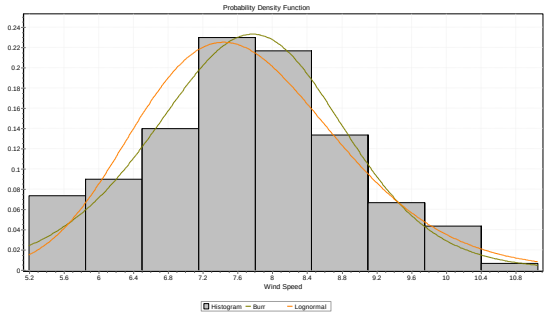

Figure 1.2: Burr and Lognormal distributions fit the wind speed data.

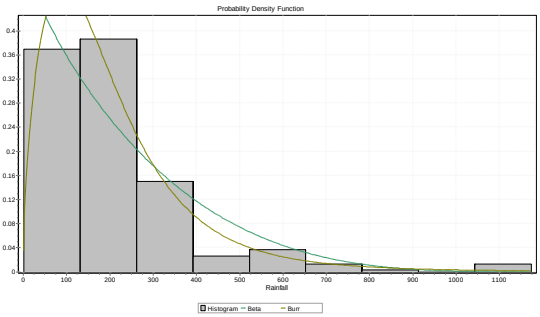

Figure 1.3: Burr and Lognormal distributions fit the rainfall data.

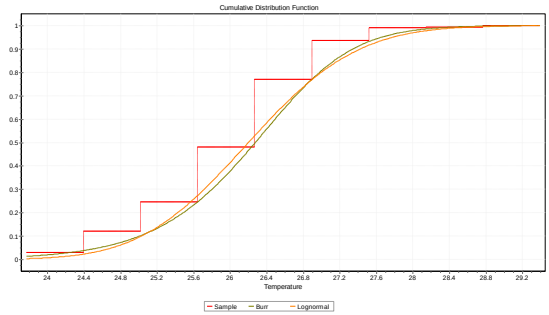

Figure 2.1: The comparison between Burr and Lognormal distributions in CDF of the temperature data.

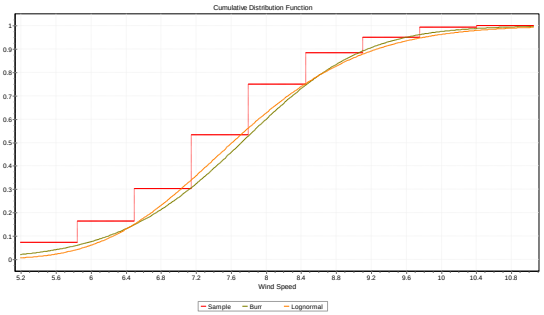

Figure 2.2: The comparison between Burr and Lognormal distributions in CDF of the wind speed data.

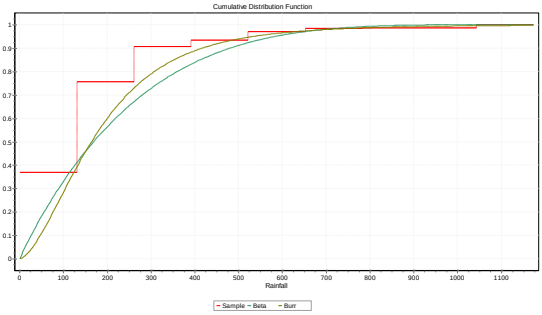

Figure 2.3: The comparison between Burr and Lognormal distributions in CDF of the rainfall data.

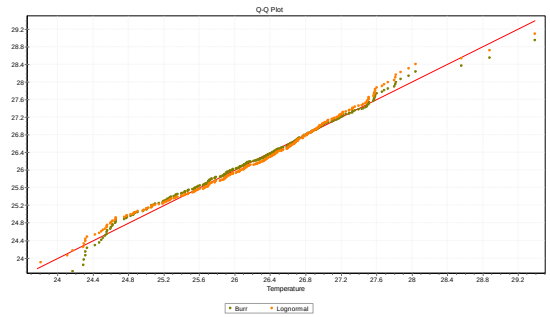

Figure 3.1: Q-Q plot of Burr and Lognormal distributions of the temperature data.

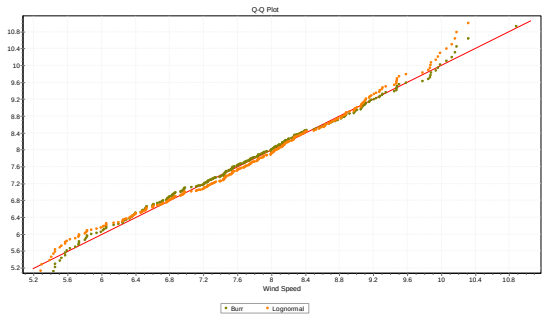

Figure 3.2: Q-Q plot of Burr and Lognormal distributions of the wind speed data.

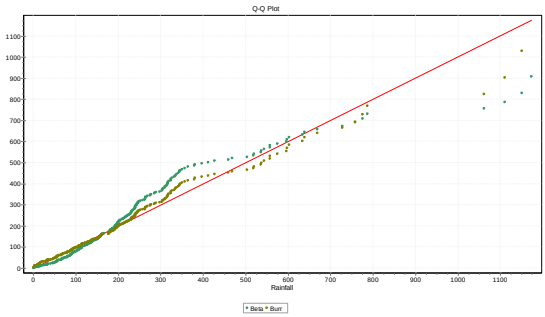

Figure 3.3: Q-Q plot of Burr and Lognormal distributions of the rainfall data.

4. Conclusion

In this analysis, five continuous probability distributions such as the Beta distribution, Burr distribution, Gamma distribution, Lognormal distribution, and Weibull distribution were evaluated and compared on three meteorological variables such as temperature, wind speed, and rainfall data for Kuala Krai Station. Two goodness-of-fit tests were attempted, which included Kolmogorov-Smirnov, and Anderson-Darling tests. The Burr distribution suits the weather patterns best, particularly when considering the three meteorological variables. The probability distribution of selected weather variables will be used to produce random weather information for Kuala Krai.

5. Acknowledgements

The authors are grateful to the Malaysian Meteorological Department for providing the daily meteorological data and to Universiti Teknologi Malaysia for funding the Research University Grant with vote no. Q.J130000.2654.17J30.

6. References

Andren, T. (2007). *Econometrics*. Ventus Publishing ApS.

Ehrlen, J., & Morris, W. F. (2015). Predicting changes in the distribution and abundance of species under environmental change. *Ecology Letters*, *18*, 303–314. https://doi.org/10.1111/ele.12410

Fu, T., Gao, H., Liang, H., & Liu, J. (2021). Spatio-temporal precipitation changes and their localized predictors in the Taihang Mountain region , North China. *Stochastic Environmental Research and*

Risk Assessment. https://doi.org/https://doi.org/10.1007/s00477-021-01970-w

Hussin, N. H., & Yusof, F. (2020). ANALYSIS OF PROBABILITY DISTRIBUTION FOR WIND SPEED DATA IN JOHOR. In *International Graduate Conference on Engineering, Science and Humanities (IGCESH2020)*.

Jamaludin, S., & Jemain, A. A. (2007). Fitting The Statistical Distributions To The Daily Rainfall Amount in Peninsular Malaysia. *Jurnal Teknologi*, *46*(C), 33–48.

Norrulashikin, S. M., Yusof, F., & Kane, I. L. (2015). An Investigation towards the Suitability of Vector Autoregressive Approach on Modeling Meteorological Data. *Modern Applied Science*, *9*(11), 89–100. https://doi.org/10.5539/mas.v9n11p89

Perkins, S. E., Pitman, A. J., Holbrook, N. J., & McAneney, J. (2007). Evaluation of the AR4 Climate Models ' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability. *Journal of Climate*, *20*, 4356–4376. https://doi.org/10.1175/JCLI4253.1

Seneviratne, S. I., Nicholls, N., Easterling, D., Goodess, C. M., Kanae, S., Kossin, J., … Zhang, X. (2012). *Changes in Climate Extremes and their Impacts on the Natural Physical Environment*. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge University Press, Cambridge, UK, and New York, NY, USA.

Yusof, F., & Mean, F. H. (2012). Use of Statistical Distribution for Drought Analysis. *Applied Mathematical Sciences*, *6*(21), 1031–1051.

Zaharim, A., Najid, S. K., Razali, A. M., & Sopian, K. (2009). Analyzing Malaysian wind speed data using statistical distribution. In *4th IASME / WSEAS International Conference on ENERGY & ENVIRONMENT (EE'09) Analyzing* (pp. 363–370).