Journal Homepage : http://dergipark.gov.tr/ejens
EJENS
European Journal of Engineering and Natural Sciences
CODEN: EJENS
DergiPark AKADEMİK

Determination of Appropriate Distribution Functions for the Wind Speed Data Using the R Language

Ismail Kirbas*
[1]*Mehmet Akif Ersoy University, Department of Computer Engineering, 15100, Merkez/Burdur, Turkey.*
**Corresponding Author email: ismailkirbas@mehmetakif.edu.tr*

Abstract

Accurate determination of the proper distribution and parameters of this distribution according to the wind characteristics of the zone is vital for wind energy investment. In determining a wind energy potential belonging to a region, meteorological wind speed measurements have a great proposition to take place within a certain statistical distribution. In our study, the wind speed data obtained from the metrology station within 1 year was evaluated and it was determined using the R language, which is an open source statistical programming language, which is better suited to distributions such as Weibull, gamma, lognormal and logistic. The Akaike Information Criterion and Schwarz-Bayesian Information Criterion (SBIC) scores were calculated as the performance parameters of the distributions and the distribution performances were compared graphically. While gamma and lognormal distributions have better results at low wind speeds, Weibull distribution achieves higher performance for higher wind speeds.

Key words

gamma, logistic, lognormal, R language, statistical distribution, Weibull, wind speed

1. INTRODUCTION

The use of renewable energy sources is increasing day by day in the production of electricity all over the world. Along with that, electricity generation from wind energy is getting more and more popular every day. The greatest difficulty of generating electricity from wind energy is the complexity of the dynamics that make up the wind and it is very difficult to predict in long term period [1]–[3].

Before the wind energy investments are made, it is necessary to make wind speed measurements related to the installation area. The long-term results obtained are analysed by using statistical methods. Weibull distribution is the most frequently used method in the analysis process. However, for places with low wind speeds, the success of the Weibull distribution remains relatively low [4]–[6].

In our study, wind speed measurements made by Antalya International Airport meteorological station between 2016 and 2017 were examined. The meteorological station records 30-min average wind speed data. However, the data obtained was converted into a 6-hour average in order to make the evaluation of the data easier. All operations on the wind data were performed using the R language and Rstudio software.

R programming language was originally written by Robert Gentleman and Ross Ihaka alumni members of the statistical department of the Auckland University in New Zealand. These two statisticians were influenced syllabically by S language which is written by Chambers, Becker and Wilson and the scheme language developed by Susman. Statisticians who have developed R language, have aimed to use open source software due to the high cost of licensing of other statistical package programs and software, and high cost of learning and teaching. Later R software developers from different parts of the world gave themselves the name "r core team". R software was published by "r core team" on 29 February 2000. Software that includes the open source code feature is used free of charge. R language is a generous medium for statistical calculations and graphs. It has the feature of data processing and storing and special operators that can be used in the calculations of data and especially matrices. It contains compatible and co-usable graphical tools that can be used for data analysis [7], [8].

The source code for R, which is part of the GNU project, is available under the GNU General Public License and is available for various operating systems. There are also various graphical user interfaces, although the R command uses a line interface. It provides a language and environment for data processing, classification, clustering, time-series analysis, classical statistical tests, linear and nonlinear modelling calculations and graphic display. Because it is open source, it is very prone to development [9].

In this study, gamma, log-normal and logistic distributions are used together with Weibull distribution which is widely used in the literature in determination of wind energy efficiency and they are graphically compared with Weibull distribution. In Figure 1, 6-hour average annual wind speed data is shown graphically. The data and the R codes used in this study can be accessed freely via the GitHub platform [10].

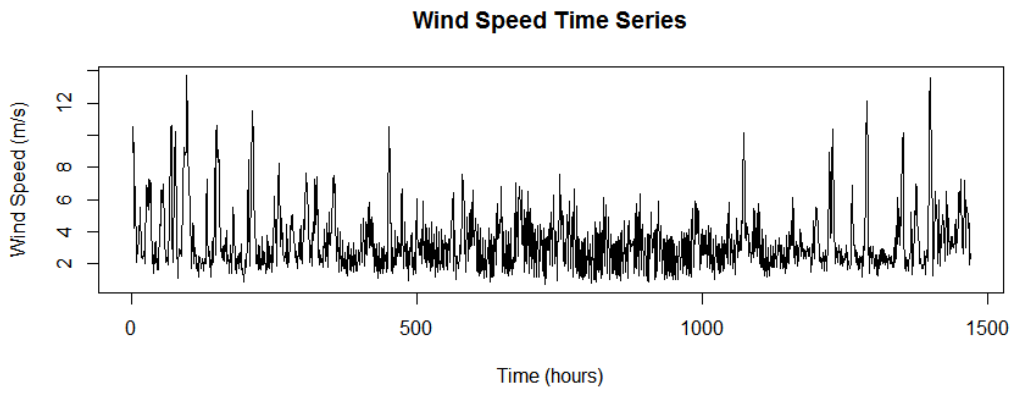

Figure 1. 6-hour average annual wind speed data from International Antalya Airport meteorological station.

Table 1 also gives some fundamental statistical information about collected data.

Table 1. Statistical results for wind speed data.

Statistical Calculation	**Value**
Average	3.3697
Standard Error Mean	0.0492
Standard Deviation	1.8858
Median	2.9200
Variance	3.5561
Skewness	1.7100
Kurtosis	3.8600
Range	12.980
Mode	2.4100
N for Mode	22
Maximum	13.720
Minimum	0.7400
Sum	4953.51
Quantity	1470

It is essential to use the correct statistical distribution model in assessing wind energy efficiency. In our study, firstly candidate models which are used in the literature have been determined in order to choose the best model. The comparative distributions were determined as Weibull, gamma, lognormal and logistic distributions. Since the Weibull distribution is the most commonly used model in wind energy analysis, the other models are compared one by one with the Weibull distribution. In the rest of our work, information about these distributions was given and mathematical and graphical comparisons were made using R language.

2. DISTRIBUTION MODELS

In this section, mathematical equations of Weibull, gamma, lognormal and logistic distributions are given and comparative performance ratios are graphically shown. In order to perform distribution analysis, fitdistrplus package is used in R language. In addition, ggplot and readr libraries are included for graphical drawings and reading data in csv format.

2.1. Weibull Distribution

The Weibull distribution, which is found by Professor Waloddi Weibull, has an important place in probability distributions. It is one of the most broadly used distributions in reliability analysis since it has the ability to characterize all the regions of the force curve. Weibull distribution is applied with two or three parameters according to the usage areas [11], [12]. Equation 1 gives Weibull probability distribution function with 3 parameters where γ is used for shape parameter, β for scale parameter and ω for location parameter.

$$f(x) = \frac{\gamma}{\beta}\left(\frac{x-\omega}{\beta}\right)^{\gamma-1} . exp\left(-\left(\frac{x-\omega}{\beta}\right)^{\gamma}\right), f(x) \geq 0, x \geq 0, \gamma > 0, \beta > 0, \omega \geq 0, \omega \leq x \leq \infty \qquad (1)$$

Two-parameter Weibull probability distribution is obtained by taking $\omega = 0$ as shown in Equation 2.

$$f(x) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} . exp\left(-\left(\frac{x}{\beta}\right)^{\gamma}\right),\ f(x) \geq 0, x \geq 0,\ \gamma > 0,\ \beta > 0 \quad (2)$$

This version of the Weibull distribution is used intensely, especially when information on wind distribution and variation in wind velocity are needed. The probability of this distribution is not symmetric but the skew is skewed and the distribution is indicated by shape and scale variables. The total likelihood of the area under this distribution equals to 1.

2.2. *Gamma Distribution*

In probability theory and statistical science, gamma distribution is a two-parameter continuous probability distribution. One of these parameters is the scale parameter θ; and the other is called the shape parameter k. If k is an integer, the gamma distribution represents the sum of random variables with k exponential distributions.

$$f(x; k, \theta) = x^{\alpha-1}\frac{e^{-x/\theta}}{\theta^k \Gamma(k)},\ x > 0,\ k > 0,\ \theta > 0 \quad (3)$$

Gamma probability distribution function is given in Equation 3 and cumulative gamma distribution function is shown in Equation 4.

$$F(x; k, \theta) = \int_0^x f(u; k, \theta)du = \frac{\gamma(k, x/\theta)}{\Gamma(k)} \quad (4)$$

Empirical and theoretical comparison for Weibull and gamma distributions is shown in Figure 2. According to histogram graph in Figure 2, gamma distribution fits data better than Weibull distribution.

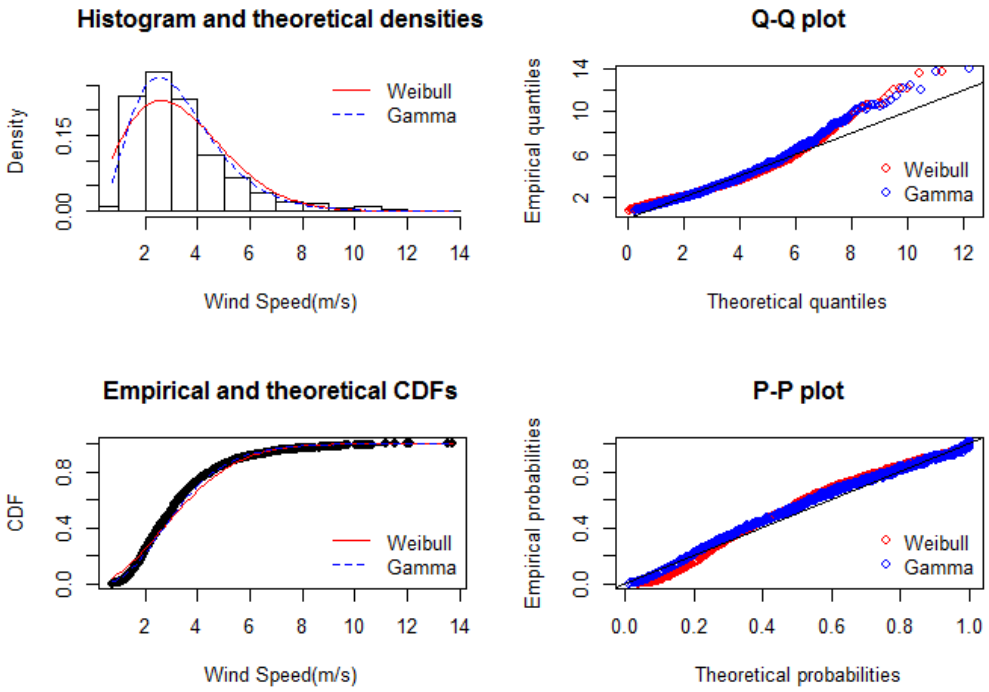

Figure 2. Empirical and theoretical comparison for Weibull and gamma distributions.

2.3. Log-Normal Distribution

Since the log-normal distribution can take many different forms, many data can be modelled by log-normal distribution. It is mainly used for economic production data and reliability analysis [13].

Log-normal probability distribution function is shown in Equation 5 and log-normal cumulative distribution function is located in Equation 6.

$$f(x) = \frac{1}{2\sigma} . \frac{1}{\sqrt{2\pi}} exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right],\ x \geq 0 \quad (5)$$

$$F(x) = P(X \leq x) = P(lnX \leq lnx) = \int_{-\infty}^{x} f(t)dt = \emptyset.\left(\frac{lnx}{\sigma\mu}\right) \quad (6)$$

Figure 3 depicts empirical and theoretical comparison for Weibull and lognormal distributions. According to histogram graph in Figure 3, lognormal distribution is more successful than Weibull distribution.

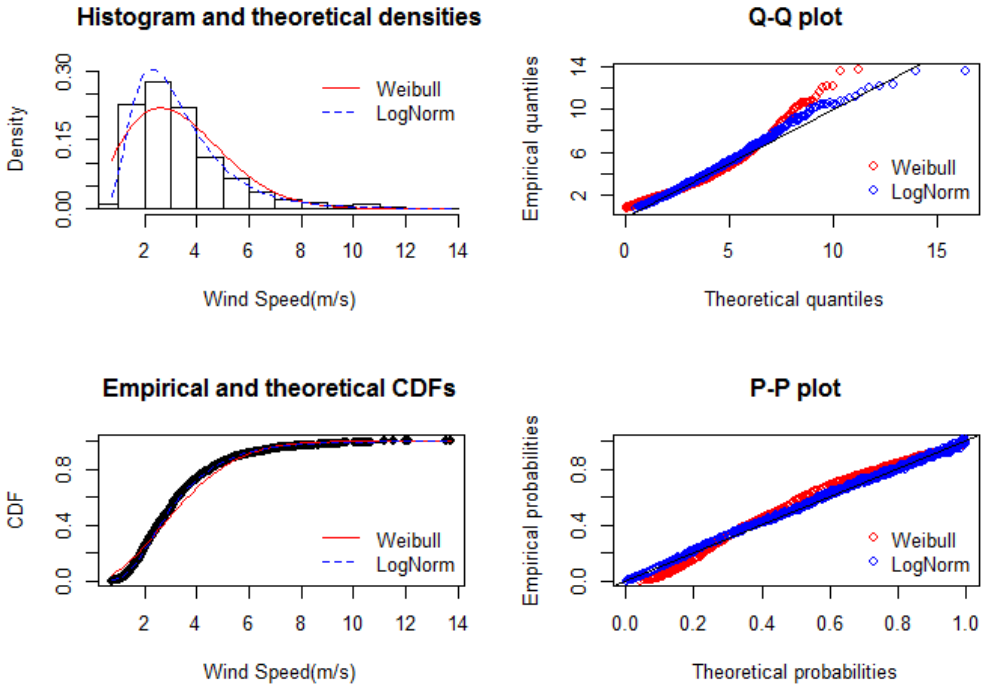

Figure 3. Empirical and theoretical comparison for Weibull and lognormal distributions.

2.4. Logistic Distribution

The logistic distribution function which is given in Equation 7 is a continuous probability distribution and also plays a role in the issues of feed-forward neural networks and logistic regression.

$$f(x;\ \mu, s) = \frac{e^{-\frac{x-\mu}{s}}}{s\left(1+e^{-\frac{x-\mu}{s}}\right)^2} = \frac{1}{4s} sech^2\left(\frac{x-\mu}{2s}\right) \quad (7)$$

Cumulative distribution function can be seen in Equation 8 and, x is the random variable, μ is the mean, and s is a scale parameter proportional to the standard deviation.

$$F(x;\mu,s) = \frac{1}{1+e^{-\frac{x-\mu}{s}}} = \frac{1}{2} + \frac{1}{2}\ tanh\left(\frac{x-\mu}{2s}\right) \quad (8)$$

Figure 4 shows theoretical and empirical comparison for Weibull and logistic distributions. Graphical results show that the Weibull distribution is more successful than the logistic distribution.

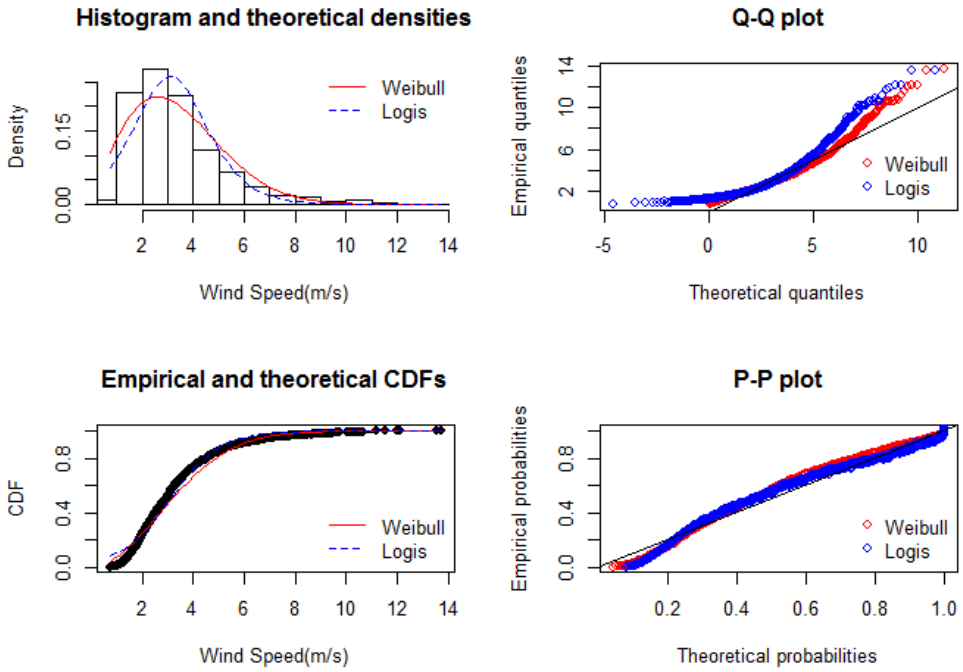

Figure 4. Empirical and theoretical comparison for Weibull and logistic distributions.

3. DISTRIBUTION EVALUATION

In literature, there are many statistical evaluation method to evaluate the developed models. In this study we choose Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, Akaike Information Criterion (AIC) and Schwarz-Bayesian Information Criterion (SBIC). Each model has been evaluated using the model evaluation methods given in Table 2 with the results obtained from the statistical model evaluations.

The Akaike information criterion (AIC) is a quality statistical relative model measure for a given set of data. When a collection of data models is given, the AIC relatively estimates each model quality. Therefore, the AIC provides a way to select the model. The Akaike criterion is based on information theory, the information given is the model data, the process is used to represent, and provides a relative estimation. Thus, the model's goodness of fit and model complexity can be understood. This criterion does not provide a model test for the null hypothesis test; If all candidate models are bad, they will not give any warning. The model with the lowest AIC value has the highest relative performance [2].

AIC value can be calculated using Equation 9 where k represents estimated number of parameters and L symbolize maximum value of the likelihood function for the model.

$$AIC = 2k - 2\ln(L) \tag{9}$$

The Bayesian Information Criterion (BIC) index imposes a penalty for increasing the number of parameters. Thus, it considers both the degree of statistical conformity and the number of parameters to be estimated. The BIC formula is given in Equation 10. Here, k denotes the number of parameters that are modelled, n represents the sample size, and finally L indicates the maximized log likelihood of the model.

$$SBIC = -2 \cdot \ln L + k \cdot \ln(n) \tag{10}$$

In Table 2, 4 different distribution models were compared according to 5 evaluation criteria and the best values were indicated as underlined.

Table 2. Distribution model evaluation results.

Metric	*Weibull*	*Gamma*	*Logistic*	*LogNormal*
Kolmogorov-Smirnov	0.0838	0.0581	0.0959	0.0292
Cramer-von Mises	3.736	1.581	2.942	0.224
Anderson-Darling	25.377	10.022	28.635	1.546
Akaike's Information Criterion	5661.043	5461.841	5841.270	5350.209
Bayesian Information Criterion	5671.629	5472.427	5851.856	5360.795

A histogram graph of 4 different distribution models is given in figure 5.

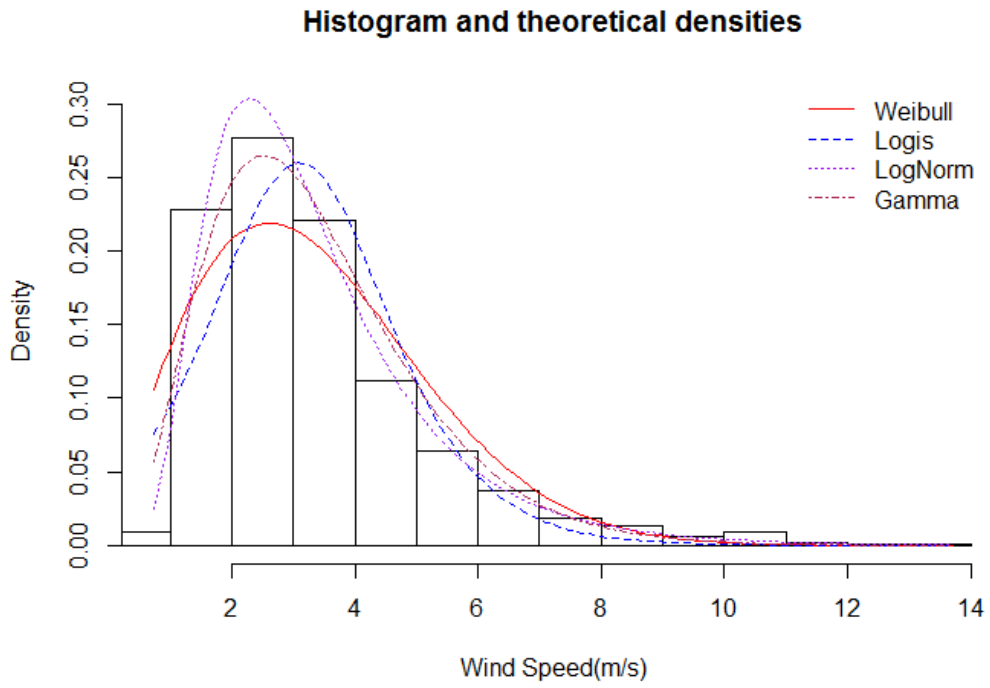

Figure 5. Graphical comparison of 4 different distribution models.

When model evaluation results in Table 2 and graphical representations in Figure 5 are evaluated together, it is seen that the most successful models for data used in the study are lognormal, gamma, Weibull and logistic distribution respectively.

4. CONCLUSION

In this study, four different distribution models were used when performing efficiency analysis before wind energy investment was made. 6-hour average wind speed data of the International Antalya Airport meteorological station was investigated as the case data. An open source software language, R, was used for statistical and graphical analysis.

The distribution models were compared according to 5 different evaluation criteria and their performance was evaluated in Table 2 and Figure 5 respectively. The results show that lognormal and gamma distributions give

better results than the Weibull distribution widely used in the literature for low wind speeds. The logistic distribution has shown the worst performance.

REFERENCES

[1] I. Kirbas and A. Kerem, "Short-Term Wind Speed Prediction Based on Artificial Neural Network Models," Meas. Control, vol. 49, no. 6, pp. 183–190, 2016.

[2] A. Kerem, I. Kirbas, and A. Saygın, "Performance Analysis of Time Series Forecasting Models for Short Term Wind Speed Prediction," presented at the International Conference on Engineering and Natural Sciences (ICENS), 2016, pp. 2733–2739.

[3] P. Bhattacharya and R. Bhattacharjee, "A Study On Weibull Dıstrıbutıon For Estimating The Parameters," J. Appl. Quant. Methods, vol. 5, no. 2, pp. 234–241, 2010.

[4] M. Kurban, Y. M. Kantar, and F. O. Hocaoğlu, "Weibull Dağılımı Kullanılarak Rüzgar Hız ve Güç Yoğunluklarının İstatistiksel Analizi," Afyon Kocatepe Univ. J. Sci., vol. 7, no. 2, pp. 205–218.

[5] W.-Y. Chang, "A Literature Review of Wind Forecasting Methods," J. Power Energy Eng., vol. 2, no. 4, pp. 161–168, 2014.

[6] T. P. Chang, "Estimation of wind energy potential using different probability density functions," Appl. Energy, vol. 88, no. 5, pp. 1848–1856, 2011.

[7] A. F. Özdemir, E. Yıldıztepe, and M. Binar, "İstatistiksel Yazılım Geliştirme Ortamı: R," presented at the XII. Akademik Bilişim Konferansı, Muğla, 2010, vol. 1, pp. 375–379.

[8] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing, 2014.

[9] P. Dalgaard, Introductory statistics with R. Springer Science & Business Media, 2008.

[10] İ. Kırbaş, "Wind Speed Distribution Dataset and R Source Codes," GitHub Page, 23-Apr-2017. [Online]. Available: https://github.com/ismkir/windSpeedDistribution/. [Accessed: 24-Apr-2017].

[11] O. Elitok, "Weibull Distributions and Its Applications," M. Sc. Thesis, Kırıkkale University Institute of Science and Technology, Kırıkkale, 2006.

[12] D. Indhumathy, C. V. Seshaiah, and K. Sukkiramathi, "Estimation of Weibull Parameters for Wind speed calculation at Kanyakumari in India," Int. J. Innov. Res. Sci. Eng. Technol., vol. 3, no. 1, pp. 8340–8345, Jan. 2014.

[13] A.-A. Bromideh, "Discriminating Between Weibull and Log-Normal Distributions Based on Kullback-Leibler Divergence," Istanb. Ünversitesi İktisat Fakültesi Ekonom. Ve Istat. Derg., no. 16, pp. 44–54, 2012.