21th International Conference on Renewable Energies and Power Quality (ICREPQ'23)
Madrid (Spain), 24th to 26th May 2023
Renewable Energy and Power Quality Journal (RE&PQJ)
ISSN 2172-038 X, Volume No.21, July 2023

Modelling Wind Turbine Power Curves based on Frank's copula

M.A. Garcia-Vaca[1], J.E. Sierra-García[2], M. Santos[3] and R. Pandit[4]

[1] Computer Science Faculty
Complutense University of Madrid
28040-Madrid, (Spain)
e-mail: magvaca@ucm.es

[2] Department of Electromechanical Engineering
University of Burgos
E-mail: jesierra@ubu.es

[3] Institute of Knowledge Technology,
Complutense University of Madrid
E-mail: msantos@ucm.es

[4] Centre for Life-cycle Engineering and Management (CLEM)
Cranfield University
E-mail: ravi.pandit@cranfield.ac.uk

Abstract. In the study of wind turbines, one of the most relevant and useful indicators is the power curve. It has been shown to be of paramount importance in evaluating turbine performance and therefore reducing operation and maintenance (O&M) costs. Various techniques can be applied to model and obtain the shape of this curve, which relates the electrical power generated by a turbine to the wind speed. Statistical copulas are used in this paper, a tool used in other fields such as econometrics, and whose potential lies in its ability to capture the complex dependency between the variables involved. In particular, the Frank copula is applied to obtain a probabilistic model of the power curve of a wind turbine. This model is compared with the Gaussian Mixture Model, a technique widely used to obtain parametric probabilistic models. As a result of this comparison, it is observed that the Frank copula model fits the power curve of the wind turbine with greater precision and reliability, which would allow its use for prediction and fault detection.

Keywords. Wind turbine, O&M, power curve, probabilistic model, statistical copulas, Frank's copula.

1. Introduction

The energy sector is considered one of the most important strategic sectors of any country, and therefore, it has a notable projection for the future. In this sense, during the last decades significant efforts have been made to make a gradual transition from the different traditional energy sources towards more sustainable and environmentally cleaner options, as well as to provide them with technologies that improve their efficiency [1-2].

Among these energy sources, wind turbines (WT) stand out, both offshore and on-land. However, in order to make electricity generation profitable with these infrastructures, it is necessary to provide them with tools that allow them to maximize their performance and detect failures, thus reducing the costs associated with their operation and maintenance (O&M).

It has been observed that one of the best indicators to evaluate the operating conditions of a turbine is its power curve [3], which relates the electrical power generated based on the wind available at that moment. It is a unique function for each specific turbine. However, modelling this curve is not a simple task since its shape is not only non-linear and complex, but also depends to a large extent on the environmental conditions of the site and the methods used to acquire the signals. Therefore, a successful approach is the use of data-driven algorithms [4-5].

Different techniques have been used to model power curves, and some surveys have been published on this topic [6-8]. To mention a few examples, Pelletier et al. [9] use a multilayer perceptron artificial neural network to model the power curve in a wind farm with 140 wind turbines. Rogers et al., in [10] proposed the use of a heteroscedastic Gaussian process model for modelling. This allows it the elimination of the need to specify a parametric functional form for the power curve and the automatic quantization of the variance in the prediction. Kusiak et al. [11], using a genetic algorithm, compares power curves with a logistic function of 4 parameters.

One of these tools are the statistical copulas. Specifically, the one proposed here, Frank's copulas, have been

exploited in various energy-related applications. To cite some works, in [12] they are used for the stochastic planning of an integrated energy system. Singh et al. [13] derive intensity-duration-frequency (IDF) curves from the bivariate analysis of rainfall frequency using this copula. In [14], they examine the development of copula models and their applications in the areas of energy, fuel cells, forestry, and environmental sciences.

In this work we will focus on the application of Frank's copulas to find a parametric probabilistic model of the power curve of a wind turbine. It will be used to estimate the expected value of the generated power, as well as its associated uncertainty. In order to compare the suitability of the proposed technique, it will be compared with one of the most widespread methods: the Gaussian mixture model.

The structure of the rest of the paper is as follows. Section 2 describes the methodology applied and the metrics used to evaluate the models. In Section 3, the data used and the pre-processing carried out on the data are presented. Section 4 is dedicated to the discussion of the results. The article ends with the conclusions and future research lines.

2. Methodology

A. Gaussian mixture model

Finite mixture models are used to describe the probability density function of a population consisting of several clusters with different underlying distributions. As the name suggests, it is assumed that the total population is a finite mixture of several independent components or modes. For continuous variables, we can assume that all clusters are modelled as Gaussian distributions, each with different parameters (mean and covariance). This technique is known as a Gaussian Mixture Model (GMM) [15-16].

For the bivariate case, let $\psi(x1, x2; \Theta)$ be the required probability density function, where Θ is the set of parameters, mean and covariance, of these Gaussian distributions. Equation (1) holds:

$$\boldsymbol{\psi(x_1, x_2; \Theta) = \sum_{k=1}^{M} \alpha^{(k)} \phi\left(x_1, x_2; \theta^{(k)}\right)} \qquad (1)$$

where $\boldsymbol{\alpha^{(k)}}$is the mixing proportion of the M clusters (the sum of all alphas will be equal to 1) and $\boldsymbol{\theta^{(k)}}$ is the matrix of mean and covariance of each k with $\boldsymbol{\phi\left(x_1, x_2; \theta^{(k)}\right)}$ the probability density of cluster *k* [16].

For the necessary calculations, MATLAB software has been used, which includes the pre-defined fitting function *fitgmdist*, which calculates the parameters of the model obtained in the element given by function *gmdistribution*.

B. Copula models: Frank copula

Statistical copulas are families of mathematical functions capable of relating dependent variables with a complex correlation between them. On the one hand, they help us to separate the marginal distributions and, on the other hand, they give information about how they are related to each other. According to Sklar's theorem [17], which establishes the pillars of copula theory, the probability density function $\boldsymbol{f(x_1, x_2)}$ for a bivariate case can be decomposed according to expression (2):

$$\boldsymbol{f(x_1, x_2) = c(u_1, u_2) f_1(x_1) f_2(x_2)} \qquad (2)$$

Where $\mathbf{f_i}$ denotes the marginal distribution function of the *i*-th variable ($\boldsymbol{x_i}$), $\boldsymbol{u_i}$ denotes the cumulative distribution function of these variables, and $\boldsymbol{c}: [\mathbf{0}, \mathbf{1}]^{\mathbf{2}} \rightarrow \mathbb{R}$ is the copula function that relates these cumulative density functions.

There are many families of parametric copulas such as Clayton, Frank, or Gumbel, among others [18]. The selection of one or another depends on the data and level of tail dependence presented. High tail dependence is equivalent to a narrowing of the scatter of the observed data around its extremes. In the case of power curve of a wind turbine, it presents a strong correlation in both extreme values of the distribution (both for high and low values). Therefore, a good choice will be the Frank copula, a type of Archimedean copula. From [19], these Frank copula functions respond to equation (3):

$$\boldsymbol{c_{Frank}(u_1, u_2, \delta) = \frac{\delta \eta e^{-\delta(u_1+u_2)}}{[\eta-(1-e^{-\delta u_1})(1-e^{-\delta u_2})]^2}} \qquad (3)$$

where $\boldsymbol{\eta = 1 - e^{\delta}}$ and $\boldsymbol{\delta}$ is the copula parameter that best fits the data. The higher this value is, the more pronounced the dependence between the two variables.

C. Evaluation metrics

In order to quantify the goodness of the fitting of the Frank copula model to the power curve, we calculate the Bayesian information criterion (BIC) and the averaged normalized root mean squared error (NRMSE) over the mean.

The BIC is commonly used to select the most appropriate parametric model from a finite range of models [20]. It is calculated from the maximum likelihood function value obtained by the considered model. For the present case, as we work with continuous variables, the value of this function is equal to the sum of the posterior probability density function value given the model parameters (θ) at every point of the dataset. That is, the BIC is given by equation (4)

$$BIC = -2 \sum_{i=1}^{N} \ln\ \mathrm{p}(x_1(i), x_2(i) | \theta) + p \cdot \ln(N) \qquad (4)$$

where N represents the total number of points and p is the number of parameters. For the case of the Frank copula, $p = 1$, since the only parameter of the model in equation (3) is δ. For the GMM, the number of parameters according to [21] is given by equation (5):

$$p = m(1 + d + \tfrac{d(d+1)}{2}) \qquad (5)$$

where m denotes the number of modes or clusters (in our case 3) and d is the data dimensions (in our case 2). That is, for the GMM, $p = 18$.

According to equation (5), it can be observed that it includes a term $p \cdot ln(N)$ that penalizes overly complex models which could lead to overfitting or loss of generalization The lower the value of BIC, the better the model considered.

The other evaluation metric used, NRMSE, provides the total error of the model, i.e., the difference between the observed and the estimated values. This quantity is calculated according to equation (6):

$$NRMSE = \frac{1}{\overline{y}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n}} \qquad (6)$$

where $\tilde{y}_i$ is the estimated value of the model for each point of the dataset, y_i is the observed value of each dataset point, $\overline{y}$ is the mean of the observed data, and n is the total number of points of the model. Like the BIC, the lower this value is, the better the model performs.

3. Dataset and pre-processing

The dataset used is available in Kaggle [22]. Data have been obtained from an onshore wind turbine located at the Yalova wind farm (Turkey). The wind turbine is a blade pitch regulated Nordex N117 with rated power of 3,6 MW. The main features are: cut-in wind speed of 3 m/s, cut-off wind speed of 25 m/s, and rated wind speed of 13 m/s. It has three blades with a total diameter of 117m.

The dataset consists essentially of two variables acquired through a SCADA system with an average acquisition time of 10 minutes over a period of one year. These variables are: generated electrical power (kW) and wind speed (m/s). For the case of study, data from August and September of 2018 are used, with a total of 8425 points. In figure 3 it can be seen the raw dataset:

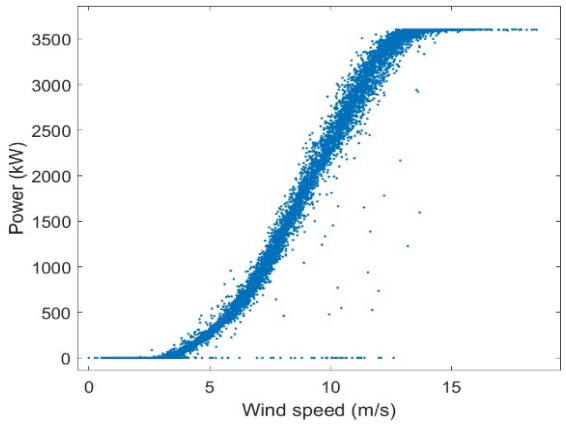

Fig.3 Raw dataset

A. *Data pre-processing*

Before proceeding with power curve modelling, data must be cleaned from anomalous values. This cleaning is carried out in two phases:

1) Rejection of data with zero or negative power, as well as data with indeterminate values.
2) Removal of outliers. These outliers are characterized by being isolated and because they do not fit in any prior pattern. Their nature is due to the fact that since the data are acquired with an average of 10 minutes, it is possible that within that time interval the turbine has changed its operating state (from on to off, or vice versa), causing an erroneous average power value. To remove these points, the power curve is divided into intervals of 0.5 m/s and its mean and standard deviation are calculated, as specified by the corresponding international standard, IEC 61400-12-1:2017 [23]. Any point that deviates ±3σ from its corresponding interval is considered an anomalous point and will be rejected.

Figure 4 shows the dataset after performing the first phase of the pre-processing (rejection of data with zero or negative values). The red line represents the mean of each interval, and the error bars represent 3σ on both sides of the mean. All points higher or lower than these bars are eliminated.

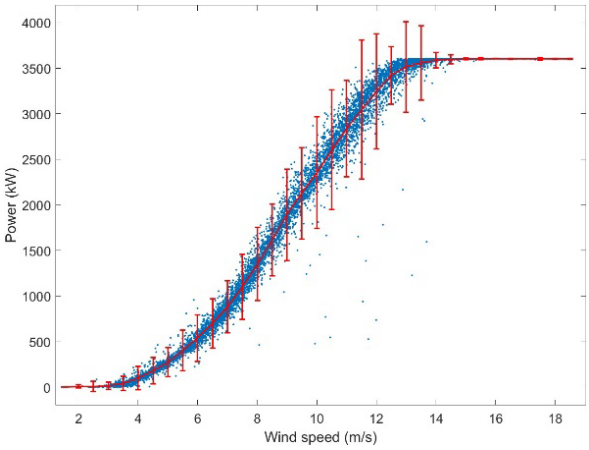

Fig. 4 Dataset after the first phase of pre-processing and its corresponding IEC curve

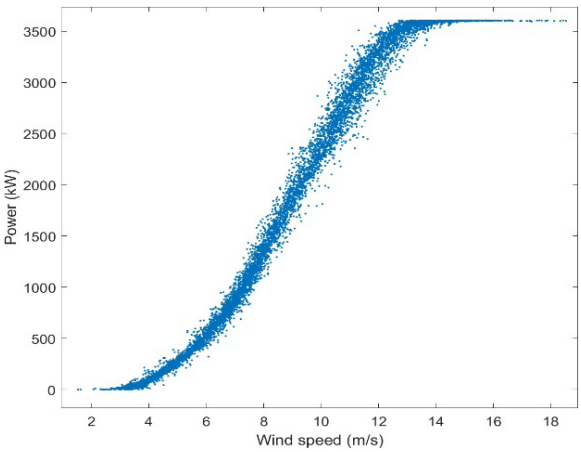

Fig. 5. Dataset used in the models

In Figure 5, the cleaned dataset after second phase of pre-processing is shown. It will be used to evaluate the

proposed models. A total of 7570 points are used in the models.

4. Discussion of the Results

For the selection of Frank's copula parameter δ (equation 5), first it is necessary to calculate the marginal distributions of the two variables, f_i, of the dataset. To do this, we use a non-parametric distribution fitting with kernel smoothing function estimation. Marginal histogram of power and wind speed are shown in figures 1 and 2, respectively. The red line indicates the fitting of the curve. The applied bandwidths are 0.32 and 7 respectively.

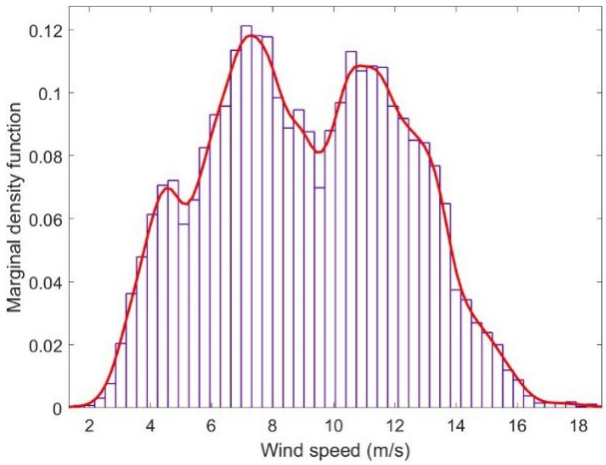

Fig. 1 Marginal distribution of wind speed

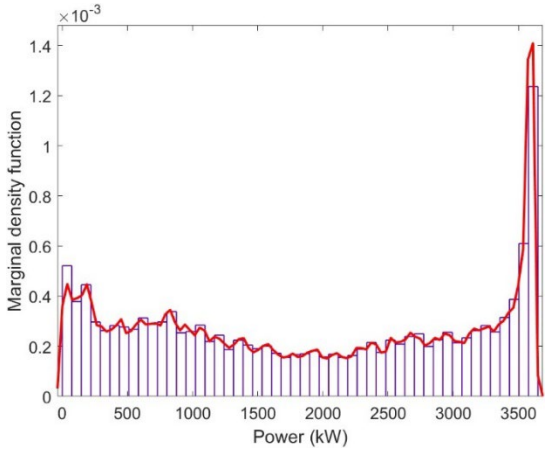

Fig 2. Marginal distribution of power.

Using these marginal distribution functions, we are now able to estimate which Frank's copula best fits the data via maximum likelihood method. As a result, a value of $\boldsymbol{\delta = 70}$ is obtained. Figure 6 shows the transformed power curve in the copula space.

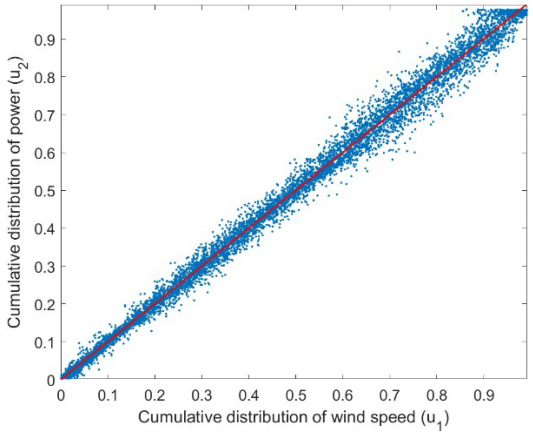

Fig. 6. Transformation of the power curve into the copula space

The higher the value of δ, the greater the dependence between the variables. Or in other words, the data will be more concentrated around the diagonal (red line in figure 6). Therefore, a high correlation in the variables is observed and proves the right choice of the Frank copula to model the power curve.

Figures (7a) and (7b) show the models obtained with the GMM and Frank's copula, respectively. The joint probability density of the power curve overlaid with the dataset is plotted.

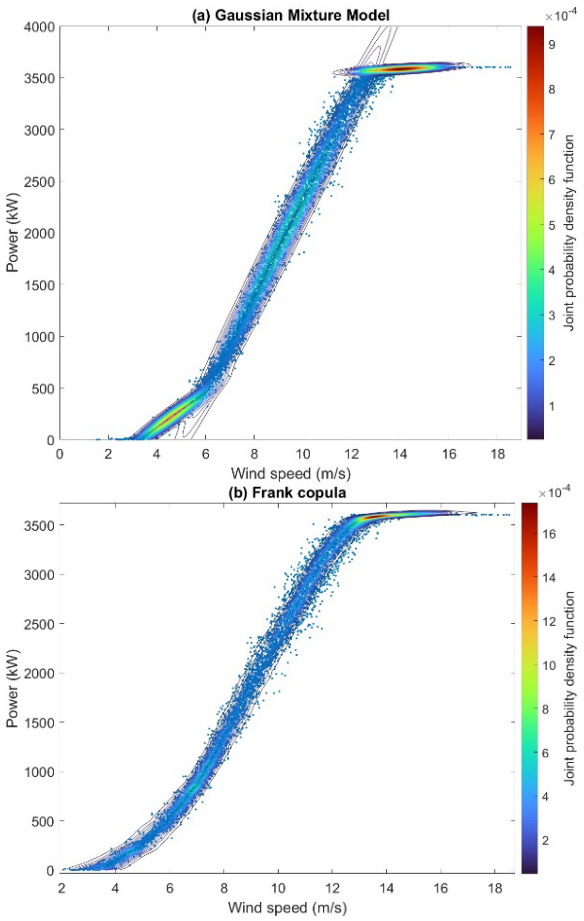

Fig. 7. Results of the models: a) GMM, b) Frank's copula

Visually, it can be seen how the copula method fits the power curve better than the GMM, which is not able to capture its shape well because it does not find a good

representation of the three operational modes of a wind turbine.

Table I shows a summary of the evaluation metrics obtained for the two models.

Table I. – Evaluation metrics

	BIC	***NRMSE***
Frank's Cópula	125500	0.084
GMM	129880	0.093

Regarding the NRMSE, an improvement of the Frank's copula over the GMM is observed. If we consider the BIC, a significant difference in the performance of the Frank's copula method is also obtained.

5. Conclusion and future work lines

In conclusion, we have seen how it is possible to apply statistical copulas to the modelling of a wind turbine power curve and how they offer us the ability to faithfully capture the nonlinear and complex nature of this curve.

As future works, other variables could be included to further improve the model, such as rotor speed, pitch blade or ambient temperature. In addition, this model could be applied to evaluate different types of turbine failures and check if it is possible to detect when a loss of turbine performance occurs.

Acknowledgement

This work has been partially supported by the Spanish Ministry of Science and Innovation under the project MCI/AEI/FEDER number PID2021-123543OB-C21.

References

[1] Strategic Energy Technology Information System (SETIS) https://setis.ec.europa.eu/index_en
[2] Sierra-García, J. E., & Santos, M. (2021). Redes neuronales y aprendizaje por refuerzo en el control de turbinas eólicas. Revista Iberoamericana de Automática e Informática industrial, 18(4), 327-335.
[3] O. Uluyol, G. Parthasarathy, W. Foslien, and K. Kim, "Power curve analytic for wind turbine performance monitoring and prognostics", Annual Conference of the PHM Society, vol. 3, no. 1, September 2011
[4] H. Long, L. Wang, Z. Zhang, Z. Song, and J. Xu, "Data-driven wind turbine power generation performance monitoring", IEEE Transactions on Industrial Electronics, vol. 62, no. 10, pp. 6627-6635, October 2015.
[5] Sierra-García, J. E., Santos, M., & Pandit, R. (2022). Wind turbine pitch reinforcement learning control improved by PID regulator and learning observer. Engineering Applications of Artificial Intelligence, 111, 104769.
[6] Y. Wang, Q. Hu, L. Li, A. M. Foley, and D. Srinivasan, "Approaches to wind power curve modeling: A review and discussion", Renewable and Sustainable Energy Reviews, vol. 116, p. 109422, December 2019.
[7] V. Sohoni, S.C. Gupta, and R.K. Nema, "A critical review on wind turbine power curve modelling techniques and their applications in wind based energy systems", Journal of Energy, July 2016.
[8] M. Lydia, S.S. Kumar, A.I. Selvakumar, and G.E.P. Kumar, "A comprehensive review on wind turbine power curve modeling techniques", Renewable and Sustainable Energy Reviews, vol. 30, pp. 452-460. February 2014
[9] F. Pelletier, C. Masson, and A. Tahan, "Wind turbine power curve modelling using artificial neural network", Renewable Energy, vol. 89, pp. 207-214, April 2016.
[10] T.J. Rogers, P. Gardner, N. Dervilis, K. Worden, A.E. Maguire, E. Papatheou, and E.J. Cross, "Probabilistic modelling of wind turbine power curves with application of heteroscedastic Gaussian process regression", Renewable Energy, vol. 148, pp. 1124-1136, April 2020.
[11] A. Kusiak, H. Zheng, and Z. Song, "Models for monitoring wind farm power", Renewable Energy, vol. 34, no 3, pp. 583-590, March 2009.
[12] S. Lin, C. Liu, Y. Shen, F. Li, D. Li, and Y. Fu, "Stochastic planning of integrated energy system via Frank-Copula function and scenario reduction", IEEE Transactions on Smart Grid, vol. 13, no 1, pp. 202-212, January 2022.
[13] V.P. Singh, and L. Zhang, "IDF curves using the Frank Archimedean copula", Journal of hydrologic engineering, vol. 12, no 6, pp. 651-662, November 2007.
[14] M.I. Bhatti, and H.Q. Do, "Recent development in copula and its applications to the energy, forestry and environmental sciences", International Journal of Hydrogen Energy, vol. 44, no 36, pp. 19453-19473, July 2019.
[15] T. Huang, H. Peng, and K. Zhang, "Model selection for Gaussian mixture models", Statistica Sinica, vol. 27, no. 1, pp. 147–69, January 2017.
[16] G. McLachlan, and D. Peel, Finite Mixture Models, Ed. John Wiley & Sons, pp.420-427, 2000.
[17] M. Sklar "Fonctions de repartition an dimensions et leurs marges", Publ. inst. statist. univ. Paris, vol. 8, pp. 229-231, 1959.
[18] R.B. Nelsen, An introduction to copulas, Ed. Springer Science & Business Media, 2007.
[19] H. Joe, Multivariate models and multivariate dependence concepts. Ed. CRC press, 1997.
[20] G. Schwarz, "Estimating the dimension of a model", The annals of statistics, vol. 6, no. 2, pp. 461-464, March 1978.
[21] W. D. Penny, "Variational Bayes for d-dimensional Gaussian mixture models", University College London, January 2001.
[22] https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset
[23] International Electrotechnical Commission, "Wind energy generation systems—Part 12-1: Power performance measurements of electricity producing wind turbines", International Electrotechnical Commission (IEC), IEC Central Office, vol. 3, pp. 2017-03, 2017