International Journal of Energy, Information and Communications
Vol.5, Issue 3 (2014), pp.23-34
http://dx.doi.org/10.14257/ijeic.2014.5.3.03

Wavelet Analysis Based Estimation of Probability Density function of Wind Data

Debanshee Datta

Department of Mechanical Engineering
Indian Institute of Technology
Banaras Hindu University
Varanasi - 221005
debansheedatta@yahoo.co.in

Abstract

Knowledge of the density as a source distribution from where the data points are sampled, either through measurements or through some simulations is essential in order to explore the further physics or any other associated rule. A popular approach of knowing the density of data points is regression analysis. However, regression analysis being a parametric method is biased and hence an unbiased method such as the non-parametric approach is looked for the estimation of density. Wavelet approach being non-parametric, the estimation of density is carried out using wavelet analysis of data points. Wavelet based density estimation is applied to construct the probability density of the wind speed data and it is proposed that wavelet based construction of shape function of wind data (wind speed) through density estimation is compassed for the first time. Results are compared with the shape function of Weibull fitted density of the same data and comparison is carried out on the basis of entropy evaluation of both methods. The paper presents in detail the method of density estimation using wavelets and evaluation of information entropy (Shannon entropy) for comparison. The complete density function of the wind data is constructed using wavelet 'sym15' at level 5.

Keywords: *Wavelet, Weibull distribution, density estimation, information entropy*

1. Introduction

Traditional statistical distribution fitting plays an important role in data interpretation. Over the past decade data mining, or knowledge discovery in databases (KDD), has become a significant area both in academia and industry. Data mining is a process of automatic extraction of novel, useful and understandable patterns from a large collection of data. Wavelet theory plays an important role in data mining because wavelets are localized and non-parametric. Wavelets have many favourable properties, such as vanishing moments, hierarchical and multi-resolution decomposition structure, linear time and space complexity of the transformations, decorrelated coefficients, and a wide variety of basis functions [1, 2]. These properties could provide considerably more efficient and effective solutions to various types of data analysis problems. First, wavelets could provide presentations of data that make the mining process more efficient and accurate. Second, wavelets could be incorporated into the kernel of many data mining algorithms. Although standard wavelet applications are mainly on data that have temporal/spatial localities (*e.g.*, time series, stream data and image data). Wavelets have also been successfully applied to diverse domains in data mining. In practice, a wide variety of wavelet-related methods have been applied to a wide range of data mining problems.

Experimental data are generally sampled from an unknown distribution, whose density is not known. Again, in any experiment, generation of data of substantial large (sample size) is not possible and hence knowledge of the density function from where the specified data is sampled is not possible to be known. However, data analyst attempts to constructs an empirical fit by using standard method of regression which is based on ordinary least squares technique. Regression method is parametric and mostly based on normal distribution. Therefore, this method if applied to construct the density function estimation will be heavily biased and will certainly be error prone. On the contrary, wavelet based method is non -parametric and is thus unbiased. Therefore, density estimation of data set of small sample size using wavelet may provide a better approach as compared to the traditional regression analysis. Wavelet transform is generally applied for a time series to extract the time and frequency information of the data set, whereas Fourier transform can give only the frequency information of the same time series. Time-series analysis by using wavelet transform can be found in detail elsewhere in [3]. Wavelet analysis also has been used in data mining by many researchers [4]. In data mining, the basic objective is to extract the hidden information [5] and construct the associated rules [5]. However, estimation of the density function of a data set in wavelet domain has not been attempted, though theoretically the mathematics of wavelet has already proved this possibility. In view of this history behind the successful application of wavelet in the domain of data analysis, here in this paper an en masse different approach of density estimation using wavelet has been presented. The remaining part of the paper has been organized in this way that, Section 2 presents the mathematical outline of wavelet and its generic properties, Section 3 presents the wavelet based methodology of density estimation, Section 4 presents the results of the density estimation of a typical wind data set and a comparison with standard Weibull probability density function fitted through the same data set and Section 5 draws a conclusion by highlighting the merits and demerits of this new methodology of density estimation. The estimated probability density function provides the knowledge of the uncertainty bounds (5th and 95th percentile) of the wind data which one can then apply to obtain the uncertainty bounds of return period. Return periods are generally applicable for assessing the safety in the design of any civil structure which faces the extreme climatic conditions such as wind load or oceanic turbulence. The results presented in this paper are generated by executing an in-house developed MATLAB code “wavwblden (version 1.0)”.

2. Mathematical Details of Wavelet

A wavelet can own many attractable properties, including the essential properties such as compact support, vanishing moments and dilating relation and other preferred properties such as smoothness and a generator of an orthonormal basis of function spaces, $L^2(R^n)$. In short, compact support guarantees the localization of wavelets; vanishing moment guarantees that wavelet processing can draw a remarkable line between the essential information and non-essential information; and dilating relation leads fast wavelet algorithms. The other properties such as smoothness and generators of orthonormal basis are preferred rather than essential. For example, Haar wavelet is the simplest wavelet which is discontinuous, while all other Daubechies wavelets are continuous. Further details of wavelets can be found elsewhere in [6]. However, in his section mathematical aspect of wavelet from its usage in data analysis is presented. For example, density estimation is one of the themes of the functional analysis of wavelets generally considered for data analysis.

2.1. Basic of Wavelet in $L^2(R)$

What is wavelet? A mother wavelet is a function $\psi(x)$ such that $\{\psi(2^j x - k), i, k \in Z\}$ is an orthonormal basis of $L^2(R)$. The basis functions are usually referred as wavelets. The term wavelet means a small wave. Compared to Fourier transform, wavelet transform provide time and frequency localizations simultaneously. Fourier transforms are designed for stationary signals because they are expanded as sine and cosine waves which extend in time forever, if the representation has certain frequency content at one time, it will have the same content for all time. Hence Fourier transform is not suitable for non-stationary signal where the signal has time varying frequency. Since FT doesn't work for non-stationary signal, researchers have developed a revised version of Fourier transform, so called as the Short Time Fourier Transform (STFT) [6]. In STFT, the signal is divided into small segments where the signal on each of these segments could be assumed as stationary. Although STFT could provide a time-frequency representation of the signal, Heisenberg's Uncertainty Principle [7] makes the choice of the segment length of a big problem for STFT: The principle states that one cannot know the exact time-frequency representation of a signal and one can only know the time intervals in which certain bands of frequencies exist. So for STFT, longer length of the segments gives improved frequency resolution and indigent time resolution while shorter segments lead to exactly opposite features. Another serious problem with STFT is that there is no inverse, *i.e.*, the original signal can not be reconstructed from the time-frequency map or the spectrogram.

Wavelet is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. This is useful for many practical signals since they usually have high frequency components for short durations (bursts) and low frequency components for long durations (trends). The signal processing application of wavelet is not the issue of the theme of this paper. Hence the fundamentals of the wavelet has been focused towards the other aspects, strictly speaking, how to estimate the probability density function of the data set having small sample size with the minimum information about the same. Wavelet is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. In wavelet based data analysis practice, the key concept behind the use of wavelets is the discrete wavelet transform (DWT) [8]. Therefore, it is required to introduce a brief discussion on DWT.

2.1.1. Dilation Equation: *Mathematically, for a function $\phi(x)$ the dilation equation is written as*

$$\phi(x) = \sum_{k=-\infty}^{\infty} a_k \phi(2x-k) \tag{1}$$

where a_k's are called the filter coefficients. The function $\phi(x)$ is called the scaling function. Under certain conditions, equation (2) presents the mathematical structure of a wavelet.

$$\psi(x) = \sum_{k=-\infty}^{\infty} (-1)^k b_k \phi(2x-k) = \sum_{k=-\infty}^{\infty} (-1)^k a_{1-k} \phi(2x-k) \tag{2}$$

Figure 1 shows the Haar wavelet and Figure 2 shows the Daubechies-2(db_2) wavelet which is supported on intervals [0, 3]. In general, db_n represents the family of Daubechies and n is the order. Generally, it can be shown that the support for db_n is on the interval [0, 2n-1], the wavelet db_n has n vanishing moments and the regularity increases with the order.

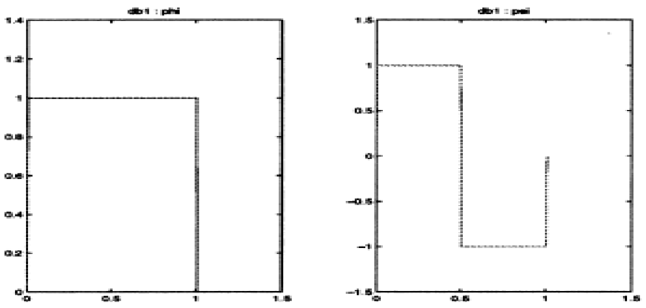

Figure 1.0. Haar Wavelet

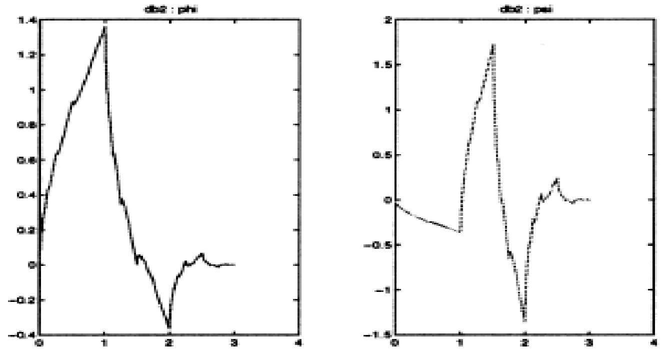

Figure 2.0. Daubechies – 2(db2) Wavelet

2.1.2. Multiresolution Analysis (MRA): The motivation of MRA is to use a sequence of embedded subspaces to approximate L^2 (R) so that a proper subspace for a specific application task can be chosen to get a balance between accuracy and efficiency. Mathematically, MRA studies the property of a sequence of closed subspaces. A direct application of multiresolution analysis is the fast discrete wavelet transform algorithm, called the **pyramid** algorithm [9]. The core idea is to progressively smooth the data using an iterative procedure and keep the detail along the way, i.e., analyze projections of **f** to W_j.

2.2. Properties of Wavelets

The analytical properties of wavelets which make them useful tools for data analysis and many other applications are as follows:

(i) **Computational Complexity:** First, the computation of wavelet transform can be very efficient. Discrete Fourier transform (DFT) requires $O(N^2)$ multiplications and fast Fourier transform also needs O(N log N) multiplications. However fast wavelet transform (based on Mallat's pyramidal algorithm) only needs O(N) multiplications. The space complexity is also linear.

(ii) Vanishing Moments: Another important property of wavelets is vanishing moments. A function f *(x)* which is supported in bounded region ω is called to have n-vanishing moments if it satisfies

$$\int_{\omega} f(x) x^{j} dx, \text{ j = 0, 1, 2,...,n} \qquad (3)$$

As an example it is know that, Haar wavelet has 1-vanishing moment and db_2 has 2-vanishing moment. The intuition of vanishing moments of wavelets is the oscillatory nature which can think to be the characterization of difference or details between a datum with the data in its neighborhood.

(iii) **Compact Support:** Each wavelet basis function is supported on a finite interval. Compact support guarantees the localization of wavelets. In other words, processing a region of data with wavelet does not affect the data out of this region.

(iv) **Uncorrelated Coefficients:** Another important aspect of wavelets is their ability to reduce temporal correlation so that the correlations of wavelet coefficients are much smaller than the correlation of the corresponding temporal process [9]. Hence, the wavelet transform will reduce the complex process in the time domain into a much simpler process in the wavelet domain.

(v) **Parseval's Theorem:** the energy, which is defined to be the square of its $\boldsymbol{L_2}$ norm, is preserved under the orthonormal wavelet transform.

3. Density Estimation using Wavelet

The basic computational scheme for the density estimation of a data set of small sample size is presented in this section. Consider estimating a density function f from independently identically distributed data {x_i | i =1, 2, 3...,n} collected or measured experimentally using wavelets. The formal wavelet expansion of the density function following [10] can be written as:

$$f(x) = \sum_{k\in Z} c_{Lk}\phi_{Lk}(x) + \sum_{j\in J_L}\sum_{k\in Z} d_{jk}\psi_{jk}(x) \tag{4}$$

where $\mathbf{J_L}$ = {m ∈ Z; m ≥ L} and the coefficients c_{Lk} and d_{jk} are given by

$$c_{Lk} = \int_{-\infty}^{\infty}\phi_{Lk}(x)\, f(x)dx \quad \text{and} \quad d_{jk} = \int_{-\infty}^{\infty}\psi_{jk}(x)\, f(x)dx$$

The wavelet basis functions at resolution level j are given by

$$\phi_{jk}(x) = 2^{j/2}\phi\left(2^{j}x - k\right)$$

$$\psi_{jk}(x) = 2^{j/2}\psi\left(2^{j}x - k\right) \tag{5}$$

The primary resolution level L determines the scale of the largest effects that can be affected by the smoothing inherent in the procedure. Now, the true coefficients c_{Lk} and d_{jk} as expressed by integral expression can be further written as

$$c_{jk} = E[\phi_{jk}(x)] \text{ and } d_{jk} = E[\psi_{jk}(x)] \tag{6}$$

where the estimators are:

$$\tilde{c}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\phi_{jk}(x_i) \tag{7}$$

$$\tilde{d}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\psi_{jk}(x_i) \tag{8}$$

These empirical coefficients are calculated for resolution levels L up to some large value J, called the finest resolution level [10]. The coefficients $\tilde{d}_{jk}$ are then threshold to compute the estimated coefficients $\hat{d}_{jk}$. The resulting density estimate can be written now as

$$\hat{f}(x) = \sum_{k} \tilde{c}_{Lk}\phi_{Lk}(x) + \sum_{j=L}^{J-1}\sum_{k} \hat{d}_{jk}\psi_{jk}(x) \tag{9}$$

It can be easily noted that n is not required to be a power of two and this theory is in principle also valid for higher dimensional data.

3.1. Algorithm used for Wavelet-Based Density Estimation

As in the regression context, the wavelets are useful in a nonparametric context, when very little information is available concerning the shape of the unknown density, or when you don't want to tell the statistical estimator what you know about the shape. The basic concept is to reduce the density estimation problem to a fixed-design regression model. More precisely the algorithm is as follows:

1. Transform the sample *X* into (*Xb*, *Yb*) data where the *Xb* are equally spaced, using a binning procedure. For each bin *i*, *Yb*(*i*) = number of *X*(*j*) within bin *i*.
2. Perform a wavelet decomposition of *Yb* viewed as a signal, using fast algorithm. Thus, the underlying *Xb* data is 1, 2... *nb* where *nb* is the number of bins.
3. Threshold the wavelet coefficients using one of the methods used for de-noising [11].
4. Reconstruct an estimate *h1* of the density function *h* from the threshold wavelet coefficients using the fast wavelet transform algorithm [12].
5. Post process the resulting function *h1*. Rescale the resulting function transforming 1, 2... *nb* into *Xb* and interpolate *h1* for each bin to calculate *hest*(*X*).

The first step of this estimation scheme depends on *nb* (the number of bins), which can be viewed as a bandwidth parameter. In density estimation, *nb* is generally small with respect to the number of observations (equal to the length of *X*), since the binning step is a pre-smoother. A typical default value is *nb* = length(*X*) / 4.

4. Results and Discussion

The density estimation of wind speed data is taken into consideration as a case study. The original data set on which the present study is focused is the yearly maximum wind speed data, in miles/hour shown in Table 1 and this data set has been quoted from Castillo (1988) [13].

Table 1. Yearly Maximum Wind Speed Data

Serial No	Wind Data (miles per hour)				
1	22.64	22.80	23.75	24.01	24.04
2	24.24	24.74	25.45	25.55	25.66
3	25.99	26.63	26.69	26.88	26.89
4	27.12	27.43	27.69	27.71	28.12
5	28.58	28.88	29.12	29.45	29.48
6	30.18	31.31	31.55	31.57	32.54
7	32.98	33.83	33.86	34.64	35.21
8	36.82	37.23	38.09	38.26	38.82
9	38.96	38.90	42.99	43.66	44.61
10	45.24	47.91	54.75	69.40	98.16

As it is known from many literatures [13, 14, and 15] that Weibull probability distribution is one of the extreme value distributions and wind speed data follows this distribution, therefore, before the density estimation, standard Weibull distribution has been fitted through this data set and the fitted probability density function is presented in

Figure 3. Mathematically Weibull probability density function is represented by equation (10), where ‘a’ denotes the location parameter and ‘b’ signifies the shape parameter.

$$y = f(x \mid a, b) = \left(\frac{b}{a}\right)\left(\frac{x}{a}\right)^{(b-1)} \exp\left[-\left(\frac{x}{a}\right)^{b}\right] \qquad (10)$$

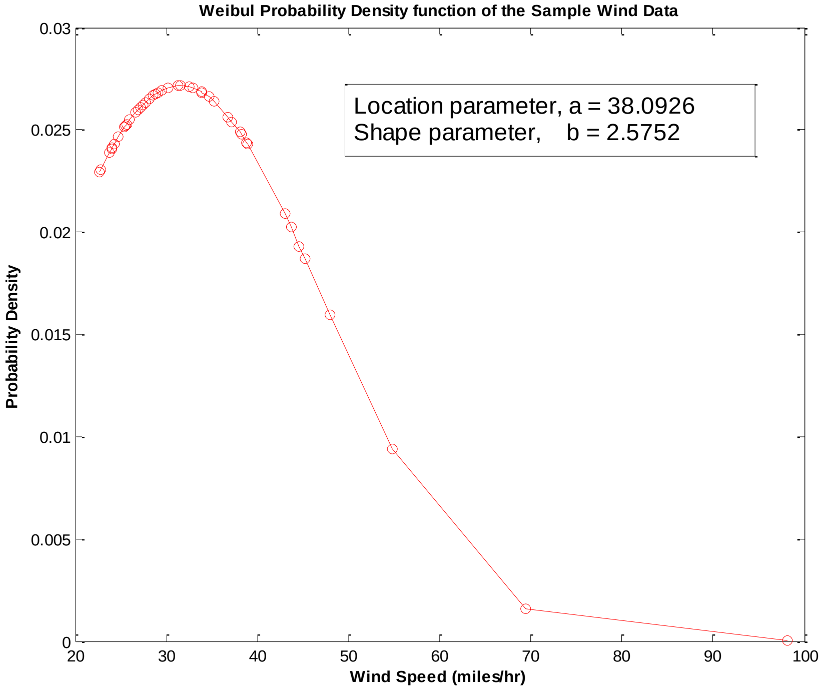

Figure 3.0. Weibull Fit of the Wind Speed Data (miles/hr)

For the sake of completeness, the statistical properties of the fitted Weibull distribution such as mean, standard deviation and entropy are computed as 33.83 miles/hr, 14.09 miles/hr and 3.69 respectively. It is also standard that if the variable ‘x’ follows Weibull distribution, then the log transformed values of that variable, *i.e.*, log(x) follows Type-I extreme value (Gumbel) distribution [16, 17]. With a view to this statement, log transformed values of the data set (Table 1) is fitted through a Gumbel distribution as shown in Figure 4. The statistical properties such as mean and standard deviation of the fitted Gumbel distributions are computed as 3.42 miles/hr and 0.498 miles/hr respectively. The estimated parameters while fitting both the Weibull and Gumbel distributions are tabulated in Table 2. Finally, using the new approach based on wavelet analysis, the probability density of the wind data set as tabulated in Table 1 is estimated. Wavelet analysis [18] based this estimation falls into the category of functional analysis and hence it is required to select first the wavelet to be used for decomposing the data set taken into account in the form of a signal [19]. As per the selection criteria of wavelets (bi-orthogonality should exist and number of vanishing moments [20] should be maximum), ‘symlet’ as mother wavelet is selected for decomposing the data set into the wavelet family. The general characteristics of the symlets wavelet are that they are compactly supported with least asymmetry and have highest number of vanishing moments for a given support width. Associated scaling filters of symlets are near linear-phase filters. The other important characteristics of symlets necessary for computation are as presented in Table 3.

Table 2 Parameters of the Fitted Distributions

Parameter	Weibull Probability Distribution (Fitted)	Type-I Extreme Value Probability Distribution (Fitted)
Location	38.0926	3.64
Shape	2.5752	0.3883

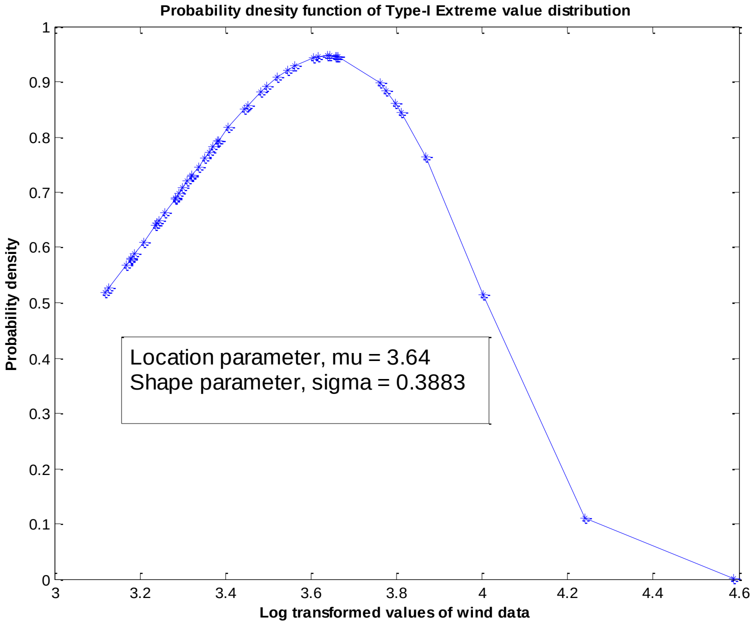

Figure 4.0. Type-I Extreme Value Probability Density Function (Gumbel Probability) of Log Transformed Values of Weibull Probability Density Function

Table 3. Important Characteristics of Symlets Wavelet

Description of item	Specification
Family	Symlets
Short Name	Sym
Order N & Examples	2, 3, higher & sym2, sym8, sym12
Orthogonal & Biorthogonal	Yes
DWT & CWT	Possible
Support width	2N – 1
Filters length	2N
Regularity	Symmetry near from
Number of vanishing moments for psi	N

With a view to the general characteristics and the important properties of symlet wavelet, the present computation of density estimation of wind data set is carried out using wavelet ‘sym15’ at level 5. Algorithm as depicted in Section 3.1 is used for computing the probability density function of the wind data (wind speed) as tabulated in Table 1. At the first stage the raw data is binned by using a suitable binning procedure (step 1 of algorithm) and the binned data is presented in Figure 5.0. The binned data, suitably normalized has been processed by wavelet decomposition (Step 2 of algorithm). The detail coefficients (coefficients contain maximum information of the original signal and of low frequency) generated by wavelet decomposition of the binned data are reconstructed after thresholding using global soft thresholding technique (step 3 of algorithm). The threshold parameters used in the computation for each detail coefficients (five detail coefficients are produced due to level 5 of the applied wavelet used for decomposition) are shown in Table 4. The basic purpose of this step is to remove any noisy events (high frequency) generally presents in detail coefficients. The approximate and the threshold detail coefficients are summed further to obtain the density of the wind data. Basically, the density estimate is the normalized sum of the signals containing the approximate and detail coefficients after coefficient thresholding. The computed density function is presented in Figure 6.0.The entropy of this density is estimated as 4.27 which is very near to that obtained through Weibull fitted distribution. Therefore, from the results of the density estimation using wavelet, it can be concluded that symlet – that is sym15 wavelet in the wavelet family can be an alternate replacement of Weibull distribution of the wind data set.

Table 4. Threshold Parameters for Coefficient Thresholding

Level	Value 1	Value 2	Value 3
1	22.64	98.16	0.006
2	22.64	98.16	0.070
3	22.64	98.16	0.064
4	22.64	98.16	0.030
5	22.64	98.16	0.006

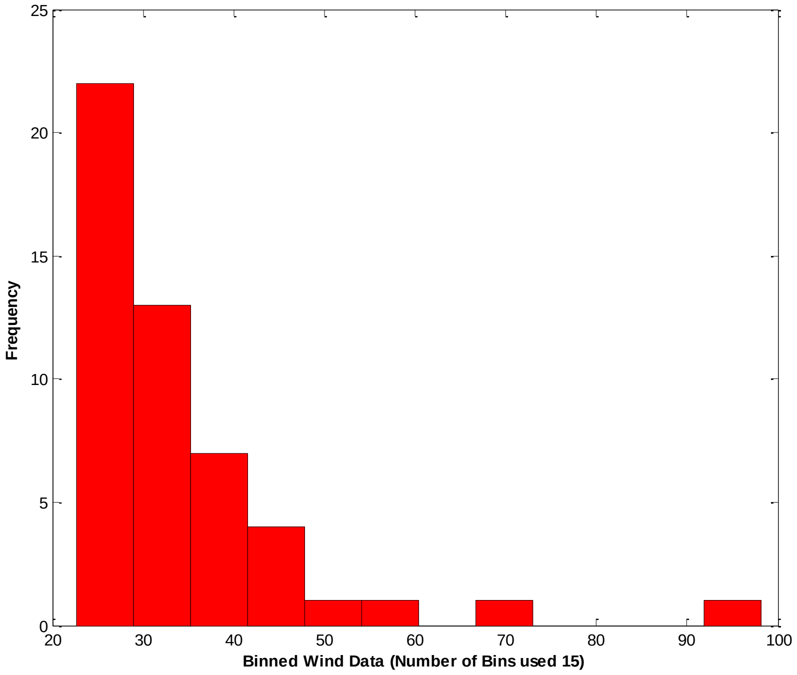

Figure 5.0. Wind Data after Binning

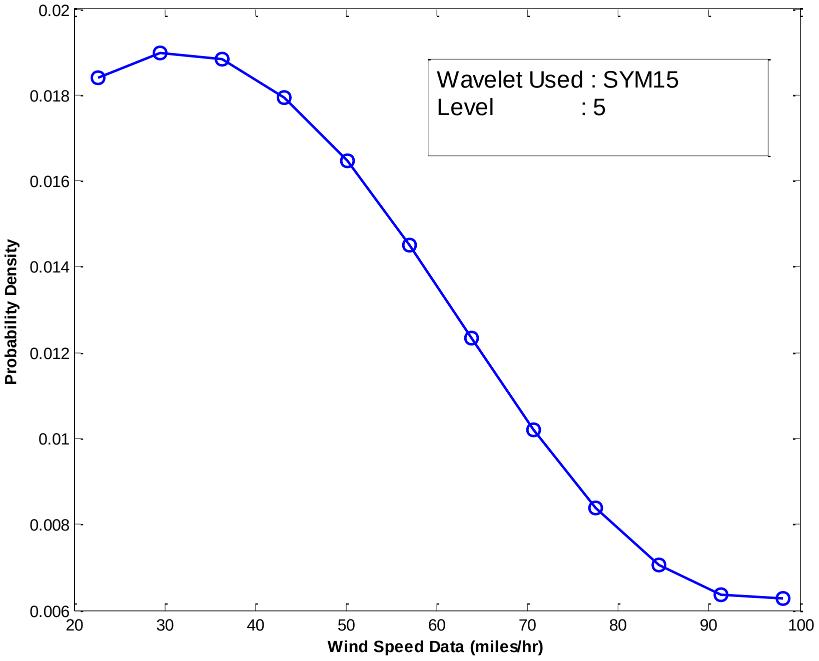

Figure 6.0. Wavelet Analysed Density Function of Wind Speed Data

5. Conclusions

Wavelet analysis based technique has been explored for estimating the probability density function of the wind data set of small sample size. The resulting wavelet based density is compared with the Weibull fitted probability density by computing the information entropy of the respective density and found to be approximately equal. It goes without saying that wavelet approaches will be of growing importance in data analysis, especially if the data belongs to a time series. It should also be mentioned that most of current works on wavelet applications in data analysis are based on density estimation. It can also be concluded that wavelets in statistics plays a major role for an innovative interpretation of the data set. In the density estimation, many other coefficient thresholding can be adopted; however selection of the thresholding requires an experience. It can also be argued that orthonormal basis may not be the best representation for noisy data even though the vanishing moments can help them to achieve denoised and reduced dimension of the data. Therefore, it is usually likely that thresholding wavelet coefficients remove useful information when they try to remove the noise or redundant information (noise can also be regarded as one kind of redundant information). To represent redundant information, it might be good to use redundant wavelet representation which is nothing but wavelet frames.

References

[1] F. Abramovich, T. Bailey and T. Sapatinas, “Wavelet analysis and its statistical applications”, JRSSD, vol. 48, **(2000)**, pp. 1–30.

[2] Daubechies. Ten Lectures on' Wavelets**.** Capital City Press, Montpelier, Vermont, **(1992)**.

[3] C. Chiann and P. A. Morettin, “A wavelet analysis for time series”, Journal of Nonparametric Statistics, vol. 10, no. 1, **(1999)**, pp. 1-46.

[4] D. Hand, H. Mannila and P. Smyth, “Principles of Data Mining”, The MIT Press, **(2001)**.

[5] I. Bruha and A. F. Famili, “Postprocessing in machine learning and data mining”, SIGKDD Exlporations, **(2000)**.

[6] R. Young, “Wavelet Theory and its Application”, Kluwer Academic Publishers, Bonston, **(1993)**.

[7] L. I. Schiff, “Qunatum Mechanics”, McGraw-Hill, Singapore, **(1968)**.

[8] P. Flandrin, “Wavelet analysis and synthesis of fractional Brownian motion”, IEEE Transactions on Information Theory, vol. 38, no. 2, **(1992)**, pp. 9 10-9 17.

[9] G. P. Nason and B. W. Silverman, “The stationary wavelet transform and some statistical applications”, Lecture Notes in Statistics, vol. 103, pp. 281-299.

[10] S. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 11, no. 7, **(1989)**, pp. 674-693.

[11] D. L. Donoho, “De-Noising by soft-thresholding”, IEEE Trans. on Inf. Theory, vol. 41, no. 3, **(1995)**, pp. 613-627.

[12] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard, “Density estimation by wavelet thesholding”, Annals of Stat., vol. 24, **(1996)**, pp. 508-539.

[13] E. Castillo, “Extreme Value Theory in Engineering”, Academic Press, INC.(London), Harcourt Brace Jovanovich Publishers, ISBN 0-12-163475-2, **(1988)**, pp. 334.

[14] J. A. Carta and P. Ramrez, “Analysis of two-component mixture Weibull statistics for estimation of wind speed distributions”, Renewable energy, vol. 32, **(2007)**, pp. 518-531.

[15] J. A. Carta and P. Ramrez, “Use of finite mixture distribution models in the analysis of wind energy in the canarian Archipelago”, Energy Conversion & Management, vol. 48, **(2007)**, pp. 281-291.

[16] K. R. Robert, “Applied Extreme Value Statistics”, BATTELLE Press, Columbia, Collier Macmillan Publishers, London, ISBN 0-02-947630-5, **(1985)**.

[17] W. J. Conover, “Practical Nonparametric Statistics”, Wiley, **(1980)**.

[18] D. L. Donoho and I. M. Johnstone, “Minimax estimation via wavelet shrinkage”, Annals of Statistics, vol. 26, no. 3, **(1998)**, pp. 879-921.

[19] A. Antoniadis, “Wavelets in statistics: a review”, J. It. Statist**.** Soc., **(1999)**.

[20] J. C. Pesquet, H. Krim and H. Carfatan, “Time-invariant orthonormal wavelet representations”, IEEE Trans. Sign. Proc., vol. 44, no. 8, **(1996)**, pp. 1964-1970.