A Comparison Between Reconstruction Methods for Generation of Synthetic Time Series Applied to Wind Speed Simulation

NEERAJ DHANRAJ BOKDE[1,2], **ANDRÉS FEIJÓO**[3], **NADHIR AL-ANSARI**[4],
AND ZAHER MUNDHER YASEEN[5]

[1]Department of Electronics and Communication Engineering, Visvesvaraya National Institute of Technology, Nagpur 440010, India
[2]Department of Engineering, Renewable Energy and Thermodynamics, Aarhus University, 8000 Aarhus, Denmark
[3]Departamento de Enxeñería Eléctrica, Universidade de Vigo, 36310 Vigo, Spain
[4]Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, 97187 Luleå, Sweden
[5]Sustainable Developments in Civil Engineering Research Group, Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam

Corresponding author: Zaher Mundher Yaseen (yaseen@tdtu.edu.vn)

ABSTRACT Wind energy is an attractive renewable sources and its prediction is highly essential for multiple applications. Over the literature, there are several studies have been focused on the related researches of synthetic wind speed data generation. In this research, two reconstruction methods are developed for synthetic wind speed time series generation. The modeling is constructed based on different processes including independent values generation from the known probability distribution function, rearrangement of random values and segmentation. They have been named as Rank-wise and Step-wise reconstruction methods. The proposed methods are explained with the help of a standard time series and the examination on wind speed time series collected from Galicia, the autonomous region in the northwest of Spain. Results evidenced the potential of the developed models over the state-of-the-art synthetic time series generation methods and demonstrated a successful validation using the means of mean and median wind speed values, autocorrelations, probability distribution parameters with their corresponding histograms and confusion matrix. Pros and cons of both methods are discussed comprehensively.

INDEX TERMS Energy sustainability, synthetic data, time series, wind speed, wind energy.

I. INTRODUCTION

Human population increment, excessive pollution, depletion of fossil fuels and growing environmental concerns lead to a raise in the demand of renewable energies. In particular, wind energy is of a great importance and potential owing to its availability and efficiency. Across the world, meteorological agencies collect wind speed data at many locations and make them available for wind energy research activities. The long period of historical data can be collected to perform the best research source. However, the major difficulty mainly associated with the unavailability of continuous time series data for a long time at specific locations.

The simulation of numerical series under the constraint of keeping some of their statistical characteristics, such as mean values or correlations, is a subject of interest in different research areas. In this sense, wind speed modeling has become a major interest in the recent years and mainly due to the increase of wind energy presence in large electrical power networks.

An assessment of wind speed simulation methods was presented in [1], where wind speed values at different locations were obtained simultaneously, satisfying some of the mentioned constraints. Auto-regressive and Markov methods were also studied in that work, for the generation of series with given auto-correlation properties.

In another researches [2], [3] have suggested several models to generate synthetic wind speed data and have used them for training instead of the original time series data.

Markov chains are used for describing a stochastic process (Markov process) in which next state is decided by the previous states depending on its probability distribution.

Research works presented in [4]–[7] have demonstrated how the Markov chain models can be used to generate synthetic wind speed time series. Markov processes have been used to generate streamflow data in [8]. Similarly, such processes have found synthetic time series generation applications in different fields including annual rainfall [9], [10] and river water flow [11]. Further, the Markov chain method has been updated with higher order schemes in [6]. Most of the research articles have presented successful outcomes of the Markov process in the wind speed data generation.

Apart from Markov processes, auto-regressive models are among the most widely accepted models in synthetic data generation processes. Auto-regressive models include auto-regressive (AR), moving average (MA), auto-regressive moving average (ARMA), auto-regressive integrated moving average (ARIMA) [12]–[14] and use auto-correlation techniques in a time series. A Bayesian method for the choice of the best model for a given series belonging to a set of auto-regressive and moving average techniques was studied in [13]. ARIMA and artificial neural networks (ANN) techniques were combined in a hybrid methodology pretending to take advantage of both methodologies in linear and nonlinear modeling, including experimental results which revealed that the proposed model can improve forecasting accuracy achieved by both methods when applied separately [14]. The applications and performance evaluation of these methods in wind speed data generation have been discussed and compared thoroughly in [1], [15]–[17]. In [1], several wind speed simulation methods were assessed, with inclusion of the correlation between series of wind speeds at several locations. Wind speed and wind power series simulation were also presented in [15], where correlations were taken into account, as well as non-Gaussian distributions and diurnal non-stationarity. The problem of day-ahead and two-day-ahead forecasting of wind speed was dealt with in [16] with the help of the method called fractional ARIMA of f-ARIMA. In [17], wind speed series were simulated for economic dispatch analysis studies, where correlations and auto-correlations were taken into account, together with a procedure of Normal-Weibull distributions conversion.

Nonetheless, there are few other methods which have been successfully examined for wind speed generation processes. These methods include the use of a Kalman filter [18]–[21], Bayesian models [22], [23], Neural networks [24] and Wavelet transform [25], [26], and also combinations of several methods.

The aim of this research is to present two methods for simulating synthetic time series. The modeling results are validated against those obtained by means of some of the methods proposed in [1]. The proposed methods are based on initiating a vector of values satisfying certain statistical properties, rearranging this vector and conferring it the desired chronological properties. Both methods have been developed for this work, although one of them has been inspired by a previous work of Iman & Conover [27], and can be considered as an adaptation of it. It is known that different prediction

horizons requires different kind of models in wind energy. Generally, for mid- and long-term planning or risk assessment of wind power integration, the probability distribution models or time-periodic model of wind speed are applicable [28]. Whereas, for short-time wind power forecasting, machine-learning models or artificial intelligent algorithms, time series models of wind speed are applicable [29]. The proposed methods are for synthetically (stochastic) generation of a time series, for a known statistical characteristics of a small time series. For sure, these methods will be useful for forecasting applications, but it cannot be differentiate into short, mid or long term, at this stage. Rather, it can be applied to all of them depending on the time series sample under observation.

The rest of the paper is structured as follows. The methodologies proposed are described in Section II. Section III discusses the results of the proposed methods when applied to the generation of synthetic time data. In Section IV, they are compared with other well established methods for wind speed time series and their validity is examined with the help of the calculation of statistical values such as mean and median ones, auto-correlation, probability distribution parameters and the corresponding histograms. A brief discussion is included in Section V and the conclusions and implications of this research are summarised in Section VI.

II. PROPOSED METHODOLOGIES

Though synthetically generated time series find much application in various areas, there are very few methods specifically devoted to synthetic time series generation. Usually, these synthetically generated time series are validated with the distribution patterns and statistical characteristics along with their seasonality and trend patterns in comparison to the original time series. While scanning the literature on synthetic time series generation methods, a very basic and common method is available, which takes independent values from a specific probability distribution and generate a new series having similar probability distribution [6], [30], [31].

Kaminsky *et al.* [6] have given a clear description on how a cumulative distribution and its inverse can be used to generate time series. With a known probability distribution of a time series, a new time series will be generated with similar probability distribution characteristics and histogram, and with acceptable mean and variance. But the power spectral density or the autocorrelation function of this synthetic time series will be significantly different from the original one. As an example, given a time series following a Weibull distribution with three parameters (scale, shape and location), with these parameters, a new time series is generated with independent values from the same Weibull distribution. The newly generated signal will be of Weibull distribution of similar scale, shape and location parameters but it will be only a sequence of random numbers without any autocorrelation similarity with the original series.

A proper rearrangement of this sequence of random numbers may lead to a generation of more accurate synthetic series and this is the motivation to introduce the Rank-wise

and Step-wise reconstruction methods, developed with this objective and presented in Sections II-A and II-B. Both methods have been designed for this purpose, although the Rank-wise one has been inspired by an interesting algorithm presented by Iman and Conover [27]. Those authors proposed an algorithm for inducing Spearman rank correlation in different distributions. Such a procedure, appropriately modified, can be used for inducing auto-correlation in a sequence, such as explained in Section II-A.

A. METHOD 1: RANK-WISE RECONSTRUCTION METHOD

The Rank-wise reconstruction method consists of reordering a set of values that have been initially generated under given conditions with regards to the distribution features. Once the series (a sequence of random numbers) has been generated, the goal is to find the permutation of this series that provides the maximum rank correlation between it and the original one. If the size of the series is n, and all the values inside are different, then there are $n!$ permutations of such a series. The calculation of the rank correlations between all these series and the original one will give $n!$ values, which can be different. The maximum of these values corresponds to the desired series. In order to easily obtain this series, the idea is to use an algorithm similar to the one proposed by Iman and Conover, which avoids the calculation of the rank correlations of all the existing permutations, and obtains the desired permutation among all, with the help of a Choleski factorization of the correlation matrix and some operations of linear algebra. The original proposal of Iman and Conover was the obtention of correlated series of different distributions, not the induction of a given auto-correlation in a series. However, with a small variation, the method can be applied to the problem dealt with in this paper. This constitutes the proposal of this Section.

The fundamentals of the proposed method are explained here. The origin of the problem lies in the need to induce a given correlation matrix to several previously independent variables. By handling the known Pearson's parametric correlation, a method based on algebraic operations including the application of a Choleski factorization to the correlation matrix, operates adequately when the distributions are Normal. However, when the distributions are of a different nature, such as Rayleigh or Weibull, the method becomes inaccurate, due to different reasons, that have been discussed in the literature [32]–[35].

Iman and Conover realized that the same method could be applied to different kind of distributions by changing the use of parametric correlation coefficients to non-parametric Spearman rank coefficients. These coefficients are not calculated as a function of the values of the series, but as a function of the positions of such values in the sequence. The values need to be reordered or classified so that the desired Spearman rank correlations are obtained. The method can be summarized by saying that it uses the following steps: generation of independent Uniform or Normal distributions, as much as the number of series to be correlated; use of the method based on the Choleski factorization mentioned

above, but for the Spearman rank correlation matrix, and applied to these distributions, which will reorder the Uniform or Normal distributions according to the desired correlations; obtention of indices for ordering the resulting elements; and finally, application of similar indices to the original series. As mentioned above, the distributions can belong to different families.

A numerical example will help understand the process. An original series, such as the following one, can be assumed: (2, 7, 4, 3, 6, 5). The generated one is as follows: (2.5, 3.5, 4, 4.5, 6, 6.5). The indices in the first series will be (1, 6, 3, 2, 5, 4), assuming that the smaller the value is, the smaller the index assigned to it is as well. However, in the case of the second series, the indices are (1, 2, 3, 4, 5, 6), in a clearly different order. Now, reordering the second series according to the indices of the first one, the result will be (2.5, 6.5, 4, 3.5, 6, 4.5). The features of the distribution do not change in this process. However, the auto-correlation changes and is now much more similar to the one corresponding to the original series. The general scheme for Rank-wise reconstruction method is shown as below.

Given: Original Wind Speed Time Series x
Output: Synthetic Wind Speed Time Series $\bar{x}$

Rank-wise Reconstruction()
    $j \leftarrow rank(x)$
    $i \leftarrow Weibull\_par(x)$
    $m(t) \leftarrow Weibull\_gen(i)$
    $\bar{x} \leftarrow Rank\_rearrange(m(t), j)$
    return Synthetic Wind Speed Time Series $\bar{x}$

Functions:
rank - to decide the rank of each entity in time series in ascending order.
Weibull_par - to calculate Weibull Parameters.
Weibull_gen - to generate a time series with known Weibull Parameters.
Rank_rearrange(a,b) - to rearrange a time series (a) in accordance with a vector of ranks (b).

For the basic understanding of the Rank-wise reconstruction method, a small and classic Box & Jenkins airline time series, collected from 1949 to 1960 on a monthly basis of an international airline [36], is examined here. Figs. 1 (a) and (b) show the original time series and the time series generated with the Rank-wise reconstruction method, respectively. Apart from this, the validity of the this method can be checked with the distribution parameters of the original and generated time series. With the maximum likelihood estimation method, the shape and scale parameters of the Weibull distribution for the original time series are found to be 2.5304 and 316.9294, respectively. Whereas, the shape and scale parameters for the synthetically generated time series are 2.3873 and 317.0714 respectively, very close to those of the original time series. Further, the correlation coefficient between original and generated time series is observed to be 0.9815 and their comparison in terms of basic statistical parameters is shown
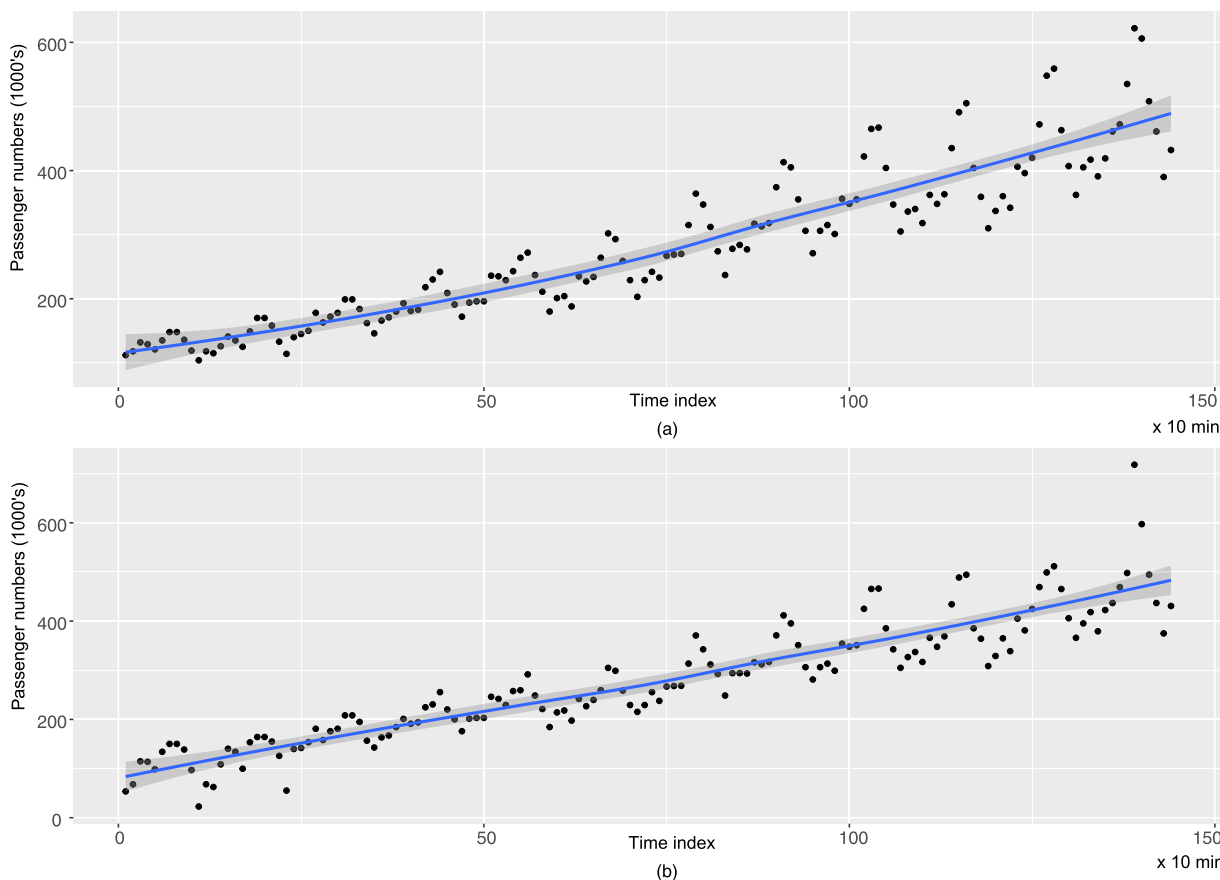
FIGURE 1. Plots showing (a) Air Passengers (original) time series, $x(t)$. (b) Time series generated with Rank-wise reconstruction method.

in Table 1, which supports the accuracy of the proposed model with statistical indicators.

The results of Rank-wise reconstruction method are further discussed in Sections III and IV.

B. METHOD 2: STEP-WISE RECONSTRUCTION METHOD

The Step-wise reconstruction method consists of reordering the set of values that has been considered in the Rank-wise method under given conditions of the distribution features. In this method, the rearranged process is performed by calculating the arithmetic difference between each time instant value of the original time series and all time instant values of the random number sequence. For each instant of the original time series, the value with minimum difference in the random number sequence is shifted to the corresponding instant which leads to generation of a rearranged time series. This rearranged time series will exhibit the pattern in which a drastic change can be observed after a specific time instant. The part of the time series prior to that time instant possesses similar patterns to that of the original time series whereas it gets deformed after that time instant. Hereafter, this time index will be named as 'knee point'. The original time series and the rearranged time series will be denoted by $x(t)$ and $x_R(t)$, respectively. The plot in Fig. 2 shows the errors

obtained between time series $x(t)$ and $x_R(t)$. The knee point at time instant '57' is the point from which significant error values are obtained and the pattern of time series $x_R(t)$ are found to be undesired. The Step-wise method handles this situation by segmenting the time series $x_R(t)$ and applying step-wise iterative operations till the desired synthetic time series $\bar{x}(t)$ is achieved. For the first iteration, time series $x_R(t)$ is segmented at the knee point into $x_{11}(t)$ and $x_{21}(t)$. Again $x_{21}(t)$ exhibits properties like another time series with random numbers.

In the second iteration, the whole operations including independent values generation from the known probability distribution, rearrangement of random values and segmentation, which were executed on $x(t)$, will be performed on time series $x_{21}(t)$, which leads to its segmentation into time series $x_{12}(t)$ and $x_{22}(t)$. The time series $x_{22}(t)$ is the random number sequence obtained in the second iteration and undergoes for the third iteration. This process of step-wise iteration is performed till a desired synthetic time series is obtained, such that it cannot be segmented further or the length of the random pattern segment will be negligible compared to the original time series. Finally, the time series generated by the sequential combination of the first segments obtained in each iteration ($x_{11}(t)$, $x_{12}(t)$, $x_{13}(t)$, ...) is considered as the

TABLE 1. Statistical characteristics of original and series generated with Rank-wise reconstruction method.

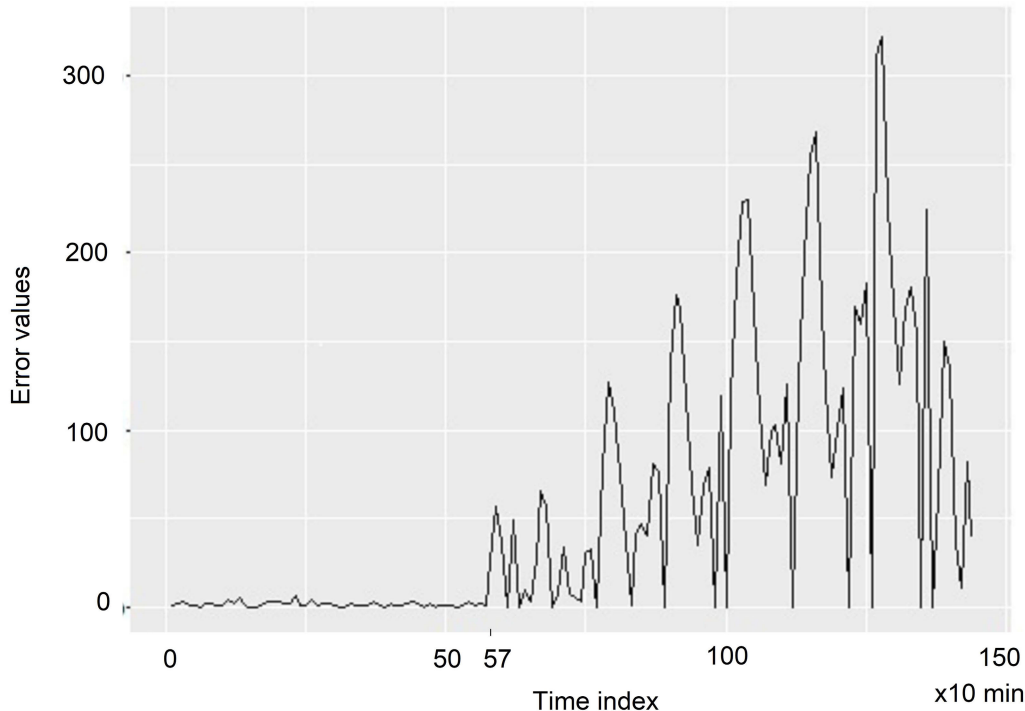| Series	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Deviation
Original	104.0	180.0	265.5	280.3	360.5	622.0	119.9
Synthetically generated	47.5	199.6	290.6	293.3	384.4	573.5	119.3

FIGURE 2. Plot showing errors between $x(t)$ and $\bar{x}(t)$ to estimate the knee point.

final synthetic time series, $\bar{x}(t)$. The time series $\bar{x}(t)$ is very similar to the original one ($x(t)$) from a seasonality and trend appearance point of view, but the statistical validation of this assert plays an important role in the decision of performance measure of the Step-wise method.

The Step-wise reconstruction method is illustrated step by step with the same Box & Jenkins airline time series [36]. Fig. 3 (c), (d) and (e) show the progress of the mentioned method in the first three iterations. In these plots, a summary related to each subplot is shown for better understanding. The mean values are shown in blue lines and the surrounding gray area represents the confidence interval which is a range of values describing the uncertainty associated with each plot. Fig. 3 (a) and (f) are the plots corresponding to original (a) and

synthetic time series (f), with very similar patterns to the original time series. Along with this, it is interesting that even though the Weibull distribution parameters of $x_{11}(t)$, $x_{12}(t)$ and $x_{13}(t)$ segments are varying and different from those of the original time series $x(t)$, those parameters of the synthetic time series ($\bar{x}(t)$) are close enough to those of the time series $x(t)$ as shown in Table 2. The general scheme for the Step-wise reconstruction is exhibited as follows.

Input: Original Wind Speed Time Series x
Variables:
knee point k
First Segment for nth iteration x_{1n}
Second Segment for nth iteration x_{2n}
Output: Synthetic Wind Speed Time Series $\bar{x}$

Step-wise Reconstruction()
$\quad n \leftarrow 1$
$\quad \bar{x} \leftarrow NULL$
$\quad\quad$ **While** $k > (0.05 * length(x))$
$\quad\quad\quad i \leftarrow Weibull\_par(x)$
$\quad\quad\quad m(t) \leftarrow Weibull\_gen(i)$
$\quad\quad\quad n(t) \leftarrow Re\_arrange(m(t))$

TABLE 2. Comparison of Weibull distribution parameters, shape and scale, of airlines time series generated at different iterations in the Step-wise reconstruction method.

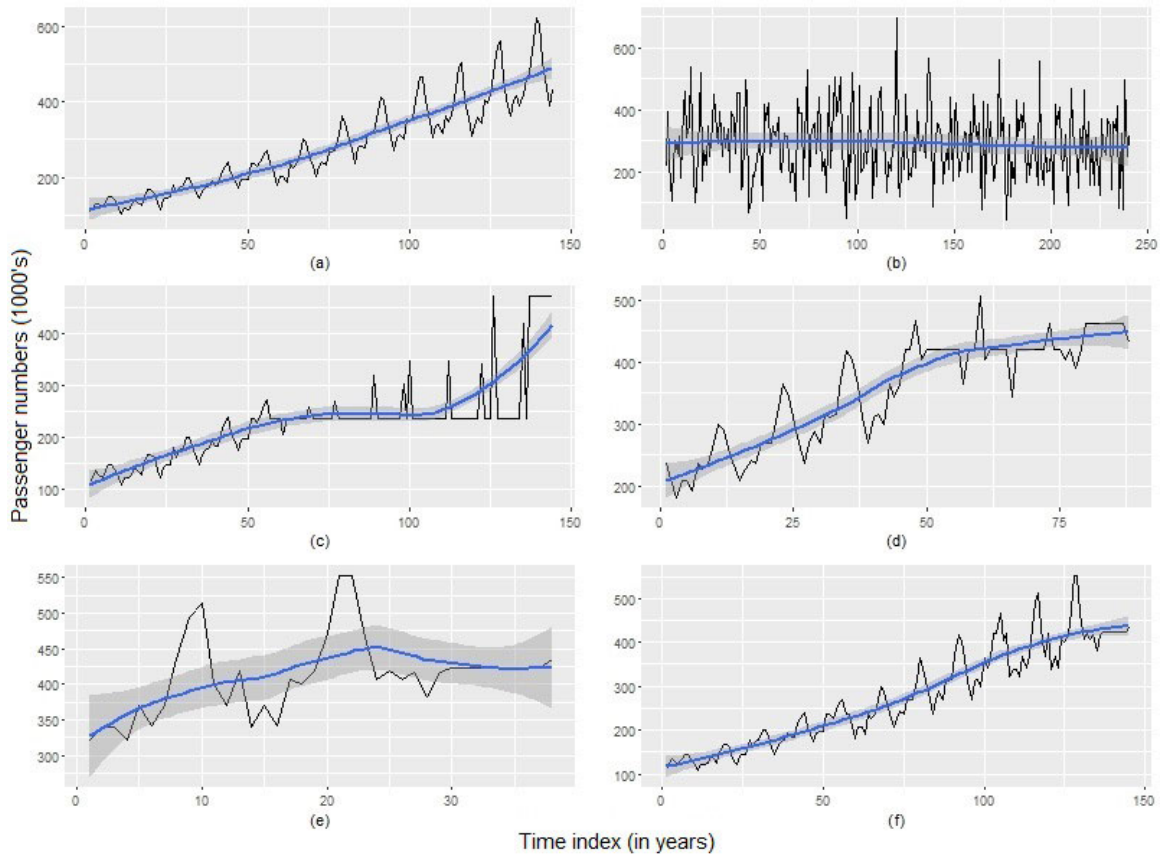| Time Series	Shape	Scale
$x(t)$	2.5304	316.9294
$x_{11}(t)$	5.7700	171.1231
$x_{12}(t)$	6.0531	282.3740
$x_{13}(t)$	6.4690	418.0920
$\bar{x}(t)$	2.7320	314.0647

FIGURE 3. Steps involved in Step-wise reconstruction method. (a) Air Passengers (original) time series, $x(t)$. (b) Time series generated with probability distribution parameters of $x(t)$. (c) (d) and (e) are the time series obtained in iteration 1, 2 and 3, respectively. (f) The synthetic time series $\bar{x}(t)$ generated in Step-wise reconstruction method.

$$k \leftarrow knee(n(t))$$
$$x_{1n} \leftarrow n(t)[1:k]$$
$$x_{2n} \leftarrow n(t)[(k+1):length(n(t))]$$
$$\bar{x} \leftarrow append(\bar{x}, x_{1n})$$
$$x \leftarrow x_{2n}$$
$$n \leftarrow n+1$$

return Synthetic Wind Speed Time Series $\bar{x}$

Functions:

Weibull_par- to calculate Weibull Parameters
Weibull_gen- to generate a time series with known Weibull Parameters
Re_arrange- to rearrange a time series with maintaining minimum error with respect to x
knee- to calculate knee point of the rearranged time series

The comparison in this section shows the performance of both proposed methods for the airline passengers dataset. The results show the comparable numbers in terms of shape and scale of the Weibull distribution with those of the original time series. In Section II, the proposed methods are applied to wind energy time series and their statistical validity is tested in Section IV.

III. APPLICATION OF WIND SPEED SIMULATION

The wind speed time series used in this study have been collected in Galicia, northwest of Spain. Galicia is one of the leading region in Spain in terms of wind energy capacity. It installed wind power density is higher than its mean value in the Spain as well as many leading countries [37]. The nature of wind in Galicia has high spatial and temporal variability within all over the year and no clear seasonal changes are observed. And, in all seasons, upwelling and nonupwelling patterns are evident [38]. In this study, wind speed mean values were measured every 10 minutes in meteorological stations of the Galician meteorological network. These time series have been collected along several years and they constitute public information. The time series used in this study corresponds to one of the stations during three months, as shown in Fig. 4. Generally, the wind speed time series are considered as part of a Weibull distribution [39], and this has been assumed here. The scale and shape parameter are 1.629 and 5.890 respectively.

In this section, the performances of the Rank-wise and Step-wise reconstruction methods are evaluated on wind speed time series and compared with state-of-the-art synthetic time series generation methods including first and second order Markov chain and auto-regressive method, $AR(1)$.
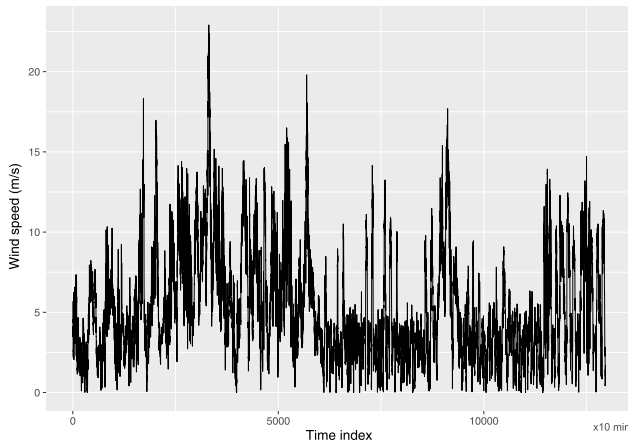
FIGURE 4. The wind speed time series collected in Galicia, Spain.

A comparison of all these methods is performed with the generation of a synthetic time series of the same length as that of the original one. Fig. 5 shows the plot of comparison for synthetic time series corresponding to all the methods under study. All the time series plots in Fig. 5 represent the 100 initial simulated values for the respective series. It can be observed that the Step-wise reconstruction method can generate the time series more accurately, but the validation of the results will be the most important factor to assess the performance of the proposed methods. The validation of the proposed methods is discussed in Section IV.

IV. STATISTICAL VALIDATION

The generated time series by means of all the methods studied are shown in Fig. 5. The patterns shown by the generated synthetic time series are similar to those of the original wind speed series. In order to clarify the performance of each method, various validation techniques are discussed in this Section. These techniques include auto-correlation, average wind speed, probability distribution function parameters and histograms. The results are validated by comparing the generated synthetic time series with the original ones for the same time intervals and lengths.

A. AVERAGE WIND SPEED

The synthetic time series can not be considered acceptable if the average values are not comparable with those of the original time series. In Table 3, all methods considered in the study are compared on the basis of mean and median of the respective synthetic time series. All them have been arranged in ascending order of their relative errors. The values in Table 3 show that both proposed methods outperform the contemporary ones.

B. AUTOCORRELATION FUNCTION

The auto-correlation function (ACF) is a relevant aspect to compare synthetic time series. With ACF, it is compared how similar the correlation of the values in a synthetic time series is to the correlation of the values in the original time series.

In Fig. 6, the ACF plots corresponding to all methods are shown. With these plots, it can be interpreted that the ACF of the proposed methods and auto-regressive methods are close to those of the original time series. Whereas, ACF plot of first and second order Markov chain methods are relatively shifted downwards and maintain magnitudes close to zero. This performance of Markov chain methods seems to be common [6], [7], [31].

In Fig. 6, an over-fitting of the ACF plot for the auto-regressive method is observed. Comparatively, a more accurate fitting is observed for the proposed methods.

C. PROBABILITY DISTRIBUTION FUNCTION

Usually, wind speed time series are considered to follow a Weibull distribution [39]. The Weibull distribution is generally expressed as a function of three parameters, i.e., shape, scale and location. The general expression for the Weibull probability distribution function in terms of shape (γ), scale (α) and location (μ) parameters is shown in 1.

$$f(x) = \frac{\gamma}{\alpha}(\frac{x-\mu}{\alpha})^{(\gamma-1)}e^{(-((x-\mu)/\alpha)^{\gamma})} \quad x \geq \mu; \gamma, \alpha > 0 \quad (1)$$

In this paper, the shape and scale parameters are calculated from the probability function using a maximum likelihood estimator. In the case of a wind speed distribution, (1) becomes simpler because the location parameter μ can be assumed to be 0. This is due to the fact that 0 is the minimum value for wind speeds. The shape and scale parameters for the original and synthetic wind speed time series can be read in Table 4. With these observations, the accuracy of the proposed methods while maintaining the probability distribution function in synthetic time series is examined. The values shown in brackets in Shape and Scale columns in Table 4 are the relative errors of the Shape and Scale values. These methods are arranged in ascending order of relative error values to indicate the performance-wise rank of all these methods.

The fourth column of Table 4 represents the analysis of variance (ANOVA) test used to compare the means of two or more samples based on the normality from the F distribution assumption. This test evaluates the null hypothesis that the samples from different groups are drawn from the populations with same mean values. In this study, a one-way ANOVA test is used to compare the results from all the methods used for time series generation. The details and application of the test are discussed in [40]–[42]. In this study, all the comparisons are significant at p-value = 0.05. This suggests the adequate selection of the methods under study.

The histograms of the original and the synthetic time series generated with all methods are shown in Fig. 7. The performance of Markov chain and auto-regression models are degraded compared with those of original time series. With the proposed methods, more accurate histograms are achieved, reflecting the accuracy and the validity of the proposed methods for the generation of synthetic wind speed time series.
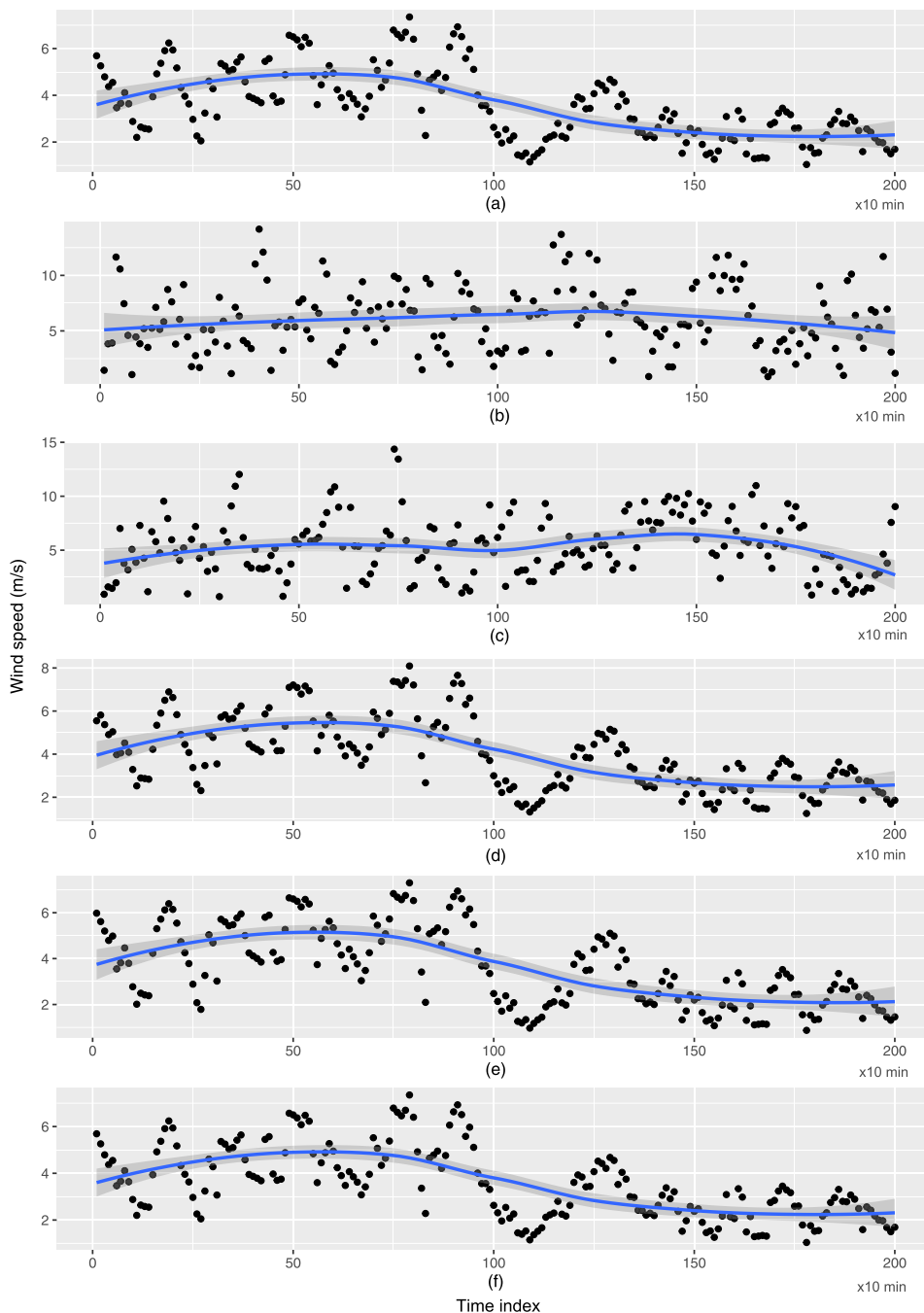
FIGURE 5. Plots showing patterns of initial 200 time-instances of synthetic time series generated with various methods. (a) Original Wind speed time series, (b) synthetic time series generated with first order Markov chain method, (c) second order Makov chain method, (d) an autoregressive method (e) rank-wise reconstruction method and (f) step-wise reconstruction method, respectively.

D. CONFUSION MATRIX

With the consideration of all prior validation schemes, it can be concluded that both proposed methods are performing better than other contemporary methods. But to further observe the accuracy of these methods, the synthetic time series are compared with the original time series at different amplitude levels. This is achieved with the classification of the original

time series into 4 steps in accordance to equal amplitude ranges as shown in Fig. 8.

The synthetic time series generated by all methods are compared within 4 steps of original time series. For each step, values in synthetic time series which match with respective step number of the original time series represent the 'True Positive' count of that particular step number. The true

TABLE 3. Comparison of mean and median of synthetic time series generated with different methods.

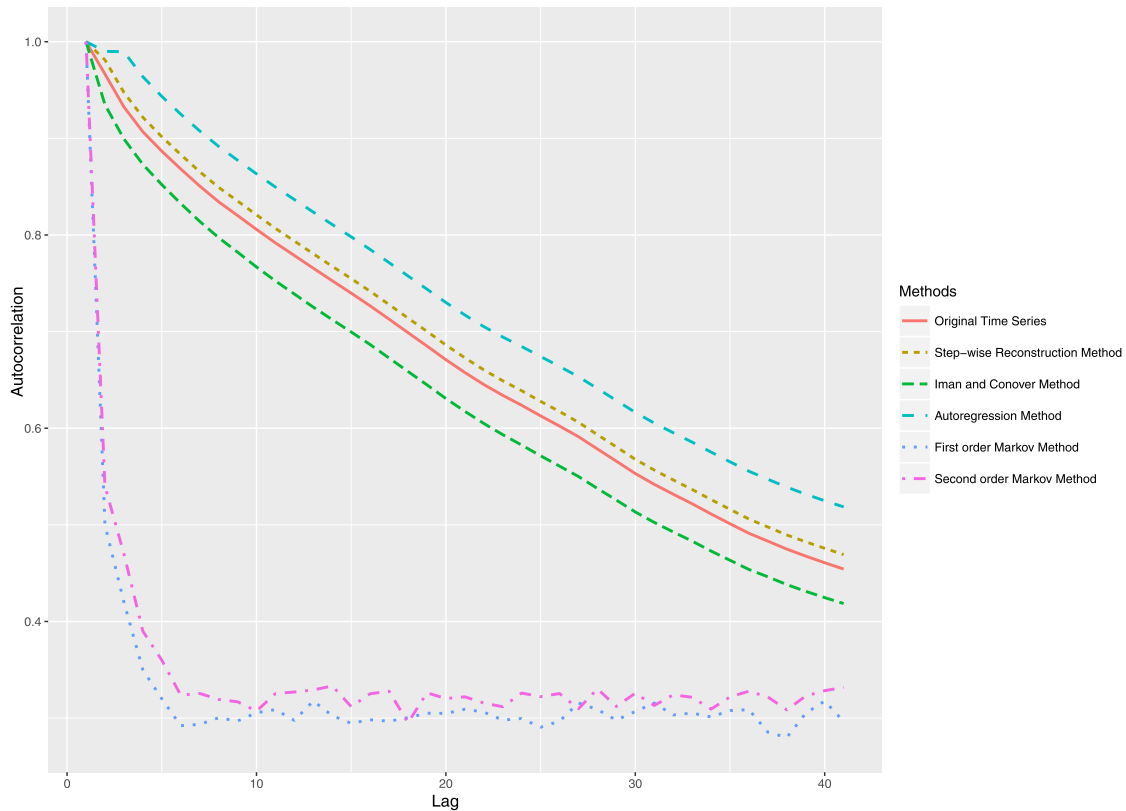| Methods	Mean	Median	Relative Error (With respect to Mean values)
Original Time Series	5.226	4.320	-
Step-wise Reconstruction	5.235	4.328	0.0007
Rank-wise Reconstruction	5.258	4.715	0.0045
Autoregressive	5.796	4.780	0.1000
Second Order Markov Chain	5.971	5.512	0.1425
First Order Markov Chain	6.028	5.787	0.1534

FIGURE 6. Comparison of ACFs for original and synthetic time series generated with different methods.

TABLE 4. Comparison of Weibull distribution parameters, shape and scale of synthetic time series generated with different methods for wind speed time series.

Methods	Shape	Scale	One-way ANOVA (P Value)
Original Time Series	1.622	5.834	-
Step-wise Reconstruction	1.624 (0.0012)	5.838 (0.0006)	<2e-16
Rank-wise Reconstruction	1.628 (0.0036)	5.876 (0.0075)	<2e-16
Auto-regressive	1.628 (0.0036)	6.477 (0.0611)	4.82e-10
Second Order Markov Chain	1.816 (0.1196)	6.792 (0.1642)	1.94e-07
First Order Markov Chain	1.996 (0.2305)	6.804 (0.1662)	2.23e-08

positive counts for all methods for each step are noted in Table 5.

In this section, accuracy assessment of the proposed methods is accomplished through the comparison of the different states (classified as discussed earlier) through the use of confusion matrix [43]. With the help of the confusion matrix, it becomes possible to determine the accuracy of both proposed methods more in detail in terms of overall accuracy, user's accuracy, producer's accuracy, F1-measure and Kappa coefficient.

Tables 6 and 7 show the confusion matrix for Step-wise and Rank-wise reconstruction methods, respectively. In these tables, the values on the diagonal correspond to the correctly observed states, whereas off-diagonal values correspond to
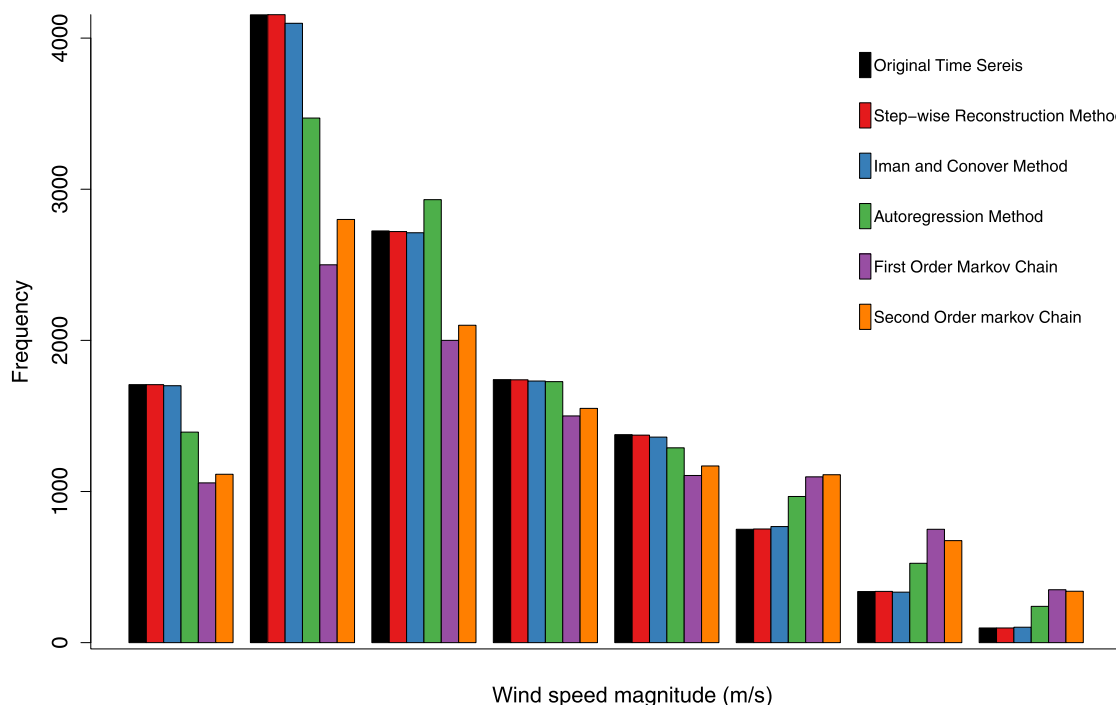
FIGURE 7. Comparison of Probability distributions for original and synthetic time series generated with different methods.
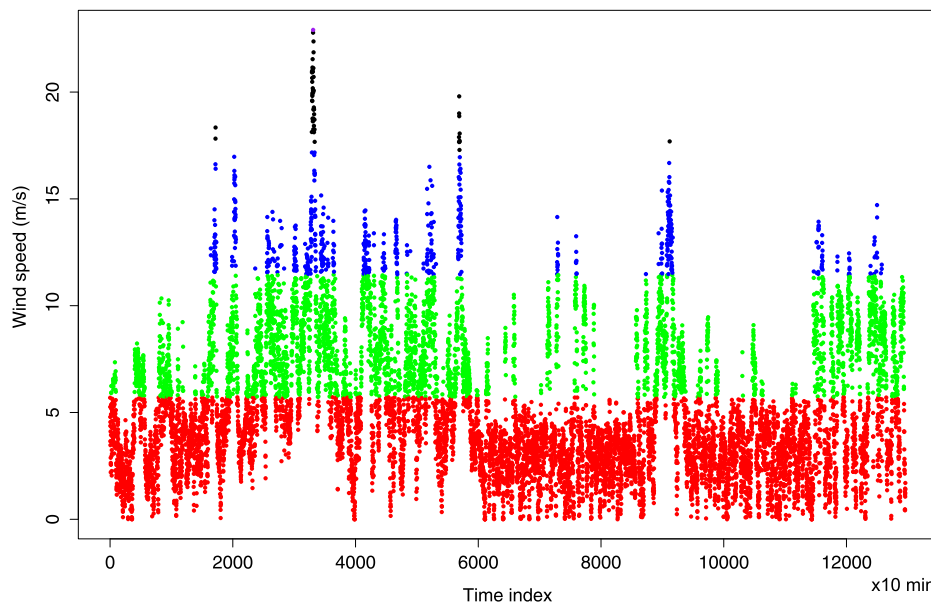

FIGURE 8. Original time series with classification in 4 steps (shown with different colors).

the wrong prediction of the states. The accuracy in classification of the states is validated with overall accuracy of 98% and 97% along with Kappa coefficients 0.97 and 0.945, respectively for Step-wise and Rank-wise methods. F1 measures for each step are 99%, 97%, 95%, and 93% for the Step-wise method and those for the Rank-wise method are 98%, 95%, 96%, and 74%. These results show that the Step-wise method achieves more accuracy compared with the Rank-wise reconstruction method.

V. DISCUSSION

The performance of the proposed methods on wind speed time series reveals that they outperform Markov chain and auto-regression methods. The validity of the proposed methods has been examined with the help of different parameters with respect to the observation, and the Step-wise reconstruction method has been observed as the most accurate. But this accuracy is achieved at the cost of more execution time. Table 8 shows the comparison of execution

TABLE 5. True Positive counts for all four step.

Methods	Step 1	Step 2	Step 3	Step 4	Total
Original Time Series	8306	3990	612	51	12959
Step-wise Reconstruction	8227	3886	579	42	12734
Rank-wise Reconstruction	8045	3770	554	30	12399
Auto-regressive	7987	3131	337	32	11487
Second Order Markov Chain	6490	2917	403	28	9838
First Order Markov Chain	6032	2607	380	22	9041

TABLE 6. Confusion matrix for step-wise reconstruction method.

		Actual					User's
		State 1	State 2	State 3	State 4	Totals	Accuracy
	State 1	8227	63	2	0	8292	99%
	State 2	65	3886	23	0	3974	97%
Predicted	State 3	4	27	579	2	612	94%
	State 4	0	0	4	42	46	91%
	Totals	8396	3976	608	44	12924	
Producer's Accuracy		99%	97%	95%	95%	Overall Accuracy - 98%	
F1- measure		99%	97%	95%	93%	Kappa Coeff. - 0.97	

TABLE 7. Confusion matrix for Rank-wise reconstruction method.

		Actual					User's
		State 1	State 2	State 3	State 4	Totals	Accuracy
	State 1	8045	149	2	0	8196	98%
	State 2	162	3770	4	0	3936	95%
Predicted	State 3	2	2	554	12	570	97%
	State 4	0	0	13	35	48	73%
	Totals	8209	3921	573	47	12750	
Producer's Accuracy		98%	96%	96%	74%	Overall Accuracy - 97%	
F1- measure		98%	95%	96%	73%	Kappa Coeff. - 0.945	

TABLE 8. Execution time comparison.

Methods	Execution time (Sec)
First Order Markov Chain	4.28
Second Order Markov Chain	6.15
Auto-regressive	0.93
Rank-wise Reconstruction	3.31
Step-wise Reconstruction	6.88

time in seconds for all the methods considered in the study.

The execution time for the Step-wise reconstruction method has been 6.88 seconds in the simulations, which is quite longer than for other methods. Whereas, the Rank-wise Reconstruction method, with the second best position according to performance of synthetic time series generation, takes 3.31 seconds to simulate wind speed time series, one of the shorter times taken in Table 8.

Hence, it can be suggested when the simulation dataset is very large and the execution time limitation is a constraint at the cost of the very little amount of accuracy. But whenever there is no time limitation, more accurate results can be achieved with the Step-wise reconstruction method. Simplicity and minimum calculations compared with other contemporary methods like Markov Chain and Auto-regression methods are the positive points for the proposed methods. Last but not least, the current research implementation demonstrated an excellent predictability

performance for the inspected case study. However, for further investigation, other meteorological stations can be inspected for the possibility to understand the generalization capacity of the proposed model. In addition, this is contributing to the base knowledge of the insightful information for wind speed prediction and and energy production in general.

VI. CONCLUSION

In this paper, two reconstruction methods are proposed for synthetic time series generation, with the aim of applying them to wind speed simulation. They are called Rank-wise and Step-wise methods. The Step-wise reconstruction method is based on independent values generation from a known probability distribution, rearrangement of random values and segmentation processes, which return a synthetic time series with statistical characteristics very close to those of the original wind speed time series. Similarly, the methodology of Rank-wise reconstruction method is based on rearrangement of random values of a known distribution. The rearrangement in the Rank-wise method is of single step, reordering the random values based on the maximum positive rank correlation between the original and the generated variables. The proposed methods have been explained with a standard sample time series and examined on wind speed time series collected in Galicia, Spain. This experiment has

compared the performance of both proposed methods with various state-of-the-art synthetic time series generation methods. The results show that the performance of the proposed methods achieve a remarkable accuracy in the statistical values of the synthetic wind speed time series and outperform auto-regressive, and first and second order Markov chain methods. Additionally, the statistical validation of the proposed methods have been performed with the mean wind speed, auto-correlation, probability distribution function and confusion matrix techniques. Also, the performance along with pros and cons of both methods have been discussed. With the evidence of results and their statistical validation, the conclusion is that the proposed methods are suitable for generation of the synthetic wind energy time series with a known distribution.

REFERENCES

[1] A. Feijóo and D. Villanueva, "Assessing wind speed simulation methods," *Renew. Sustain. Energy Rev.*, vol. 56, pp. 473–483, Apr. 2016.

[2] W. Yuan-Kang, L. Ching-Ying, T. Shao-Hong, and S.-N. Yu, "Actual experience on the short-term wind power forecasting at Penghu—From an island perspective," in *Proc. Int. Conf. Power Syst. Technol.*, Oct. 2010, pp. 1–8.

[3] M. J. Woods, C. J. Russell, R. J. Davy, and P. A. Coppin, "Simulation of wind power at several locations using a measured time-series of wind speed," *IEEE Trans. Power Syst.*, vol. 28, no. 1, pp. 219–226, Feb. 2013.

[4] E. D. H. Cheng, "Wind data generator: A knowledge-based expert system," *J. Wind Eng. Ind. Aerodyn.*, vol. 38, nos. 2–3, pp. 101–108 199).

[5] D. Jones and M. H. Lorenz, "An application of a Markov chain noise model to wind generator simulation," *Math. Comput. Simul.*, vol. 28, no. 5, pp. 391–402 1986.

[6] F. C. Kaminsky, R. H. Kirchhoff, C. Y. Syu, and J. F. Manwell, "A comparison of alternative approaches for the synthetic generation of a wind speed time series," *J. Sol. Energy Eng.*, vol. 113, pp. 280–289, Nov. 1991.

[7] A. D. Sahin and Z. Sen, "First-order Markov chain approach to wind speed modelling," *J. Wind Eng. Ind. Aerodynamics*, vol. 89, nos. 3–4, pp. 263–269 2001.

[8] H. A. Thomas and M. B. Fiering, "Mathematical synthesis of streamflow sequences for the analysis of river basin by simulation," *Des. Water Resource Syst.*, vol. 89 pp. 459–493, 1962.

[9] J. Borwein, P. Howlett, and J. Piantadosi, "Modelling and simulation of seasonal rainfall using the principle of maximum entropy," *Entropy*, vol. 16, no. 2, pp. 747–769, 2014.

[10] J. Piantadosi, J. Boland, and P. Howlett, "Generating synthetic rainfall on various Timescales—Daily, monthly and yearly," *Environ. Model. Assessment*, vol. 14, no. 4, pp. 431–438, 2009.

[11] S. Ahmad, I. H. Khan, and B. Parida, "Performance of stochastic approaches for forecasting river water quality," *Water Res.*, vol. 35, no. 18, pp. 4261–4266, 2001.

[12] H. Madsen, *Time Series Analysis*. Boca Raton, FL, USA: CRC Press, 2007.

[13] R. L. Kashyap, "Optimal choice of AR and MA parts in autoregressive moving average models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-2, no. 2, pp. 99–104, Mar. 1982.

[14] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159–175, Jan. 2003.

[15] B. G. Browna, R. W. Katzb, and A. H. Murphy, "Time series models to simulate and forecast wind speed and wind power," *J. Climate Appl. Meteorol.*, vol. 23, no. 8, pp. 1184–1195, 1984.

[16] R. G. Kavasseri and K. Seetharaman, "Day-ahead wind speed forecasting using f-ARIMA models," *Renew. Energy*, vol. 34, no. 5, pp. 1388–1393, 2009.

[17] D. Villanueva, A. Feijoo, and J. L. Pazos, "Simulation of correlated wind speed data for economic dispatch evaluation," *IEEE Trans. Sustain. Energy*, vol. 3, no. 1, pp. 142–149, Jan. 2012.

[18] E. A. Bossanyi, "Short-term wind prediction using Kalman filters," *Wind Eng.*, vol. 9, no. 1, pp. 1–8, 1985.

[19] K. Chen and J. Yu, "Short-term wind speed prediction using an unscented Kalman filter based state-space support vector regression approach," *Appl. Energy*, vol. 113, pp. 690–705, Jan. 2014.

[20] C. D. Zuluaga, M. A. Álvarez, and E. Giraldo, "Short-term wind speed prediction based on robust Kalman filtering: An experimental comparison," *Appl. Energy*, vol. 156, pp. 321–330, Oct. 2015.

[21] P. Louka, G. Galanis, N. Siebert, G. Kariniotakis, P. Katsafados, and I. Pytharoulis, and G. Kallos, "Improvements in wind speed forecasts for wind power prediction purposes using Kalman filtering," *J. Wind Eng. Ind. Aerodynamics*, vol. 96, no. 12, pp. 2348–2362, 2008.

[22] G. Li and J. Shi, "Application of Bayesian model averaging in modeling long-term wind speed distributions," *Renew. Energy*, vol. 35, no. 6, pp. 1192–1202, 2010.

[23] G. Li and J. Shi, "On comparing three artificial neural networks for wind speed forecasting," *Appl. Energy*, vol. 87, no. 7, pp. 2313–2320, 2010.

[24] G. Li, J. Shi, and J. Zhou, "Bayesian adaptive combination of short-term wind speed forecasts from neural network models," *Renew. Energy*, vol. 36, no. 1, pp. 352–359, 2011.

[25] D. Kaur, T. T. Lie, N. K. Nair, and B. Vallès, "Wind speed forecasting using hybrid wavelet transform-ARMA techniques," *AIMS Energy*, vol. 3, pp. 13–24, Jan. 2015.

[26] H. Liu, H.-Q. Tian, C. Chen, and Y.-F. Li, "A hybrid statistical method to predict wind speed and wind power," *Renew. Energy*, vol. 35, no. 8, pp. 1857–1861, 2010.

[27] R. L. Iman and W. J. Conover, "A distribution-free approach to inducing rank correlation among input variables," *Commun. Statist.-Simul. Comput.*, vol. 11, no. 3, pp. 311–334, 1982.

[28] J. Wang, X. Xiong, H. Li, and X. Lu, "Time-periodic model of wind speed and its application in risk evaluation of wind-power-integrated power systems," *IET Gener. Transmiss. Distrib.*, vol. 13, no. 1, pp. 46–54, 2018.

[29] N. Bokde, A. Feijóo, D. Villanueva, and K. Kulat, "A review on hybrid empirical mode decomposition models for wind speed and wind power prediction," *Energies*, vol. 12, no. 2, p. 254, 2019.

[30] A. Naimo, "A novel approach to generate synthetic wind data," *Proc.-Social Behav. Sci.*, vol. 108, pp. 187–196, Jan. 2014.

[31] A. Shamshad, M. A. Bawadi, W. M. A. Hussin, T. A. Majid, and S. A. M. Sanusi, "First and second order Markov chain models for synthetic generation of wind speed time series," *Energy*, vol. 30, no. 5, pp. 693–708, 2005.

[32] J. P. Kleijnen and W. van Groenendaal, *Simulation: A Statistical Perspective*. Hoboken, NJ, USA: Wiley, 1992.

[33] M. E. Johnson, *Multivariate Statistical Simulation A Guide to Selecting and Generating Continuous Multivariate Distributions*, vol. 192. Hoboken, NJ, USA: Wiley, 1987.

[34] L. Lee, "Multivariate distributions having Weibull properties," *J. Multivariate Anal.*, vol. 9, no. 2, pp. 267–277, 1979.

[35] D. R. Jensen, "A generalization of the multivariate Rayleigh distribution," *Indian J. Statist. Ser. A*, vol. 32, no. 2, pp. 193–208, 1970.

[36] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

[37] *Asociación Empresarial Eólica*. Accessed: Jul. 21, 2019. [Online]. Available: https://aeeolica.org

[38] R. Torres, E. D. Barton, P. Miller, and E. Fanjul, "Spatial patterns of wind and sea surface temperature in the Galician upwelling region," *J. Geophys. Res. Oceans*, vol. 108, no. C4, pp. 1–14, 2003.

[39] L. L. Freris, *Wind Energy Conversion Systems*. Upper Saddle River, NJ, USA: Prentice-Hall, 1990.

[40] R. A. Armstrong, S. V. Slade, and F. Eperjesi, "An introduction to analysis of variance (ANOVA) with special reference to data from clinical experiments in optometry," *Ophthalmic Physiol. Opt.*, vol. 20, no. 3, pp. 235–241, 2000.

[41] R. A. Armstrong, F. Eperjesi, and B. Gilmartin, "The application of analysis of variance (ANOVA) to different experimental designs in optometry," *Ophthalmic Physiol. Opt.*, vol. 22, no. 3, pp. 248–256, 2002.

[42] Ø. Langsrud, "ANOVA for unbalanced data: Use type II instead of type III sums of squares," *Statist. Comput.*, vol. 13, no. 2, pp. 163–167, 2003.

[43] K. M. Ting, "Confusion matrix," in *Encyclopedia of Machine Learning*. Boston, MA, USA: Springer, 2011, p. 209.

NEERAJ DHANRAJ BOKDE received the M.E. degree in embedded systems from the EEE Department, BITS Pilani, Pilani Campus, India, in 2014, and the Ph.D. degree from the Electronics and Communication Department, Visvesvaraya National Institute of Technology, Nagpur, India. He is currently a Postdoctoral with the Department of Engineering, Renewable Energy and Thermodynamics, Aarhus University, Denmark. His research interests include time series analysis, data mining, and prediction methodologies.

ANDRÉS FEIJÓO received the M.Sc. degree in electrical engineering from the Universidade de Santiago de Compostela, Spain, in 1990, the Ph.D. degree in electrical engineering from the Departamento de Enxeñería Eléctrica, Universidade de Vigo, Spain, in 1998, and the B.Sc. degree in mathematics from the Universidad Nacional de Educación a Distancia, Spain, in 2018. He is currently an Associate Professor with the Departamento de Enxeñería Eléctrica, Universidade de Vigo. His research interests include the field of renewables and especially wind energy.

NADHIR AL-ANSARI received the B.Sc. and M.Sc. degrees from the University of Baghdad, in 1968 and 1972, respectively, and the Ph.D. degree in water resources engineering from Dundee University, in 1976. He was with Baghdad University, from 1976 to 1995, and with Al-Bayt University, Jordan, from 1995 to 2007. He served as several academic administrative post, such as the Dean and the Head of Department. He has supervised more than 66 postgraduate students at Iraq, Jordan, U.K., Australia universities. He is currently a Professor with the Department of Civil, Environmental and Natural Resources Engineering, Luleå University of Technology, Sweden. He executed more than 60 major research projects in Iraq, Jordan, and U.K. He has published more than 424 articles in international/national journals, chapters in books, and 13 books. He holds one patent on physical methods for the separation of iron oxides. His research interests include geology, water resources, and environment. He was a member of several scientific societies, e.g., the International Association of Hydrological Sciences, the Chartered Institution of Water and Environment Management, the Network of Iraqi Scientists Abroad, and the Founder and President of the Iraqi Scientific Society for Water Resources. He was also a member of the Editorial Board of ten international journals. He was a recipient of several scientific and educational awards, among them is the British Council on its 70th Anniversary awarded him top five scientists in Cultural Relations.

ZAHER MUNDHER YASEEN received the master's and Ph.D. degrees from the National University of Malaysia (UKM), Malaysia. He has an Excellent Expertise in machine learning, advanced data analytics, and environmental sciences. He is currently a Senior Lecturer and a Senior Researcher in civil engineering with Ton Duc Thang University. He has published more than 80 articles in international journals with a Google Scholar H-Index of 17 and a total of 921 citations. His major research interests include hydrology engineering, water resources engineering, hydrological processes modeling, and environmental engineering and climate.

· · ·