An Effective Wind Power Prediction using Latent Regression Models

1,2 Benamar Bouyeddou, 3 Fouzi Harrou, 4 Ahmed Saidi, 3 Ying Sun
1 Abou Bekr Belkaid University, STIC Lab., Department of Telecommunications, Tlemcen, Algeria
2 University of Saida-Dr Moulay Tahar, Department of Electronics, Faculty of Technology, Saida, Algeria
3 King Abdullah University of Science and Technology, CEMSE Division
Thuwal, 23955-6900, Saudi Arabia, fouzi.harrou@kaust.edu.sa
4 Ahmed Draia University, Department of Science and Technology, Laboratory of Sustainable Development and Computer Science, Adrar 01000, Algeria

Abstract—Wind power is considered one of the most promising renewable energies. Efficient prediction of wind power will support in efficiently integrating wind power in the power grid. However, the major challenge in wind power is its high fluctuation and intermittent nature, making it challenging to predict. This paper investigated and compared the performance of two commonly latent variable regression methods, namely principal component regression (PCR) and partial least squares regression (PLSR), for predicting wind power. Actual measurements recorded every 10 minutes from an actual wind turbine are used to demonstrate the prediction precision of the investigated techniques. The result showed that the prediction performances of PCR and PLSR are relatively comparable. The investigated models in this study can represent a helpful tool for model-based anomaly detection in wind turbines.

Index Terms—Power prediction, Wind turbine, regression methods, Latent variable models.

I. INTRODUCTION

Wind power production’s main critical and challenging problem is its intermittent volatility, mainly to weather conditions [1], [2]; making the integration of wind turbines into the power grid not an easy task [3], [4]. Hence, accurately predicting wind power is of great interest to cope with the impacts of wind power fluctuation on power system operation. Furthermore, it is helpful for safety and managing the power grid, delivery, storage. Accurate wind power prediction plays a core role in efficiently and safely integrating wind power in a power grid. Essentially, the necessity for a precise wind power prediction has been raised with the increase of installed capacity. Thus, several prediction methodologies have been designed in the literature and can generally be grouped into two classes that are physical and statistical-based methods [5]. Generally speaking, the essence of physical models consists in the application of atmospheric motion equations to estimate future trends of meteorological measurements, then forecasting wind power by using some forecasted meteorological variables (e.g., wind speed) [6], [7]. Importantly, a physical model is generally performed using numerical weather estimation, which is implemented into two steps: prediction of wind speed, then transformation of this wind speed to wind power [8]. This mechanism is usually accomplished by using wind power curve which could be modeled using parametric or non-parametric methods [9]. Although physical models provide acceptable features in predicting the long-term trend of wind variance, they are costly to develop and time-consuming and provide low prediction accuracy for a local area. For more details refer to [9]. On the other hand, statistical models use data mining methods to construct a model that expresses the interaction between wind power and other input variables [10]–[12].

Over last decades, several researches focused on applying artificial intelligence techniques for wind power prediction [13]–[15]. In [13], the k-nearest neighbor classifier (kNN) is used for predicting wind power in short-term horizon using multi-tupled meteorological input measurements. Based on genetic programming and ensemble of neural networks, authors in [15] proposed a robust approach to predict wind power. In li2020wind, a wind power prediction method is proposed by combining the advantages of wavelet decomposition, SVM, and atomic search algorithm. In this hybrid approach, wavelet decomposition has been used to preprocess the input wind power measurement, and the Atomic search algorithm is applied to optimize the SVM for improved wind power prediction. The study in li2021improving designed an optimized SVR model to achieve more reliable wind power prediction. Importantly, in this approach, cuckoo search arithmetic-based optimization is used to determine the SVR model’s optimal hyper-parameters. Data collected from a French wind farm has been used to show the prediction efficiency of this approach. In [16], a method integrating a data mining technique and enhanced SVM algorithm is introduced for a short-term of wind power. This method used data mining to explore the correlation linking wind powers and wind speed and correct the invalid original data. However, this method is not verified for long-term prediction [16]. The authors in [17] propose an approach for forecasting wind power based on machine learning and kernel density estimation. This approach use information from nearby measurements and meteorological forecast values to enhance prediction precision. In [18], ensemble learning-driven methods, including Random Forest (RF) and Boosted Trees, have been implemented to predict wind power. Measurements from four wind turbines placed in
Turkey and France have been adopted to verify the prediction accuracy of these techniques. Results reveal the suitable performance of the ensemble methods in predicting wind power and their superior performance compared to Support Vector Regression and Gaussian process regression. In addition, the findings show that incorporating lagged variables improves the prediction accuracy of the ensemble models. In [19], a hybrid generative adversarial network (GAN)-driven model is proposed for short-term wind power forecasting. Specifically, the prediction problem has been addressed as a min-max game using GAN. The authors in shahid2020novel introduced an approach to predict wind power by embedding wavelet kernels with long short-term memory (LSTM). This approach, called WN-LSTM, aims to catch the dynamic behavior of wind power measurements. Prediction results using actual data from wind farms in Europe recommends using WN-LSTM compared to other existing machine learning techniques. Recently, a deep learning-driven approach called staked independently recurrent autoencoder (IRAЕ) has been presented in [20] to effectively predict wind power. At first, the variational mode decomposition approach is applied to decompose data into subsequences to reduce noise in data. Then, IRAЕ is utilized for wind power prediction. This approach exhibited better prediction compared to some commonly used prediction models. The authors in [21] proposed an autoregressive dynamic adaptive (ARDA) model to predict wind power in real-time. This ARDA approach is dynamically adapted to catch data variations. This approach exhibited enhanced performance compared to ARIMA and LSTM models.

Precise prediction of wind power is crucial to sustainably integrate the wind power in a power grid. The overarching aim of this work is to design an approach enabling an efficient forecast of wind power production using times series SCADA data from a wind turbine. This paper investigated the performance of two latent variables regression methods, i.e., partial least squares (PLS) and principal components regression (PCR) [22]–[24]. To the best of our knowledge, these latent variables regression (LVR) models have not been widely exploited for wind power prediction. Multivariate data-derived models, e.g., PCR and PLS, are frequently implemented to correlated multivariable by performing regression on a reduced number of uncorrelated variables (i.e., latent variables), which consists of linear combinations of the original variables [25]–[27]. When implementing LVR models, only fewer latent variables are utilized instead of using raw measurements. Generally, LVR models result in well-conditioned parameter estimations and reliable model predictions. One of the main contributions of this paper is bringing this LVR method to the attention of the renewable energy community and show how it can be applied for predicting wind power. Real measurements recorded every 10 minutes from an actual wind turbine are adopted to show the prediction quality of the investigated techniques.

The wind power data used in this study is described in Section II. Then, the regression models are briefly reviewed in Section III. In Section IV, the prediction results are discussed. Lastly, conclusions are presented in Section V.

II. WIND TURBINE DATASETS

The wind power data studied in this article is gathered from a high-wind-speed wind turbine Senvion MM82 in France. The principal characteristics of this wind turbine are given in Table I. The hub heights of this wind turbine of 80 meters make it a choice for sites with height restrictions (Table I). The dataset is recorded every ten minutes. The recorded measurements contain twelve input variables and the active power as the response variable.

<table>
<thead>
<tr>
<th>TABLE I WIND TURBINE MAIN CHARACTERISTICS.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Senvion MM82</strong></td>
</tr>
<tr>
<td>Rated power</td>
</tr>
<tr>
<td>Rotor diameter</td>
</tr>
<tr>
<td>Hub height</td>
</tr>
<tr>
<td>Number of blades</td>
</tr>
<tr>
<td>Rotor speed, max</td>
</tr>
<tr>
<td>Generator Speed, max</td>
</tr>
<tr>
<td>Cut-in wind speed</td>
</tr>
<tr>
<td>Rated wind speed</td>
</tr>
<tr>
<td>Cut-out wind speed</td>
</tr>
</tbody>
</table>

The training measurements gathered from September 1st, 2013 to May 14, 2014, is employed for constructing the prediction models. Table II presents summary statistics of the training dataset. Table II indicates that several variables were negatively skewed with relatively large kurtosis.

III. METHODS

This section describes the prediction methodologies PLSR and PCR used for predicting wind power.

A. Principal Component Regression (PCR)

The PCR approach is implemented into two stages: first, we apply a principal component analysis on the matrix of the explanatory variables X. Then, Ordinary Least Squares regression is applied to link the retained principal component and the response variable [28], [29] (Figure 1).

By using PCA, the data matrix X can be expressed as a sum of the approximated matrix, Ū, and residual data, E.

$$ X = TW^T = \sum_{i=1}^{k} t_i w_i^T + \sum_{i=k+1}^{m} t_i w_i^T = \hat{X} + E $$  \hspace{1cm} (1)

where \( T \in \mathbb{R}^{n \times m} \) denotes a matrix of the principal components (PCs) and \( W \in \mathbb{R}^{m \times m} \) refers to the loading matrix. In the case of cross-correlated data, X, the first ‘k’ PCs (where \( k < m \)) are enough to preserve pertinent information in the input data. Here cumulative percentage variance (CPV) procedure is utilized to select the number of PCs to retain in the model [30]. Let \( \hat{T} \in \mathbb{R}^{n \times l} \) is the matrix of the \( k \) retained
TABLE II
SUMMARY STATISTICS OF THE TRAINING MEASUREMENTS.

<table>
<thead>
<tr>
<th></th>
<th>mean</th>
<th>std</th>
<th>min</th>
<th>Q.25</th>
<th>Q.5</th>
<th>Q.75</th>
<th>max</th>
<th>skewness</th>
<th>kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wind speed (m/s)</td>
<td>5.697</td>
<td>2.512</td>
<td>2.31</td>
<td>4.29</td>
<td>5.70</td>
<td>7.12</td>
<td>19.44</td>
<td>0.27</td>
<td>3.88</td>
</tr>
<tr>
<td>Torque (Nm)</td>
<td>2206.73</td>
<td>2310.09</td>
<td>-734.41</td>
<td>350.24</td>
<td>1625.31</td>
<td>3223.94</td>
<td>10875.70</td>
<td>1.37</td>
<td>4.72</td>
</tr>
<tr>
<td>Generator converter speed (rpm)</td>
<td>1137.29</td>
<td>596.29</td>
<td>1.86</td>
<td>971.80</td>
<td>1257.59</td>
<td>1640.26</td>
<td>1804.78</td>
<td>-0.81</td>
<td>2.43</td>
</tr>
<tr>
<td>Generator speed (rpm)</td>
<td>1135.81</td>
<td>596.57</td>
<td>-0.07</td>
<td>970.07</td>
<td>1256.11</td>
<td>1639.08</td>
<td>1803.15</td>
<td>-0.81</td>
<td>2.43</td>
</tr>
<tr>
<td>Converter torque (Nm)</td>
<td>2342.14</td>
<td>2265.35</td>
<td>-9.32</td>
<td>641.63</td>
<td>1838.41</td>
<td>3288.11</td>
<td>10915.70</td>
<td>1.36</td>
<td>4.82</td>
</tr>
<tr>
<td>Rotor speed (rpm)</td>
<td>10.83</td>
<td>5.70</td>
<td>0</td>
<td>9.23</td>
<td>11.96</td>
<td>15.64</td>
<td>17.21</td>
<td>-0.81</td>
<td>2.42</td>
</tr>
<tr>
<td>Pitch angle (deg)</td>
<td>11.01</td>
<td>25.23</td>
<td>-37.06</td>
<td>-0.98</td>
<td>-0.95</td>
<td>0.11</td>
<td>92.68</td>
<td>2.10</td>
<td>6.38</td>
</tr>
<tr>
<td>Gearbox oil sump temperature (°C)</td>
<td>52.95</td>
<td>8.84</td>
<td>11.47</td>
<td>52.08</td>
<td>56.14</td>
<td>57.89</td>
<td>63.16</td>
<td>-2.53</td>
<td>9.69</td>
</tr>
<tr>
<td>Absolute wind direction corrected (deg)</td>
<td>172.43</td>
<td>100.22</td>
<td>0.01</td>
<td>66.39</td>
<td>190.51</td>
<td>249.15</td>
<td>359.99</td>
<td>-0.11</td>
<td>1.84</td>
</tr>
<tr>
<td>Rotor bearing temperature (°C)</td>
<td>27.13</td>
<td>6.18</td>
<td>5.07</td>
<td>23.53</td>
<td>27.94</td>
<td>31.30</td>
<td>41.51</td>
<td>-0.57</td>
<td>3.25</td>
</tr>
<tr>
<td>Generator bearing 1 temperature (°C)</td>
<td>39.70</td>
<td>7.24</td>
<td>6.09</td>
<td>37.61</td>
<td>41.40</td>
<td>44.28</td>
<td>61.56</td>
<td>-1.78</td>
<td>6.87</td>
</tr>
<tr>
<td>Generator bearing 2 temperature (°C)</td>
<td>39.17</td>
<td>7.36</td>
<td>4.28</td>
<td>37.23</td>
<td>40.73</td>
<td>43.66</td>
<td>60.52</td>
<td>-1.90</td>
<td>7.30</td>
</tr>
<tr>
<td>Active power (kW)</td>
<td>564.01</td>
<td>626.66</td>
<td>-12.38</td>
<td>79.15</td>
<td>320.11</td>
<td>873.02</td>
<td>2266.90</td>
<td>1.20</td>
<td>3.34</td>
</tr>
</tbody>
</table>

Fig. 1. A conceptual representation of PCR model.

principal components. In the PCR model, the linear regression between \( \hat{T} \) and the response variable \( y \) is obtained via solving the optimization problem (2).

\[
\hat{\beta} = \arg \min_{\beta} \left( \| \hat{T} \beta - y \|_2^2 \right)
\]  

(2)

The least squares solution is expressed as:

\[
\hat{\beta} = \left( \hat{T}^T \hat{T} \right)^{-1} \hat{T}^T y.
\]  

(3)

**B. Partial Least Square (PLS)**

Consider the matrix of the explanatory variables \( X \in \mathbb{R}^{n \times m} \) and the response matrix \( Y \in \mathbb{R}^{n \times p} \). The core idea of PLS is extracting the latent variables iteratively via the maximization of the covariance of the extracted latent variables. PLS comprises two models: the inner model and outer model [31]–[33] (Figure 2). The outer model is given as:

\[
\begin{align*}
X &= \sum_{i=1}^{l} tp_i^T = TP^T + G \\
Y &= \sum_{i=1}^{q} uq_i^T = UQ^T + F,
\end{align*}
\]  

(4)

where, \( T \in \mathbb{R}^{n \times l} \) and \( U \in \mathbb{R}^{n \times q} \) denotes a matrix of the transformed uncorrelated variables. The loading matrices of input and output space are \( P \in \mathbb{R}^{m \times l} \) and \( Q \in \mathbb{R}^{p \times q} \), respectively. Here, \( G \) and \( F \) refer to the model residuals. The cross-validation procedure is applied to determine the number of PCs, \( l \). The maintained PCs of the input and output space, i.e., \( T \) and \( U \), are linked through the following linear inner model:

\[
U = T\beta + E,
\]  

(5)

where \( \beta \) is a regression matrix, and \( E \) is a residual part. The input \( Y \) can now be expressed as:

\[
Y = T\beta Q^T + F^*.
\]  

(6)

The construction of the PLS model is performed in a sequential manner [34], [35].

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The wind power prediction based on regression models (e.g., PCR and PLSR) is performed into three stages as summarized in Figure 3. Using training data, prediction models
have been constructed. Here, we used 5-fold cross-validation to train the models and avoid overfitting. Then, these models are employed for predicting the response based on unseen input data. The quality of the predicted is judged by computing R-squared ($R^2$), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

\[
MAE = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t|, \tag{7}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2}, \tag{8}
\]

\[
R^2 = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - mean(Y))^2}. \tag{9}
\]

A. Preliminary data analysis

Generally speaking, wind turbines are constructed to work within an interval of wind speeds and with maximum power. This power value is the nominal value of the generator, but it is rarely achieved. The operation of the wind turbine relies on wind speed. This wind speed will constantly vary, as shown in Figure 4(a), which represents the evolution of wind speed and power production from the 2.05 MW Senvion MM82 wind turbine. The wind power curve (Figure 4(a)) is usually used to detect abnormalities in wind turbines by comparing the empirical wind power curve with the theoretical one provided by the constructor. Figure 4(b) illustrates the plot of rotor speed against wind speed. One can see that the rotor speed is increasing in accordance with wind speed until reaching the cut-in wind speed. Another important curve to characterize the operating condition of a wind turbine is the pitch curve which plots wind speed against pitch angle (Figure 4(c)). Similar to the power curve analysis, deviations in characteristic operation in the pitch curve can be identified, and further analysis for anomaly detection can be performed. Figure 4(c) displays a wind speed variation as a function of blade pitch angle degree under normal conditions.

B. Models design

The training measurements are utilized to design prediction methods. Here, we used the 5-fold cross-validation procedure to find the model parameters. One crucial step in PCR and PLSR models consists in the selection of the number of PCs. To this end, the cumulative percentage variance (CPV) method is adopted because of its simplicity and efficiency. Figure 5(a-b) shows respectively the CPV explained in the input data $X$ for both PLSR and PCR, as well as the CPV explained in the output for the PLSR model. From Figure 5(a-b), it can be seen that two PCs are sufficient to describe around 99% of the variability in $X$ in the two models. We used two PCs to build the PCR model, while three PCs are used for the PLSR model.

Table III summarizes the prediction accuracy of the PLSR and PCR models when using the testing dataset. As shown in Table III and Figure 6, the measured wind power prediction match well with the predicted power from the PLSR and PCR models.

<table>
<thead>
<tr>
<th></th>
<th>$R^2$</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCR</td>
<td>0.949</td>
<td>146.102</td>
<td>101.601</td>
</tr>
<tr>
<td>PLSR</td>
<td>0.949</td>
<td>146.007</td>
<td>101.577</td>
</tr>
</tbody>
</table>

The prediction results of the two models based on testing data are illustrated in Figure IV and Table V. The results in Table III indicate that the prediction performances of PCR and PLSR are relatively comparable. Also, this result confirms that the LVR models are relatively appropriate in predicting wind power.

V. Conclusion

A reliable prediction of wind power production may be a tool for facilitating wind turbine health monitoring. This paper presented a comparative study between two latent variable regression approaches (i.e., PCR and PLSR) for predicting wind power using SCADA dataset. The data used to verify the quality of these two methods has been collected from a 2.05 MW Senvion MM82 wind turbine. The conducted experiments demonstrate the PCR and PLSR achieve acceptable prediction performance of wind power. These two models achieved an $R^2$ of 93% for predicting wind power. This study demonstrated the feasibility of using LVR methods to predict wind power.

As future work, we plan to use these models to develop statistical monitoring schemes able to detect faults in wind tur-
bines. Specifically, LVR models will be combined with monitoring charts, such as CUSUM and generalized likelihood tests to detect sensors and process anomalies in wind turbines [22]. LVR models will generate residuals that monitoring charts will check to uncover potential anomalies in the inspected wind turbine [36].

ACKNOWLEDGEMENT

This publication is based upon work supported by King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR) under Award No: OSR-2019-CRG7-3800.

REFERENCES


