Wind Power Generation Prediction on a Large Real Life Dataset

Roshil Paudyal\textsuperscript{1}, Mugizi Robert Rwebangira\textsuperscript{2}, Mandoye Ndoye\textsuperscript{3}, roshil.paudyal@bison.howard.edu, rweba@scs.howard.edu, mndoye@mytu.tuskegee.edu

\textsuperscript{1}Department of Mathematics, Howard University, Washington, D.C. 20059 USA
\textsuperscript{2}Systems and Computer Science, Howard University, Washington, D.C. 20059 USA
\textsuperscript{3}Department of Electrical Engineering, Tuskegee University, Tuskegee, AL 36088

Abstract—The intermittent nature of wind power generation is becoming more of a problem as the percentage of wind energy used in the grid is increasing. We propose a data driven approach using machine learning methods to predict daily wind power generation output. A novel aspect of this work is the verification of the algorithm on a massive dataset with more than 500,000 observations.

Keywords—Wind Power; Renewable energy; Machine Learning; Logistic regression

I. INTRODUCTION

Wind power accounts for a relatively small percentage of energy in the US distribution grid despite the fact that it constitutes one of the most promising sources of clean and renewable energy. When the percentage of wind energy in the grid is negligible, control room operators can schedule these resources without facing any serious issues. However, the percentage of wind energy usage within the grid has been steadily increasing, and there is now a pressing need for more accurate forecasts of wind power production, which can then be exploited to make better informed scheduling decisions. The integration of wind energy in the power grid is a very difficult task because of its intermittent nature. One of the main challenges is the lack of predictability of the amount of power from wind turbines. This increases the spinning reserve requirements and unanticipated ramp events, causing elevated production costs and decreased reliability. Accurate and reliable methods for forecasting wind power generation are essential if wind power is to become a staple in countries energy diet. In this work we address this problem. One of our main contribution is that we are able to verify our algorithms on a dataset with 500,000 observations, which is far larger than most previous work.

II. RELATED WORK

Several researchers have proposed algorithms for wind power forecasting. Just to mention a few of them, Mabel and Fernandez [6] proposed using artificial neural networks [ANN] for wind power prediction. Their data covers only a 2 year time span. Świątek and Dutka [4] also propose a neural network approach. Their data covers less than a 2 year time period. Finally Lei et. al [7] also give a comprehensive bibliography of various approaches. Many of them were preliminary and only used limited data or were test runs.

III. THE DATA

Our data, retrieved from the Bonneville Power Administration [4] has one of the largest observations of weather and power output available. The data records Barometric Pressure, Humidity, Temperature, Wind Speed, Wind Direction, Peak Wind Speed and Peak Wind Direction at the station at each 5 minute interval from 01/01/2009 to 01/01/2014 or a period of more than 5 years.

Below is an example of what the data looks like:

We also have the corresponding power output at each time period.

IV. ALGORITHM

As a preliminary step, we are just trying to predict the wind power output as high or low value using a logistic regression model. In order to train the model, we convert a continuous-valued power output variable to a binary one. We use the median of the power output as a criteria to judge the power output values as high or low, i.e. a value higher than the median (1063 MW) was given the binary value 1, and a value lower than that was given the binary value 0.

We then used Pressure, Humidity, Temperature and Wind Speed as predictors and trained the logistic model to output a binary high/low value given those parameters. One of the issues with the data was that while it contained individual files describing the weather at individual stations, the power output provided was the sum of power output by all stations, which made it challenging to assess the contributions of the individual stations to the overall power output. So we devised a step-by-step approach, which first takes the weather data at a single station as a predictor, and then averages the weather
data across all the stations, and takes that as a predictor of high or low power output.

V. RESULTS

In Fig 1 we show the output of logistic regression when we use a randomly chosen station’s weather data as the predictor.

Especially interesting to note is the correlation of the various weather indicator’s with the power output. As we expect the correlation is relatively small, and even negative for pressure.

Then we look at the raw accuracy of the prediction, again with the weather at one randomly chose station:

<table>
<thead>
<tr>
<th></th>
<th>Predicted high</th>
<th>Predicted low</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed high</td>
<td>4890</td>
<td>2707</td>
<td>64.36</td>
</tr>
<tr>
<td>Observed low</td>
<td>7727</td>
<td>5224</td>
<td>59.67</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td>61.40</td>
</tr>
</tbody>
</table>

We immediately note that the correlation is much higher between wind speed and power output, which is what we’d expect.

Looking at the raw accuracy of the predictions:

<table>
<thead>
<tr>
<th></th>
<th>Predicted high</th>
<th>Predicted low</th>
<th>% Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed high</td>
<td>5502</td>
<td>2101</td>
<td>72.36</td>
</tr>
<tr>
<td>Observed low</td>
<td>9909</td>
<td>3036</td>
<td>76.54</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td>75.00</td>
</tr>
</tbody>
</table>

The accuracy has increased from 61% when using only one station to 75%, which is state of the art in comparison with results reported in the literature.

VI. CONCLUSION

We can draw a couple of conclusions from these preliminary experiments.

(1) The massive amount of observations directly lead to more accurate prediction and are effectively utilized (in comparison with other researchers who had access to less data).

(2) Even though taking the average of the weather across different stations is not a perfect solution it still leads to remarkably improved prediction accuracy.
A more principled way of combining the weather observations from different stations is highly desirable. One way of doing this might be by taking an average of the weather variables weighted by their correlation with the power output.

REFERENCES


[5] Lawrence Livermore National Labs, Predicting Wind Power with Great Accuracy, Science and Technology Review, April/May 2014
