# A Wind Power Forecasting Problem

Gaurav Kapoor, Behnam Montazeri Najafabadi, Robert J. Yates

## INTRODUCTION

We plan to evaluate the Kaggle project 'A wind power forecasting problem: predicting hourly power generation up to 48 hours ahead at 7 wind farms'. The details of this project can be found at https://www.kaggle.com/c/GEF2012-wind-forecasting. This project has been sponsored by IEEE Power & Energy Society and a total prize pool of $7,500 have been dedicated to the first 3 teams. Moreover selected teams will be invited to IEEE PES General Meeting 2013 in Vancouver, Canada to present their methodologies and results.

Cristina Archer and Mark Z. Jacobson [1] estimated that the total wind energy that can be extracted in a practical and cost-competitive manner is between 72 and 170 TW but predicting future energy generation precisely for an specific wind power generation field is very important. The ability to predict generated wind power would help in proper design of the power network and achieving full compliance of customers energy demands. In this project, we are predicting power generation at seven wind farms. We are given the 48-hour ahead forecast for wind speed, wind direction, and the zonal and meridional wind components at each of the farms. The outputs are normalized wind power measurements between 0 and 1 for each of the seven wind farms. From 2009/07/01 to 2010/09/30 is a model identification and training period. From 2010/10/01 to 2010/12/31 is the evaluation period. In this period we are given generated power for the first 36 hours slots and we need to predict the power for the next missing 48-hour periods. This pattern of 72 hours (36 hours of given and 48 hours of missing data) is repeated over the entire evaluation period. Such predictions can be extremely challenging as the features change over time. The weather and climate forecasts are our main resource for predicting wind power generation despite the difficulties for meteorologists to create accurate weather forecasts. As a result, the collected the data has complicated features with sudden jumps that make it difficult to make accurate predictions for wind power generation.

## PROBLEM MODELING

We present detailed study of the predictive learning algorithms that are based on time-series regression and feature selection for the wind farms. Figure (1-a) shows the generated power at the wind farm on January 1st, 2010 and figure (1-b) shows the generated power at the wind farm over the month of January 2010. We developed four models to predict the generated power over the 48 hours periods. For our first two models, we use univariate series without using the given wind data (speed and direction) from the wind farms. However, we try to optimize our next two models by making use of this wind farm data.

Figure (2) shows the auto-correlation function (ACF) of the generated power at the wind farm. The ACF has a peak at the lag of one hour and not surprisingly decreases as lag-hour is increasing. Another important observation is that there are local maximas around lag-hour 24 and 48. This is expected because the generated power at each hour should be highly correlated to the same hour 1 day ago and 2 days ago.

### Plain off-the-shelf ARIMA models

We first notice that we can't use plain off-the-shelf ARIMA model [2]. We have to forecast data 48 periods ahead and a plain off-the-shelf ARIMA model such as the one given below will lead to a wide confidence interval and a poor prediction

$$y_t = \frac{\phi_0 + \phi_1 y_{t-1} + \ldots \phi_p y_{t-p}}{\theta_0 a_t + \theta_1 a_{t-1} + \ldots \theta_q a_{t-q}}$$

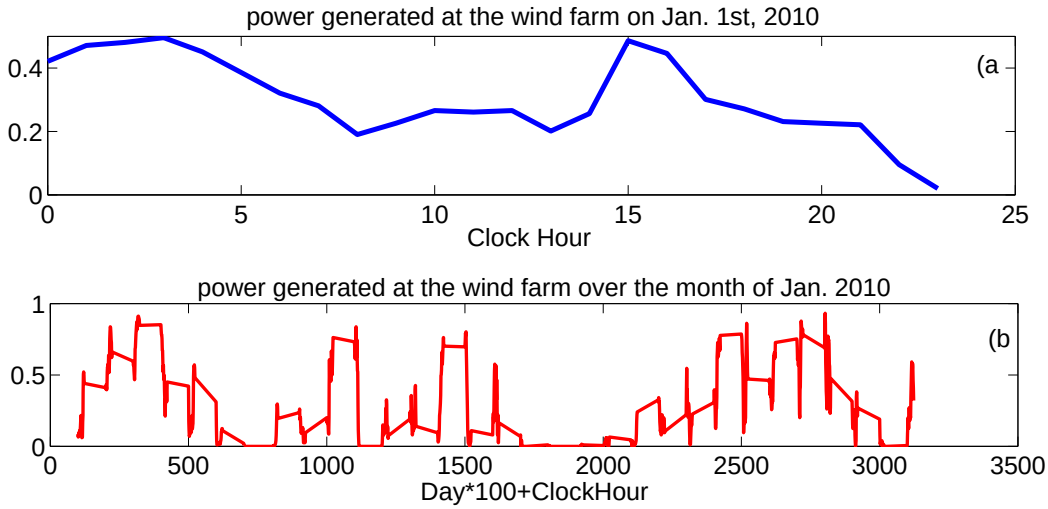


**FIGURE 1.** **a)** Generated wind power at the wind farm on January 1st, 2010. **b)** Generated wind power at the wind farm over January 2010.

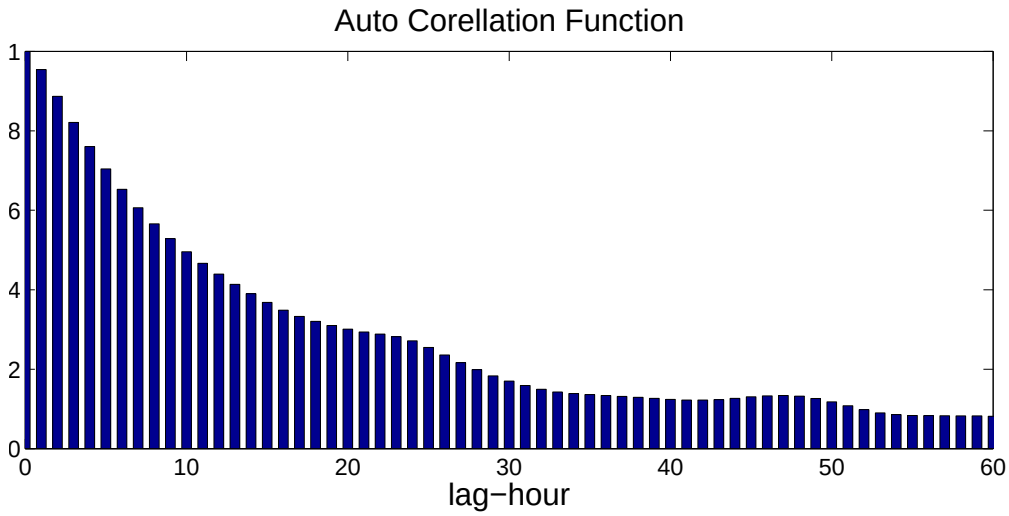


**FIGURE 2.** Auto-correlation function for the generated power.

Hence, we have to take advantage of inherent structure in the system and come up with a new model as described below: As mentioned above, we see that for any observation the maximum correlation is at lag 1 along with local maxima at lag 24 and lag 48. These observations can easily be used to forecast wind power for one period ahead. However, for forecasting two period ahead we will not use forecasted value at lag 1 as an input, rather we will make another model with last known value and 24-hours back value. Specifically,

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-24} \\ y_{t+l} &= \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t+l-24} \qquad 1 \le l \le 23 \\ y_{t+l} &= \phi_0 + \phi_1 \hat{y}_{t-24} + \phi_2 y_{t+l-48} \qquad 24 \le l \le 47 \end{aligned}$$

Thus, when we have to forecast one period ahead, we used the last value and the value at lag 24. When we need to forecast between period 2 and 24, we used the last **actual known** value (instead of the forecasted value at period 1) and the value with lag 24. This helps us to avoid propagation of 1-period ahead forecasted error and is not problematic

as the correlation between $y_t$ with its few lagged values is good as suggested by ACF plot. For forecasting period between 24 and 48 hours, we will use the forecasted value at period 24 and the actual value with 48 hours lag. This held us to take advantage of high correlation at lag 48 where we have known value instead of forecasted value.

**Regularization:** Instead of doing plain regression of $y_t$ on its lagged value $y_{t-1}, y_{t-24}, \ldots$, we decided to use regularized regression. The regularized regression is $\hat{\phi} = \arg\min_{\phi} ||y_t - \phi^T y_{t-l}||^2 + \frac{\lambda}{2}\phi^T\phi$ where the value of $\lambda$ is decided using cross validation; however, we decided to use $\hat{\phi} = \arg\min_{\phi} ||y_t - \phi^T y_{t-l}||^2$ subject to $0 \leq \phi_1 \leq C$ and $0 \leq \phi_2 \leq C$. The value of C was used as 1 and no penalty was added for intercept term.

## Locally Weighted Regression Model

The next possible model that we can think of for time series data prediction is Locally Weighted Regression [3] based on the generated power data itself,

$$\begin{aligned} y_{predicted} &= \frac{\sum_i w^{(i)} y^{(i)}}{\sum_i w^{(i)}} \\ w^{(i)} &= \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right) \end{aligned}$$

Where we calculate the predicted value at each time instance based on the value of 48 vectors. As we mentioned above, the data over each specific time period has correlation with the similar periods which is called the seasonality of the data. We can make use of this hidden characteristic of the data and make the locally wighted regression for the data and predict the power based on the similar instances of the data in the past. In the prediction period, we are given 36 hours of the generated power and we are asked to predict for the next 48 hours. Therefore, we divide the model identification data in chunk of 36+48 periods and use the similarity of first 36 period to forecast the next 48 period using locally wighted regression model. In order to find the right value for the bandwidth parameter, we use cross validation method to optimally compute it.

## ARIMAX Model with Wind Speed

Other than the historic time series data generated power data, we are provided with wind speed and direction features which can be used in training and prediction process. Specifically we are provided with four features wind: speed, direction, zonal and meridional components of the wind. As an important stepping stone, we need to intelligently choose the features that are closely correlated to that generated power (this process is feature selection). Since these features may have inner dependencies and using all of the features together may lead to over fitting. Therefore we looked into the correlation of wind features and generated power in the training set. We observed a high correlation between wind power and wind speed, however wind direction, zonal, meridional components were not found to have significant correlation to the power. So we include the wind speed as part of linear regression model with the new model as follows:

$$\begin{aligned} y_t &= \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-24} + \phi_3 ws_t \\ y_{t+l} &= \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t+l-24} + \phi_3 ws_{t+l} \qquad 1 \leq l \leq 23 \\ y_{t+l} &= \phi_0 + \phi_1 \hat{y}_{t-24} + \phi_2 y_{t+l-48} + \phi_3 ws_{t+l} \qquad 24 \leq l \leq 47 \end{aligned}$$

## Locally Weighted Regression with Wind Speed

Instead of relying only on univariate wind power, we now use forecasted wind speed, direction, zonal component, meridional component in our model. The locally weighted regression model automatically discards the features which

have low correlation to the generated power so we don't need to worry about the low correlated features. Again in this model we are using cross validation to find the bandwidth parameter. In this model $y_{predicted}$ formula stays the same, the only difference is the extra term in $w^i$ formula

$$w^{(i)} = \exp\left(-\frac{(x^{(i)}-x)^2}{2\tau^2}\right) + \theta \exp\left(-\frac{(s^{(i)}-s)^2}{2\tau^2}\right)$$

Where $x \in R^{36}$ vector of wind power $s \in R^{36}$ vector of wind speed $\theta = 0.5$.

## RESULTS AND DISCUSSIONS

Figure (3) shows the predicted power over a specific 48 hours period for wind farm 1 with four models.

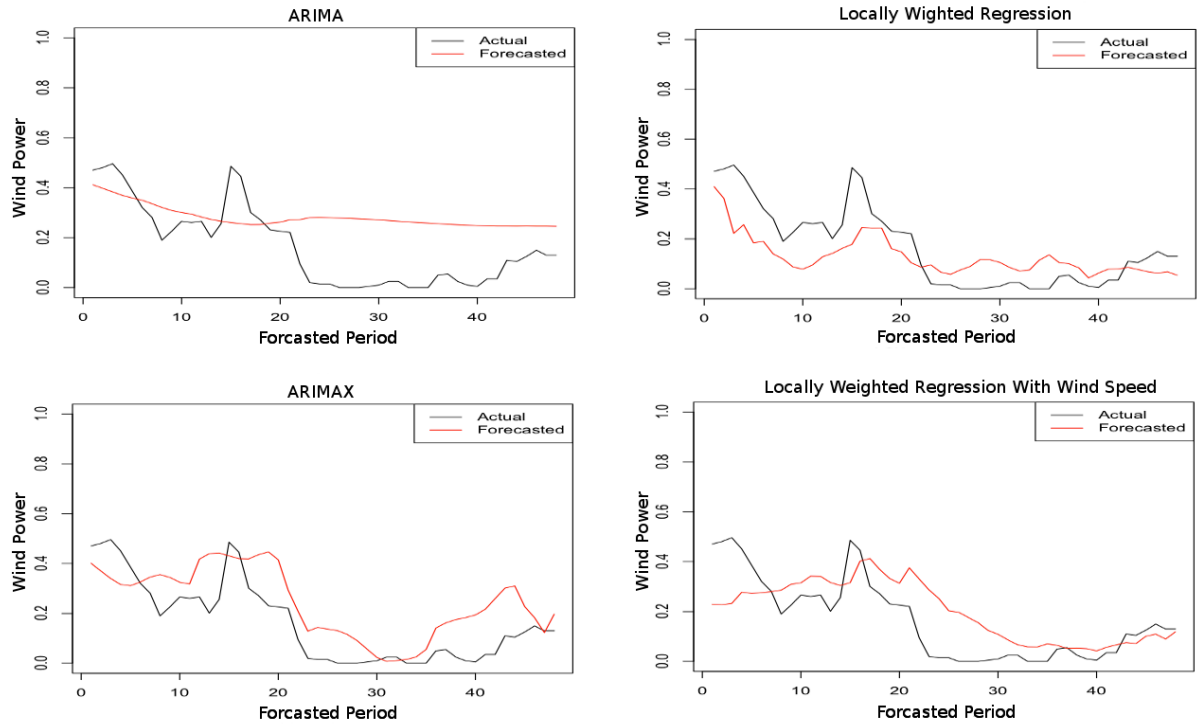


**FIGURE 3.** Predicted power vs. actual generated power over 48 hours period with four models.

The results for the ARIMA model show that this model is trying to smooth out the sudden jumps of the data and give an averaged out prediction. The prediction for the first 24 hours is quiet reasonable, but for the second 24 hours period the result is poor. The reason is that the model uses the first 24 hours predicted power to predict over the second 24 hours.

The cross validated locally weighted regression does a great job in predicting the sudden jumps and following the trend of this specific 48 hours period. However, there are few periods where this method drastically fails in prediction. The method tries to minimize the error over the given 36 hours which sometimes lead to a worse prediction for the next 48 hours since the weather conditions can dramatically change over these periods.

The predictions for the ARIMAX and Locally Weighted Regression with Wind Speed models are much closer to the actual data. We are making use of the the wind speed feature which is highly correlated (up to 80% correlation) to the generated power. These two models successfully predict the sudden peaks and valleys of the data over the time periods in a day that power generation are the highest and the lowest.

Figure (4) shows the root mean square error (RMSE) of the prediction from 2011/01/01 to 2012/06/28 for four models for all seven wind farms along with the best RMSE submitted to the Kaggle.com for this contest. As we can

see, performance of ARIMAX model is comparable to the best entries submitted to Kaggle. All four models give a better result than the result of the benchmark method currently being used for wind farm prediction.

RMSE Results

| Model | RMSE |
|---|---|
| Best Kaggle Entry | 0.15 |
| ARIMAX Model with Wind Speed | 0.19 |
| Locally Weighted Regression with Wind Speed | 0.21 |
| ARIMA Model | 0.26 |
| Locally Weighted Regression | 0.33 |
| Benchmark | 0.35 |

**FIGURE 4.** Root Mean Square Error of the predictions along with the result of the best entry submitted to Kaggle contest and benchmark result from the persistent prediction methods.

## SUMMARY AND CONCLUSION

In this project, we predicted wind power generation at seven wind farms. The model identification and training period is from July 1, 2009 to September 30, 2010. and the evaluation period for the plots is from October 1, 2010 to December 31, 2010 while the evaluation period for Kaggle submissions is from 2011/01/01 to 2012/06/28. In this period we are given generated power for the first 36 hour slots and we need tovpredict the power for the next missing 48 hour periods. As any other climate and meteoric related data, the wind power data is highly noisy and prediction on this type of data can be extremely challenging.

In this project we used two time series prediction models (ARIMA, Locally Weighted Regression) which make predictions based on the historic features of the generated power itself. As we saw that although the ARIMA model tries to average out the prediction, its overall performance is better than the Locally Weighted Regression model. The ARIMIA model trains itself based on the hourly generated power over all of the training set. In contrast, the Locally Weighted Regression gives more regression weight to the training instances that are similar to the 36 hour given data in the prediction intervals. However, this often leads to huge error in the prediction for Locally Weighted Regression.

We improved these models by adding wind speed features that we chose in our feature selection procedure through correlation to the generated power. These new ARIMAX and Locally Weighted Regression with Wind Speed models which had a quite reasonable performance improvement over ARIMA and Locally Weighted Regression and beat the benchmark method by almost a factor of 2.

## REFERENCES


1. A. Ananthaswamy, and M. Le, *New Scientist* (2012).
2. T. C. Mills, *Time Series Techniques for Economists*, Cambridge University Press, Publisher City, 1990.
3. A. Ng, *CS229 Lecture notes*, 2012.