One Day Ahead Wind Speed Prediction - Final Report

Team Name: Wind Predictor

Bo Yang (by57@cornell.edu)

Introduction

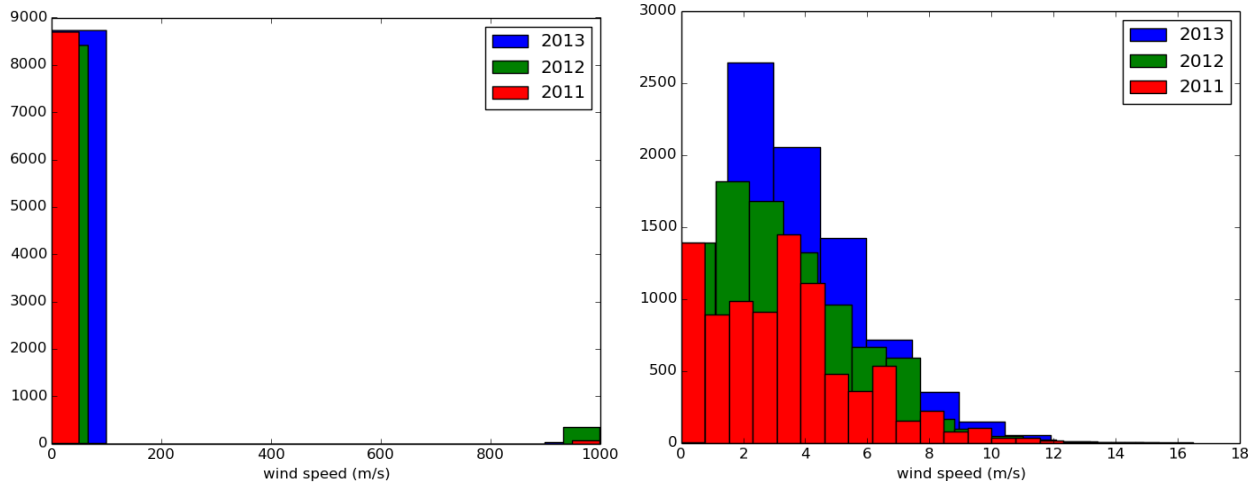
The goal of the project is to develop a statistical model to predict one day ahead wind speed.
Wind prediction is important for both energy and environment concerns. As a renewable energy, the main challenge in implementing wind power into the grid is the intermittency, which could bring a lot of problems such as market designs, capacity of transmission system, energy storage, and optimal reductions in greenhouse gas emissions. With regard to atmospheric environment, wind speed and other factors interact with the physical features of the landscape to determine the movement and dispersal of air pollutants, which are related to health and regulation issues. If future wind information could be accurately predicted, many of these problems could be overcome and/or addressed. For instance, a good wind speed forecast can help to develop well-functioning day-ahead energy markets. As for air quality, the potential impact on nearby regions caused by some certain air pollution sources at different locations can be evaluated by using future wind data.
Statistical approaches are generally good for short term and very short-term predictions (seconds to 6 hours ahead), but not so good for the medium (6 hours to 1 day ahead) and long term (1 day to 1 week or more ahead). On the other hand, the physical approach, Numeric Weather Prediction (NWP), can predict medium term and long term wind by solving complex physical-mathematical models with many data such as temperature, pressure, surface roughness, and buildings. But supercomputers are needed for this approach. And the simulations were run once or twice a day due to the difficulty of obtaining information in short time. Therefore, it will be very useful if good statistical methods for medium or even long term prediction could be developed because they are much cheaper and faster than the physical approach.
In order to achieve the project goal, the hourly wind data history of Ithaca airport will be used. In this final report, a simple Two-Year-Based model performance was firstly evaluated, which was proposed by El-Fouly et al. (2006). Secondly, a model was proposed and developed based on the previous days for comparison. Different loss functions and regularizers were tested. The third part of the report was to briefly introduce the widely used Auto-regression Moving Average (ARMA) model. Then model capabilities and limitations were discussed.

Data preview

Three years of wind speeds (Jan 1 2011 - Dec 31 2013) at the Ithaca Airport (ITH) were used to develop the one day ahead wind speed prediction models. Like other measurement database, there are missing data. Fortunately, shown from the Figure 1 (a), there are only tiny amount of missing data hours with regard to the whole year. Data of each year have similar histograms and Year 2012 has more missing data. In order to implement time series models, the missing data were filled by using the average value of the two nearest available data. The histogram after cleaning was shown in Figure 1 (b).

Data Features

The hourly wind speed data are classic time series signals. And it is generally difficult to assume a certain pattern. Unfortunately, a simple Auto-Regression model using past few hours data cannot be directly applied for this one day ahead prediction because there are 24 hours wind speeds need be predicted. Wind speeds depend on many factors including temperature, pressure, cloud, and terrain, which means, for some specific locations, wind speeds variation may have some sort of common trend over specific periods of time. These periods might be much longer than one hour.
Figure 2 showed a continuous 72 hours wind speeds at ITH in 2011 and 2013. The general trends are similar



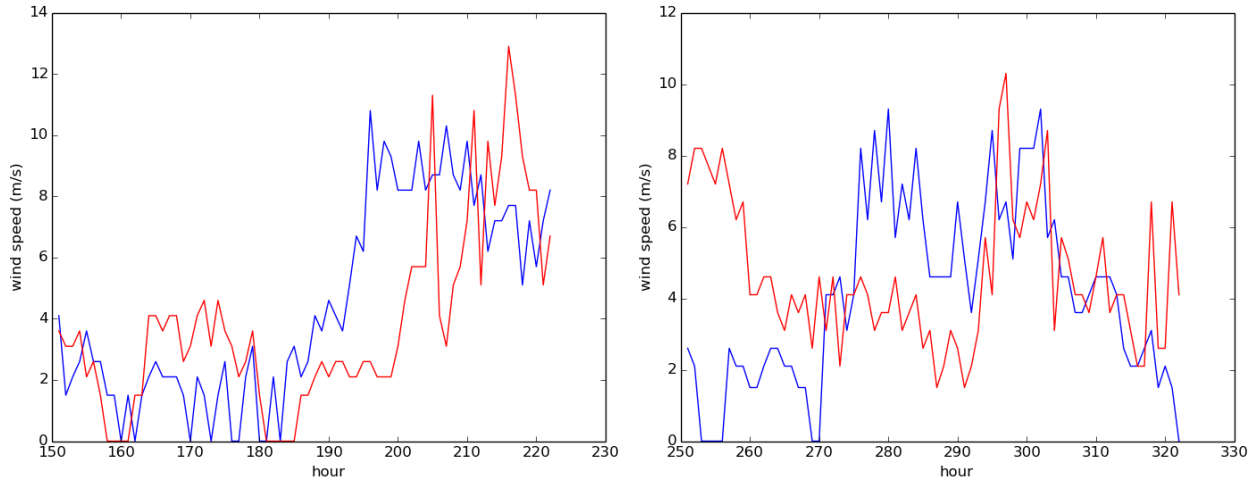
(a) With missing data

(b) Filling missing data

Figure 1: Histogram of wind speed of 2011, 2012, 2013 at Ithaca airport (ITH)

from hour 150 to hour 180, illustrated in Figure 2 (a). In Figure 2 (b), the trends of the two years are also similar from hour 295 to hour 315. The wind speeds at the same hours of the past two years could be used as features for a linear model. This is the basic idea of the Two-Year-Based model proposed by El-Fouly et al. (2006).

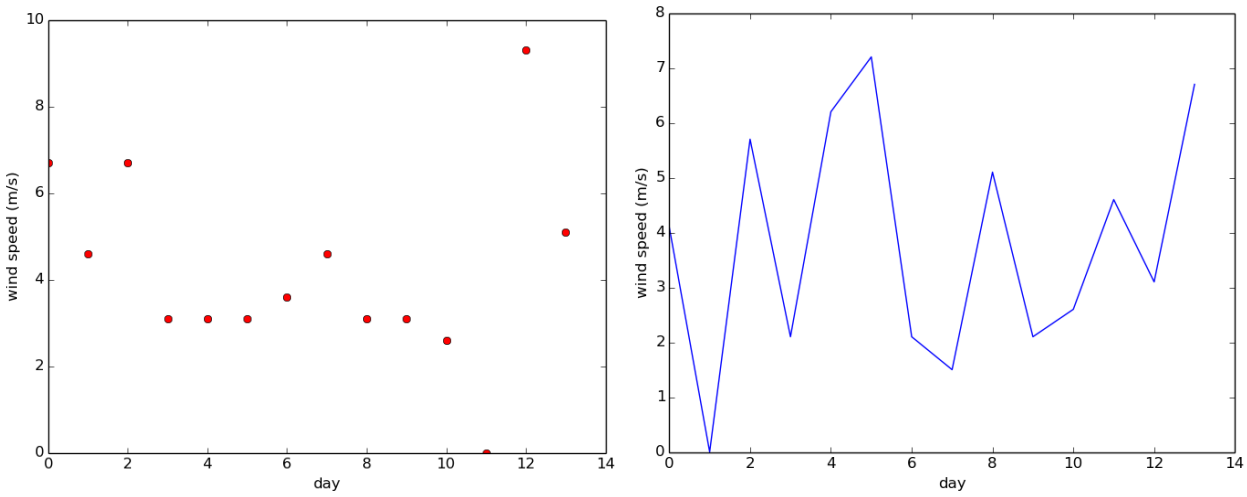
We may also look at the wind speeds at the same hour of continuous days. For example, the wind speeds at the same hour from Day 3 to Day 10 are similar, illustrated in Figure 3 (a). Also, the wind speed roughly oscillated around a certain value in Figure 3 (b). Then, the wind speeds at the same hours of several past days could be applied as features for a linear model. This is the main idea of the model developed in this project.



(a)

(b)

Figure 2: Wind speeds for 72 hours; Year 2011: blue lines; Year 2013: red lines



(a) Year 2011

(b) Year 2012

Figure 3: Wind speeds of the same hour on continuous days

Model 1: Two-Year-Based model

A simple Two-Year-Based model proposed by El-Fouly et al. (2006) uses data of n points from the current year and two previous years' series. The model is then used with the next n data points from the two previous year series to predict the next n points for the current wind speed series. TheTwo-Year-Based model can be represented by the following formula:

$$\hat{Y}_t(n+i) = w_1 X_{t-1}(n+i) + w_2 X_{t-2}(n+i) + w_3, (i = 1, 2, ..., n)$$

X_{t-1}, X_{t-2} are the corresponding wind speed values of the two past years. We try to predict the one day ahead wind speed, then $n = 24$.
The training set contains hourly wind speeds of 259 days, which were randomly sampled from one year. The wind speeds of the same 259 days were taken from the other two years. The training data were around 70% of one year. The training data set is slightly conservative, but we don't want to over-fit the model. Then, 53 days data (15% of one year) were randomly sampled from the rest of data to form a validation set, which could be used to decide good model(s). The rest of 53 days data form the test set.

Model 2: Two-Day-Based model

Similar to the Two-Year-Based model, the data of h points from the current day and two previous days. The model is then used with the next h data points from the two previous days to predict the next h points for the current wind speed series. In order to generate enough numbers of rows for the covariate matrix. Five continuous days data were used to fit the model. The Two-Day-Based model can be represented by the following formula:

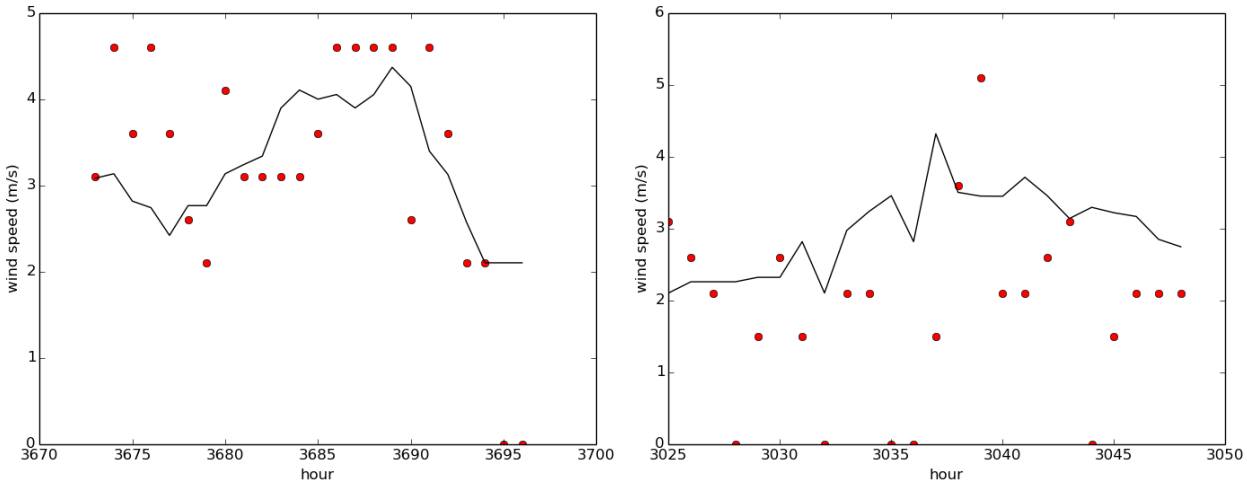
$$\hat{Y}_t(j) = w_{d1} X_{t-h}(j) + w_{d2} X_{t-2h}(j) + w_{d3}, (j = 1, 2, ..., 3h-1)$$

X_{t-h}, X_{t-2h} are the corresponding wind speed values of the two past days. We try to predict the one day ahead wind speed, then $h = 24$. Five days hourly data were used to form a 72 hours (rows) covariate matrix X. The first column of X is hourly wind speeds from $(t-h)$ to $(t+2h-1)$, and the second column is from $(t-2h)$ to $(t+h-1)$. The vector of $\hat{Y}$ is from t to $(t+3h-1)$. Only the last 24 data of $\hat{Y}$, from $(t+2h+1)$ to $(t+3h)$ were the prediction. This model configuration used Four previous days data from $(t-2h)$ to $(t+2h-1)$, but the covariate matrix columns were formed in the Two-Day base.
For each year, the training set contains hourly wind speeds of 51 groups of 255 days, which were randomly

sampled from that year. One group contains five continuous days. The training data were around 70% of one year. The training data set is slightly conservative, but we don't want to over-fit the model. Then, 11 groups (55 days) data (15% of one year) were randomly sampled from the rest of data to form a validation set, which could be used to decide good model(s). The rest of 11 groups (55 days) data form the test set.

Results of the Two-Year-Based model

From the training set, 259 linear models were obtained by using the quadratic loss function. The non-negative regularizer was also applied because the wind speeds are all positive real numbers. The root-mean-square error (RMSE) was calculated for each model by using the validation set. The best model, $w = [0.212, 0.103719, 2.09972]$, has the least average error (around 2.3 m/s) for the whole validation set (53 days). The standard deviation of this model's error of the whole validation set is around 0.9 m/s. Figure 4 showed the model performance for two days from the test set. For one day ahead wind speed prediction problem, this model is not too bad, but not good enough, either.



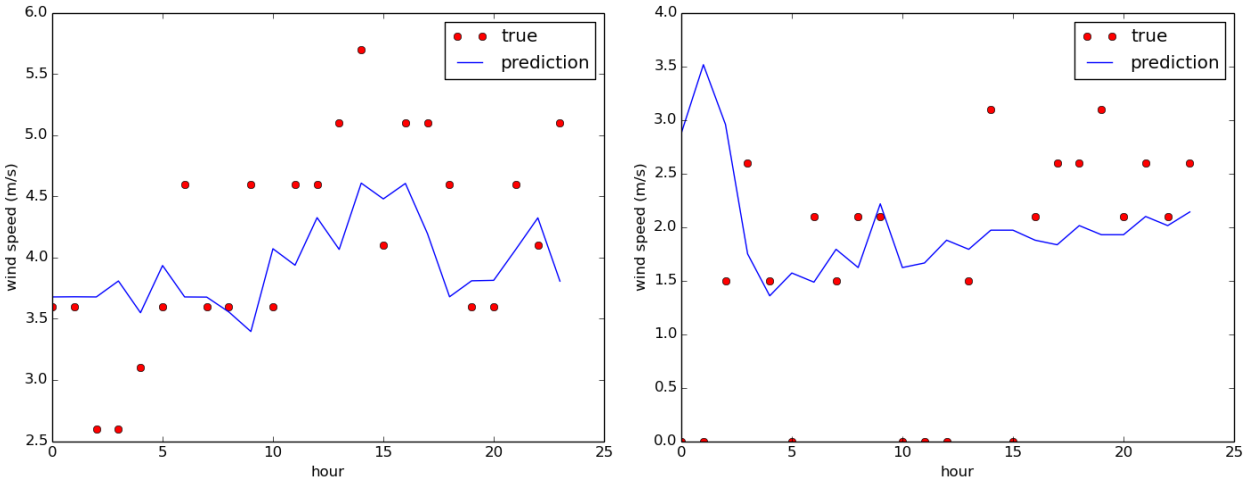
(a) (b)

Figure 4: Two-Year-Based model performance

Results of the Two-Day-Based model

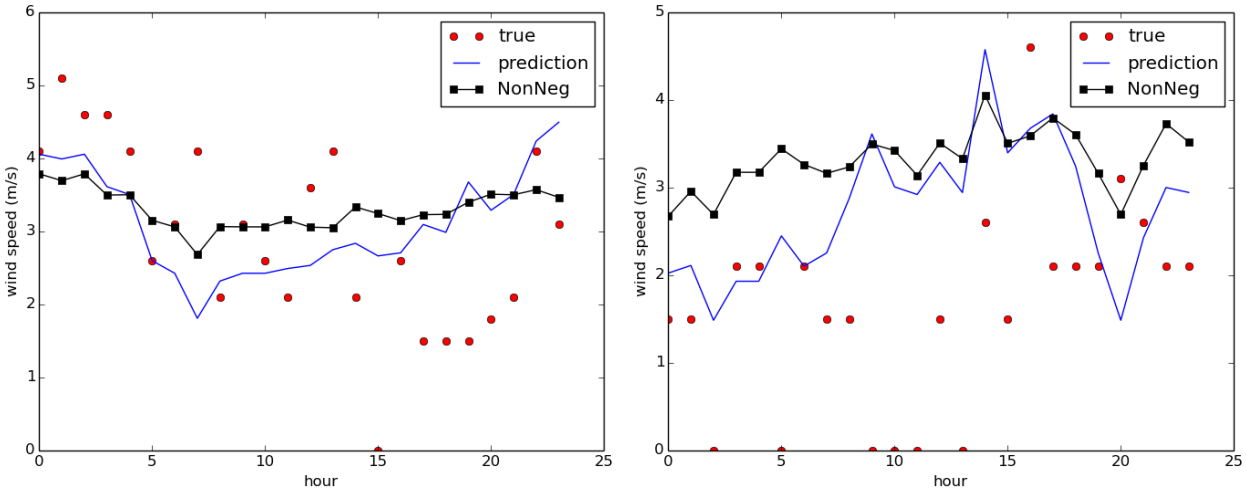
From the training set, 51 linear models were obtained by using both the quadratic loss and huber loss functions. Models were fitted by using the non-negative regularizer and without regularizer.
The root-mean-square error (RMSE) was calculated for each model by using the validation set. The best model using quadratice loss, without using regularizer, $w_d = [0.028, -0.089, 3.061]$, has the least average error (around 1.5 m/s) for the whole validation set (11 groups). The standard deviation of this model's error of the whole validation set is around 0.4 m/s. Figure 5 showed the model performance for two days from the test set. For one day ahead wind speed prediction problem, this model is generallly better than the Two-Year-Based model.
There is risk of getting negative wind speed prediction without using regularizers. By adding the non-negative regularizer, the best model using quadratic loss, $w_{nn} = [0.0, 0.147, 2.459]$, has the least average error (around 1.6 m/s) for the whole validation set . The standard deviation of this model's error of the whole validation set is around 0.5 m/s. Figure 6 showed the model performance is similar to the model without using regularizers, but the model can be more robust with regularizers.
From the wind speed histogram in Figure 1, there are some comparatively high wind speed hour/days, which means other loss function. And by using Huber loss with the non-negative regularizer, the model fitting time in this case is shorter than the quadratic loss with the non-negative regularizer. The best model using Huber

loss and non-negative regularizer, $w_{nn} = [0.299, 0.278, 1.039]$, has the least average error (around 1.8 m/s) for the whole validation set (11 groups). The standard deviation of this model's error of the whole validation set is around 0.5 m/s. Figure 7 showed the model performance is similar to the quadratic loss.



(a) (b)

Figure 5: Two-Day-Based model performance without regularizers

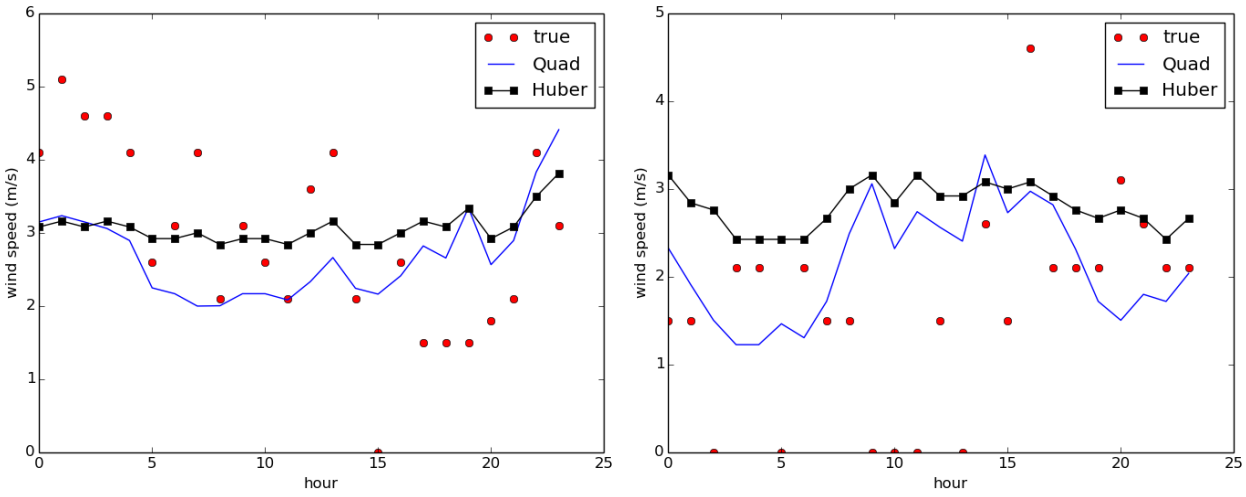


(a) (b)

Figure 6: Two-Day-Based model performance with Non-negative regularizer

Auto-Regression Moving Average Idea

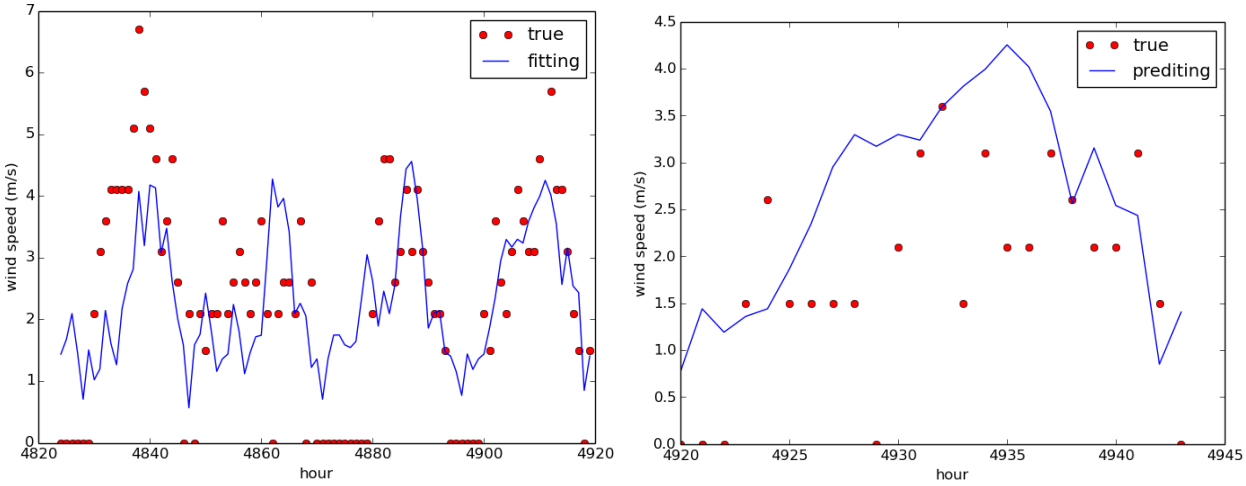
By fitting the wind speed using four previous days, a very promising trend can be seen in Figure 8 (a), and the prediction of the next day is also impressive, shown in Figure 8 (b).
In a simple Auto-Regression (AR) model, the current hour data can be modeled by several previous hours data. But as mentioned earlier, the 24 hours prediction is difficult.
In R script and Python, there are implemented Auto-Regression Moving Average (ARMA) method. However, using this method for predicting the future 24 hours wind speed is not really a "big" data problem because



(a) (b)

Figure 7: Two-Day-Based model performance comparison between Quadratic and Huber losses

the latest few days data can easily be obtained. It will become a "big" data problem if we are looking at future wind speed at many different geographic locations, for example, to predict wind speed and direction contour of the next 24 hours of New York State. Then we may take advantage of the ARMA model and predict wind speed and wind direction at each airport, analyze the geographic correlations among these airports. The ARMA method performance can be found in many literature, so it will not be expanded here.



(a) 5 days fitting (b) the next day prediction

Figure 8: Auto-Regression Model Idea

Discussion

The Two-Day-Based model performance is generally better than the Two-Year-Based model. The model performance would be better if it could be updated by fitting new data in the test set, which means to predict the next 24 hours wind speed by using the latest history (several previous days).
The model developed in this project could be used as a sub-tool for the ARMA model to predict wind speed because the prediction goal is to get the hourly data of the next day on the current afternoon, which means

some of the latest data on the current day were not recorded yet. In this case, both the Two-Year-Based model and the Two-Day-Based model are useful because the latest hourly wind speeds on the current day are not required.

Reference

1. Website of the Data: https://mesonet.agron.iastate.edu/request/download.phtml
2. Kavasseri, R. G., & Seetharaman, K. (2009). Day-ahead wind speed forecasting using f-ARIMA models. Renewable Energy, 34(5), 1388-1393.
3. El-Fouly, T. H., El-Saadany, E. F., & Salama, M. M. (2008). One day ahead prediction of wind speed and direction. IEEE Transactions on Energy Conversion, 23(1), 191-201.
4. El-Fouly, T. H. M., El-Saadany, E. F., & Salama, M. M. A. (2006, June). One day ahead prediction of wind speed using annual trends. In 2006 IEEE Power Engineering Society General Meeting (pp. 7-pp). IEEE.
5. Khan, A. A., & Shahidehpour, M. (2009, March). One day ahead wind speed forecasting using wavelets. In Power Systems Conference and Exposition, 2009. PSCE'09. IEEE/PES (pp. 1-5). IEEE.