IOE 565 Course Project Report: Wind Speed Prediction

By Team No.11:

Yuheng Cai (yhcai) **Qianyu Chen (chienyee)**

Chang Li (chanli) **Jianyao Zhu (jianyzhu)**

1. Introduction

Within the last several decades, the wind power is receiving increasing attention all over the world due to its environmental-friendly and renewable nature. According to 2016 U.S. Wind Industry Market Reports by American Wind Energy Association, the U.S. had an installed capacity of 82,183 MW by the end of 2016[1]. One important subfield of wind energy researches is the prediction of wind speed and wind direction basing on the previous data. However, it is a huge challenge to give accurate wind speed forecasts because of its stochastic nature and unpredictable environmental shocks.

The main purpose of this project is to develop a proper time series model for the wind speed measured in each of the four meteorological stations and then make one-step-ahead, two-step-ahead and three-step-ahead predictions for the hourly wind speed in December 2008. The original dataset provides 17,544 hourly observations of the wind speed and wind direction at each of the four stations from the beginning of 2007 to the end of 2008. Other coordinative information is the location, including latitude, longitude, and altitudes of the four meteorological stations are also provided.

The remaining sections of this report are data visualization, data preprocessing, fitting models with deterministic seasonal trend, fitting ARMA models, wind speed prediction and conclusion. We will give a glance of the whole dataset, pick up our target data and time periods, test and adjust time series for stationarity, decompose the model into deterministic part and stochastic part, fit regression model and ARMA model respectively for the two parts, and based on the models we select, predict the wind speed in December 2018 as described in the purpose of the project.

2. Data Visualization And Target Data Selection

Before we explore the suitable models for predicting wind speed, the preliminary stage of the statistical analysis includes checking the correlation of the wind speed as well as wind direction for the four anemometers, using the coordinative data to pick our target datasets, detecting trend

[1] "AWEA 2016 Fourth Quarter Market Report". AWEA. American Wind Energy Association. Retrieved 9 February 2017. (https://www.awea.org/resources/publications-and-reports/market-reports/2016-u-s-wind-industry-market-reports)

and seasonality in the original time series of each anemometer's data points and finally determine the suitable periods for prediction.

Firstly, we try to subset the whole dataset to find out the reasonable variables for predicting the wind speed measured at one station. Our assumption here is:

Assumption 1: the wind speed data at each station is an independent time series. Only the wind speed data at the same station will be used for analysis and prediction.

In order to demonstrate **assumption 1**, we plot the correlations among available variables to see if we should include wind directions for wind speed prediction, or if we should include the data from another anemometer for the specific anemometer we research into. **Figure 1** shows there does exist positive correlations among wind speed measured at different anemometers. But **Figure 2** indicates the correlations among wind direction detected at different anemometers are not significant. The reasons lie behind the differences could be the disparate locations of the four anemometers.

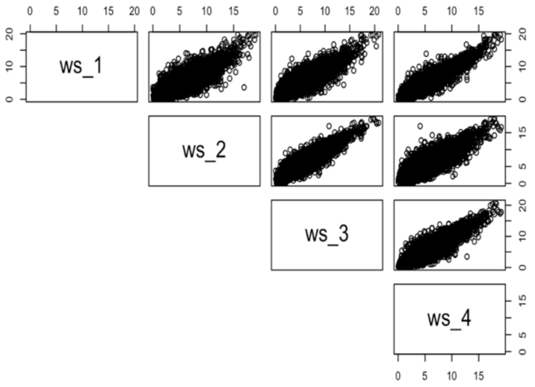


Figure 1: Correlation Matrix of Wind Speed at Different Anemometers

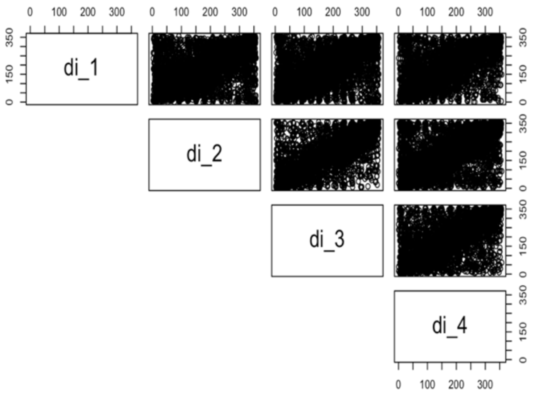


Figure 2: Correlation Matrix of Wind Direction at Different Anemometers

Based on these correlation plots, we can reasonably infer these four anemometers might be in the same area so that the wind speeds are similar, while their wind directions are not because they are not parallel in either one of the three space dimensions. Then we use the coordinate location dataset to testify our assumption. **Figure 3** shows the relative locations in the three-dimension space of the four anemometers.

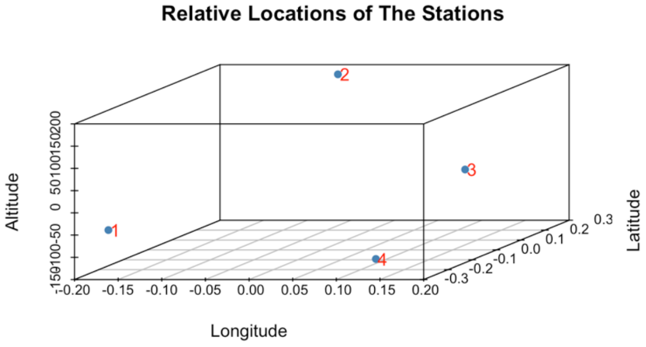


Figure 3: Relative Locations Plot

Since the four anemometers have very different longitudes, latitudes and altitudes, we will only regard the wind speed from each anemometer itself as a time series and make predictions according to the previous wind speed measured at the same anemometer. One may argue that if we detect the wind direction at station i is to station j, and we know the current wind speed at station i, then we can assume the current wind speed at station i will change the wind speed at station j in one hour. But these four stations are distant from each other, there are not enough pieces of evidence to show whether and how much the wind speed increase or decrease, or if the wind direction changes during the journey from one station to another. Even we assume the wind does not change direction and never accelerate or decelerate, the effect of wind speed/direction of one station to another could still be insignificant because the stations have different altitudes. Therefore, wind directions will not be considered for predicting the wind speed in the future. Similarly, wind speed detected in other stations will not be considered for modeling and prediction, either.

Next, we would like to detect the seasonality and trend in the time series of wind speed at each station. Our assumption is:

Assumption 2: The time series of wind speed at each station may have strong daily, monthly and quarterly seasonality.

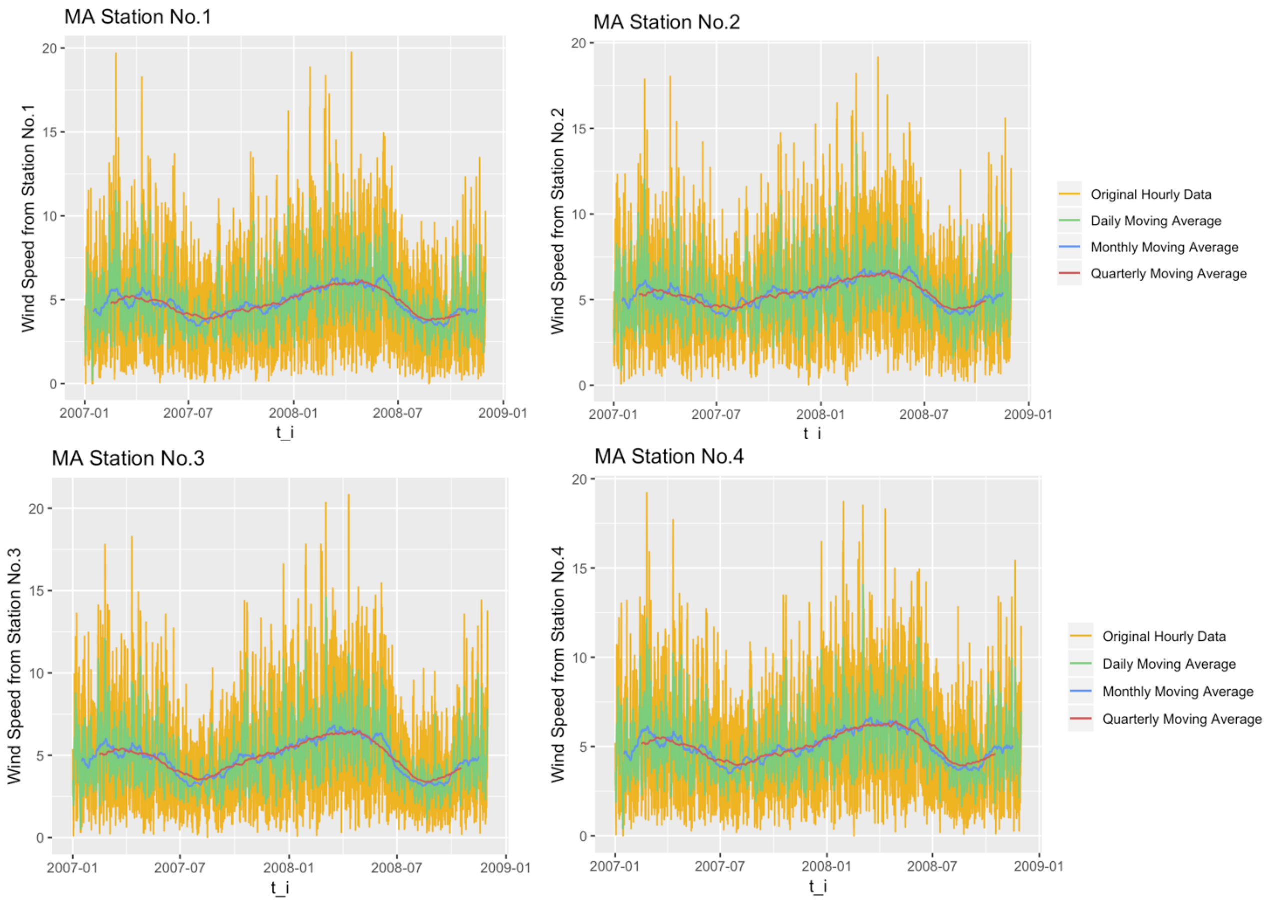


Figure 4: Time Series Plots For Historical Data At Each Station

Figure 4 is the set of time series plots for the hourly, daily, monthly and quarterly wind speeds from the four stations from 2007 to 2008 excluding the predicting period. We could see they may all exhibit seasonality. In the short term, every 24 hours could be a cycle for wind speed. In the long term, the wind speed is higher in the summer time and lower in the fall and winter. **Figure 5** is the set of quarterly decomposition plots of all the four time series. Here we notice the peak values of wind speed appear in summer times, while the peak values of 2008 are higher than those of 2007. This phenomenon indicates we may use only partial data points to get precise predictions.

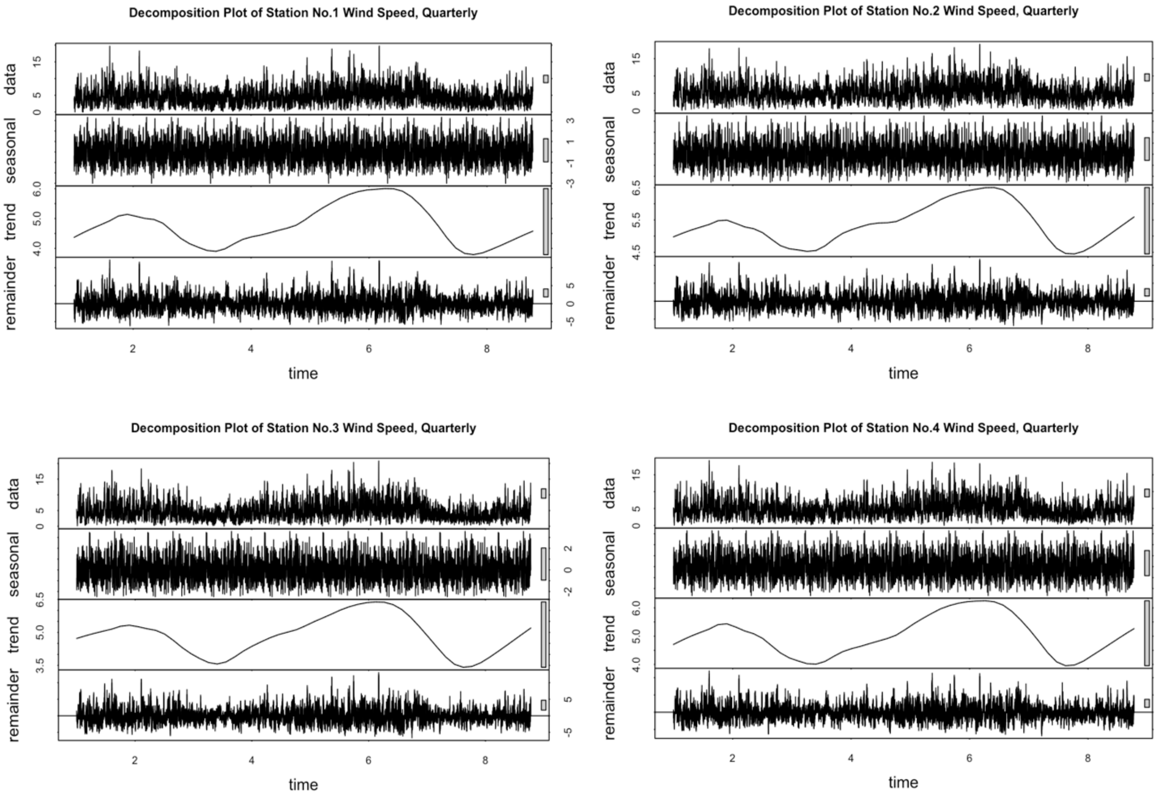


Figure 5: Wind Speed Time Series Decomposition At Each Station, Quarterly

As we found the original wind speed time series are stochastic process, we want to cut off the period for a better prediction. The corresponding assumption is:

Assumption 3: data points from 1:00 am, July 1^{st}, 2008 to 0:00 am, December 1^{st}, 2008 are adequate to work as training dataset to predict the hourly wind speed in December 2008.

Based on **Figure 6** below, the datapoints of wind speed are collected from fall to winter, which share similar relatively lower wind speed during a year. The variation of wind speed in this period is not so large as that in any longer period. Meanwhile, the period is both consecutive and long enough for accurate predictions.

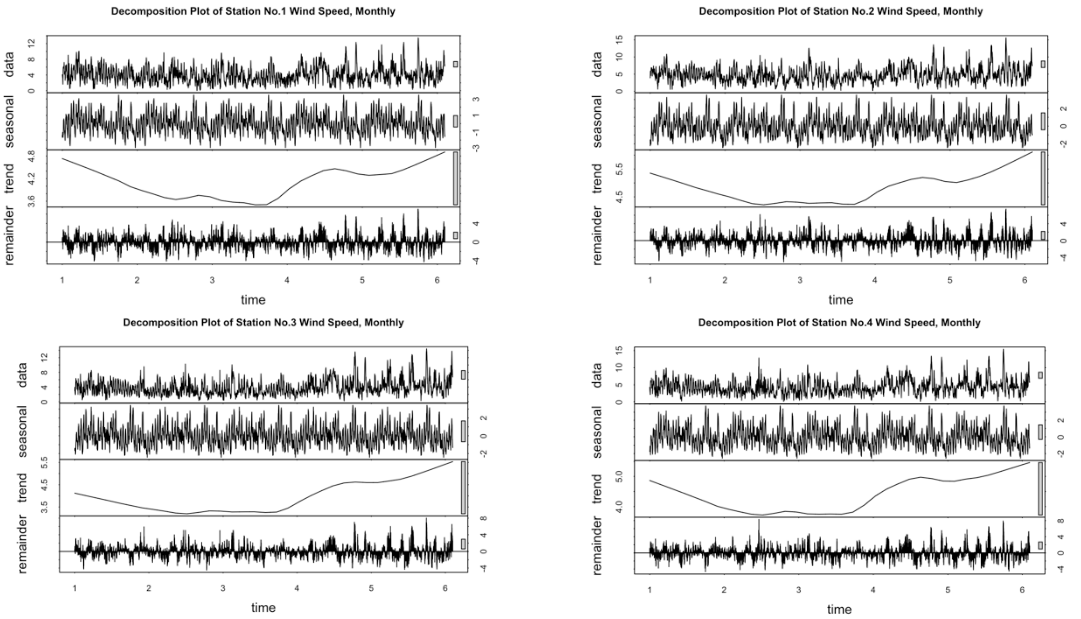


Figure 6: Wind Speed Time Series Decomposition At Each Station, Quarterly

3. Data Preprocessing

According to the data visualization and analysis in the previous sections, we will use the hourly wind speed data from 1:00 am, July 1st, 2008 to 0:00 am, December 1st, 2008 from each anemometer as training dataset to predict the hourly wind speed from 1 am, December 1st, 2008 to 0:00 am, January 1st, 2009 at the same anemometer. The prediction part is also the test dataset. For further analysis, we will regard the wind speed data from each station as an independent time series. Since **Figure 4** and **Figure 6** show wind speed at all the four sations exhibit daily cycles and seasonal patterns, time series based on the wind speed are non-stationary. The data need to be preprocessed before predicting wind speed in December 2018.

Theoretically, there are many methods which can make time series to be stationary, including but not limited to transforming and differencing. For example, we tried to log-transform and double differencing the wind speed data to achieve the stationarity of time series. Although the time series appear stationarity, data preprocessing method we used is more theoretically solid-founded. Then a more suitable method is derived in the following parts.

Since the wind speed presents obvious daily cycles, we subtract the hourly mean collected at the same points of time during a day between July and November from the original hourly wind speed to remove the daily seasonality fixed effect. Then we also clear away the monthly pattern by discounting the corresponding mean of the same points of time during a one-month interval. These two transformations could be represented by the matrices below:

$$\begin{bmatrix} s_{1,1} & \cdots & s_{153,1} \\ \vdots & \ddots & \vdots \\ s_{1,24} & \cdots & s_{153,24} \end{bmatrix} \xrightarrow{remove\ hourly\ fixed\ effects} \begin{bmatrix} s_{1,1} - \overline{h_1} & \cdots & s_{153,1} - \overline{h_1} \\ \vdots & \ddots & \vdots \\ s_{1,24} - \overline{h_{24}} & \cdots & s_{153,24} - \overline{h_{24}} \end{bmatrix}$$

$$\begin{bmatrix} s_{1,1} & \cdots & s_{153,1} \\ \vdots & \ddots & \vdots \\ s_{1,24} & \cdots & s_{153,24} \end{bmatrix} \xrightarrow{remove\ montly\ fixed\ effects} \begin{bmatrix} s_{1,1} - \overline{m_7} & \cdots & s_{153,1} - \overline{m_{11}} \\ \vdots & \ddots & \vdots \\ s_{1,24} - \overline{m_7} & \cdots & s_{153,24} - \overline{m_{11}} \end{bmatrix}$$

$s_{i,j}$ represents the wind speed at i-th day, and j-th hour (in the real observations, j = 24 is represented by j = 0 in (i+1)-th day), where $i = 1, \dots, 153, j = 1, \dots, 24$. $\overline{h_\mathrm{j}}$ denotes the mean of j-th hourly wind speed during a day. $\overline{m_\mathrm{k}}$ denotes the mean of k-th-month's wind speed, where $k = 7, \dots, 11$. For $i = 1, \dots, 31$, the cleaned data is $s_{i,j} - \overline{m_7}$; for $i = 32, \dots, 62$, the cleaned data is $s_{i,j} - \overline{m_8}$; for $i = 63, \dots, 92$, the cleaned data is $s_{i,j} - \overline{m_9}$; for $i = 93, \dots, 123$, the cleaned data is $s_{i,j} - \overline{m_{10}}$; and for $i = 124, \dots, 153$, the cleaned data is $s_{i,j} - \overline{m_{11}}$. For instance, the clean data point which is collected at 12:00 pm, August 2nd, 2008 is $s_{33,12} - \overline{h_{12}} - \overline{m_8}$.

Figure 7 below shows the effect of data preprocess is significant by comparing with the patterns in the original data in **Figure 4**. The mean of data become steady throughout time but the variance still varies by time. Making further transformation is necessary to make the data stationary. This result also appears in the autocorrelation function plots and partical autocorrelation function plots in **Figure 8**. With both ACF and PACF showing significant values, we believe ARMA models will serve for the future needs.

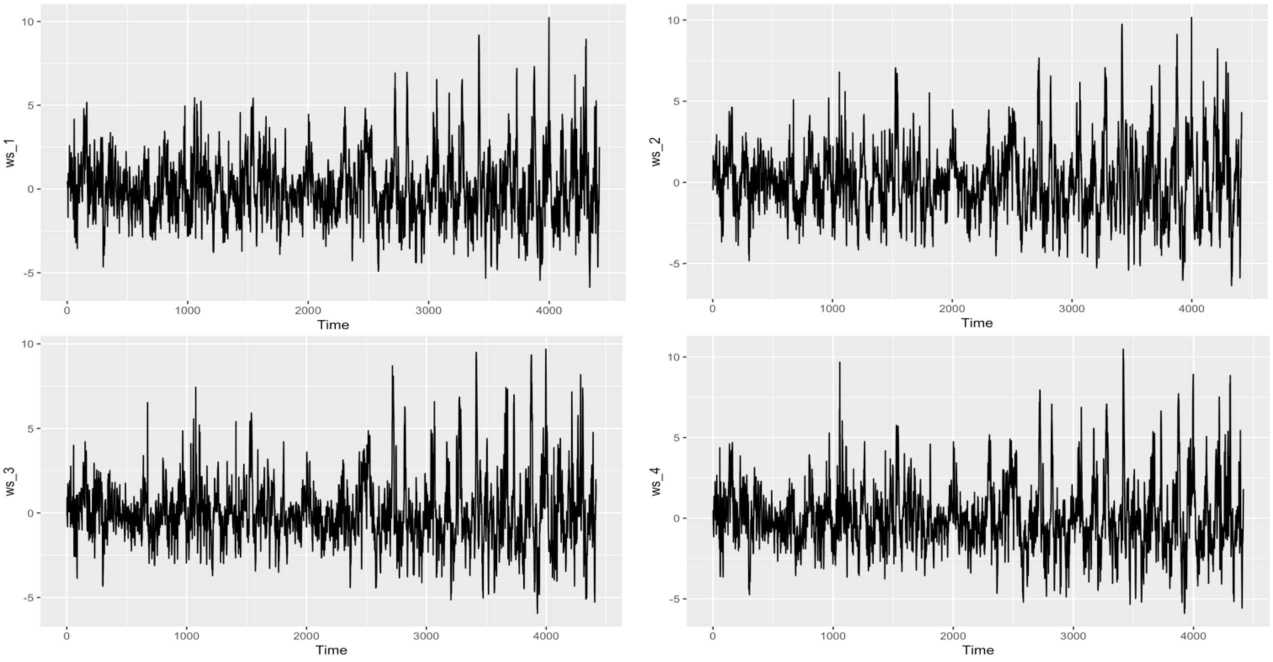


Figure 7: Wind Speed From Four Stations After Treatment, 07/08 – 11/08

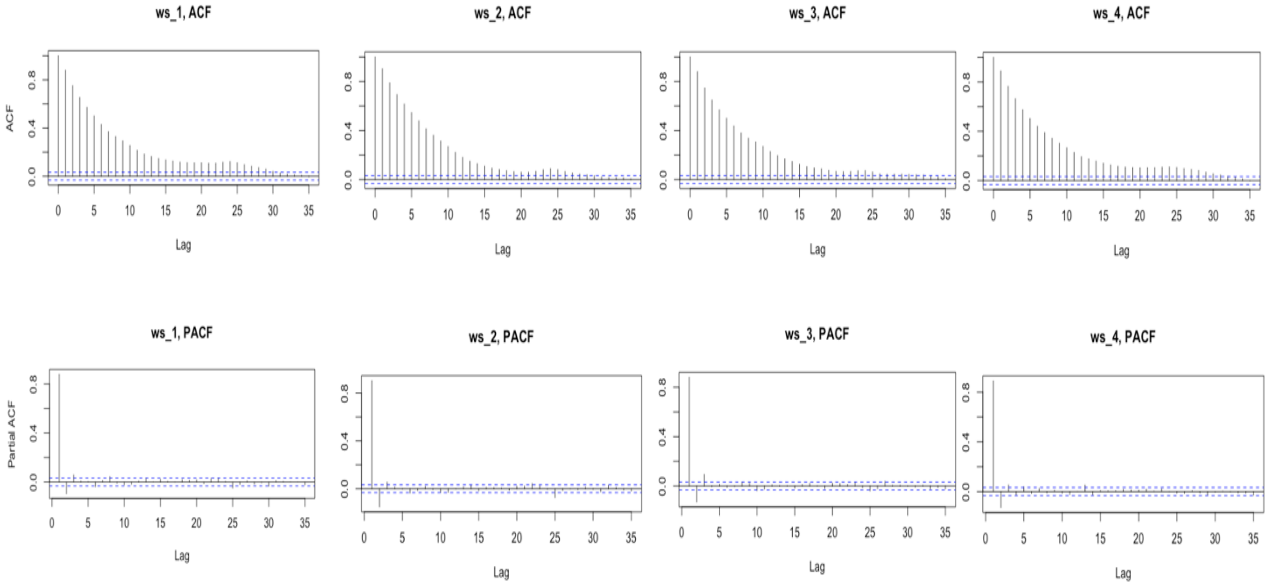


Figure 8: ACF And PACF Plots of Four Stations, 07/08 – 11/08

Thus, we decide to apply a deterministic trend model on the treated data to model the seasonality. After decomposing the model into the deterministic part and the stochastic part, we will applying ARMA(n,m) model on the stochastic part because the seasonality has been decomposed. The following two parts discuss the model we fit for the four wind speed time series and the corresponding results.

4. Fit Models With Deterministic Trend

This section intends to decompose the time series into deterministic trend and stochastic part. Considering the nature of wind speed and previous analysis, we use two period parameters to formulate: $\mathrm{p}_1 = 12$ and $\mathrm{p}_2 = 24$. And thus the corresponding $\omega_1 = \frac{2\pi}{p_1}$ and $\omega_2 = \frac{2\pi}{p_2}$ are the dominant frequency for the model. Denote the treated wind speed time series as $\mathrm{y_t}$, the intercection term as α_0, and the residuals as $\mathrm{X_t}$, we get the model:

$$\mathrm{y_t} = \alpha_0 + \sum_{j=1}^{2} \left[\beta_{1j} \sin(\omega_j t) + \beta_{2j} \cos(\omega_j t)\right] + X_t$$

$$where\ \mathrm{X_t} = ARMA(n,m), with\ residual\ a_t$$

To get the most suitable model, we select β_{1j} and β_{2j} to minimize $\{\mathrm{y_t} - \alpha_0 - \sum_{j=1}^{2} \left[\beta_{1j} \sin(\omega_j t) + \beta_{2j} \cos(\omega_j t)\right]\}^2$. **Table 1** shows the coefficients for fitting the treated wind speed at the four stations into the model.

Table 1: Deterministic Seasonal Trend Model Coefficients for Treated Wind Speed

	α_0	β_{11}	β_{12}	β_{21}	β_{22}
WS_1	-3.28e-07	-.0195	-.0729	-.0829	.0839
WS_2	-3.20e-07	-.0678	-.0396	-.0111	.0415
WS_3	-2.81e-07	-.0617	-.0261	-.0001	.0762
WS_4	-3.11e-07	-.0377	-.0534	-.0309	.0602

5. Fit ARMA Models

Then we fit ARMA models to the stochastic part in the deterministic seasonal trend model in section 4. The fundamental rule of picking up the best AMRA model is to choose the adequate model with the smallest orders. We use the ARMA(n,n-1) strategy for stepwise model selection, and compare parsimonious model and full model at each step by conducting the F test. As long as we find out the suitable order (n*, n*-1) for ARMA model, we also try out (n*-1, n*-1) to see if it is adequate as well as leading to independent residuals. If so, we use ARMA(n*-1, n*-1) following the fundamental rule. Otherwise, we will stick with ARMA(n*, n*-1). To check the independence of residuals of the ARMA model, we use Box-Ljung test, Portmanteau test and Bartlett band test.

Table 2: ARMA Models And Test Results for Treated Wind Speed

	WS_1	WS_2	WS_3	WS_4
ARMA Order	(4, 4)	(5, 4)	(4, 3)	(2, 1)
Coefficients ϕ	$\phi_1 = .5073$ $\phi_2 = -.3793$ $\phi_3 = -.3113$ $\phi_4 = -.2571$	$\phi_1 = 1.0951$ $\phi_2 = -.9914$ $\phi_3 = .3864$ $\phi_4 = .3184$ $\phi_5 = -.0320$	$\phi_1 = .9637$ $\phi_2 = -.9914$ $\phi_3 = .2730$ $\phi_4 = .3266$	$\phi_1 = .4438$ $\phi_2 = .3717$
Coefficients θ	$\theta_1 = .4653$ $\theta_2 = .6809$ $\theta_3 = .3144$ $\theta_4 = -.0330$	$\theta_1 = -.0711$ $\theta_2 = .7611$ $\theta_3 = .4393$ $\theta_4 = -.0034$	$\theta_1 = .0394$ $\theta_2 = .6707$ $\theta_3 = .5163$	$\theta_1 = .05642$
Box Test	Pass	Pass	Pass	Pass
Portmnteau Test	Pass	Pass	Pass	Pass

Table 2 shows the final models and corresponding coefficients we pick for the four stations and their autocorrelation statuses by various testing methods. **Figure 9** shows the diagnostic plots of the corresponding optimal ARMA models. All the p-values derived from the Box-Ljung test is greater than 95%, which indicate that with 95% significant level we conclude that the residuals are uncorrelated to each other. The Portmnteau Test also show the same result as

the Q calculated is smaller than the corresponding chi-square critical value under 95% confidence level.

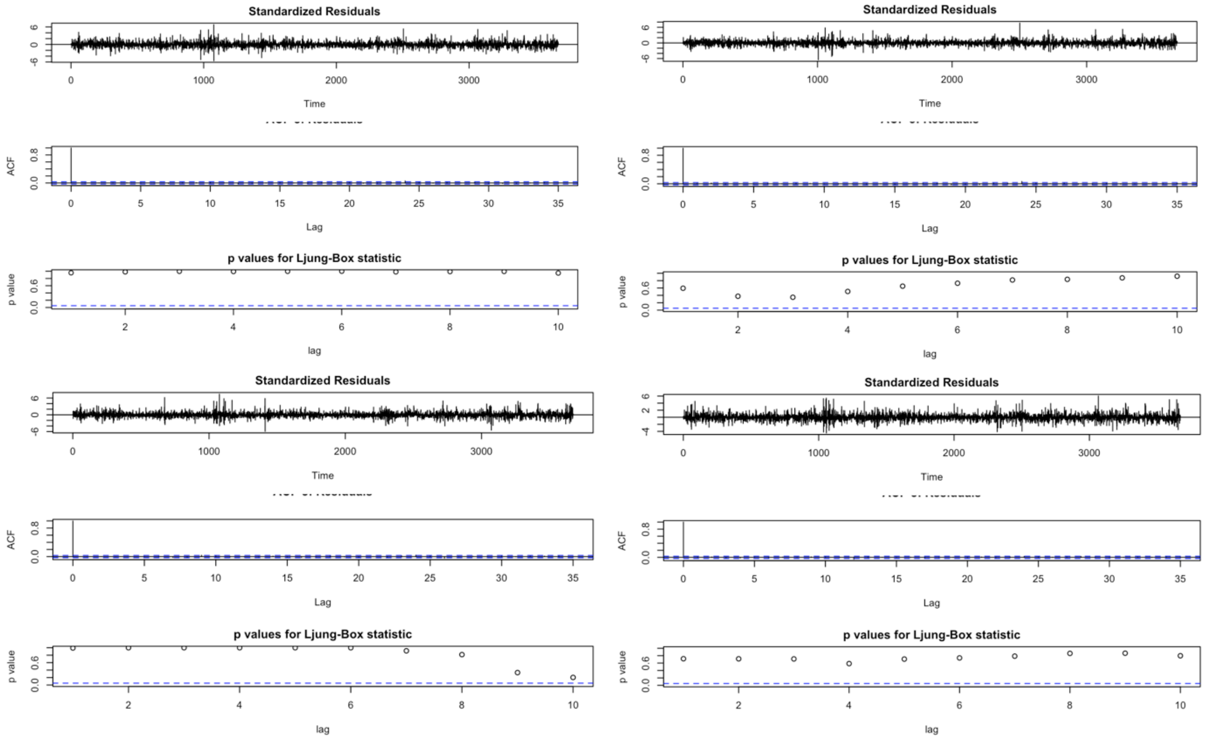


Figure 9: Diagnostic Plots

Figure 10 shows the Bartlett Band Test results for the picked ARMA model residuals. All fall within the bands. Therefore, we can conclude the residuals are uncorrelated. To further test whether the stochastic process is stationary, we report the roots of the optimal ARMA model for wind speed at each station in **Table 3**. The norms of roots are all less than 1, which indicates that the four models all capture the stationary patterns of wind speed processed data.

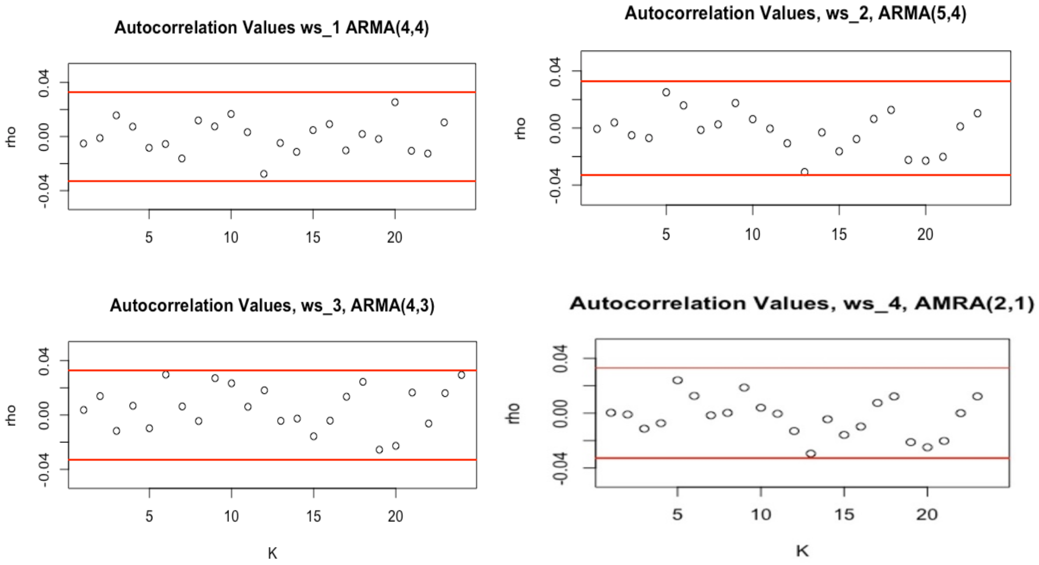


Figure 10: Bartlett Band Test Results

Table 3: Roots of The Optimal ARMA Models (λ Values)

	WS_1	**WS_2**	**WS_3**	**WS_4**
Order	(4, 4)	(5, 4)	(4, 3)	(2, 1)
λ	$\lambda_1 = -.412$ $\lambda_2 = -.871$ $\lambda_{3,4}$ $= .024 \pm .846\mathrm{i}$	$\lambda_1 = -.988$ $\lambda_2 = -.209$ $\lambda_3 = .888$ $\lambda_{4,5} = .263 \pm .963\mathrm{i}$	$\lambda_1 = -.395$ $\lambda_2 = .853$ $\lambda_{3,4}$ $= .253 \pm .951\mathrm{i}$	$\lambda_1 = -.427$ $\lambda_2 = .871$

Figure 11 gives a visualization of the roots. With real number and imaginary number being axes, all the roots lie within the circle of radius equal to 1. Thus the residuals of treated wind speed after applying our models are stationary for all the stations.

In addition, the autocorrelation plots in **Figure 12** show that all the lags are not beyong the absolute critical line for all of the four stations. All the measurements give us strong evidences of stationarity from the ARMA model derived, hence the model for each station is optimal and will be applied for predicting the wind speed in December, 2008.

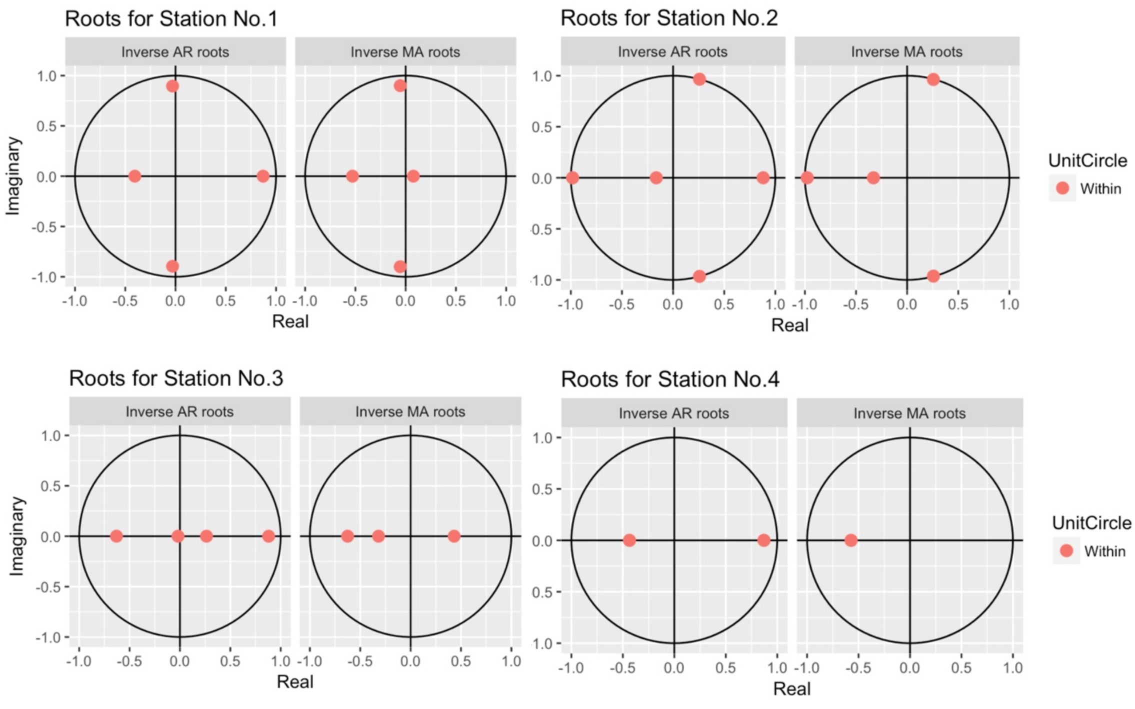
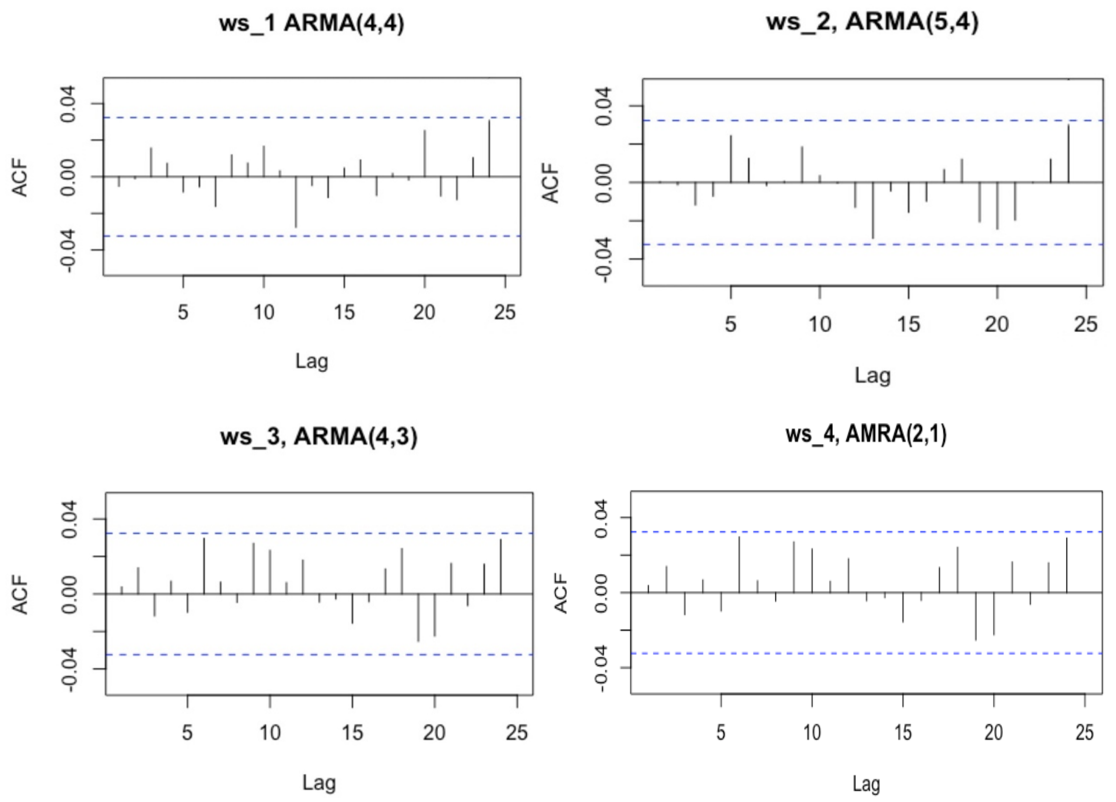


Figure 11: AR Roots And MA Roots Plot


ws_1 ARMA(4,4)

ws_2, ARMA(5,4)

ws_3, ARMA(4,3)

ws_4, AMRA(2,1)

Figure 12: ACF Plots

6. Predictions Results

After fitting the suitable models for each station's data, we make one-step-ahead, two-step-ahead and three-step-ahead predictions for the hourly wind speed in December 2008 at each station. The basic idea of prediction is predicting X_t first with the ARMA models, adding back the deterministic trend to get y_t, which is the treated wind speed. After getting y_t, we add back the monthly fixed effect and daily fixed effect we removed in the first place. All the wind speed predicted with steps ahead overlap with the actual wind speed recored in December 2008, and it seems that our prediction method forecast the future wind speed accurately in some extent. Now we get the final results of wind speed in the prediction period. The specific prediction results are available in "Group11-ForecastingResults.xlsx". Comparing our prediction results with the test dataset, we show the total prediction errors measured by RMSE in **Table 3**. We also plot the predicted wind speed values and the real observations in **Figure 13** for clear comparison.

Table 4: Prediction Errors

Total RMSE	**WS_1**	**WS_2**	**WS_3**	**WS_4**
One-Step-Ahead	1.5072	1.3797	1.5039	1.4027
Two-Step-Ahead	1.5055	1.3772	1.4980	1.3941
Three-Step-Ahead	1.5518	1.4430	1.5568	1.4494

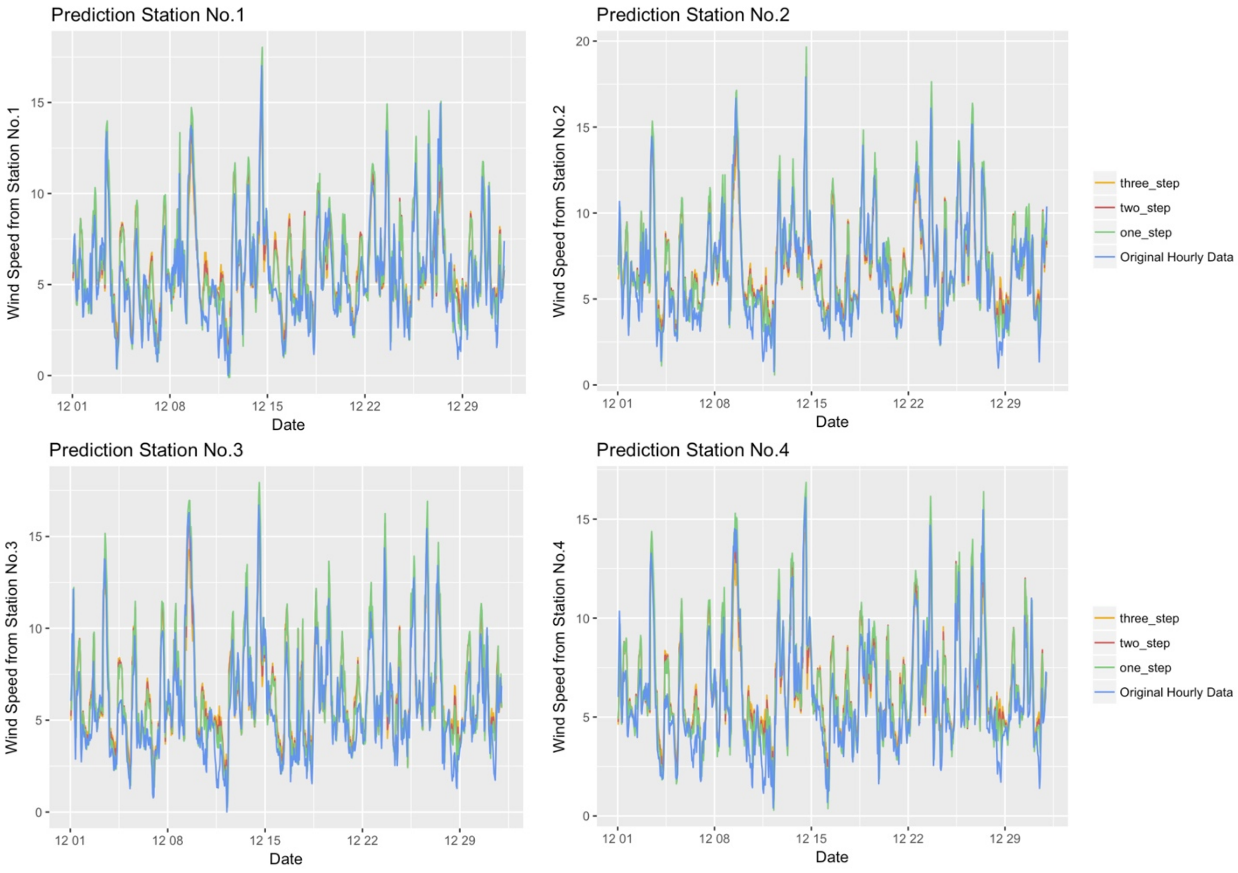


Figure 13: Predictions And Real Observations

7. Conclusion

This project focuses on predicting the hourly wind speed for a whole month by analyzing the wind speed, wind direction from the previous one year and eleven months, and the locations of four anemometers. In order to get a satisfying model, we first preprocessed the provided data. We excluded the wind direction data according to the location data provided, and also because of the lack of information about wind speed and direction changes during the wind movements. We treated wind speed data from each station as separate time series. To increase the prediction accuracy, we used the time series of hourly wind speed from the previous five months.

In order to remove the seasonality lies in the wind speed, we removed the corresponding mean for same time points in each 24-hour interval and the mean for same time points in each 1-month interval from the original hourly wind speed. To further discriminate the deterministic trend and stochastic part, we applied a deterministic seasonal trend model with periods equal to 12 and 24. Then we used ARMA to model the stochastic part and testified the residuals of the stochastic part after using ARMA model are stationary. The final forecast results were derived by adding up the prediction of the stochastic part, the deterministic seasonal trend, and the previously removed hourly mean and monthly mean. The prediction results were pretty reasonable, and the RMSE's were acceptable.

Through this project, we practiced the time series modeling and predicting techniques we learned in class, we also explored more related theoretical knowledge and programming techniques in the meantime. For example, how to deal with non-stationary time series, how to select a model for time series with seasonality, and how to pick the most moderate and effective subsets for modeling and predicting from the whole dataset.

We also believe there could be more aspects to explore for analyzing this problem. For instance, the relationship between wind speed and wind direction among different anemometers. But since the main purpose for this project is the prediction part, we did not spend too much time on those possible aspects.