# GR5221 Final Project Report

## Forecasting Spatial-Temporal Climate Data

Columbia University

Jingyou Jiang (UNI: jj3192)

Prof. Gabriel Young

Spring 2022

## I. Introduction

Forecasting windspeed and solar radiation has become increasingly important in fields such as engineering and finance. In the Wind-Energy Marketplace, the forecasted windspeed becomes highly valued information, especially to the governing companies who own the windmills. In short, companies must project how much energy output a potential buyer requires before setting up an energy deal. This could then lead to the windspeed futures being above or below market expectations.

In this project, we will use the dataset (WindSpeed_Month_Ave.csv), consisting of monthly spatial-temporal observations for windspeed over the span: from January 1979 to December 2018. The training data is constructed from averaging daily measurements, i.e., one case is based on roughly 30 days. Averaging the cases helps to smooth the data and reduces the computational burden of fitting the full set of observations. The resulting dataset consists of n = 480 months over the timespan and 916 positions (columns) from X1 to X916. Each case $X_t(l_1, l_2), t = 1,2, \dots ,480$ is defined as a function of latitude ($l_1$) and longitude ($l_2$), representing the observation domain of windspeed mainly over Mexico and surrounding states like Texas, New Mexico, and Oklahoma, as displayed in Figure 1.

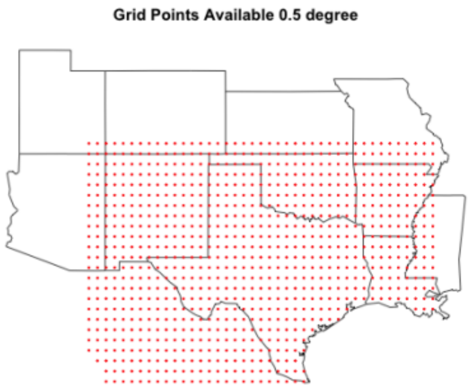
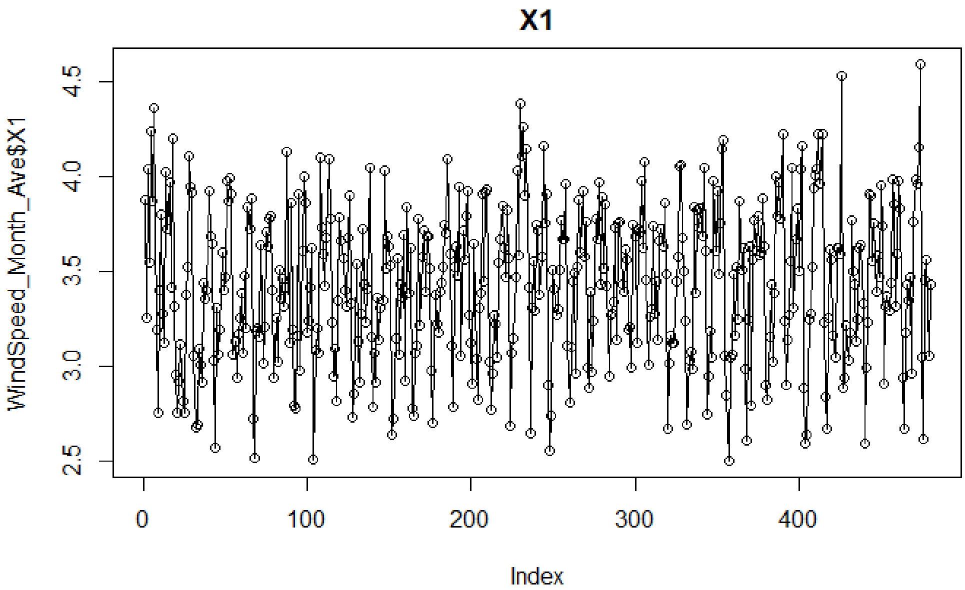



**Figure 1**

In this project, we are not going to evaluate the influence of longitude and latitude on the windspeed for each position, and we consider each position as an independent case. Taking position X1 as an example, we can visualize its average windspeed for 480 months. Graph of X1 windspeed over time:

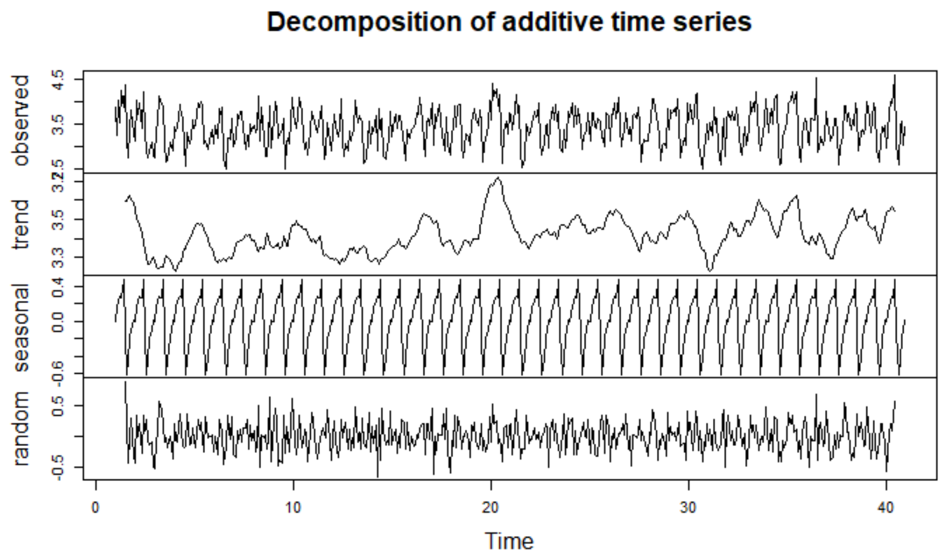



This project is going to build a Time Series model that could generally fit well and forecast windspeed first six months ahead over all 916 positions in the dataset.

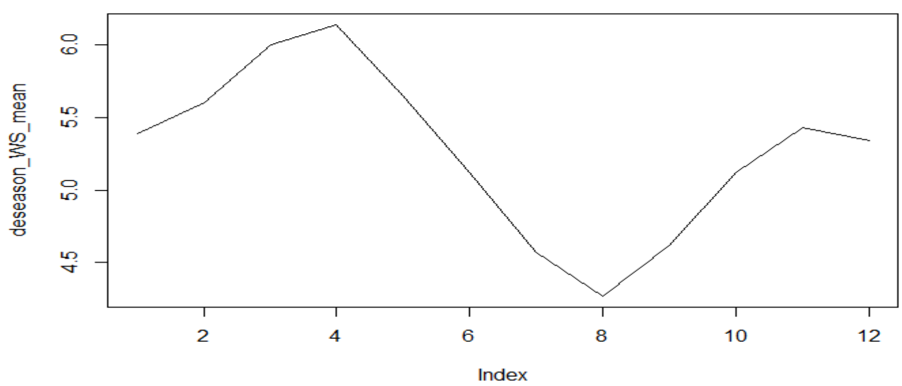
# II. Statistical Method & Model(s)

## 1. EDA

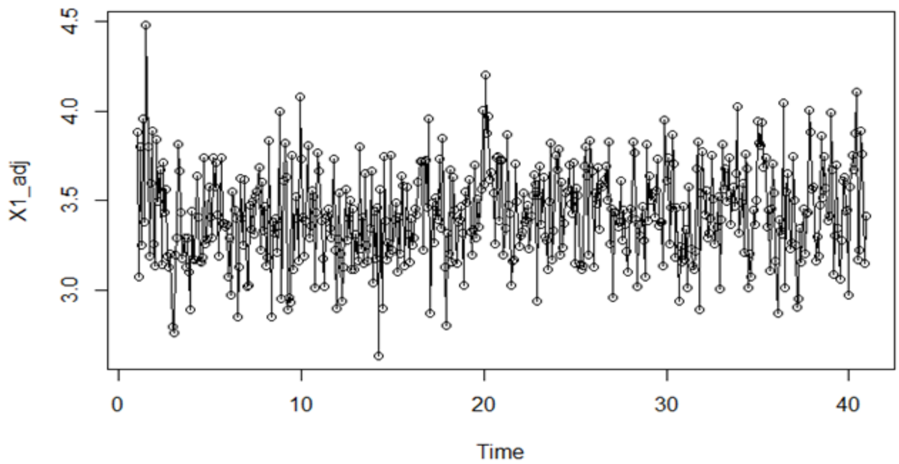
To analyze if the data have a seasonal pattern, we need to split the data into several components to adapt a time series model. First, we convert our data into a time series format to analyze by using the *ts( )* function in R, setting the frequency of 12 months. Then, decompose our data into four important parts: observed data, seasonal, trend, and residuals to see any useful information, using X1 as example.




From the plot above, there is no stable increasing or decreasing trend but a clear pattern of seasonal change over the time. Moreover, we evaluate the average windspeed over all position for each month and display in the plot below:

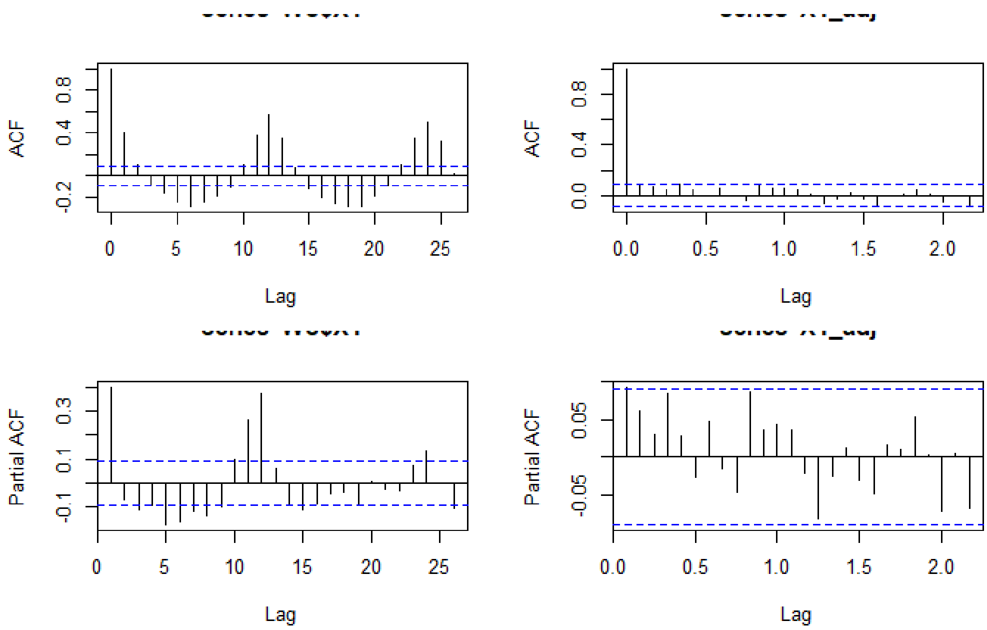


There is a similar pattern shown as the seasonal pattern of decomposed X1 data. It shows there is an obvious increasing trend from January to April and a steep decrease from April to August, then gradually increasing from August to November. Thus, we could use the decomposed data to build the model. And here is a plot for the X1 position after the decomposition.




To check the effectiveness of decomposition, we can also compare the ACF and PACF plots of X1 before and after the seasonal decomposition.

**ACF & PACF:**



From the plot above we can see obvious improvement after subtracting the seasonal pattern of X1 position. Thus, we can conclude the seasonality of our data and our de-seasonal data have stationary residuals.

## 2. ARIMA Model

Before modeling, we extract first 360 months of windspeed over all positions as the training data and the last 120 months as the testing data so that we split the data into 75%-25% for train and test. We first try the ARIMA(3,0,1) model on the training data of position X1 and get the model summary with coefficients:

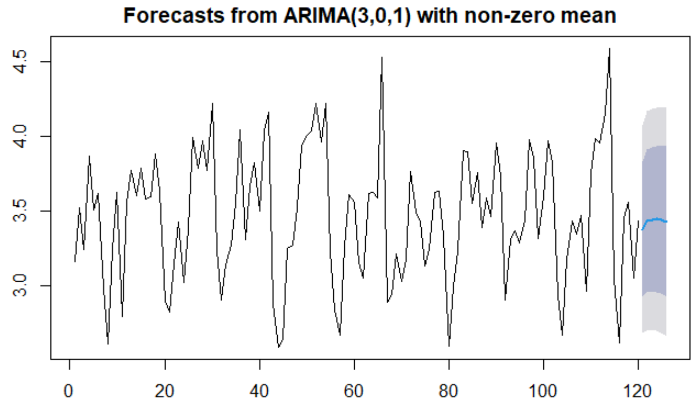
```
Series: train_X1
ARIMA(3,0,1) with non-zero mean

Coefficients:
         ar1      ar2      ar3      ma1    mean
      1.0746  -0.2198  -0.1728  -0.7241  3.4138
s.e.  0.0777   0.0818   0.0550   0.0617  0.0162

sigma^2 = 0.1254:  log likelihood = -134.79
AIC=281.58   AICc=281.82   BIC=304.9

Training set error measures:
                      ME      RMSE       MAE       MPE     MAPE      MASE         ACF1
Training set 0.001763509 0.3516373 0.2837504 -1.042939 8.490224 0.8023499 -0.008430768
```

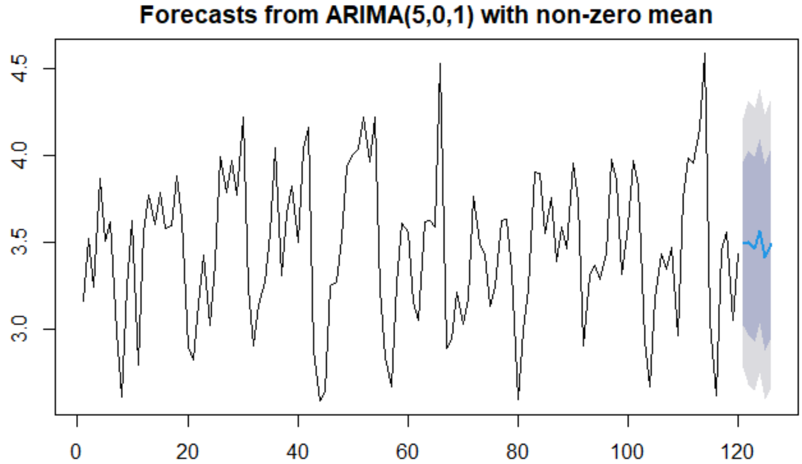
And the forecasted windspeed of next six months for X1 position are [Jan 2019-3.375800, Feb 2019-3.432467, Mar 2019-3.439464, Apr 2019-3.443856, May 2019-3.437244, June 2019-3.427965] when tested with the test data sets with the graph:



The tested model performs relatively good with RMSE=0.352 on position X1. Then, we decide to run the auto.arima( ) on all the position to acquire most frequently appeared model parameters as our best model selection that can be applied among all positions to forecast their windspeed for the next 6 months. To accomplish this, we build a for loop to run the ARIMA model among all positions and get the most frequently appeared parameters are [5, 1, 0, 0, 1, 0, 0], which stands for the model **ARIMA (5,0,1)*(0,0,0)**. Then, we can apply this model to serval different positions to check its availability and representative.

To compare the model directly with previous methods, we first use X1 position to continue our analysis. Then, we get our model with RMSE=0.2433 and a forecasting plot for the next 6 months:

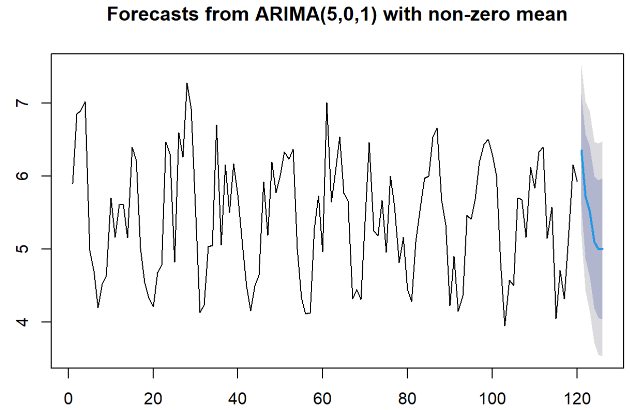
$$\boldsymbol{X_t = -0.4539X_{t-1} + 0.2158X_{t-2} - 0.1558X_{t-3} - 0.0650X_{t-4}}$$
$$\boldsymbol{- 0.2390X_{t-5} + Z_t + 0.9942Z_t + 3.4602}$$




The forecasted windspeed of next three months on X1 with confidence intervals:

| Forecast | Lo 80 | Hi 80 | Lo 95 | Hi 95 |
|---|---|---|---|---|
| 3.492533 | 3.023336 | 3.961729 | 2.774958 | 4.210107 |
| 3.498491 | 2.965942 | 4.031040 | 2.684027 | 4.312955 |
| 3.458310 | 2.925602 | 3.991019 | 2.643604 | 4.273017 |
| 3.564452 | 3.031592 | 4.097312 | 2.749514 | 4.379391 |
| 3.411672 | 2.874648 | 3.948696 | 2.590365 | 4.232979 |
| 3.494803 | 2.948746 | 4.040860 | 2.659681 | 4.329925 |

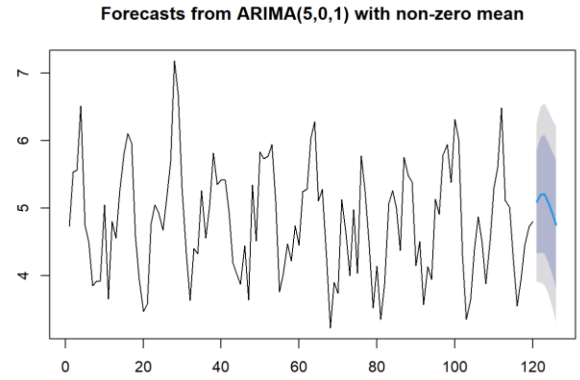
To further check the availability of our model, we randomly sample 3 points from all positions to evaluate and visualize their forecasted windspeed for the next 6 months. The chosen sample in our report are positions X728, X263, and X751. Here are the plots for these sampled cases:




```
Coefficients:
          ar1     ar2     ar3      ar4      ar5     ma1  intercept
      -0.4955  0.6268  0.0228  -0.5454  -0.2637  1.0000     5.4121
s.e.   0.0880  0.0861  0.1037   0.0879   0.0909  0.0802     0.0659

sigma^2 estimated as 0.3496:  log likelihood = -109.13,  aic = 234.26
```

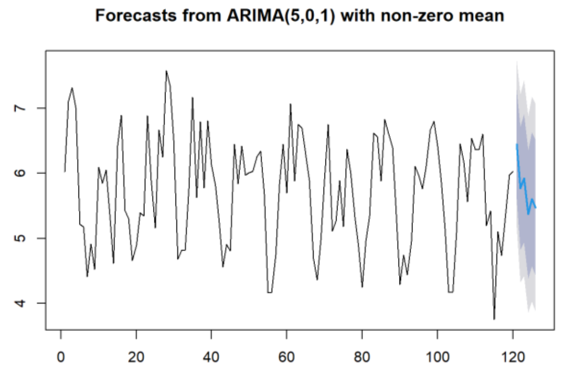
**X728**




```
Coefficients:
         ar1      ar2      ar3     ar4      ar5      ma1  intercept
      1.1477  -0.2373  -0.3798  0.1797  -0.1146  -0.6866     4.7993
s.e.  0.1387   0.1447   0.1355  0.1398   0.0990   0.1096     0.0438

sigma^2 estimated as 0.3637:  log likelihood = -110.23,  aic = 236.46
```

**X263**




```
Coefficients:
          ar1     ar2      ar3      ar4      ar5     ma1  intercept
      -0.3798  0.5396  -0.1925  -0.4458  -0.0399  0.9032     5.7105
s.e.   0.2015  0.1131   0.1028   0.0962   0.1195  0.1832     0.0753

sigma^2 estimated as 0.4267:  log likelihood = -119.8,  aic = 255.6
```

**X751**

We can see the results using the ARIMA model performs well on these random cases while forecasting the windspeed of next 6 months.

# III. Results & Conclusion

In this project, I convert the windspeed data into time series format and visualize the decomposition of the time series data. It illustrates an obvious pattern of seasonality in a period of 12 months. Then I check the de-seasonal data with ACF and PACF plots to get stationary residuals after eliminating seasonality. Through the project, I mainly focus on the analysis of position X1 to build and improve my time series model. And I assume every position is independent with each other.

In this project, I mainly use the seasonal ARIMA model to analyze each position. After running the model on all positions, we get the most fit model **SARIMA (5,0,1)*(0,0,0)** as our final model to forecast the windspeed for the next six months: January 2019 to June 2019. We check our model with X1 and other 3 randomly selected positions: X728, X263, and X751 in our report. We successfully obtain the forecast graphs and model summary with coefficients and test error and it illustrate the model performs well on our selected points.

Thus, we can conclude the model **SARIMA (5,0,1)*(0,0,0)** is the best model to forecast the windspeed in this dataset and we can apply this model to forecast next 6 months windspeed among all cites. The forecasted windspeed of our selected positions for the next 6 months are shown below.

| Position | Jan 2019 | Feb 2019 | Mar 2019 | Apr 2019 | May 2019 | June 2019 |
|---|---|---|---|---|---|---|
| X1 | 3.492533 | 3.498491 | 3.458310 | 3.564452 | 3.411672 | 3.494803 |
| X728 | 6.352045 | 5.722804 | 5.523135 | 5.094271 | 4.996948 | 5.003797 |
| X263 | 5.090700 | 5.196958 | 5.212280 | 5.077323 | 4.921382 | 4.754636 |
| X751 | 6.445614 | 5.770715 | 5.922999 | 5.368139 | 5.603182 | 5.469434 |

# IV. Appendix

## 1. Model Selection

```
Model Parameter Selection

```{r}
order<-matrix(NA,nrow=n,ncol=7)
freq_order<-rep(NA,7)
for (i in 1:n){
  model<-auto.arima(ws[,i])
  order[i,]<-model$arma
}
for (i in 1:7){
  freq_order[i]<-as.numeric(names(which.max(table(as.data.frame(order)[,i]))))
}
freq_order
```

 [1] 5 1 0 0 1 0 0
```

## 2. Model Validation Testing

```
> accuracy(X1_best_fore,test=test_X1)
                     ME      RMSE       MAE       MPE   MAPE     MASE      ACF1
Training set -0.1428604 0.3006796 0.2895806 -4.700392 8.6356 2.414996 0.2642448
```

## 3. Residual Check