Weather Stations Competition Report

Jun Zhang 301327368

Xin(Sarah) Hu 301306560

Xiyue(Emma) Liu 301297119

Harprit Rhandhawa 301297404

November 12, 2019

1 Introduction

As part of the STAT 440 requirement, we as a group, were required to carry out analysis of a set of data. The aim of this module is to predict aspects of the missing information from partially observed information which includes wind direction, wind speed, temperature, dew point and pressure. There are 29,000 predictions in total. Upon analysis, we obtained predictions with our best score of 0.19 in the private leaderboard.

2 Data Exploration

The data set analyzed was based on the weather data in the NOAA-UK dataset. The dataset involves 76 weather stations in the United Kingdom and include 1,880,893 observations over a 5 year period. The variables in the data set include the station number, month, day, hour, minute, wind direction, windspeed, temperature, dew point, and pressure. There are various missing values in the dataset. For the variables days and hours, values are only observed in the year 2012. In the year 2008, wind direction and wind speed appear to have large, continuous chunks of missing data. This similar pattern is also visible in the year 2010.

3 Methods

3.1 Time Series Imputation

The first method we tried to use is the ‘na.locf’ function to replace each NaN with the most recent non-NaN value prior to it. The score we obtained for this “Last Observation Carried Forward” is around 0.24815. Next, we tried the carried backward method which uses ‘na.locf’ function with the extra argument ‘fromLast = TRUE’ to carried the backward observations and replace into the missing values, and the score is 0.23939. Taking the average of the two nearest non-NaN values of both carried forward and backward will provide us a score of 0.20929.

To improve the prediction of missing values, new methods are brought in. For example, linear interpolation is a good way to replace missing values using ‘na.interpolation’. For further improvement, Kalman Smoothing is used on structural time series models for the imputation of missing values. Parameter “StructTS” is for using a structural model fitted by maximum likelihood. The score of taking averages of these methods is around 0.20419.

3.2 Linear Regression

Before applying other methods, we first reduce the original dataset to a smaller one without any missing values. Then, we splitted it into our local training and testing sets. After fitting linear regressions, we found that there are no linear relationships between the predictors, besides the variables of temperature and dew point. From the linear regression output table, we observed a R-squared value of 0.8 between temperature and dew point; however, by performing on our local test set, we concluded that a linear regression on other variables tends to have a bad performance in terms of mean absolute errors.

3.3 Stacking Different Models With Linear Regression

Similarly, we first fitted a model on wind direction using predictors of wind speed, temperature, and pressure. Here, we excluded the variable of dew point to avoid collinearity since we observed that there is a high correlation between temperature and dew point. After testing the results on our local test set, we can see from Figure 1 that with the methods of linear regression, k-nearest neighbors, regression trees, and support vector machine, we would have the mean absolute errors around 0.76 which is a pretty bad result. Even with applying ensemble methods to stack the models together, we would not improve the predictions much better since the models do not have good performances individually.

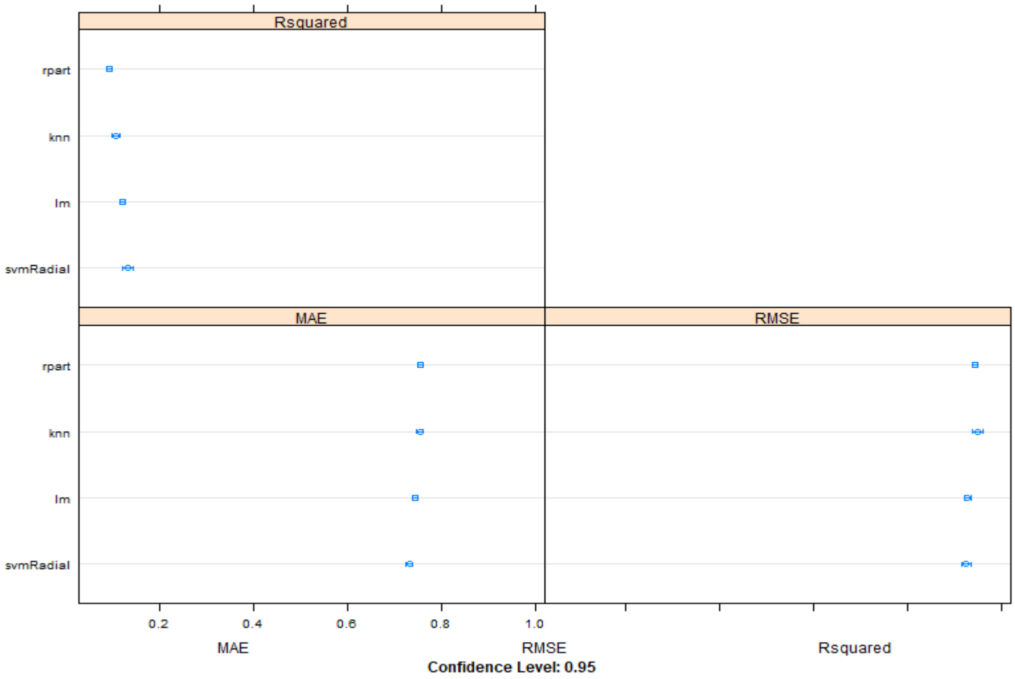


Figure 1: Performance of Different Models With the Response of Wind Direction

3.4 Time Series Imputation with Moving Averages

As we explored that the time series imputation tends to do a better job to predict the missing values, we started to think that fitting a moving average curve to the data may provide a better prediction. For example, as shown in Figure 2, the curve in black indicates the first 800 values of wind direction at station USAF 992700. It is obvious that from the graph, the line is not connected since there are many missing values. Then, we considered to fit some different moving average curves such as simple and exponential moving averages to fill the missing values. In R, under the package "imputeTS," there is a function that automatically does that for us.

Since our original dataset has over 1.8 million observations, we first divided it into 76 subsets based on their weather station numbers. One advantage of splitting the data is that if we were to fit a moving average using the original data directly, we might mistakenly used the data from the other weather stations. Since we were interested in predicting the 29000 missing values from the provided test set, we first took the indices of them and found them in the subsets that we just created. Then, for each of those 29000 indices, we searched the values backward and forward within a window width. After testing different sizes of the widow width on our local test set, we found that using a width of 20 with both backward and forward 10 numbers, provides a better prediction in terms of mean absolute errors. Also, we believed that using exponential weighted moving averages gives us a better prediction after several tests. Unfortunately, this cannot apply to all 29000 missing values. As Figure 3 indicates, we recorded the number of successful replacements as 1, the number of unsuccessful replacements as 2 and 3. Additionally, 2 means the values within the width are all missing, and 3 indicates there is only one non-missing value within the range. Even when we expanded the search width to more than 20, we were still not able to calculate the moving averages since most of the values were missing through many rows and some were not observed throughout a year. In total, we observed 3163 out of 29000 unsuccessful replacements.

The biggest challenge for us in this module is that we could not find some better ways to deal with those 3163 missing values. We tried to use our previous time series imputation methods to predict them, the mean absolute errors were always between 0.188 and 0.2. At the end, we stacked the results together which led us to get 0.1874 in the public leaderboard.

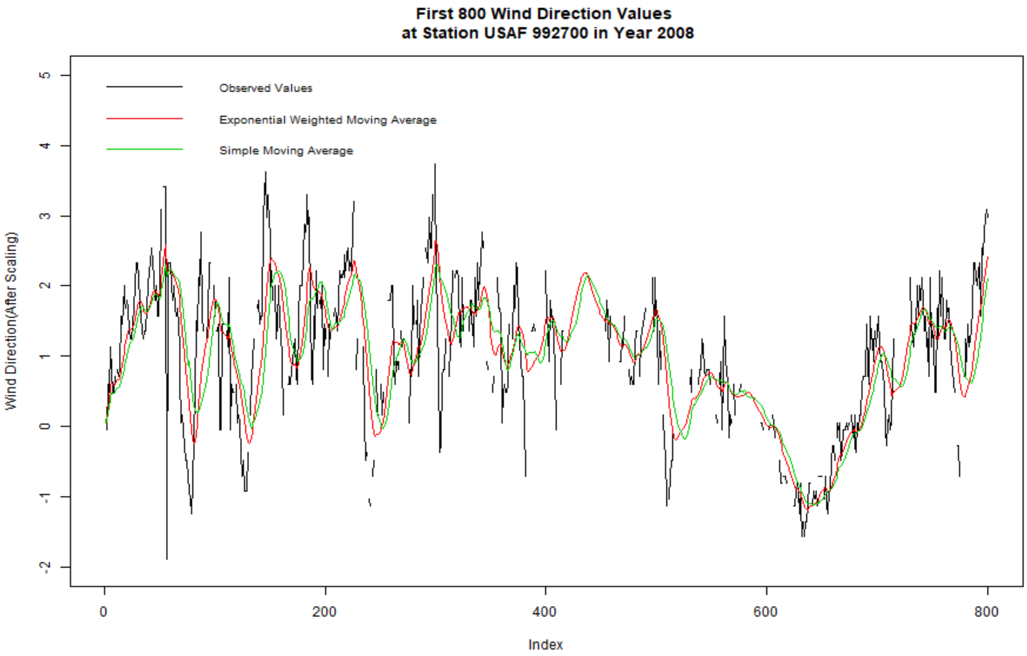
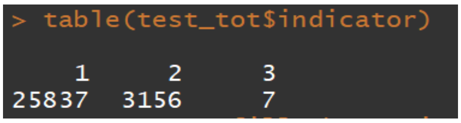


Figure 2: Time Series Plot of Wind Directions



```
> table(test_tot$indicator)

    1      2      3
25837   3156      7
```

Figure 3: Some Indicators

4 Future Ideas

Towards the end of this module, we came up with some interesting ideas on handling those 3163 missing values. If the values are not observed through many rows in one weather station, we might be able to draw some connections from its nearest stations.

From the additional "ish-history" CSV file, we found that it contains the coordinates of each weather station. We mapped these locations into a plot, as shown in Figure 4. There are 76 points in the graph and each point represents a weather station, even though some colours are used more than once. We were thinking that if we cannot handle those 3163 missing values properly, we might be able to replace them with the data from the nearest stations. Interestingly, we found some functions in R to determine the k-neighbors by the shortest distance between two points. In Figure 5, we can see the two nearest stations of the first six stations, where the first column indicates the number of stations which are sorted by their names. For example, if we were to observe that the weather station 6 has wind direction values missing for the first half of the year, then we can try to find the average values from stations 7 and 38 within that period.

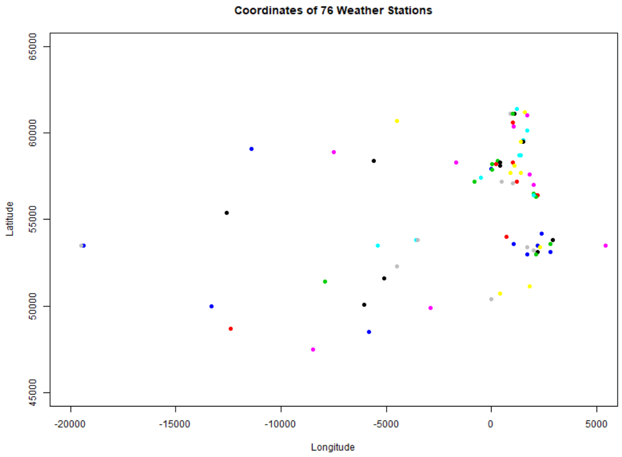
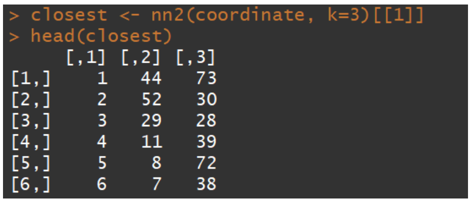


Figure 4: Coordinates of 76 Stations



```
> closest <- nn2(coordinate, k=3)[[1]]
> head(closest)
     [,1] [,2] [,3]
[1,]    1   44   73
[2,]    2   52   30
[3,]    3   29   28
[4,]    4   11   39
[5,]    5    8   72
[6,]    6    7   38
```

Figure 5: Neighbors of First 6 Stations

5 Conclusion

To put it in a nutshell, we started out with some time series imputation methods such as push forward, mean interpolation, linear interpolation, etc. And the scores we obtained for those methods were between 0.204 and 0.25. Then, we tried to use different statistical models like linear regression and k-nearest neighbors to predict the missing values. The performances of these models were not as good as the time series imputations. Our final step was to implement the moving average model with a fixed width, which produced our submitted score of 0.19. Although there were many ways we could have done to improve our score, we understand the concept and the use of different time series imputations better throughout this module.