Forecasting Spatial-Temporal Climate Data

Xirui Guo(xg2357)

Introduction:

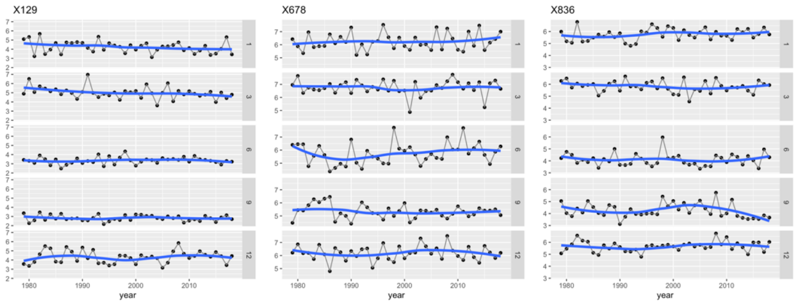
This project has three supported datasets: "WindSpeed_Month_Ave.csv"(shown later as Dataset1), "WS_month_lat_lon.csv"(shown later as Dataset2), and "lat_lon_index_key.csv" (shown later as Dataset3). Dataset1 has 480 rows to represent 480 different months from January 1979 to December 2018 and 918 columns which consist of 2 date-related columns and 916 different sites primarily over Texas, New Mexico, and Oklahoma. Dataset 2 is a 511680×4 tidy data frame that shows the wind speed for different (latitude, longitude) during the same period as Dataset1. Since 511680÷480=1066, Dataset2 shows more sites than Dataset1. Dataset3 is the smallest dataset with a structure of 916×3. In this dataset, people can get the longitude and latitude corresponding to each site. If people join Dataset2 and Dataset3 by the location coordinate, it can generate Dataset1 and show the location of each site.

The goal of the report is to predict the wind speed in the future one, two, three, four, five, and six months. This report will assume that the wind speed at each site is independent, so it is possible to only use Dataset1 in the analysis process. It will generate 916 different time series models for these 916 different sites. However, since the forecast process follows similar steps, this report will randomly choose three sites and show the whole process from data training to get the result of the prediction.

Exploratory Data Analysis:

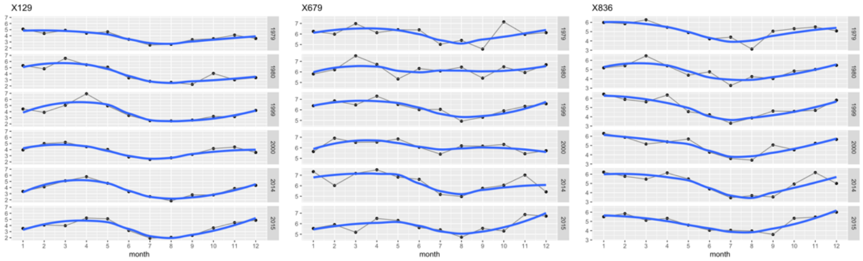
When using $set.seed(1)$ and sample function to randomly get three numbers from 1 to 916 without replacement, people can get 129, 679, and 836. In the below data exploratory process, this report mainly focuses on site X129, X679, and X836.

- The relationship among same month in different years



By extracting each January, March, June, September, and December from 1979 to 2018 and using loess smoother, the speed winds in different months barely change between each year. So, it is reasonable to recognize the dataset has no significant trend.

- The relationship among different months within one year



By extracting the wind speeds from 1979, 1980, 1999, 2000, 2014, and 2015, the data represent a curve change in its loess smoother. The first six months have a higher wind speed than the remaining months. From June to August, the speed decreases and after that, the speed becomes faster. Through observation, the data may have sin, cosine, or a combination of sin and cosine relationships. Data will be found a suitable linear regression model for de-seasonality in the data training part.

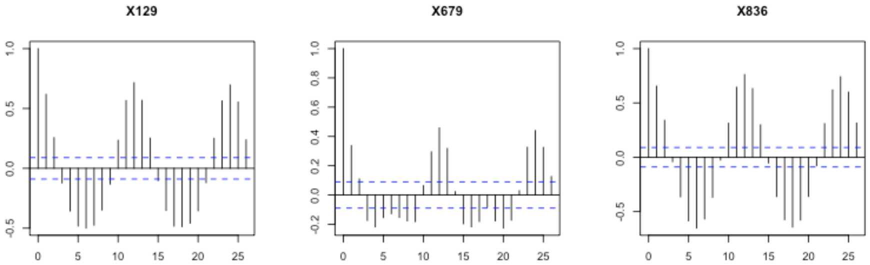
Data Training:

- Adding two new columns: “Date” and “Index”

The “Date” column, shown as date type column, combines the “year” and the “month” columns in Daraset1. It is useful for drawing the time series lines chart when people use the package ggplot2. The “Index” column can assist to generate self-defined function in code.

- De-seasonality

In the EDA part, data are found to have seasonality. Drawing ACF plots to check it:



These ACF plots confirm the appearance of seasonality. Since the loess smoothers of data show regular fluctuating curve, this report considers trigonometric functions as de-seasonal models:

model 1: $X = \sin(k * t)$

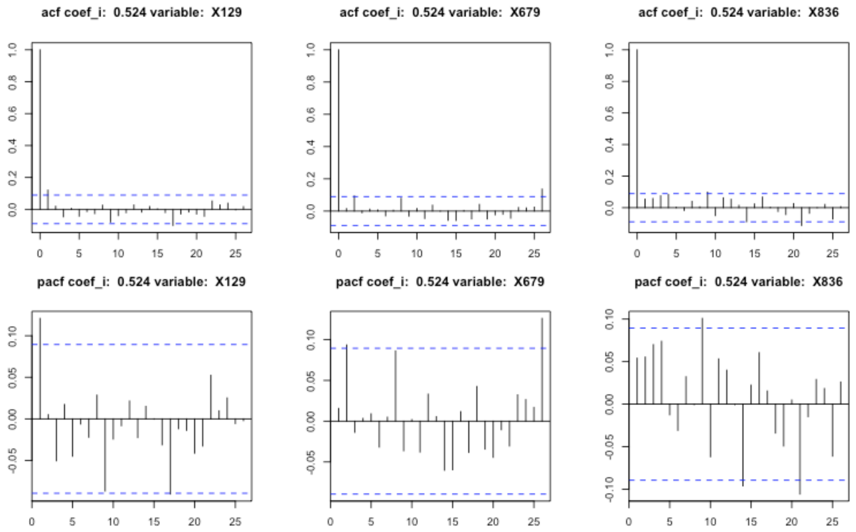
model 2: $X = \cos(k * t)$

model 3: $X = \sin(k * t) + \cos(k * t)$

model 4: $X = \sin(k * t) + \cos(k * t) + t^2$

model 5: $X = \sin(k * t) + \sin(2 * k * t) + \cos(k * t)$

Choosing hyperparameter k in the range [0, 1] with break 0.001 and the chosen k can let the model has the smallest residual standard error. Model 5 has the expectable ACF and PACF plots:



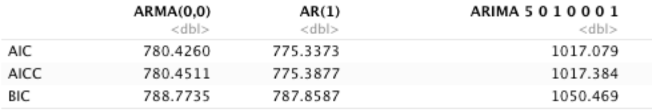
According to these plots, people can find for site X129, its ACF is exponentially close to zero and its PACF is cut off at h=1, consider using $AR(1)$; for both site X679 and site X836, ACF is exponentially close to zero and PACF is cut off at h=0, consider using ARMA(0,0).

Model selection:

$AR(1)$ and $ARMA(0,0)$ are two possible time series models for the three sites. Also, R supports a strong model selection syntax $auto.arima()$ which can automatically find the best parameters in a certain range of the input with the smallest AIC/AICC/BIC. The output of $auto.arima(data)\$arma$ represents (AR, MA, SAR, SMA, period, d, D) which are the components of the complete model $arima(AR, d, MA)(SAR, D, SMA)[period]$. In the model selection part, this report will compare three different models: $AR(1), ARMA(0,0)$, and the model selected by $auto.arima()$. One thing that needs to pay attention is AR(1) and ARMA(0,0) use the residual from the $X = sin\,(0.524 * t) + sin\,(2 * 0.524 * t) + cos\,(0.524 * t)$, but the model selected by syntax uses the original data because the syntax can recognize the best period parameter. The indicators for evaluating the quality of the model are AIC, AICC, and

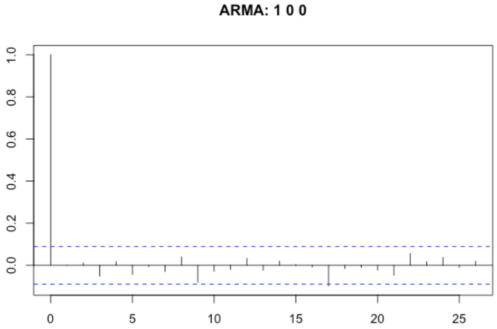
BIC. The smallest value for the indicator, the better performance for the model. The report will mainly focus on three columns "X129", "X679" and "X836" and represent the details

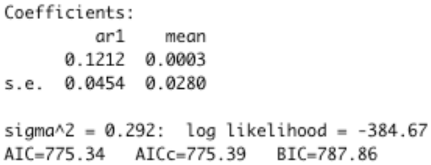
- "X129" (the 131st column in the Dataset1)



	ARMA(0,0) <dbl>	AR(1) <dbl>	ARIMA 5 0 1 0 0 0 1 <dbl>
AIC	780.4260	775.3373	1017.079
AICC	780.4511	775.3877	1017.384
BIC	788.7735	787.8587	1050.469

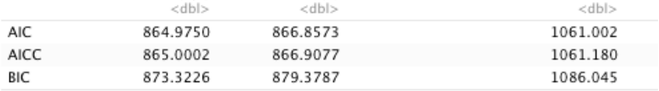
The last column represents the model selected by $auto.arima()$. The model parameters show in order $[AR, d, MA, SAR, D, SMA, period]$. For site X129, the syntax chooses the mode $ARMA(5,1)$. According to the corresponding ACF and PACF plot, AR(1) is the chosen model. $AR(1)$ has the smallest value among AIC, AICC and BIC. Also, the below ACF plot checks the model is good. Although $AR(1)$ has the smallest value in indicators, the AIC value of $ARMA(0,0)$ has a tiny difference from $AR(1)$. In this case, $AR(1)$ still is the final model for forecasting the future wind speed. The $AR(1)$ model coefficients are ar1= 0.1212 and mean=0.0003.





```
Coefficients:
         ar1    mean
      0.1212  0.0003
s.e.  0.0454  0.0280

sigma^2 = 0.292:  log likelihood = -384.67
AIC=775.34   AICc=775.39   BIC=787.86
```

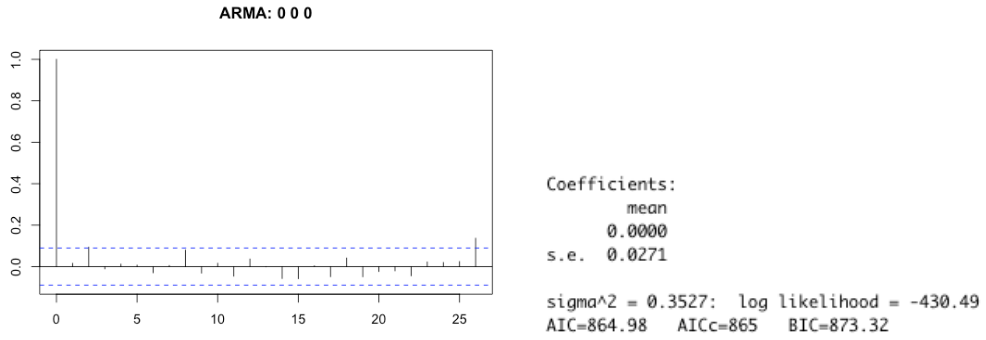
- "X679" (the 681st column in the Dataset1)



	ARMA(0,0) <dbl>	AR(1) <dbl>	ARIMA 3 0 1 0 0 0 1 <dbl>
AIC	864.9750	866.8573	1061.002
AICC	865.0002	866.9077	1061.180
BIC	873.3226	879.3787	1086.045

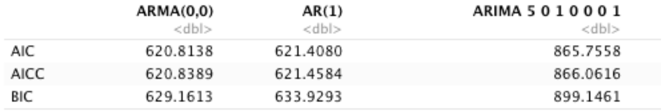
$auto.arima()$ chooses $ARMA(3,1)$ in the site X679. $ARMA(0,0)$ is the expectable model used in site X679 from previous observations. After comparing the three models,

$ARMA(0,0)$ has the smallest value of three indicators. The below ACF plot also confirm it is a suitable model. The only parameter mean in this model is also 0. People also need to notice that the difference in AIC between $AR(1)$ and $ARMA(0,0)$ is small.



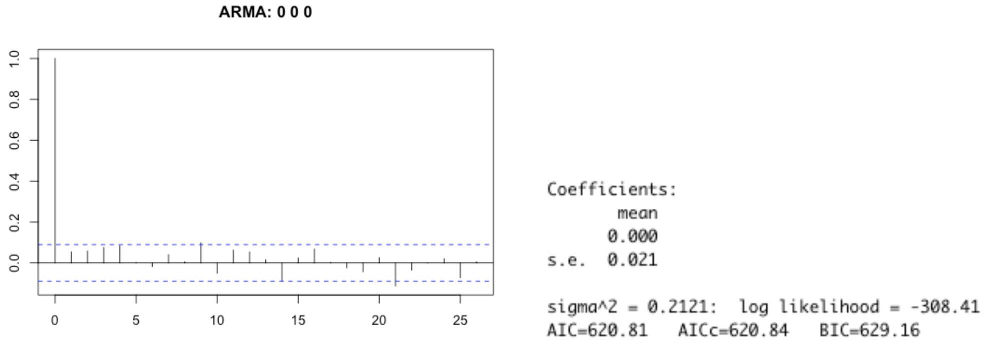
```
Coefficients:
        mean
      0.0000
s.e.  0.0271

sigma^2 = 0.3527:  log likelihood = -430.49
AIC=864.98   AICc=865   BIC=873.32
```

- “X836” (the 838th column in the Dataset1)



	ARMA(0,0) <dbl>	AR(1) <dbl>	ARIMA 5 0 1 0 0 0 1 <dbl>
AIC	620.8138	621.4080	865.7558
AICC	620.8389	621.4584	866.0616
BIC	629.1613	633.9293	899.1461

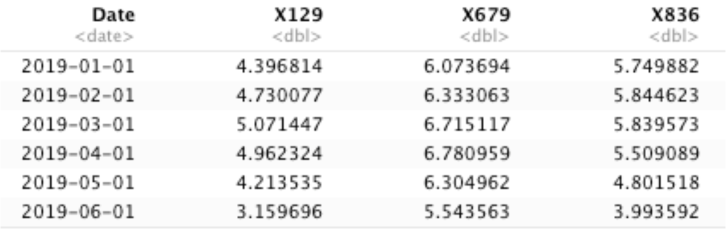
$ARMA(5,1)$ is the syntax selected model. People select $ARMA(0,0)$ by observing the ACF and PACF plot of the residual after using trigonometric functions to delimitate the seasonality of data. Form both indicator comparison data frame and the ACF plot, $ARMA(0,0)$ is an efficient time series model for site X836. Same as site X679, the AIC, AICC, and BIC value for $AR(1)$ is close to the $ARMA(0,0)$.



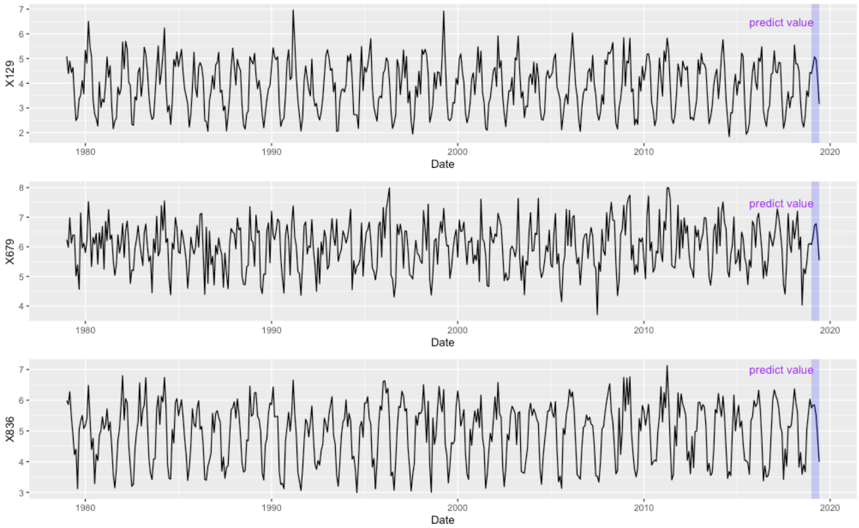
```
Coefficients:
       mean
      0.000
s.e.  0.021

sigma^2 = 0.2121:  log likelihood = -308.41
AIC=620.81   AICc=620.84   BIC=629.16
```

Conclusion:

In the model selection part, site X129, X679, and X836, the corresponding forecasting model are $AR(1)$, $ARMA(0,0)$, and $ARMA(0,0)$. $ARMA(0,0)$ represents the white noise process. The residual of the de-seasonal model is white noise, indicating that the model fits well, and the residual part is pure random data that cannot be captured. Thus, for site X679 and X836, the future 6 months' wind speed can directly use $X = \sin(0.524 * t) + \sin(2 * 0.524 * t) + \cos(0.524 * t)$ with t equal to 481-486 to predict. Since the residual of the de-seasonal model is $AR(1)$ on site X129, the predicted result should consist of two parts: the results from the de-seasonal model add the predicted results from $AR(1)$. The predicted results and corresponding plots show:



Date <date>	X129 <dbl>	X679 <dbl>	X836 <dbl>
2019-01-01	4.396814	6.073694	5.749882
2019-02-01	4.730077	6.333063	5.844623
2019-03-01	5.071447	6.715117	5.839573
2019-04-01	4.962324	6.780959	5.509089
2019-05-01	4.213535	6.304962	4.801518
2019-06-01	3.159696	5.543563	3.993592



The plots check these predicted results are possible. One thing that needs to notice is the AIC has a small difference between $ARMA(0,0)$ and $AR(1)$. If people want to generate one model for these three sites, $ARMA(0,0)$ is suitable because it would reduce the workload for prediction. The predictions for other sites repeat the data training and model selection steps and can easily get the results. So, this report treats site X129, X679, and X836 as representatives and show details for the predictions.