University of L'Aquila

Department of Information Engineering and Computer Science and Mathematics

Masters Degree in Applied Data Science

TIME SERIES WITH APPLICATIONS ON BIG DATA

Windspeed Forecast in Delhi

Hassanat Oluwatobi Awodipe
280020

Contents

Chapter 1

Introduction

In this project, the daily climate time series data for Delhi is analysed using different **ARIMA** models in **GretL**. The best model is then identified and used to create a forecast.

The variables of interest are dates and wind speed (measured in km/h). The data was collected from the 1st of January to the 24th of April in 2017. Although this is daily data, it is assumed that the season is constant since it is for four months.

Firstly, the wind speed is plotted against time to identify the data trend. Time is the independent variable while windspeed is the dependent variable. To make the windspeed series stationary, the data is transformed by taking the first difference. Next, possible orders of p and q are selected using a correlogram and the *armax* function. This makes it easy to create the models.

Using only observations from the Train set, each of the models is fitted. The models are then tested with the Test data and used to produce a forecast for wind speed. The forecasted value is compared with the actual windspeed value to determine each model's RMSE (Root Mean Squared Error) and choose the best model.

Finally, the best model is used to create a prediction for the next 20 days.

The data used can be downloaded from Kaggle.

Chapter 2

Preliminary Analysis

Here the general trend of the data is shown. It is seen that the windspeed has a general upward trend between January and April however there are periods of low wind speed particularly in February. It may also be seen that the highest windspeed was in the middle of April.

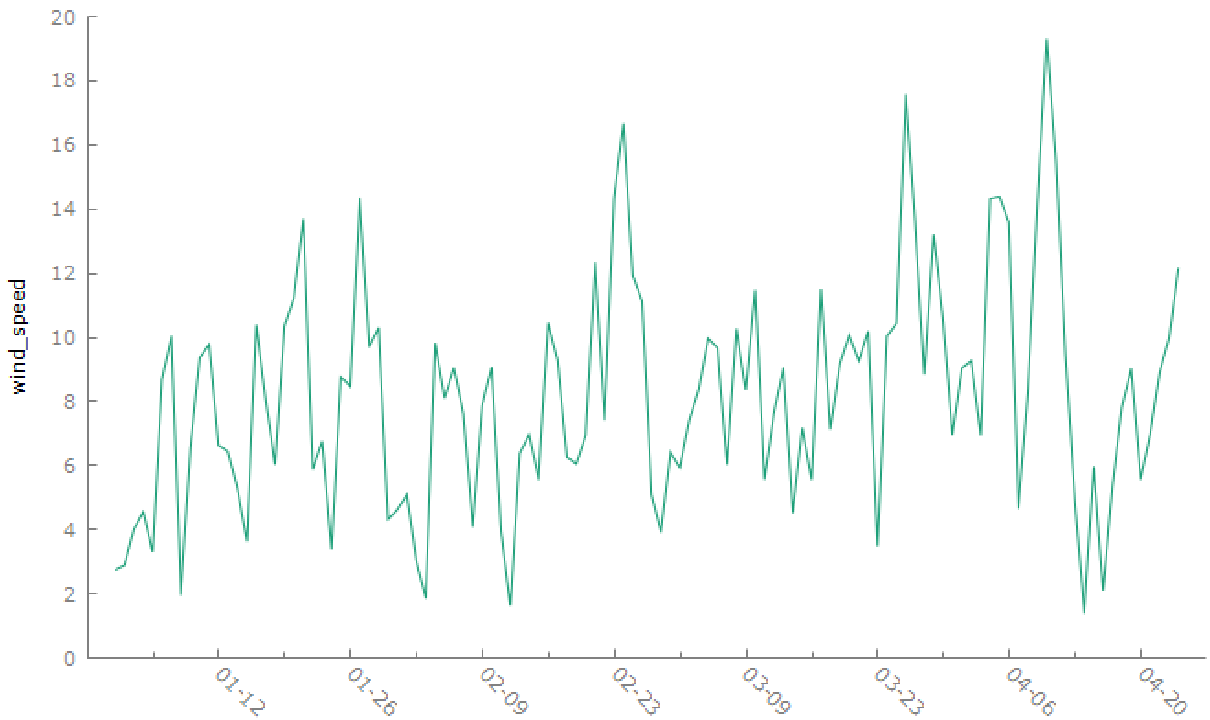


Figure 2.1: Windspeed against Time

Due to this systematic increase over time, the first difference is taken so that the data can achieve stationarity i.e. fluctuate around a constant mean. This makes it easier to model and forecast.

Chapter 3

Data Transformation

In the previous chapter, a trend was recognised in the data. A key assumption for time series models is that they must be stationary hence the first difference of the data has to be calculated.

Given a time series x_t, and the lag operator $\Delta = 1 - L$

The first difference is defined as

$$\Delta x_t = x_t - x_{t-1}$$

where

Δx_t is the differenced series

x_t is the value of the windspeed at time t

x_{t-1} is windspeed at $t - 1$

Below is the graph of windspeed against time after the first difference is taken.

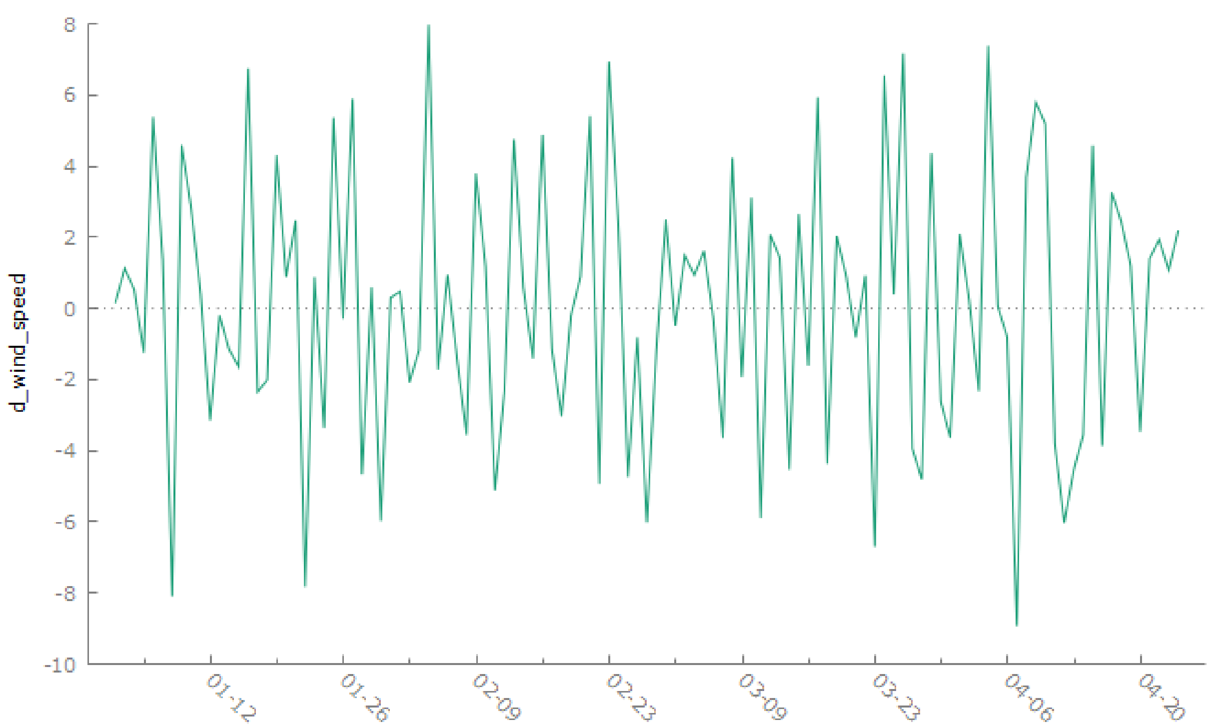


Figure 3.1: first difference of windspeed against time

It is noticed that the windspeed appears stationary around 0. The summary statistics for the first difference series are given below and it can be noticed that the mean of the data is different from 0.

Table 3.1: Summary statistics for the first difference of windspeed

```
Summary statistics, using the observations 2017-01-01 - 2017-04-24
for the variable 'd_wind_speed' (113 valid observations)

  Mean                          0.083306
  Median                         0.39170
  Minimum                        -8.9278
  Maximum                         7.9655
  Standard deviation              3.7396
```

In addition, the Augmented Dicker Fuller (ADF) test was applied to the first differenced series. This is a test used to determine whether a time series data is stationary. The ADF test has a null hypothesis H0 which states that the time series has a unit root i.e non-stationary. The Python function *adfuller* from the *statsmodel* package was used to carry out this test and the result below was obtained.

Table 3.2: ADF test results

	adf_result	values
0	Test Statistic	-7.3679
1	p-value	9.1335e-11
2	Number of Lags Used	5
3	Number of Observations Used	107
4	Critical Values	{'1%': -3.493, '5%': -2.889, '10%': -2.5814}
5	IC Best	533.1258

The test statistic is less than the critical value at 1%, 5% and 10% significance values indicating that the null hypothesis can be rejected. Furthermore, the p-value is less than 0.05 significance level. Therefore, the first differenced series is stationary, and the data requires no further transformation. The number of lags used, 5, was chosen to minimise the Information Criteria (IC) value of 533 which in this case represents the Akaike IC (**AIC**) value.

Chapter 4

Model Identification

Firstly, the correlogram with a maximum lag length of 10 is plotted for the first differenced data. This shows the graphs for the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF). In the graph below, there is a spike at lag length 1 then it sharply diminishes at 2 for both ACF and PACF suggesting that $q=1$ and $p=1$.

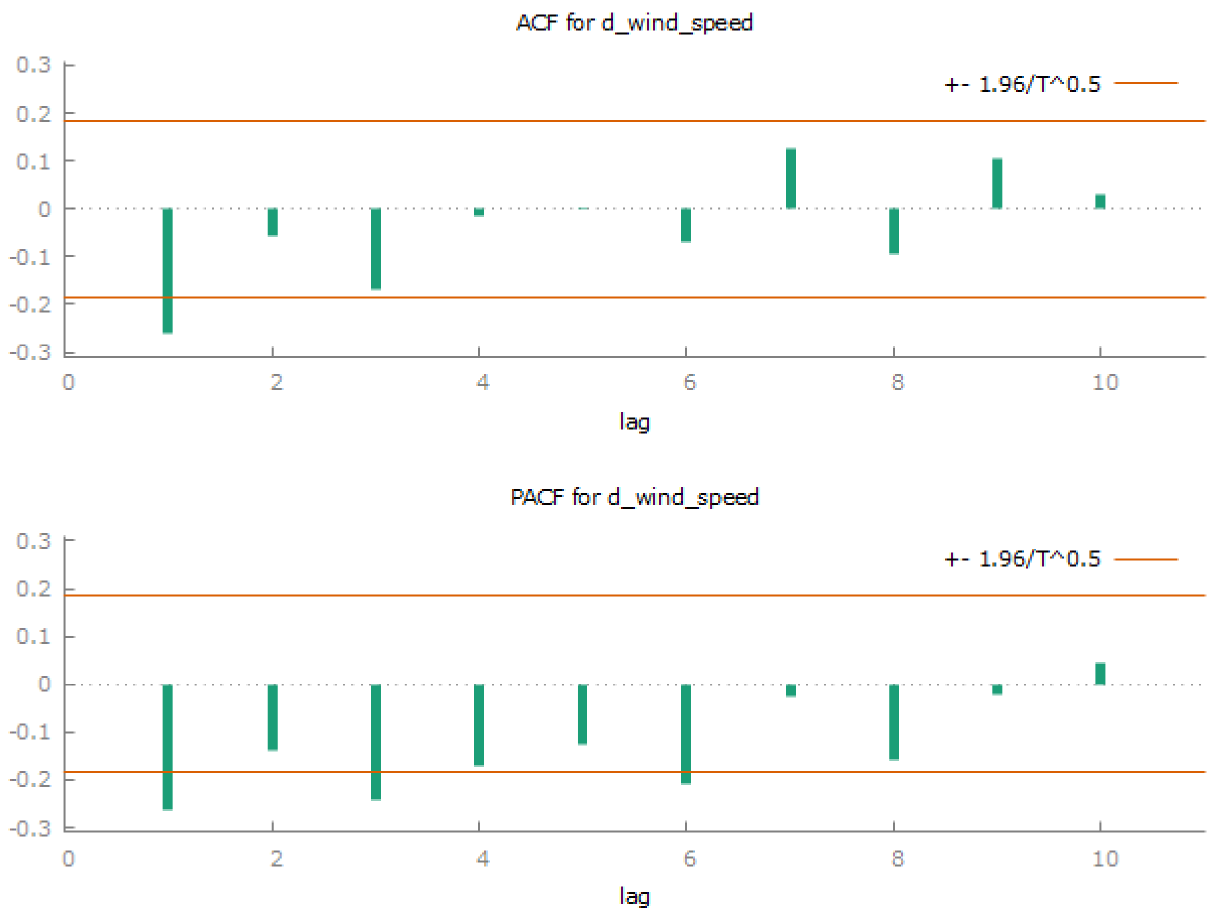


Figure 4.1: Correlogram for the first difference of windspeed

4.1 Information Criteria

Sometimes, visualising the correlogram is not enough to determine the ARIMA models to build. The GretL software has the armax function which can be used to check the Information Criteria (IC). This uses three different estimators to achieve a balance between model fit and complexity. The estimators are **AIC**, Bayesian IC (**BIC**) and Hannan-Quinn Criterion (**HQC**). The minimum values of **AIC**, **BIC** and **HQC** are usually chosen as the best model.

Based on the correlogram, the parameters for the *armax* function was set as: $(0 \leq p \leq 3, 0 \leq q \leq 3)$ and the following table was produced

Table 4.1: Information Criteria according to Lag Length

```
? armax(3, 3, d_wind_speed, null, 1, 1, 0, 1, 0)
===============================================
 Information Criteria of ARMAX(p,q) for d_wind_speed
-----------------------------------------------
 p, q          AIC          BIC          HQC
-----------------------------------------------
 0, 0     621.7660     627.2208     623.9795
 0, 1     609.8551     618.0373     613.1754
 0, 2     594.6586     605.5682     599.0856
 0, 3     591.3316     604.9686     596.8654
 1, 0     615.7644     623.9466     619.0846
 1, 1     591.1318*    602.0413*    595.5588*
 1, 2     593.0435     606.6804     598.5772
 1, 3     593.1997     609.5641     599.8402
 2, 0     615.6510     626.5606     620.0780
 2, 1     592.9886     606.6256     598.5224
 2, 2     595.0509     611.4152     601.6914
 2, 3     595.1964     614.2881     602.9436
 3, 0     610.8831     624.5200     616.4168
 3, 1     593.0676     609.4320     599.7081
 3, 2     595.0554     614.1471     602.8026
 3, 3     597.0478     618.8669     605.9018
===============================================
* indicates best models.
'9999.9999' suggests failures to estimate the models.
```

It can be observed that the *armax* function indicates $p=1$ and $q=1$ as the best model with minimum values of **AIC**, **BIC** and **HQC**. It was earlier established that $d=1$ since the first difference was applied once to make the series stationary so we have the model as

$$\Delta x_t \sim ARMA(1,1) \Rightarrow x_t \sim ARIMA(1,1,1)$$

Hence the model has the following equation:

$$(1 - \phi_1 L)(1 - L)x_t = c + (1 + \theta_1 L)u_t \tag{4.1}$$

Furthermore, two other models are considered; **AR**(1) and **MA**(1) to understand how the AutoRegressive and Moving Average parts work together to explain a given time series.

ARIMA(0,1,1)

$$(1 - L)x_t = c + (1 + \theta_1 L)u_t \tag{4.2}$$

ARIMA(1,1,0)

$$(1 - \phi_1 L)(1 - L)x_t = c + u_t \tag{4.3}$$

Chapter 5

Model Estimation

For this process, the data is divided into two parts:

Train set: 80% of total data (2017-01-01 to 2017-03-01)

Test set: remaining 20% (2017-04-01 to 2017-04-24)

5.1. First model: **ARIMA**(1,1,1)

Below is the result of applying the model to the data. The coefficients θ_1 and φ_1 and the constant are all significantly different from 0 with p-values < 0.05

*Table 5.1.1: Model 1 **ARIMA**(1,1,1) result*

```
Model 5: ARIMA, using observations 2017-01-02:2017-03-31 (T = 89)
Estimated using AS 197 (exact ML)
Dependent variable: (1-L) wind_speed
Standard errors based on Hessian

             coefficient   std. error      z       p-value
  ---------------------------------------------------------
  const       0.0423140    0.0175430      2.412   0.0159    **
  phi_1       0.323923     0.102340       3.165   0.0015    ***
  theta_1    -1.00000      0.0313546    -31.89    3.31e-223 ***

Mean dependent var    0.047264   S.D. dependent var    3.671858
Mean of innovations   0.161578   S.D. of innovations   2.969894
R-squared             0.187668   Adjusted R-squared    0.178331
Log-likelihood       -225.0835   Akaike criterion      458.1671
Schwarz criterion     468.1216   Hannan-Quinn          462.1795
```

So we can substitute their values into equation 4.1

$$(1 - 0.324L)(1 - L)x_t = 0.042 + (1 - L)u_t$$

5.1.1 Model Checking

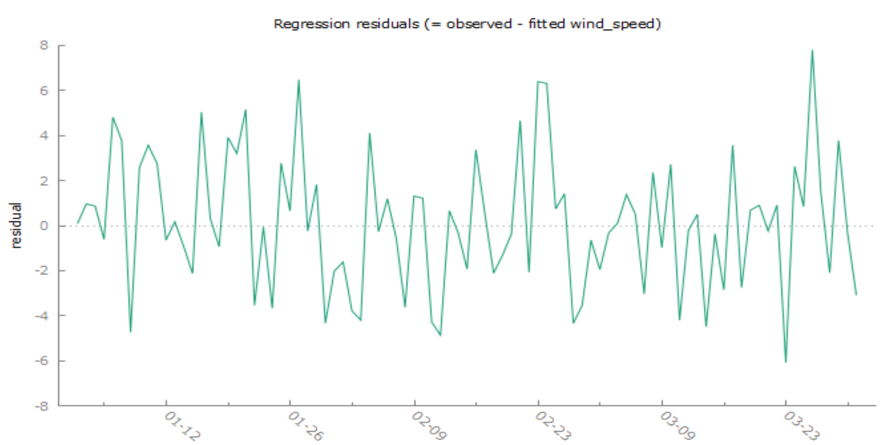


Figure 5.1.1: Residuals plotted against time

It is necessary to check the behaviour of the residuals to understand how good the model is. The residuals are analyzed to determine if it is a Gaussian White Noise (WN). In *fig 5.1.1* above, the residuals are plotted against time. It can be observed that the residuals are stationary with mean around 0.

Secondly, the correlogram below is plotted and it shows no obvious autocorrelation. All the values are close to 0 at all lags and lie within the 95% confidence interval. Hence, it can be concluded that the residuals behave like WN

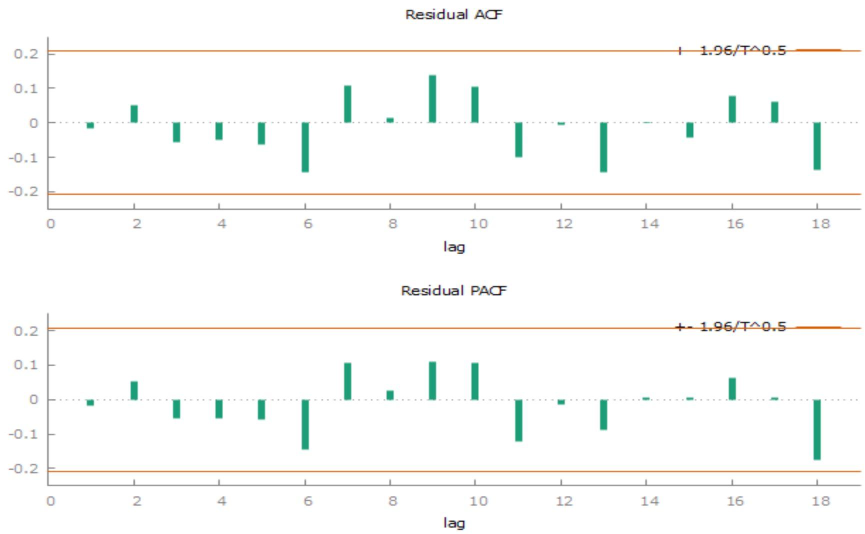


Figure 2.1.2: Residual Correlogram

Thirdly, the normality of the residuals are checked using a histogram and Q-Q plot.

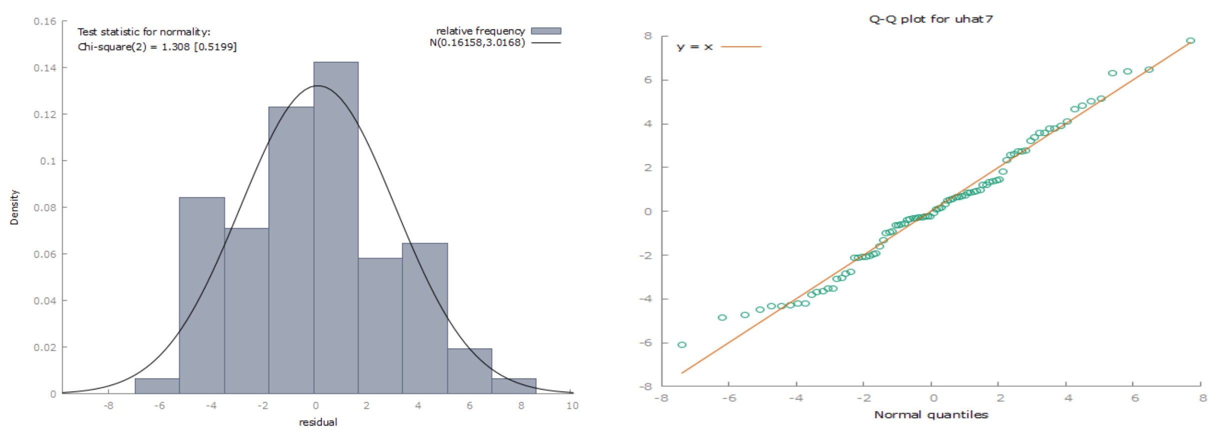


Figure 5.1.3: Normality tests for the residuals

The histogram in *fig 5.1.3* shows a p-value of 0.5199 which is greater than 0.05 indicating that the null hypothesis (no significant autocorrelation) is not rejected. In the q-q plot, most values fall on the red line. This means that the residuals follow a Gaussian distribution.

However, creating a plot of the residuals versus the actual series below shows that there seems to be a correlation. But since the residuals passes the other tests, this can be considered a good model.

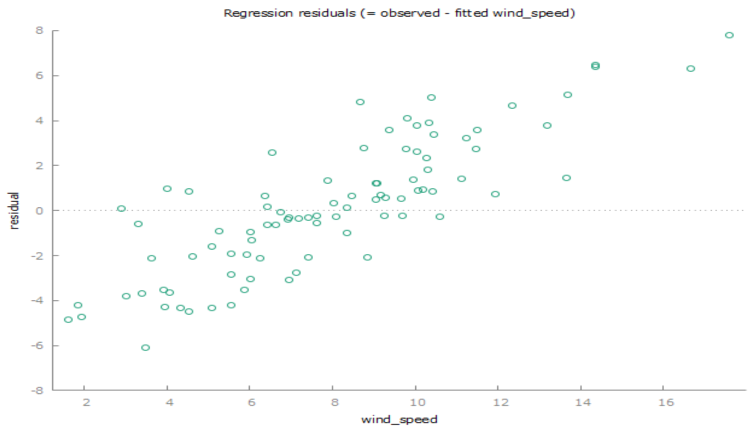


Figure 5.1.4: Residuals plotted against windspeed

5.2. Second model: **ARIMA**(0,1,1)

Applying an **MA**(1) model to the data yields the following result

Table 5.2.1: Model 2- ARIMA(0,1,1)

```
Model 3: ARIMA, using observations 2017-01-02:2017-03-31 (T = 89)
Estimated using AS 197 (exact ML)
Dependent variable: (1-L) wind_speed
Standard errors based on Hessian

             coefficient   std. error      z        p-value
  ---------------------------------------------------------
  const        0.0421542   0.0126700      3.327    0.0009     ***
  theta_1     -1.00000     0.0378550    -26.42     8.84e-154  ***

Mean dependent var    0.047264   S.D. dependent var    3.671858
Mean of innovations   0.207986   S.D. of innovations   3.122642
R-squared             0.109050   Adjusted R-squared    0.109050
Log-likelihood       -229.8779   Akaike criterion      465.7558
Schwarz criterion     473.2217   Hannan-Quinn          468.7651
```

Substitute the coefficient values into (4.2):

$$(1 - L)x_t = 0.042 + (1 - L)u_t$$

5.2.1. Model Checking

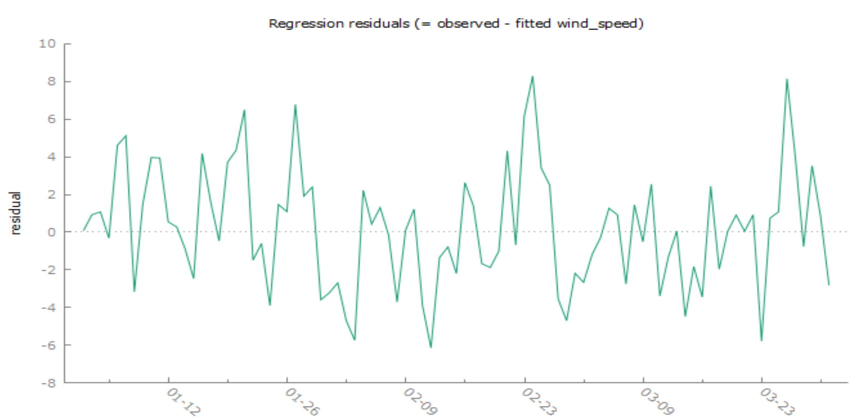


Figure 5.2.1: Residuals plotted against time

Taking a closer look at the plot of the residuals above, it appears to have a constant mean and variance.

However, at lag 1 in the correlogram, there is a value outside the confidence interval indicating that not all the residuals behave like white noise.

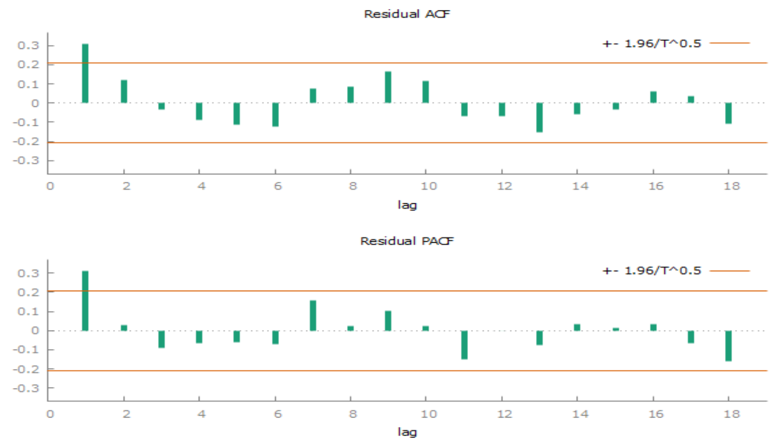


Figure 5.2.2: Correlogram of residuals

Observing the normality of the residuals below shows that they follow a Gaussian distribution and the p-value is greater than 0.05.

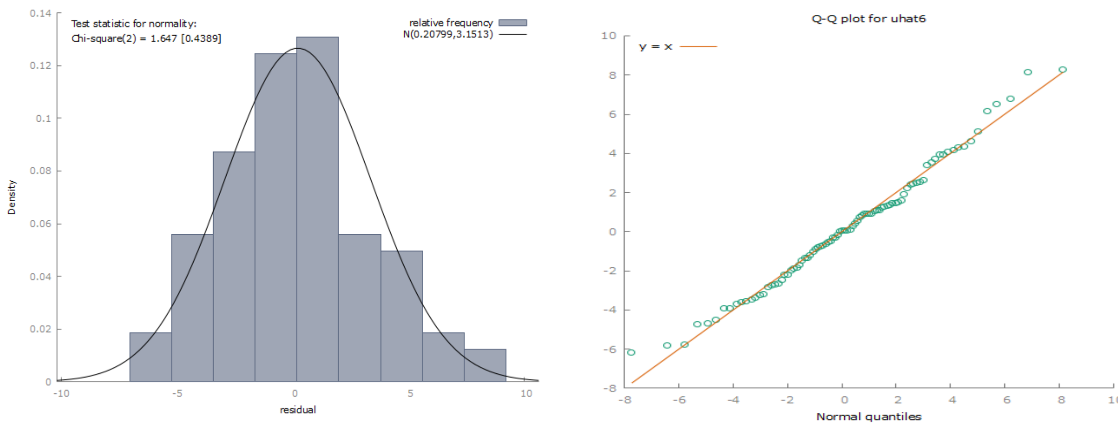


Figure 5.2.3: Normality tests for the residuals

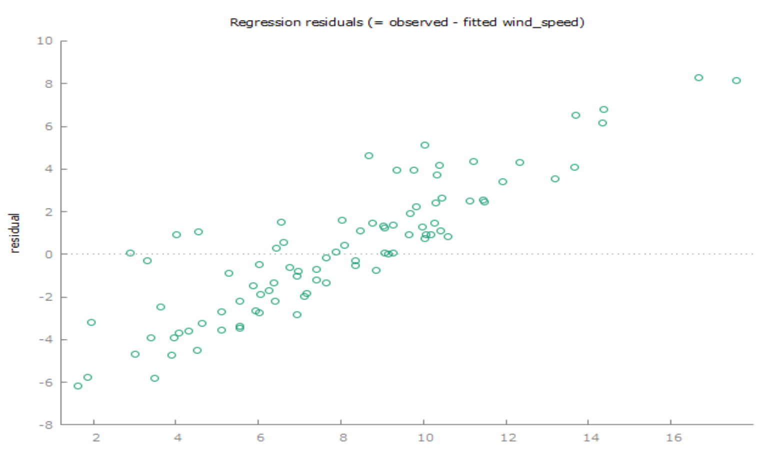


Figure 5.2.4: Residuals plotted against windspeed

However in the residuals versus the windspeed plotted above, there appears to be a correlation. The model does not pass the check for Residual correlogram and correlation with the actual series but it performs well for the other tests so it is a fairly good prediction model.

5.3. Third model: **ARIMA**(1,1,0)

Applying the **AR**(1) model to the data yields the result below

*Table 5.3.1: Model 3 - **ARIMA**(1,1,0)*

```
Model 9: ARIMA, using observations 2017-01-02:2017-03-31 (T = 89)
Estimated using AS 197 (exact ML)
Dependent variable: (1-L) wind_speed
Standard errors based on Hessian

             coefficient   std. error      z      p-value
  --------------------------------------------------------
  const       0.0578589    0.267667      0.2162   0.8289
  phi_1      -0.354815     0.0988847    -3.588    0.0003   ***

Mean dependent var    0.047264   S.D. dependent var    3.671858
Mean of innovations   0.000303   S.D. of innovations   3.410940
R-squared             0.145096   Adjusted R-squared    0.145096
Log-likelihood       -235.5547   Akaike criterion      477.1095
Schwarz criterion     484.5754   Hannan-Quinn          480.1187
```

Substitute the coefficient values into (4.3):

$$(1 + 0.355L)(1 - L)x_t = 0.058 + u_t$$

5.3.1. Model Checking

Once again, the characteristics of the residuals are checked to determine if it is a Gaussian WN. Firstly, the residuals are plotted against time.

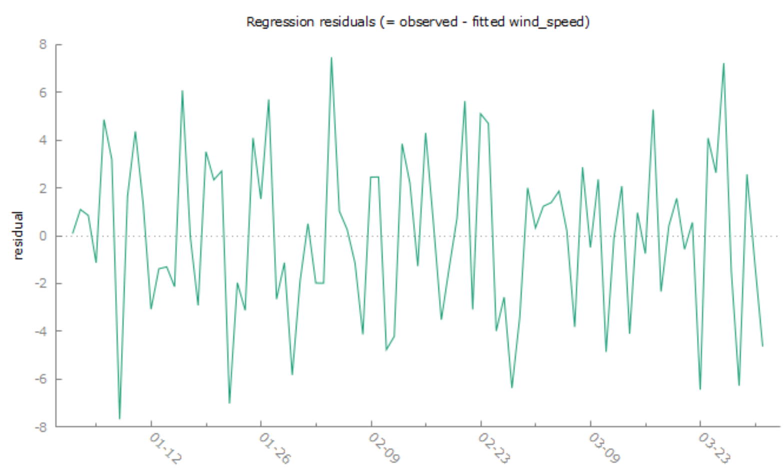


Figure 5.3.1: Residuals plotted against time

Next, the correlogram is observed below. In the Residual ACF, all values lie within the confidence interval but this is not the case in the PACF.

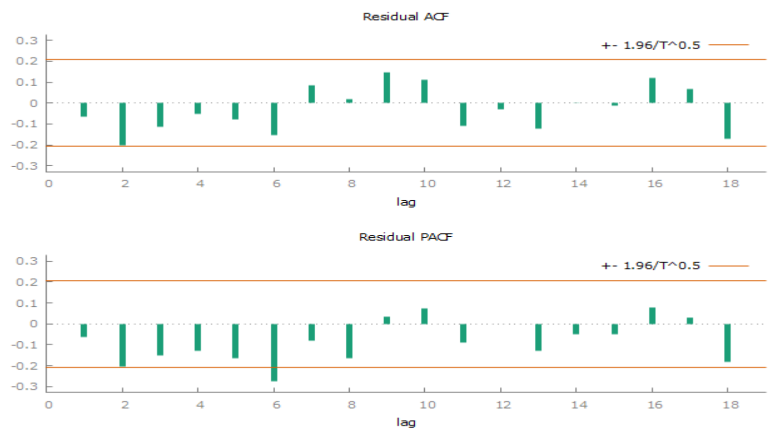


Figure 5.3.2: Correlogram of residuals

Looking at the results of the normality tests indicates that the residual is a Gaussain WN. The p-value is greater than 0.05 and majority of the residuals lie on the red line in the q-q plot.

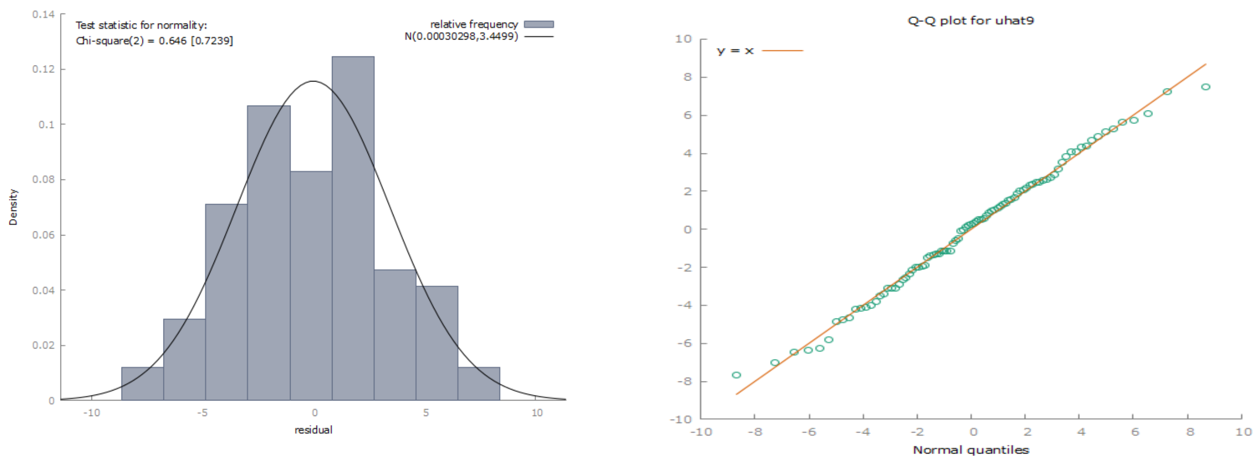


Figure 5.3.3: Normality tests for the residuals

Lastly, the residuals are plotted against the windspeed and unlike the previous models, there are no obvious correlations. Hence it can be concluded that this is a good prediction model.

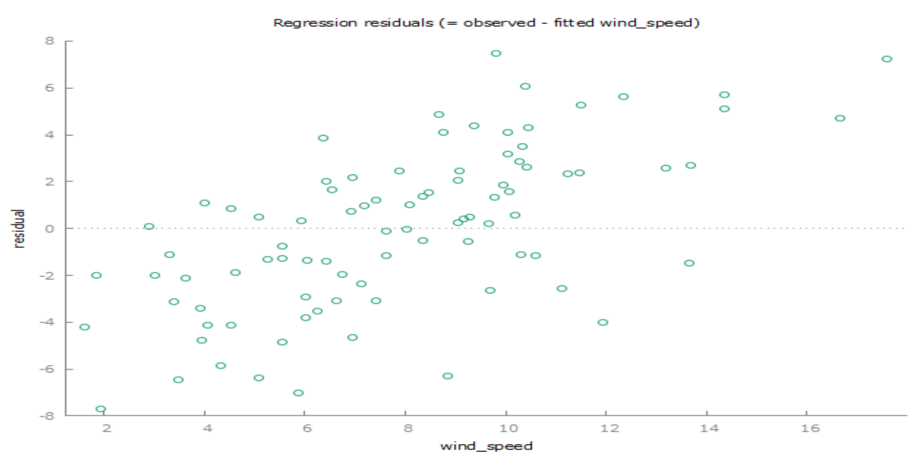


Figure 5.3.4: Residuals plotted against windspeed

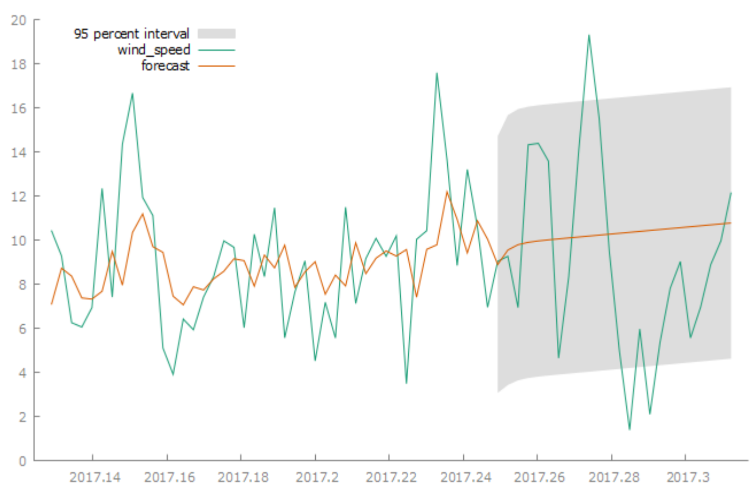
Chapter 6

Model Selection

In the previous chapter, three models were defined and fitted to the data. In this chapter, the best model is selected. The prediction of April's values is obtained by applying the models to the test set and the model with the lowest RMSE value is the best model.

Below are the graphs that compare the predicted values and the actual values and a table showing the forecast statistics for each model.

6.1 **ARIMA**(1,1,1) Forecast



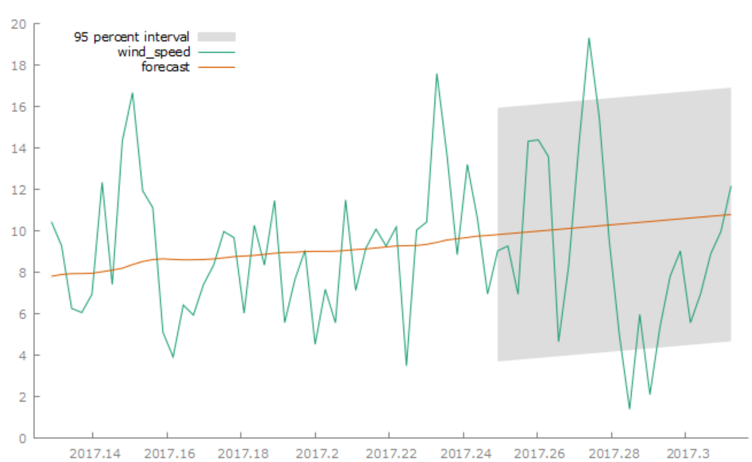
Forecast evaluation statistics using 24 observations

Mean Error	-1.102
Root Mean Squared Error	4.5597
Mean Absolute Error	3.7987
Mean Percentage Error	-63.456
Mean Absolute Percentage Error	81.004
Theil's U2	1.1619
Bias proportion, UM	0.05841
Regression proportion, UR	0.072055
Disturbance proportion, UD	0.86953

*Figure 3.1: **ARIMA**(1,1,1) forecast result*

The above graph shows that only a few points lie outside the confidence interval. The negative mean error indicates that a lower windspeed value was mostly forecasted.

6.2. **ARIMA**(0,1,1) Forecast



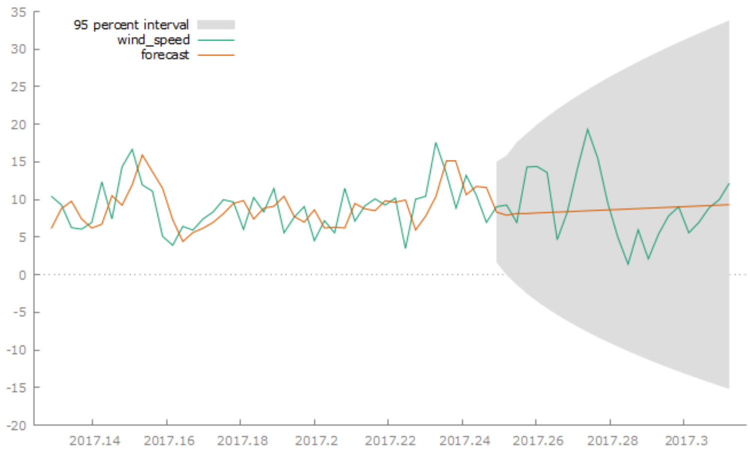
Forecast evaluation statistics using 24 observations

Mean Error	-1.1715
Root Mean Squared Error	4.5674
Mean Absolute Error	3.8441
Mean Percentage Error	-64.29
Mean Absolute Percentage Error	81.621
Theil's U2	1.1647
Bias proportion, UM	0.065784
Regression proportion, UR	0.099958
Disturbance proportion, UD	0.83426

*Figure 6.2: **ARIMA**(0,1,1) forecast result*

Fig 6.2 is very similar to Fig 6.1 however the latter is better because it has a lower RMSE of 4.5597. **ARIMA**(0,1,1) had an RMSE of 4.5674**.**

6.3. **ARIMA**(1,1,0) Forecast




```
Forecast evaluation statistics using 24 observations

Mean Error                           0.47734
Root Mean Squared Error              4.4778
Mean Absolute Error                  3.4942
Mean Percentage Error              -38.282
Mean Absolute Percentage Error      65.689
Theil's U2                           0.86131
Bias proportion, UM                  0.011364
Regression proportion, UR            0.12714
Disturbance proportion, UD           0.8615
```

Figure 6.3: ARIMA(1,1,0) forecast result

The third model produces a better result than the other two models. Its RMSE value of **4.478** is the lowest of all models. This suggests that the **MA** part of the model can affect prediction performance. Hence, **ARIMA**(1,1,0) is identified as the best model and will be used for forecasting.

Chapter 7

Forecast and Conclusion

The data is restored to the full range (2017-01-01 to 2017-04-24) and the model **ARIMA**(1,1,0) is applied.

*Table 7.1: Best model **ARIMA**(1,1,0)*

```
Model 10: ARIMA, using observations 2017-01-02:2017-04-24 (T = 113)
Estimated using AS 197 (exact ML)
Dependent variable: (1-L) wind_speed
Standard errors based on Hessian

             coefficient   std. error      z       p-value
  ---------------------------------------------------------
  const       0.0793046    0.268567      0.2953   0.7678
  phi_1      -0.260645     0.0904446    -2.882    0.0040   ***

Mean dependent var   0.083306   S.D. dependent var   3.739620
Mean of innovations  0.000144   S.D. of innovations  3.592408
R-squared            0.189298   Adjusted R-squared   0.189298
Log-likelihood      -304.8822   Akaike criterion     615.7644
Schwarz criterion    623.9466   Hannan-Quinn         619.0846
```

Then the windspeed for the next 20 days is predicted.

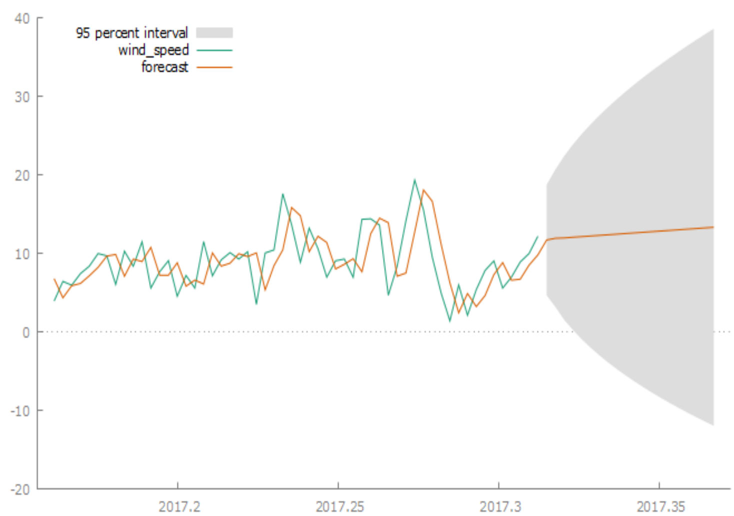


Figure 7.1: Graph of actual vs predicted windspeed

2017-04-25	11.6851
2017-04-26	11.9081
2017-04-27	11.9499
2017-04-28	12.0390
2017-04-29	12.1157
2017-04-30	12.1957
2017-05-01	12.2748
2017-05-02	12.3542
2017-05-03	12.4335
2017-05-04	12.5128
2017-05-05	12.5921
2017-05-06	12.6714
2017-05-07	12.7507
2017-05-08	12.8300
2017-05-09	12.9093
2017-05-10	12.9886
2017-05-11	13.0679
2017-05-12	13.1472
2017-05-13	13.2265
2017-05-14	13.3058

Table 7.2: Windspeed forecast

It can be observed that the windspeed values are expected to increase gradually over the next 20 days. However, this will likely be affected by other climatic factors such as rainfall and humidity which the forecast did not consider so the increase will not be a linear one and there will be periods of windspeed decrease.