MOVING AVERAGE MODEL OF TIME SERIES

A. Introduction of Time Series

Time Series is a collection of data points collected at constant time intervals and is usually analyzed to determine long-term trends and to predict the future or perform other forms of analysis.

Time Series is time dependent and along with an increasing or decreasing trend, some Time Series have some form of seasonality trends which we can analysis.

B. Introduction of Moving Average

A Moving Average is a calculation to analyze data points by creating series of averages of different subsets of the full data set. Given a series of numbers and a fixed subset size, the first element of the moving average is obtained by taking the average of the initial fixed subset of the number series.

1) Simple Moving Average: The simplest form of a moving average, appropriately known as a simple moving average (SMA) [1], is calculated by taking the arithmetic mean of a given set of values. In other words, a set of numbers, are added together and then divided by the number in the set.

The way it moves is that the new values become available while the oldest data points must be dropped from the set. Since new data points come in to replace them, the data set is constantly "moving" to account for new data. This method of calculation ensures that only the current information is being accounted for.

2) Exponential Moving Average: In order to avoid the problem that each point in the data series uses the same weighting, regardless of where it occurs in the sequence, we can use a new kind of moving average calculation method - Exponential Moving Average. In this way, the most recent data is more significant than the older data and have a greater influence on the final result. Below is the EMA equation [2]:

$$EMA = (P * \alpha) + (Previous\ EMA * (1 - \alpha))$$

$$P = Current\ Price$$

$$\alpha = Smoothing\ Factor = \frac{2}{1 + N}$$

$$N = Number\ of\ Time\ Periods$$

There is no value available to use as the previous EMA and this small problem can be solved by starting the calculation with a simple moving average and continuing on with the above formula from there.

3) Weighted Moving Average: Exponential Moving Average is a kind of Weighted Moving Average since it places more emphasis on recent data.

Weighted Moving Average is a weighted average of the last n prices, where the weighting decreases with each previous price. This is a similar concept to the EMA, but the calculation for the WMA is different [3].

$$(Price\ X\ weighting\ factor) + (Price\ previous\ period\ X\ weighting\ factor - 1)$$

For example, if there are four prices you want a weighted moving average of, then the most recent weighting could be 4/10, the period before could have a weight of 3/10, the period prior to that could have a weighting of 2/10, and so on. 10 is a randomly picked number, and a weight of 4/10 means the most recent price will account for 40% of the value of the WMA. The price three periods ago only accounts of 10% of the WMA value.

C. Loading and Handling Time Series in Pandas

Pandas has proven very successful as a tool for working with time series data. Using timedelta64 dtypes, we can easily manipulate the time series data.

After reading the csv file using pandas, we transfer the data type of 'date_time' column to datetime64[ns] and set it as index to help us analysis.

Since the original data is too large for time series analysis, we resample the 'uber_price_per_second' to 2 hours period using the mean of them.

D. Stationary

A Time Series is said to be Stationary if its statistical properties such as mean, variance remain constant over time. Since most of the Time Series models work on the assumption that the Time Series is stationary. If a Time Series has a particular behavior over time, there is a very high probability that it will follow the same in the future.

Also, the theories related to stationary series are more mature and easier to implement as compared to non-stationary series.

1) Evaluation: For practical purposes, we can assume the series to be stationary if it has constant statistical properties over time. For example:

a) Constant Mean: The mean of the series should not be a function of time rather should be a constant.

b) Constant Variance: The variance of the series should not a be a function of time. This property is known as homoscedasticity.

c) An auto covariance that does not depend on time: The covariance of the i th term and the (i + m) th term should not be a function of time.

When we plot the price per second of Uber price, we cannot observe the stationary directly. So, we define a function to test time series' stationary. Since we resample the data as 2-hour period, we set the window as 12 because we want to take one day for 24 hours. Then we will plot the original data, rolling mean and rolling standard of the time series we want to test. Also, in order to know the stationary in a more statistical way, we also use the Dickey-Fuller test. In statistics, the Dickey–Fuller test tests the null hypothesis that a unit root is present in an autoregressive model. The alternative hypothesis is stationarity. The regression model of Dickey–Fuller test can be written as [4]

$$\Delta y_t = (\rho - 1) y_{t-1} + \mu_t = \delta y_{t-1} + \mu_t$$

where Δ is the first difference operator. This model can be estimated and testing for a unit root is equivalent to testing $\delta = 0$ (where $\delta = \rho - 1$). Since the test is done over the residual term rather than raw data, it is not possible to use standard t-distribution to provide critical values. Therefore, this statistic t has a specific distribution simply known as the Dickey–Fuller table.

The test results comprise of a test statistic and some critical values for difference confidence levels. If the test statistic is less than the critical value, we can reject the null hypothesis and say that the series is stationary.

The stationary plot of our Uber price is shown as Fig.1:

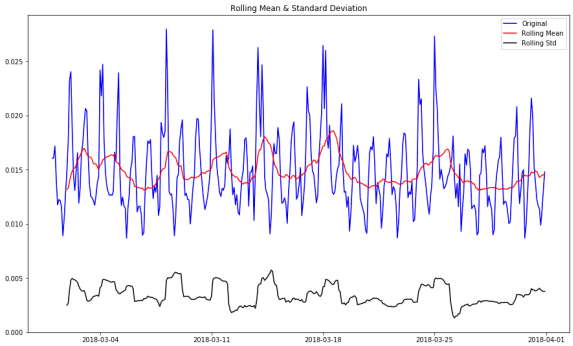


Fig.1 stationary of Uber price

The stationary result of our Uber price is shown as Table.1:

TABLE 1. UBER PRICE STATIONARY

Test Statistic	-4.938296
p-value	0.000029
#Lags Used	17.000000
Number of Observations Used	354.000000
Critical Value (1%)	-3.448958
Critical Value (5%)	-2.869739
Critical Value (10%)	-2.571138

The Test Statistic is -4.938296, which is less than Critical Value (1%) -3.448958. So, we have 99% confidence to say that the data is stationary.

However, we can still see some Seasonality like fluctuation in our dataset. So, we will do some effort to make our data more stable.

Usually, if a Time Series is non-stationarity, there are usually two reasons:

The first one is trend, it means varying mean over time. For example, the price grows over time. The second one is seasonality variations at specific time-frames. For example, people might have a tendency to buy cars in a particular month because of pay increment or festivals. Then we need to eliminate the impact of these two elements.

E. Eliminating Trend and Seasonality

1) Reduce trend: we can try to reduce trend using transformation which penalize higher values more than smaller values. We take the log transform method and then use moving average to remove the trend from the log data. After this, we test the stationary again, the result is shown below as Table.2:

TABLE 2. LOG MOVING AVERAGE

Test Statistic	-6.530094e+00
p-value	9.919489e-09
#Lags Used	1.700000e+01
Number of Observations Used	3.430000e+02
Critical Value (1%)	-3.449560e+00
Critical Value (5%)	-2.870004e+00
Critical Value (10%)	-2.571279e+00

The test statistic is smaller than the 1% critical values so we can say with 99% confidence that this is a stationary series. Also, the p value reduces a lot.

Also, we try Weighted Moving Average to see if there is any improvement and the stationary result is shown below as Table.3:

TABLE 3. LOG WEIGHTED MOVING AVERAGE

Test Statistic	-5.936612e+00
p-value	2.312900e-07
#Lags Used	1.700000e+01
Number of Observations Used	3.540000e+02
Critical Value (1%)	-3.448958e+00
Critical Value (5%)	-2.869739e+00
Critical Value (10%)	-2.571138e+00

The stationary does improve but not much and for the predict purpose, we will use simple moving average method which is efficient and easy to remove the trend.

2) Reduce seasonality: There are two ways of removing trend and seasonality:

a) Differencing: Taking the difference with a particular time lag.

In this technique, we take the difference of the observation at a particular instant with that at the previous instant. So, we shift one unit of our log data and test the stationary of the data as Table.4:

TABLE 4. DIFFERENCING

Test Statistic	-5.632004
p-value	0.000001
#Lags Used	17.000000
Number of Observations Used	353.000000
Critical Value (1%)	-3.449011
Critical Value (5%)	-2.869763
Critical Value (10%)	-2.571151

Also, the plot looks much better as Fig.2:

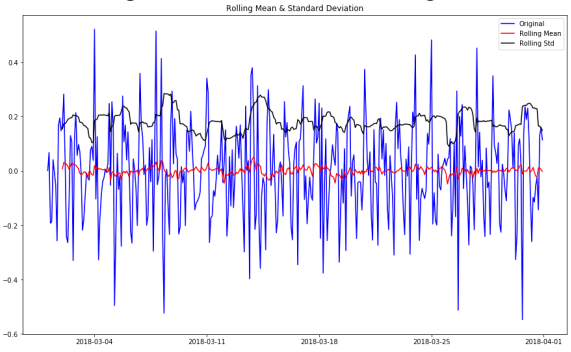


Fig.2 stationary of Uber price after differencing

P value reduces a lot and the rolling mean is closer to 0, which proves our data becomes more stable

b) Decomposition: Modeling both trend and seasonality and removing them from the model. In this approach, both trend and seasonality are modeled separately and the remaining part of the series is returned.

We use seasonal_decompose from statsmodels.tsa.seasonal to split the trend, seasonal and residual from our data and plot them together as Fig.3:

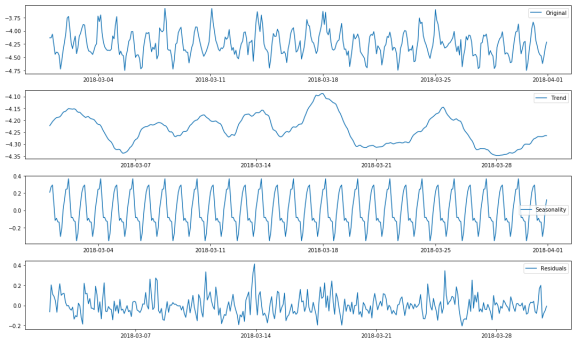


Fig.3 Decomposing

Then we test the stationary of our data after decomposing and show the result in Table.5:

TABLE 5. DECOMPOSING

Test Statistic	-8.768752e+00
p-value	2.554897e-14
#Lags Used	9.000000e+00
Number of Observations Used	3.380000e+02
Critical Value (1%)	-3.449846e+00
Critical Value (5%)	-2.870129e+00
Critical Value (10%)	-2.571346e+00

The result is not as good as Differencing, so for our project, we will take Differencing method.

E. Forecasting a Time Series

We will make model on the Time Series after differencing. Now we get a series with significant dependence among values. In this case we need to use some statistical models like ARIMA to forecast the data. ARIMA stands for Auto-Regressive Integrated Moving Averages.

The ARIMA forecasting for a stationary time series is nothing but a linear (like a linear regression) equation.

Given a time series of data X_t where t is an integer index and the X_t are real numbers, an $ARMA\ (p', q)$ model is given by [5]

$$X_t - \alpha X_{t-1} - \cdots - \alpha_{p'} X_{t-p'} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

The predictors depend on the parameters (p, d, q) of the ARIMA model:

P stands for Number of AR (Auto-Regressive) terms, AR terms are just lags of dependent variable.

Q stands for Number of MA (Moving Average) terms. MA terms are lagged forecast errors in prediction equation.

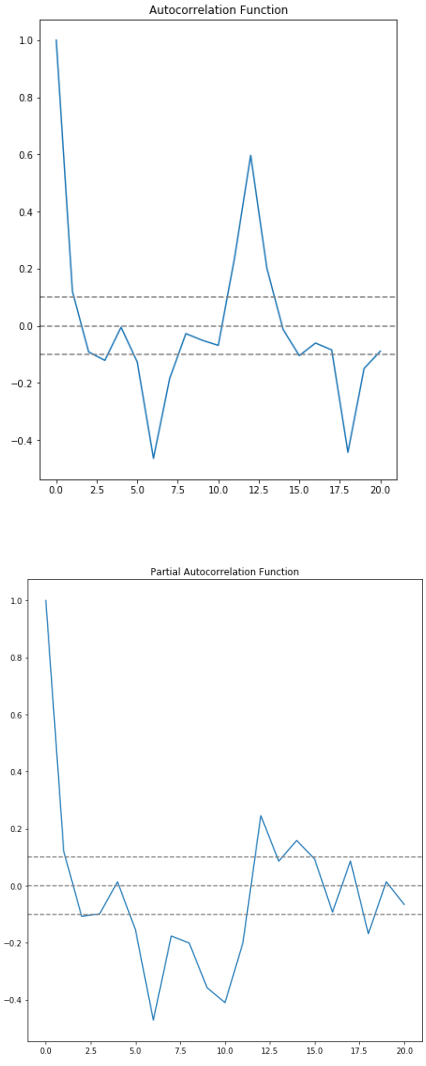
D stands for number of differences.

An importance concern here is how to determine the value of 'p' and 'q'. We use Autocorrelation Function (ACF) plot to determine q and Partial Autocorrelation Function (PACF) plot to determine q.

Autocorrelation Function (ACF) is a measure of the correlation between the Time Series with a lagged version of itself. For instance, at lag 5, ACF would compare series at time instant 't_1'…'t_2' with series at instant 't_{1-5}…t_{2-5} (t_{1-5} and t_2 being end points).

Partial Autocorrelation Function (PACF) measures the correlation between the Time Series with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons. For example, at lag 5, it will check the correlation but remove the effects already explained by lags 1 to 4.

We import acf, pacf from statsmodels.tsa.stattools and get the plots below as Fig.4:



In these two plots, the two dotted lines on either side of 0 are the confidence intervals. p can be determined as the lag value where the PACF chart crosses the upper confidence interval for the first time. If you notice closely, in this case p=2. q can be determined as the lag value where the ACF chart crosses the upper confidence interval for the first time. If you notice closely, in this case q=2.

We make 3 different ARIMA models considering individual as well as combined effects.

1) AR Model

We take (p, d, q) as (2, 1, 0) to be used in AR Model and here is the result of Fig.5

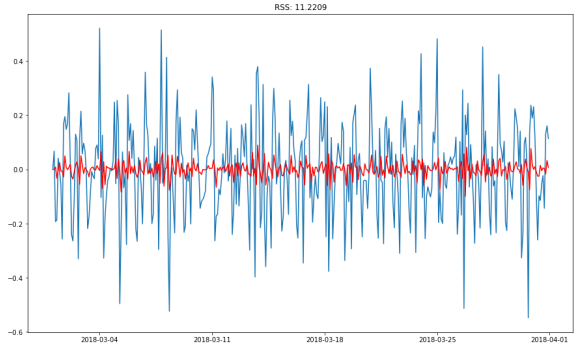


Fig.5 AR Model

2) MR Model

RSS is 11.2209 of the AR Model which is really not good, so we try MA Model using (p, d, q) as (0, 1, 2) and show the result is Fig.6:

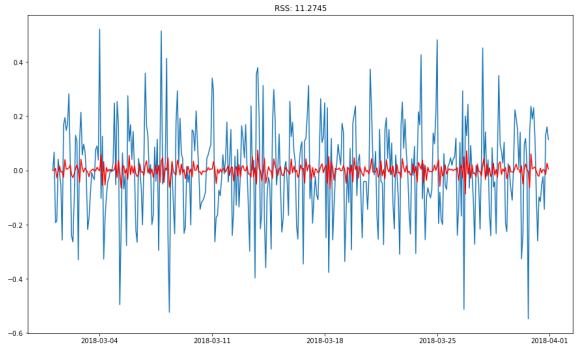


Fig.6 MR Model

3) Combined Model

RSS is 11.2745 of the MR Model which is really not good as well, so we try to combine MA Model and AR Model using (p, d, q) as (2, 1, 2) and show the result is Fig.7:

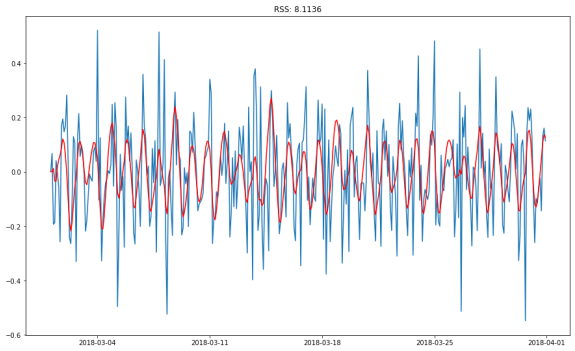


Fig.7 Combination Model

RSS is 8.1336 of the Combined Model which is much better now, so we select Combined Model to predict.

3) Predict

First, we need to take these values back to the original scale and we step store the predicted results as a separate series and observe it.

```
date_time
2018-03-01 02:00:00    0.000159
2018-03-01 04:00:00    0.000275
2018-03-01 06:00:00    0.012257
2018-03-01 08:00:00   -0.034082
2018-03-01 10:00:00   -0.034680
Freq: 2H, dtype: float64
```

Notice that these start from '2018-03-01 02:00:00' and not the first hour. This is because we took a lag by 1 and first element doesn't have anything before it to subtract from.

The way to convert the differencing to log scale is to add these differences consecutively to the base number. An easy way to do it is to first determine the cumulative sum at index and then add it to the base number. The cumulative sum can be found as:

```
date_time
2018-03-01 02:00:00    0.000159
2018-03-01 04:00:00    0.000435
2018-03-01 06:00:00    0.012691
2018-03-01 08:00:00   -0.021390
2018-03-01 10:00:00   -0.056071
Freq: 2H, dtype: float64
```

Next, we need to add them to base number. For this we create a series with all values as base number and add the differences to it.

```
date_time
2018-03-01 00:00:00   -4.131946
2018-03-01 02:00:00   -4.131787
2018-03-01 04:00:00   -4.131512
2018-03-01 06:00:00   -4.119255
2018-03-01 08:00:00   -4.153337
Freq: 2H, dtype: float64
```

Last step is to take the exponent and compare with the original series. We get the prediction below as Fig.8:

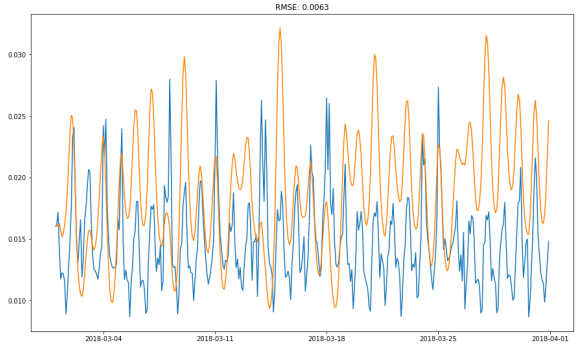


Fig.8 Prediction Result

Finally, we have a forecast at the original scale. The RSS is 0.0063 which is not a very good forecast but we can predict some basic trend of Uber price. Next, we will try to find more accurate prediction method to apply.

Source:
https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/

[1]: https://en.wikipedia.org/wiki/Moving_average
Moving average. Wikipedia

[2]:https://www.investopedia.com/university/movingaverage/movingaverages1.asp#ixzz5D2bBkCmk
Moving Averages: What Are They? . Investopedia

[3]: https://www.thebalance.com/simple-exponential-and-weighted-moving-averages-1031196
Simple, Exponential and Weighted Moving Averages. The Balance

[4]:https://en.wikipedia.org/wiki/Dickey%E2%80%93Fuller_test
Dickey–Fuller test. Wikipedia

[5]:https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
Autoregressive integrated moving average. Wikipedia