IoT analytics: Forecasting

Wen-Han Hu(whu24)

Data preparation:

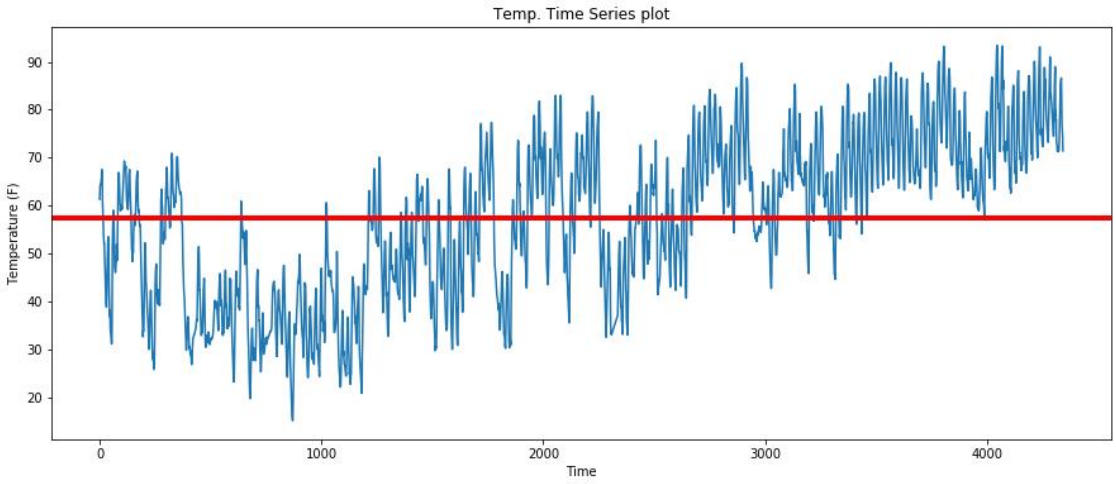
We only need to use the 'Temperature' attribute in the dataset in this project. As the result, we only extract the 'Temperature' from the dataset and overview the 'Temperature' attribute to have a brief understanding of our data.

	Temperature
count	4345.000000
mean	57.343014
std	16.497043
min	15.100000
25%	43.700000
50%	59.500000
75%	69.600000
max	93.400000

As the figure above, we have total number of 4345 'Temperature' data, we will separate the first 70% of our training data the remaining 30% as our test data.

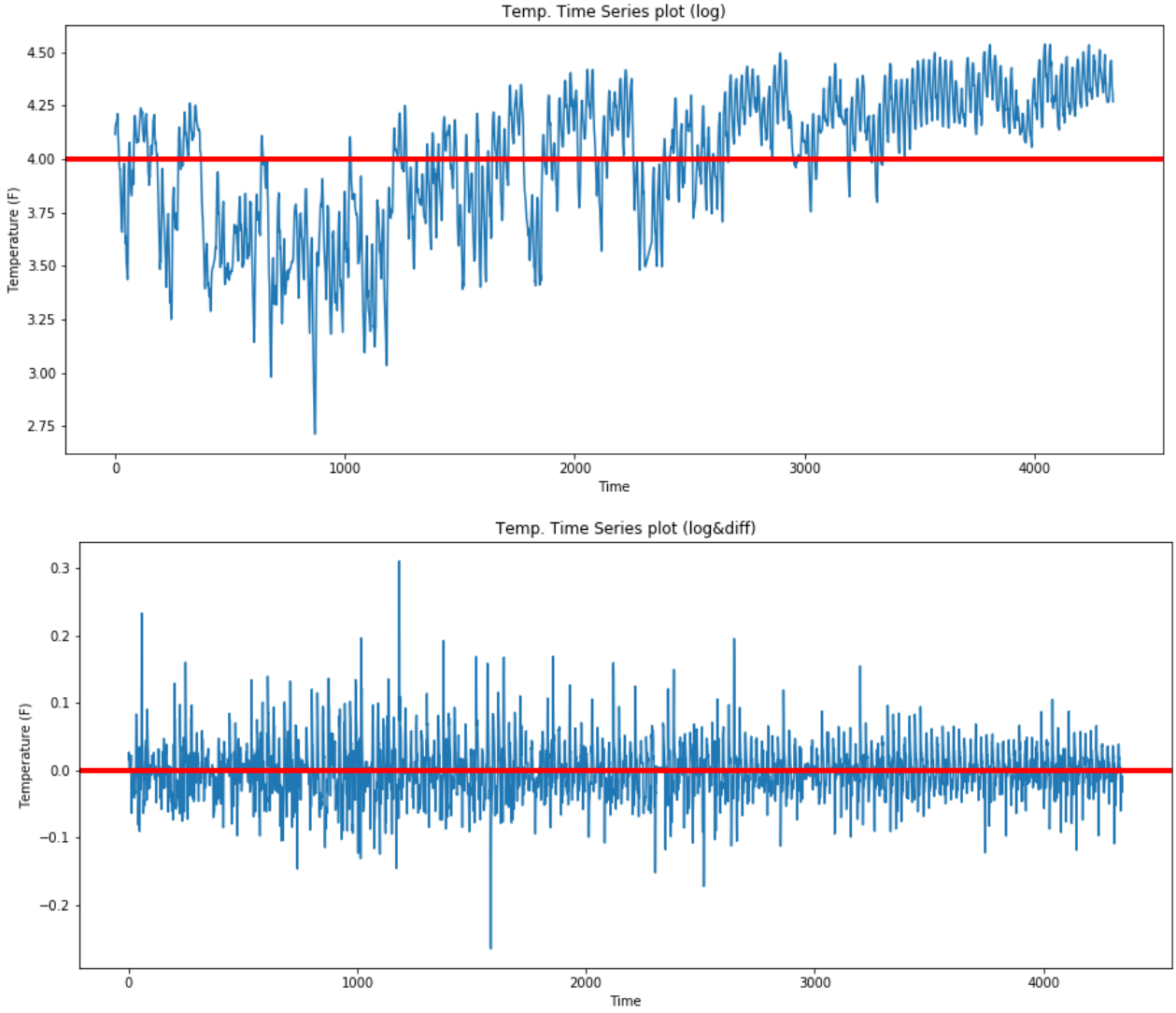
Task 1. Check for stationarity

In order to check the time series data if stationary, we need to evaluate its mean, variance and seasonality.



The red line indicates the overall mean of the 'Temperature' data, we can observe that the mean slightly changes overall as well as the variance. Hence, we can use log transformation and first order differencing to remove the trend and variable variance to

make the data stationary.



After transformation, we can observe that the time series data has constant mean and variance without trend. To be more objective to evaluate the time series if stationary, we can apply some statistical tests, such as augmented Dickey-Fuller test, to validate the time series data.

```
ADF Statistic: -4.194491
p-value: 0.000673
Critical Values:
        1%: -3.432
        5%: -2.862
        10%: -2.567
```

Running the temperature data prints the test statistic value of -4. The more negative this statistic, the more likely we are to reject the null hypothesis (we have a stationary dataset). As part of the output, we get a look-up table to help determine the

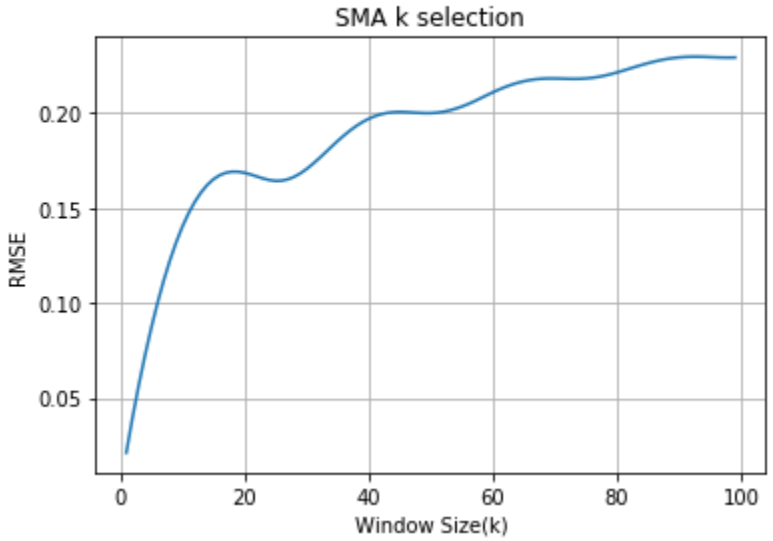
ADF statistic. We can see that our statistic value of -4 is less than the value of -3.432 at 1%. This suggests that we can reject the null hypothesis with a significance level of less than 1% (i.e. a low probability that the result is a statistical fluke).

Rejecting the null hypothesis means that the process has no unit root, and in turn that the time series is stationary or does not have time-dependent structure

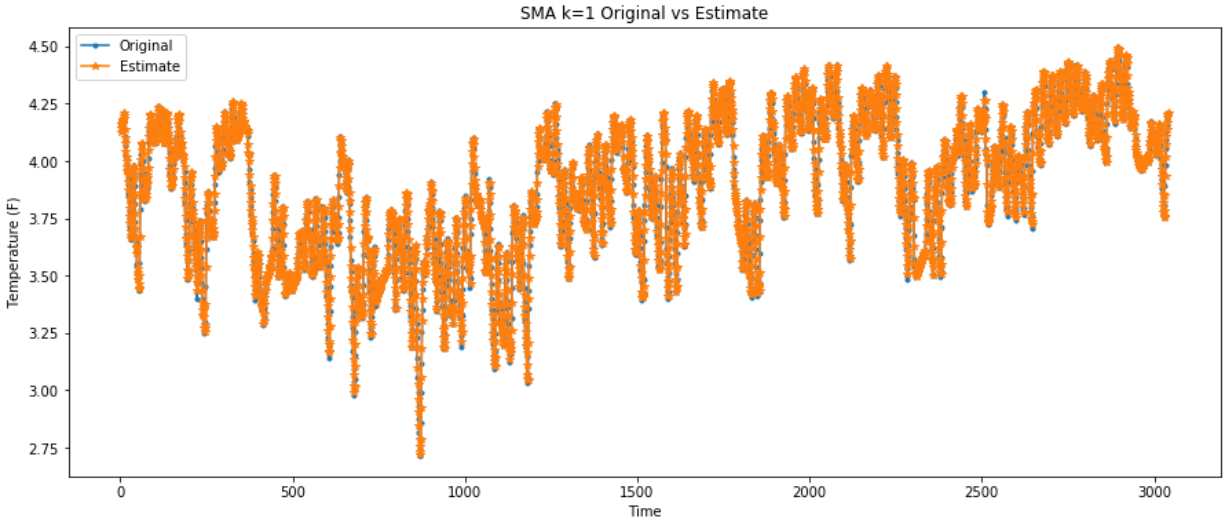
Task 2. Fit a simple moving average model (using the training set)

In order to apply simple moving average model, we need to find the k with minimum RMSE.

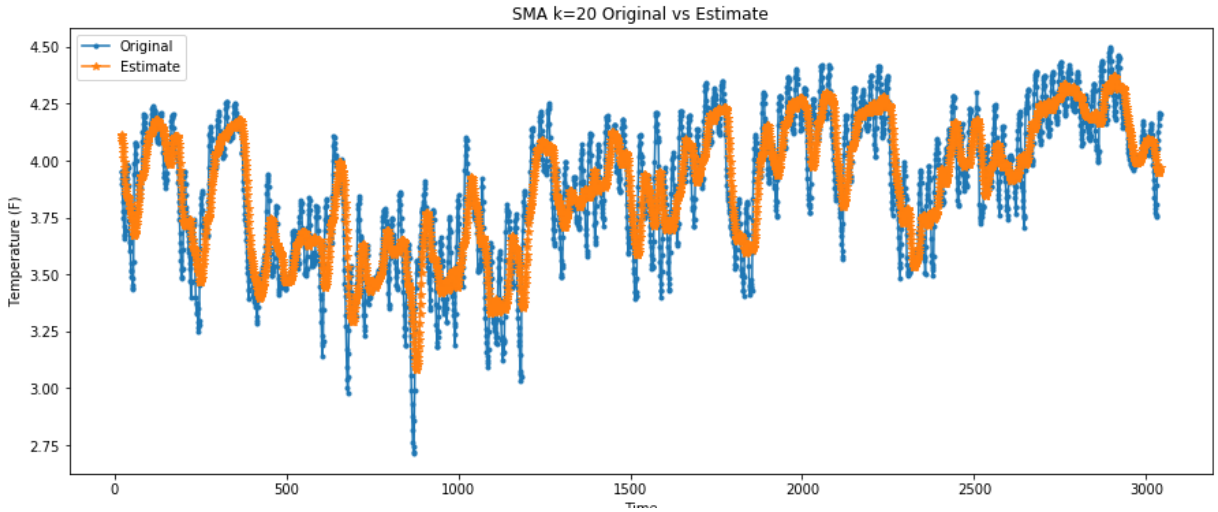
```
Minumum RMSE value is 0.021129338585416366, when k = 1
```



With the results above, the lowest RMSE is 0.021129 when k = 1. Hence, we can decide our k is 1 and plot the original data versus predicted data.



The blue line presents the original data and the orange line indicate the predicted data. We can see the most of them are overlapped. In order to have a different example, the following plot with k = 20 which is not the best k selection with minimum RMSE.



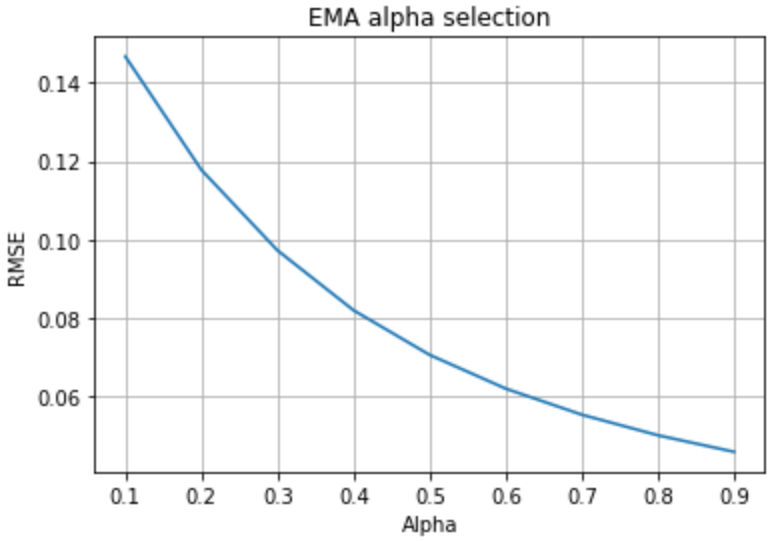
These two graphs indicate that the least RMSE has the best model fitting.

Due to the SMA k = 1, we can know that the best prediction of our time series data is to use the previous data point to predict the future temperature.

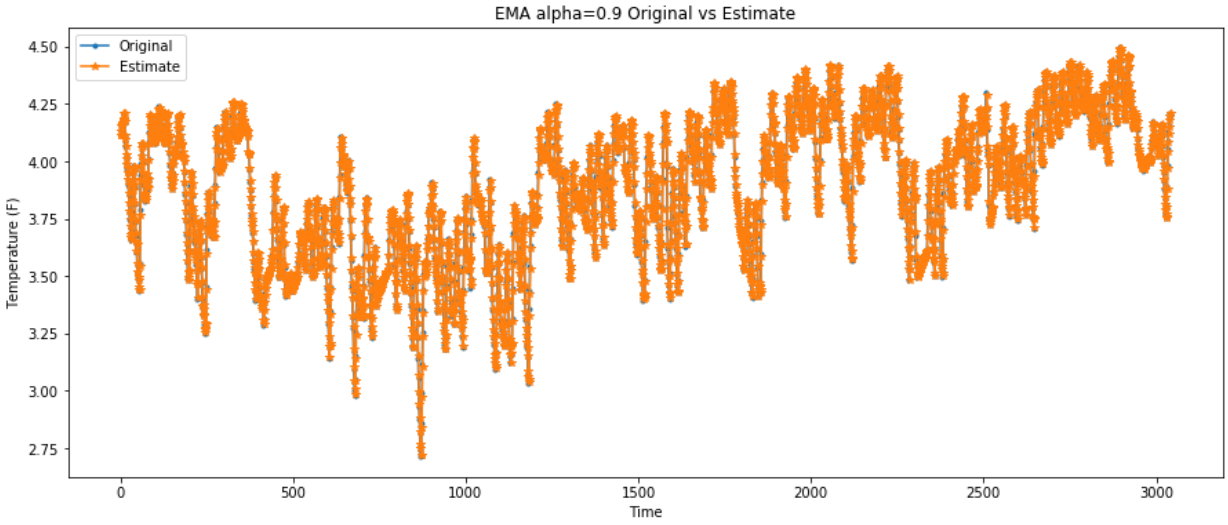
Task 3. Fit an exponential smoothing model (use the training set)

In order to apply exponential smoothing model, we need to find the alpha with minimum RMSE.

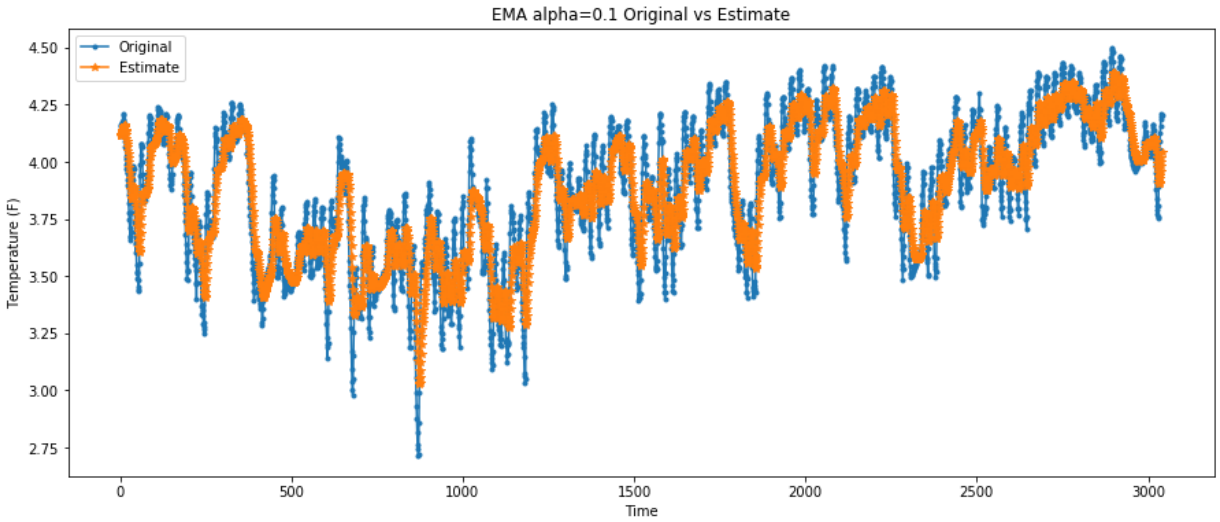
```
Minumum RMSE value is 0.045682336467215 14, when alpha = 0.9
```



With the results above, the lowest RMSE is 0.04568 when alpha = 0.9. Hence, we can decide our alpha is 0.9 and plot the original data versus predicted data.



The blue line presents the original data and the orange line indicate the predicted data. We can see the most of them are overlapped. In order to have a different example, the following plot with alpha = 0.1 which is not the best alpha selection with minimum RMSE.



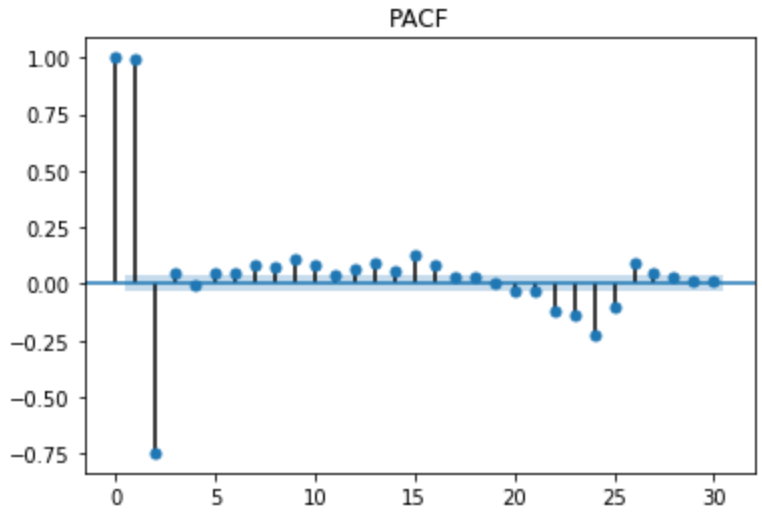
These two graphs indicate that the least RMSE has the best model fitting as the results we had from the SMA model. Because we have too intensive data points so that it is difficult to observe that the SMA model, which has smaller RMSE then EMA model, should be slightly better than EM model.

We can further confirm that the prediction closely depends on the prior data by the SMA k =1 and EMA alpha = 0.9 which is given more weight on the previous actual data point.

Task 4. Fit an AR(p) model (use the training set)

In order to apply autoregressive model, we need to find the lag k at which PACF cuts off. The cut-off point is 0.15 which is the same value as describe in the textbook.

```
p value using PACF is 3
```



We calculate the PACF coefficients until we get the first zero coefficient and fix the order *p* of the AR(*p*) model to the lag prior to getting the zero coefficient. As in the case of the autocorrelation, a partial autocorrelation value is zero if it falls within the confidence interval bounds, or in the absence of these bounds, its absolute value is less than 0.15. We see that at $p = 3$, $p < 0.15$, and therefore we select an AR(3).

The following show the parameters of AR(3) model:

ARMA Model Results

Dep. Variable:	Temperature	No. Observations:	3041
Model:	ARMA(3, 0)	Log Likelihood	6523.739
Method:	css-mle	S.D. of innovations	0.028
Date:	Sat, 02 Nov 2019	AIC	-13037.478
Time:	16:41:35	BIC	-13007.378
Sample:	0	HQIC	-13026.658

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.8883	0.032	122.306	0.000	3.826	3.951
ar.L1.Temperature	1.7526	0.018	96.759	0.000	1.717	1.788
ar.L2.Temperature	-0.8075	0.034	-24.086	0.000	-0.873	-0.742
ar.L3.Temperature	0.0388	0.018	2.142	0.032	0.003	0.074

Roots

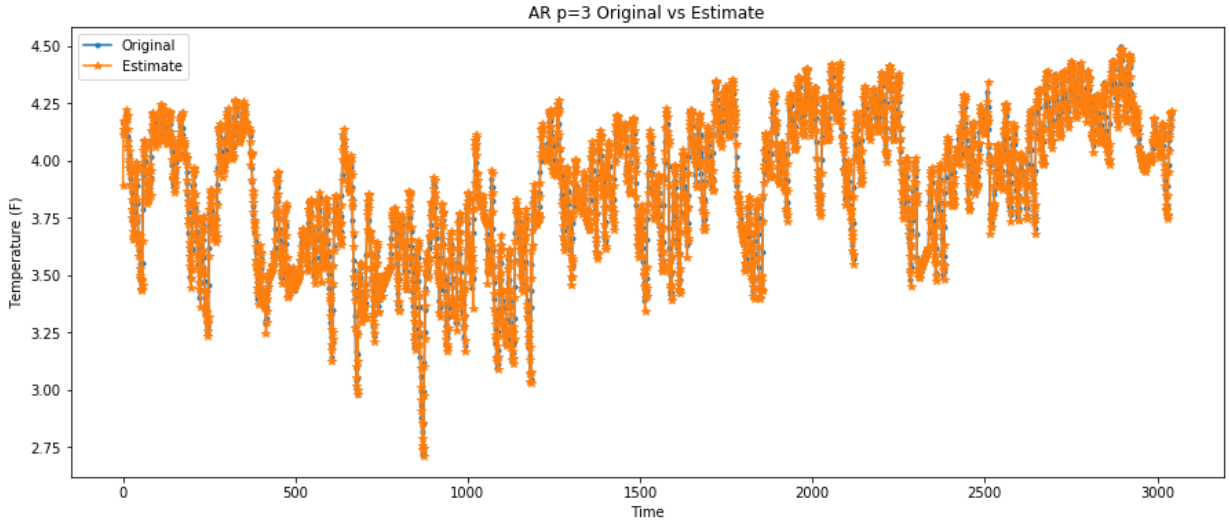
	Real	Imaginary	Modulus	Frequency
AR.1	1.0809	+0.0000j	1.0809	0.0000
AR.2	1.2932	+0.0000j	1.2932	0.0000
AR.3	18.4331	+0.0000j	18.4331	0.0000

With the coefficients above, then we can calculate the RMSE of original data and predicted data.

Thus, we have the following expression.

$$X_t = 3.8883 + 1.7526X_{t-1} - 0.8075X_{t-2} + 0.0388X_{t-3}$$

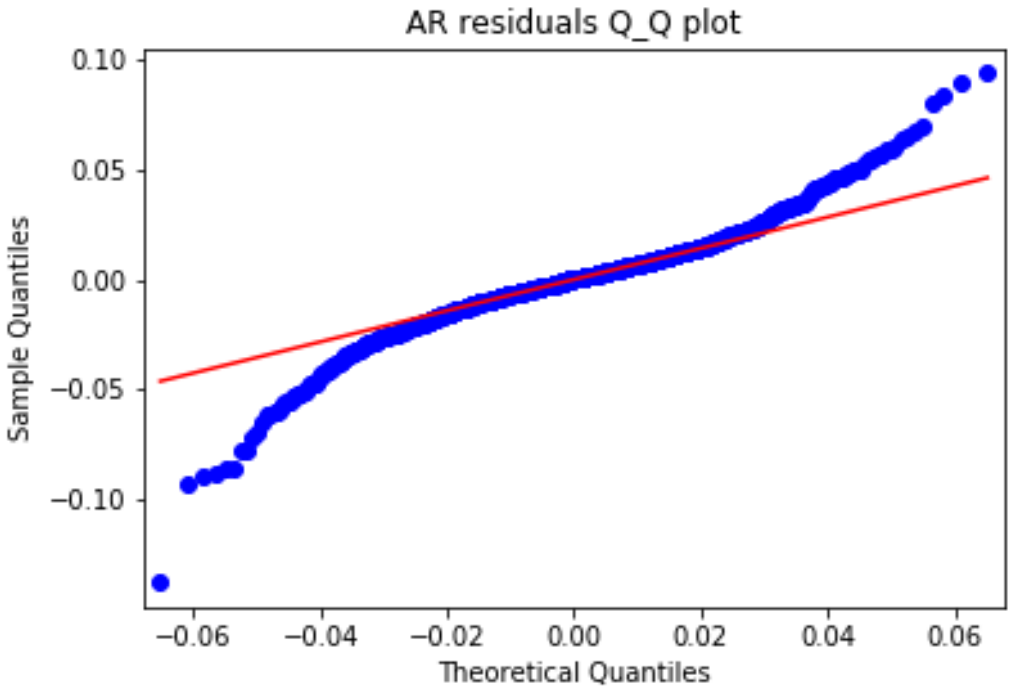
```
RSME of AR model with train data: 0.028591347296862888
```



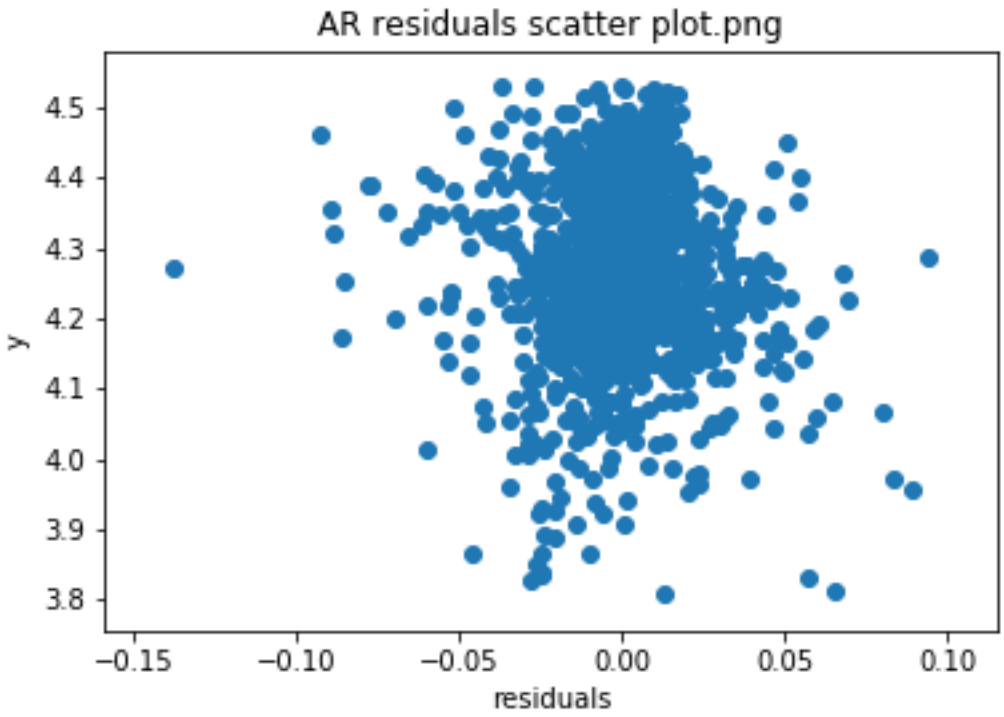
The RMSE is 0.0285913 of our AR(3) model which has better performance than EMA but less than SMA model.

AR(3) residual analysis:

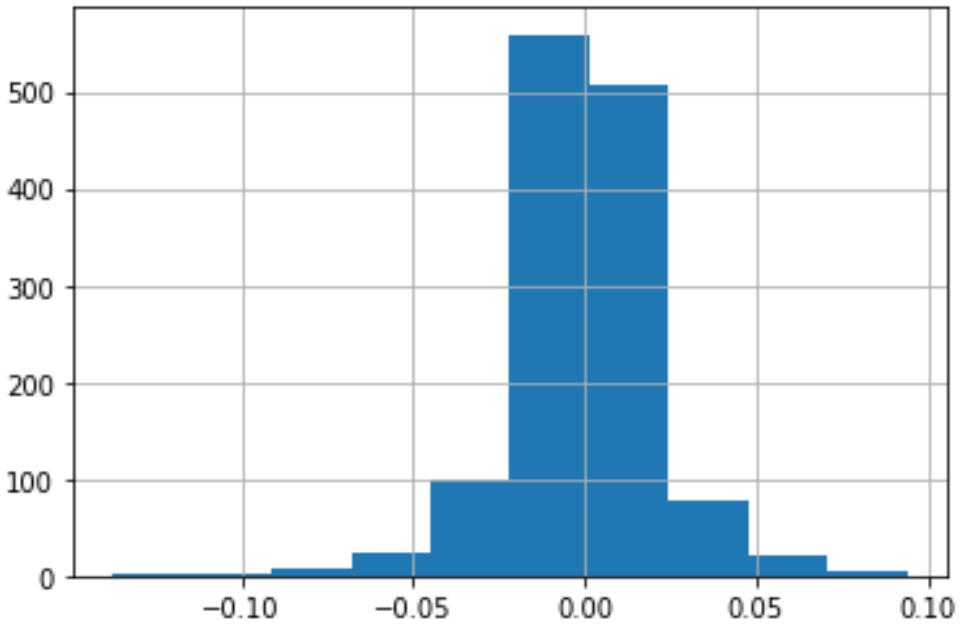
Q_Q plot:



Residual scatter plot:



Residual histogram:



We can easily observe that the residuals are not normal distributed by the Q-Q plot, even though there is no trend in the scatter plot. But, let's still run the chi-squared test to determine the p-value.

```
Running chisquare test.........
We have p-value is 9.80328652614438e-35
Significant Result, the null hypothesis rejected
```

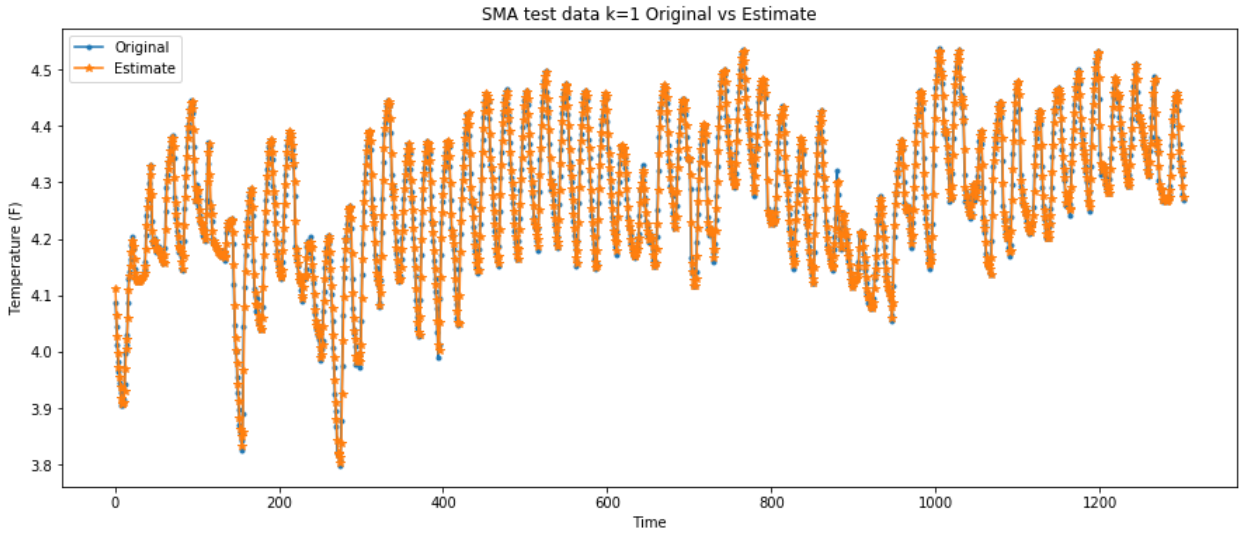
The reason why the residuals are not normal distributed is the model not well explain the data. That is because the AR model may be not sufficient to fit our data, we can try ARIMA to see if there is improved. (However, this is beyond the scope of this project so that we stop here for this analysis.)

p.s. This part of comment has been confirmed with professor.

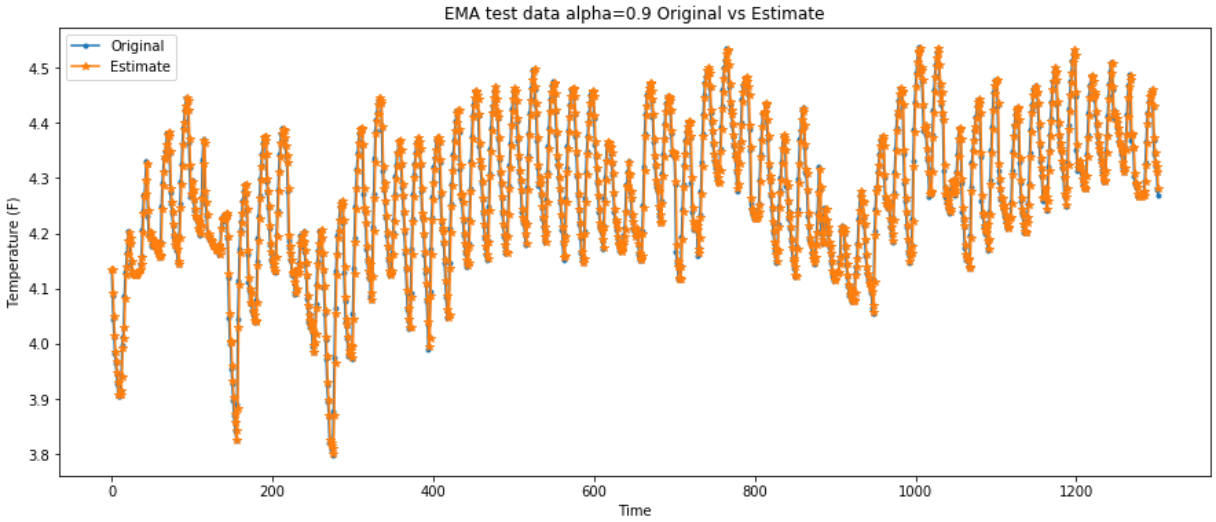
Task 5. Comparison of all the models (use the testing set)

After we decide all the parameters of SMA, EMA and AR model, now we can apply these parameters, k =1, alpha = 0.9, p = 3 separately for each model on our test data.

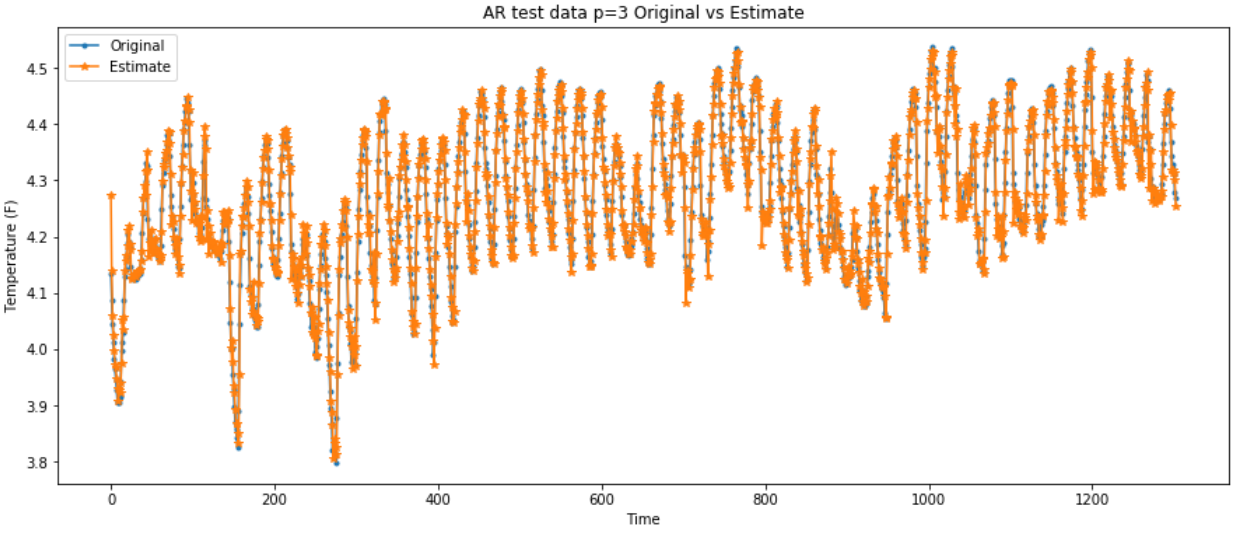
```
RSME of SMA model with test data: 0.015068503431487195
```



RSME of EM model with test data: 0.032512157130479516



RSME of AR model with test data: 0.020567843898572604



	SMA, k=1	EMA, alpha=0.9	AR(3)
Train	0.021129338585416366	0.04568233646721514	0.028591347296862888
Test	0.015068503431487195	0.032512157130479516	0.020567843898572604

The SMA model in our case has the best performance with the minimum RMSE which mean our prediction is closer comparing to EMA and AR(3) model. The result can be contributed to this data is very intensive with the temperature so that there is not much difference between the current point with previous point and the difference with next point. As the result, the data is closely relied on the prior value of data and SMA is the model which is best for this kind of data.