HW2a

YL5090

Yuqiong Li

SEG0- Write up

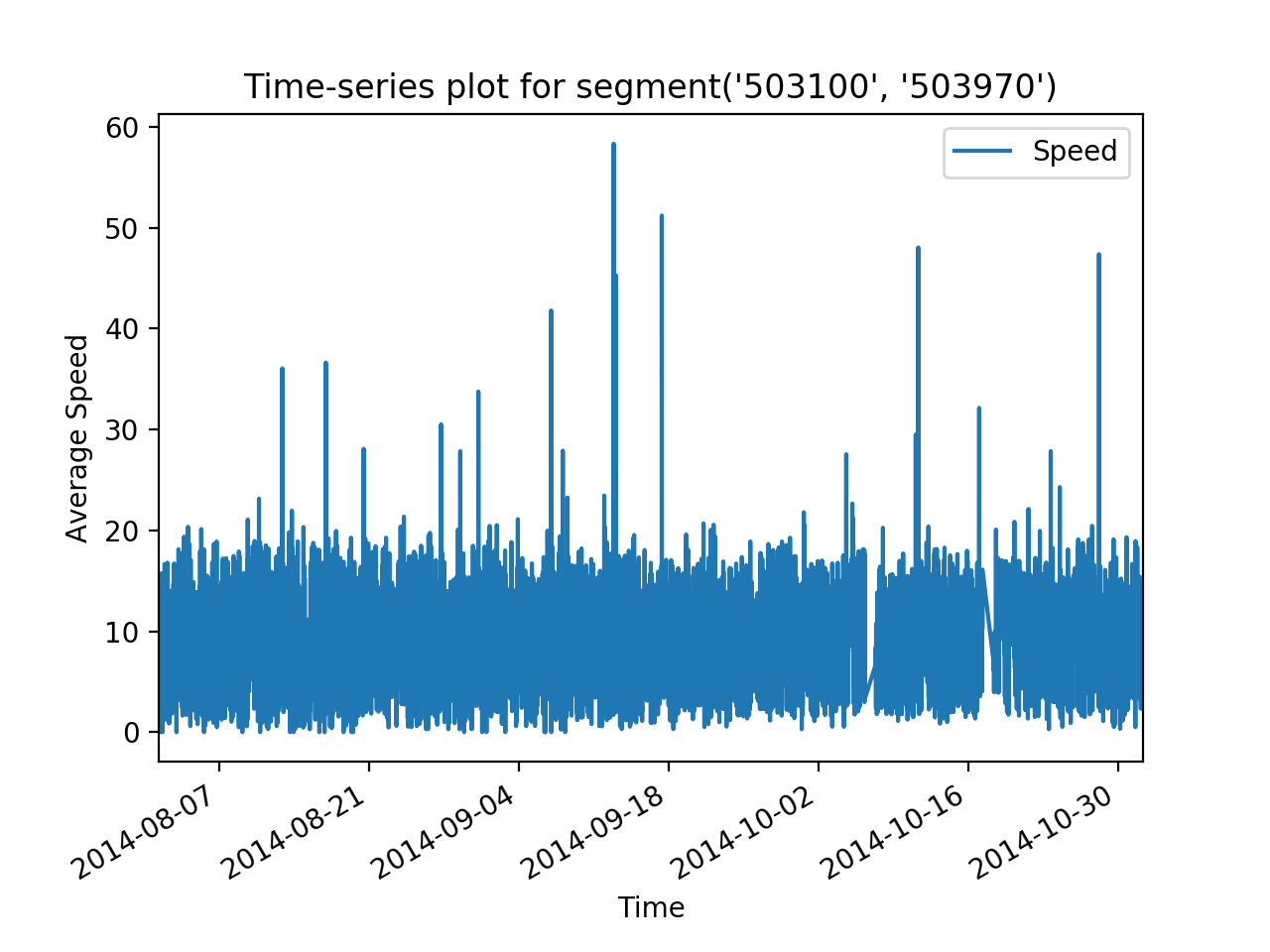
# **Preprocessing:**

1. Order the speed data as time
2. Remove outliers
   1. This step is done by removing all points that are 3 standard deviation away from sample mean.
3. For NaN values, fill it with the last valid value before it
4. The circle of seasonality is 144. Remove seasonality by subtracting the value 144 time stamp away. Then remove the first 144 values as they are nan now.

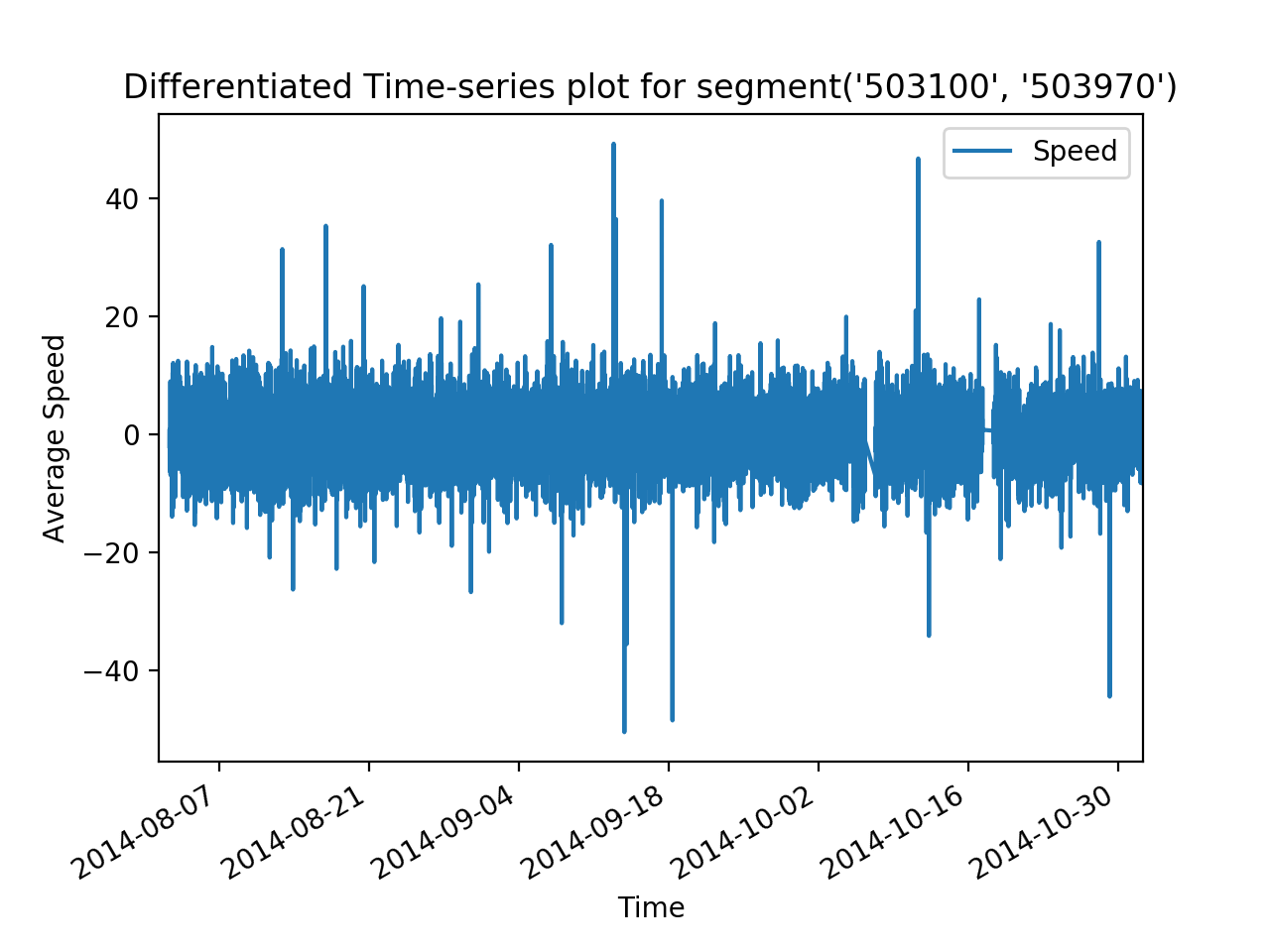
Identification:

After cleaning data, now identify an ARIMA(p, d, q) model.

Step 1: First plot the time-series



After differentiating by 144 lag, now plot again



Use the Augmented Dickey-Fuller test to check existence of unit root:

ADF statistic: -13.944134

p-value: 0.000000

Critical Values:

1%: -3.431

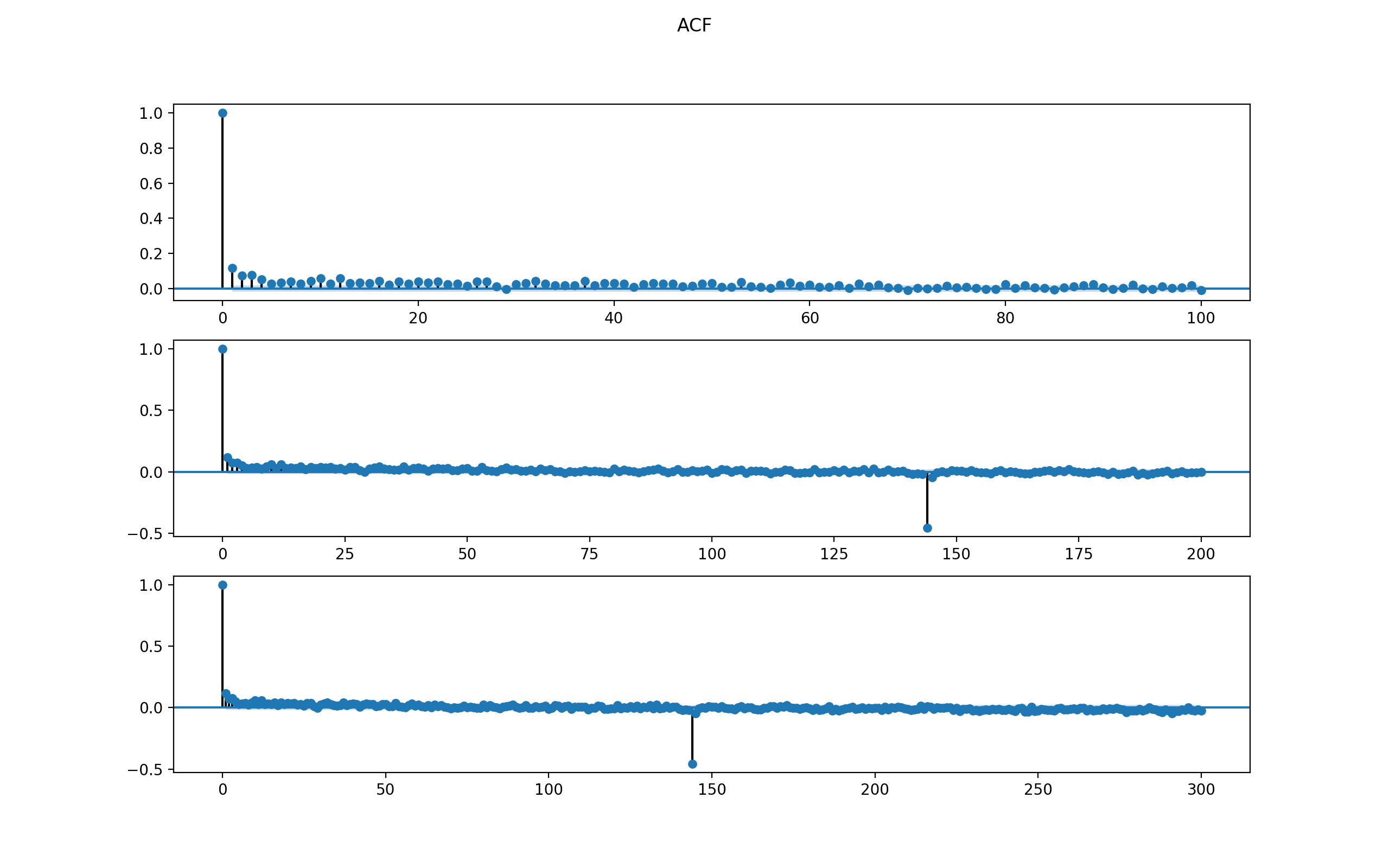
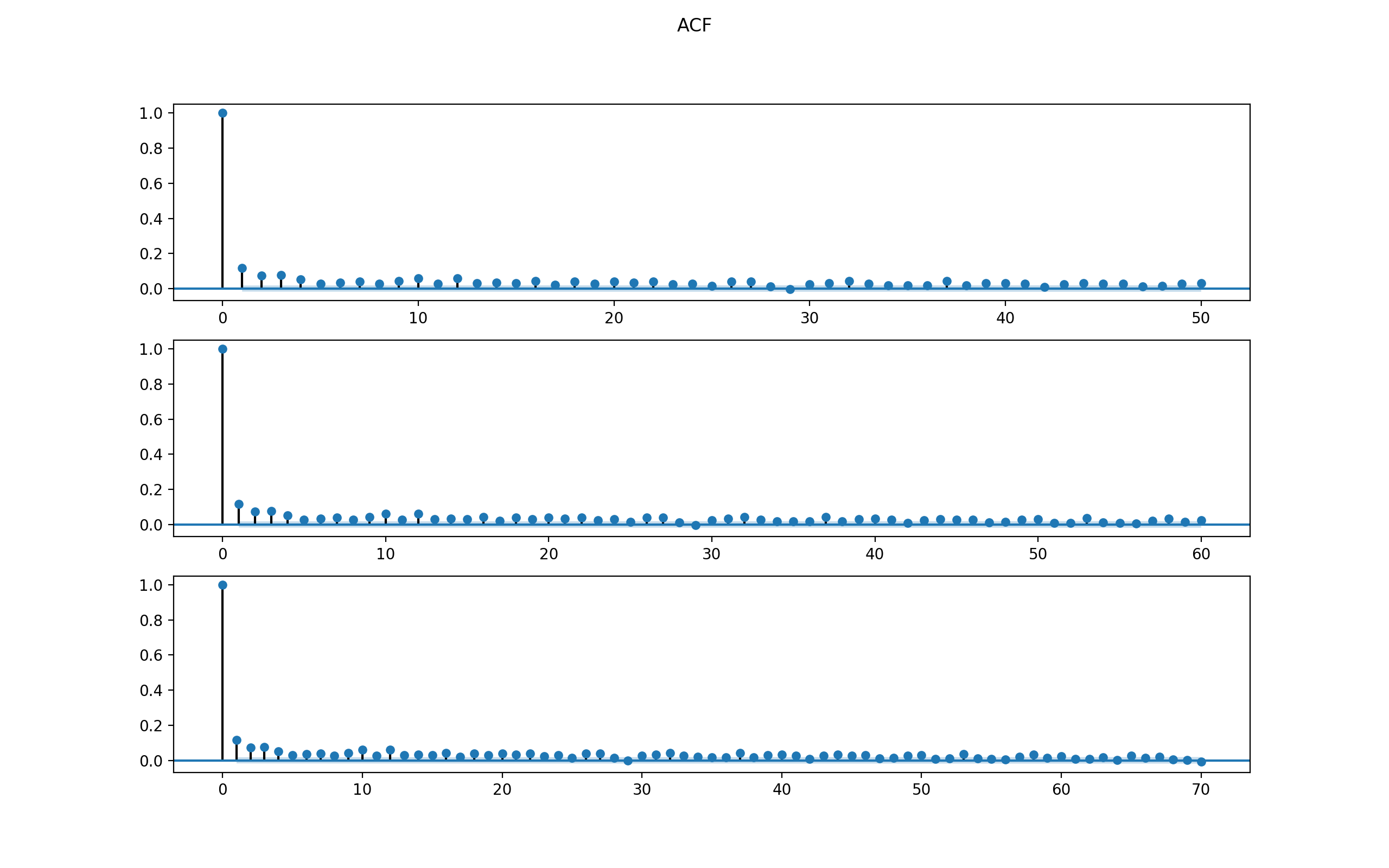
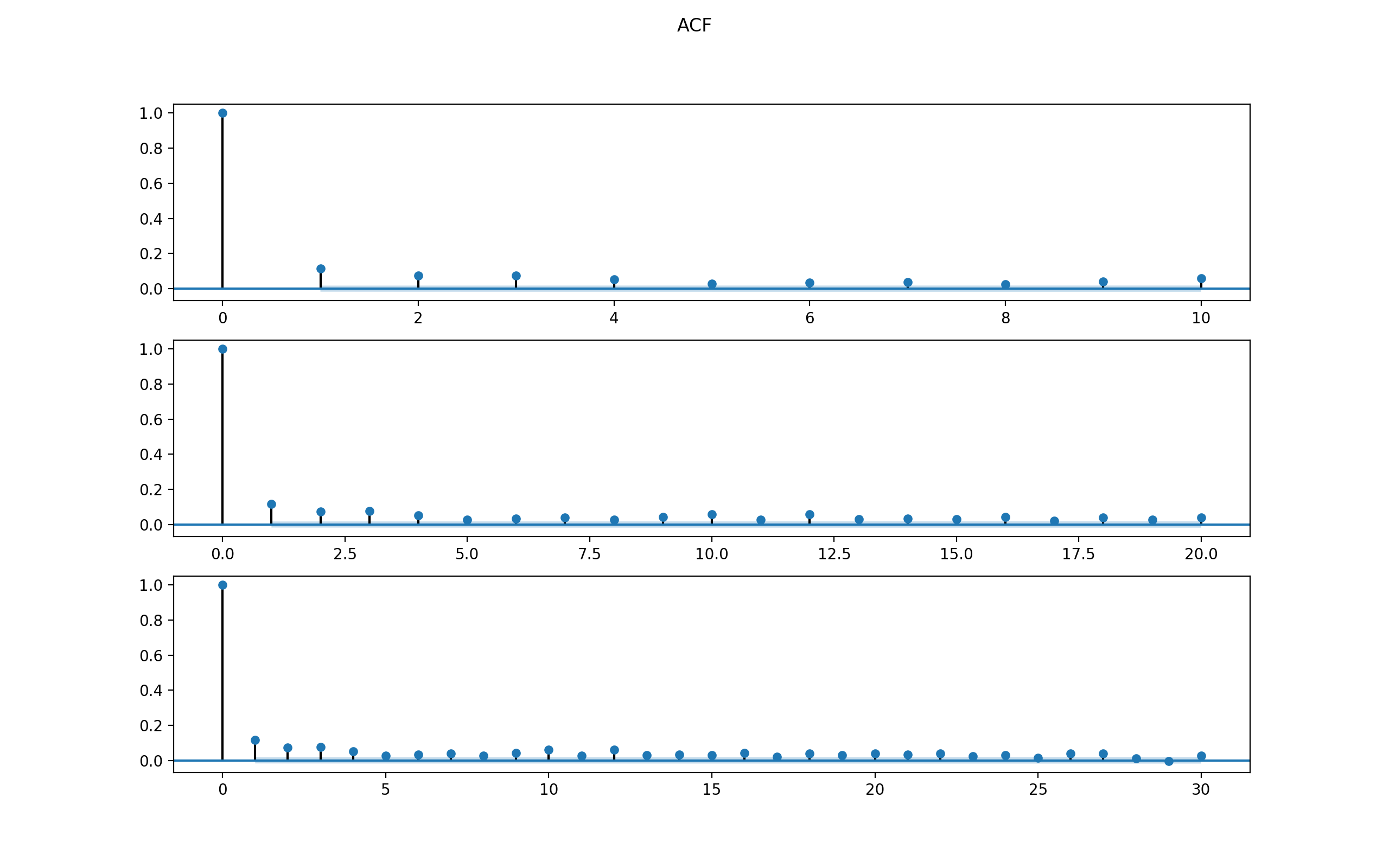
5%: -2.862

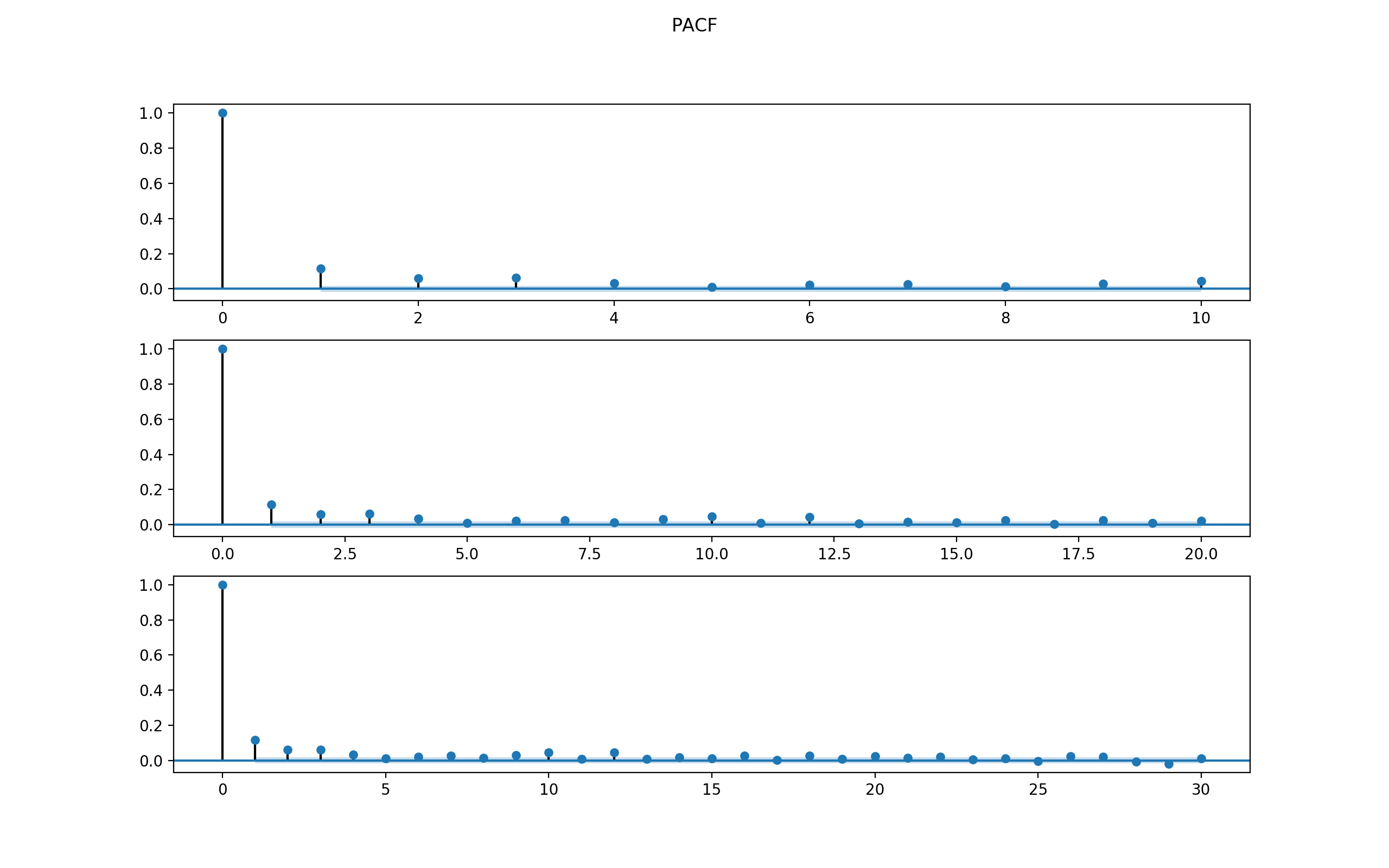
10%: -2.567

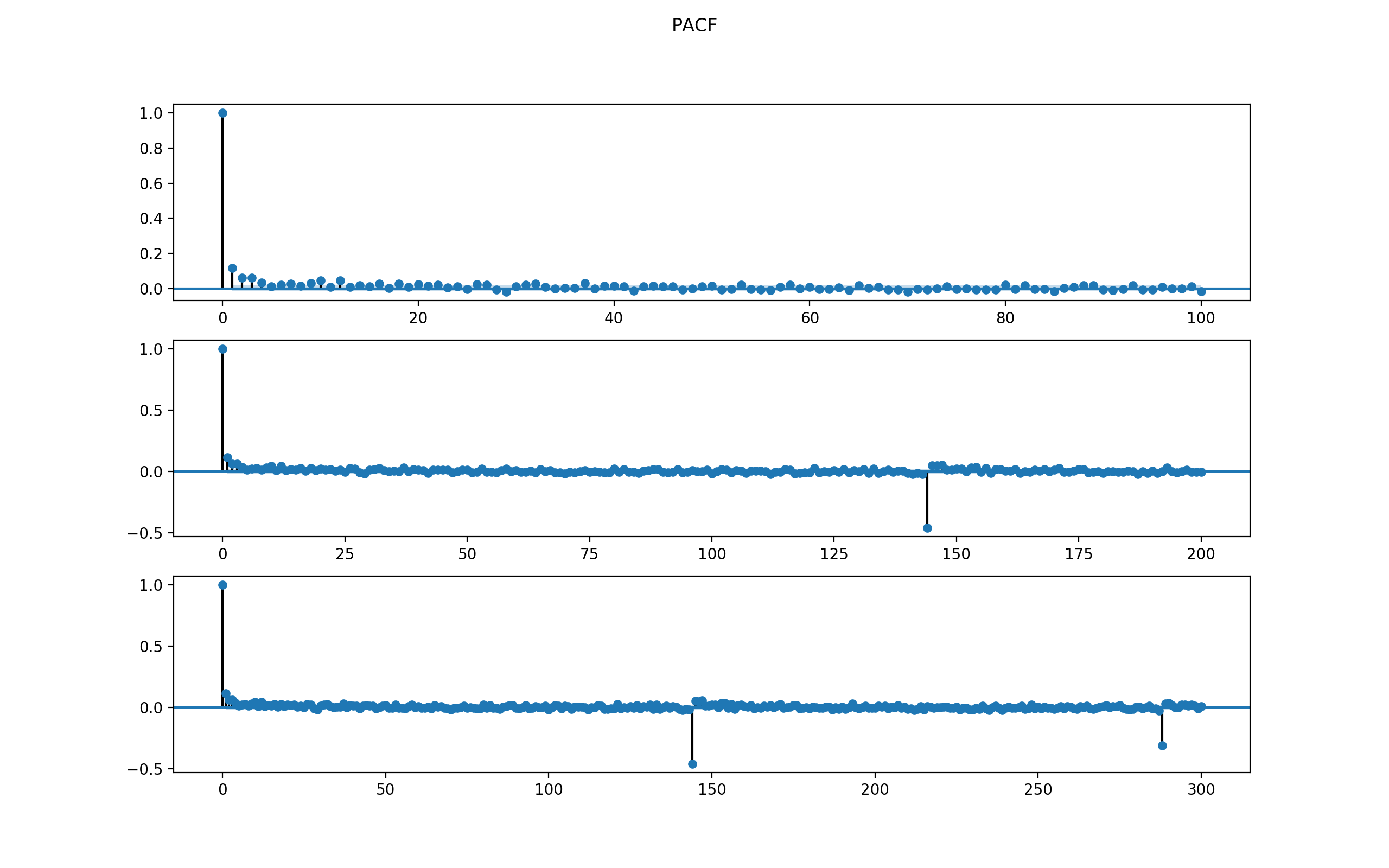
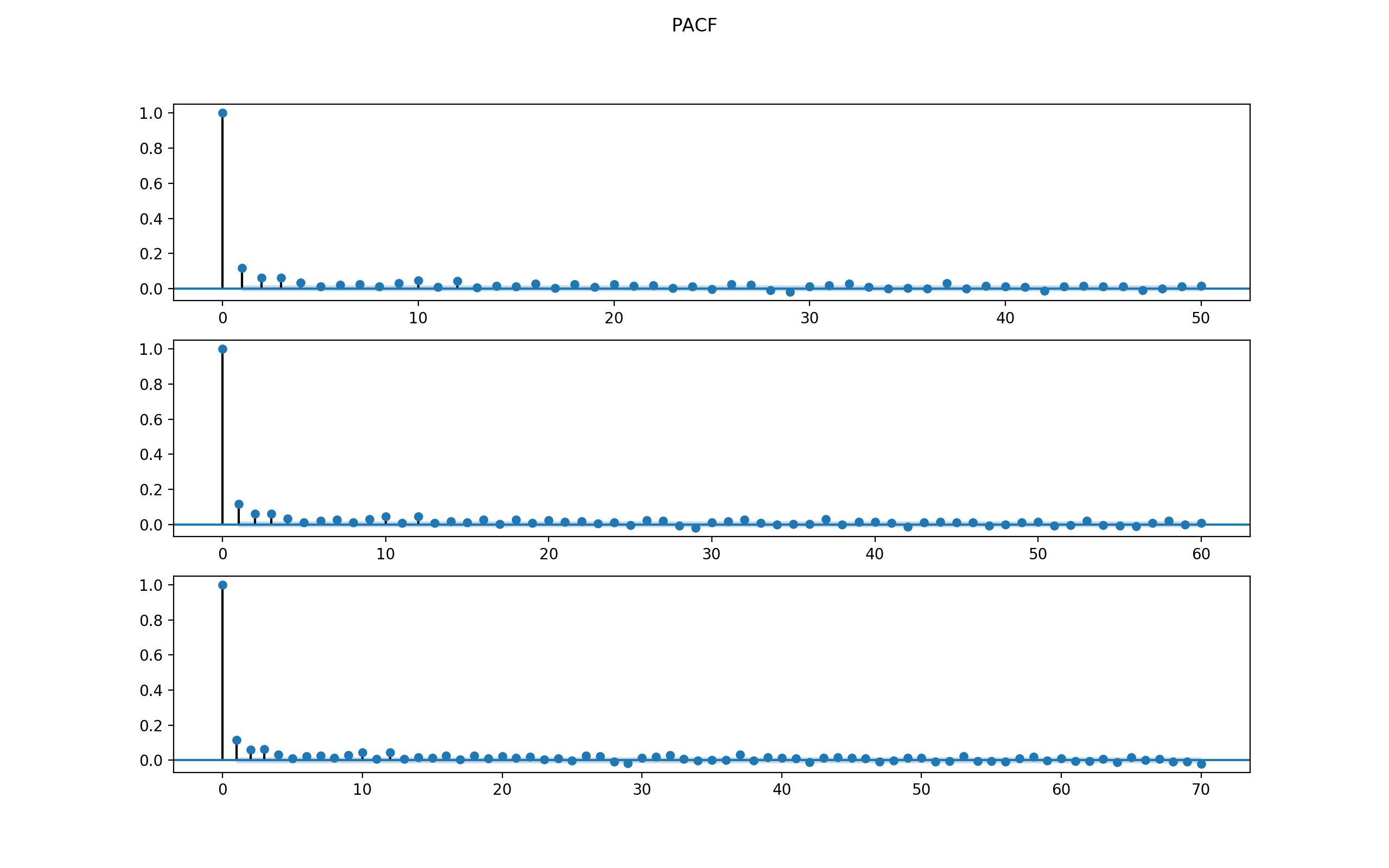
Conclusion:

1. After differentiating the data doesn’t seem to fit any upward or downward trend.
2. Augmented Dickey-Fuller test rejects null hypothesis. Thus no unit root. No time dependent structure.
3. Thus, now conclude stationary. Set d = 0.
4. This makes sense, as there is no need to believe average bus speed follows a growing trend in three months, unless there is a change in technology.

# Step 2: Identify AR or MA process by plotting ACF and PACF for different lag values.







Observations:

1. ACF dies off, no cut off. Becomes zero after lag = 4 or so.
2. PACF dies off, no cut off. Becomes zero after lag = 3 or so.

Thus, the model might be a mixture of AR and MA, with p being at most 4 and q being at most 3.

Start from simple model for first stage filtering

* AR() of p being 1 to 4,
* MA() of q being 1 to 3,
* and also ARMA() combination of (p, q) from the above range.

After filtering and selecting the lowest AIC and BIC for each family as well as the simplest model possible, decide to further compare three models below:

ARMA Model Results

==============================================================================

Dep. Variable: Speed No. Observations: 8544

Model: ARMA(3, 0) Log Likelihood -25961.539

Method: css-mle S.D. of innovations 5.051

Date: Mon, 11 Dec 2017 AIC 51933.077

Time: 17:24:29 BIC 51968.342

Sample: 08-02-2014 HQIC 51945.108

- 09-30-2014

===============================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------

const 0.0037 0.070 0.053 0.958 -0.133 0.140

ar.L1.Speed 0.1061 0.011 9.815 0.000 0.085 0.127

ar.L2.Speed 0.0544 0.011 5.011 0.000 0.033 0.076

ar.L3.Speed 0.0533 0.011 4.937 0.000 0.032 0.075

Roots

=============================================================================

Real Imaginary Modulus Frequency

-----------------------------------------------------------------------------

AR.1 2.1412 -0.0000j 2.1412 -0.0000

AR.2 -1.5802 -2.5016j 2.9589 -0.3397

AR.3 -1.5802 +2.5016j 2.9589 0.3397

ARMA Model Results

==============================================================================

Dep. Variable: Speed No. Observations: 8544

Model: ARMA(0, 3) Log Likelihood -25967.173

Method: css-mle S.D. of innovations 5.055

Date: Mon, 11 Dec 2017 AIC 51944.346

Time: 17:24:30 BIC 51979.611

Sample: 08-02-2014 HQIC 51956.377

- 09-30-2014

===============================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------

const 0.0036 0.067 0.054 0.957 -0.127 0.134

ma.L1.Speed 0.1037 0.011 9.589 0.000 0.083 0.125

ma.L2.Speed 0.0608 0.011 5.788 0.000 0.040 0.081

ma.L3.Speed 0.0548 0.011 5.206 0.000 0.034 0.075

Roots

=============================================================================

Real Imaginary Modulus Frequency

-----------------------------------------------------------------------------

MA.1 0.8372 -2.4189j 2.5597 -0.1970

MA.2 0.8372 +2.4189j 2.5597 0.1970

MA.3 -2.7833 -0.0000j 2.7833 -0.5000

ARMA Model Results

==============================================================================

Dep. Variable: Speed No. Observations: 8544

Model: ARMA(4, 3) Log Likelihood -25909.220

Method: css-mle S.D. of innovations 5.019

Date: Mon, 11 Dec 2017 AIC 51836.440

Time: 17:25:56 BIC 51899.917

Sample: 08-02-2014 HQIC 51858.096

- 09-30-2014

===============================================================================

coef std err z P>|z| [0.025 0.975]

-------------------------------------------------------------------------------

const 0.0046 0.123 0.038 0.970 -0.237 0.246

ar.L1.Speed -0.1590 0.015 -10.655 0.000 -0.188 -0.130

ar.L2.Speed 0.2101 0.009 23.637 0.000 0.193 0.227

ar.L3.Speed 0.9489 0.006 157.278 0.000 0.937 0.961

ar.L4.Speed -0.0673 0.012 -5.748 0.000 -0.090 -0.044

ma.L1.Speed 0.2646 0.010 26.628 0.000 0.245 0.284

ma.L2.Speed -0.1594 0.012 -13.181 0.000 -0.183 -0.136

ma.L3.Speed -0.9518 0.010 -95.816 0.000 -0.971 -0.932

Roots

=============================================================================

Real Imaginary Modulus Frequency

-----------------------------------------------------------------------------

AR.1 -0.6124 -0.8002j 1.0077 -0.3540

AR.2 -0.6124 +0.8002j 1.0077 0.3540

AR.3 1.0232 -0.0000j 1.0232 -0.0000

AR.4 14.3045 -0.0000j 14.3045 -0.0000

MA.1 -0.6088 -0.7937j 1.0003 -0.3541

MA.2 -0.6088 +0.7937j 1.0003 0.3541

MA.3 1.0501 -0.0000j 1.0501 -0.0000

-----------------------------------------------------------------------------

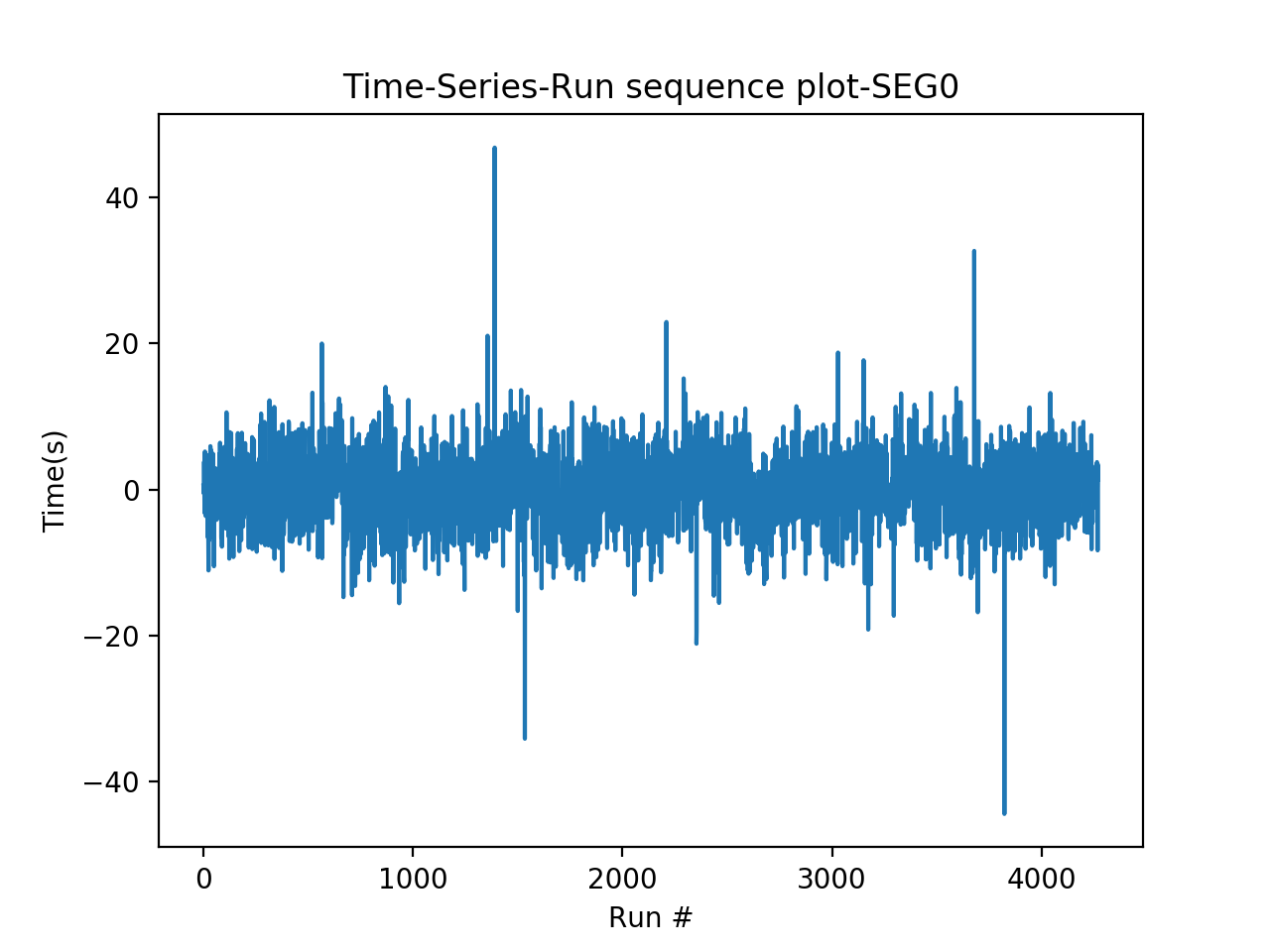
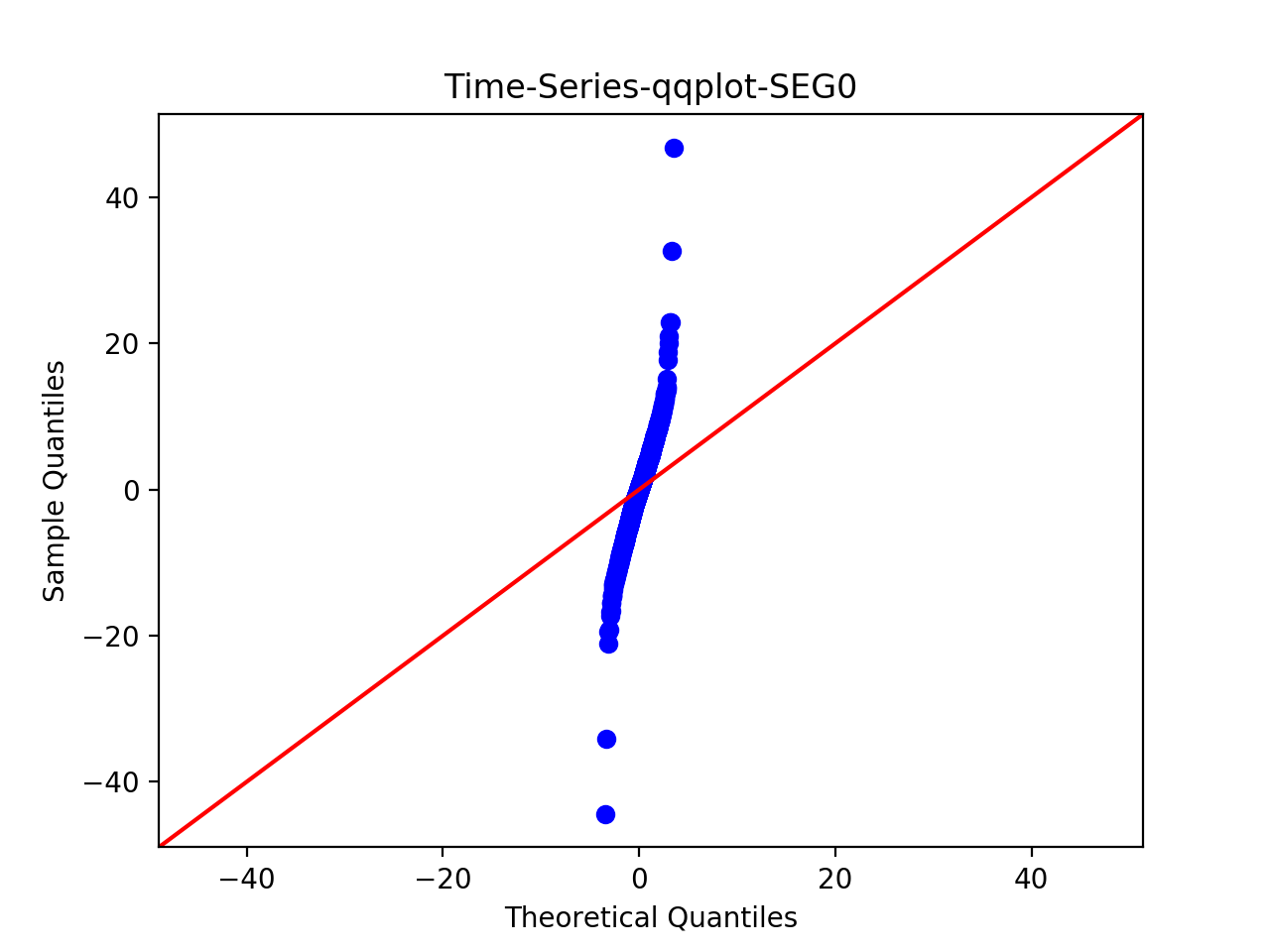
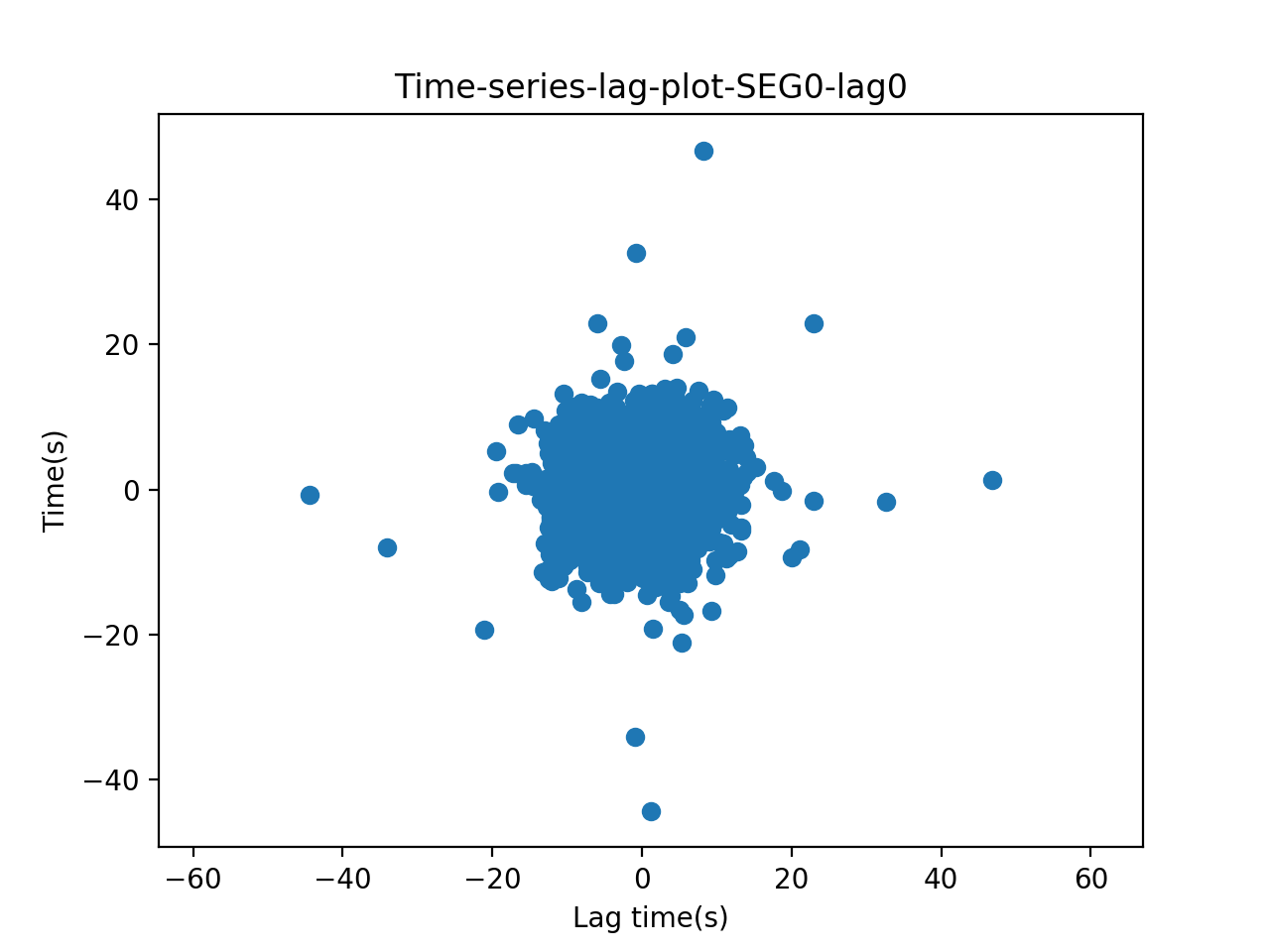
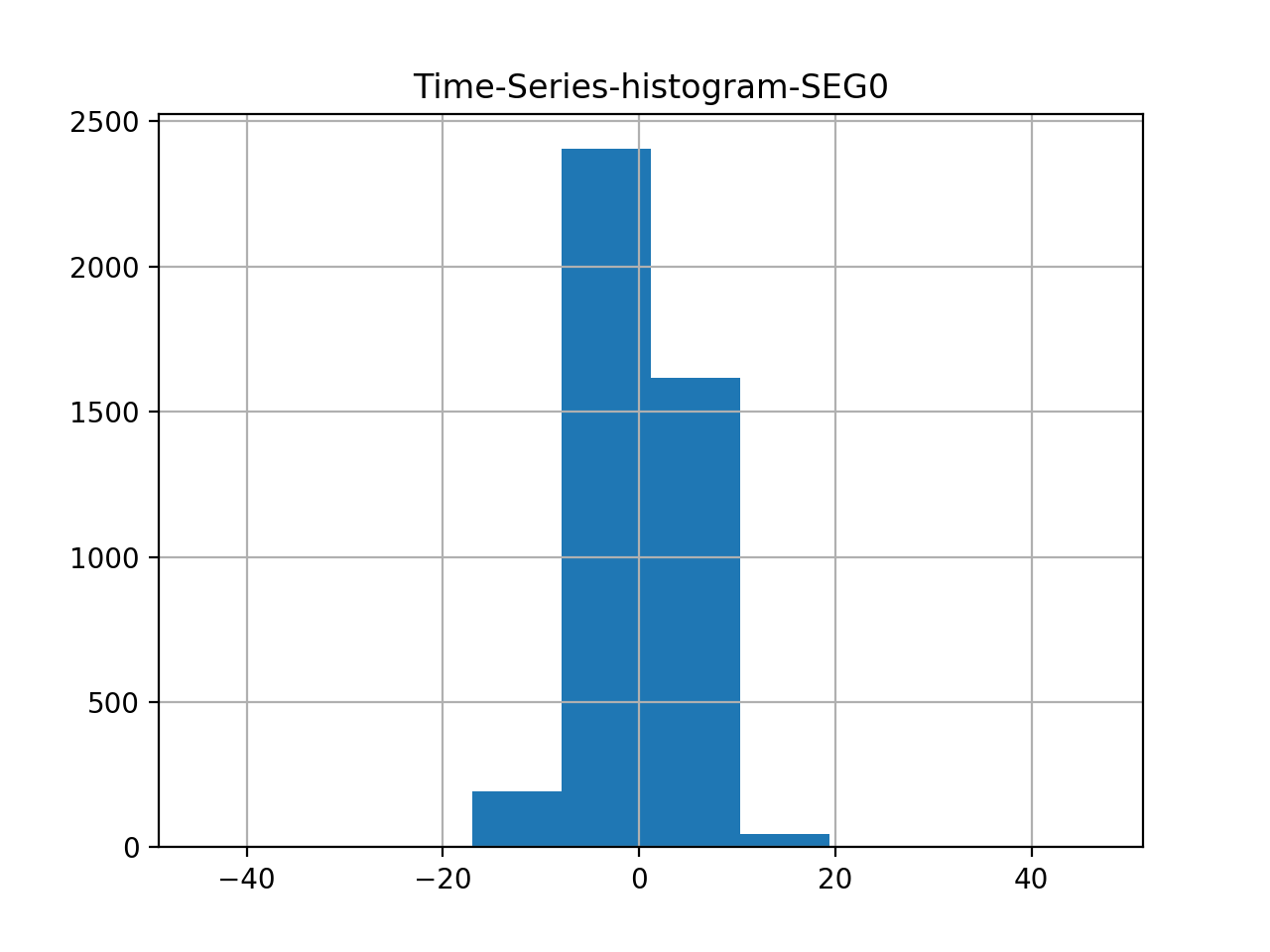
The final chosen model is ARMA(4, 3) because 1) it has lowest BIC and AIC among the three and 2) the ACF and PACF plots support the number of parameters.

Check the residuals to see if the model makes sense.

For ARMA(4, 3) model for SEG0, the 4-plot of residuals are below.

The MSE is 95731.5455165382

The conclusion is that the residuals are not normal. Thus ARMA(4,3) might not be a good fit.



Use Chi-square goodness of fit and set number of bins = 200, results:

Power\_divergenceResult(statistic=44549.608419668788, pvalue=0.0)

The p-value is smaller than 0.05, thus reject null. Thus there is a significant difference between the two data series – thus the model is not a good fit.

# Step 3: Test for OLS regression model, using mllib.

Data transformation: to run regression, since previous fit an ARMA(4,3) model, now use past 4 value as features.

For training data, the mean Squared Error = 21.8713948424

For test data, Mean Squared Error = 25.5037410686

For out-of-sample prediction, after binning there are multiple zeros in the predicted value and also observed values. This indicates an obvious mismatch, because where number of observed values are zero, the probability of predicted value occurring there should also be zero.

Thus, use MLlib for chi-square test returns the following results:

>>> pearson.pValue

0.0

>>> pearson.statistic

inf

Conclude that regression on a single past value is not a good fit, although the MSE is reasonable.

# Step 4: Test for Ridge regression model, using mllib.

Also use past 3 day’s value as features.

For training data, the mean Squared Error = 21.8713948424

For test data, Mean Squared Error = 25.5037410686

Thus, use MLlib for chi-square test returns the following results:

>>> pearson.pValue

0.0

>>> pearson.statistic

inf

Conclude that ridge regression on a single past value is also not a good fit.

# Step 5: Test for Lasso regression model, using mllib.

Again using past three day’s value as feature.

For training data, the mean Squared Error = 21.8713948424

For test data, Mean Squared Error = 25.5037410686

Thus, use MLlib for chi-square test returns the following results:

>>> pearson.pValue

0.0

>>> pearson.statistic

inf

Conclude that Lasso regression on a single past value is also not a good fit.

Conclusion for Step 3 – Step 5: the MSE for OLS, Ridge and Lasso look almost identical with small MSE = 25.5037410686 on test data.

ARMA(4, 3) model has a higher MSE around 95731.