HW2a

YL5090

Yuqiong Li

SEG4- Write up

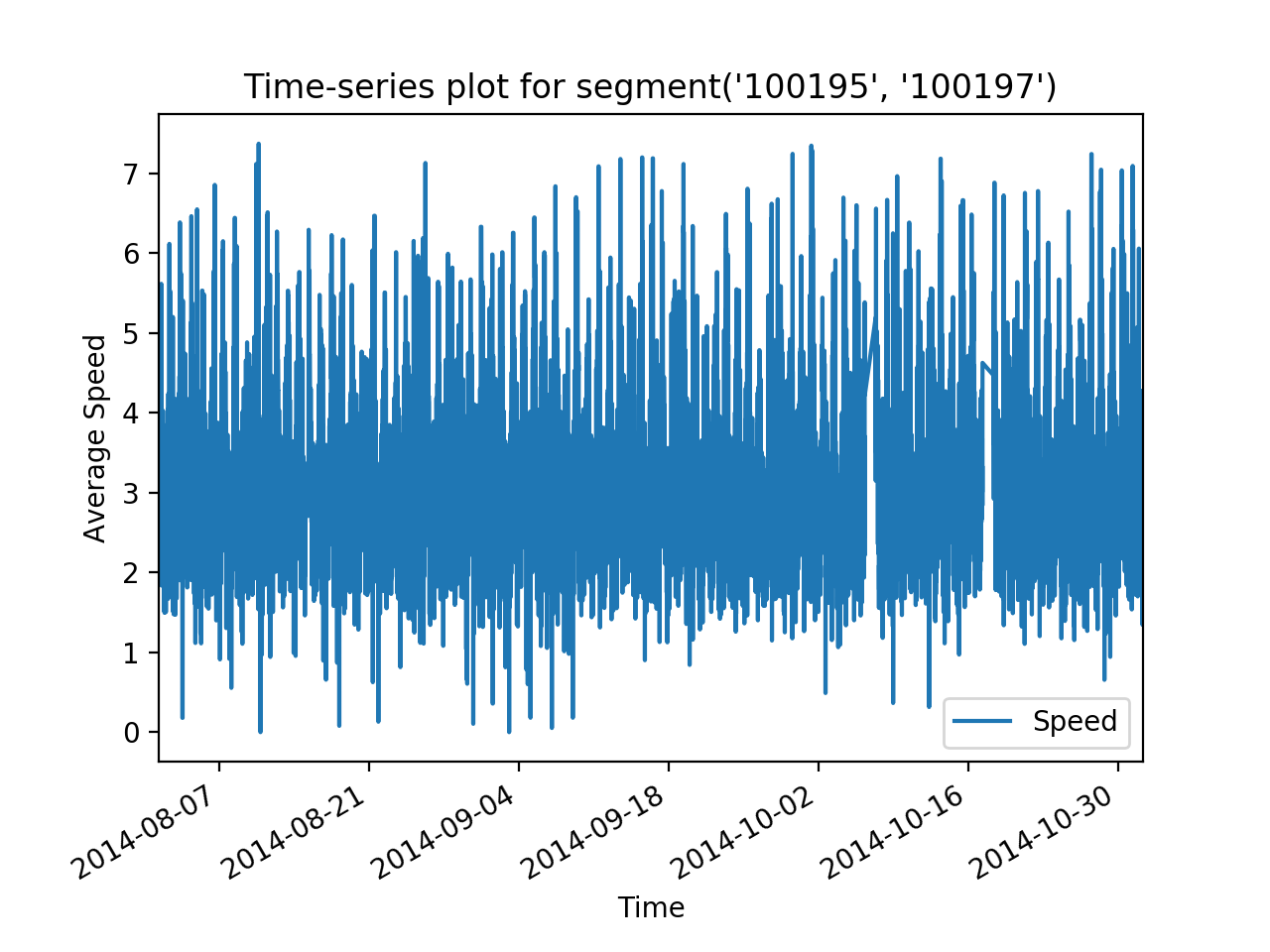
# **Preprocessing:**

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.

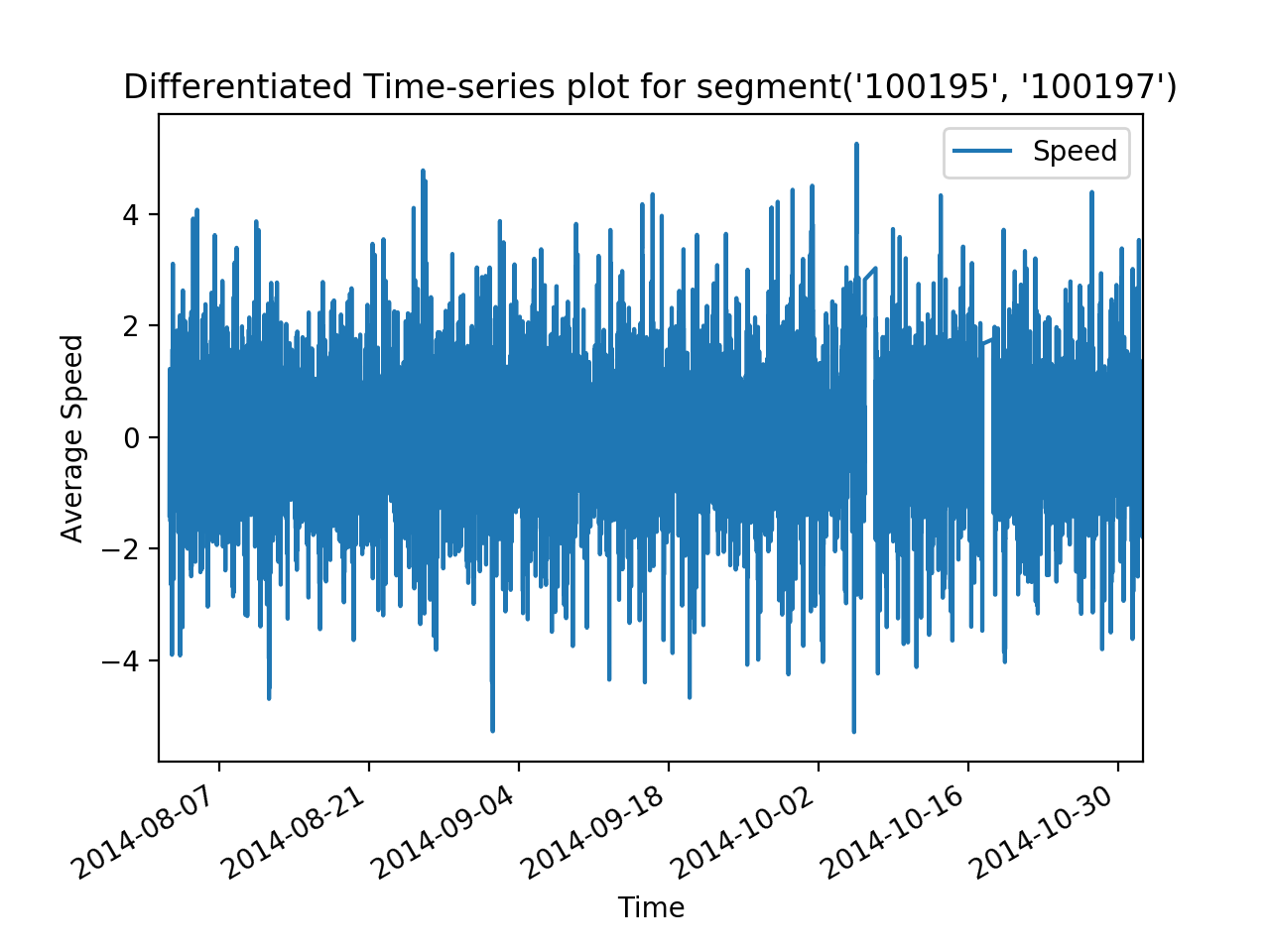
Identification:

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

Step 1: First plot the time-series



After differentiating:



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

ADF statistic: - 22.624604

p-value: 0.000000

Critical Values:

1%: -3.431

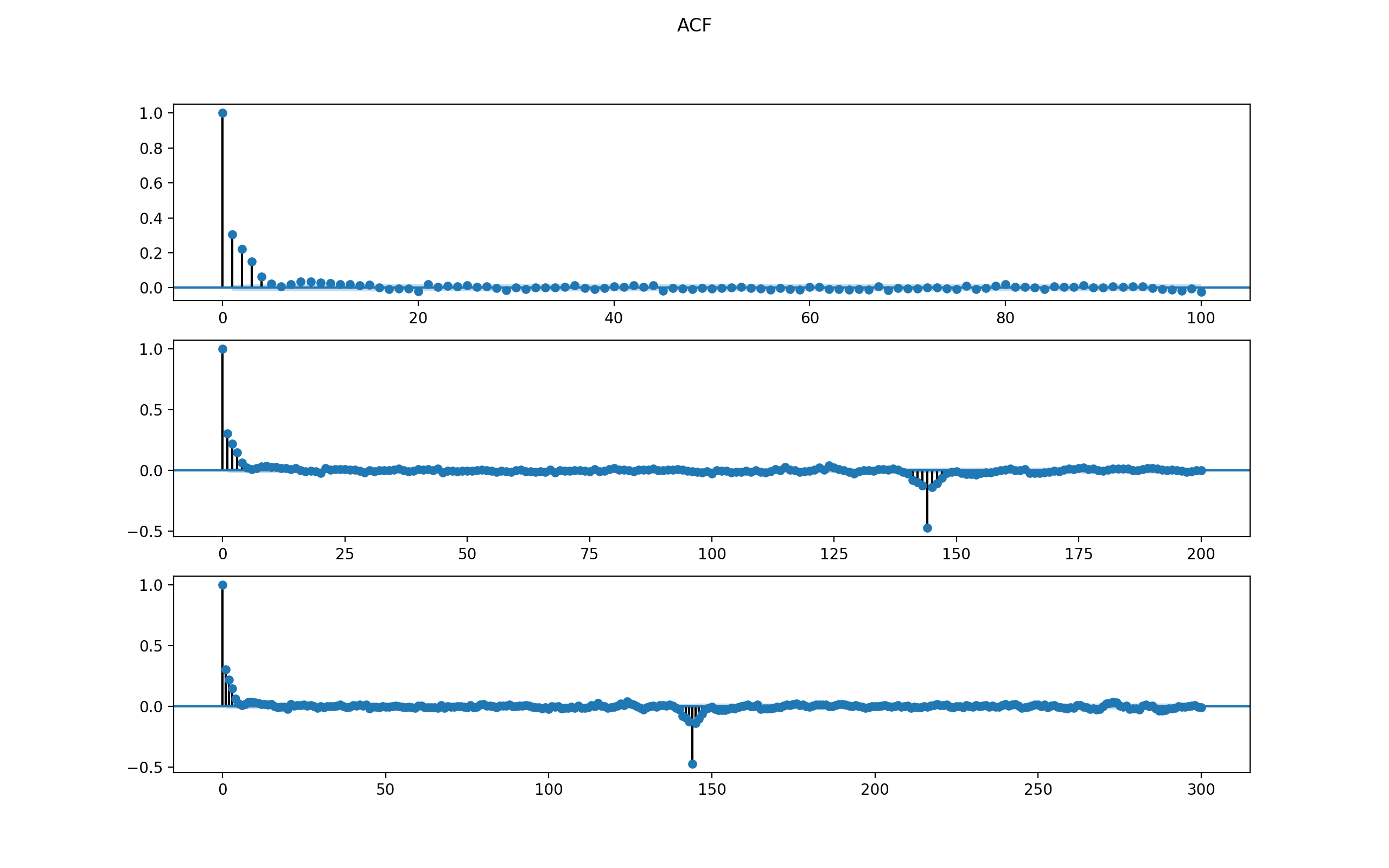
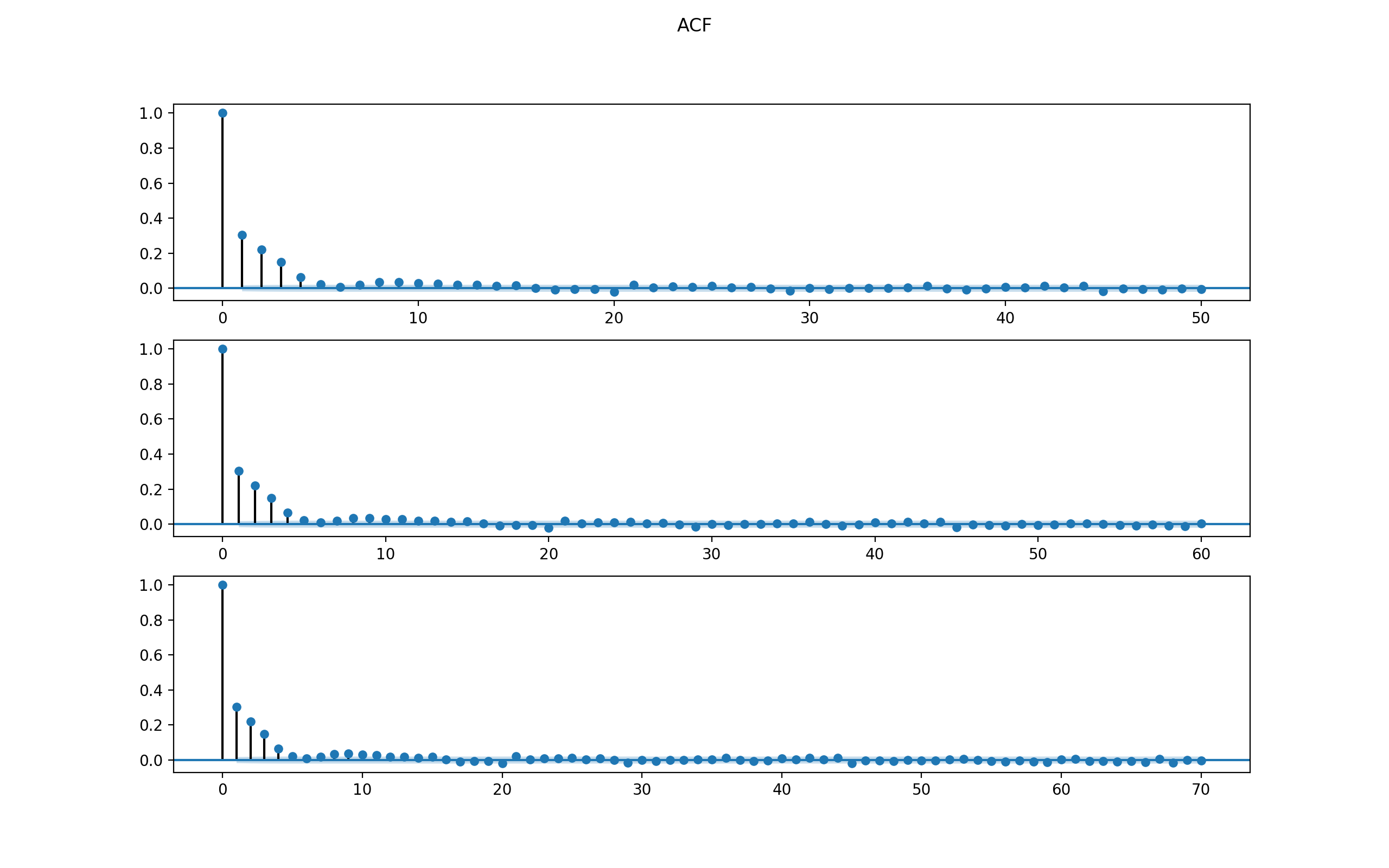
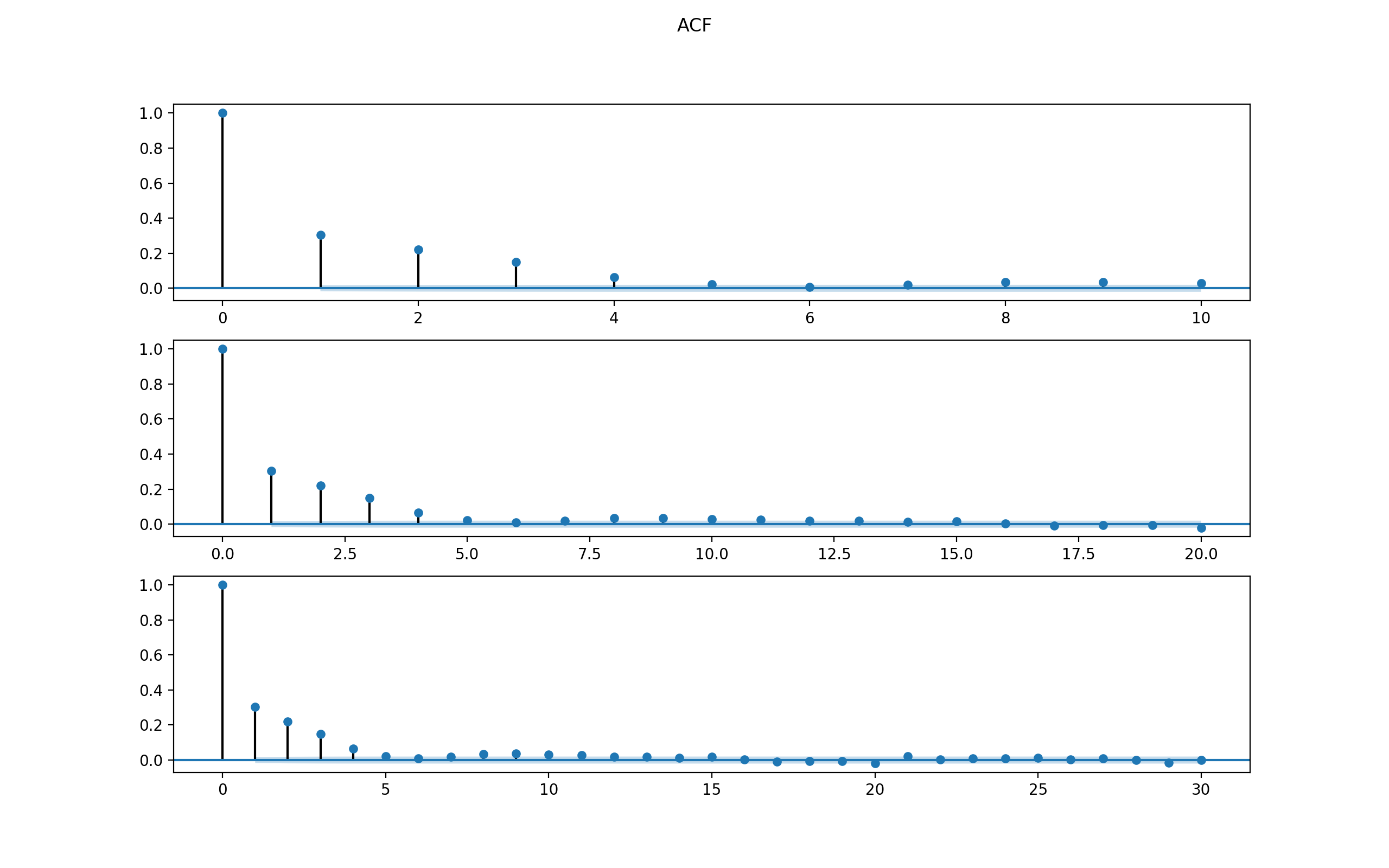
5%: -2.862

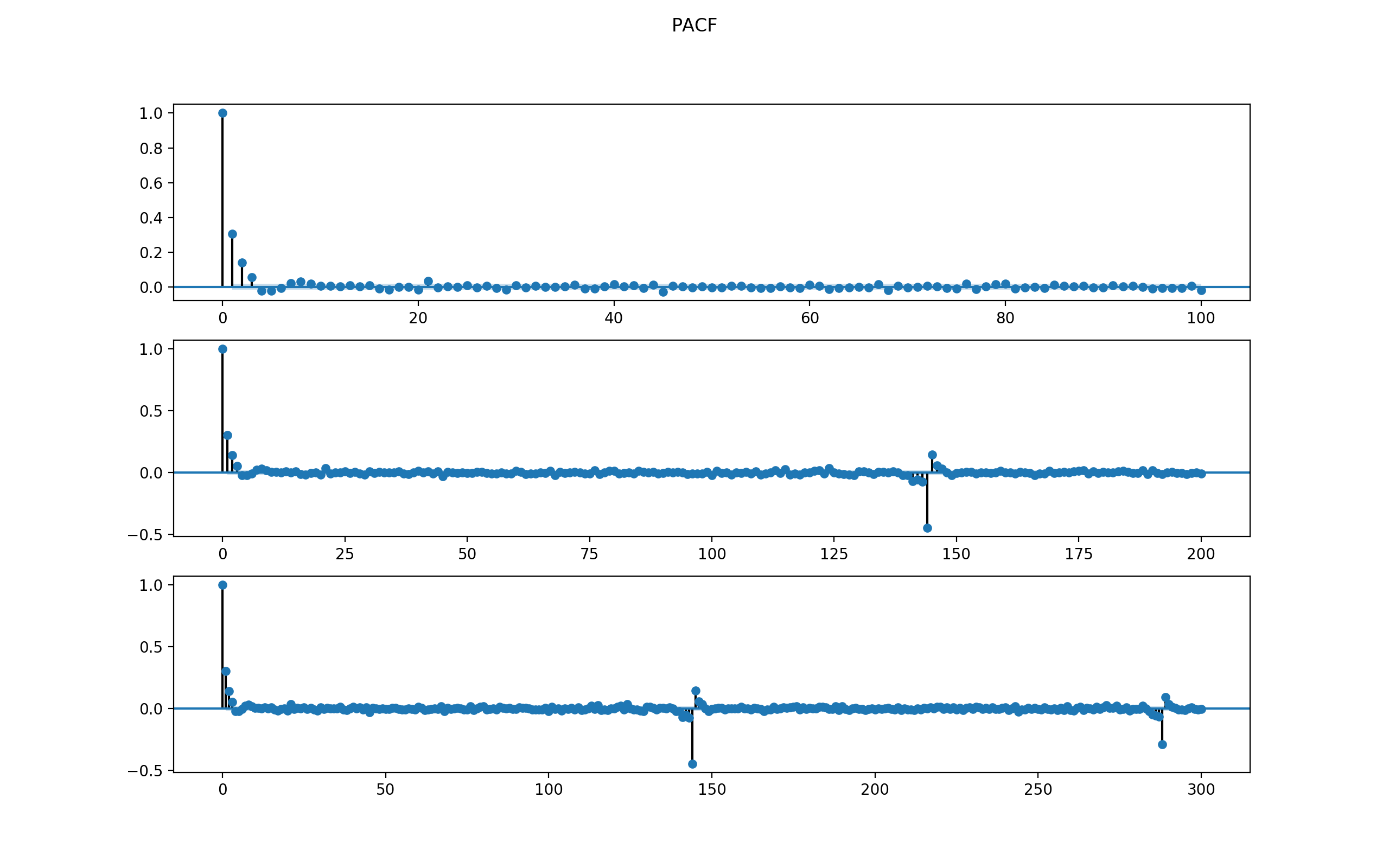
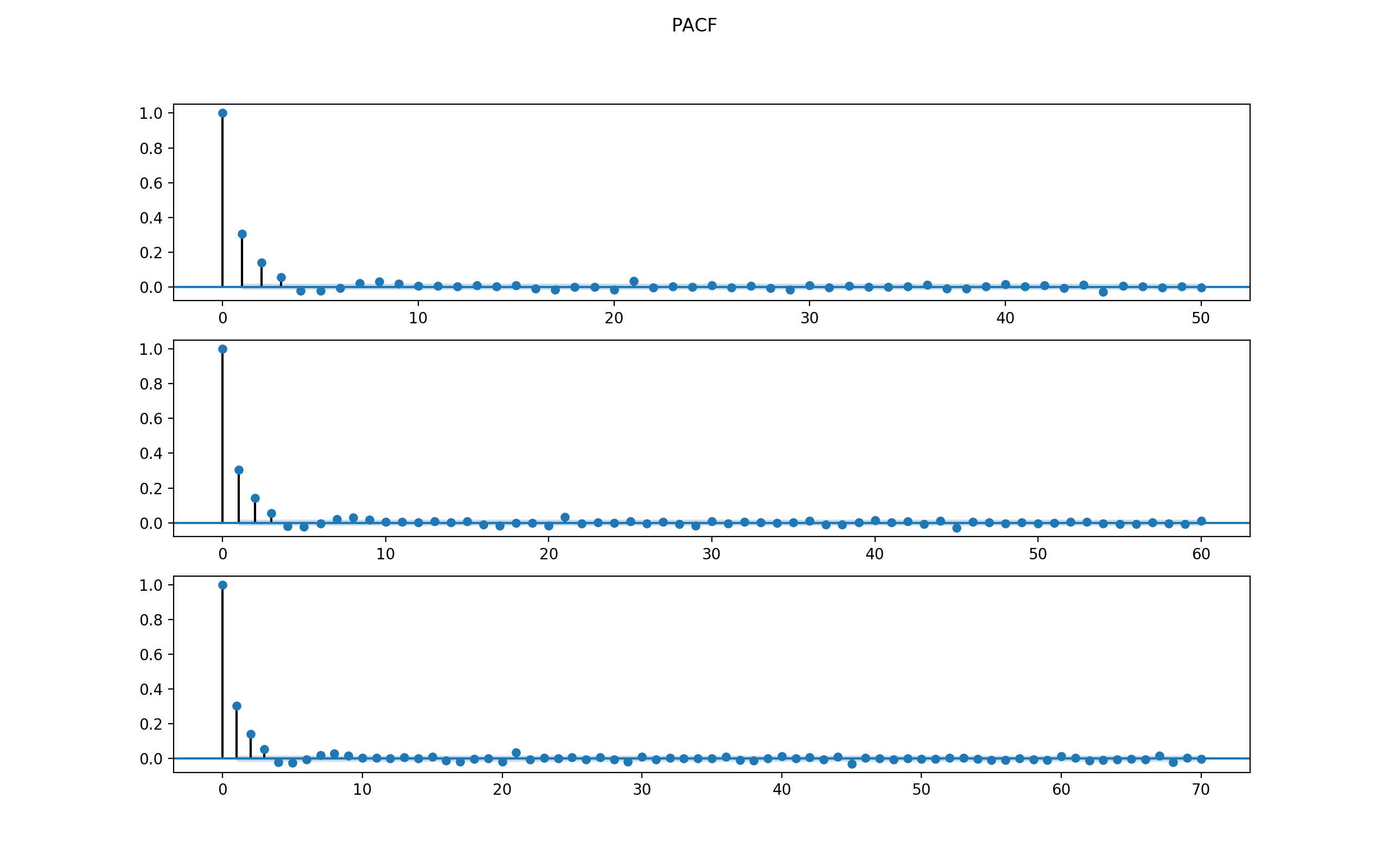
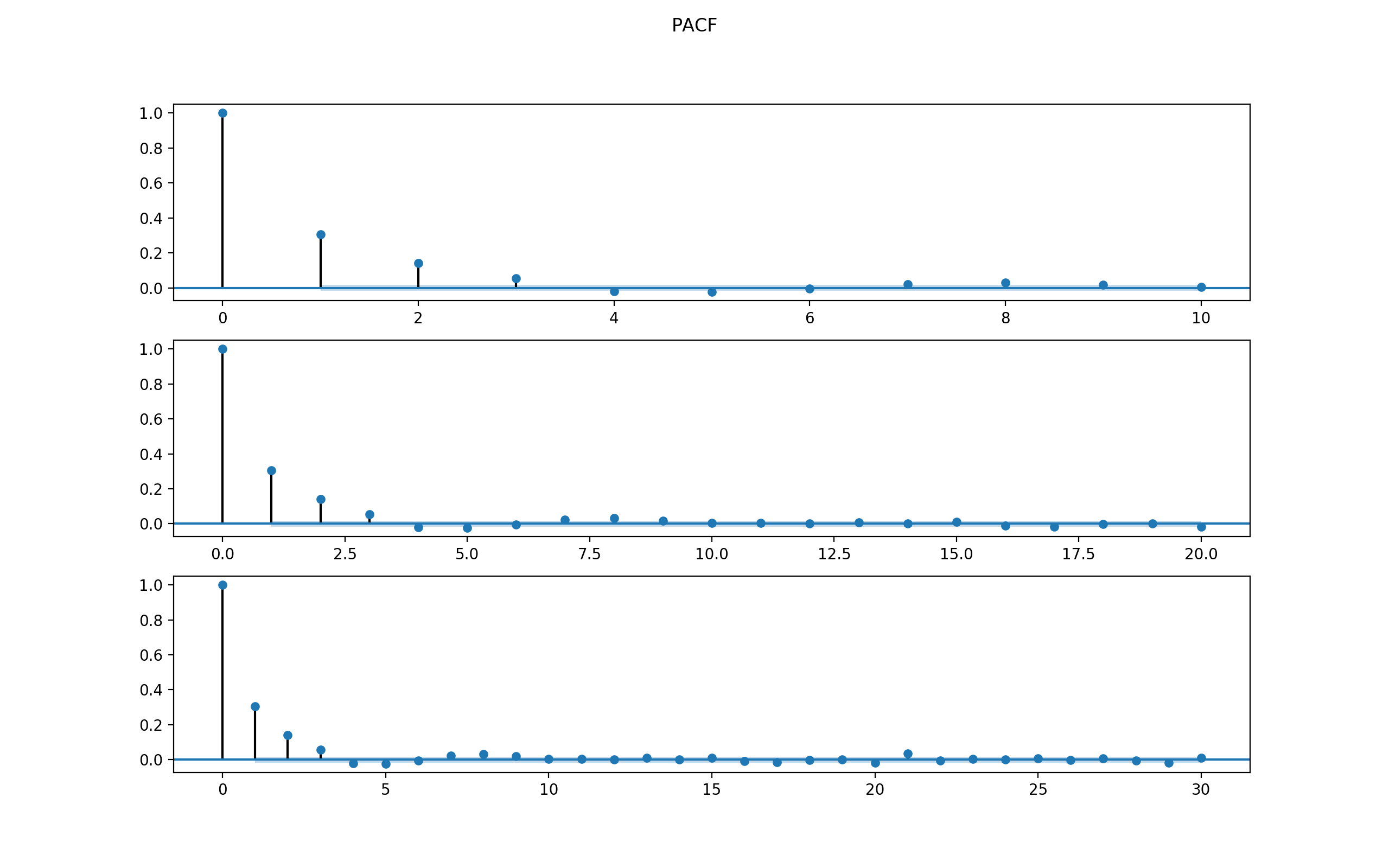
10%: -2.567

Conclusion:

1. Differentiated 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.

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

Observations:



1. ACF dies off after lag 4 or so, indicating AR(4)
2. PACF dies off after lag 3 or so, indicating MA(3).

Thus, the model might be an AR(4), MA(3), or a mixture of ARMA. Test all possible p from 1 to 4, q from 1 to 3, and combination of (p, p) from this range.

ARMA Model Results

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

Dep. Variable: Speed No. Observations: 8544

Model: ARMA(3, 0) Log Likelihood -12268.006

Method: css-mle S.D. of innovations 1.017

Date: Mon, 11 Dec 2017 AIC 24546.011

Time: 16:50:46 BIC 24581.276

Sample: 08-02-2014 HQIC 24558.042

- 09-30-2014

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

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

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

const 0.0003 0.020 0.014 0.989 -0.038 0.039

ar.L1.Speed 0.2383 0.011 22.083 0.000 0.217 0.259

ar.L2.Speed 0.1315 0.011 11.948 0.000 0.110 0.153

ar.L3.Speed 0.0725 0.011 6.717 0.000 0.051 0.094

Roots

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

Real Imaginary Modulus Frequency

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

AR.1 1.5878 -0.0000j 1.5878 -0.0000

AR.2 -1.7008 -2.4075j 2.9477 -0.3479

AR.3 -1.7008 +2.4075j 2.9477 0.3479

ARMA Model Results

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

Dep. Variable: Speed No. Observations: 8544

Model: ARMA(0, 4) Log Likelihood -12268.011

Method: css-mle S.D. of innovations 1.017

Date: Mon, 11 Dec 2017 AIC 24548.021

Time: 16:50:48 BIC 24590.339

Sample: 08-02-2014 HQIC 24562.459

- 09-30-2014

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

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

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

const 0.0003 0.018 0.016 0.987 -0.035 0.036

ma.L1.Speed 0.2390 0.011 22.154 0.000 0.218 0.260

ma.L2.Speed 0.1901 0.011 17.389 0.000 0.169 0.212

ma.L3.Speed 0.1535 0.011 14.120 0.000 0.132 0.175

ma.L4.Speed 0.0567 0.011 5.328 0.000 0.036 0.078

Roots

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

Real Imaginary Modulus Frequency

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

MA.1 0.6794 -1.5935j 1.7323 -0.1859

MA.2 0.6794 +1.5935j 1.7323 0.1859

MA.3 -2.0335 -1.3206j 2.4247 -0.4083

MA.4 -2.0335 +1.3206j 2.4247 0.4083

ARMA Model Results

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

Dep. Variable: Speed No. Observations: 8544

Model: ARMA(1, 3) Log Likelihood -12265.401

Method: css-mle S.D. of innovations 1.017

Date: Mon, 11 Dec 2017 AIC 24542.802

Time: 16:50:53 BIC 24585.120

Sample: 08-02-2014 HQIC 24557.240

- 09-30-2014

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

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

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

const 0.0003 0.019 0.015 0.988 -0.036 0.037

ar.L1.Speed 0.4126 0.063 6.517 0.000 0.289 0.537

ma.L1.Speed -0.1728 0.063 -2.721 0.007 -0.297 -0.048

ma.L2.Speed 0.0948 0.018 5.318 0.000 0.060 0.130

ma.L3.Speed 0.0779 0.018 4.409 0.000 0.043 0.112

Roots

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

Real Imaginary Modulus Frequency

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

AR.1 2.4234 +0.0000j 2.4234 0.0000

MA.1 0.9825 -1.7522j 2.0088 -0.1687

MA.2 0.9825 +1.7522j 2.0088 0.1687

MA.3 -3.1827 -0.0000j 3.1827 -0.5000

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

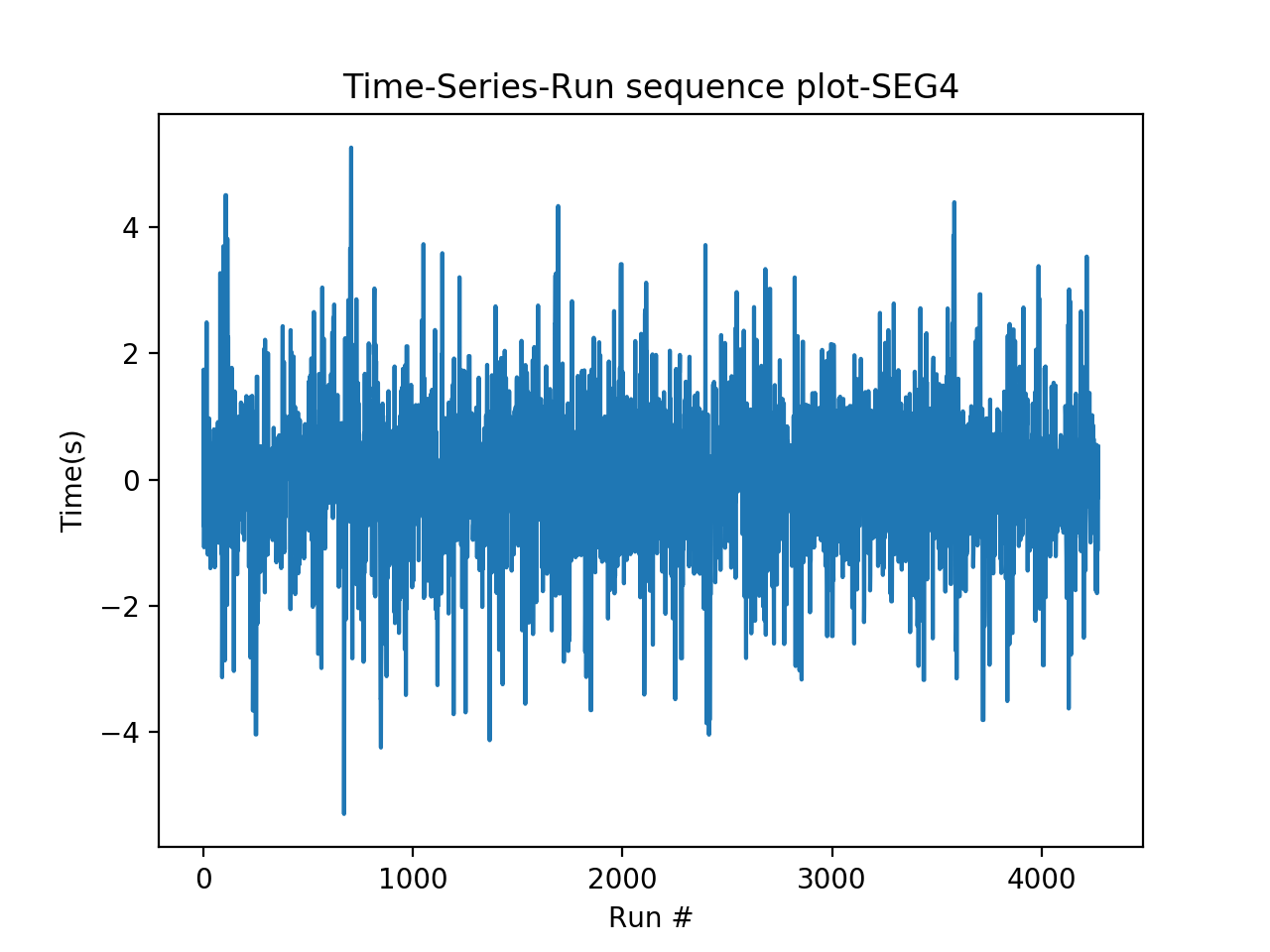
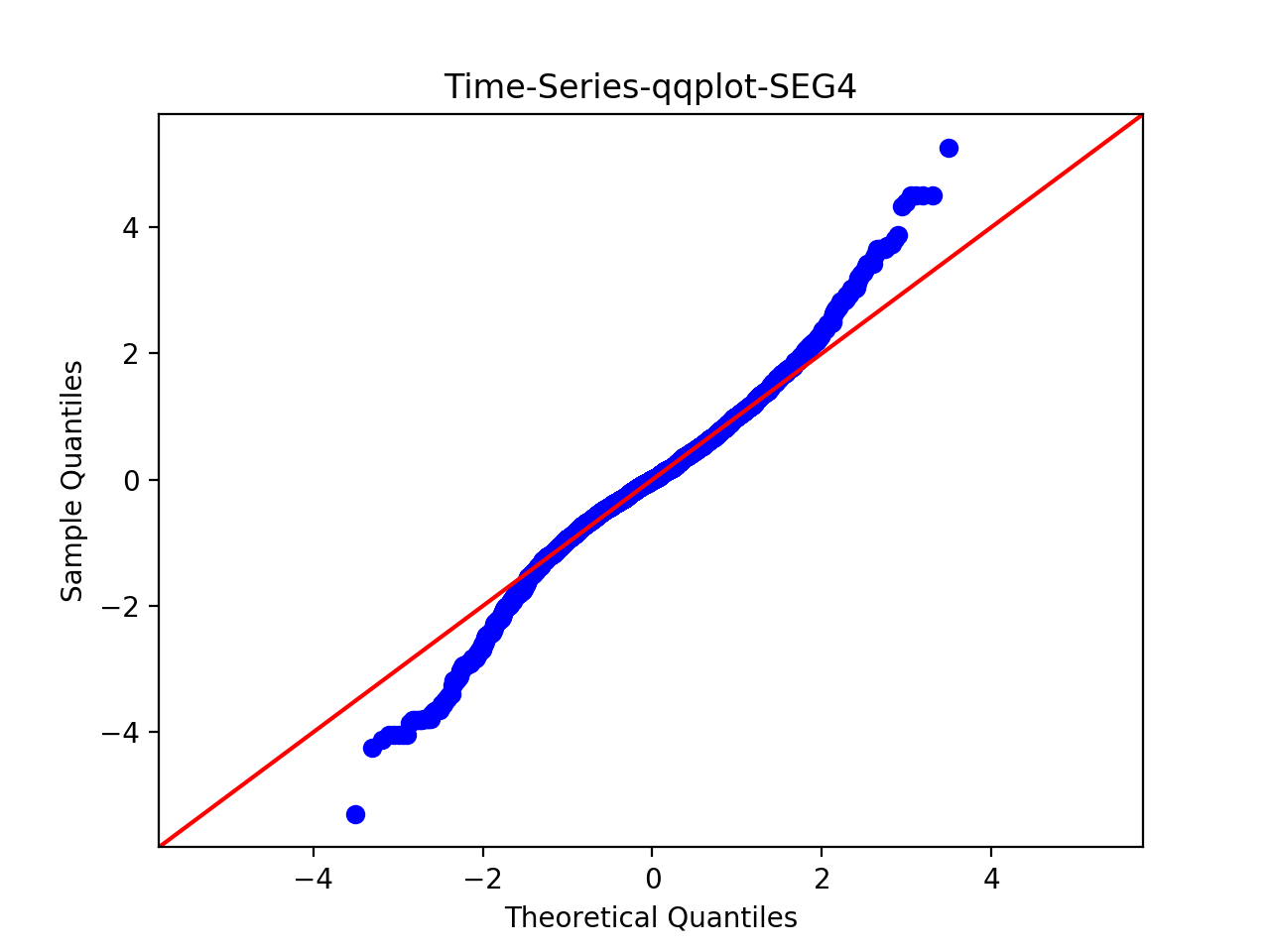
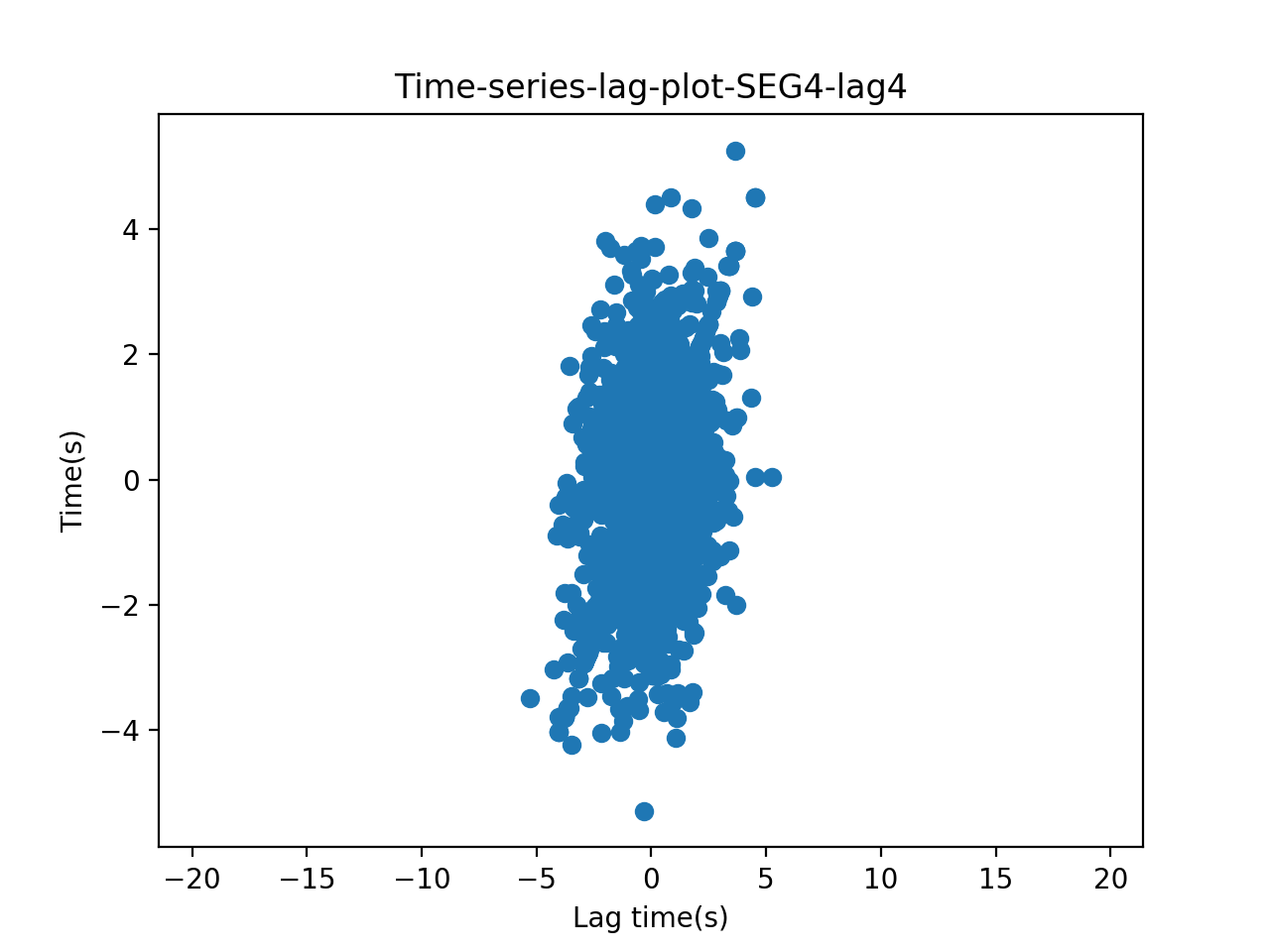
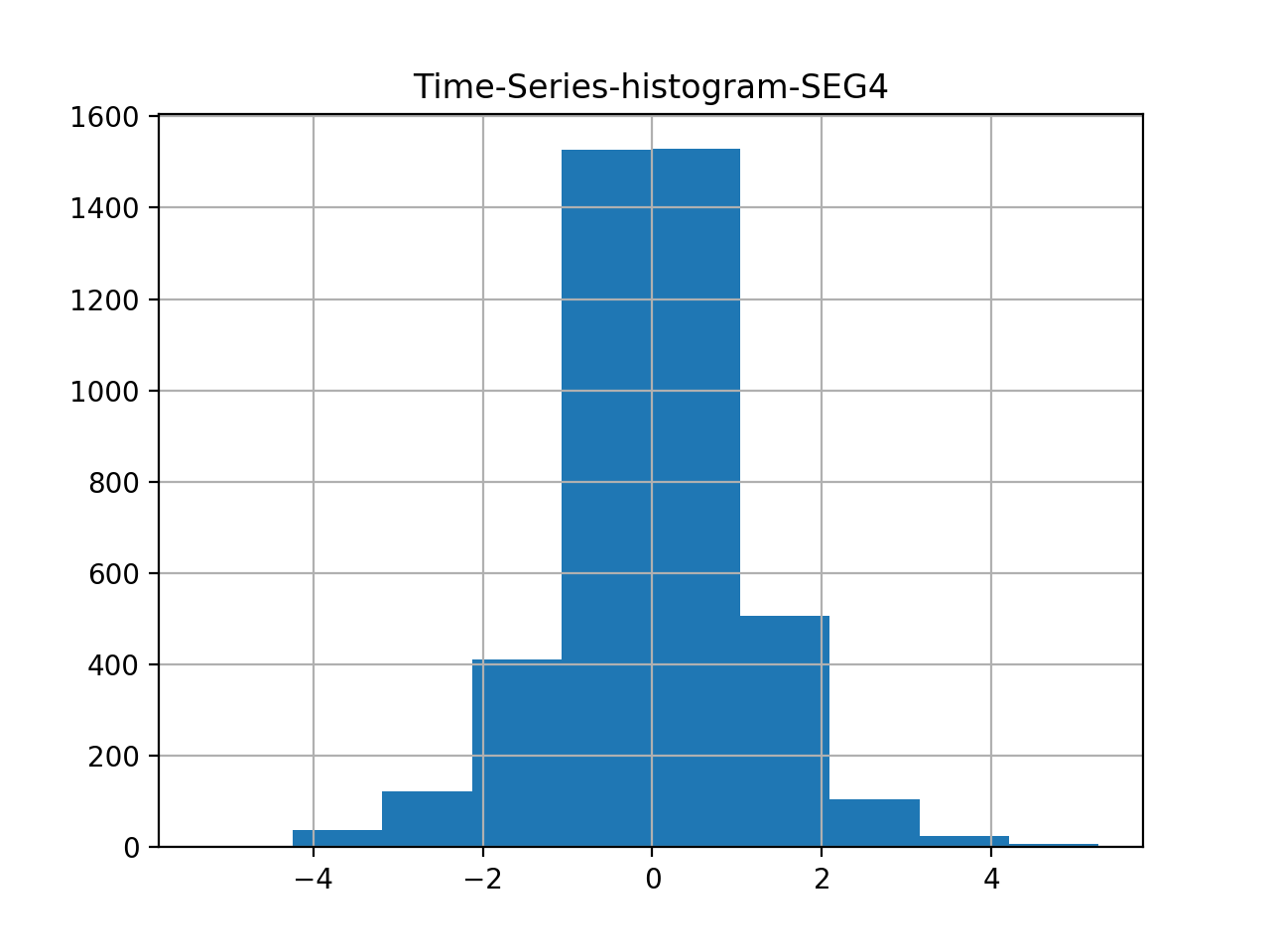
The final chosen model is AR(3) because 1) the PACF supports AR(3) and 2) it has smallest BIC compared to other models and 3) it also has the fewest parameters.

Check the residuals to see if the model makes sense.

For AR(3) model for SEG4, the 4-plot of residuals are below.

The MSE is 5416.079025446667. Seems to be a good fit compared to other modelling segments.

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



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

Power\_divergenceResult(statistic=11253.0539801578, 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 this model might not be a good fit. However the QQ plot looks ok.

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

Data transformation: to run regression, since previous fit an AR(1) model, now use past day’s value as features.

For training data, the mean Squared Error = 0.929070477989

For test data, Mean Squared Error = 0.943797669086

This seems to be a good fit.

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, but the MSE value is small.

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

Again using past three day’s value as feature.

For training data, the mean Squared Error = 0.929070477989

For test data, Mean Squared Error = 0.943797669086

The goodness of fit value:

>>> pearson.pValue

0.0

>>> pearson.statistic

inf

Conclude that ridge regression on a single past value is also not a good fit, although the MSE values are small.

# 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 = 0.929070477989

For test data, Mean Squared Error = 0.943797669086

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:

Ridge, OLS, and Lasso seem to generate similar results. The MSE for test set is 0. 943797669086

AR(1) model, the MSE for test data is 5416.079025446667.