DATS 6450 Time Series Analysis & Modeling

Professor Reza Jafari

Final Report

KONG Deyu

Dec. 13, 2021

Table of Content

Table of Content 2

Abstract 3

Introduction 3

About The Dataset 4

Part 1. Data Cleaning 5

Part 2. Data Overview 6

Part 3. Base Models 9

Part 4. SARIMA Modelling 12

Summary and Conclusion 18

Appendix 19

Abstract

This project focus on analyze Metro Interstate Traffic Volume data retrieved from UCI Machine Learning Depository. After duplicates and outliers are detected and removed, we will subset the data set, and perform decomposition and stationarity check. After that we will develop base models and evaluate their forecast performance. Lastly, we will build a SARIMA model, derive prediction and forecast function, and examine how its forecast fair with base models.

Introduction

After cleaning duplicates and reindexing the data with date time, we will subset the dataset and impute missing traffic volume observations with drift method. STL decomposition will be used to detect strength of trend and seasonality. Base models including average, naïve, drift, SES, Holt linear and Holt-Winters methods will be developed, and have forecast compared with linear regression models. SARIMA model will be formulated by properly difference the series both seasonally and non-seasonally. We will then use GPAC and ACF/PACF to determine the order of AR and MA structure to fit a multiplicative SARIMA model with LM algorithm.

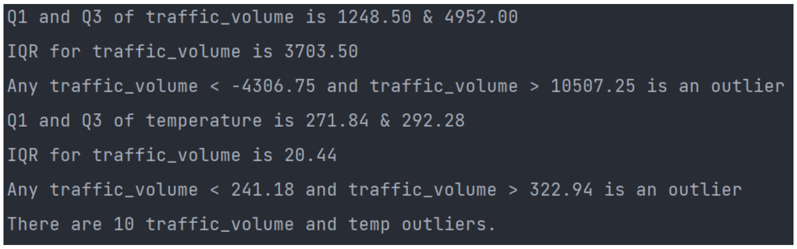
About The Dataset

The Metro Interstate Traffic Volume data retrieved from UCI Machine Learning Repository[1]. It is a time series data that contains 48,204 hourly observations for the traffic volume westbound I-94 in Minnesota from 2012-2018. Includes temperature, perspiration, weather types and holiday features. The target variable will be the traffic volume.

[1] https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume

Part 1. Data Cleaning

The dataset is being cleaned by removing duplicates based on ‘date_time’ column, and a total of 7629 duplicates are removed. Then the dataset is reindexed based on cleaned date time. IQR method is used to detect and remove outliers from ‘temp’ and ‘traffic_volume’ with following result:



```
Q1 and Q3 of traffic_volume is 1248.50 & 4952.00
IQR for traffic_volume is 3703.50
Any traffic_volume < -4306.75 and traffic_volume > 10507.25 is an outlier
Q1 and Q3 of temperature is 271.84 & 292.28
IQR for traffic_volume is 20.44
Any traffic_volume < 241.18 and traffic_volume > 322.94 is an outlier
There are 10 traffic_volume and temp outliers.
```

Fig 1. Result from IQR Outlier Detection

The dataset does not contain incomplete rows (all rows are column filled). But there are missing observations during this time-span of study. Considering the time-dependency nature of time series study, we will use a subset of data from Janurary 1st, 2016 onward where missing data are scarce. However, we found there are still 1012 missing observations in this subset, we will impute them using pd.Dataframe.interpolate method, which like drift methods, impute the missing data by average of non-missing observation before and after.

Part 2. Data Overview

First, we should have a look of the data by plotting the series against time as follows:

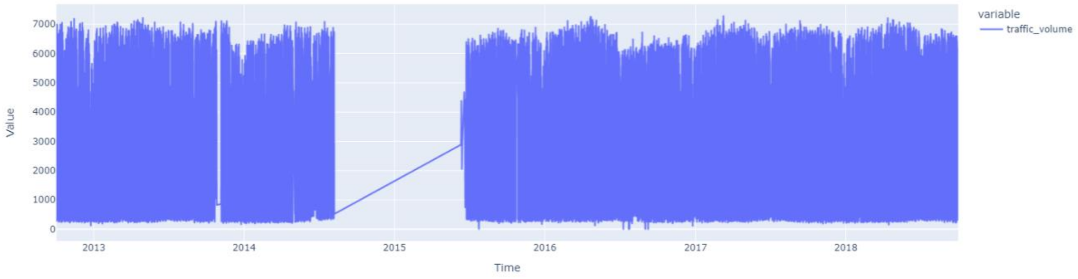


Fig 2(a). Line Plot of Traffic Volume from 2013 to 2018

We can see there is a massive gap of missing data an entire year around late 2014 to 2015. We can also spot missing observations here and there. For time series specified models with time dependency assumption, we will only use the data subset after 2016 where missing data is scarce.

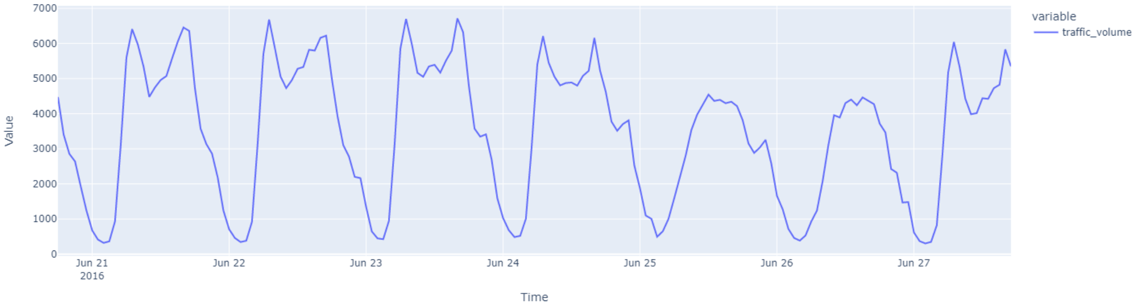


Fig 2(b). Hourly Line Plot of Traffic Volume

Here displays a randomly sliced time frame of weekly traffic (168 hours). We can spot a weekly seasonality when traffics are lower on the weekends. Daily patterns are significant as well, with morning and evening rush hours reoccurring. These findings make good candidates for seasonality modeling.

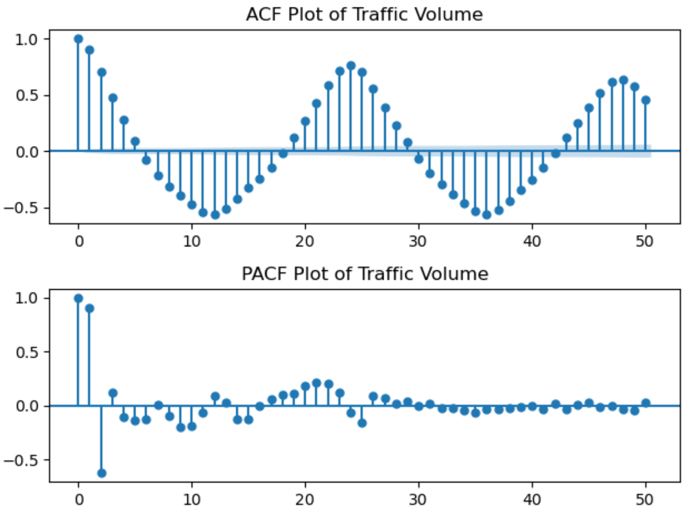
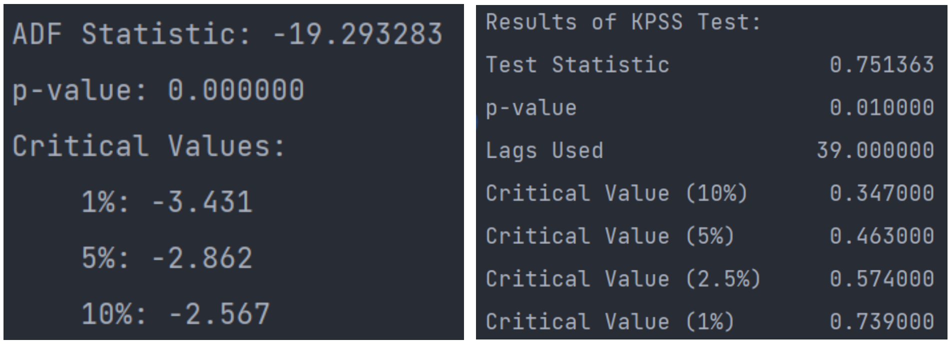


Fig 3. ACF/PACF Plot of Traffic Volume, lags=50

ACF/PACF Plot shows that the series might not be stationary, and needs to be differenced both seasonally and non-seasonally.



```
ADF Statistic: -19.293283
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
```

```
Results of KPSS Test:
Test Statistic              0.751363
p-value                     0.010000
Lags Used                  39.000000
Critical Value (10%)        0.347000
Critical Value (5%)         0.463000
Critical Value (2.5%)       0.574000
Critical Value (1%)         0.739000
```

Fig 4. ADF/KPSS Test Result of Traffic Volume

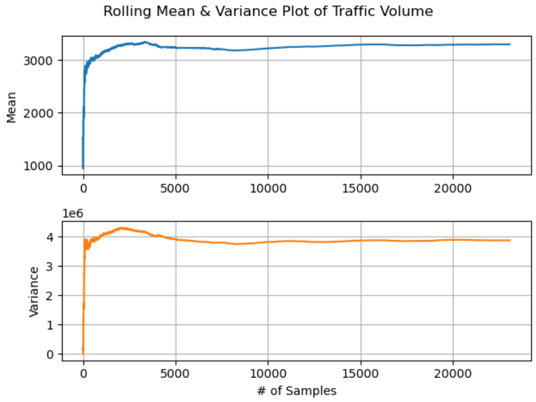


Fig 5. Rolling Mean and Variance of Traffic Volume

We can see that the traffic volume data passed ADF test but failed KPSS test. We should conclude that this series is not stationary. Correlation coefficient heatmap between numerical variables are generated as below:

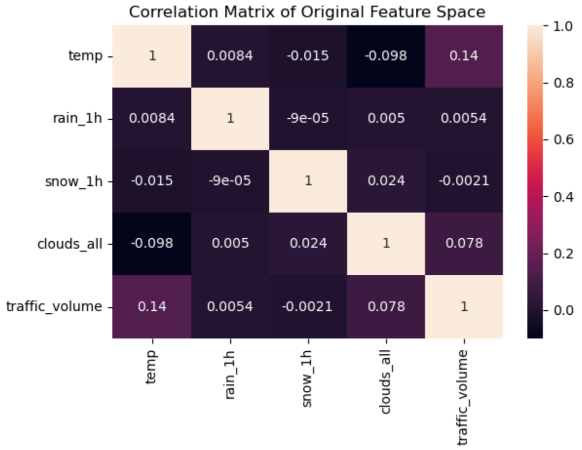


Fig 6. Correlation Heatmap

We can see that all 4 features are not significantly correlated with each other, or with traffic volume.

Now we can use STL decomposition to detect the strength of trend and seasonality of the data. After decomposition, the strength of trend is 0.528, the strength of seasonality is 0.933. It is obvious that this time series is highly seasonal.

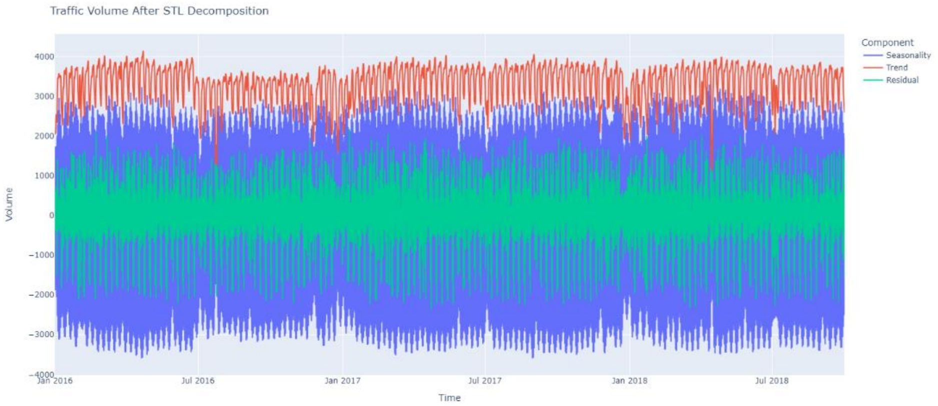


Fig 7. STL Decomposition

Part 3. Base Models

In this part, we will split 5% of the data (Considering this is hourly data, larger testing set does not make much sense) for testing and construct base models using average, naïve, drift and simple exponential smoothing (SES). Holt-Linear, Holt-Winters and OLS regression model will also be considered and evaluated.

For OLS regression, we further convert weather types into dummy variables to examine the effect of these weathers on traffic. All weather types except clear days are converted to avoid multi-collinearity. We shall conduct PCA for feature reduction first:

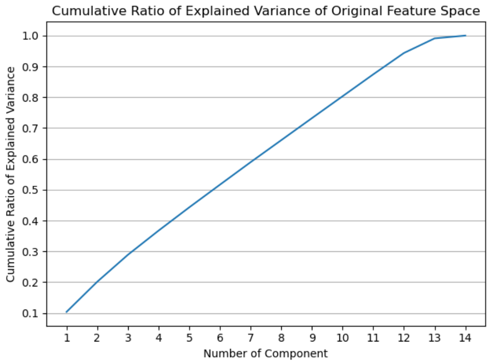
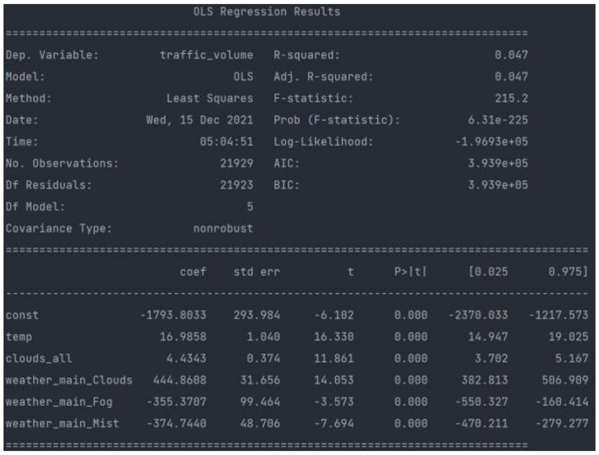


Fig 8. Cumulative Ratio of Explained Variance

We can see that 12 components are needed if we want to have at least 90% of variance represented. We also calculated the condition number of this feature space being 3.4, suggesting the degree co-linearity is limited. After performing a backward selection process on OLS models, we have the following model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         traffic_volume   R-squared:                       0.047
Model:                            OLS   Adj. R-squared:                  0.047
Method:                 Least Squares   F-statistic:                     215.2
Date:                Wed, 15 Dec 2021   Prob (F-statistic):          6.31e-225
Time:                        05:04:51   Log-Likelihood:            -1.9693e+05
No. Observations:               21929   AIC:                         3.939e+05
Df Residuals:                   21923   BIC:                         3.939e+05
Df Model:                           5
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const               -1793.8033    293.984     -6.102      0.000   -2370.033   -1217.573
temp                   16.9858      1.040     16.330      0.000      14.947      19.025
clouds_all              4.4343      0.374     11.861      0.000       3.702       5.167
weather_main_Clouds   444.8608     31.656     14.053      0.000     382.813     506.909
weather_main_Fog     -355.3707     99.464     -3.573      0.000    -550.327    -160.414
weather_main_Mist    -374.7440     48.706     -7.694      0.000    -470.211    -279.277
==============================================================================
```

Fig 9. OLS Regression Summary

Apart from temperature, cloud, fog and mist have significant impact on traffic. We can see from adjusted R^2 that only 4.7% of data have been explained by this linear model, and we have a large BIC, suggesting the model is weak.

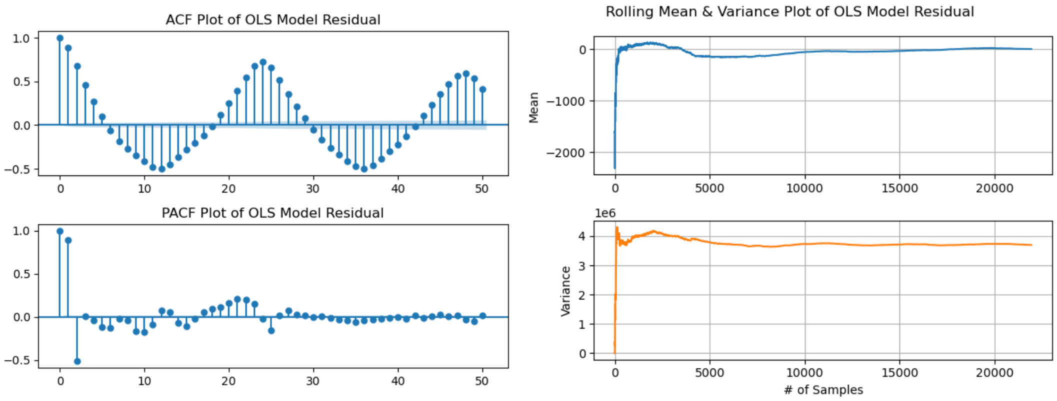


Fig 10. OLS Model Residual Analysis

As expected, the residual from OLS model is not white. Variations in residual mean and variance are still visible.

Models using average, naïve, drift, SES, Holt linear and Holt-Winters models are constructed. Forecasts along with RSME from all models are generated, graphed and displayed as follows:

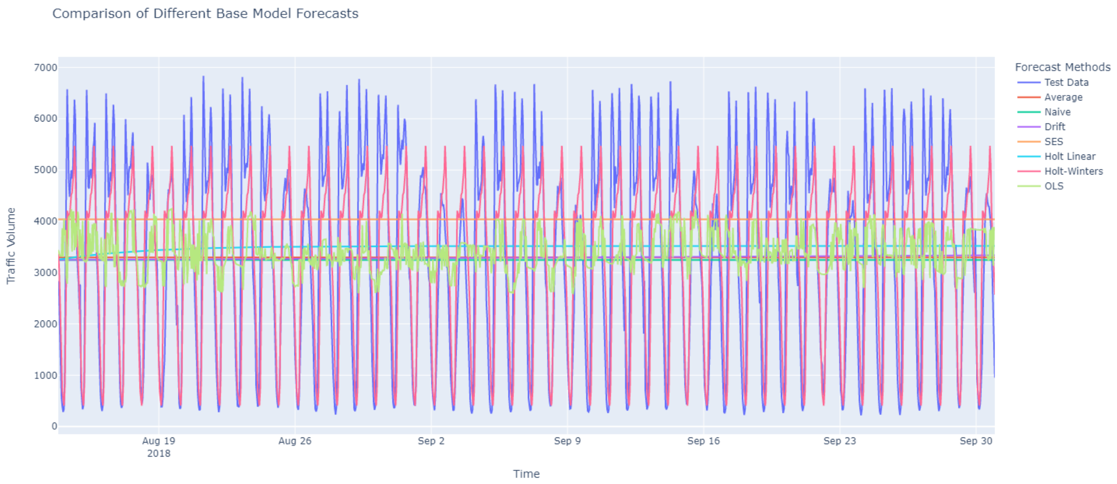
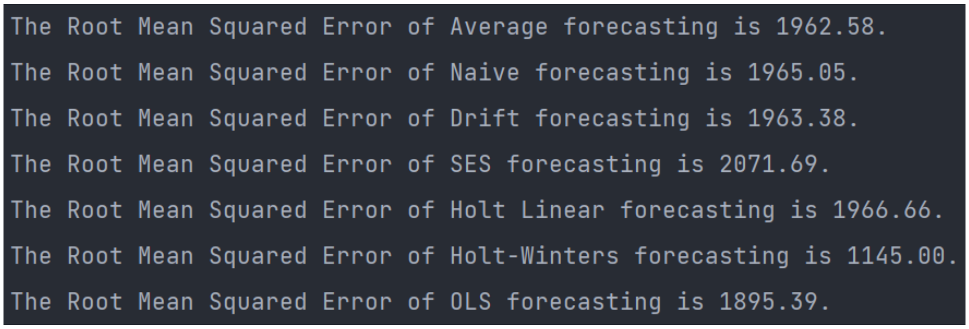


Fig 11. Model Forecast Results Compared with Test Data



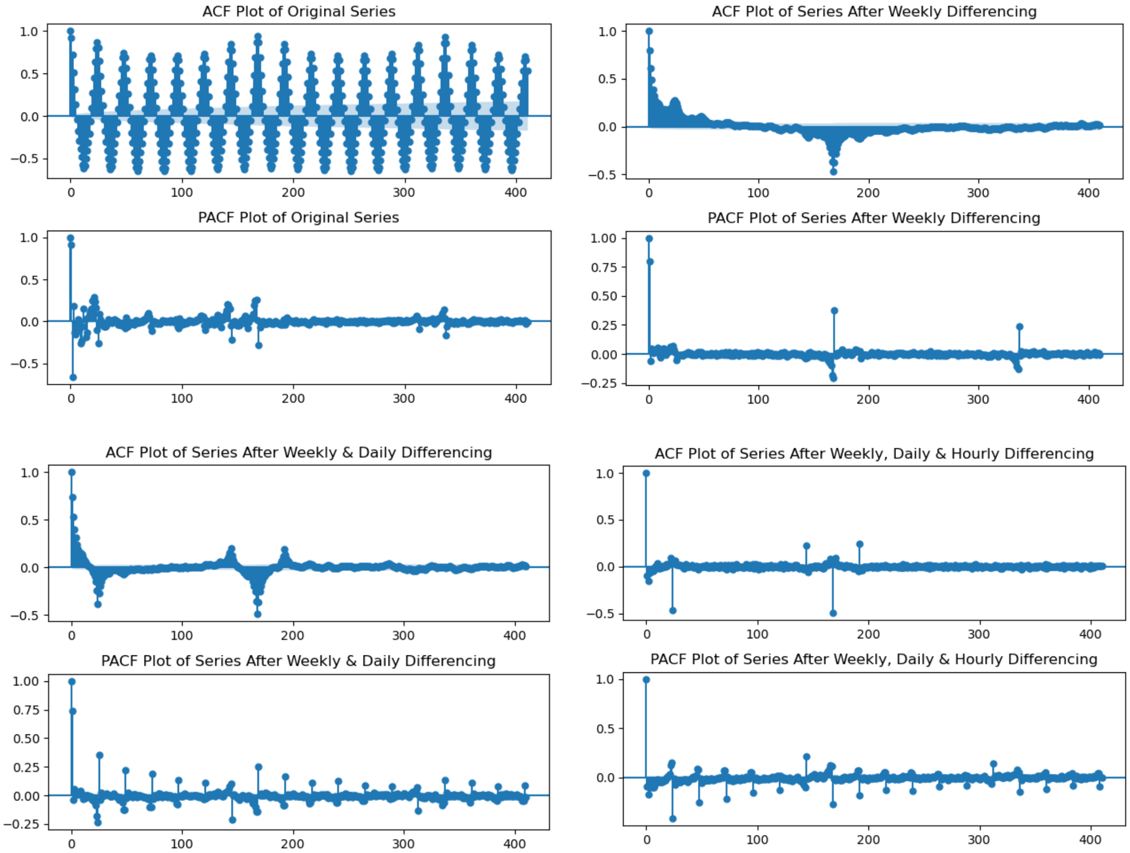
```
The Root Mean Squared Error of Average forecasting is 1962.58.
The Root Mean Squared Error of Naive forecasting is 1965.05.
The Root Mean Squared Error of Drift forecasting is 1963.38.
The Root Mean Squared Error of SES forecasting is 2071.69.
The Root Mean Squared Error of Holt Linear forecasting is 1966.66.
The Root Mean Squared Error of Holt-Winters forecasting is 1145.00.
The Root Mean Squared Error of OLS forecasting is 1895.39.
```

Fig 12. Model Forecast Results

We can see from both forecast plot and RMSE that Holt-Winters gives the best model so far.

Part 4. SARIMA Modelling

The time series is standardized and made stationary by differencing. So far, we observed daily (season of 24) and weekly (season of 168) seasonality.



ACF Plot of Original Series

ACF Plot of Series After Weekly Differencing

PACF Plot of Original Series

PACF Plot of Series After Weekly Differencing

ACF Plot of Series After Weekly & Daily Differencing

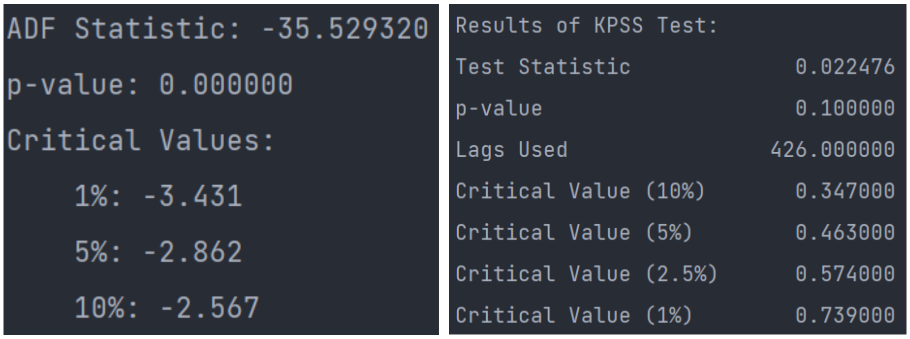
ACF Plot of Series After Weekly, Daily & Hourly Differencing

PACF Plot of Series After Weekly & Daily Differencing

PACF Plot of Series After Weekly, Daily & Hourly Differencing

Fig 13. Differencing Process

After Weekly (order of 1), Daily (order of 1), and Hourly (order of 1) differencing, ACF/PACF Plots start to make sense. ADF/KPSS test results as follows:

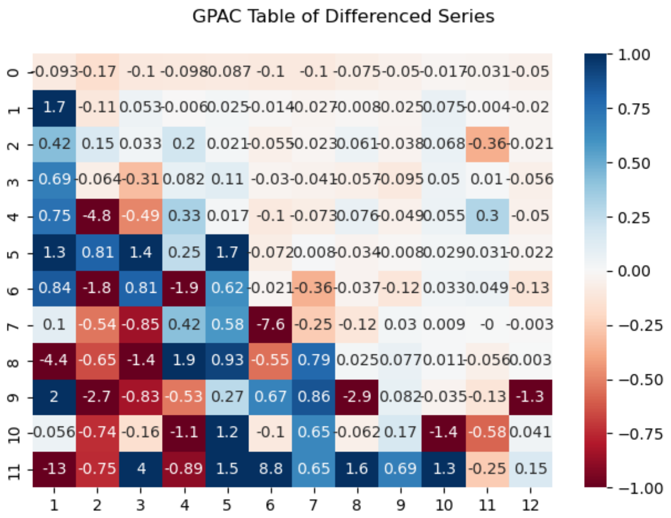


```
ADF Statistic: -35.529320
p-value: 0.000000
Critical Values:
    1%: -3.431
    5%: -2.862
    10%: -2.567
```

```
Results of KPSS Test:
Test Statistic               0.022476
p-value                      0.100000
Lags Used                  426.000000
Critical Value (10%)         0.347000
Critical Value (5%)          0.463000
Critical Value (2.5%)        0.574000
Critical Value (1%)          0.739000
```

Fig 14. ADF/KPSS Result After Differencing

After differencing, the series passed both ADF and KPSS test, and we can say it is stationary. The GPAC of differenced series is as follows:


GPAC Table of Differenced Series

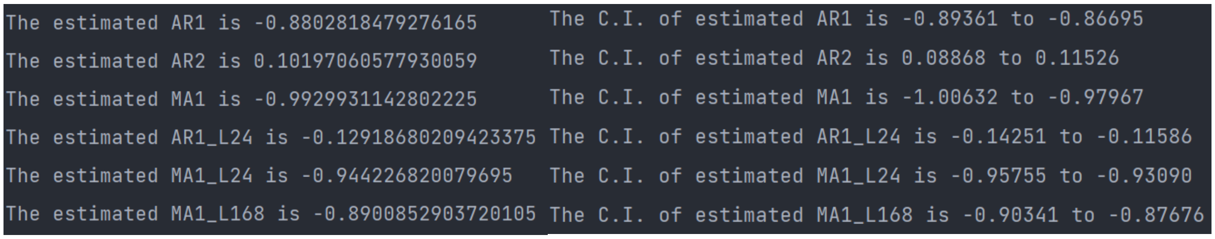
	1	2	3	4	5	6	7	8	9	10	11	12
0	-0.093	-0.17	-0.1	-0.098	0.087	-0.1	-0.1	-0.075	-0.05	-0.017	0.031	-0.05
1	1.7	-0.11	0.053	-0.006	0.025	-0.014	0.027	0.008	0.025	0.075	-0.004	-0.02
2	0.42	0.15	0.033	0.2	0.021	-0.055	0.023	0.061	-0.038	0.068	-0.36	-0.021
3	0.69	-0.064	-0.31	0.082	0.11	-0.03	-0.041	-0.057	0.095	0.05	0.01	-0.056
4	0.75	-4.8	-0.49	0.33	0.017	-0.1	-0.073	0.076	-0.049	0.055	0.3	-0.05
5	1.3	0.81	1.4	0.25	1.7	-0.072	0.008	-0.034	0.008	0.029	0.031	-0.022
6	0.84	-1.8	0.81	-1.9	0.62	-0.021	-0.36	-0.037	-0.12	0.033	0.049	-0.13
7	0.1	-0.54	-0.85	0.42	0.58	-7.6	-0.25	-0.12	0.03	0.009	-0	-0.003
8	-4.4	-0.65	-1.4	1.9	0.93	-0.55	0.79	0.025	0.077	0.011	-0.056	0.003
9	2	-2.7	-0.83	-0.53	0.27	0.67	0.86	-2.9	0.082	-0.035	-0.13	-1.3
10	-0.056	-0.74	-0.16	-1.1	1.2	-0.1	0.65	-0.062	0.17	-1.4	-0.58	0.041
11	-13	-0.75	4	-0.89	1.5	8.8	0.65	1.6	0.69	1.3	-0.25	0.15

Fig 14. ADF/KPSS Result After Differencing

From the GPAC table we can spot ARMA(2, 1) in hourly level; from ACF/PACF plot we can spot ARMA(0, 1) or ARMA(1, 1) in daily level and ARMA(0, 1) in weekly level.

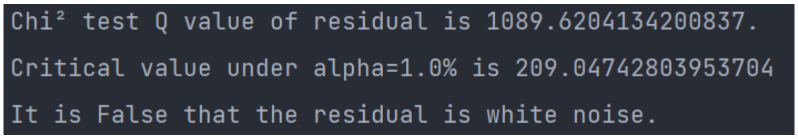
Our tentative model is: ARIMA(2, 1, 1) × ARIMA$(1, 1, 1)_{24}$ × ARIMA$(0, 1, 1)_{168}$

We will fit this multiplicative model with LM algorithm and estimate the parameters.



```
The estimated AR1 is -0.8802818479276165       The C.I. of estimated AR1 is -0.89361 to -0.86695
The estimated AR2 is 0.10197060577930059       The C.I. of estimated AR2 is 0.08868 to 0.11526
The estimated MA1 is -0.9929931142802225       The C.I. of estimated MA1 is -1.00632 to -0.97967
The estimated AR1_L24 is -0.12918680209423375  The C.I. of estimated AR1_L24 is -0.14251 to -0.11586
The estimated MA1_L24 is -0.944226820079695    The C.I. of estimated MA1_L24 is -0.95755 to -0.93090
The estimated MA1_L168 is -0.8900852903720105  The C.I. of estimated MA1_L168 is -0.90341 to -0.87676
```

Fig 15. Parameters Estimated with LM Algorithm



```
Chi² test Q value of residual is 1089.6204134200837.
Critical value under alpha=1.0% is 209.04742803953704
It is False that the residual is white noise.
```

Fig 16. Model Residual χ^2 Test Result

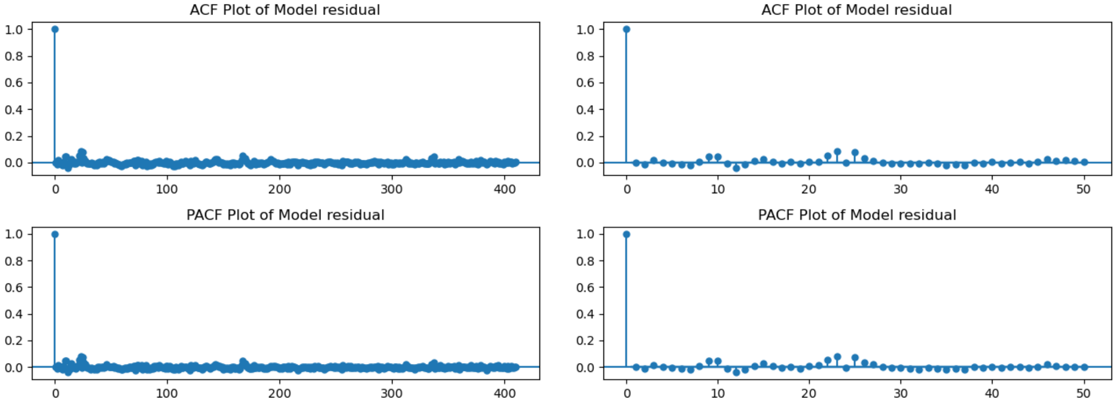
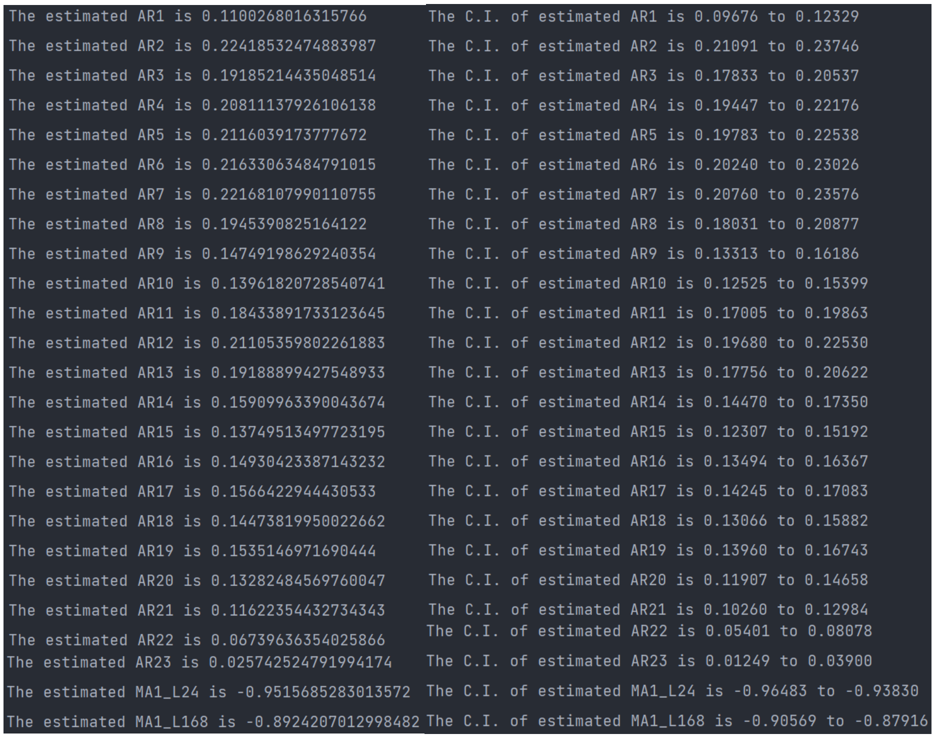


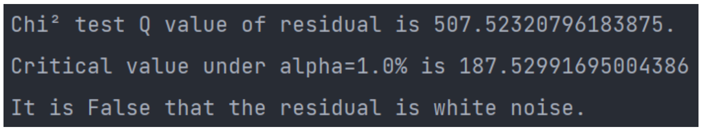
Fig 16. ACF/PACF of Model Residual Using Lag 410 and 50

The model residual is not white, but we are not far away. From the ACF/PACF plot we can see there are hourly structures before lag 24 not yet considered by the model, so we shall increase the non-seasonal AR to order 23 and adjust the model by removing non-significant parameters by then. The results are as follows:



```
The estimated AR1 is 0.1100268016315766        The C.I. of estimated AR1 is 0.09676 to 0.12329
The estimated AR2 is 0.22418532474883987       The C.I. of estimated AR2 is 0.21091 to 0.23746
The estimated AR3 is 0.19185214435048514       The C.I. of estimated AR3 is 0.17833 to 0.20537
The estimated AR4 is 0.20811137926106138       The C.I. of estimated AR4 is 0.19447 to 0.22176
The estimated AR5 is 0.2116039173777672        The C.I. of estimated AR5 is 0.19783 to 0.22538
The estimated AR6 is 0.21633063484791015       The C.I. of estimated AR6 is 0.20240 to 0.23026
The estimated AR7 is 0.22168107990110755       The C.I. of estimated AR7 is 0.20760 to 0.23576
The estimated AR8 is 0.1945390825164122        The C.I. of estimated AR8 is 0.18031 to 0.20877
The estimated AR9 is 0.14749198629240354       The C.I. of estimated AR9 is 0.13313 to 0.16186
The estimated AR10 is 0.13961820728540741      The C.I. of estimated AR10 is 0.12525 to 0.15399
The estimated AR11 is 0.18433891733123645      The C.I. of estimated AR11 is 0.17005 to 0.19863
The estimated AR12 is 0.21105359802261883      The C.I. of estimated AR12 is 0.19680 to 0.22530
The estimated AR13 is 0.19188899427548933      The C.I. of estimated AR13 is 0.17756 to 0.20622
The estimated AR14 is 0.15909963390043674      The C.I. of estimated AR14 is 0.14470 to 0.17350
The estimated AR15 is 0.13749513497723195      The C.I. of estimated AR15 is 0.12307 to 0.15192
The estimated AR16 is 0.14930423387143232      The C.I. of estimated AR16 is 0.13494 to 0.16367
The estimated AR17 is 0.1566422944430533       The C.I. of estimated AR17 is 0.14245 to 0.17083
The estimated AR18 is 0.14473819950022662      The C.I. of estimated AR18 is 0.13066 to 0.15882
The estimated AR19 is 0.1535146971690444       The C.I. of estimated AR19 is 0.13960 to 0.16743
The estimated AR20 is 0.13282484569760047      The C.I. of estimated AR20 is 0.11907 to 0.14658
The estimated AR21 is 0.11622354432734343      The C.I. of estimated AR21 is 0.10260 to 0.12984
The estimated AR22 is 0.06739636354025866      The C.I. of estimated AR22 is 0.05401 to 0.08078
The estimated AR23 is 0.025742524791994174     The C.I. of estimated AR23 is 0.01249 to 0.03900
The estimated MA1_L24 is -0.9515685283013572   The C.I. of estimated MA1_L24 is -0.96483 to -0.93830
The estimated MA1_L168 is -0.8924207012998482  The C.I. of estimated MA1_L168 is -0.90569 to -0.87916
```

Fig 17. Parameters Estimated in Updated Model



```
Chi² test Q value of residual is 507.52320796183875.
Critical value under alpha=1.0% is 187.52991695004386
It is False that the residual is white noise.
```

Fig 18. Updated Model Residual χ^2 Test Result

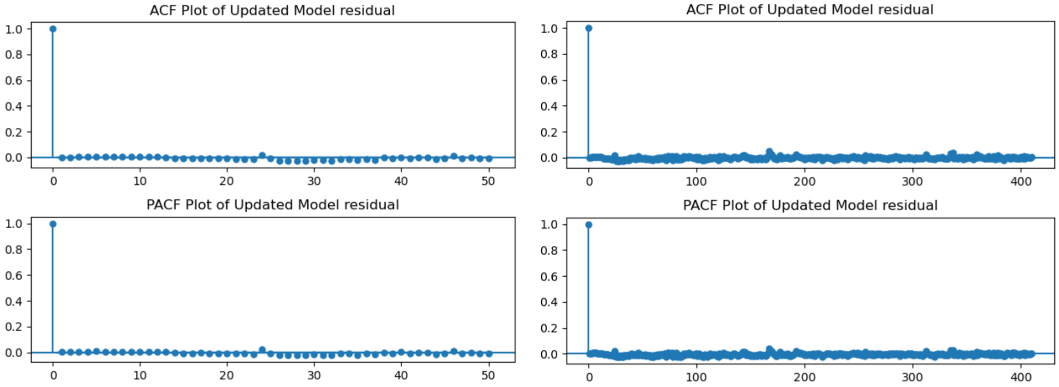


Fig 19. ACF/PACF of Updated Model Residual Using Lag 410 and 50

From the ACF/PACF plot we can see we made the model residual almost white noise. Even though it did not pass the χ^2 test, we made quite a lot improvement compared with our first model.

We will use this model: ARIMA(23, 1, 1) × ARIMA(0, 1, 1)$_{24}$ × ARIMA(0, 1, 1)$_{168}$ for further analysis.

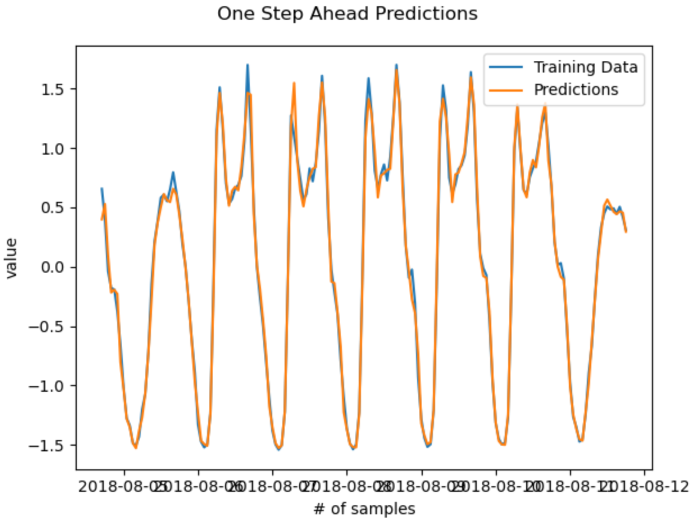


Fig 17. One Step Ahead Prediction Compared with Training Data, Last 170 Samples

One step ahead prediction is derived using estimated parameter in our updated model. Considering the large sample size, we graph only the last 170 observations as illustration. We can see that the model performs well in one step ahead predictions, capturing most of the volatility in the data, while missing certain structures like morning or evening rush hour peak are often underestimated.

Forecasts along with RSME from this model is generated, graphed and displayed as follows:

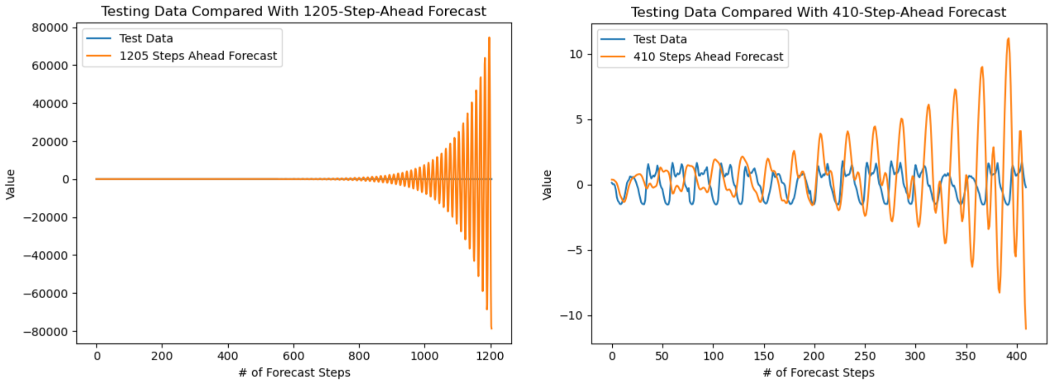


Fig 18(a). SARIMA Updated Model Forecast Results Compared with Test Data

We can see the forecast values explode after 200 steps, and could not give useable result. If we use simpler (the first) model, we have forecast as follows:

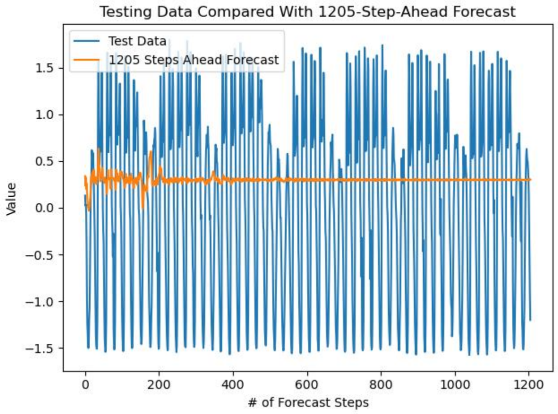


Fig 18(b). SARIMA Simple Model Forecast Results Compared with Test Data

Now we have a stationary forecast that converges towards the mean of test data. RMSE of this forecast is 3898, which could not compete with OLS or Holt-Winters methods.

Summary and Conclusion

Overall, our SARIMA failed to make prediction residual white noise, and failed to beat base models. Our best forecast comes from Holt-Winters model. Multiplicative SARIMA models are computational expensive to fit (especially considering LM algorithm is an efficient algorithm already), our very first model took 4 minutes to finish estimation, while through the process the most complicated model with much more parameters took more than 40 minutes to compute! But we did not walk away without a gain. The LM estimation algorithm we develop could consider more complicated multiplicative model then the current build in statsmodels package could, and the toolbox we built along the process forms a complete pipeline for series simulation, parameter estimation, prediction plotting, residual analysis and forecast generation, which could be transferred to future time series studies.

Appendix

KONG Deyu_6450_FTP.zip