IE5202 Project 2

Dong Zheng
A0119545B

In this project, we would like to predict traffic volume in 2018, given 5 years historical data of hourly traffic volume and weather information.

1: Data Exploration

First of all we plot the given training data set from 2012 to 2017 as Figure below.

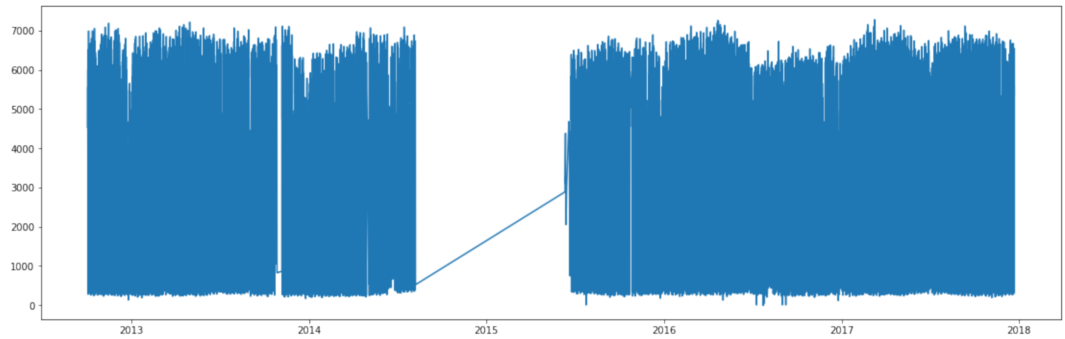


Figure 1.1 Training data from 2012 to 2017

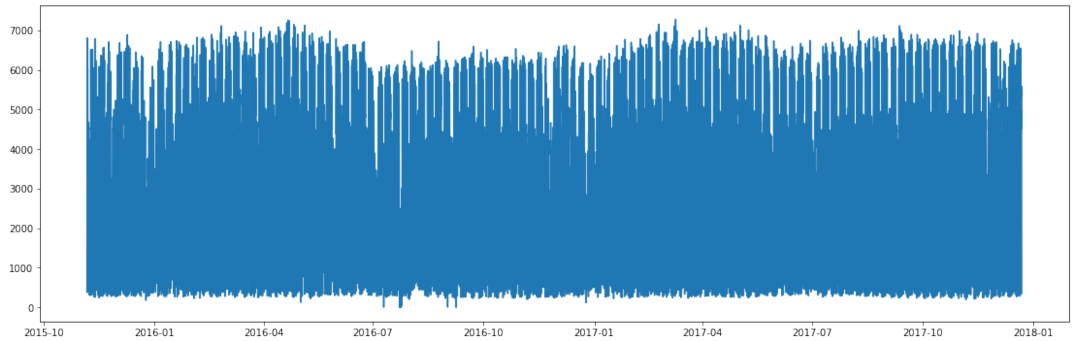
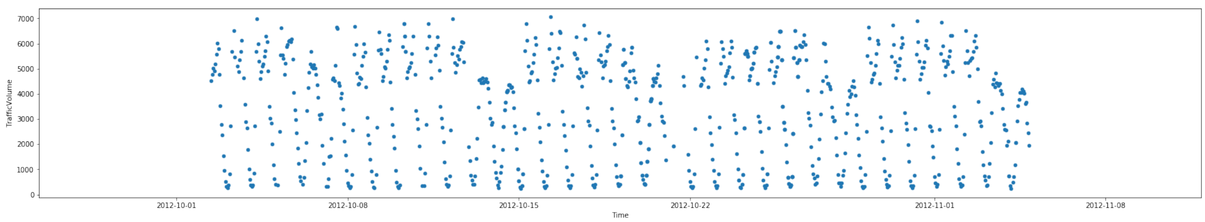


Figure 1.2 Training data from 2015 to 2017

Noticing that there is a big gap of missing data from 2014-08-01 to 2015-08-01 (Figure 1.1), we will first focus on the right part of the training data set (Figure 1.1) because it is nearer to the test data set. Zoom into a smaller time interval, you can observe that there are obvious daily and weekly seasonal components, and the trend movement is trivial.



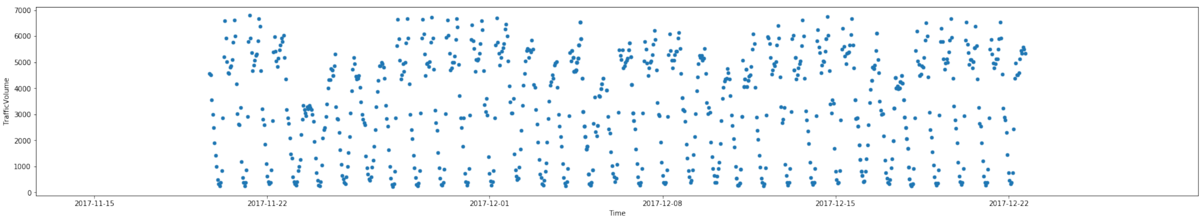


Figure 1.3 Weekly training data in 2012 and 2017

To further investigate seasonality, new columns of [year, month, week, hour] information are generated for each observation in the data. From the following box plots of traffic volumes within each year, month, week and hour. We can see that hourly and weekly seasonal patterns are significant, while year-wise trend and monthly seasonal effects are not significant.

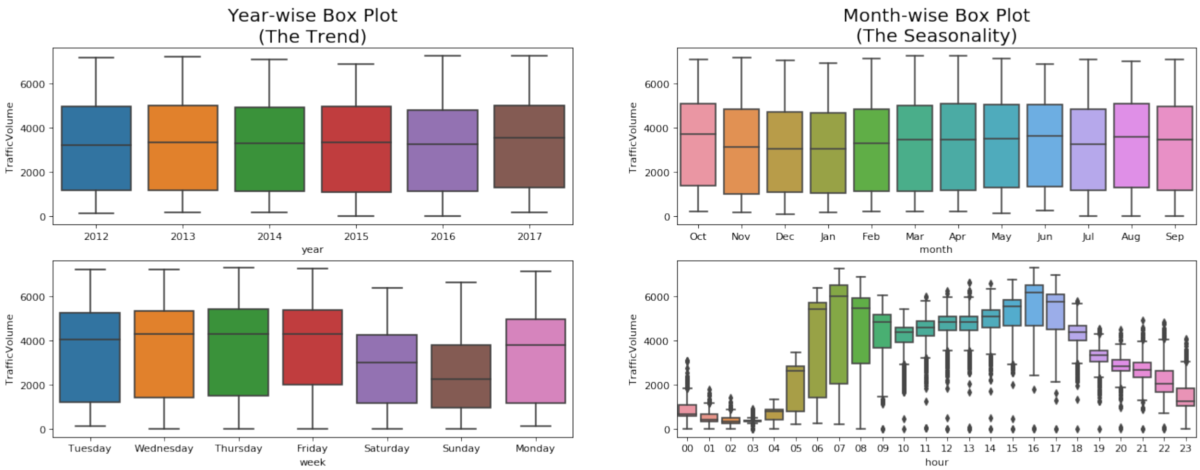


Figure 1.4 box plots of traffic volumes by year, month, week and hour

In the training dataset, some hourly data are missing and some other hourly data are duplicated if weather changes. These ununiform time indexes will cause problem in some time-series models, such as exponential smoothing model. Therefore, we have done some data pre-processing and data cleaning. Index duplicated data are aggregated by their mean, and missing hourly data are interpolated by their preceding and succeeding observations. Moreover, Holiday information is also converted to Boolean data type for later use.

Table 1.5 Missing and duplicated data within an hour

None	278.23	0	0	1	Clear	sky is clear	10/3/2012 6:00	5673
None	278.12	0	0	1	Clear	sky is clear	10/3/2012 8:00	6511
None	282.48	0	0	1	Clear	sky is clear	10/3/2012 9:00	5471
None	291.97	0	0	1	Clear	sky is clear	10/3/2012 12:00	5097
None	281.25	0	0	99	Rain	light rain	10/10/2012 7:00	6793
None	281.25	0	0	99	Drizzle	light intensity drizzle	10/10/2012 7:00	6793
None	280.1	0	0	99	Rain	light rain	10/10/2012 8:00	6283
None	280.1	0	0	99	Drizzle	light intensity drizzle	10/10/2012 8:00	6283
None	279.61	0	0	99	Rain	light rain	10/10/2012 9:00	5680
None	279.61	0	0	99	Drizzle	light intensity drizzle	10/10/2012 9:00	5680

2: Regression on Time

After data exploration, in this section we need to build a regression model by only using 'Time' and the response value 'TrafficVolume'. Two different methods used for model building are seasonal factor approach and trigonometric functions approach.

2.1 Seasonal Factor Model

Recall from previous section that the data exhibits seasonal variation on daily and weekly basis. Therefore, a regression model of the following form can be used:

$$yt = TRt + SNt + \varepsilon t;$$

The linear trend part (TRt) can be expressed as $TR_t = \beta_0 + \beta_1 t$, where t is the time delta to the time zero, corresponding to 'T_difference' column of the training dataframe. The seasonality part (SNt) has three components: hour, week and month. Then, we can plug in the following formula to train the OLS regression model:

TrafficVolume ~T_difference+C(hour)+C(week)+C(month)

The regression result is in Appendix 1 and shows that model has a R-squared of 0.838, which indicates that 83.8% of the variability of the response variable can be explained by the seasonal factor model.

2.2 Trigonometric Function Model

Another common way to model seasonal time series data is to use the trigonometric functions. Here we use collections of periodic functions with 4 different types of frequencies. Hence Sin1, Cos1, Sin2, Cos2, Sin3, Cos3, Sin4, Cos4 values are calculated and stored in the dataframe. We use period (L) of 24 hours and fit above trigonometric values into the model and estimated coefficients are as below:

Table 2.2.1 Coefficients for trigonometric function model

	coef	std err	t	P>|t|	[0.025	0.975]
Sin1	-699.6669	8.522	-82.105	0.000	-716.370	-682.964
Cos1	-2133.0414	8.521	-250.323	0.000	-2149.744	-2116.339
Sin2	-328.8191	8.521	-38.589	0.000	-345.521	-312.117
Cos2	-638.6561	8.522	-74.945	0.000	-655.359	-621.953
Sin3	-347.2996	8.521	-40.756	0.000	-364.002	-330.597

Cos3	468.4080	8.521	54.969	0.000	451.705	485.111
Sin4	-11.7784	8.521	-1.382	0.167	-28.481	4.924
Cos4	153.8245	8.521	18.051	0.000	137.122	170.527

We can see that most of the trigonometric terms are significant as their p-values are small. The R-squared and Adj. R-squared of the trigonometric model are 0.823 (Appendix 2), which is similar to 0.838 of the seasonal factor model. We can expect trigonometric model's R-square is less than that of seasonal factor model, because trigonometric model has less parameters in the model. Here we achieved similar R-square value with less by using Trigonometric Function Model here.

2.3 Model Diagnostic

After we trained our Regression on Time model, we need to diagnose the models and check if the model assumptions are satisfied. Regression assumptions made for error terms are 1. Zero mean 2. Constant variance 3. No correlation with X 4. No autocorrelation 5. Normally distributed. Here we will use residuals from seasonal factor model to do the diagnosis.

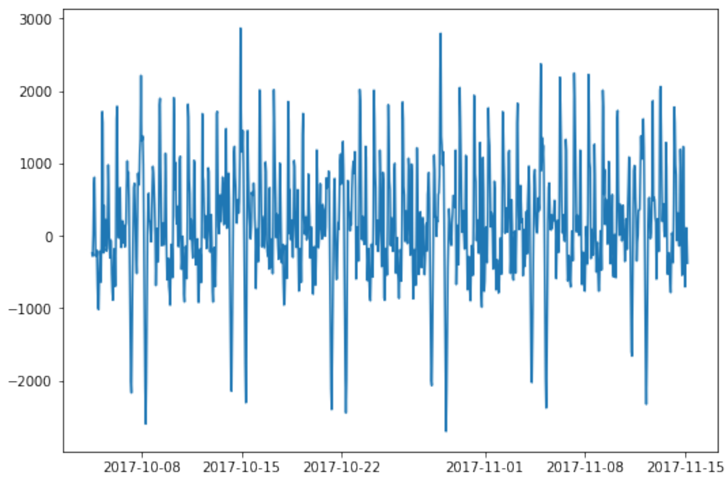


Figure 2.3.1 Residual plot for seasonal factor model

First, we plot the residuals of the training data set as below to check assumption 1, 2 & 4. From the residual plot, we can see that the mean of the error term is around zero and variance are in general constant. However, we can observe that there is an obvious seasonal pattern in the residual plot. Thus, assumption 4 is violated. We need to adopt other method in the later section to tackle autocorrelation property of the residuals.

Secondly, to check the normality of the residuals, quantile-quantile (QQ) plot and histogram of normalized residuals are displayed as below:

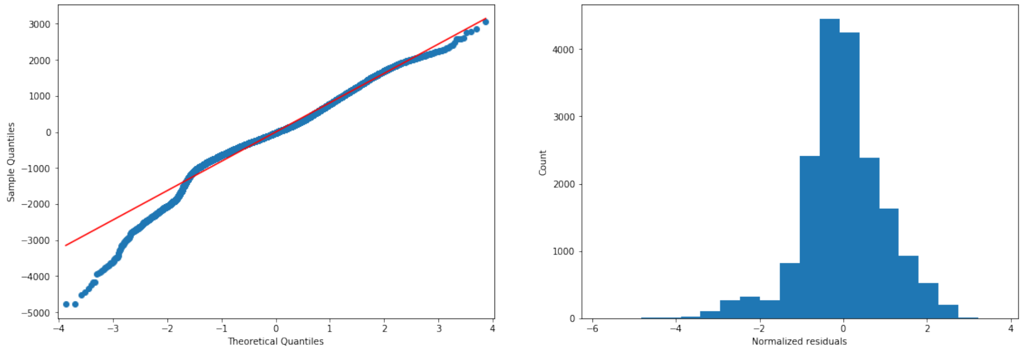


Figure 2.3.2 Histogram and QQ plot for normalized residuals

From the right side of the histogram and QQ plot, we can see that residuals are quite normally distributed. However, on the left extreme, the residuals deviate far from normal distribution. It may be due to the non-negative constrain of traffic volume, while our regression model may predict negative result for extreme low traffic volume.

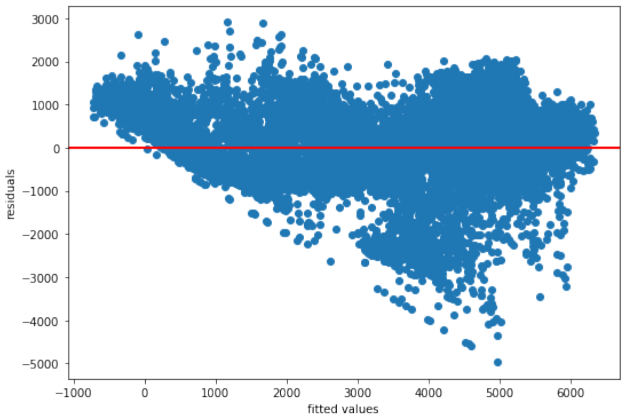


Figure 2.3.3 Residuals versus fitted value

The residuals against fitted value plot shows that when the fitted values are less than zero, residuals are always positive, and as the fitted value goes larger, residuals are more likely to become negative. It implies some correlations between fitted values and residuals. Therefore, we need to improve the current inadequate model or find a better approach in the later part.

3: Exponential smoothing

In this section, we will explore different kinds of exponential smoothing models, which includes simple exponential smoothing, double exponential smoothing and Holt-Winters method. We will use training dataset to select the best model and then use the selected model to make predictions.

Firstly, a simple exponential smoothing model can be expressed as $Yt = \beta 0 + \epsilon t$. Since it only assumes a constant trend, it is not able to catch the trend and seasonal information as shown in the results obtained from statsmodels is shown as below:

Table 3.1 Results for simple exponential smoothing

Dep. Variable:	TrafficVolume	**No. Observations:**	1000
Model:	SimpleExpSmoothing	**SSE**	760388237.529
Optimized:	True	**AIC**	13545.584
Trend:	None	**BIC**	13555.400
Seasonal:	None	**AICC**	13545.625

The optimal alpha obtained is 0.995, which implies that the forecasting value will be highly relevant to the latest observation only. It will be hard to predict data with seasonality.

Secondly, we fit training data in to a double exponential smoothing model: $Yt = \beta 0 + \beta 1t + \epsilon t$, the Holt's Trend Modelling results are as below:

Table 3.2 Results for double exponential smoothing

Dep. Variable:	TrafficVolume	**No. Observations:**	1000
Model:	Holt	**SSE**	637642390.091
Optimized:	True	**AIC**	13373.533
Trend:	Additive	**BIC**	13393.164
Seasonal:	None	**AICC**	13373.617

Again, the level smoothing parameter ($\beta 0$) and growth rate smoothing parameter ($\beta 1$) obtained are both 0.995. It means that both level and trend are affected drastically by the latest observation. The Holt's Trend model can capture the trend information but is not able to capture the seasonality. Therefore it is not suitable for this data set.

Lastly, we have Holt-Winters Method of the form: $Yt = \beta 0 + \beta 1t + SNt + \epsilon t$. Besides three smoothing parameters $\beta 0$, $\beta 1$, SNt, we also need to decide the parameter seasonal_period (L) manually. Here we will try L = 24 or 24*7, which corresponding to number of hours in a day and week respectively. When L is 24, the level, trend and seasonal smoothing coefficient obtained are 0.95, 0.0001, and 0.04 respectively. This implies the tanning data has some seasonal component while trend component is not obvious. While seasonal periods is 24*7, these three coefficients are 0.3975, 0.0025796, 0.0039728. We need to compare their forecasting power in the validation data set as below.

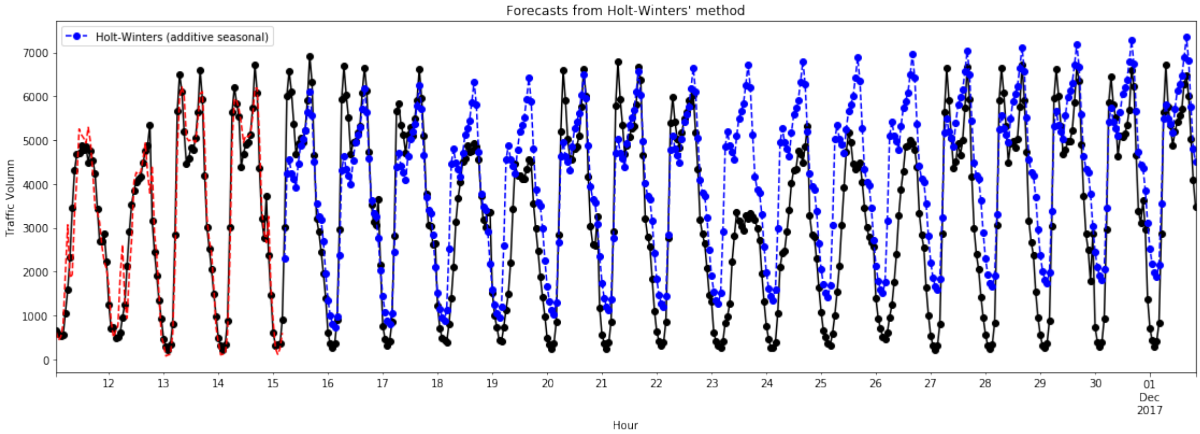


Figure 3.1 Predicted vs. True value for Holt-Winters method (L=24)

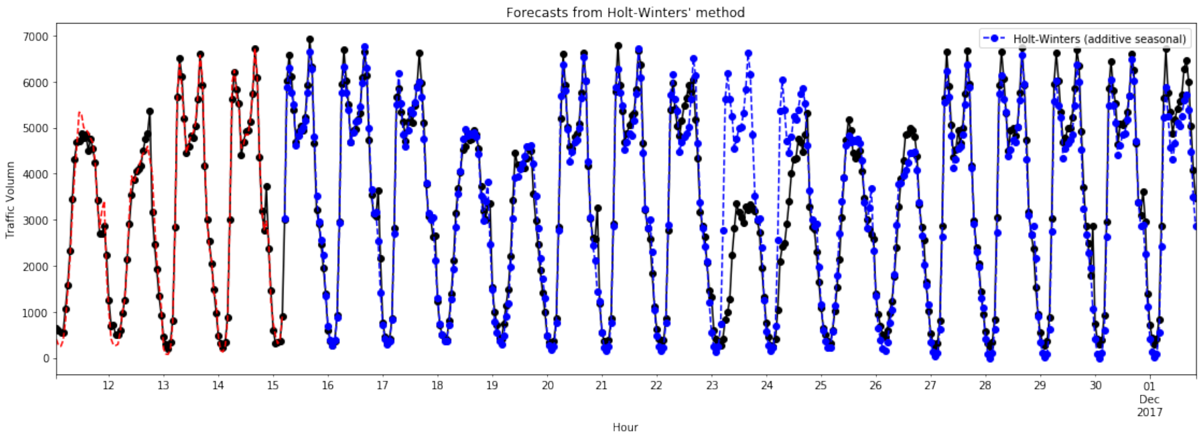


*Figure 3.2 Predicted vs. True value for Holt-Winters method (L=24*7)*

The blue line displays the forecasted values from Holt-Winters' model, and the black line represents true values. When seasonal_periods is 24, we observe from Figure 3.2 that the forecasted value start to deviate from the true value after a few weeks. From Figure 3.3 where seasonal_periods is 24*7, we can see forecasted results are close to true observations, except 25th Dec, which is a public holiday.

Therefore, we choose Holt-Winters Model with seasonal period equals to 24*7 as our final exponential smoothing model. From table 3.2 below, we can see that Holt-Winters Model (L=24*7) has the lowest SSE and AIC.

Table 3.2 Comparing model SSE and AIC for all exponential smoothing models

	Simple Exponential smoothing	Double Exponential Smoothing	Holt-Winters 24	Holt-Winters 24*7

SSE	7.60388237e8	6.37642390e8	2.760015e8	4.6020454e7
AIC	13545	13373	12584	11080

*Table 3.3 Results for Holt-Winters method (L=24*7)*

Dep. Variable:	TrafficVolume	**No. Observations:**	1000
Model:	ExponentialSmoothing	**SSE**	46020454.315
Optimized:	True	**AIC**	11080.841
Trend:	Additive	**BIC**	11924.975
Seasonal:	Additive	**AICC**	11154.659

In summary, exponential smoothing models are good at explaining historical data set, but its prediction power for unknown future are quite limited. It is also unable to consider special case such as holidays. Therefore, we will explore more advance models in the next section.

4: Free form forecasting

In this section, we will use a combination of multivariable regression model and SARIMA model to make predictions of the missing values in test data.

4.1 Methodology

Observed from the model diagnostic part in Step 1, residuals display some seasonal patterns. Therefore, ARIMA could be a good choice to model the residuals. Since other information in training data such as weather and holidays date could also possibly influence traffic volume, we would like to incorporate them into our multivariable regression model. The final multivariable regression model picked by best subset selection is:

TrafficVolume~C(hour)+C(week)+C(month)+WeatherMain+Temp+CloudsAll+year+IsHoliday

4.2 Time Series Stationarity

The result for above regression model (Appendix 4) shows that the model's adjusted-R square value is 0.841, which is slightly higher than regression on time model in Section 1. The trained model is then applied to the testing data set to get the initial predicted values. Afterwards, the training residuals from this multivariable regression model are obtained and plotted as below:

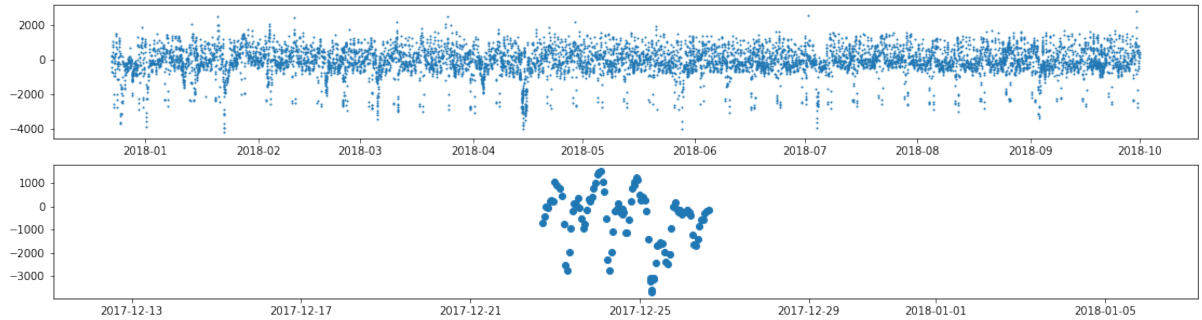


Figure 4.1 Original residuals for stationarity check

The residual data is obviously non-stationary, as its variance changes over time. A first order non-seasonal differencing is then applied to the residual data.

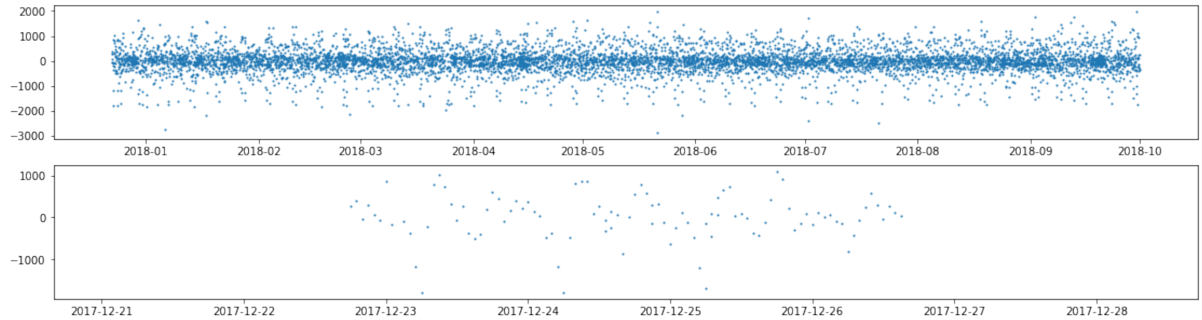


Figure 4.2 Residuals after first order non-seasonal differencing

4.3 ACF, PACF and Parameter Tuning

Now data appears more stationary after one non-seasonal differencing. Next, ACF and PACF of the differentiated residual are plotted to identify parameter p and q in ARIMA model.

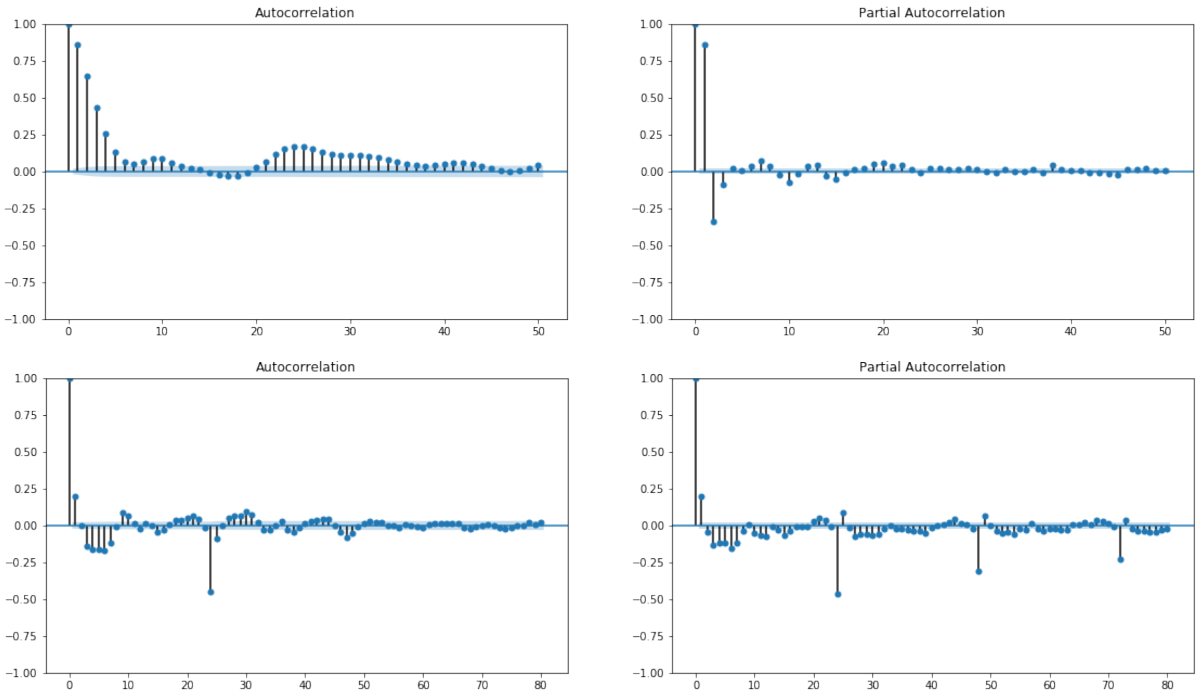


Figure 4.3: ACF and PACF after first-order regular differencing

Figure 4.4: ACF and PACF after first-order seasonal differencing

Observed from the non-seasonal autocorrelation plot (ACF) and Partial Auto-Correlation Function (PACF), ACP dies down and PACF cut off at 3, suggesting a MA(3) model. For seasonal peaks, PACF dies down and ACF cuts off after one peaks at 24, suggesting an seasonal AR(2) model. In order to validate our observation and find best ARIMA Model hyperparameters. We grid search the optimal p, q combination (Table 4.1).

Table 4.1 Grid Search Results for p,q

AIC	p	q
117479.33	1	0
117479.449	1	1
116695.07	1	2
116442.888	1	3
117469.663	2	0
116427.549	2	1
116368.428	2	2
116379.013	2	3
117305.486	3	0
116360.709	3	1

(P, D, Q, L)	AIC
(0, 0, 1, 24)	115906.4
(0, 0, 2, 24)	115490.2
(0, 0, 3, 24)	115169.7
(0, 1, 1, 24)	116601.8
(0, 1, 2, 24)	116803.3
(0, 1, 3, 24)	116443.8
(1, 0, 1, 24)	115901.5
(1, 0, 2, 24)	115489.2
(1, 0, 3, 24)	115151.9
(1, 1, 1, 24)	115685.5
(1, 1, 2, 24)	116719.1
(1, 1, 3, 24)	116447.4
(2, 0, 1, 24)	115551.6
(2, 0, 2, 24)	115539.4
(2, 0, 3, 24)	115171.1
(2, 1, 1, 24)	116786.7
(2, 1, 2, 24)	115842.1
(2, 1, 3, 24)	116442

The grid search result indicates the optimal (p,d,q) is (3,1,1). This combination has the lowest AIC value of 116360. This result is consistent with our previous ACF and PACF plots. For the seasonal hyper-parameters, the gird search result shows a combination of (0,0,3,24) has the lowest AIC value. Therefore, a SARIMA(3,1,1)(0,0,3)24 is built on training data set and then applied on test data set for prediction.

Table 4.2 Results for SARIMA(3,1,1)x(0,0,3)24 Model

Dep. Variable:	y	**No. Observations:**	7870
Model:	SARIMAX(3, 1, 1)x(0, 1, [1, 2, 3], 24)	**Log Likelihood**	-58149.326
Date:	Wed, 10 Nov 2021	**AIC**	116314.653
Time:	23:43:25	**BIC**	116370.394
Sample:	0	**HQIC**	116333.751
	- 7870		
Covariance Type:	opg		

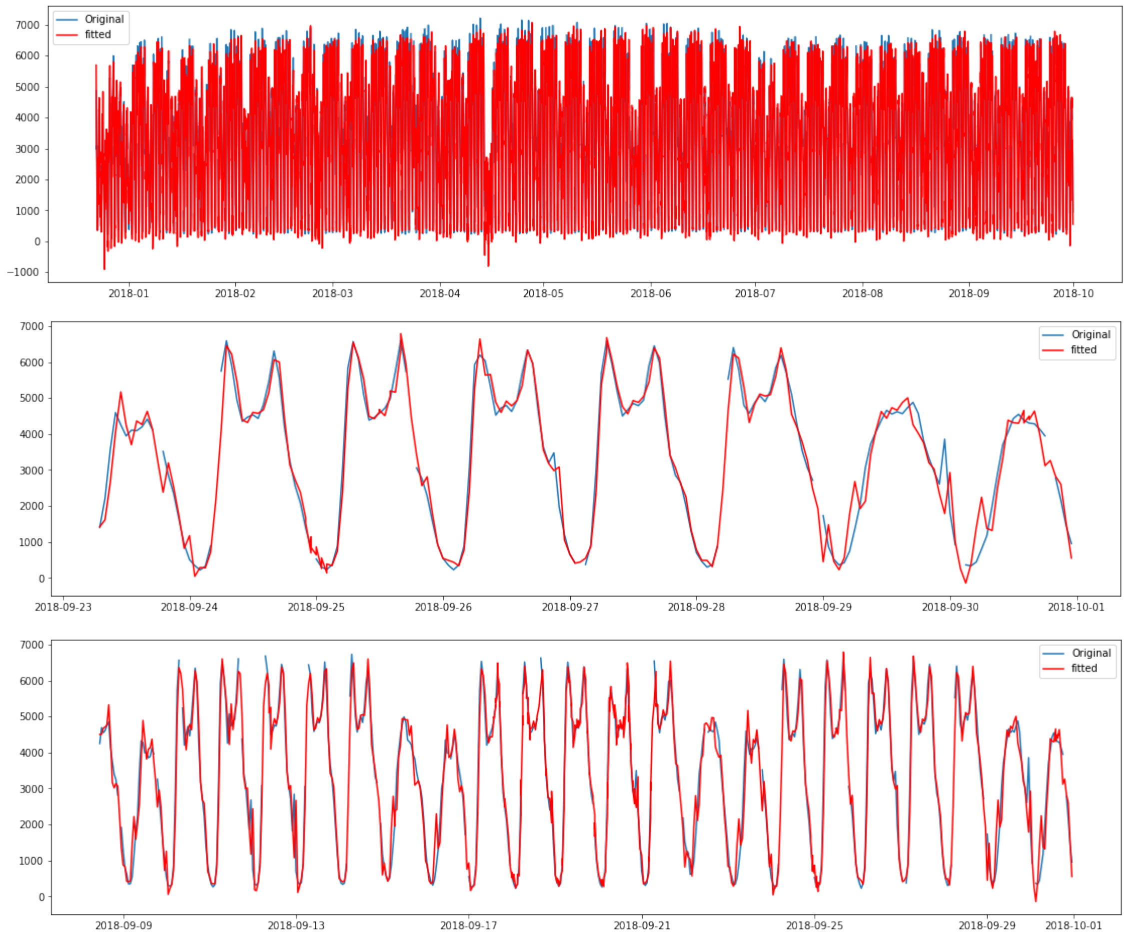
	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.1152	0.007	150.275	0.000	1.101	1.130
ar.L2	-0.2327	0.013	-18.468	0.000	-0.257	-0.208
ar.L3	-0.0920	0.010	-9.017	0.000	-0.112	-0.072
ma.L1	-1.0000	0.099	-10.082	0.000	-1.194	-0.806
ma.S.L24	-0.9486	0.101	-9.403	0.000	-1.146	-0.751
ma.S.L48	-0.0532	0.016	-3.351	0.001	-0.084	-0.022
ma.S.L72	0.0018	0.011	0.160	0.873	-0.020	0.024
sigma2	1.568e+05	6.45e-07	2.43e+11	0.000	1.57e+05	1.57e+05

Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	4423.15
Prob(Q):	0.91	**Prob(JB):**	0.00
Heteroskedasticity (H):	0.85	**Skew:**	-0.48
Prob(H) (two-sided):	0.00	**Kurtosis:**	6.55

4.4 Model Forecasting

We use One-step out-of-sample forecast iteratively to forecast the residuals of missing value in the test data set, which means that only one prediction is made once at a time, and only data before the prediction time are used. After each prediction, the forecasted results are recorded in the dataframe and SARIMA model is also updated for next prediction. One-step forecasting usually performs better than multi-step out-of-sample forecasting. It also allow us making full use of test data before the prediction time point to train the model. The final predicted traffic volume is a sum of predicted values from multivariable regression model and predicted residuals from SARIMA model, plotted as below:

Figure 4.5: Fitted Value vs. True Values in Test Data



4.5 Model Diagnostics

After obtaining the final prediction values, the diagnostic plots of the combined model in test data set are draw below:

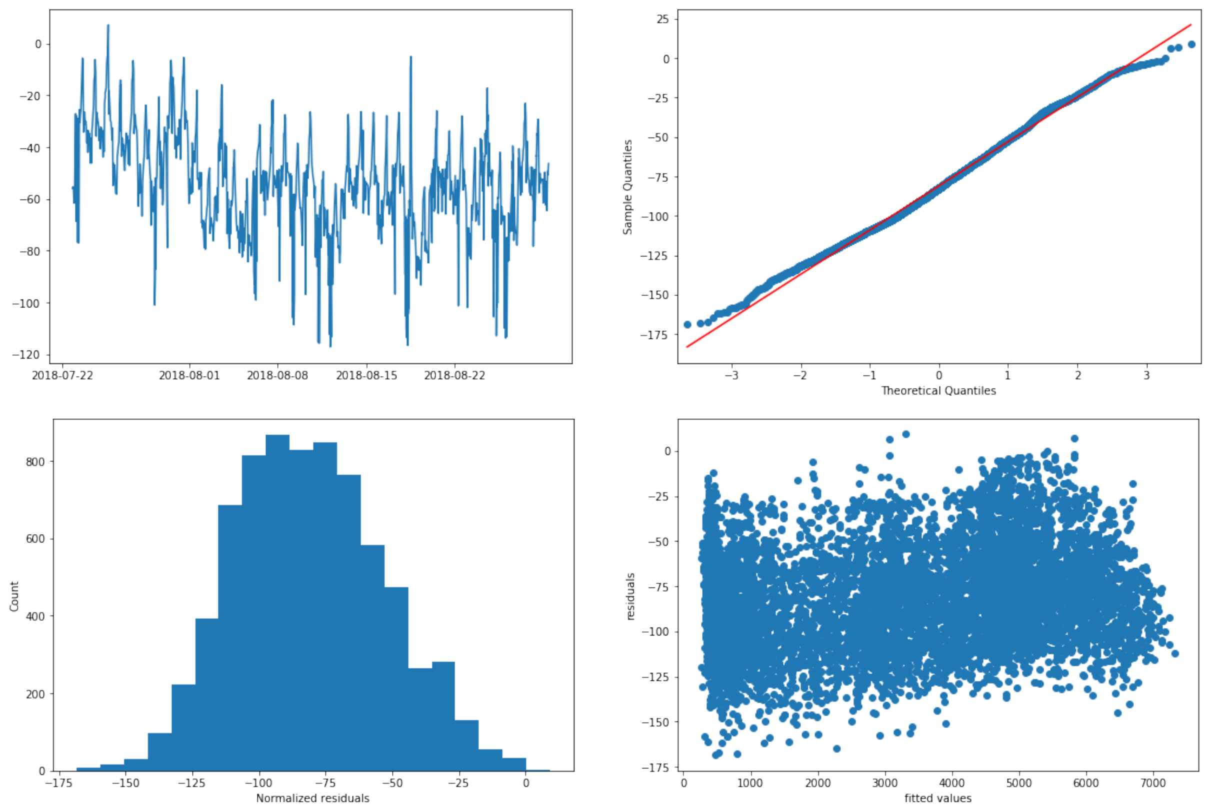


Figure 4.5: Diagnostic plots for final result's residuals

From QQ plot and histogram of the residuals plots, we can see that residuals are quite normally distributed. However, it can be observed from 1st graphs that the mean value of the residuals is -75 (less than zero), which indicates that our model is very likely to overestimate the traffic volumes. Also observed from the residuals vs fitted values plot, when the fitted values are big, our model could predict more accurately. One possible explanation is some unknown factors could make prediction for low traffic volume less accurate.

4.6 Future Improvement

There are also some potential ways that could further optimize our multivariable regression model and SARIMA model. Firstly we can decomposite time series into different components (level, trend, seasonal) and then model them separately. Secondly, employing Augmented Dickey-Fuller (ADF) test after visual checking of stationarity. Fourthly, considering interaction effects between different variables when building regression model. Lastly, tuning seasonal hyperparameters P, D, Q and even L values systematically could further optimizes model performance in test dataset.

Appendix

Appendix 1. Regression results for Seasonal Factor Model

Dep. Variable:	TrafficVolume	**R-squared:**	0.838
Model:	OLS	**Adj. R-squared:**	0.838
Method:	Least Squares	**F-statistic:**	2240.
Date:	Sun, 07 Nov 2021	**Prob (F-statistic):**	0.00
Time:	12:18:58	**Log-Likelihood:**	-1.5097e+05
No. Observations:	18664	**AIC:**	3.020e+05
Df Residuals:	18620	**BIC:**	3.024e+05

Appendix 2. Regression results for Trigonometric Factor Model

Dep. Variable:	TrafficVolume	**R-squared:**	0.823
Model:	OLS	**Adj. R-squared:**	0.823
Method:	Least Squares	**F-statistic:**	3464.
Date:	Sat, 13 Nov 2021	**Prob (F-statistic):**	0.00
Time:	12:56:04	**Log-Likelihood:**	-1.5181e+05
No. Observations:	18664	**AIC:**	3.037e+05
Df Residuals:	18638	**BIC:**	3.039e+05

Appendix 3. Results for triple exponential smoothing (L=24hrs)

Dep. Variable:	TrafficVolume	**No. Observations:**	1000
Model:	ExponentialSmoothing	**SSE**	276001574.456
Optimized:	True	**AIC**	12584.162
Trend:	Additive	**BIC**	12721.579
Seasonal:	Additive	**AICC**	12586.081

Appendix 4. Regression results for Multivariable Regression Model

Dep. Variable:	TrafficVolume	**R-squared:**	0.841
Model:	OLS	**Adj. R-squared:**	0.840
Method:	Least Squares	**F-statistic:**	1636.
Date:	Wed, 10 Nov 2021	**Prob (F-statistic):**	0.00
Time:	18:40:46	**Log-Likelihood:**	-1.3318e+05
No. Observations:	16471	**AIC:**	2.665e+05