Metro Interstate Traffic Volume

STAT 429 Final Project Report

Date: 4 May 2022

Team Members:

Harsh Baberwal
Mayank Agarwal
Yash Kalyani

Abstract

Using the Metro Interstate traffic volume data, we are trying to analyse the traffic volume based on a number of factors like temperature, holidays, rainfall, etc. This report includes comparison of different methods. Methods used include Multiple Linear Regression, SARIMA models, Neural Network and Spectral Analysis. For each method, diagnostics have been performed and suitable transformations have been done. We found that the data is weekly seasonal. There is weak yearly seasonality. The significant factors affecting traffic volume were found to be holidays, temperature, rainfall, weather, and date. Neural network autoregression model is used to forecast traffic volume for the next 30 days. Spectral analysis on our time series gives us 405 days and 6.98 days as the most significant periods. These periods coincide with our findings of weekly seasonality and a weak yearly seasonality.

Contents

Chapter 1

Introduction

1.1 Objective

In this project, we are analyzing the metro interstate traffic volume data.

Using this dataset, we are trying to predict the traffic volume based on a number of factors. Since the same dataset will be used for Part A, B, and C, We will compare different models and test their performance.

In Part A, we perform Exploratory Data Analysis to understand the predictors and how they affect the traffic volume. Next we will use Linear Regression to create a prediction model. The model will be based on Multiple Linear Regression. While implementing the prediction mode, we perform several methods for variable selection and model diagnostics. Future points will then be forecasted using the selected model.

In Part B, we compare different SARIMA models based on the ACF/PACF. Using AIC and BIC as model selection criteria, a suitable model will be selected and residual analysis will be conducted. Future points will then be forecasted using the selected model.

Finally, in Part C, Spectral Analysis and Neural Networks will be used to forecast future points.

1.2 Data Description

The dataset can be found on the UCI machine learning repository[1] and was posted on the website on 7 June, 2019. It contains Hourly Interstate 94 Westbound traffic volume for MN DoT ATR station 301, roughly midway between Minneapolis and St Paul, MN. Hourly weather features and holidays are also included for impacts on traffic volume.

It was previously used in a Talk on anomaly detection[2].

The original dataset consists of 48204 rows and 9 columns. For this project, we choose data after 2014 and filter data corresponding to 20:00 hour of each day. This enables us to be consistent with our analysis. We have also removed a few outliers which are present in the data. We have also added a few columns which represent the dummy variables for weather. The predictors were chosen on basis of possible correlations with the dependent variable, i.e., traffic volume data. These predictors such as rain, snow, temperature, etc. were suspected to affect the metro state traffic volume. Along with these, a dummy variable for indicating holidays is included along with date as a predictor.

Hence, the final dataset contains 1108 rows and 18 columns. The columns in the dataset are:

- **Traffic volume** = This is westbound traffic volume on I-94 roughly midway between Minneapolis and St Paul, MN
- **Date** = Date
- **holiday** = This includes all US National Holidays (including regional holiday like Minnesota State Fair)
- **temp** = Average Temperature in Kelvin
- **rain_1h** = Amount of rain(mm) that occurred in the hour
- **snow_1h** = Amount of snow(mm) that occurred in the hour
- **clouds_all** = Percentage of cloud cover
- **weather_main_Clear** = if weather is clear (1) or not (0)
- **weather_main_Clear** = if weather is clear (1) or not (0)
- **weather_main_Clouds** = if clouds are present (1) or not (0)
- **weather_main_Drizzle** = if it is drizzling (1) or not (0)
- **weather_main_Fog** = if there is Fog (1) or not (0)
- **weather_main_Haze** = if there is Haze (1) or not (0)
- **weather_main_Mist** = if there is Mist (1) or not (0)
- **weather_main_Rain** = if it is Raining (1) or not (0)
- **weather_main_Smoke** = if there is Smoke (1) or not (0)
- **weather_main_Snow** = if it is snowing (1) or not (0)
- **weather_main_Thunderstorm** = if there is s Thunderstorm (1) or not (0)

Chapter 2

Exploratory Data Analysis

We perform preliminary analysis on our data, to analyze our data to find its main characteristics. In our analysis we are trying to find a time series model to predict the traffic volume based on the predictors.

On initial exploration of our data, we found outliers in the data. We removed those outliers as they could affect our analysis. The categorical variables like weather and holiday were converted into dummy variables. We also found missing values, so they were imputed using Kalman Smoothing.

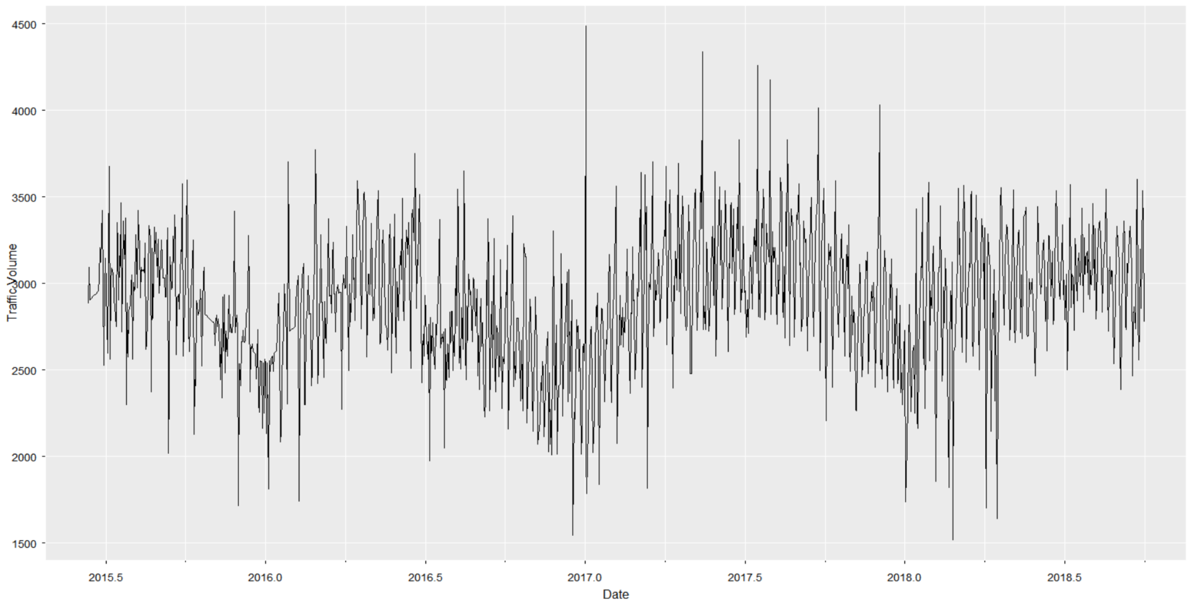
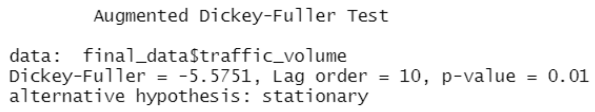


Figure 2.1: Time Series Plot

We plot a time series plot with our target variable (Traffic Volume) (Figure 2.1). The time series plot shows that there is no significant trend in our data. Also by looking at our time series plot we can say that variance is nearly the same. The variance of our data is neither increasing nor decreasing.

On average there are 2906 vehicles during 20:00 hours each day where our data is collected. The minimum number of vehicles present during that hour is 1520 and at the peak hour traffic there are 4490 vehicles present. The median traffic volume during the hour is 2924.

We can observe certain pattern in our data from the time series plot. At the start and end of a year the traffic volume decreases. Almost at the middle of the year the volume of traffic seems to be the highest. This pattern is observed year over year.



```
        Augmented Dickey-Fuller Test

data:  final_data$traffic_volume
Dickey-Fuller = -5.5751, Lag order = 10, p-value = 0.01
alternative hypothesis: stationary
```

Figure 2.2: Dickey-Fuller Test

We perform Augmented Dicky Fuller Test (ADF) to test whether the given time series is stationary or not. This is a unit root test. The Null hypothesis assumes the presence of a unit root. We get a p-value < 0.05. Thus, our series is stationary.

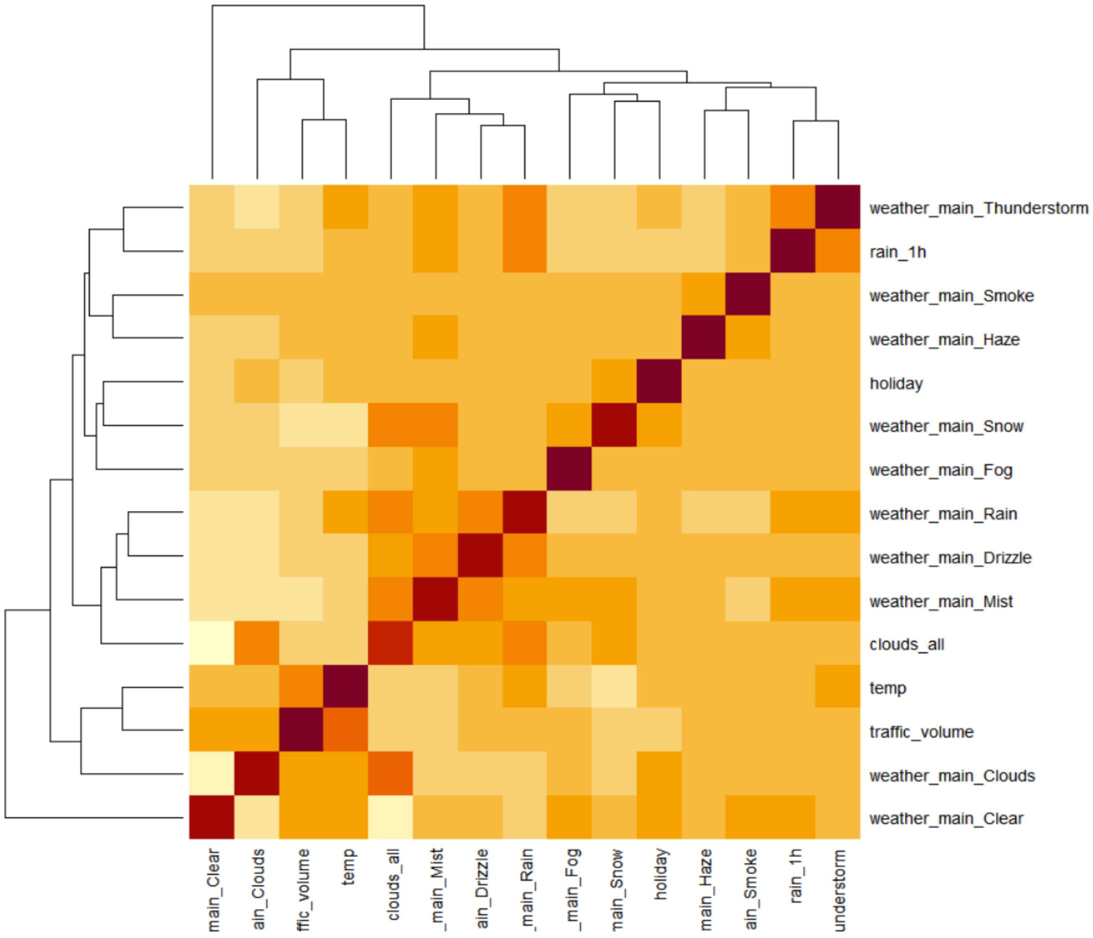


Figure 2.3: Correlation Heatmap

From the heatmap (Figure 2.3) we observe that the only variable showing the most correlation with traffic volume is temperature predictor. We also see that Snow and temperature are negatively correlated. Clouds and Weather_main_clear are also negatively correlated, as the presence of clouds will affect how clear the weather is.

Chapter 3

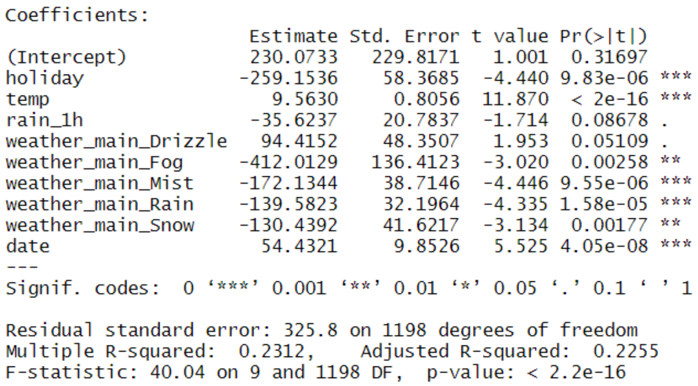
Methods

3.1 Regression Analysis

In Part A, regression analysis is conducted. We will be implementing Multiple Linear Regression. It is represented as -

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + e_i$,where y is our response variable, and β are the coefficients of x.

We use Backward elimination as model selection procedure and AIC as our selection criteria. Upon performing Backward elimination, we are left with the following predictors: holiday, temp, rain_1h, weather_main_Drizzle, weather_main_Fog, weather_main_Mist, weather_main_Rain, weather_main_Snow, and date.



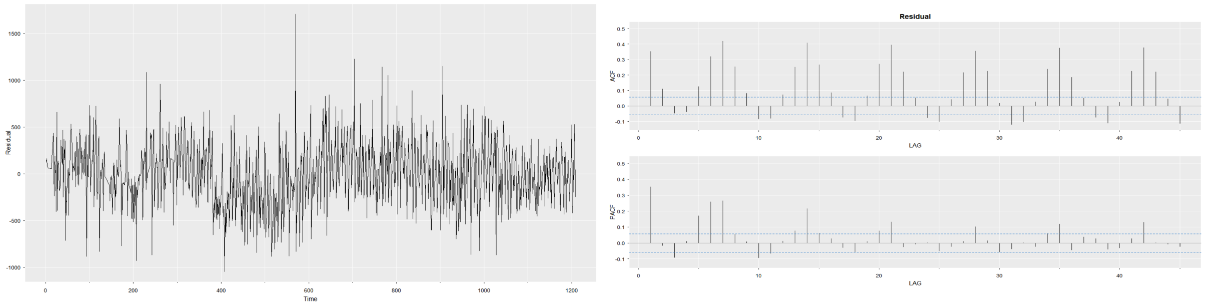
```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            230.0733   229.8171   1.001  0.31697
holiday               -259.1536    58.3685  -4.440 9.83e-06 ***
temp                     9.5630     0.8056  11.870  < 2e-16 ***
rain_1h                -35.6237    20.7837  -1.714  0.08678 .
weather_main_Drizzle    94.4152    48.3507   1.953  0.05109 .
weather_main_Fog      -412.0129   136.4123  -3.020  0.00258 **
weather_main_Mist     -172.1344    38.7146  -4.446 9.55e-06 ***
weather_main_Rain     -139.5823    32.1964  -4.335 1.58e-05 ***
weather_main_Snow     -130.4392    41.6217  -3.134  0.00177 **
date                    54.4321     9.8526   5.525 4.05e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325.8 on 1198 degrees of freedom
Multiple R-squared:  0.2312,    Adjusted R-squared:  0.2255
F-statistic: 40.04 on 9 and 1198 DF,  p-value: < 2.2e-16
```

Figure 3.1: Regression Model

We achieve R-squared = 0.2312, AIC = 17420.06, and BIC = 17476.12.

We then perform residual analysis. From the time series plot of residuals(Figure 3.2a), we notice possible seasonality patterns, but no trend or variance increase. So we check the ACF/PACF plots (Figure 3.2b). From the ACF plot, it is clear that there is seasonal persistence. The seasonal lag is 7, which shows that we have weekly seasonality.

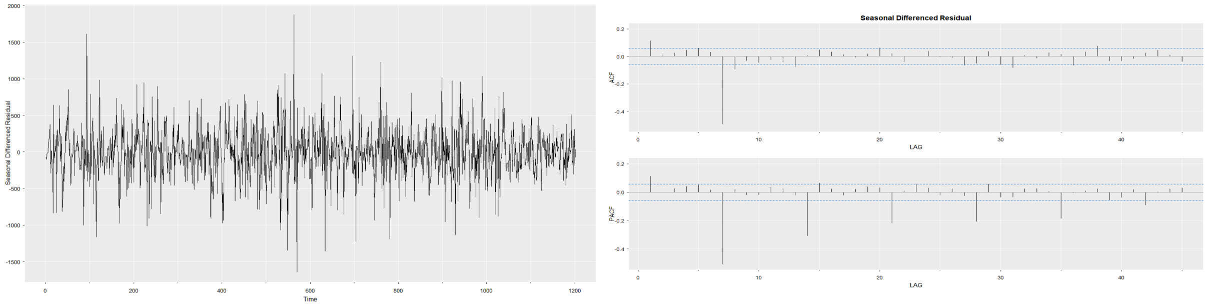
So we perform seasonal differencing. After seasonal differencing, we plot the time series again and look at its ACF/PACF (Figure 3.3b). We see that for the seasonal part, ACF



(a) Residual Plot

(b) ACF/PACF of Residuals

Figure 3.2: Residual Plot and ACF/PACF



(a) Seasonal Differenced Residual Plot

(b) ACF/PACF of Seasonal Differenced Residuals

Figure 3.3: Seasonal Differenced Residual Plot and ACF/PACF

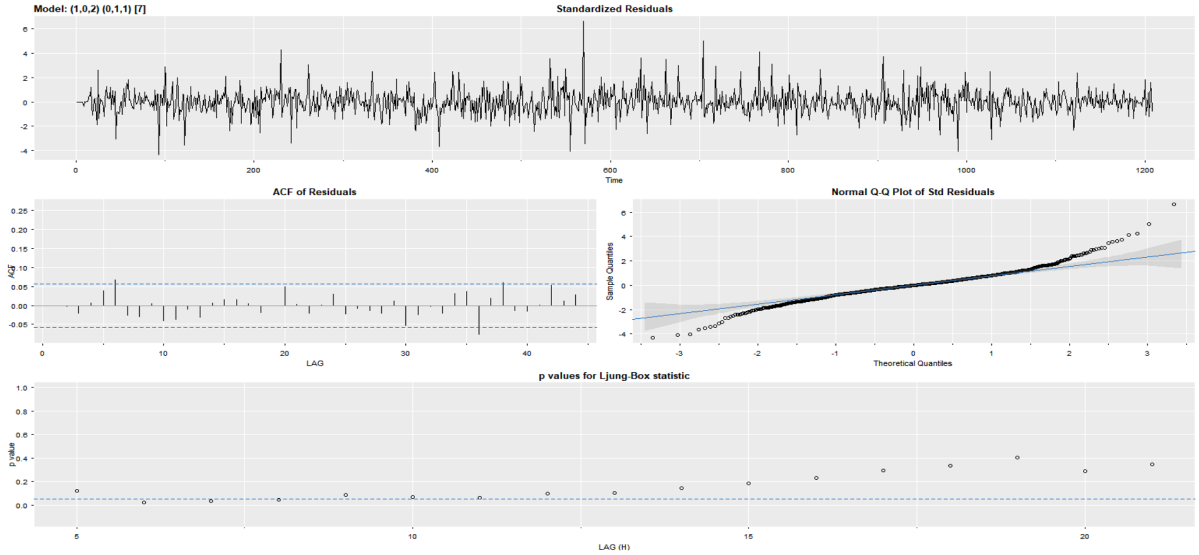
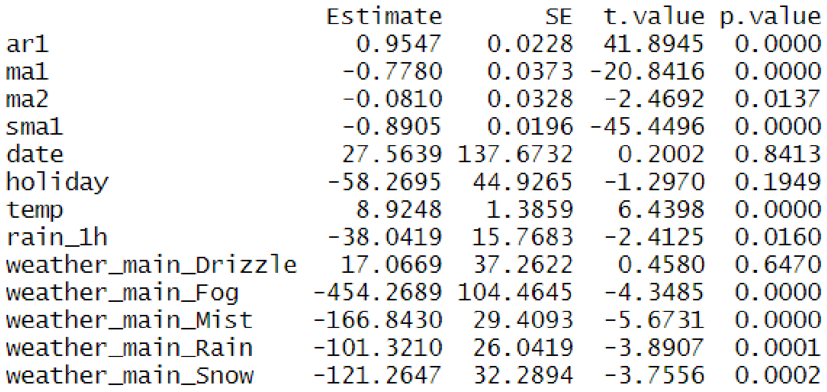


Figure 3.4: Residual Analysis of Final Model

cuts off, and PACF tails off. So MA(1) models are chosen for the seasonal part. For the non-seasonal part, both ACF and PACF tails off. So ARMA models are chosen for the non-seasonal part. After comparing multiple models, we select the one with the least AIC values and all estimates significant. Thus, we can model the residuals as ARIMA$(1, 0, 2) \times (0, 1, 1)_{s=7}$.



```
                      Estimate       SE   t.value p.value
ar1                     0.9547   0.0228   41.8945  0.0000
ma1                    -0.7780   0.0373  -20.8416  0.0000
ma2                    -0.0810   0.0328   -2.4692  0.0137
sma1                   -0.8905   0.0196  -45.4496  0.0000
date                   27.5639 137.6732    0.2002  0.8413
holiday               -58.2695  44.9265   -1.2970  0.1949
temp                    8.9248   1.3859    6.4398  0.0000
rain_1h               -38.0419  15.7683   -2.4125  0.0160
weather_main_Drizzle   17.0669  37.2622    0.4580  0.6470
weather_main_Fog     -454.2689 104.4645   -4.3485  0.0000
weather_main_Mist    -166.8430  29.4093   -5.6731  0.0000
weather_main_Rain    -101.3210  26.0419   -3.8907  0.0001
weather_main_Snow    -121.2647  32.2894   -3.7556  0.0002
```

Figure 3.5: Final Model

From the residual analysis of this model, we see that QQ plot is nearly a straight line, so assumptions of normality are satisfied. The time series plot of residuals don't show any possible patterns. The ACF plot of residuals don't show any possible correlation. From the t-table, we see that the estimates of AR1, MA1, MA2, and SMA1 are significant.

Thus, our model equation can be finally written as $y_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + ... + \beta_p z_{tp} + X_t$,where y is our response variable, and β are the coefficients of z. Here, X_t is modeled as ARIMA$(1,0,2) \times (0,1,1)_{s=7}$. Thus,we can write X_t as $(1 - B^7)(1 - \phi B)X_t = (1 + \theta_1 B + \theta_2 B^2)(1 + \Theta B^7)W_t$.

Our Regression equation is

$$y_t = \beta_1\text{date} + \beta_2\text{holiday } + \beta_3\text{temp} + \beta_4\text{rain\_1h} + \beta_5\text{weather\_main\_Drizzle} \\ + \beta_6\text{weather\_main\_Fog} + \beta_7\text{weather\_main\_Mist} + \beta_8\text{weather\_main\_Rain} + \beta_9\text{weather\_main\_Snow} + X_t$$

,where y is Traffic Volume, and β are the coefficients (available in Figure 3.5). Here, X_t is modeled as ARIMA$(1,0,2) \times (0,1,1)_{s=7}$. Thus, we can write

$(1 - B^7)(1 - 0.9547B)X_t = (1 - 0.7780B - 0.0810B^2)(1 - 0.8905B^7)W_t$.

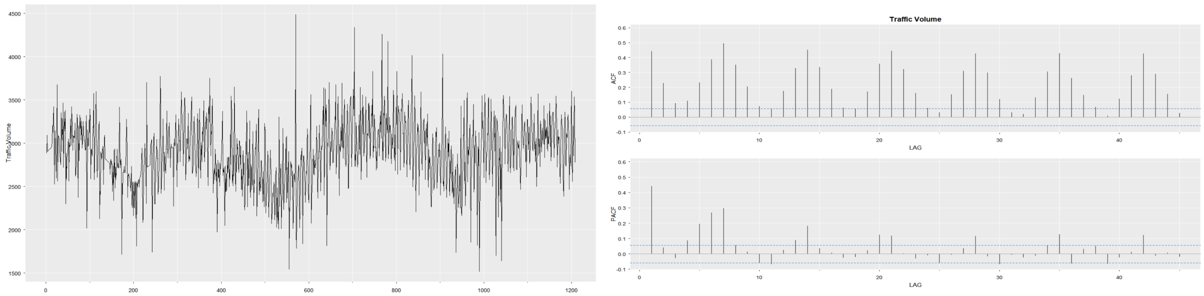
We achieve a final AIC value = 13.97, and BIC = 14.03.

3.2 SARIMA Model

We now fit a SARIMA model to the time series. We first plot the data, and look at its ACF/PACF plot.

From the ACF plot (Figure 3.6b), it is clear that there is seasonal persistence in the data at every seventh lag. Hence, we perform seasonal differencing and build the autocorrelation plots again.

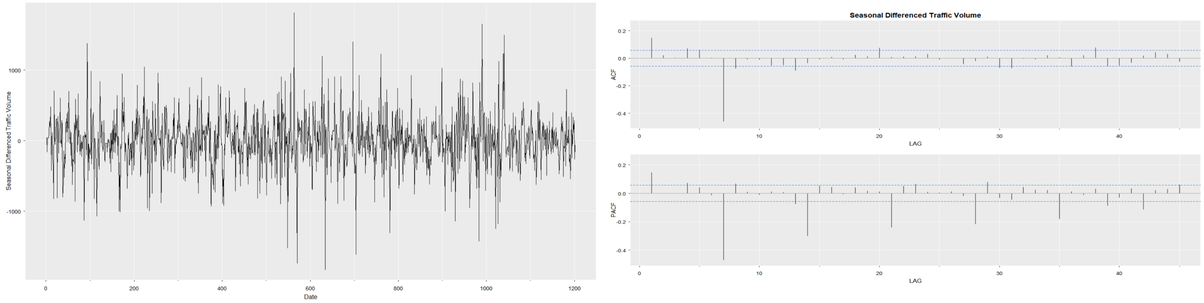
On checking the ACF and PACF of the seasonally differenced series (Figure 3.7b), we can see that for the seasonal part ACF cuts off after the first seasoanal lag while PACF tails off. With this, we can say the seasonal part of the data has AR order (P) of 0 and MA order (Q) of 1. Additionally, in the non-seasonal part, both ACF and PACF cuts off after



(a) Time Series Plot

(b) ACF/PACF plot

Figure 3.6: Time Series Plot and ACF/PACF



(a) Seasonal Differenced Plot

(b) ACF/PACF of Seasonal Differenced Series

Figure 3.7: Seasonal Differenced Plot and ACF/PACF

first lag. Hence, we can take this to be an ARMA model.

Since, we cannot identify the underlying order for AR and MA part, we start with both as 1 and iteratively choose the best model basis estimate significance, residual analysis and AIC/BIC.

Model	AIC	BIC
ARIMA$(1,0,1) \times (0,1,1)_7$	14.14	14.16
ARIMA$(1,0,2) \times (0,1,1)_7$	14.13	14.16
ARIMA$(2,0,1) \times (0,1,1)_7$	14.13	14.15
ARIMA$(2,0,2) \times (0,1,1)_7$	14.13	14.16

Table 3.1: Performance of Models

On iteration, since the AR2 and MA1 estimates were significant, we end up with a seasonal ARIMA model with non-seasonal order (p,d,q) as (2,0,1) and seasonal order (P,D,Q) as (0,1,1) with seasonality of 7. On comparing AIC and BIC (Table 3.1) as well, the chosen model performs the best.

Furthermore, on residual analysis (Figure 3.8), we found that the residuals are approximately normal with most of the p-values in the Ljung-Box test above 0.05 and the ACF values almost all being under the expected large sample distribution values for gaussian white noise.

Thus, our final model is ARIMA$(2,0,1) \times (0,1,1)_7$. We can estimate traffic volume as

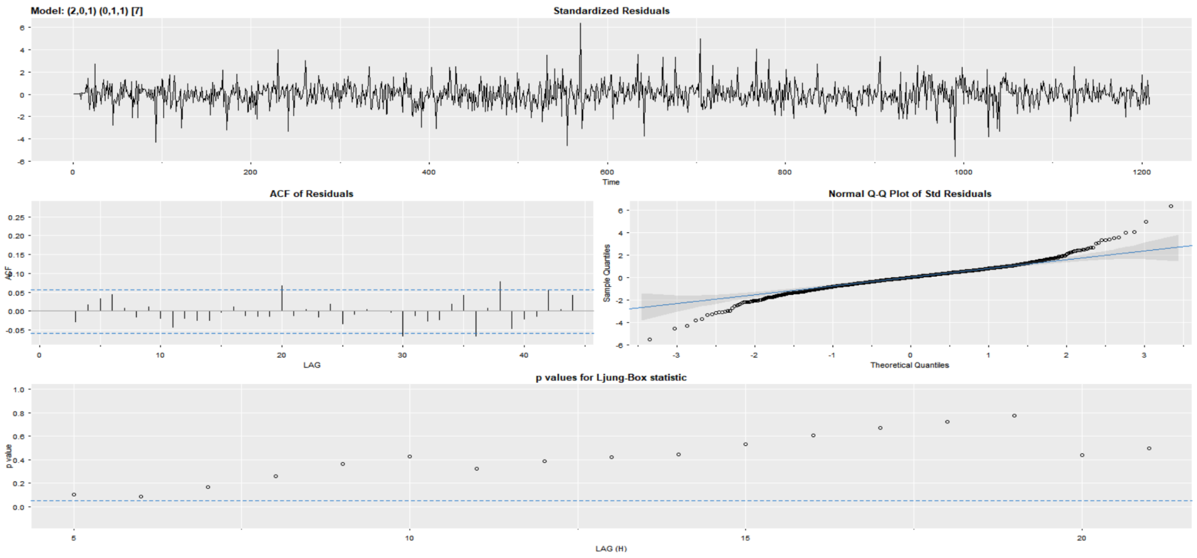


Figure 3.8: Residual Analysis of Final SARIMA Model

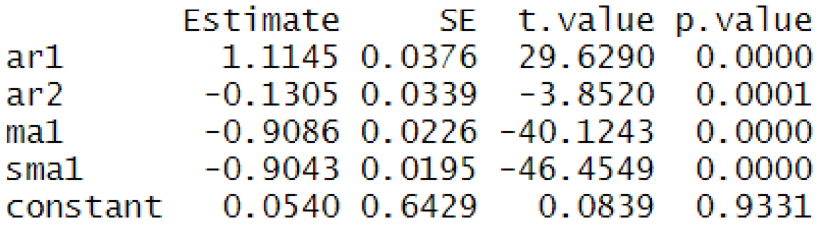

```
         Estimate     SE  t.value p.value
ar1        1.1145 0.0376  29.6290  0.0000
ar2       -0.1305 0.0339  -3.8520  0.0001
ma1       -0.9086 0.0226 -40.1243  0.0000
sma1      -0.9043 0.0195 -46.4549  0.0000
constant   0.0540 0.6429   0.0839  0.9331
```

Figure 3.9: Final SARIMA Model

$(1-B^7)(1-\phi_1 B-\phi_2 B^2)Y_t=(1+\theta B)(1+\Theta B^7)$. Using the estimates from Figure 3.9, we get

$(1-B^7)(1-1.1145B+0.1305B^2)Y_t=(1-0.9086B)(1-0.9043B^7)$.

Future 5 values are forecasted and presented in Figure 3.10.

3.3 Advanced Topics

We explore our data using some advanced topics.

3.3.1 Neural Network Autoregression Model

Artificial neural networks are forecasting methods that allow complex nonlinear relationships between the response variable (traffic volume) and its predictors. The inputs to the neural network are lagged values of the time series. NNAR (p,k) indicates there are p lagged inputs and k nodes in the hidden layer.

NNAR (30,16) means that the last 30 observations are used as predictors and there are 16 neurons in the hidden layer.

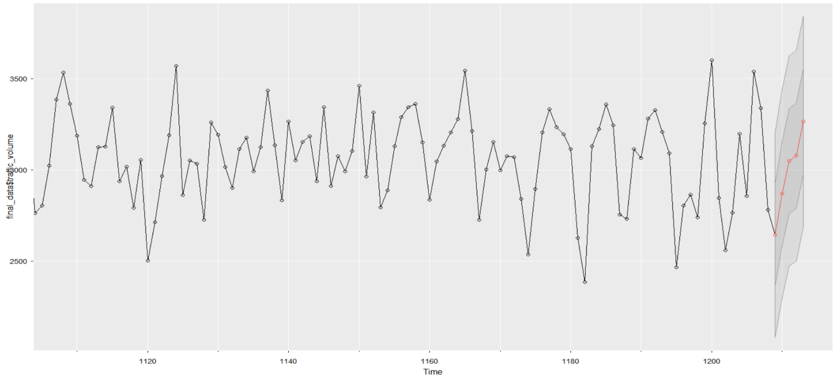


Figure 3.10: Forecasting using SARIMA model

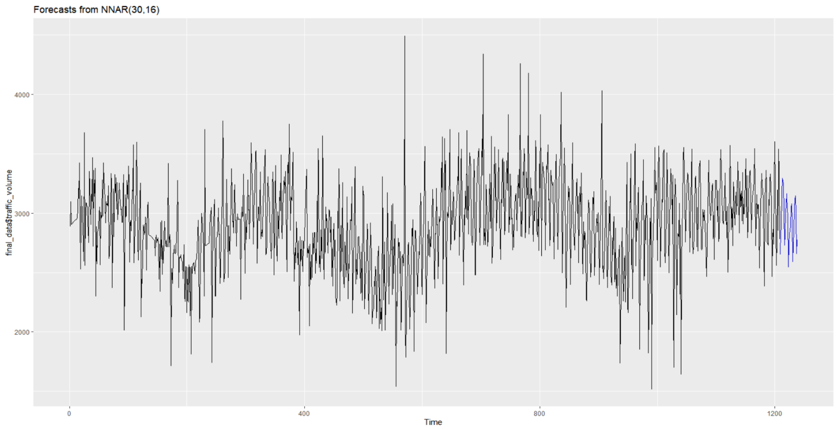


Figure 3.11: Forecast using Neural Network

3.3.2 Spectral Analysis

There is a presence of periodic behavior in time series and it can be quite complex. Spectral analysis is used to find the underlying periodicities. Spectral analysis is done in the frequency domain. The decomposition of a time series into underlying sine and cosine functions of different frequencies allows us to determine those frequencies that are strong or important.

The periodogram distributes the variance over frequency. The main drawback of raw periodogram is that it packs spikes close to each other. We smooth the periodogram to better distinguish the peaks. The smoothing should not be excessive as it might blur the features that we want to find out.

The periodogram shows that peaks are occurring at around '0' frequency. We find the frequency of the highest three spectrum values from the periodogram. Inverse of the frequency gives us the periodicities. The strongest periods are 405 and 6.98 days. The confidence intervals for these predominant periods are (1759796, 11166723) and (1508974, 9575143). The lower bounds are greater than 0, so they are significant.

The periodicity of 405 days means that there is some cyclic pattern observed yearly. This same pattern is observed when we look at the time series plot of traffic volume. The value

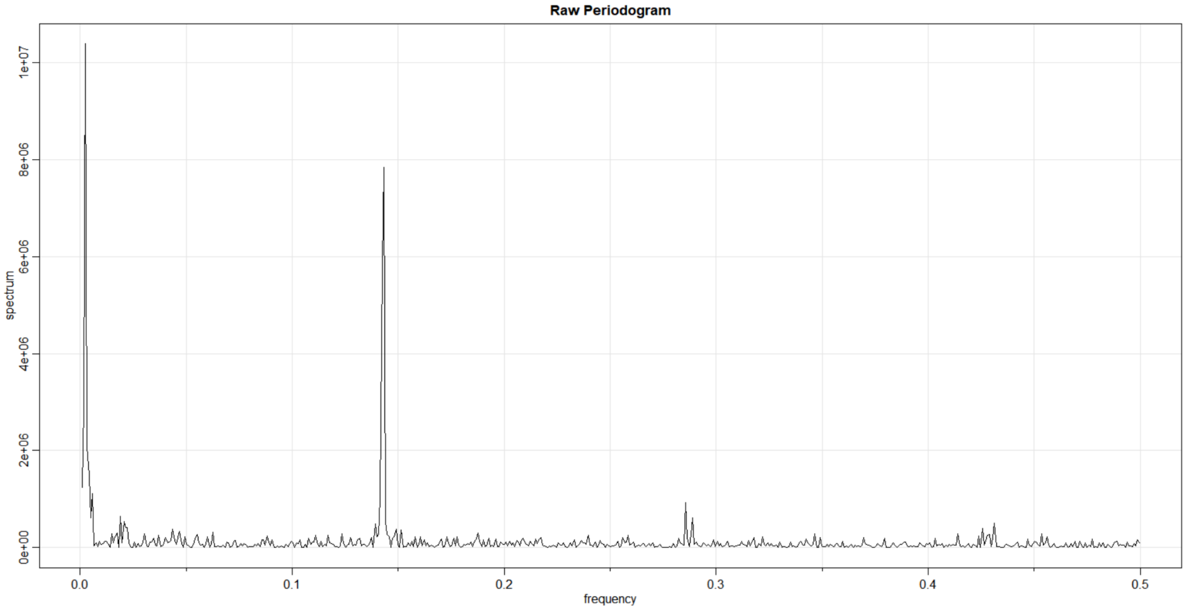


Figure 3.12: Periodogram

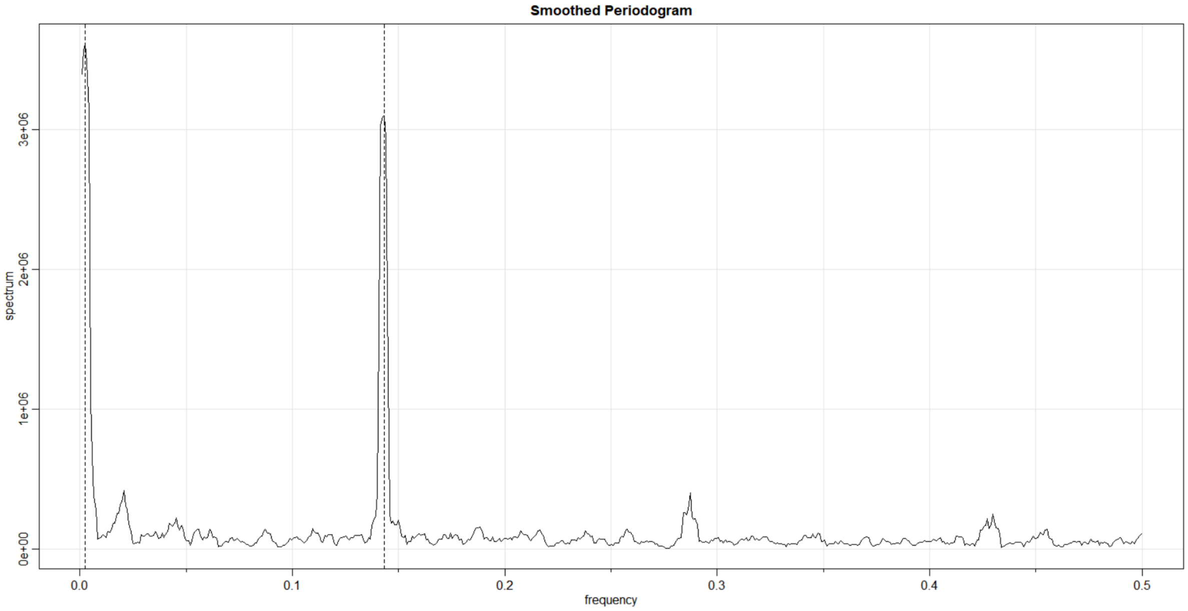


Figure 3.13: Smoothed Periodogram

of our second periodicity, 6.982 matches with the weekly seasonal persistence (7 days) we observe in the ACF plots.

Chapter 4

Results and Discussion

In the time series analysis of the metropolitan traffic volume data, we identified the underlying relationship in the data. Initially, we built time series and autocorrelation plots to analyze and understand the data. From this, we found seasonal persistence in the data every seven days. This validates the understanding of the traffic volume showcasing patterns week over week.

We then used multiple linear regression followed by time series analysis of the auto-correlated errors, to model the traffic volume time series along with correlation analysis. With this, we identified the statistically significant predictors and achieved an AIC of 13.97. The significant predictors in the model aligned with our understanding of the prevalence of weather conditions such as fog, mist, etc., being significant predictors along with others such as holidays.

We also fit the traffic volume data to a seasonal ARIMA model, wherein we finalized the one with statistically significant estimates' p-values and lowest AIC of 14.13. Furthermore, for this model, we identified a seasonal component of ARIMA(0,1,1), along with the non-seasonal part being an ARIMA(2,0,1). This could be understood as the current traffic volume being for the hour in question being predicted by the traffic volume of the past 2 days and from 7 days back, with some degree of error.

The spectral analysis also revealed two important periods. It confirmed our weekly cycle and also showed the presence of yearly cycles, which we suspected from the time series plot.

This study was limited due to presence of outliers and a lot of missing values, which were imputed using Kalman Smoothing and State Space Models. We also looked at only a particular hour of the day. This study could be expanded by including data of different hours to provide a more detailed inference of data.