Time Series analysis and forecast on Seoul Metropolitan Subway Passenger data between 2019 to 2021 September to explore how the pandemic has affected number of Seoul Metro passengers.

통계학과 2017150470 한승헌

0. Abstract

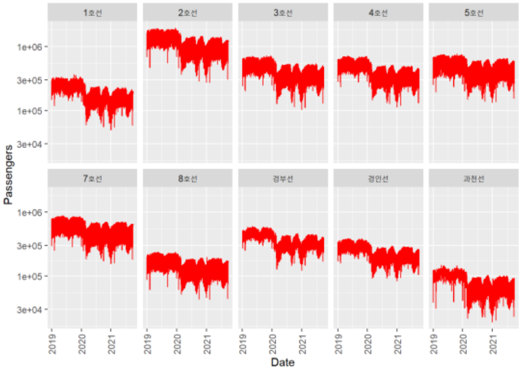
In this project, the goal is to assess impact of COVID on Seoul metro and forecast future monthly passenger number via time series analysis. The assessment focused on time period between Jan 2019 to Sep 2021. Time series modelling could be used to deconstruct data to properly obtain information on characteristics of the process. Because the passenger number data had significant decline after the COVID-19 had landed, it was evidently non-stationary. For most of time series analysis, stationarity condition is paramount because the characteristics of the process has to be time-invariant to extend an analysis to forecasting. Hence, first order differencing was applied to original data set. Afterwards, various time series models including ARIMA(p,d,q), harmonic regression ARIMA, regression model with trend/seasonal components and time series regression model with two relevant predictors, were constructed. For TS regression with predictors, reproduction number and daily COVID infection cases were implemented. Once the model was formed, I conducted model diagnostics, modifications. With adequately modified model was formed, two months forecasting was carried out. Forecasting results were then compared to actual data (Average daily Passenger number in October and November, 2021) to assess a quality of fitness.

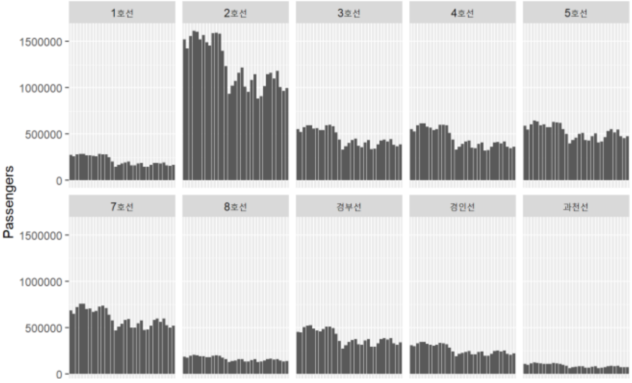
1. Introduction: preliminary analysis on original dataset

Objective is to measure effects of the Covid-19 pandemic on Metro passenger numbers numerically. Specifically, the major aim is to measure economical and social impact of the pandemic indirectly, via exploring fluctuations in daily metro passenger numbers before and after covid; notable decrease in passenger number would indicate major shrink in social interactions and revenue of small business owners. After constructing adequate models, I will anticipate future passenger numbers using forecasting techniques.

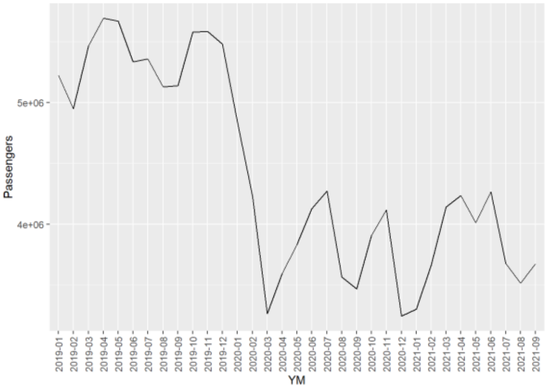
From the raw dataset provided by Seoul metro, I filtered ten lines that had the most number of passengers from Jan, 2019 to Sep, 2021. Using filtered data set, I calculated average daily passenger number for each month. For example, if average daily passenger number is 3 million for October, 2021, it means on average, 3 million people daily used Seoul metro in October.

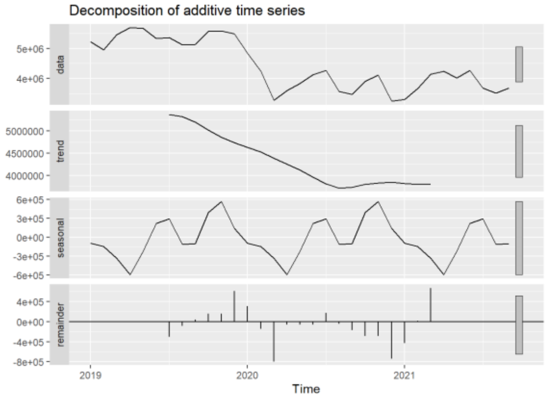
- Trend & Seasonality





Overall, all top 10[1] metro lines have decreasing trend throughout 2019 to 2021. Metro line 2 with the biggest passenger number amongst 10 lines, show clear decline. Also, there seems to be clear seasonality in passenger: Passenger number fluctuates periodically.



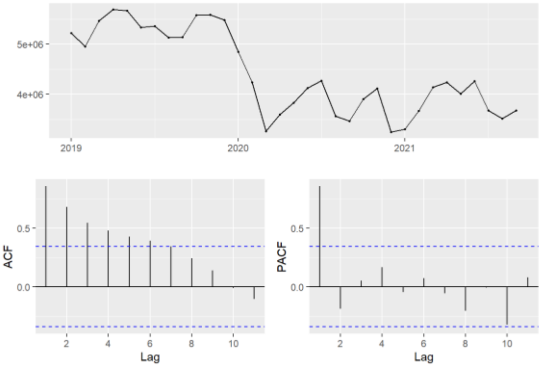


The seasonality could be noticed with sum of passenger number[2] of ten metro lines. There is a decline in passenger number during the summer season, followed by increase in the winter season. Also, there is a huge decline in passenger number after January of 2020. The seasonality witnessed in 2019 is again repeated throughout 2020 and 2021 (Decline during the summer and increase during the winter). However, the average number of passenger per day for each month never recovers to the level of before-pandemic period. The additive decomposition (Because the variance is fairly stable) of data set confirms decreasing trend and existence of seasonality.

[1] Ten Metro lines with the highest average daily passenger number throughout 2019~2021

[2] Average daily passenger numbers of each month for entire TOP 10 metro lines

- Stationarity

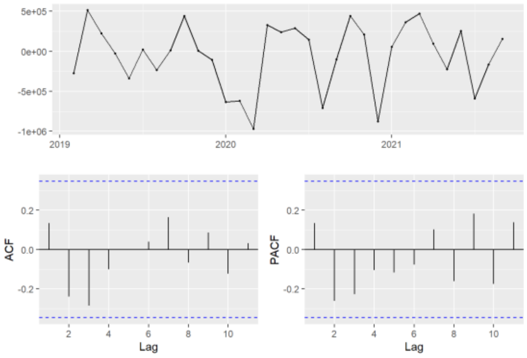


```
## #######################
## # KPSS Unit Root Test #
## #######################
##
## Test is of type: mu with 3 lags.
##
## Value of test-statistic is: 0.7015
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

As shown in the decomposition, there is a decreasing trend. Furthermore, ACF of the data is decreasing very slowly, which is a major sign of non-stationarity. KPSS unit root test rejects null hypothesis (H_0: Series is stationary) under 5% significant level. In a nutshell, as the original dataset has non-constant mean and autocorrelation, an adequate modification is needed for further processes.

- Modifications

To remedy non-constant mean issue, differencing is often implemented. Since the data has seasonality, different combinations of differencing have been applied (seasonal differencing followed by lag=1 differencing, lag=1 differencing and lag=12 differencing). Through thorough evaluation of time series plot and ACF graph of each differenced series, the lag=1 first order differenced series has been chosen as the best modification method.

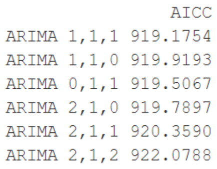


```
## #######################
## # KPSS Unit Root Test #
## #######################
##
## Test is of type: mu with 3 lags.
##
## Value of test-statistic is: 0.0718
##
## Critical value for a significance level of:
##                 10pct  5pct 2.5pct  1pct
## critical values 0.347 0.463  0.574 0.739
```

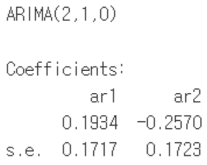
Differenced series now satisfies stationarity: No distinct trend, ACF all smaller than critical values. KPSS test now does not reject null hypothesis under 5% significant level. Therefore, for models that require original series be stationary, differencing would be necessary prior to model fitting process.

2. Fitting the model with ARIMA model

- The Diagnostics & Modifications



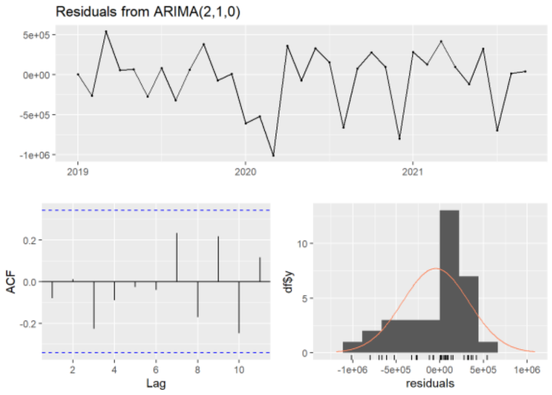
```
                   AICC
ARIMA 1,1,1 919.1754
ARIMA 1,1,0 919.9193
ARIMA 0,1,1 919.5067
ARIMA 2,1,0 919.7897
ARIMA 2,1,1 920.3590
ARIMA 2,1,2 922.0788
```

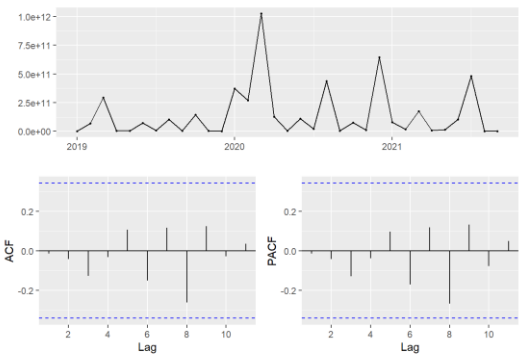


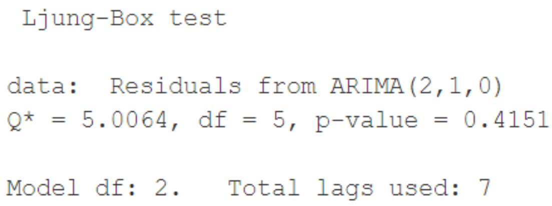
```
ARIMA(2,1,0)

Coefficients:
         ar1      ar2
      0.1934  -0.2570
s.e.  0.1717   0.1723
```

Different types of ARIMA model could be used. Since the ACF and PACF curve of differenced data did not show any sign of AR(p) of MA(q), different combinations of ARIMA(p,1,q) should be tested. Comparing AICC value of different models, ARIMA (1,1,1) or ARIMA(0,1,1) seemed like the best model. However, as the MA coefficient was 1 it was an inadequate model[3]. Thus, ARIMA (2,1,0) seems like the best option.



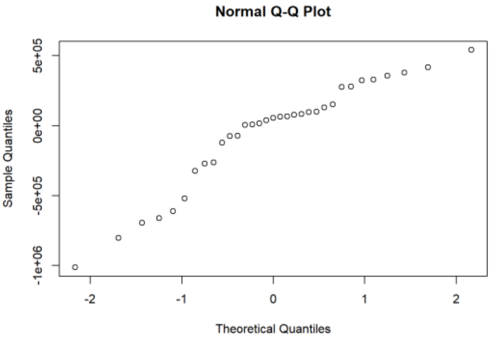




```
 Ljung-Box test

data:  Residuals from ARIMA(2,1,0)
Q* = 5.0064, df = 5, p-value = 0.4151

Model df: 2.   Total lags used: 7
```

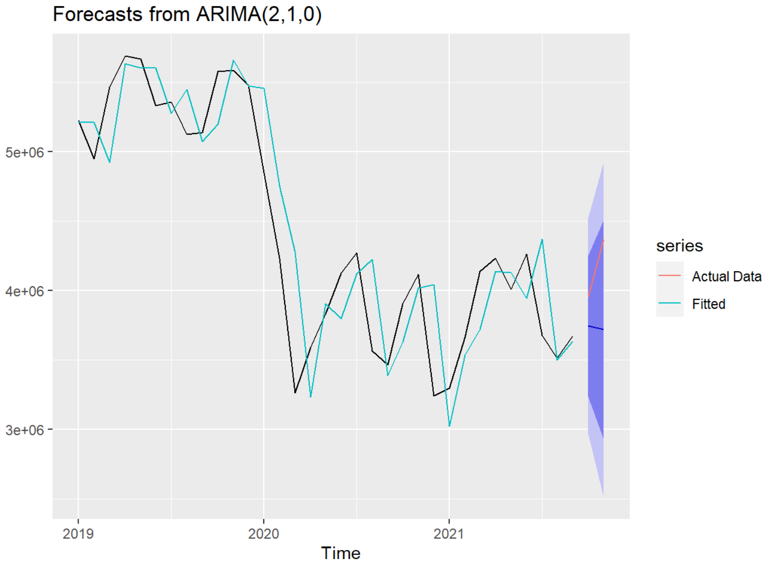


Residual plot of chosen model seems to be stationary: constant mean, variance, and no sign of autocorrelation. There seems to be no conditional heteroscedasticity as ACF and PACF graph of

[3] MA(1) coefficient =1 is a sign of not-invertible model,

residual squared seems to suggest no sign of autocorrelation[4]. Though the data does not seem to follow normality, as shown in the normal Q-Q plot, the model generally has a good fit[5]. Thus, ARIMA (2,1,0) model could be used for forecasting.

- **Forecasts**

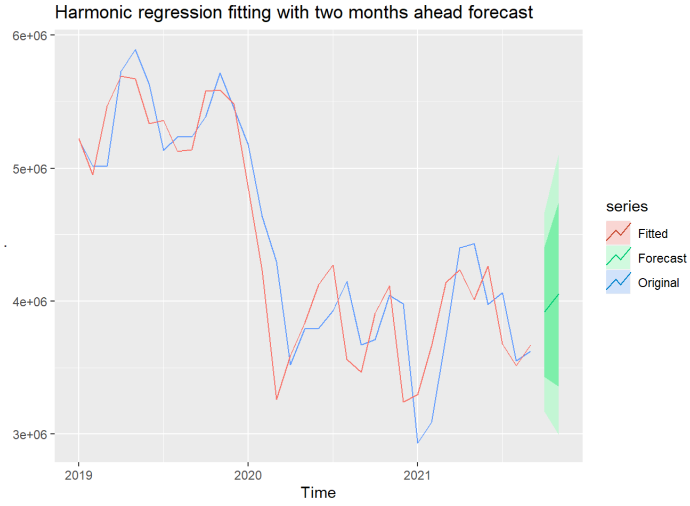


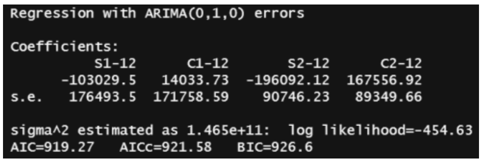
Chosen ARIMA model fits the original data very well. Seasonality has been accurately estimated. Two months forecast showed stagnancy in number of passengers per day. However, actual data from October and November shows extremely steep increase. Such outcome might be due to implementation of "With Corona" policy which extended metro hours by an hour or so.

Another possible ARIMA model is dynamic harmonic regression model which uses Fourier terms as predictors.

[4] Ljung-Box Test can not reject null hypothesis: There is no autocorrelation in the model.

[5] As long as the residual is random, normality assumption need not to be considered for ARIMA





```
Regression with ARIMA(0,1,0) errors

Coefficients:
            S1-12     C1-12       S2-12      C2-12
      -103029.5  14033.73  -196092.12  167556.92
s.e.   176493.5  171758.59   90746.23   89349.66

sigma^2 estimated as 1.465e+11:  log likelihood=-454.63
AIC=919.27   AICc=921.58   BIC=926.6
```

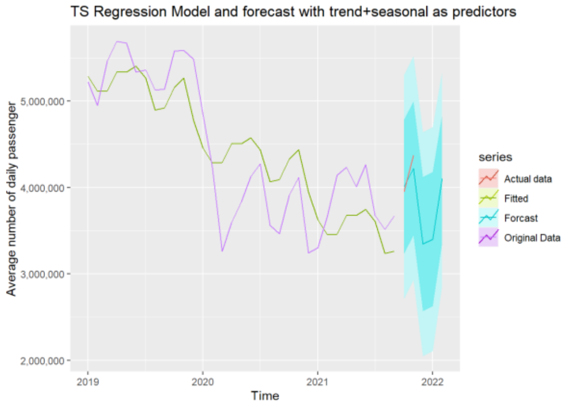
Different options for k was tested and k=2 returned the smallest AIC value. The fitness of model is very accurate. Thanks to Fourier terms, seasonality of the data has been very precisely estimated. Increase in average daily passenger number is expected according to the forecast.

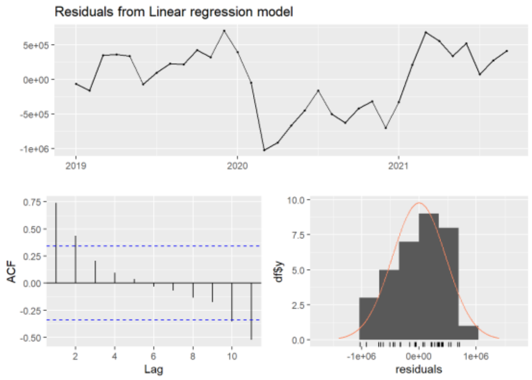
3. Fitting model with tslm

- Fitting

Another way of decomposing and analyzing time series is tslm. By trend/seasonal factors and regular input variables could be considered as predictors.

- Diagnostics & Modifications & Forecast





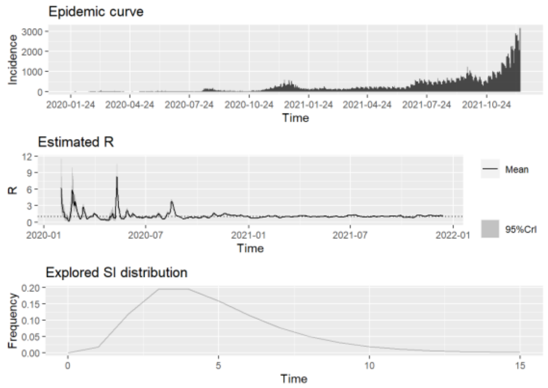
Using trend and seasonal component of tslm function in R, the series was fitted. Compared to ARIMA(2,1,0) model the seasonal fluctuation was less accurately estimated. Nonetheless, the fitted model generally follows decreasing trend. The residual plot satisfies stationarity: Constant mean and ACF dropping to zero very quickly.

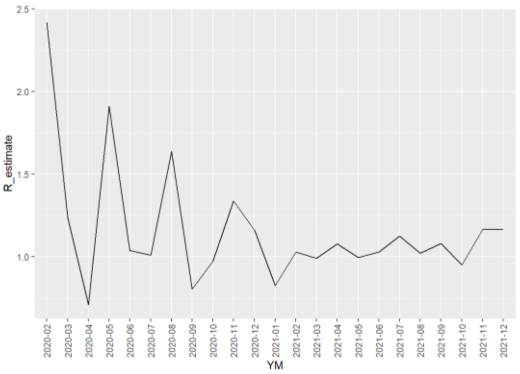
Interestingly, forecast given by tslm model is very similar to an actual data. It forecasts drastic increase for the first two months and a decline in the following month.

- Introducing new predictors

As mentioned, tslm can also fit regular regression model. Intuitively, I anticipated that degree of COVID-19 spread would have significant relation with number of metro passengers for two reasons. First, one is less likely to use public transportation when he/she feels there is a great chance of getting infected. Second, as the authority is likely to enforce strict quarantine policies when there is a major outbreak, citizens are not likely to get out of their home. To measure fear level for infection, I used number of daily reported infection cases and weekly average of reproduction number (R, 감염재생산지수). It was fairly easy to scrape infection cases online. However, I had to calculate reproduction number using "EpiEstim", a R package.

Calculation of reproduction number requires probability density function of gamma distribution. Korean Disease Control and Prevention Agency (KDCA) uses mean=4.8 and standard deviation=2.3 for PDF. Based on the given information and data set of daily reported infection cases, I was able to calculate daily reproduction number.





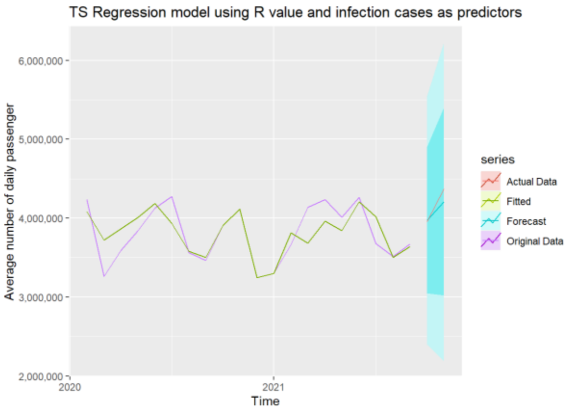
Traditionally, R value bigger than 1 is considered major outbreak of a disease. In regards to COVID pandemic in Korea, she had went through quite a hardship during the first half of 2020 as the R value reached as high as 2.4 in Feb, 2020. The second graph shows monthly average reproduction number[6][7] from Feb, 2020 to Dec, 2020. R value fluctuated severely in 2020. In 2021, R value remained very stable.

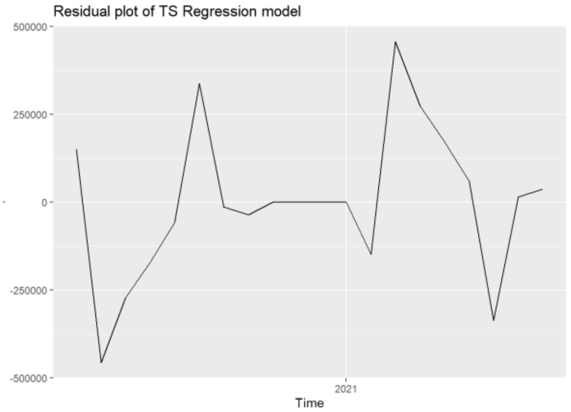
[6] <PUBLIC HEALTH WEEKLY REPORT, KDCA>

[7] Average of daily reproduction number for each month

Considering close psychological relation between fear of pandemic and willingness to use public transportation, it would be worthwhile attempt to fit passenger number with R value and daily reported infection cases.

Before conducting an analysis, prior data processing was necessary. Since there is no data for R value in 2019, I would have to narrow my original dataset to 2020 and 2021.





Notice that the dataset from Feb 2020 to Sep 2021 follows stationary process with constant mean and variance. Thus, no differencing or log transformation was required. Also, since the data still has seasonality, I added seasonal component to tslm function.

The result is astonishing. Regression model fits the original data very well. For forecasting, I used actual R production value and reported infection cases for October and November. Then, forecast was almost the same as the real data.

4. Discussion

Covid-19 has made some critical damages to our society in various ways. Due to pandemic, untact social interaction has become a new culture. As a result, citizens do not travel or leave the door as much as they did before. Such change was evident in metro passenger number data. Average passenger number in 2020 and 2021 decreased significantly. Based on a TS regression model forecast which used trend and seasonal component, even after 5 months, it is likely that the passenger number would not recover back to pre-covid range.

Also, it was discovered that passenger number has seasonality: There is a major decline during summer and increase during winter. This may be due to summer vacation when citizens take a long vacation which results in decrease in number of commuters.

Regarding the models, dynamic harmonic regression model which incorporates ARIMA and Fourier terms has fitted the seasonality the best. This was due to the wriggling nature of Fourier terms (sin and cos functions). Though the model was very accurate, it has predictor variables that are very hard to interpret as it uses series of trigonometric functions.

The Time series regression model is useful as it can incorporate different kinds of data as predictors. For example, fitting response variable with relevant predictors avoids chance of committing confounding bias. Nonetheless, it is limited because the forecasting technique requires future value of predicts be commuted. In other words, to forecast response variable, another forecast of predictors is needed and such method would increase bias. However, forecast was very precise when actual values of predictors were commuted, so there is a trade-off.

The regular ARIMA(p,d,q) model could be a compromise between dynamic harmonic regression model and TS regression model. As it uses series of previous observations and errors, it can fit seasonality fairly well. Also, it does not require prediction of predictors.

- Weaknesses

Transportation data is a very tricky data. There are numerous confounding variables even after taking various available into account. For example, due to high price of petrol, commuters might use metro instead of their own vehicle which would lead to unanticipated increase in passenger number. Therefore, timeseries regression based on R-value and infection cases is likely to have confounding bias.

I believe two years worth of accumulated data is not sufficient enough to build robust model. Due to such reason, most of my models (ARIMA, tslm) did not satisfy normality assumption. Thus, much bigger sample of data would be required to construct much more precise models.