Analysing Daily Passenger counts in NYC Yellow Taxis

Gaurav Singh
UID: 305353434
December 6, 2022

Abstract

New York City's (NYC) Yellow Taxis are iconic to the city. Most of us have seen it in movies or real life or might have used it at some point in time. They used to be very popular in the last decade but something has changed since then and the passenger counts are decreasing every year. This paper analyses the daily passenger counts for NYC Yellow Taxis for the period of 2017 to 2019 and aims to predict the daily passenger counts given historic data using Time series Models. This analysis might be useful for the Government and the Taxi drivers to estimate the risk of extinction for this iconic service.

**The code for this project can be found here: [1]

1 Introduction

Being one of the most populous cities in the United States, New York City has millions of taxi trips completed every month. As of now there are three most popular Taxi services available for the people to choose from - Yellow taxis, Green Taxis and Uber rides [2]. Out of these Yellow medallion taxis are the oldest and most iconic which can be hailed from anywhere in NYC and are allowed to drop off passengers at any spot. Their number is quite limited though as they are managed by the State government.

The Green taxis and Uber rides are a newer edition which seem to be preferred more by people and the drivers. One of the main reason why the Yellow medallion taxis are getting extinct is because the Yellow Taxi drivers have fixed rates depending on the area as compared to their competitor - Uber which gives better incentives to the drivers depending on area, distance, congestion and time of day. Also, it is more convenient for drivers to pick up passengers from fixed spots instead of roaming around. These reasons are explained well in this post [3]. Interestingly there were more than $\sim 13,500$ such taxis around 10 years back but now less than 9,000 of them are left.

The goal of this project is to analyse the daily passenger counts for the Yellow taxis, identify the hidden patterns and cycles associated with these passenger counts and can Time series models like ARIMA, predict these counts well for a 60 day period. This paper also aims to investigate if there is really a need to revive these iconic Yellow taxis or they are just doing fine?

2 Dataset

The dataset used is provided by NYC Taxi and Limousine Commission [4] and is open-sourced. The data contains number of rides completed by a Taxi and the number of passengers corresponding to the trip is reported by the taxi driver. The dataset is provided in chunks on monthly basis for every trip. This project uses data from 01/01/2017 to 12/31/2019 i.e. 36 months of data. The whole data is merged into a single file and grouped by date to aggregate the daily passenger counts. Full consolidated dataset has 1095 days of data out of which last 61 days (11/01/2019 to 12/31/2019) are kept for testing and is not used for any analysis. COVID pandemic time (2020 on-wards) is not taken into consideration given at that time there were only 700 taxis in service.

3 Exploratory Data Analysis

Looking at the time series data from Fig. 3, it looks like data is quite noisy, discontinuous with a strong presence of cycles and seasonality. This is evident because repeating patterns are observed in the plot. It's observed that overall trend of the data is decreasing and a linear trend would be a good estimate for this decrease which is shown using the blue line in Fig. 3. The fitted trend has a slope of -200.13 and a intercept of 547,356 as shown in Table 1, which shows that in the beginning of 2017 there were $\sim 550K$ daily passengers but the mean of the train data is $\sim 438K$ which shows that over the time the passengers are decreasing, precisely around 200 passengers/day.

Fig. 1(a) shows the total yearly passenger counts and the decay is quite linear from 2017 to 2019 which supports the choice of linear trend. Monthly aggregates from Fig. 1(b), shows that Spring season is favourable for Yellow taxis with large number of passengers as compared to Winter season where the counts are quite scarce. This makes sense because NYC sees quite cold winters with snowfall which restricts people from going out. Weekly aggregates from Fig. 1(c) indicates that Monday and Sunday are not favourable for taxi business whereas there is a linear growth in passengers as the week progresses from Tuesday to Saturday. Further inspection on the individual dates shows that the mid-March week which marks the beginning of Spring in NYC seems to be most profitable for taxi drivers and Holiday weeks like Christmas and Independence week seems to be least profitable with lowest no. of passengers as shown in Table 2

From the ACF plot in Fig. 4, it looks like ACF is oscillating and gradually decreasing whereas the PACF in Fig. 5 seems to cut off at a lag 2, which hints that this time series can be modelled well by an AR(1) process. Each lag in the ACF and PACF plots represents 1 day in a year.

4 Data Preparation before Time Series Modelling

Before any time series modelling it is absolutely necessary that it's almost stationary and that it's free from any dominating trends and cycles. Augmented Dickey-Fuller Test conducted on train data has a p-value of 0.01 which shows that despite the series being discontinuous with strong presence of cycles and trends, it's Stationary with low confidence. The first thing is to remove the linear trend from the train data which gave the residuals as shown in Fig. 6. Looking at the residuals, it looks much more stationary but the cyclic patterns are still intact in it. To identify such cycles, AR spectral estimation and Daniell window (span = 3), smoothened Spectral analysis is conducted and the Frequency densities are plotted as shown in Fig. 7.

From the both the AR and non-parametric spectrum, it's evident that there is a presence of lot of low frequency (large period) cycles, but the major concern is with the sharp sudden spikes in the plot which carry too much power and dominate the time series. 6 such cycles are identified with weekly cycle(7 day) carrying the most power, followed by Half-yearly(182 day), Yearly(365 day), One and a half month (45 day), Monthly(30 day) and Half-weekly cycles (in decreasing order of power contribution). Out of these 6 dominant cycles, 5 cycles having most power are removed from the residuals. The method opted for multiple cycle removal is the removal of mean of the corresponding days of cycle from the residuals. The cycles are removed sequentially one at a time and the means are computed only from the train data residuals left after each successive removal. It is observed that removing a large period cycle before a shorter period cycle introduced some new cycles whose period is hard to estimate. This was not the case if shorter cycles are removed before the longer ones. The 5 cycles are shown in Fig. 8 where mean residuals are plotted for each day. The weekly cycle resembles in shape with the total weekly distribution shown in Fig. 1(c).

The cycle removed residuals are shown in Fig. 10 and it looks like they are quite free from any trends or cycles as such and look more like white noise. ADF test on these residuals has a p-value of 0.001 which shows the left out residuals are much more stationary with higher confidence. It was observed that removal of these 5 cycles introduced some other period cycles as shown in Fig. 9 but those are not removed given their less relative densities. ACF and PACF of these final residuals as shown in Fig. 11 and Fig. 12 respectively, which strengthen the hypothesis that AR(1) might be a good fit given ACF is tailing off much better as compared to raw data and PACF still cuts off at lag 2.

5 Modelling Approaches

From the ACF and PACF plots it looks like the final residuals can be modelled by an AR(1) time series model. Also ARMA models can be helpful to fit this kind of noisy and complex data given the original time series had both ACF and PACF have somewhat oscillating values. ARIMA models are not considered given the time series is stationary from the beginning and that d parameter would be 0.

To find the potential p and q parameters for the ARMA(p, q) model, grid search is used with p, q $\in [1, 10]$ where p is the order of Autoregressive (AR) model and q is the order of Moving Average (MA) model. p and q values are restricted below 10 to avoid overfitting and that runtimes are very high for higher order models. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) scores are calculated for every model. The AIC scores are shown as a heatmap in Fig. 13(a) which supports ARMA(9, 9) as a good candidate with lowest AIC score of 22.55. From the heatmap it is observed that variance of AIC scores is very low and choosing such a high order model can lead to overfitting. To avoid this, harmonic mean of AIC and BIC scores is plotted, as shown in Fig. 13(b) and ARMA(6, 4) is chosen as another potential candidate based on lowest harmonic mean score of 22.58. BIC score penalizes model complexity and thus helps in choosing simpler models.

6 Results and Discussion

The first model, ARMA(9, 9), is fitted on final train residuals and is shown in Fig. 14 which is extremely choppy with sudden spikes. The train residual RMSE is 27848.66. One thing noticed is that, the shape of the predictions is matching with the actual residuals but there is some positive constant missing which makes the prediction values small. One of the possible reasons for this anomaly is that it's a complex model and relies a lot on the first 9 values. Observing Fig. 14, it looks like the model did poorly in the start which gave the predictions a poor momentum, making the further predictions underwhelming by some constant. To mitigate this, different positive constants are added to the predictions till the train RMSE decreased. A constant of +9000 is added and the residual predictions are shown in Fig. 15 which decreased train RMSE by 1390.2. The full train data predictions are shown in Fig. 16 which again is quite discontinuous with sharp peaks and looks like an overfit with RMSE of 26458.46. The test predictions on last two months of hold data is shown in Fig. 17 and has a RMSE of 39954.88.

The second model, ARMA(6, 4) is simpler compared to first one and the fit on train residuals is shown in Fig. 18 which looks a bit smooth and is less over-fitting when compared to ARMA(9, 9). The train residual RMSE is 63430.52 which is quite high but looking at predictions in Fig. 18, the shape of the prediction is quite good, it's just the issue of a missing constant. A positive constant of +59000 is added to reduce the RMSE by 40348.29 and the final residual fit is shown in Fig. 19. The reasoning done for ARMA(9, 9) about why this happens still holds here, given this is also a complex model but the constant required here is quite large. One of the reasons for this is that the AR coefficients(0.02, 0.08, 0.04...) of ARMA (9, 9) are small thus they offered less shift and wrong momentum as compared to ARMA(6, 4) which has large AR coefficients (2.31, -2.32, 0.97...) thus drifting the start by a large error. The full train data predictions are shown in Fig. 20 which again is less discontinuous as compared to ARMA(9, 9) but still looks like an overfit with RMSE of 23082.23. The test predictions on last two months of hold data is shown in Fig. 21 has a RMSE of 37890.48 which is an improvement over ARMA(9, 9) showing that simpler models for this data are working better.

The simplest model ARMA(1, 0), which is found from the ACF and PACF plots earlier, gets the best results. The residual predictions are smoother as compared to previous complex models with less signs of overfitting as shown in Fig. 22 and has a RMSE of 20177.38. The best part is that there is no need for level adjustment given adding or removing a constant from the predictions increases the RMSE. The train data predictions fit much better as compared to other models with less sharp peaks and variations as shown in Fig. 23 and has a train RMSE of 20177.38 which is the least obtained error. The test RMSE obtained from this model is the smallest with a value of 34880.39 and the test data predictions are shown in Fig. 24. The AR(1) coefficient has a value of 0.519 which is not a large one, explains why the predictions are not continuous and has less effect of X_{t-1}. All the results are consolidated in Table 2. Given the mean of data is around 438K passengers/day the test RMSE of AR(1) is quite well within the permissible limits.

The question arises that since simplest of the models did well, will there be any further improvements with even more simpler models? The answer to that is NO, for this data. The Table 3 emphasises on this. Various simpler models like predicting just the train mean for test data, simple linear regression, linear regression with 7-day cycle (having most power) removed, model which gives previous day value as prediction, are tried but the Test RMSEs observed are way too large when compared with AR(1). Also the modelling process involved removing of cycles but is it really needed for this data? The answer to this is YES, its important given it makes the residuals much more stationary making them more suitable for time series modelling. AR(1) model on residuals without any cycle removal has the Test RMSE of 36742.08 which is quite large when compared the best RMSE achieved.

7 Future Improvements

From the test predictions, the AR(1) model did well, achieving quite reasonable RMSE but the results are not that good if the predictions in Fig. 24 are analysed in depth. The model doesn't do well if there is a significant difference between the previous days passenger counts. This can be seen especially in the period of mid-November to the first week of December. ARMA model can't account for the sudden dips and rises which are not seen in history. One approach is that outliers can be removed in the data processing phase and data can be smoothened a bit before modelling. Another thing is that daily passenger counts depends a lot on different factors like the day, date and weekday of the year, if there is some change in local laws and regulations, daily temperatures, domestic/foreign travel policies given NYC is a tourist place, etc. Thus a better approach for this time series would be to use Multi-variate time series modelling to account for better logical variations in the data and in turn better predictions.

Acknowledgements

I would like to thank the students of Time Series Analysis course, Fall, 2022 and Prof. Rick Schoenberg for providing guidance for this project.

References

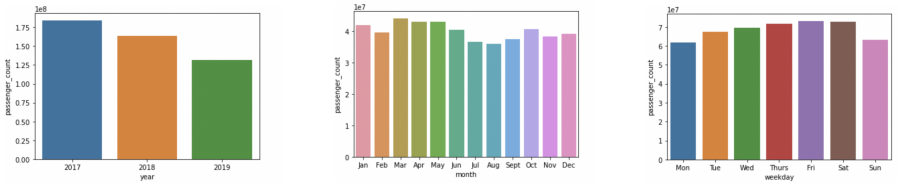
[1] Gaurav Singh. *Analysing Daily Passenger counts in NYC Yellow Taxis*. `https://github.com/gauravSingh30/TimeSeriesAnalysis-NYCYellowTaxis`. 28-Nov-2022.

[2] Wikipedia. *Taxis of New York City*. https://en.wikipedia.org/wiki/Taxis$_o f_N ew_Y ork_C ity$. 22-Nov-2022.

[3] Kirsten Fleming. *Congestion pricing killing NYC's iconic yellow cabs*. https://nypost.com/2022/10/12/congestion-pricing-will-kill-nycs-iconic-yellow-cabs/. 12-Oct-2022.

[4] NYC Taxi and Limousine Commission. *TLC Trip Record Data*. `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. 15-Nov-2022.

Appendix



(a) (b) (c)

Figure 1: Aggregated Passenger Counts (a) Yearly (b) Monthly (c) Weekday

	month_date	passenger_count
78	Dec-25	641107.0
176	Jul-4	819967.0
79	Dec-26	917287.0
77	Dec-24	921746.0
346	Sept-2	957732.0
357	Sept-3	967124.0
177	Jul-5	995328.0
173	Jul-3	1005275.0
335	Sept-1	1018126.0
84	Dec-30	1018691.0

(a)

	month_date	passenger_count
242	Mar-9	1582415.0
213	Mar-10	1551799.0
103	Feb-2	1550036.0
25	Apr-5	1541837.0
26	Apr-6	1532067.0
268	May-4	1531779.0
223	Mar-2	1530383.0
138	Jan-26	1526901.0
234	Mar-3	1523628.0
269	May-5	1522002.0

(b)

Figure 2: Top-10 Dates with (a) Least passengers (b) Most passengers

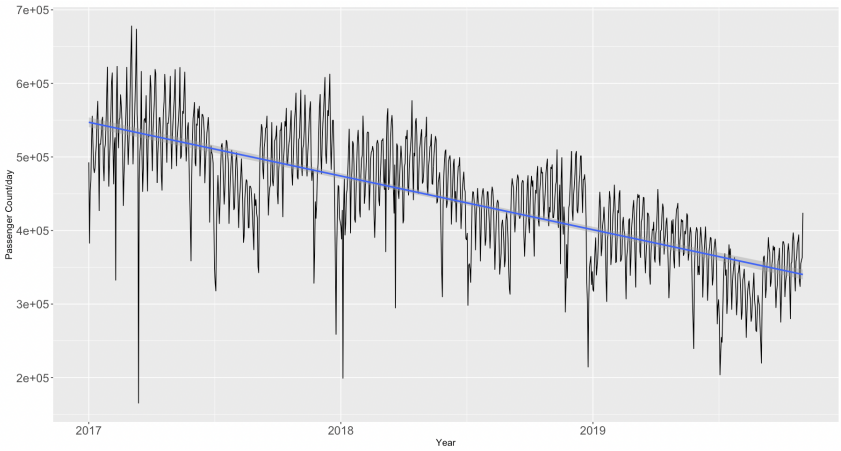


Figure 3: Raw Train Data + Linear trend

	Estimate	Std. Error	t value	Pr(> |t|)
Intercept	547356.006	3401.829	160.90	<2e-16 ***
Time	-200.138	5.689	-35.18	<2e-16 ***

Table 1: Linear Trend Summary

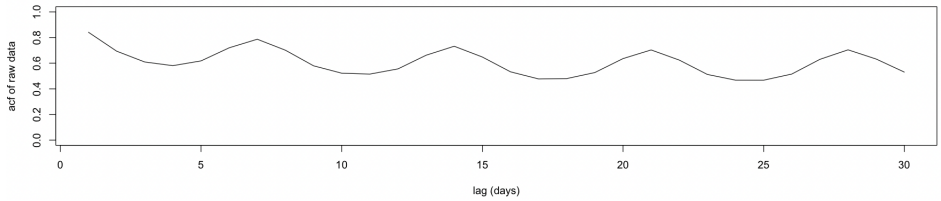


Figure 4: ACF plot for raw data

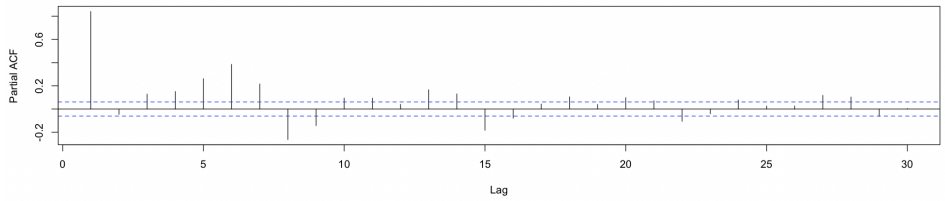


Figure 5: PACF plot for raw data

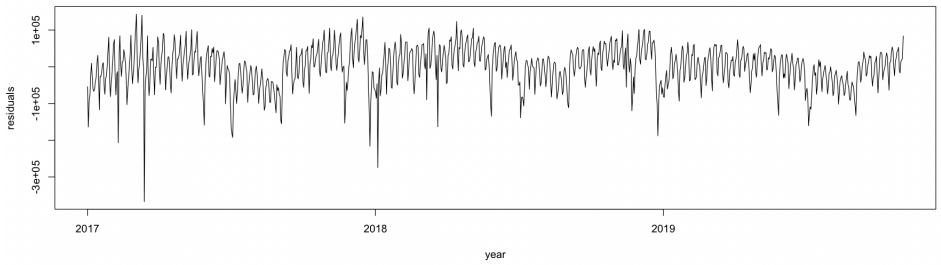


Figure 6: Residuals after linear detrending

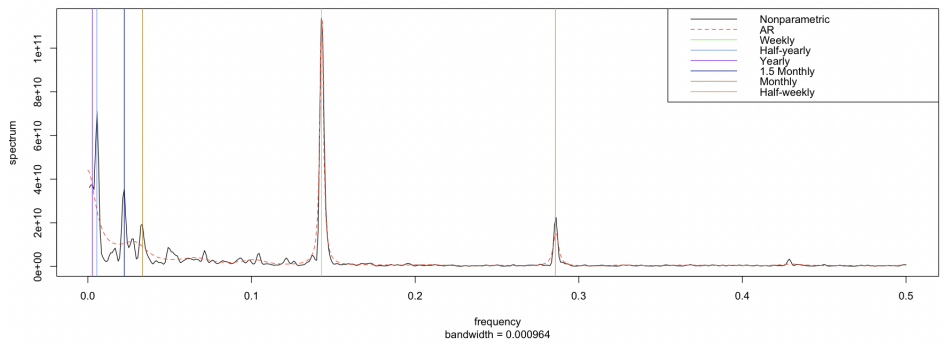
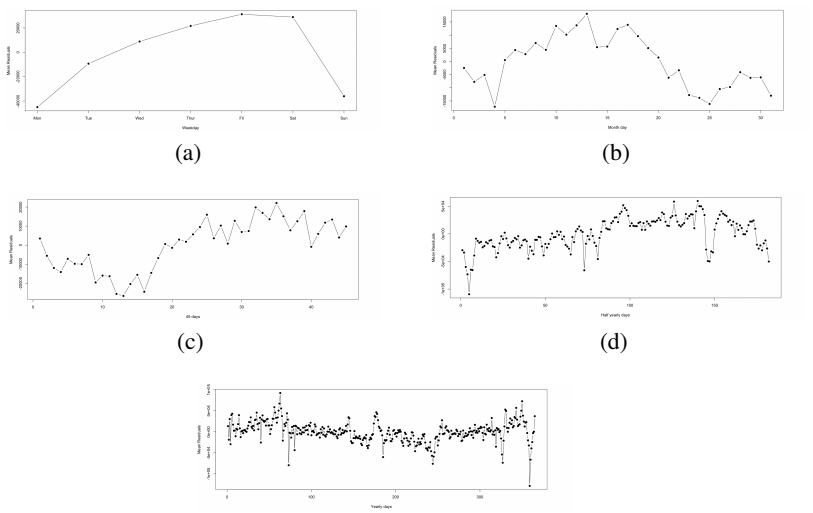


Figure 7: Frequency densities for detrended residuals

Model(p, d, q)	Adjustment	Train RMSE	Test RMSE
ARIMA(9, 0, 9)	+9000	26458.46	39954.88
ARIMA(6, 0, 4)	+59000	23082.23	37890.48
ARIMA(1, 0, 0)	0	20177.38	34880.39

Table 2: Model Evaluations



(a) (b) (c) (d) (e)

Figure 8: (a) Weekly cycle (b) Monthly cycle (c) 45-day cycle (d) Half-yearly cycle (e) Yearly cycle

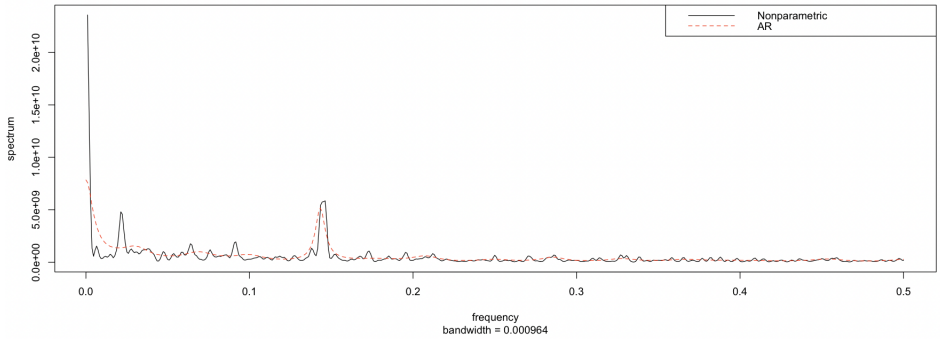


Figure 9: Frequency densities for cycle removed residuals

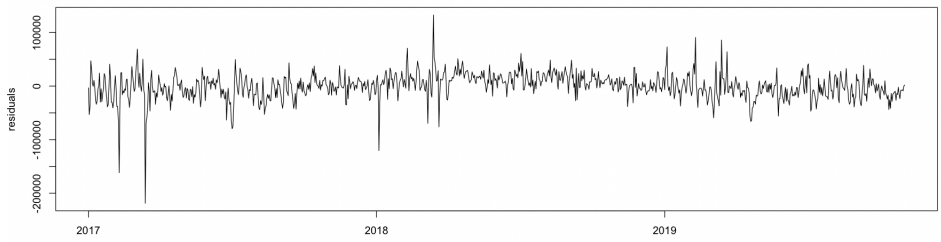


Figure 10: Final Residuals

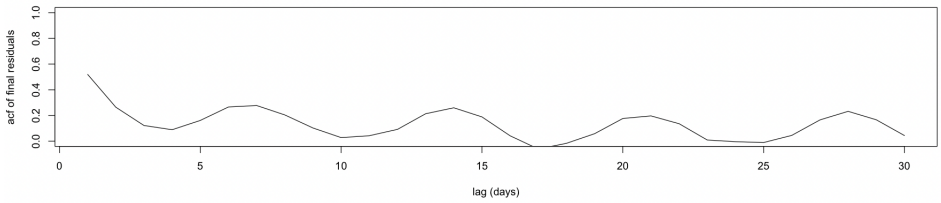


Figure 11: Final Residuals: ACF

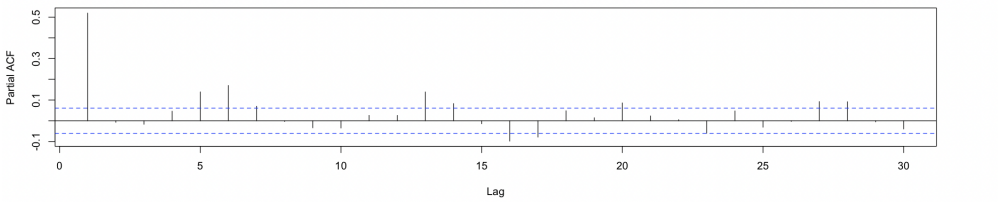
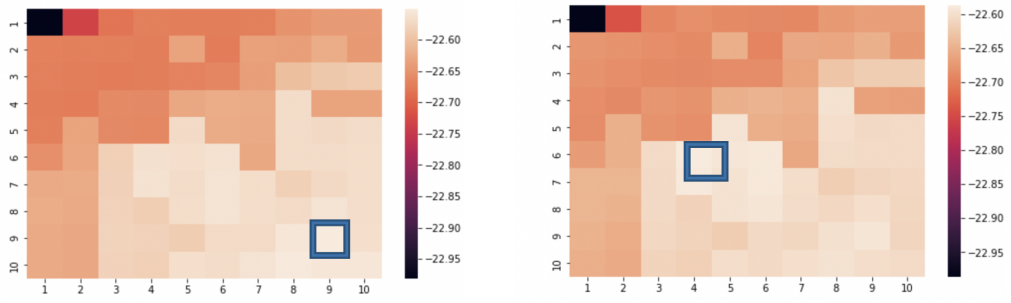


Figure 12: Final Residuals: PACF



(a) (b)

Figure 13: (a) AIC scores (b) Harmonic mean of AIC and BIC scores

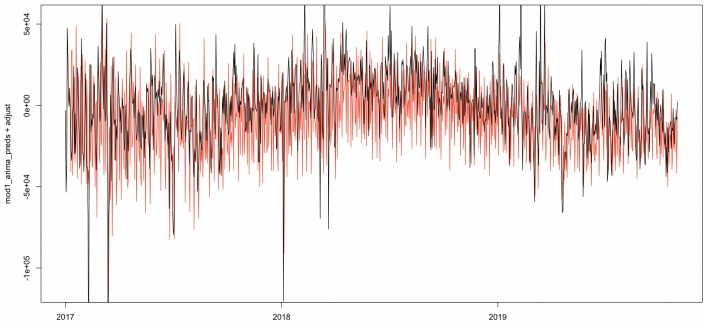


Figure 14: ARMA(9, 9): Train Residuals prediction (red) without level adjustment

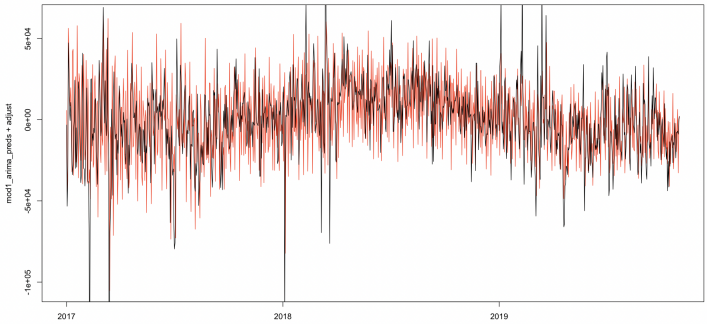


Figure 15: ARMA(9, 9): Train Residuals prediction (red) with level adjustment

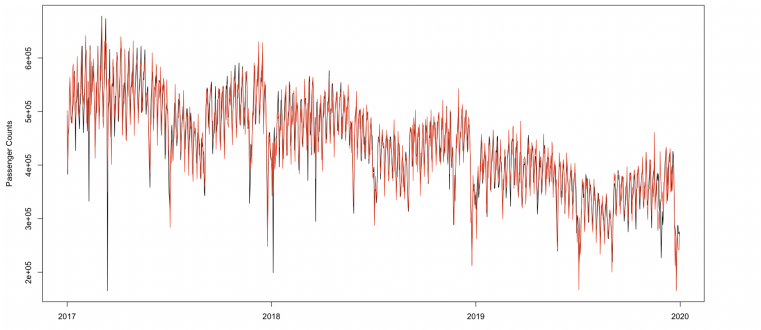


Figure 16: ARMA(9, 9): Train data predictions (red) with level adjustment

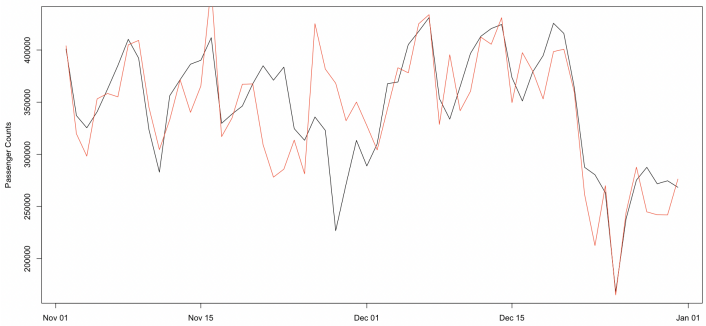


Figure 17: ARMA(9, 9): Test data predictions (red) with level adjustment

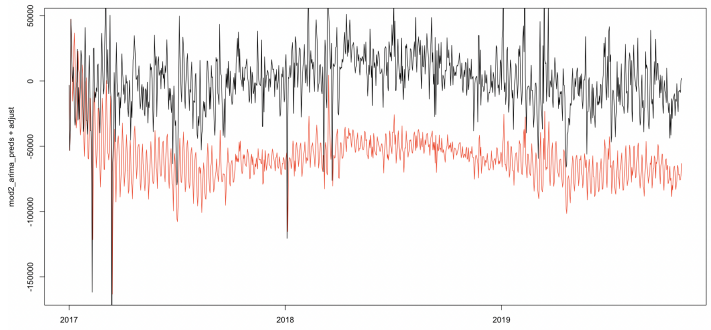


Figure 18: ARMA(6, 4): Train Residuals prediction (red) without level adjustment

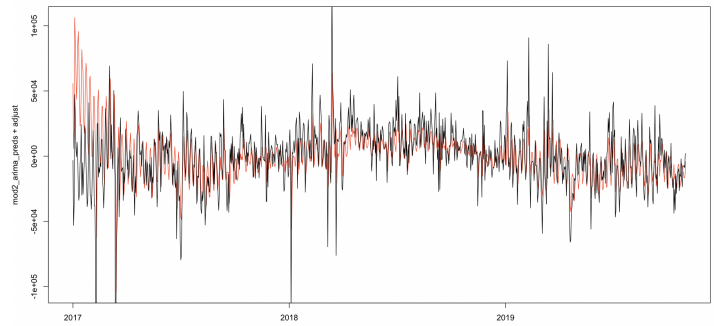


Figure 19: ARMA(6, 4): Train Residuals prediction (red) with level adjustment

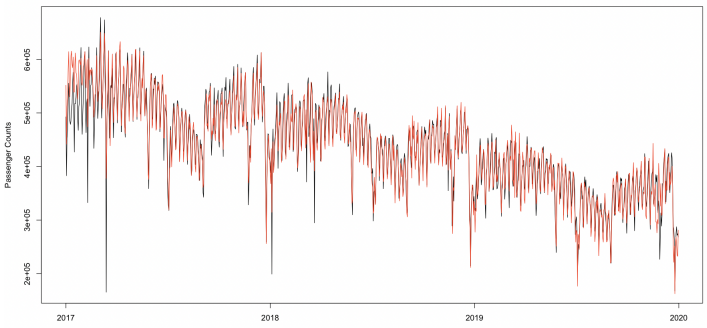


Figure 20: ARMA(6, 4): Train data predictions (red) with level adjustment

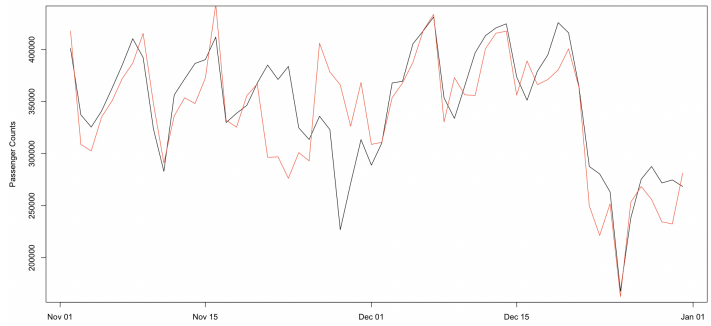


Figure 21: ARMA(6, 4): Test data predictions (red) with level adjustment

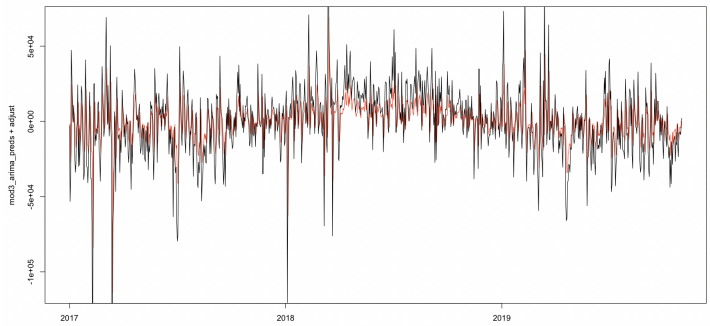


Figure 22: ARMA(1, 0): Train Residuals prediction (red)

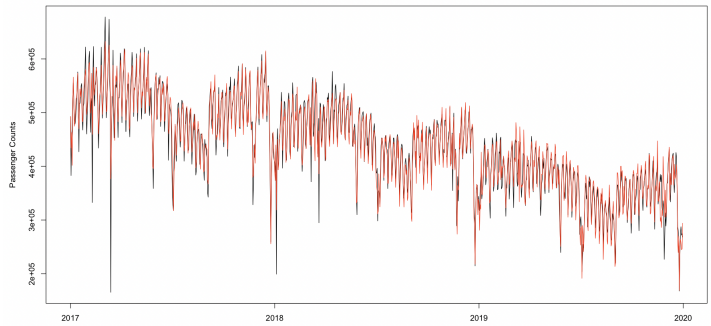


Figure 23: ARMA(1, 0): Train data predictions (red)

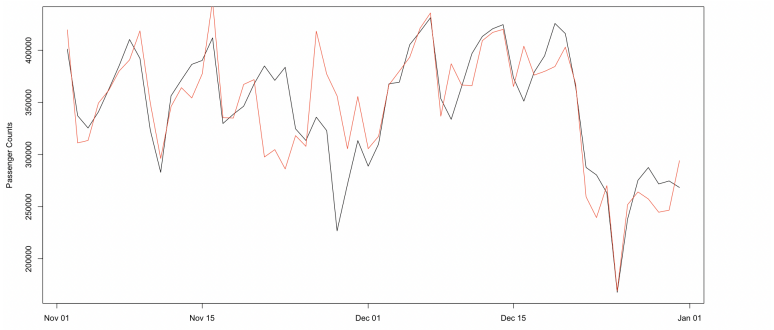


Figure 24: ARMA(1, 0): Test data predictions (red)

Model	Test RMSE
Mean of Train Data	113420.61
Linear Regression(LR)	56739.59
LR + 7-day cycle	52615.37
$X_t = X_{t-1}$	38643.50
AR(1) without cycle removal	36742.08

Table 3: Comparison with simpler models