STAT 485

Group 14
2023-12-10

I. Introduction

Our analysis revolves around the Air Traffic Passenger Statistics dataset sourced from DataSF, a comprehensive collection capturing air traffic activities within San Francisco. Spanning from 1999 to 2023 and updated quarterly, this dataset offers a rich repository of insights into air travel dynamics. With 15 variables and 34,878 observations, it delves into various aspects of air traffic, encompassing passenger statistics, flight counts, and other variables essential to understanding this domain.

Among the many variables present, we'll primarily focus on the passenger number, leveraging their significance in delineating trends and patterns within the dataset. Notably, our preliminary analysis using a line chart for airline passengers has revealed a distinct seasonal trend, hinting at the suitability of time series models to capture the underlying dynamics.

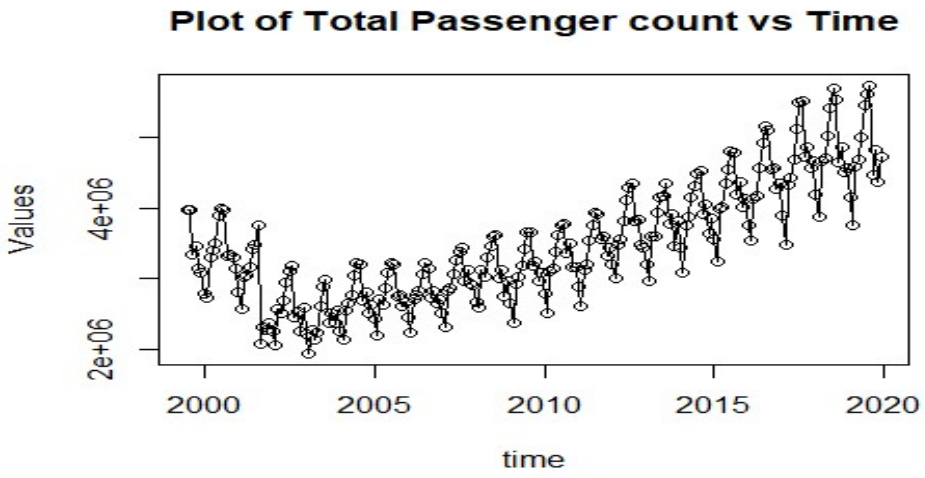
Our primary goal is a thorough exploration of the air traffic dataset, employing various statistical methods and specialized time series modeling techniques. We aim to achieve two key objectives:

Firstly, we seek to develop robust predictive models capable of forecasting future trends in monthly and yearly passenger numbers, as well as flight counts, over the upcoming 3 to 5 years. These models will leverage the dataset's historical information to generate accurate projections. Additionally, we aim to delve into the impact of seasonality on the airline industry. Employing time series analysis techniques tailored to this dataset, our exploration focuses on unraveling the cyclical nature of air traffic trends. By decoding seasonal variations, we aim to gain deeper insights into their influence on passenger behavior and flight demands.

II. Analysis

In our effort to analyze the dataset effectively, we began by organizing the data according to months. This strategic grouping facilitates a more detailed examination of trends and patterns, offering a comprehensive understanding of the data's temporal dynamics.

To ensure the accuracy and relevance of our analysis, a critical step involved the extraction of pre-COVID data. Recognizing the unique and disruptive nature of the COVID-19 pandemic, we opted to focus on a period preceding the pandemic. This deliberate choice stems from the understanding that data during the COVID era might not accurately represent usual or standard scenarios due to unprecedented global circumstances. By isolating the pre-COVID data, our goal is to uncover insights that reflect normal conditions, providing a more reliable foundation for our analysis.



1. **White Noise assumption - Residual analysis:**

To investigate if there was seasonality in our dataset, we started by performing the residual analysis to identify if the stochastic component was white noise. The unobserved process or the predicted residual behaved like independent random variables, the true stochastic component would behave similarly.

In analyzing the residuals, we first plotted the residuals over time in Exhibit [2.1]. The graph showed a clear upward trend with empty upper left and bottom right corners. Indicating a possibility of linear or quadratic patterns. The plot of the residuals against the fitted trend of seasonal means in Exhibit [2.2] also showed an apparent positive linear pattern. These failed to align with our assumptions of white noise residuals, where it should show no perceivable trends or patterns.

Next, we analyzed the standardized residuals for normality. Exhibit [2.3] was a histogram of the residuals from the means model above, which was asymmetrical and skewed to the left. The quantile-quantile (QQ) plot in Exhibit [2.4] further showed a considerable deviation from the 45-degree reference line in the first and the third normal quartiles. Our residuals did not appear to be normally distributed.

Lastly, we plotted a correlogram in Exhibit [2.5] to test the autocorrelation function of the standardized residuals. The autocorrelations at all lags exceeded two standard errors above zero. With all the evidence found above, we could conclude that the stochastic component was not a white noise process.

2. **Detrend the time series:**

The graphs of total monthly passengers over time showed possible linear or quadratic trends. We first fitted the linear regression function for the time series to get function Y = -215367900 + 108895*Xt. The plot of the residuals against fitted trend in Exhibit [3.1] showed a possible quadratic

pattern. Next, we plotted a histogram of predicted residuals in Exhibit[3.2] and a QQ-plot in Exhibit[3.3] to check for normality. The histogram was left-skewed and the mode was not at zero, while the QQ-plot had some deviations from the 45-degree line.

In addition, the acf plot in Exhibit[3.4] displayed a relatively high positive autocorrelation of 0.780 at lag 1. This suggests that the residual is strongly correlated with the previous residual value at lag 1. At lag 12, there was a relatively high positive autocorrelation of 0.765, indicating a potential seasonal pattern in the residuals. For the pacf plot in Exhibit[3.5], there was a relatively high positive partial autocorrelation of 0.780, showing a strong direct relationship between the current residual value and its lag 1 value, after removing the effects of intervening lags. At lag 7, there was a significant positive partial autocorrelation of 0.577. This suggested a direct relationship between the current residual value and its lag 7 value, indicating potential trend at this lag. By calculating the standard deviation of residuals of the linear model (i.e. 538284.1) , the autocorrelations had large-sample standard deviations, which explained the discrepancy.

For the quadratic model, we generated the quadratic regression function of the time series Yt = 3.998e+10 - 3.989e+07Xt + 9.951e+03Xt^2. Exhibit [3.5] of the residuals against fitted trend showed a cluster at the beginning. Next, the plot of the histogram of predicted residuals in Exhibit[3.6] showed a bell shape. The QQ-plot in Exhibit[3.7] additionally indicated that most values lied on the 45-degree line, confirming normality. Moreover, there was a positive autocorrelation of 0.684 for the acf of predicted residuals in Exhibit[3.8], indicating a moderate positive correlation between the current residual value and its lag 1 value at lag 1. The autocorrelations gradually decreased as the lag increased, with values ranging from 0.438 to -0.573. The autocorrelations after lag 10 tended to be smaller and fluctuated around 0, suggesting a weaker or no significant correlation between the residuals and their further lagged values. By calculating the standard deviation of residuals of the linear model (i.e. 438413.1), the autocorrelations had large-sample standard deviations, which explained the discrepancy.

Detrending the data allowed us to see any potential subtrends. First, we detrended the time series by a linear model. From the acf using a linear model in Exhibit[3.9], there was still a quadratic curve of the trend. The autocorrelations appeared to fluctuate around zero, with some positive and negative values at various lags. Autocorrelation values closer to 1 or -1 indicated a stronger linear relationship, while values closer to 0 suggested a weaker or no linear relationship. From the pacf of the detrended using linear model in Exhibit[3.10], the partial autocorrelation at lag 1 was 0.0833. This value represented the correlation between the series and itself at the same lag. It indicated a weak positive relationship between the current observation and itself, which was expected since it was the same value.

The partial autocorrelations at lag 2 to lag 10 were all positive and gradually increased. This suggested a positive relationship between the current observation and its lagged values at these lags. The increasing values indicated that the impact of the lagged values on the current observation became stronger as the lag increased. From the above acf and pacf, the first 9 lags were the same.

Based on the analysis of the time series data from 2020 to 2023, which was after the pandemic period, it was evident that the total number of air passengers had been steadily increasing over time. However, choosing a quadratic model to represent this trend would not be appropriate as a quadratic curve implied a rapid and unrealistic increase of passengers at certain points in time, which did not align with real-world scenarios. For instance, it was illogical to expect a large number of passengers in the past, which the quadratic model indicated. Therefore, while a linear model could not capture all existing trends, it was more reasonable to detrend an upward linear model from our time series for a more realistic approach.

The residuals of the linear detrended model showed their ACF exponentially decay, while their PACF somewhat fluctuated over 0, thus, the process was possibly an autoregressive process or an autoregressive-moving average process.

3. Model specification/selection:

Exhibit [4.0] displayed the extended ACF of the residuals, where there was a cluster of "x" marks at the beginning, suggesting a high-order AR or a MA component. There were some significant lags at high order, indicating a possible need for differencing. Based on the models we learned in this course, we would test the first-order and second-order models.

We first fitted the detrended time series to the first-order autoregressive process AR(1). The plot for model predicted residuals over time in Exhibit [4.1] now fluctuated over mean, but with a clear seasonal pattern. The residuals of Exhibit [4.2] plotted against the fitted trend tended to hang together and fluctuated around 0 from -50,000 to 50,000. The residuals were also mostly cluttered in earlier times while having a wider range in later times. These violated the white noise assumptions.

The histogram in Exhibit [4.3] was not symmetric and the QQ plot in Exhibit [4.4] deviated a lot from the reference line. The graphs rejected the normal distribution assumption of the white noise process. The ACF of the residuals in Exhibit [4.5] was different from 0 only at lag 6 and lag 12, while the PACF in Exhibit [4.6] was different from 0 at lag 6, 12, and 18, indicating seasonal patterns.

We then fitted the time series to AR(2). The residual analysis for AR(2) showed very similar results to AR(1), where the residuals over time had seasonal patterns and were cluttered, while the standardized residuals were not normally distributed. The ACF and PACF of the residuals showed similar behaviour to the AR(1) model. As there was a seasonal pattern, implying a non-stationary time series, the AR models were insufficient to fit our time series. Since the time series seemed to be non-stationary, thus, our next potential improvements were SARIMA models or ARIMA(p,d,q)x(P,D,Q)12, applying seasonal differencing.

To find the best model, we used the dynamics method, where we found the parameters by trial and error. The best model that was parsimonious, we arrived at after trying all possible first-order methods was ARIMA(1,0,1)(0,1,1)12. The ACF and PACF in Exhibit [4.11] and [4.12] showed that they were within two standard errors, except for lag 1.25. This might be due to the randomness. The EACF in Exhibit [4.13] also fitted with our assumption of the model. The plot of residuals against Time shown in Exhibit [4.14] illustrated a random fluctuation around 0, without any discernible trend. The histogram and QQ plot in Exhibit [4.15] and [4.16] confirmed the normality. Thus, we could conclude this model was a good fit for our time series.

Our next criteria to choose the best model was based on the accuracy of the prediction. As COVID-19 was a rare event, we considered it as an outlier. Based on the linear trend, it was reasonable to assume the air traffic industry to continue the current upward trend in the next three years. If we took COVID-19 into account, the passengers were also likely to resume travelling after the pandemic as there was a fast recovery from 2020 to 2023, suggesting the air traffic industry to recover from the recession and follow the linear trend. The prediction of ARIMA(1,0,1)(0,1,1)12 in Exhibit [4.18] seemed to be much more accurate and realistic, aligning with our assumptions. As such, we chose ARIMA(1,0,1)(0,1,1)12.

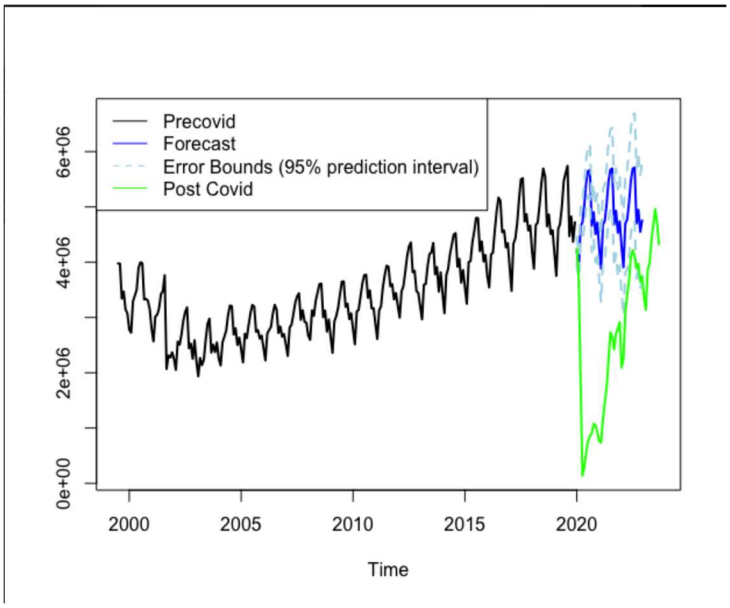


Exhibit 4.18

III. Conclusion

The comprehensive analysis of the Air Traffic Passenger Statistics in San Francisco has provided us with valuable insights into the dynamics of the local air traffic. Focusing on the number of passengers in the pre-COVID period, we have constructed a reliable model that forecasts the future air traffic trend in scenarios where the pandemic did not happen.

Our analysis showed significant seasonal patterns in the time series, resulting in a non-stationary process. After our exploration of various AR and seasonal ARIMA models, we discovered the ARIMA(1,0,1)(0,1,1)12 model to be the best fit based on the course materials and the principle of parsimony. This model has effectively accounted for the seasonal patterns and its residuals have shown a reasonable alignment with the white noise assumption. Most importantly, its prediction was robust to the outlier caused by the pandemic. The model suggested a continuation of the pre-pandemic trend in air traffic, assuming the industry's recovery as observed from 2020 to 2023.

In conclusion, based on these insights, airline companies can safely maintain their operations despite the recent recession to accommodate future air traffic demands without worrying about losses. However, it is important to remain attentive of potential disruptions or significant changes in travel patterns and passenger behaviors, as such time series analysis and prediction are only reliable in the short-term period.

APPENDIX

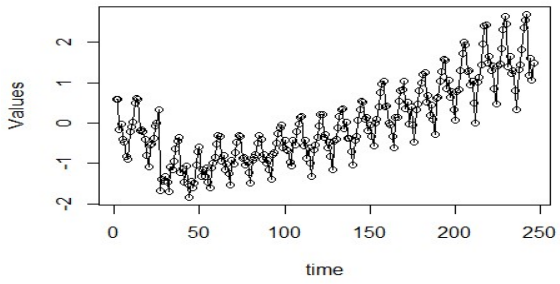


Exhibit [2.1]

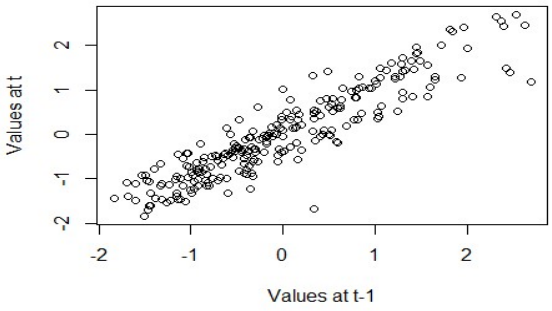


Exhibit [2.2]

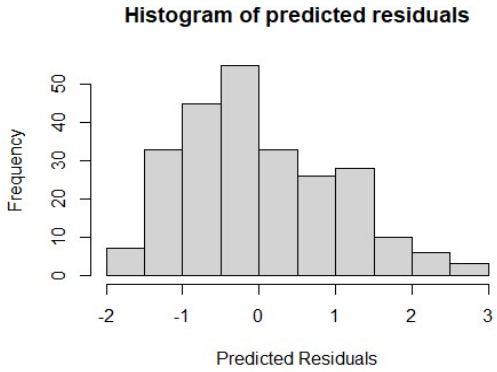


Exhibit [2.3]

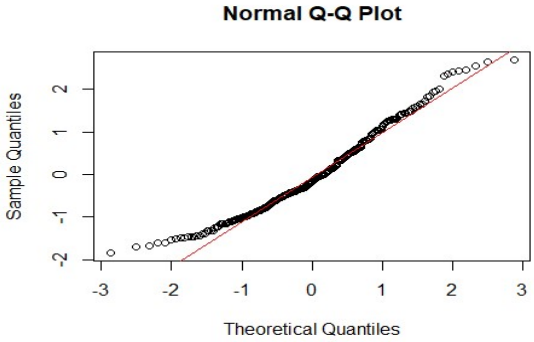


Exhibit [2.4]

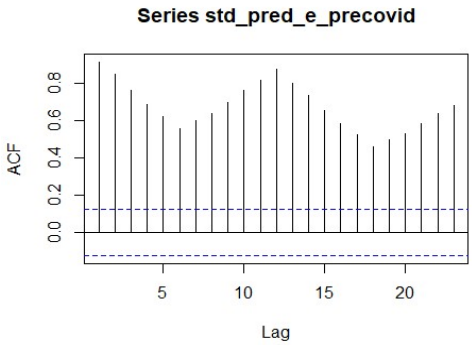


Exhibit [2.5]

```
## Call:
## lm(formula = prets ~ time(prets))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1051431  -384590   -97468   279971  1609092
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -215367900   11679047  -18.44   <2e-16 ***
## time(prets)     108895       5811   18.74   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 539400 on 244 degrees of freedom
## Multiple R-squared:   0.59,  Adjusted R-squared:  0.5883
## F-statistic: 351.1 on 1 and 244 DF,  p-value: < 2.2e-16
```

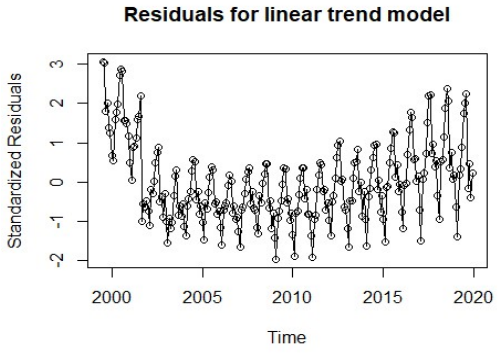


Exhibit [3.0]

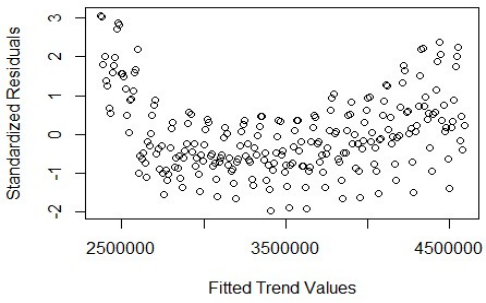


Exhibit [3.1]

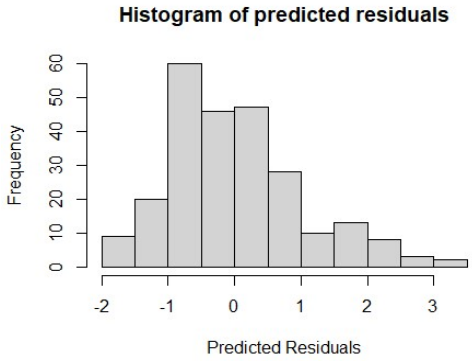


Exhibit [3.2]

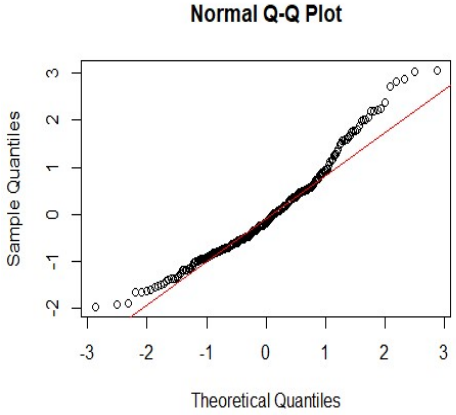


Exhibit [3.3]

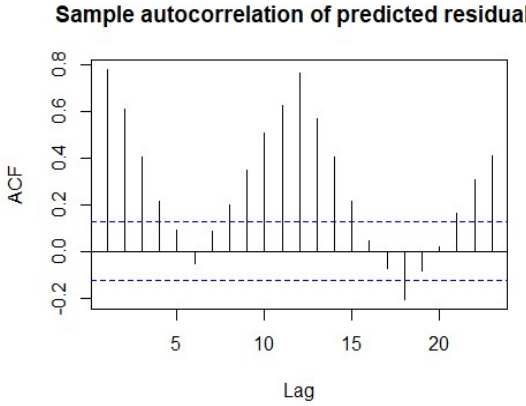


Exhibit [3.4]

```
## Autocorrelations of series 'rstudent(model1)', by lag
##
##     1     2     3     4     5     6     7     8     9    10    11
## 0.780 0.609 0.405 0.217 0.089 -0.052 0.084 0.197 0.350 0.505 0.623
##    12    13    14    15    16    17    18    19    20    21    22
## 0.765 0.571 0.407 0.215 0.046 -0.075 -0.207 -0.084 0.020 0.162 0.308
```

```
##     23
##  0.410
acf(prets,plot=FALSE)            #Print r_k for series (check: different from
above)
##
## Autocorrelations of series 'prets', by lag
##
## 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167
##  0.912  0.847  0.760  0.683  0.621  0.554  0.599  0.637  0.697  0.760  0.813
## 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
##  0.877  0.799  0.735  0.652  0.580  0.522  0.457  0.496  0.529  0.582  0.637
## 1.9167
##  0.680
## lm(formula = prets ~ t + t2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1273525  -295632   -38152   302521  1025149
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.998e+10  3.619e+09   11.04   <2e-16 ***
## t           -3.989e+07  3.602e+06  -11.07   <2e-16 ***
## t2           9.951e+03  8.961e+02   11.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 440200 on 243 degrees of freedom
## Multiple R-squared:  0.728,  Adjusted R-squared:  0.7258
## F-statistic: 325.2 on 2 and 243 DF,  p-value: < 2.2e-16

##      1      2      3      4      5      6      7      8      9     10     11
##  0.684  0.438  0.138 -0.140 -0.312 -0.507 -0.287 -0.113  0.127  0.371  0.548
##     12     13     14     15     16     17     18     19     20     21     22
##  0.774  0.500  0.272 -0.001 -0.244 -0.401 -0.573 -0.368 -0.202  0.024  0.255
##     23
##  0.411
## Autocorrelations of series 'prets', by lag
##
## 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167
##  0.912  0.847  0.760  0.683  0.621  0.554  0.599  0.637  0.697  0.760  0.813
## 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
##  0.877  0.799  0.735  0.652  0.580  0.522  0.457  0.496  0.529  0.582  0.637
## 1.9167
##  0.680
## Autocorrelations of series 'detrend_tslm', by lag
##
## 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333
##  0.781  0.610  0.404  0.215  0.087 -0.054  0.083  0.195  0.349  0.506
```

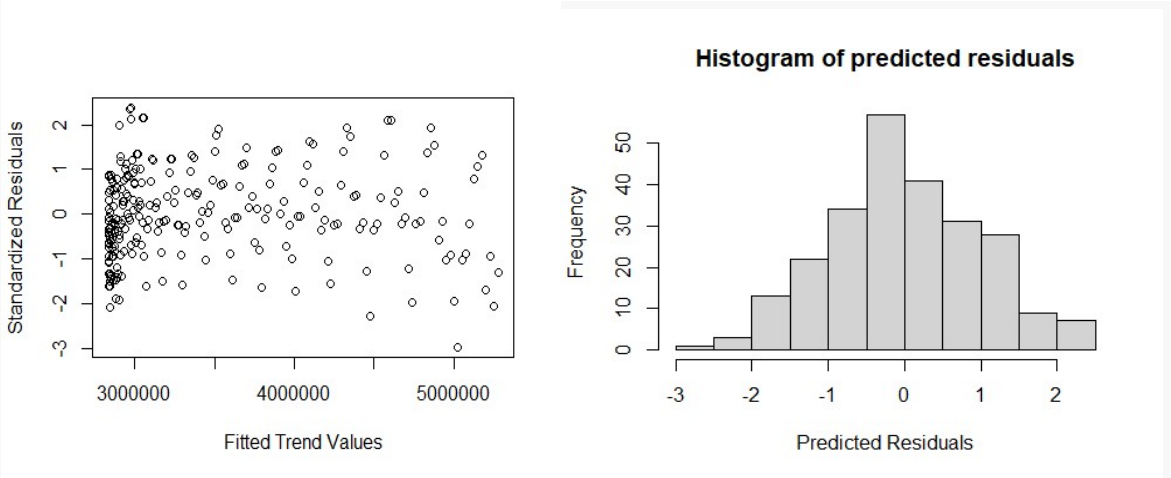


Exhibit [3.5] *Exhibit [3.6]*

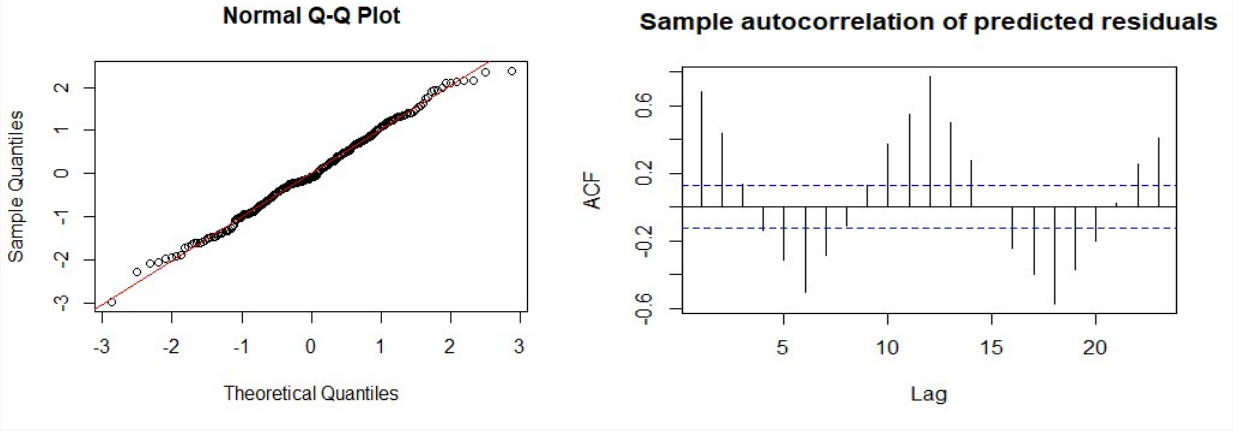


Exhibit [3.7] *Exhibit [3.8]*

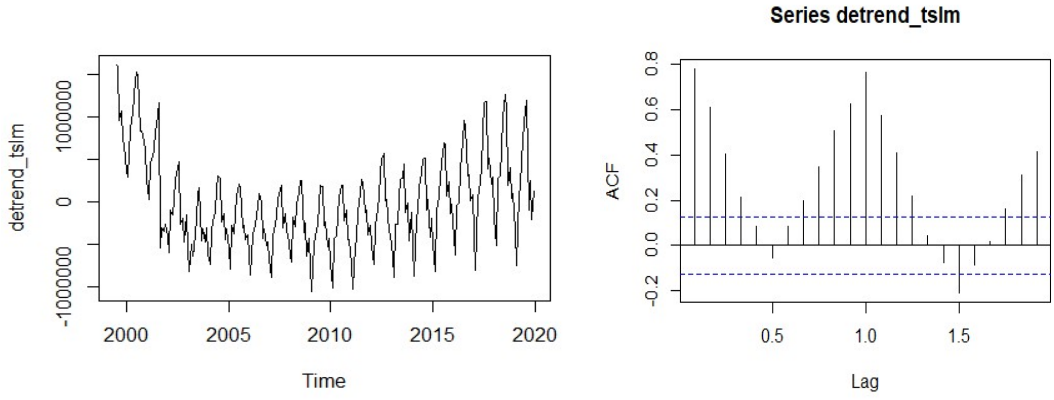
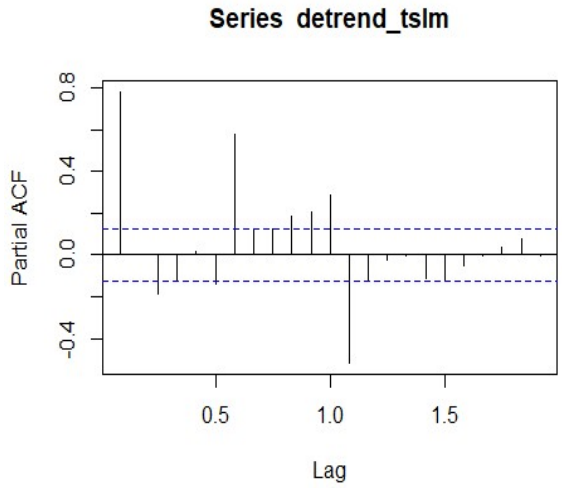
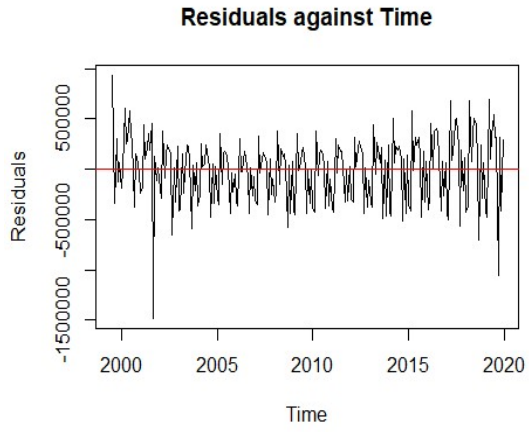


Exhibit [3.9] *Exhibit [3.10.]*




```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x x x o o o x x x x  x  x  x
## 1 o x x o o o o o o x o  x  o  x
## 2 o x o o o o x o o x o  x  o  x
## 3 x x o o o o x o o o o  x  x  o
## 4 x x o o o x x o o o o  x  o  o
## 5 x o o x o x x x o o o  x  o  x
## 6 x o x x o x o x o o x  x  x  o
## 7 x x x x x x o o o x o  x  o  o
## Call:
## arima(x = detrend_tslm, order = c(1, 0, 0))
##
## Coefficients:
##          ar1  intercept
##       0.8069   28548.13
## s.e.  0.0389  105956.81
##
## sigma^2 estimated as 1.06e+11:  log likelihood = -3472.17,  aic = 6948.34
```



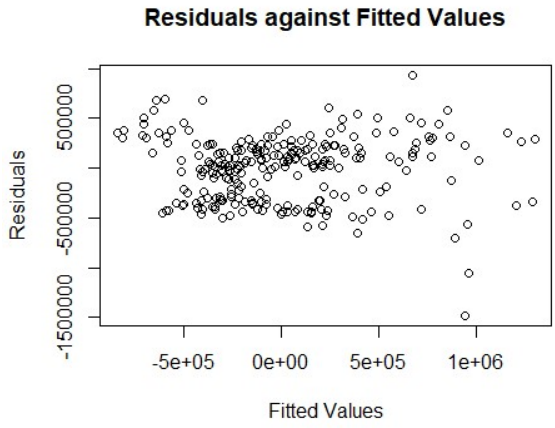


Exhibit [4.1]

Exhibit [4.2]

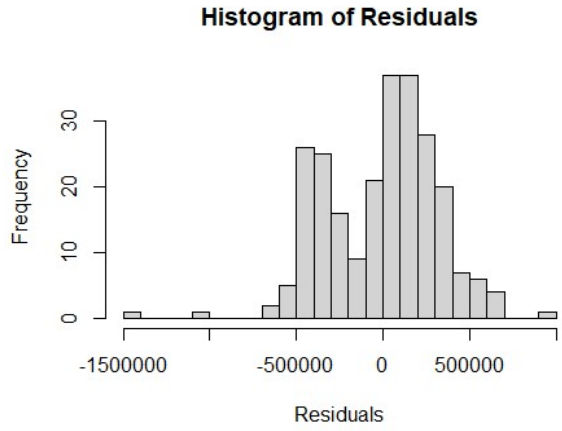


Exhibit [4.3]

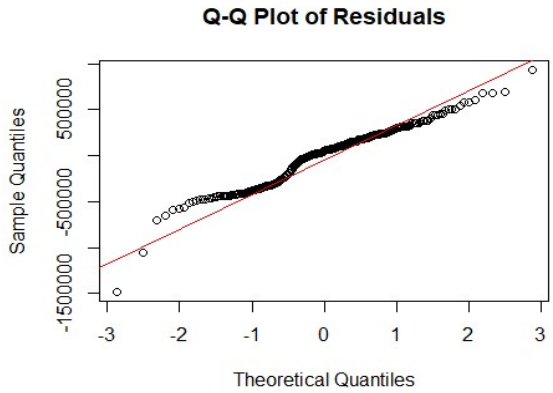


Exhibit [4.4]

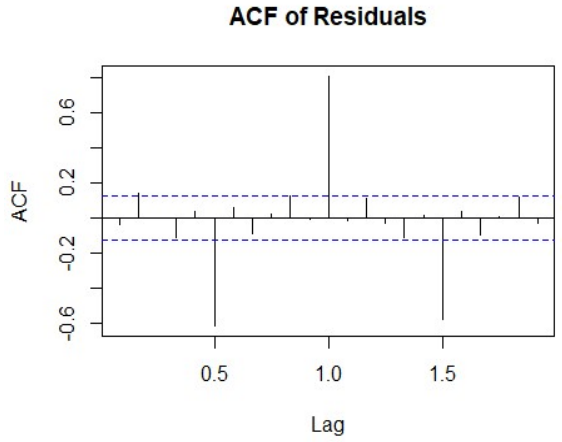


Exhibit [4.5]

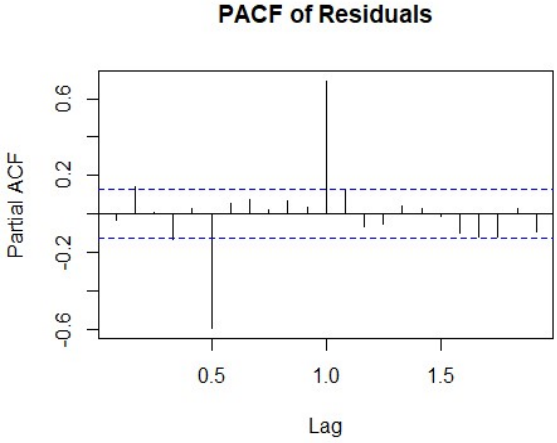
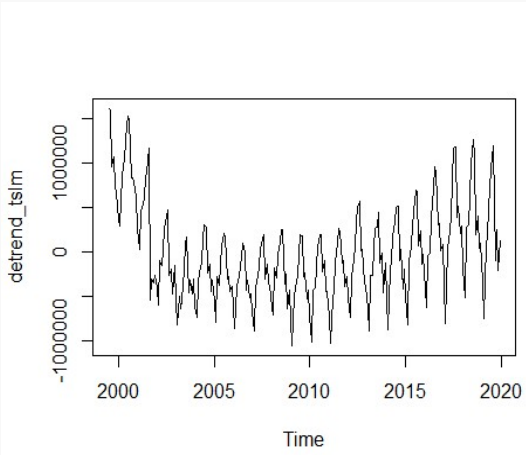
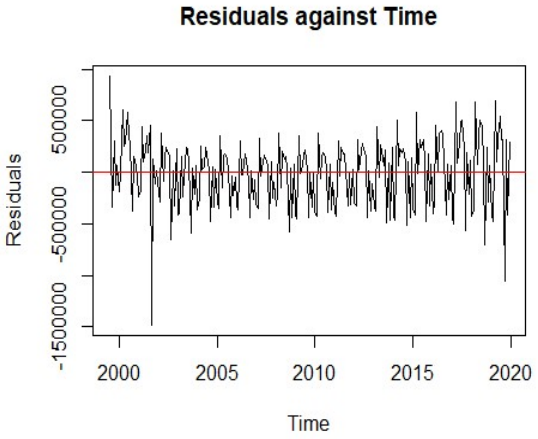


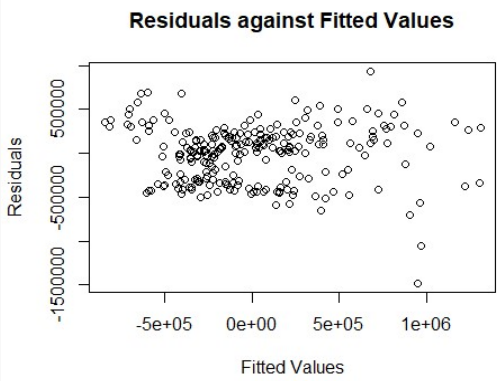
Exhibit [4.6]

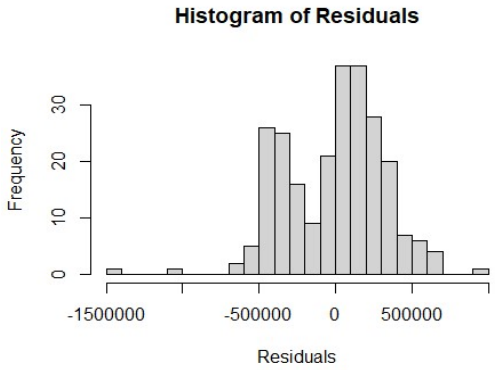
```
## Autocorrelations of series 'residuals(AR)', by lag
##
## 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167
## -0.034  0.145 -0.001 -0.113  0.037 -0.611  0.064 -0.086  0.023  0.130 -0.009
## 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
##  0.809 -0.016  0.112 -0.030 -0.109  0.015 -0.576  0.041 -0.094  0.006  0.124
## 1.9167
## -0.030
##
## Autocorrelations of series 'prets', by lag
##
## 0.0833 0.1667 0.2500 0.3333 0.4167 0.5000 0.5833 0.6667 0.7500 0.8333 0.9167
##  0.912  0.847  0.760  0.683  0.621  0.554  0.599  0.637  0.697  0.760  0.813
## 1.0000 1.0833 1.1667 1.2500 1.3333 1.4167 1.5000 1.5833 1.6667 1.7500 1.8333
##  0.877  0.799  0.735  0.652  0.580  0.522  0.457  0.496  0.529  0.582  0.637
## 1.9167
##  0.680
```

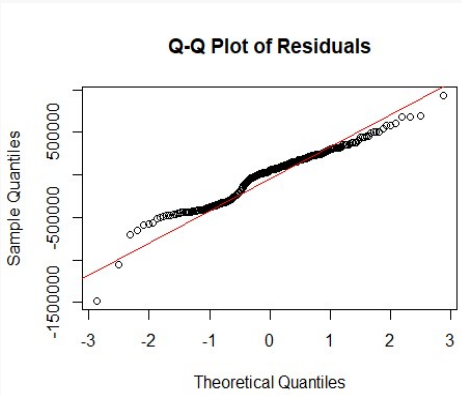
```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x x x x x x x x x  x  x  x
## 1 o x o o o x o o o o o  x  o  x
## 2 x x o o o x x o o o o  x  x  x
## 3 x x o o o x o x o o o  x  o  x
## 4 x x o o o x o x o o o  x  o  x
## 5 x x o x o x x o o o o  x  x  x
## 6 o x o x o x o o o o o  x  o  x
## 7 x x x x x x o o o x o  x  o  x
## Call:
## arima(x = detrend_tslm, order = c(2, 0, 0))
##
## Coefficients:
##          ar1     ar2  intercept
##       0.7926  0.0184   31849.29
## s.e.  0.0637  0.0649  108481.50
##
## sigma^2 estimated as 1.06e+11:  log likelihood = -3472.13,  aic = 695
```

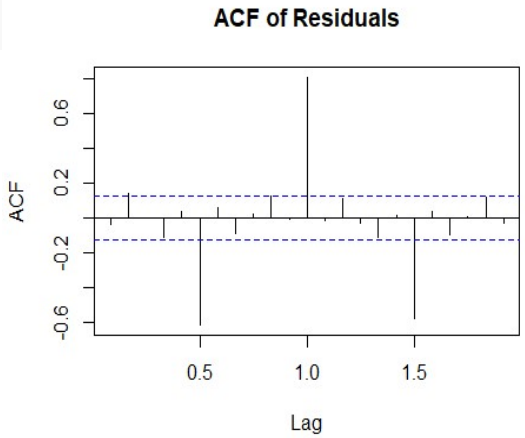


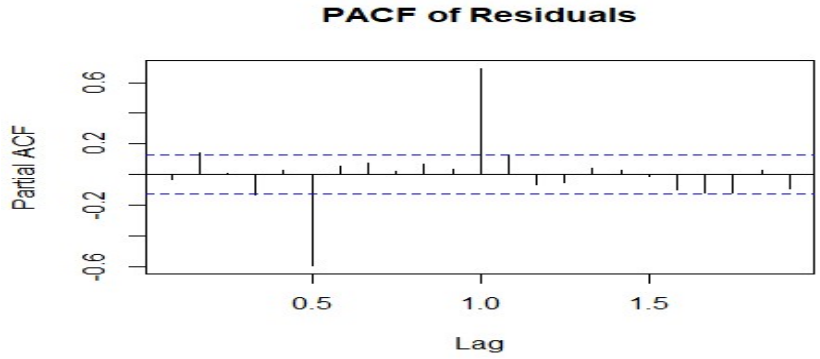






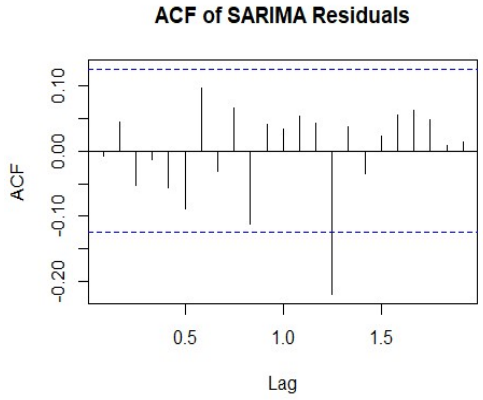







```
## Call:
## arima(x = prets, order = c(1, 0, 1), seasonal = list(order = c(0, 1, 1), period
= 12))
##
## Coefficients:
##          ar1      ma1     sma1
##       0.9723  -0.3107  -0.5345
## s.e.  0.0163   0.0660   0.0687
##
## sigma^2 estimated as 1.481e+10:  log likelihood = -3074.71,  aic = 6155.42
##
## Training set error measures:
## Warning in trainingaccuracy(object, test, d, D): test elements must be within
## sample
```

```
##                ME RMSE MAE MPE MAPE
## Training set NaN  NaN NaN NaN  NaN
```



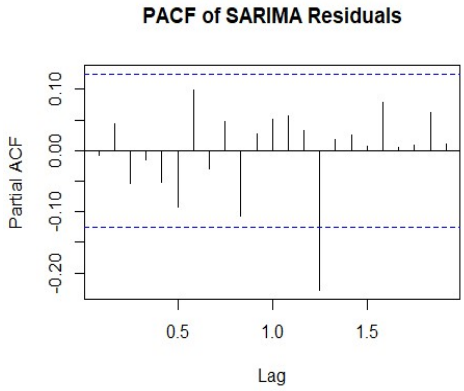
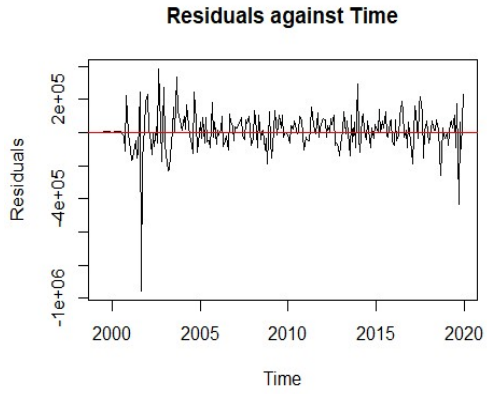


Exhibit [4.11]

Exhibit []4.12

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 o o o o o o o o o o  o  o  o
## 1 x o o o o o o o o o  o  o  o
## 2 x x o o o o o o o o  o  o  o
## 3 x x o o o o o o o o  o  o  o
## 4 x x o x o o o o o o  o  o  o
## 5 x o x x x o o o o o  o  o  o
## 6 x x x o x o o o o o  o  o  o
## 7 x x x o x x o o o o  o  o  o
```

Exhibit [4.13]



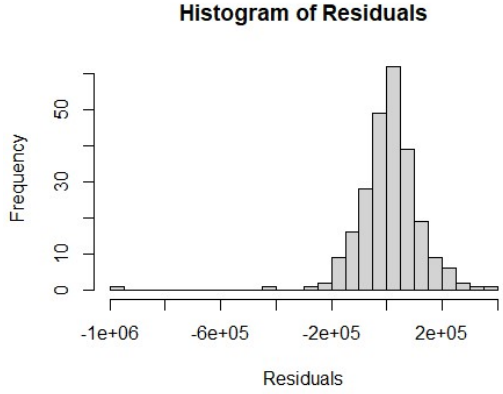


Exhibit [4.14]

Exhibit [4.15]

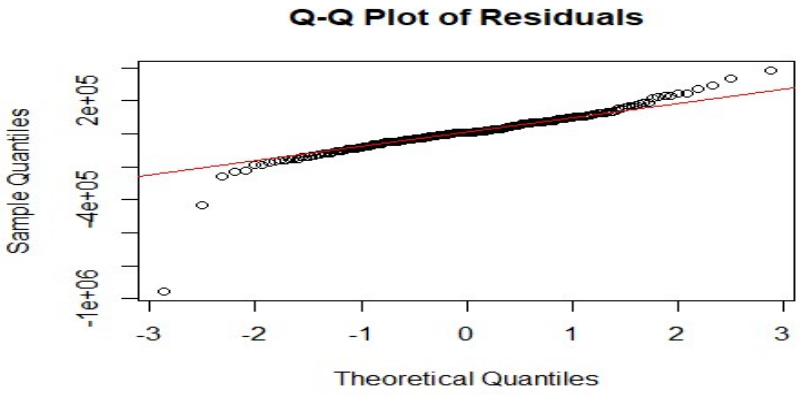

Q-Q Plot of Residuals
Sample Quantiles
-1e+06
-4e+05
2e+05
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Exhibit [4.16]