Passenger Rate:

Times Series Analysis

PSTAT 174

Professor Bapat

# 

# Sarah Foster, Kalvin Goode,Tony Khoury, Jia Ma, Timothy Nguyen, UCSB

**Table of Contents 1**

**Abstract 2**

**Introduction 2**

**Exploratory Analysis 2**

Preliminary Data Exploration 2

Decomposition Model 4

**Data Transformation 4**

**Model Identification and Estimation 6**

Preliminary Model Identification 6

Model Selection 7

Model Estimation 8

**Diagnostics 9**

Normality Check 9

Detection of Serial Correlation 10

Heteroscedasticity 10

**Forecasting 11**

**Future Study 13**

**Conclusion 13**

**References 14**

**Appendix 14**

**Abstract**

The objective of our project is to predict the number of monthly international airline passengers. Having this information is vital for international airlines so that they may make the necessary decisions to grow their business. Using time series analysis, we built an appropriate SARIMA model with the least amount of parameters and the lowest AICc score based on our stationary data. This stationary data was acquired through transforming and differencing our original data to eliminate its trend and seasonality. In addition, our model went through several diagnostic checks, and we forecast the number of international airline passengers for an entire year. After checking the validity from the original dataset, our predicted values are within 95% confidence interval.

**I. Introduction**

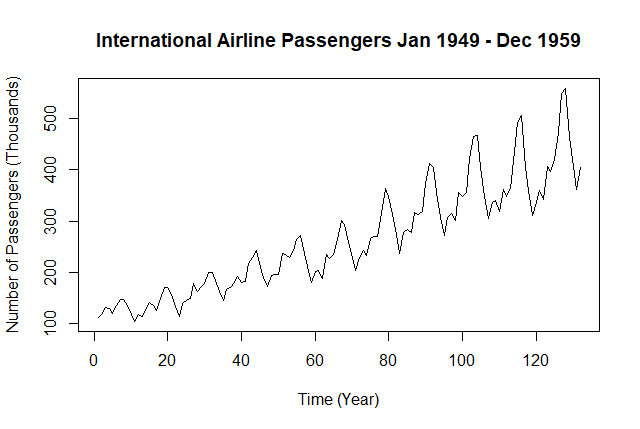
The invention of air travel has changed the world dramatically. People can easily travel to other countries without much effort. Airline industries have been consistently growing worldwide as air travel becomes more common. In addition, the airline industry is also a competitive business environment. Many airlines tried to reduce travel cost by adding available passengers per flight. One important factor of airline to the economic benefit is shortening the connections between cities. This increases the flow of goods, people and capital to boost the overall development of the world. In this report, we are interested in the rate of passengers in international air travel. This data is useful as the airline companies can estimate potential future growth in a short period, so that they can adjust the ticket price and the fuel consumption accordingly. The data we found has an interval of 12 years between 1949 and 1960. The data is taken from Time Series Data Library. Plotting the data set in R, we see that it has an upward trend with increasing variance. In addition, it is obvious that there is a seasonal pattern for each year. Therefore, we difference the data at lag 12 with a log transformation and then lag 1, and the modified data has a constant variable throughout all years. Using time series techniques learned from class, we are able to have a some suitable models to the rate of passengers. After valuation on models with different test, we decide the best model for the international passenger rate is SARIMA(3,1,1)X(1,1,0)12. Lastly, we forecast the rate of passenger for the whole year in 1960. Our true data points for the year of 1960 fall within our 95% confidence interval based on our predicted data; therefore, we conclude our model is satisfied.O

**II. Exploratory Analysis**

**Preliminary Data Exploration**

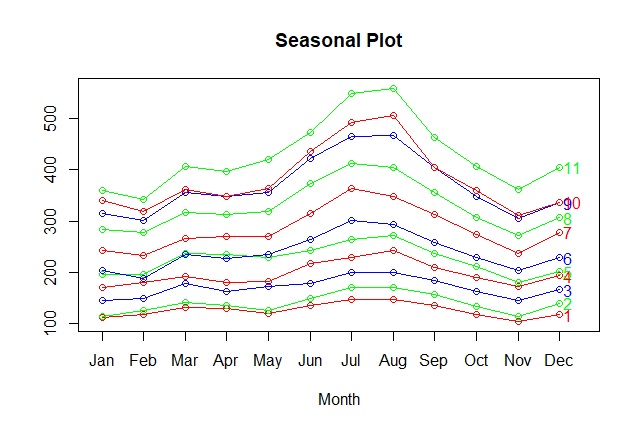
The dataset that we will be analyzing is the monthly total number of international airline passengers from January 1949 to December 1960. The two variables are the monthly dates and the number of passengers in thousands. There are 144 observations in total, but we will withhold the last 12 observations as test data to compare with our forecasted values. Therefore, we will be using 132 training data points for our time series analysis.

The first step is to put the data in a time series form and then plot it to check for stationarity. The time series plot with 132 observations is shown below in **Figure 2.1**.



**Figure 2.1**

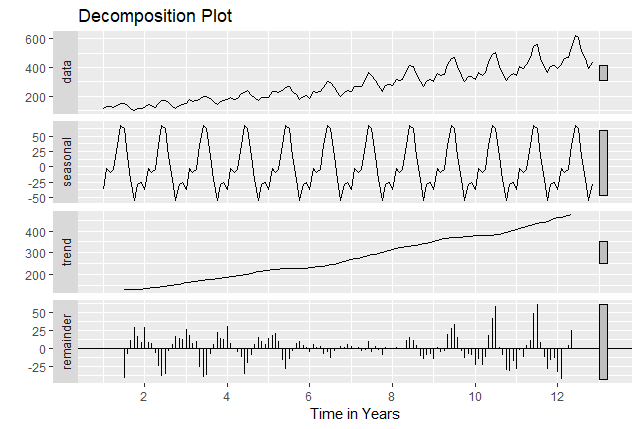
Based on the plot, we can observe that there is an upward trend. Additionally, the spikes that occur at equal time intervals indicate that seasonality is another factor that is contributing to non-stationarity. To prove that there is a seasonal pattern in the time series plot, we draw a seasonal plot that is shown below in **Figure 2.2.**

****

**Figure 2.2**

From the seasonal plot, we can observe that there are generally more international airline passengers in July and August and less in February and November. This makes sense because summer vacation is during June, July, and August, and that is around the time when families travel together. In February and November, students are still in school and the weather is colder, therefore not many people will travel.

**Decomposition Model**   
 The decomposition model Yt = mt + st + St is used to further show that there is an upward trend and seasonality in the time series dataset. The model contains: Yt, the dataset, mt, the trend component, st, the seasonal component and St, the stationary process. The decomposition plot is shown below in **Figure 2.3**.

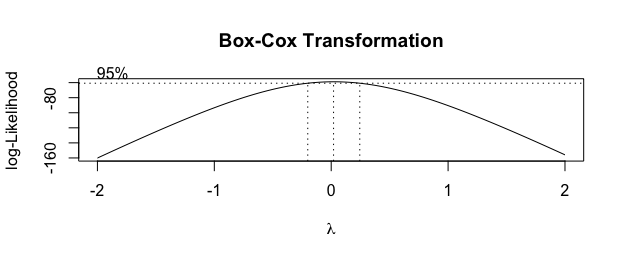
****

**Figure 2.3**

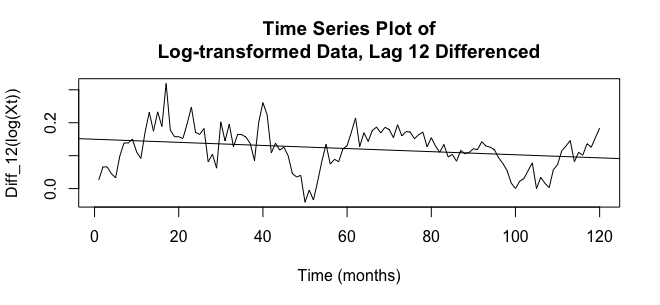
From the decomposition plot, we can clearly see the existence of seasonality since there are high and low spikes on the graph. Additionally, we can also see that there is an upward trend in our dataset. With the seasonal and decomposition plot, we can conclude that the time series dataset is not stationary. To make it stationary, we need to take transformations and difference the data.

**III. Data Transformation**

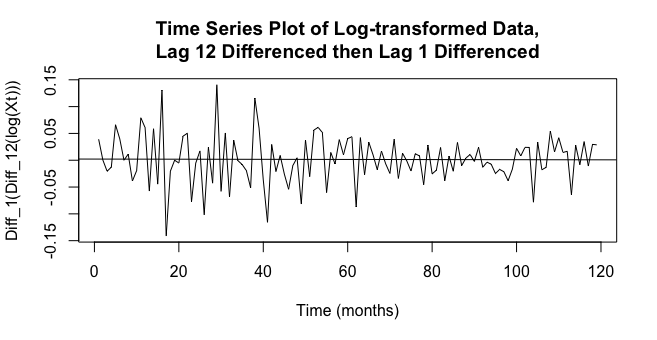
We first applied a Box Cox power transformation to the training data to obtain a λ. Plotting the Box Cox graph (**Figure 3.1**), we observe that 0 is in the confidence interval. Accordingly, we apply a log transform to the training data.

**Figure 3.1**

Next, we wanted to remove seasonality from the data, so we differenced the transformed data. Because the seasonality of the data is 12 due to a year having 12 months, we chose a lag of 12 for the differencing. The seasonality of 12 can also be seen in the decomposition plot. As shown in **Figure 3.2**, the lag 12 differencing removed the seasonality from the data and the variance is 0.00404; however, we draw a best fit line through the differenced data to reveal a downward trend.

**Figure 3.2**

To remove this trend, we apply a lag 1 difference. **Figure 3.3** shows the resulting twice differenced data along with a horizontal best fit line. This data has a variance of 0.00203, lower than 0.00404 and thus indicative that the second difference was appropriate. We apply yet another lag 1 difference to the twice differenced data and obtain a variance of 0.00537, which is a substantial increase from 0.00203 indicating that the second lag 1 difference is unnecessary. Thus, we move forward with the twice differenced data shown in **Figure 3.3**.

**Figure 3.3**

**IV. Model Identification and Estimation**

Our data on international airline passengers from 1949 to 1959 has a seasonal component so a *SARIMA* model would be best for analyzing and forecasting our data. A general *SARIMA* model follows the below structure:

*SARIMA (p, d, q) x (P, D, Q)s*

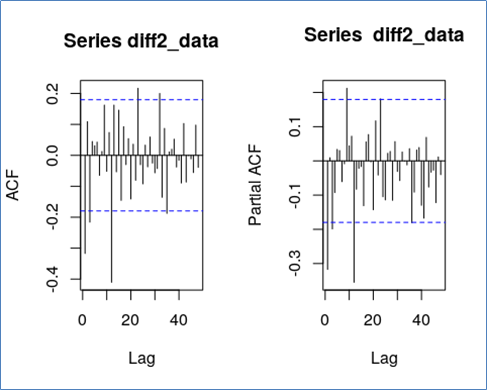
where p, d, and q make up the non-seasonal component and P, D, and Q make up the seasonal component. Specifying further, *p* = order of non-seasonal AR, *d* = differencing of non-seasonal component, *q* = order of non-seasonal MA, *P* = order of seasonal AR, *D* = differencing of seasonal component, *D* = order of seasonal MA. Lastly, *S* represents the seasonal time period of our model.

For our data, *S* = 12 because the seasonality trend repeats every 12 months. We also know that *d* = *D* = 1 because we differenced the seasonality component once with lag 12 and then we also differenced the non-seasonal component with lag 1 to remove the apparent downward trend. Next, we used ACF and PACF plots to determine the orders of *p, q, P,* and *Q*.

### Preliminary Model Identification

We are able to plot the ACF and PACF of our data as we have transformed and differenced our data to make it a stationary time series. First, we determine seasonal terms *P* and *Q* by looking at the seasonal lags. Since our seasonality is 12 months, we look at lags with a factor of 12 (12, 24, 36, …) in our ACF and PACF plots. Our ACF plot cuts off at lag 12 ( Q = 1) and our PACF plot cuts off after lag 12 ( P = 1).

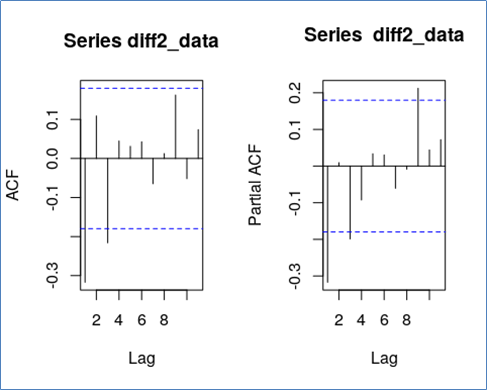
**ACF and PACF plots of Deseasonalized Transformed Data**



**Figure 4.1**

Next, we look at zoomed in ACF and PACF plots from lag 1 to 11 to determine non-seasonal terms *p* and *q.* Both plots cut off after lag 3 which means we have an ARMA model with *max(p,q)* = 3. From this, *p* and *q*  can both take values anywhere from 0 to 3. Now, we will test all possible models with P and Q fixed at 1, and combinations for p and q from 0 to 3. We end up with 16 possible models.

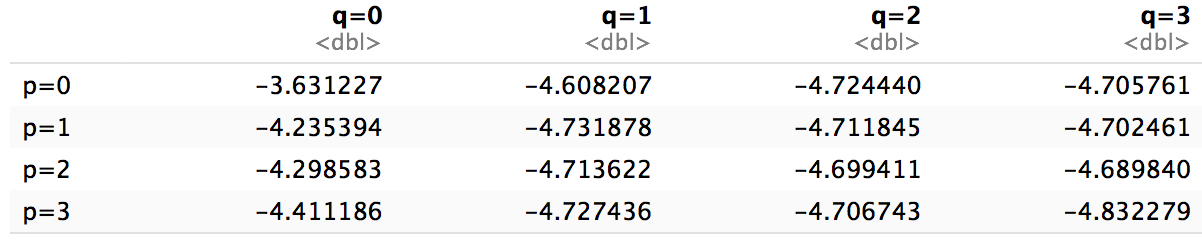
**Zoomed in ACF and PACF plots from Lag 1-11**

**Figure 4.2**

### Model Selection

From the 16 preliminary models identified, we can test what the best model is by using the small-sample-size corrected Akaike’s information criterion (AICc). When using AICc, we want a model with the low AIC and a small number of parameters. Our smallest calculated AICc comes from the model with *p = q =* 3, but this model has a high number of parameters, so we may dismiss it by principle of parsimony. The next 3 models with lowest AICc values are p = q = 1, p =3 q =1, p = 0 q=2, respectively. Therefore, our three possible models are: *SARIMA (1, 1, 1 ) x (1, 1, 0)12, SARIMA (3, 1, 1 ) x (1, 1, 0)12* and *SARIMA (0, 1, 2 ) x (1, 1, 0)12.*

NOTE: As shown by the models chosen, Q = 0 instead of Q = 1 as discussed in the previous section. This change is due to the fact that no matter the model chosen with Q = 1, there always appeared to be a unit root in the non-seasonal or seasonal MA(q) part of the model. Therefore, we modified Q to be 0 and were able to find proper models.



**Model Estimation**

The next step is to estimate the coefficients of these three best models; we can do this by using the MLE method. The table below displays the coefficients for each model.

|  |  |  |  |
| --- | --- | --- | --- |
|  | *SARIMA (1, 1, 1 ) x (1, 1, 0)12* | *SARIMA (3, 1, 1 ) x (1, 1, 0)12* | *SARIMA (0, 1, 2 ) x (1, 1, 0)12* |
| AR(1) | -0.3499 | -0.3634 |  |
| AR(2) |  | -0.0921 |  |
| AR(3) |  | -0.2173 |  |
| MA(1) | -0.9999 | -0.9703 | -1.3714 |
| MA(2) |  |  | 0.3856 |
| SAR(1) | 0.0721 | -0.0731 | -0.6827 |

***Model 1 :* *SARIMA (1, 1, 1 ) x (1, 1, 0)12***

(1 − 0.3499B)(1 − 0.0721B12)Xt = (1 − 0.999B)Zt,

where Zt ∼N (0, 0.003021)

***Model 2* : *SARIMA (3, 1, 1 ) x (1, 1, 0)12***

(1 − 0.3634B - 0.0921B2 - 0.2173B3)(1 + 0.0731B12)Xt = (1+0.9703B)Zt,

where Zt ∼N (0, 0.002925)

***Model 3* : *SARIMA (0, 1, 2 ) x (1, 1, 0)12***

(1 + 0.6827B12)Xt = (1+1.3714B - 0.3856B2)Zt,

where Zt ∼N (0, 0.003044)

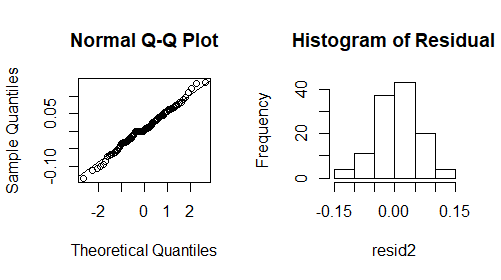
We can test causality and invertibility of each model by plotting their roots on a unit circle; if all red points are outside the unit circle, the model is causal and invertible. Our first model’s MA(1) root falls right on the unit circle so this is most likely a unit root. Our third model’s roots look good, but the absolute value of the MA(1) coefficient is greater than 0, so this indicates the model is not invertible (figures in the Appendix). This leaves us with the second model which from the plot of its roots is causal and invertible.

**V. Diagnostics**

From the three models that we have identified, only Model 2 is both casual and invertible, therefore we will proceed in validating the assumptions of that model. The diagnostics that we need to check are for the normality of error, serial correlation, and heteroscedasticity.

**Normality Checking**

To check for normality, we will analyze the histogram and the normal Q-Q plot, and perform the Shapiro Wilk Test at α = 0.05. The normal Q-Q plot and histogram are shown below in **Figure 5.1**.



**Figure 5.1**

In the normal Q-Q plot, we can see that most of the points lie on the straight 45 degrees line, which means that the errors are normal. In the histogram, the graph has a bell shape, which means that it is a normal distribution.

The Shapiro-Wilk Test will be performed to further prove the normality of errors. **Table 5.1** below shows the result of the Shapiro-Wilk Test.

*H0* : Residuals are normal  
*H1* : Residuals are not normal

|  |  |  |
| --- | --- | --- |
|  | **W. Statistics** | **P-value** |
| **Model 2** | 0.99023 | 0.5609 |

**Table 5.1**

In **Table 5.1**, we can see that the p-value for model 2 is 0.5609, which is greater than α = 0.05. The W. Statistic is also not too small; therefore, we do not reject our null hypothesis. The errors are normal. This also means that the residual is approximately an independent and identically distributed Gaussian.

**Detection of Serial Correlation**

Generally, time residuals are found to be serially correlated with their own lagged values. We do not want any correlation between the residuals in our final model, therefore we will test our models using the Ljung-Box Test and the Box-Pierce Test. The results of our tests is shown below in **Table 5.2**.

*H0* : Residuals are serially uncorrelated

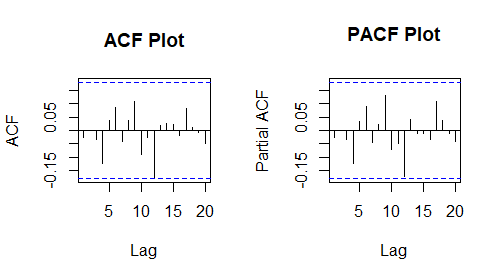
*H1* : Residuals are not serially uncorrelated

|  |  |
| --- | --- |
|  | **Model 2 P-value** |
| **Ljung-Box** | 0.2133 |
| **Box-Pierce** | 0.2772 |

**Table 5.2**

In **Table 5.2**, we can see that the p-value for both of the tests are greater than α = 0.05, therefore we do not reject our null hypothesis. The residuals are serially uncorrelated.

**Heteroscedasticity**

Heteroscedasticity is a violation of the constant error variance assumption. It occurs if variance of error is changing by time. We want our residuals to have constant variance. To check for heteroscedasticity, we can analyze the ACF and PACF plots of the squared residuals in our model. They should lie within 95% of the White Noise limits, or else it is a sign of heteroscedasticity. **Figure 5.2** below shows the ACF and PACF plot.

**Figure 5.2**

In both of the plots, we can see that most of the values lie within the bound (denoted by the blue dotted lines). This means that the constant variance assumption is not violated, therefore there is no heteroscedasticity.

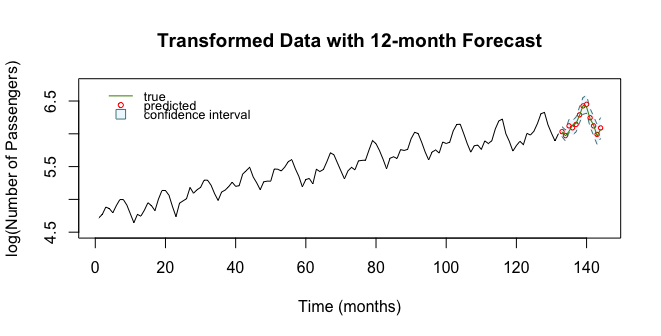
Since the model passed all of the diagnostic checks, it will be used as our final model. Let Xt be a stationary time series, X*t* = ∇∇*12* log(Y).

***Final* *Model* : *SARIMA (3, 1, 1 ) x (1, 1, 0)12***

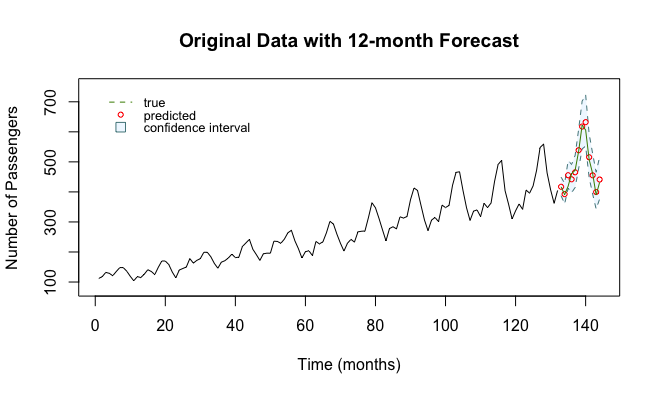
(1 − 0.3634B - 0.0921B2 - 0.2173B3)(1 + 0.0731B12)Xt = (1+0.9703B)Zt,

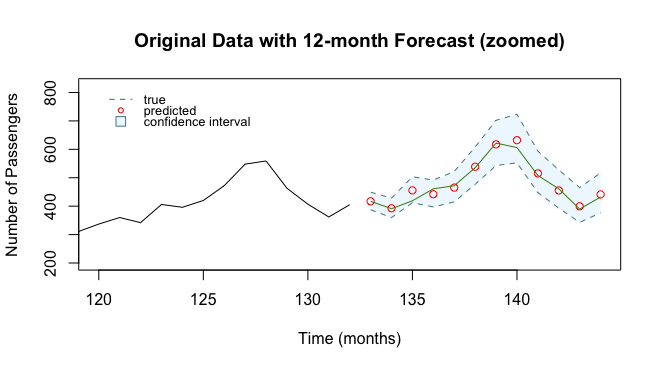
where Zt ∼N (0, 0.002925)

**VI. Forecasting**

Having validated our best model, we can now use it to forecast future values. We will forecast 12 months into the future, then compare our predicted values to the true values from the holdout data. **Figure 6.1** shows our model’s 12-month forecast on the transformed data, and **Figure 6.2** shows the 12-month forecast on the original data as does **Figure 6.3**, zoomed.

**Figure 6.1**

**Figure 6.2**

**Figure 6.3**

Based on these forecasting results, we can say that our model successfully captured seasonality and trend in the data. We see our predicted values gradually peak at the 7th and 8th forecasted values before falling, following the seasonality of the data. Our most egregious prediction yielded a deviation from the true answer of 36.55 passengers, and our average deviation from the true answer was 10.86 passengers, providing an amenable average deviation rate of 2.31% from the ground truth. Furthermore, all true values lie well within our 95% confidence interval. Thus, we consider our final model successful in forecasting and successful overall.

**VII. Future Study**

The forecasted values of our time series model are very close to the values form the original data, but our model still may have certain limitations and things it does not account for. As our data shows, international airline travel is very seasonal with a peak number of passengers in July and a minimum amount in November. Nonetheless, this seasonality can still be affected by certain factors that we cannot account for. A spike in airline ticket prices due to an increase in the cost of oil could cause a dramatic decrease in the number of international airline passengers for that time period. Sudden events such as these would cause our model to not be as accurate. To increase the accuracy of our forecasts we would need a larger sample of data. One that might include events such as the one discussed above. This would strengthen our model by making it more representative of the population of data. For the time being, our final model is a good predictor of future values while the data is following a similar trend. We are able to forecast values of international airline passengers for an entire year with 95% confidence. These predicted values can be of great use for airlines to continue to grow their industry.

**VIII. Conclusion**

Our goal was to use data on the number of international airline passengers from 1949 to 1960 per month to construct a time series and predict the number of international airline passengers for the next 12 months. In our time series, there was a noticeable trend upward each year because the US population increases each year which in turn means more people will fly internationally each year. There was also a significant seasonality component: there was always an increase in international passengers in the summer months and low international passengers in the winter months. This is mainly because school is not in session during the summer months. This allows families to go on long vacations internationally in the summer; typically, parents will not plan to take a two-week vacation while their child is in school.

We make our data stationary by transforming our time series and differencing it, and then we test several models and run diagnostic checking to make sure our final model is causal and invertible. Let X*t* be a stationary time series, X*t* = ∇∇*12* log(Y),

***Final* *Model* : *SARIMA (3, 1, 1 ) x (1, 1, 0)12***

(1 − 0.3634B - 0.0921B2 - 0.2173B3)(1 + 0.0731B12)Xt = (1+0.9703B)Zt,

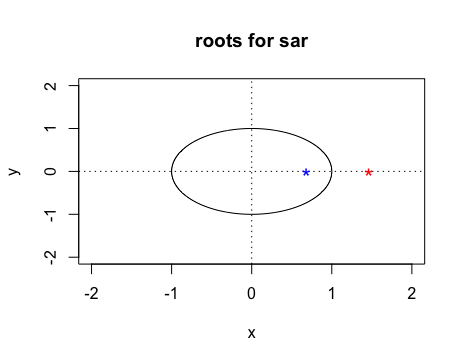
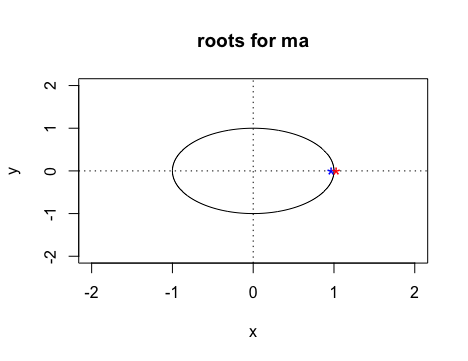
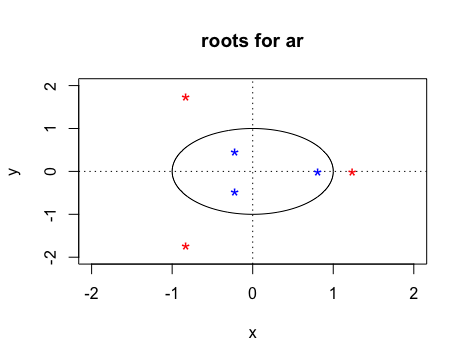
where Zt ∼N (0, 0.002925)

After finding our best fit model, we forecasted the monthly number of international airline passengers for the last 12 observations, January 1st, 1960 to December 31st, 1960. Our forecasted values were within our 95% confidence interval and they matched up to the real data points of those months very well. This proves that our final model is reasonable for this time series.

**IX. References**

1. DataMarket,<https://datamarket.com/data/set/22u3/international-airline-passengers-monthly-totals-in-thousands-jan-49-dec-60#!ds=22u3&display=line>

**X. Appendix**

****

**R code**

library(tseries)

library(forecast)

library(tseries)

library(astsa)

library(TSA)

library(GeneCycle)

library(ggplot2)

#international-airline-passengers.csv contains airline passenger totals indexed by month

data\_dir<-'/Users/timmy/Documents/school/year-4/fall/pstat175\_survival\_analysis/final\_project/data'

data\_path <- file.path(data\_dir, 'international-airline-passengers.csv')

airline <- read.csv(data\_path)

colnames(airline) <- c('month', 'passengers')

# splitting into train (first 90%) and test (last (10%)) set

train\_indices <- c(1:132) # leaving out last year

train\_airline <- airline[train\_indices,]

test\_airline <- airline[-train\_indices,]

# plotting time series

data <- train\_airline$passengers

plot(ts(data))

# plot shows seasonality, an upward trend, and increasing variance over time

par(mfrow=c(2,1))

# seasonal plot

seasonplot(ts(data),12,col=rainbow(3),year.labels=TRUE, main="Seasonal Plot")

x = window(ts(data), start(c(1949,1),end=c(1959,12)))

ggseasonplot(x)

# decomposition plot

decom = decompose(ts(data))

autoplot(decom, main="Decomposition Plot", xlab= "Time in Years")

# apply lag-12 difference on log-transformed data

diff\_data <- diff(log(data), lag = 12)

plot(ts(diff\_data))

abline(lm(diff\_data ~ as.numeric(1:length(diff\_data))))

# apparent downward trend

var(diff\_data) # 0.004036354

# lag-1 difference on previously differenced data

diff2\_data <- diff(diff\_data, lag = 1)

plot(ts(diff2\_data))

abline(lm(diff2\_data ~ as.numeric(1:length(diff2\_data))))

# horizontal line -> no trend

var(diff2\_data) # 0.002028451

# look at ACF and PACF for seasonal and non-seasonal preliminary model building

par(mfrow = c(1,2))

acf(diff2\_data, lag.max = 11)

pacf(diff2\_data, lag.max = 11)

acf(diff2\_data, lag.max = 48)

pacf(diff2\_data, lag.max = 48)

# model building,use AIC first

AICc <- numeric()

for(p in 0:3) {

for (q in 0:3) {

AICc <- c(AICc, sarima(diff2\_data, p, 1, q, 1,1,0, 12, details = FALSE)$AICc) } }

AICc <- matrix(AICc, nrow = 4, byrow = TRUE)

rownames(AICc) <- c("p=0", "p=1", "p=2", "p=3")

colnames(AICc) <- c("q=0", "q=1", "q=2", "q=3")

AICc <- data.frame(AICc)

aicc <- setNames(AICc, c("q=0", "q=1", "q=2", "q=3"))

# lowest AICc p = q= 3; dismiss due to many parameters

# the next 3 models with lowest AICc values are p = q = 1, p =3 q =1, p = 0 q=2

# fit and estimation based on MLE method

# model 1: SARIMA (1,1,1,1,1,0)

sarima1 <- arima(diff2\_data, order = c(1,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

plot.roots(NULL, polyroot(c(1, -0.3499)), main = "roots for ar")

plot.roots(NULL, polyroot(c(1, -0.9999)), main = "roots for ma")

plot.roots(NULL, polyroot(c(1, -0.6788)), main = "roots for sar")

# appears to be a unit root in ma part of model

#model 2: sarima(3,1,1,1,1,0)

sarima2 <- arima(diff2\_data, order = c(3,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

plot.roots(NULL, polyroot(c(1,-0.3634,-0.0921, -0.2173)), main = "roots for ar")

plot.roots(NULL, polyroot(c(1, -0.9703 )), main = "roots for ma")

plot.roots(NULL, polyroot(c(1, -0.6818)), main = "roots for sar")

# roots look good!

# model 3:SARIMA (0,1,2,1,1,0)

sarima3 <- arima(diff2\_data, order = c(0,1,2), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

plot.roots(NULL, polyroot(c(1, -1.3714, 0.3856)), main = "roots for ma")

plot.roots(NULL, polyroot(c(1, -0.6827)), main = "roots for sar")

# roots look good, but not invertible because absolute value of one ma root greater than 1

# test diagnostics for model 2 -> our best model

resid2 <- residuals(sarima2)

hist(resid2)

qqnorm(resid2)

qqline(resid2)

shap2 <- matrix(c(shapiro.test(resid2)$statistic, shapiro.test(resid2)$p.value))

Box.test(resid2, type = "Box-Pierce", lag = 12, fitdf = 4)

Box.test(resid2, type = "Ljung-Box", lag = 12, fitdf = 4)

shapiro.test(resid2)

# all tests pass because p > 0.05

acf(resid2, main = "ACF Plot")

pacf(resid2, main = "PACF Plot")

# this model is feasible and we move on to forecasting

# forecasting using model 2

# sarima.for(data, 12, 3,1,1,1,1,0,12)

trans <-log(data)

# based on final model with transformation

fit = arima(trans, order = c(3,1,1), seasonal = list(order = c(1,1,0), period = 12), method = "ML")

pred1 <-predict(fit, n.ahead = 12)

u.pred1 <- pred1$pred + 2\*pred1$se

l.pred1 <- pred1$pred - 2\*pred1$se

ts.plot(trans, xlim = c(1, length(trans) + 12), ylim = c(4.5,6.75), ylab='log(Number of Passengers)', xlab='Time (months)')

polygon(c((length(data3)+1):(length(data)+12), rev((length(data3)+1):(length(data)+12))), c(u.pred1, rev(l.pred1)),

col = 'aliceblue', border = NA)

lines(u.pred1, col = "cadetblue4", lty='dashed')

lines(l.pred1, col = "cadetblue4", lty='dashed')

points((length(trans)+1):(length(trans)+12), pred1$pred, col = "red", cex = 0.5)

lines((length(data3)+1):(length(data)+12), log(test\_airline$passengers), col = 'chartreuse4')

legend('topleft', legend = c('true', 'predicted', 'confidence interval'), col = c('chartreuse4','red','cadetblue4'), lty=c('solid', NA, NA), pch=c(NA, 1, 22), pt.bg=c(NA,NA,'aliceblue'), pt.cex=c(NA,0.7,1.5), box.lty=0, cex=0.8, inset=.05)

title('Transformed Data with 12-month Forecast')

# based on original

pred2 <-exp(pred1$pred)

u.pred2 <- exp(u.pred1)

l.pred2 <- exp(l.pred1)

data3 <-ts(data)

ts.plot(data3, xlim = c(1, length(data3)+12), ylim = c(80,750), ylab='Number of Passengers',xlab='Time (months)')

polygon(c((length(data3)+1):(length(data)+12), rev((length(data3)+1):(length(data)+12))), c(u.pred2, rev(l.pred2)),

col = 'aliceblue', border = NA)

lines(u.pred2, col = "cadetblue4", lty='dashed')

lines(l.pred2, col = "cadetblue4", lty='dashed')

points((length(trans)+1):(length(trans)+12), pred2, col = "red", cex = 0.7)

lines((length(data3)+1):(length(data)+12), test\_airline$passengers, col = 'chartreuse4')

legend('topleft', legend = c('true', 'predicted', 'confidence interval'), col = c('chartreuse4','red','cadetblue4'), lty=c('solid', NA, NA), pch=c(NA, 1, 22), pt.bg=c(NA,NA,'aliceblue'), pt.cex=c(NA,0.7,1.5), box.lty=0, cex=0.8, inset=.05)

title('Original Data with 12-month Forecast')

# zoom in on original data

ts.plot(data3, xlim = c(length(data3)-12, length(data3)+12), ylim= c(200, max(u.pred2)+100),ylab='Number of Passengers',xlab='Time (months)')

polygon(c((length(data3)+1):(length(data)+12), rev((length(data3)+1):(length(data)+12))), c(u.pred2, rev(l.pred2)),

col = 'aliceblue', border = NA)

lines((length(data3)+1):(length(data)+12), u.pred2, col = "cadetblue4", lty='dashed')

lines((length(data3)+1):(length(data)+12), l.pred2, col = "cadetblue4", lty='dashed')

points((length(data3)+1):(length(data)+12), pred2, col = "red")

lines((length(data3)+1):(length(data)+12), test\_airline$passengers, col = 'chartreuse4')

legend('topleft', legend = c('true', 'predicted', 'confidence interval'), col = c('chartreuse4','red','cadetblue4'), lty=c('solid', NA, NA), pch=c(NA, 1, 22), pt.bg=c(NA,NA,'aliceblue'), pt.cex=c(NA,0.7,1.5), box.lty=0, cex=0.8, inset=.05)

title('Original Data with 12-month Forecast (zoomed)')