PSTAT 174 FINAL PROJECT

Monthly Armed Robberies in Boston (1966-1975)

-

Time Series Analysis

Instructor: Sudeep Bapat

Authors: *William Berman*
Anthony Castillo
Tatiana Moreno
Anton Sunico
Adam Trujillo

Table of Contents

1 Executive Summary

In recent decades, statistical modelling of crime has led to a better understanding of its nature and origins. In particular, through the statistical modelling of various criminal activities, one can even attempt to predict and insure against crimes as a means of risk management. For instance, in the data seen in this report, we analyze a decades worth of armed robbery data from Boston. This data is measured on a monthly basis from January 1966 to October 1975. The purpose of this project is to predict twelve future monthly values of the number of Boston robberies given a decades worth of data in the form of a time series and the prerequisite means of analyzing that data. We will forecast Boston's armed robbery rates from November 1975 - October 1976, and model these values from an ARIMA model that we derived from the original time series by transforming and differencing the original data. Our predictions are within the 95% level of confidence, and they are modelled close to the data in the original time series.

1.1 Introduction

Agents of law enforcement have used various computational and analytical methodologies to understand the frequency and origins of many types of crimes for many decades. Thus, it is important for us to understand these methodologies and put them into practice. Crime prevention has been a major issue in politics in all of American history. The data we found provides us a snapshot view of crime in one period of time, that from January 1966 to October 1975. We chose this data set because it provides a very concise view into one of the most pivotal times in modern politics and has a sufficiently large sample size for the purpose of analysis (that being n=118 months of data).

Plotting the dataset, we clearly see an upward trend of robberies in Boston and an increase in its variance as time progresses. Although the decomposition plot shows that there may be a seasonality component present, delving deeper into further analysis elucidates how a seasonality component is not significant for our data. We eliminate the increasing variance by taking the cube root of the data set. To remove the upward trend, we differenced the now transformed data once. Since further differencing increases the variance, the model is only differenced once. After these transformations, we end up with a data set with a variance that is significantly less than the original model, indicating an improvement our representation of the data set.

Based on the ACF and PACF plots and the AIC and BIC model selection criteria, we were able to identify two possible models to represent the data: ARIMA(1, 1, 2) and ARIMA(2, 1, 2). Additional diagnostic checks, such as the serial correlation check, Box-Pierce, and Ljung-Box were done to confirm that our models are sufficient. Heteroscedasticity was checked through the ACF and PACF plots of the residuals. After coming to the conclusion that both models ARIMA(1, 1, 2) and ARIMA(2, 1, 2) were stationary, invertible and passed all diagnostic tests, we concluded the most suitable model would be ARIMA(1, 1, 2) due to its lower AIC and BIC values as well as its lower number of parameters. Finally, we forecasted monthly crime numbers 12 months in advance and compared these forecasted values to the true data from the Massachusetts Crime Rates 1960 - 2016 Public Records document listed in our references below and saw the 95% confidence intervals for our forecasted values contained the true number of armed robberies in Boston from November 1975 to October 1976. Therefore, this reaffirmed the adequacy of our final model.

2 Data Analysis

2.1 Introductory Data Exploration

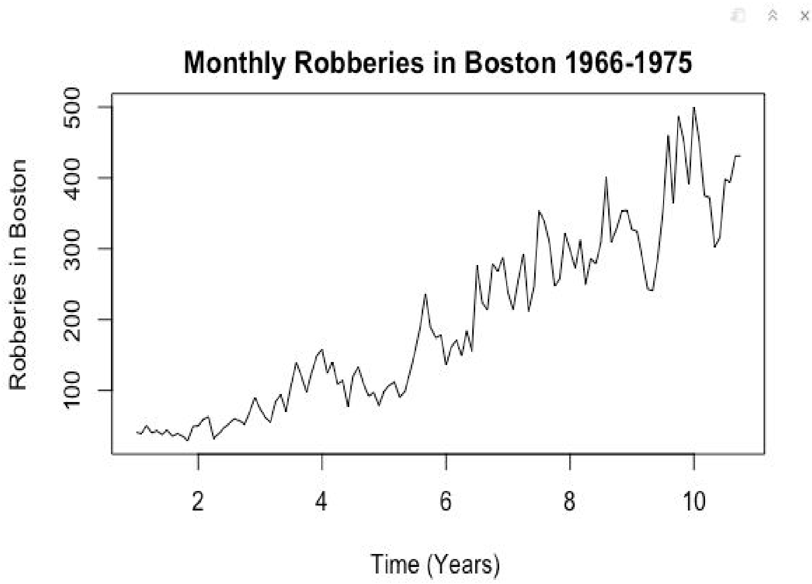


FIGURE 2.1.1

In the plot to the left (Figure 2.1.1) we see an upward trend in the monthly robberies in Boston data, and also see a slightly increasing variance over time. To fix this trend we will have to difference the data, and test for stationarity. Furthermore, in this plot, we do see an increasing variance situation over time; therefore, we must use the Box-Cox transformation to generate a stationary time series. Because there does not necessarily seem to be consistent peaks over time in the original data, we do not detect a component of seasonality in this graph; however, we will do further analysis to prove our initial hypotheses.

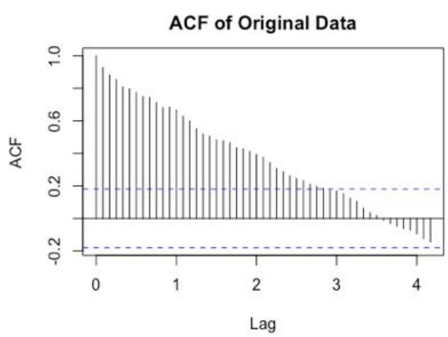


FIGURE 2.1.2

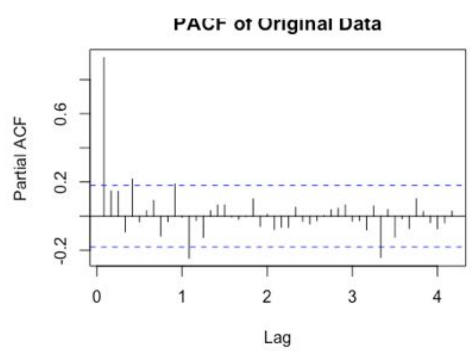


FIGURE 2.1.3

Looking at the ACF of our original data above, we do not necessarily see evidence supporting a seasonality component in our data, as there are only miniscule oscillations in the graph. The evident decreasing trend in the ACF graph may be present due to some trend in our original data. Therefore, we must do more analysis to see if it is significant to account for seasonality and trend in our final model.

2.2 Decomposition Model

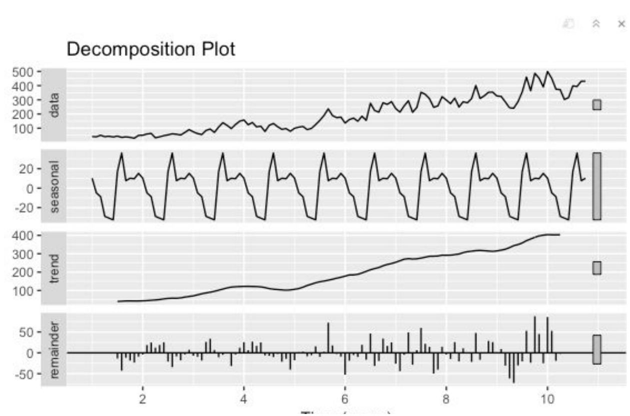
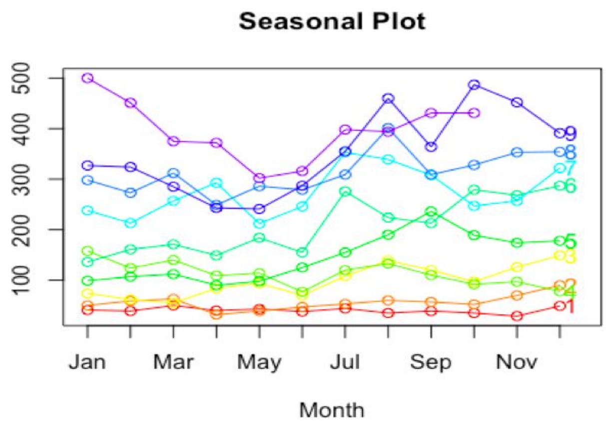


FIGURE 2.2.1

We graphed a decomposition plot with the formula $X_t = m_t + s_t + S_t$, where m_t is the trend, s_t is the seasonality and S_t is the stationarity to further analyze any existence of seasonality or trend in our robbery data X_t. Here in the decomposition plot under "trend" we see an obvious upward trend which reaffirms our initial proposition that we would have to difference the data due to an upward trend in our original data (Figure 2.1.1). The decomposition plot also confirms the initial observation (Figure 2.1.1) of possible increasing variance over time, so transforming the y-variable will likely be necessary. To handle this situation, we will first be performing a Box-Cox transformation. Additionally, there may be some pattern in the "seasonality" component of this decomposition plot. However, we must delve deeper into checking if the seasonality component is a significant.



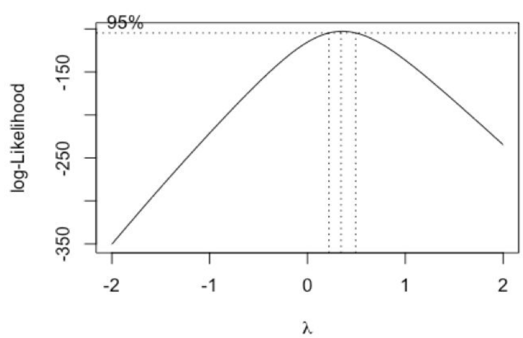
The seasonal plot to the left does not elucidate a clear image of a seasonality component present in our initial data. All we are able to detect is a slight decreasing trend in the earlier months from January to May then possibly increasing from May to July. Because the image from Figure 2.2.2 does not show a clear seasonality issue with our data, we will not assume there is a seasonality component in our data. Ultimately, there is not enough evidence proving a seasonality component is present. Thus we will proceed assuming no seasonality is present.

FIGURE 2.2.2

3 Data Transformation

3.1 Stabilizing Variance of the Number of Robberies

Because our goal is to construct a stationary time series from our original data, we must stabilize the increasing variance in our model over time. Thus, we must perform a Box Cox transformation to fix the increasing variance.



We use the Box-Cox transformation to find the best λ for transformation. In the Box Cox graph (Figure 3.1.1), we see that 0 is not contained in the interval for λ. Therefore, a log transformation would not be appropriate in this case. However, we see λ could possibly be equal to 1/3 or 1/2, as we see in the interval below. Therefore we will plot the square root of the number of robberies as well as the cube root of the number of robberies and analyze both graphs to determine which λ is best suited for our model.

FIGURE 3.1.1

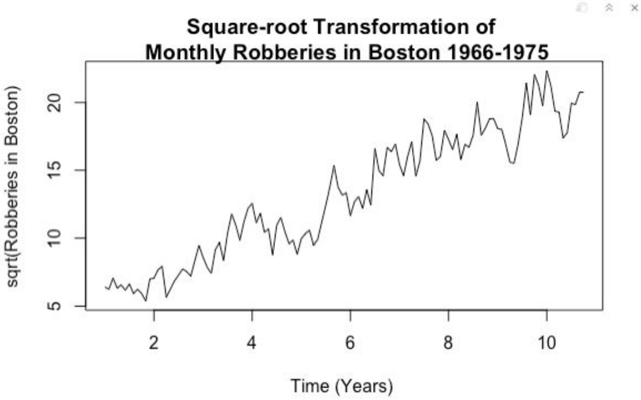


FIGURE 3.1.2

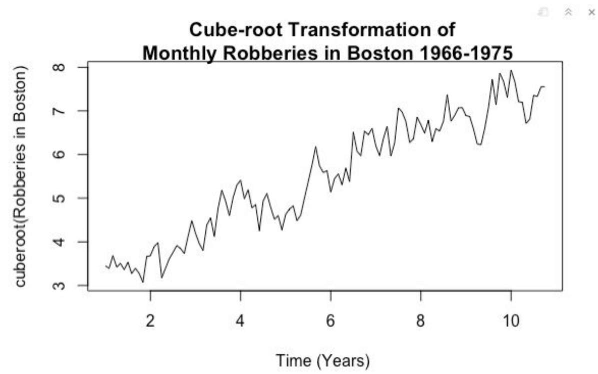


FIGURE 3.1.3

After determining the cube-root transformation of monthly robberies has a lower variance than the square-root transformation the untransformed original time series data, and the Box Cox value is approximately 0.3, we decided the best suited transformation for our model would be the cube-root of the number of robberies.

3.2 Removing Upward Trend by Differencing

The cube-root transformation of monthly robberies model has an evident upward trend, which we will remove by differencing. We start by differencing the data once. Doing so decreases the variance of the data, which is one of our goals. Differencing the data once more, however, increases the variance instead. This means that differencing the data once is best for our final model, and any additional differencing to the data is unnecessary. In plotting the differenced data, we see that the graph resembles Gaussian white noise, and is thus stationary.

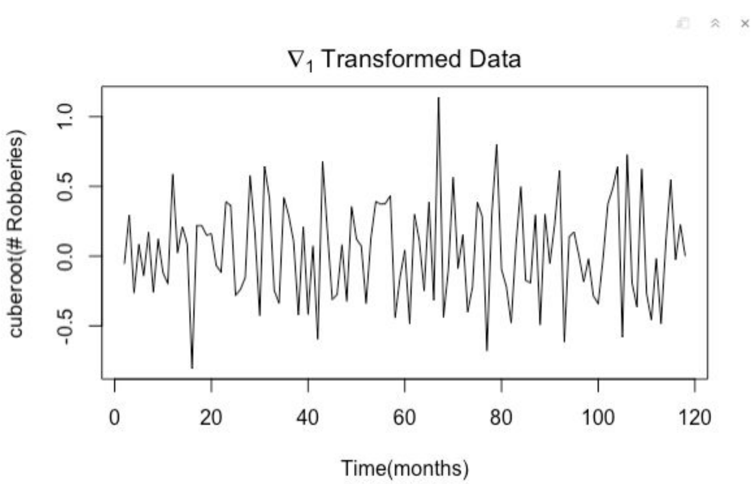


FIGURE 3.2.1

Looking at the figure 3.2.1 to the left we see a non-constant and non- increasing variance along with no visible trend. Therefore, the cube-root transformation of the number of robberies differenced at lag 1 to remove the trend looks to be an appropriate stationary time series.

4 Model Identification and Estimation

4.1 Model Identification

We see that our data suggests an ARIMA model due to the necessity of differencing our data to remove the upward trend. An ARIMA model has the model notation ARIMA(p,d,q), where p represents the model's AR component, q represents the model's MA component, and d represents the amount of times that the data is differenced. In order to make our data stationary, we differenced our cube-root transformed data at lag 1 in order to account for the upward trend component. Therefore, in the ARIMA model we see d = 1. Now we must find the other values of p, and q to find the best model.

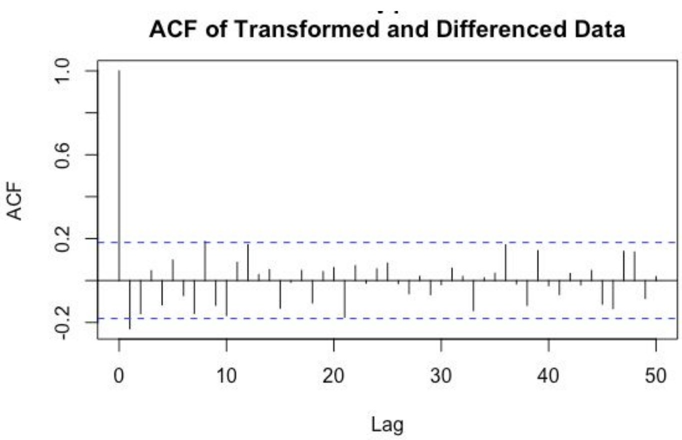


FIGURE 4.1.1

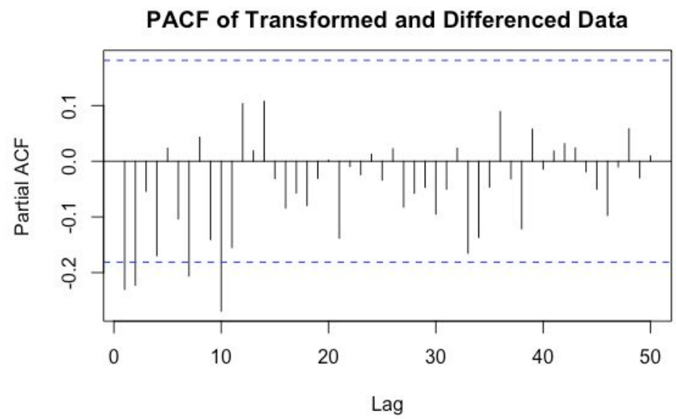


FIGURE 4.1.2

In the ACF plot above (Figure 4.1.1) we see the ACF seems to cut off at lag = 1, and looking at the PACF plot above (Figure 4.1.2), we see that the PACF tails off, which suggests that MA(1) may be a suitable model. The PACF plot can also be interpreted to be cutting of at lag 10, indicating an ARMA(1, 10) model to be sufficient. We will ultimately determine the best model by checking to see which model has the minimum value for AIC and BIC for p and q values ranging from 0 to 10.

4.2 Model Selection

We took all possible model parameters that our analysis indicated and analyzed the information criterion for all combinations. The following ARIMA(1,1,2) model resulted in the minimum AIC and minimum BIC, and the ARIMA(2,1,2) model was the next model that minimized the BIC. Thus, these are the two models we chose to continue with model estimation. The values of ARIMA(1,1,2) and ARIMA(2,1,2) AIC and BIC are provided in the table below.

Model 1: ARIMA(1,1,2) -

p	**d**	**q**	**AIC**	**BIC**
1	1	2	88.56592	99.58028

TABLE 4.2.1 AIC, BIC values for ARIMA(1,1,2)

Model 2: ARIMA(2,1,2) -

p	**d**	**q**	**AIC**	**BIC**
2	1	2	90.33802	104.10597

TABLE 4.2.2 AIC, BIC values for ARIMA(2,1,2)

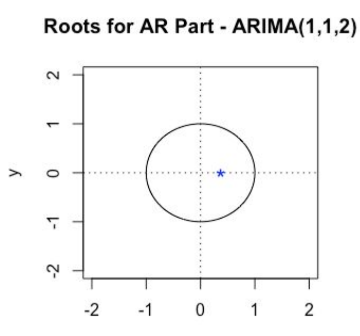
4.3 Model Estimation

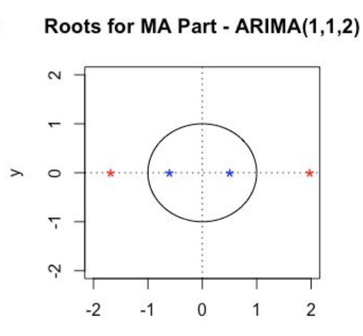
To proceed with the model estimation of ARIMA(1,1,2) and ARIMA(2,1,2) using the maximum likelihood estimation procedure we determined the estimated coefficients of both models in order to determine if any of these values were close enough to zero to see if any contribution of an AR of MA component were not necessary; however, the coefficients were not close to zero. We also must use these coefficients in order to proceed with checking both models' invertibility and causality.

	Model 1: ARIMA(1,1,2)	**Model 2**: ARIMA(2,1,2)
AR(1)	-0.3718	-0.2991
AR(2)	-	0.1702
MA(1)	0.0913	0.0266
MA(2)	-0.302	-0.4532

TABLE 4.3.1 Coefficients of Both ARIMA Models

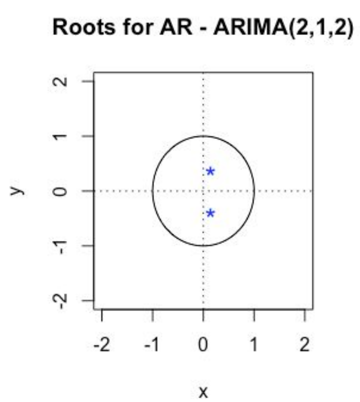
Now we plot the roots of ARIMA(1,1,2) and ARIMA(2,1,2) in order to check for stationarity and invertibility.

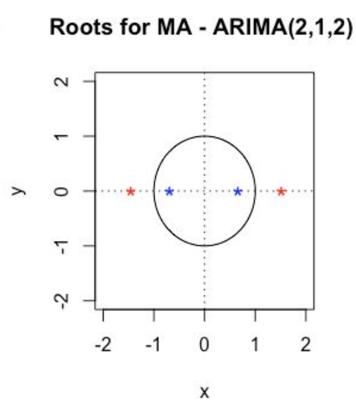




Model 1: ARIMA(1,1,2)

Because all the roots (red stars) for the AR, and MA components in the ARIMA(1,1,2) model lie outside the unit circle (some so far outside they are not seen in the constraints of the graph) we conclude Model 1 is causal and invertible.





Model 2: ARIMA(2,1,2)

Because all the roots (red stars) for the AR and MA components in the ARIMA(2,1,2) model lie outside the unit circle (some so far outside they are not seen in the constraints of the graph) we conclude Model 2 is causal and invertible.

5 Diagnostic Checking

Since we have concluded ARIMA(1,1,2) and ARIMA(2,1,2) are stationary and invertible by using the estimation of these models' parameters, we then may proceed to diagnostic checking process of both models.

5.1 Normality Checking

In order to check the normality of error terms in both models, we must first plot a histogram and check if these errors follow a normal distribution. Furthermore, we will plot a Normal-Q-Q plot and perform the Shapiro Wilk test at α = 0.05 as further evidence to reaffirm the normality of error terms.

Model 1: ARIMA(1,1,2)

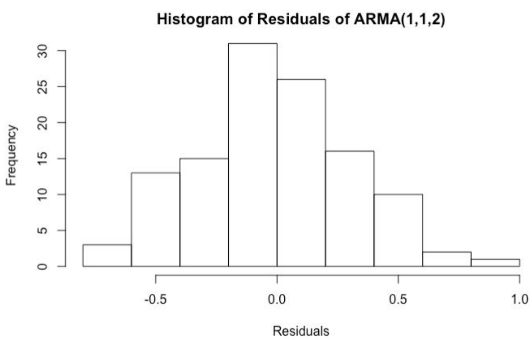


FIGURE 5.1.1

Model 2: ARIMA(2,1,2)

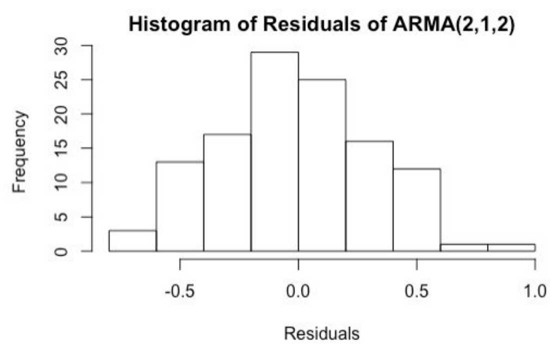


FIGURE 5.1.2

Looking at Figure 5.1.1 and Figure 5.1.2 above we see the residuals of both models seem to follow a normal distribution centered at 0.

Model 1: ARIMA(1,1,2)

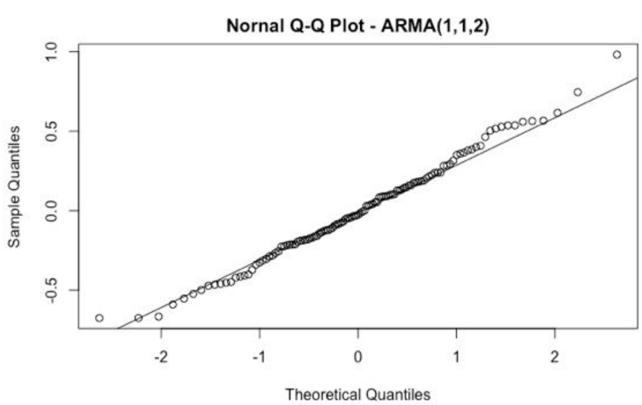


FIGURE 5.1.3

Model 2: ARIMA(2,1,2)

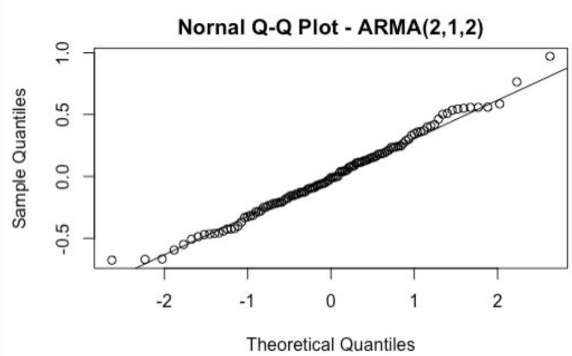


FIGURE 5.1.2

In Figure 5.1.3 and Figure 5.1.2 above we see the points all seem to fall on the line, meaning the residuals are normal. However, we do see very slight divergence near the ends. Thus, we will perform the Shapiro-Wilk test to come to our final conclusion of whether or not the residuals are normally distributed.

Shapiro Test Results: H_0= residuals are normal distributed vs. H_1 = residuals are not normally distributed

	Model 1: ARIMA(1,1,2)	Model 2: ARIMA(2,1,2)
W-Statistic	0.99125	0.9909143
P-Value	0.6673572	0.6362329

TABLE 5.1.4

For ARIMA(1,1,2), the Shapiro-Wilk test returns a p-value of 0.6673572, which is greater than 0.05. Therefore we fail to reject the assumption of normality for Model 1 and conclude the residuals are normal for this model. For ARIMA(2,1,2), the Shapiro-Wilk test returns a p-value of 0.6362329, which is greater than 0.05. Therefore we fail to reject the assumption of normality for Model 2 and conclude the residuals are normal for this model as well.

5.2 Independence (Serial Correlation) Checking

To ensure our residuals are not serially correlated we will perform the Box-Pierce and Ljung-Box test, with the null and alternative hypotheses as stated below.

H_0= residuals are serially uncorrelated (independent) vs. H_1 = residuals are serially correlated (dependent)

	Model 1: ARIMA(1,1,2)	Model 2: ARIMA(2,1,2)
Box-Pierce P-Value	0.1857596	0.1931732
Ljung-Box P-Value	0.1373411	0.1427492

TABLE 5.2.1

Because both the Box-Pierce P-Value = 0.1857596 and the Ljung-Box P-Value = 0.1373411 of the model ARIMA(1,1,2) > 0.05, we fail to reject our null hypothesis and conclude the ARIMA(1,1,2) residuals are not serially correlated. Furthermore, since the Box-Pierce P-Value = 0.1931732 and the Ljung-Box P-Value = 0.1427492 of Model 2 > 0.05, we fail to reject our null hypothesis and conclude the ARIMA(2,1,2) residuals are also not serially correlated as well. Thus, we proceed with diagnostic checking for both models.

5.3 Checking for Constant Variance

It is imperative to check for heteroscedasticity in our models to reaffirm our model estimation and prediction is accurate. Thus, to check for a violation of the constant variance of errors in our models we analyzed the ACF and PACF plots of the residuals of both models to see if the values lied within the 95% confidence interval bounds as evidence supporting the residuals have constant variance.

Model 1: ARIMA(1,1,2)

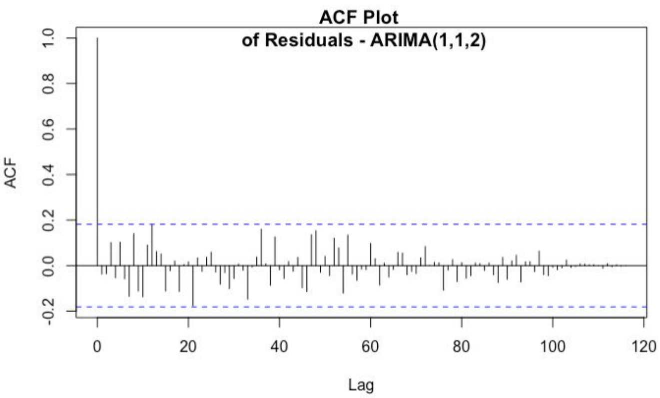


FIGURE 5.3.1

Model 1: ARIMA(1,1,2)

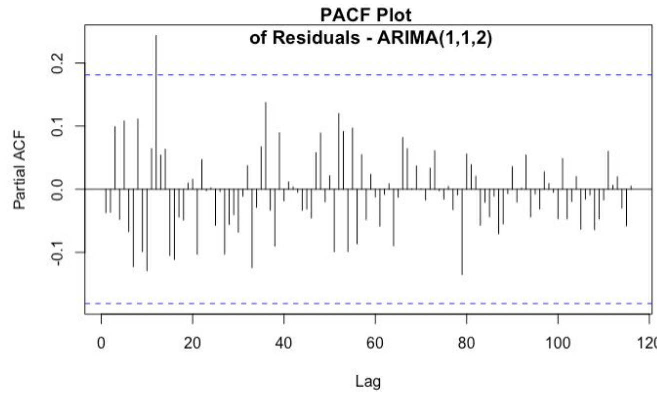


FIGURE 5.3.2

In Figure 5.3.1, we see the residuals of Model 1 are all within the 95% confidence interval limits, much like the corresponding PACF (exempting one line at Lag 12). We assume this value is some sort of outlier in our data. Therefore, there is no violation of the constant variance of errors in the model ARIMA(1,1,2).

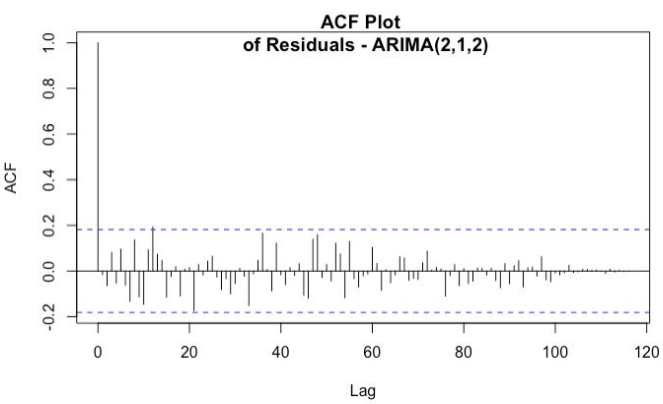


FIGURE 5.3.3

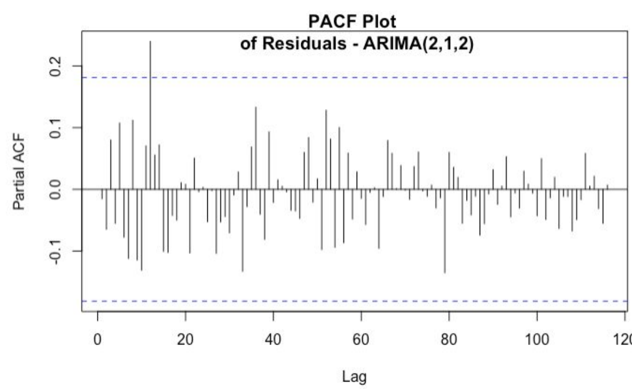


FIGURE 5.3.4

In Figure 5.3.3 the residuals of Model 1 are all within the 95% confidence interval limits. The PACF plot above has all values within the 95% bounds as well with one line exceeding the bound. Once again, we assume this value is some sort of outlier in our data and proceed to assume there is no heteroscedasticity in the model ARIMA(2,1,2).

Because both models, ARIMA(1,1,2) and ARIMA(2,1,2) both pass all diagnostic tests, we choose ARIMA(1,1,2) as our final model due to it having a lower number of parameters as well as lower AIC and BIC values.

Final Model: ARIMA(1,1,2)

$$(1 + 0.37B)X_t = (1 + 0.09B - 0.30B^2)Z_t$$

Where $Z_t \sim N(0, 0.1064543)$

6 Forecasting

The dataset only included the monthly number of robberies in Boston from January 1966 to October 1975. Therefore, we decided to forecast the number of robberies from November 1975 - October 1976. Thus, we will have the predicted number of robberies for the next 12 months. In order to reaffirm the accuracy of our forecasted values we will be comparing the number of predicted robberies from November 1975 - October 1976 to the Massachusetts Crime Rates 1960 - 2016 Public Records document.

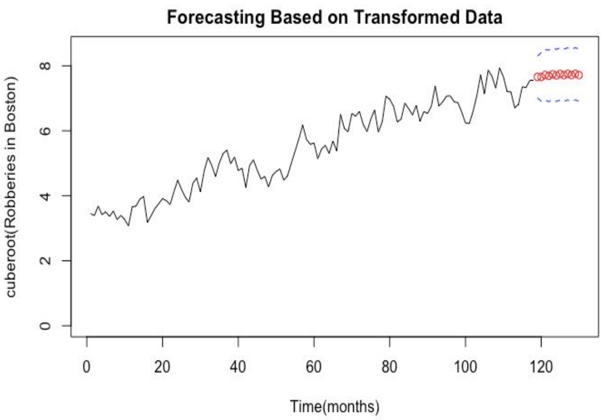


FIGURE 7.1.1

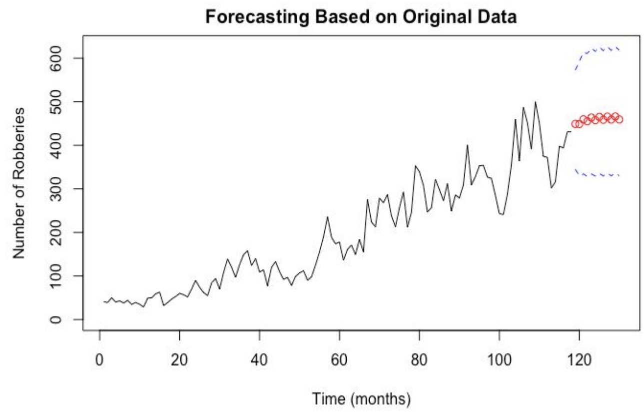


FIGURE 7.1.2

Figure 7.1.1 above shows the forecasted number of robberies based on our transformed data which was the cube-root of the original time series. The red dots on the graph are representative of the twelve forecasted values, while the blue dotted lines are indicative of the 95% confidence interval for those forecasted values. Figure 7.2.2 has the forecasted values of our initial time series data set and is what we are interested in analyzing. Below is a zoomed in plot of the forecasted values for our original dataset.

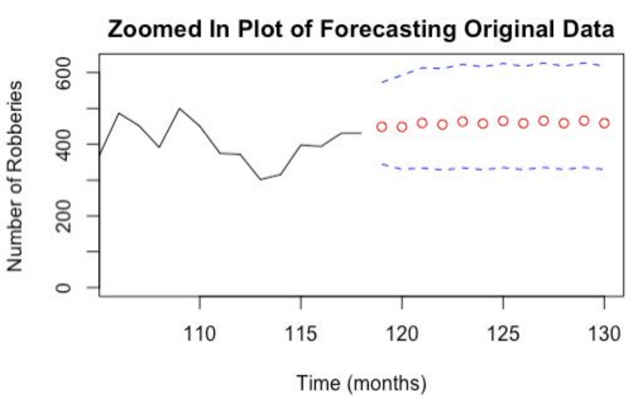


FIGURE 7.1.3

In Figure 7.1.3 to the left we see the forecasted values of the number of robberies from November 1975 - October 1976 seem to lie within approximately 300 to 600 armed robberies for this time period. Checking the Massachusetts Crime Rates 1960 - 2016 Public Records document we see that the number of robberies from 1975-1976 lies within this interval. Thus, this further proves the accuracy of our final model.

7 Future Study

Despite the accuracy of our forecasted values we do notice there are some limitations to our analysis. Actual contextual political and demographic data (ie- incomes, election results) are limited since we are analyzing only a decades worth of data on a monthly basis. The time series provides us with robbery numbers over the course of a decade, but what it does not provide are insights into political trends and demographic changes over time. These are absolutely necessary to understand why the number of robberies increased by so much over time. In the course of a decade, only so many social trends and electoral cycles will occur, so having data from before and after this time series would provide more context for even more in-depth analysis.

Furthermore, the time series itself is also limited as much as the context is because any notion of seasonality only emerges biennially, and even that cannot be proven as the SARIMA models that would describe such a model do not pass all the requisite diagnostic checks. Thus, the model we decided would most accurately describe the time series ended up not having a seasonality component. If the data were broken into weekly intervals or if two decades of data were offered, then perhaps an even more precise analysis could occur.

Despite these limitations, we were able to gain insight into criminology in Boston. In the decade we studied, we saw a clear upward trend in the number of robberies in Boston. However, additional data we sourced from Bostonian public records indicates a decrease in the number of robberies going all the way into 2016. This apex in robbery numbers implies many things mostly related mass social change in Boston at that time. For instance, by quickly glancing at public records data cited below, we see a spike in Boston's population from 1966-1975. This spike in population was not seen again until recent decades, and from this population change, one can infer that a large number of these robberies came from this population change. Through this analysis, we can continue to ask these kinds of questions so we can solve the issues of violence in society.

8 Conclusion

Our main objective was to form a time series model that could forecast the number of armed robberies in Boston between 1966-1975. Through this exercise, we were able to predict Boston armed robbery rates 12 months in advance from November 1975 to October 1976. From analyzing the graph of our original data and the decomposition plot, we were able to see a clear upward trend in the number of robberies, and we reasonably concluded that the final time series model follows an ARIMA(1,1,2) model. A great challenge we faced was determining if there was a seasonality component in our initial data after noticing a pattern in the seasonality portion of the decomposition plot. However, we primarily addressed the possible presence of a seasonality component by fitting various SARIMA models to our data. However, in each instance we were unable to find a SARIMA model that passed all of the diagnostic tests as well as was stationary and invertible. This issue made us believe our data was overdifferenced and we should follow the evidence in the seasonality plot that indicated there was no obvious seasonality component and continue with an ARIMA model. Ultimately, we concluded there was a lack of statistically significant seasonality in our time series. When we assumed there was no clear seasonality in our original time series and transformed our data as well as differenced our data once to account for the upward trend, various ARIMA models were stationary and invertible as well as passed all diagnostics tests. After forecasting a year's worth of the number of armed robberies in Boston and comparing these forecasted values to the Boston Crime Rates 1960 - 2016 Public Records document we see that the actual number of armed robberies all lied within the 95% confidence interval of our forecasted plot. Therefore, due to our forecasting accuracy we are able to conclude the ARIMA(1,1,2) model is sufficient, and a seasonality component was not statistically significant.

Therefore, our final model where X_t is the transformed and differenced data $X_t = \nabla_1 Y_t^{1/3}$. is as follows:

$$\text{ARIMA}(1,1,2)$$
$$(1 + 0.37B)X_t = (1 + 0.09B - 0.3B^2)Z_t$$
$$\text{Where } Z_t \sim N(0, 0.1064543)$$

We thank Professor Bapat for providing us with the requisite information to complete this time series project.

9 References

1. *Data Market,* https://datamarket.com/data/set/22ob/monthly-boston-armed-robberies-jan1966-oct1975-deutsch-and-alt-1977#!ds=22ob&display=line

2. *Massachusetts Crime Rates 1960 - 2016 - Public Records,* http://www.disastercenter.com/crime/macrime.htm

10 Appendix (RCode)

````
```{r}
library(ggplot2)
library(forecast)
```
```{r}
rob <- read.table("monthly.boston.robberies.txt", header = F)
rob.ts <- ts(rob, frequency = 12)
plot(rob.ts, xlab = "Time (Years) ", ylab = "Robberies in Boston", main = "Monthly Robberies
in Boston 1966-1975")
```
```{r}
acf(rob.ts, lag.max=50)
title("ACF of Original Data", line = 1)
pacf(rob.ts, lag.max=50)
title("PACF of Original Data", line = 1)
```
```{r}
#decomposition plot of original data
decom.plot = decompose(rob.ts)
autoplot(decom.plot, main = "Decomposition Plot", xlab = "Time (years)")
```
```{r}
#checking for seasonality component
seasonplot(rob.ts, 12, col=rainbow(12) ,year.labels=TRUE, main="Seasonal Plot")
```
```{r}
#Stabilizing the increasing variance over time
# Box-Cox Transformation
require(MASS)
bcTransform <- boxcox(rob.ts~as.numeric(1:length(rob.ts)))
```
```{r}
#square root transformation
plot(sqrt(rob.ts), xlab = "Time (Years) ", ylab = "sqrt(Robberies in Boston)", main =
"Square-root Transformation of \n Monthly Robberies in Boston 1966-1975")
````
````

````
```
```{r}
#cube root transformation
plot((rob.ts)^(1/3), xlab = "Time (Years) ", ylab = "cuberoot(Robberies in Boston)", main =
"Cube-root Transformation of \n Monthly Robberies in Boston 1966-1975")
```
```{r}
#Comparing the Variances of the initial time series with square-root and cube root
transformations
var(rob.ts)
var(sqrt(rob.ts))
var((rob.ts)^(1/3))
# cube root transformation has the smallest variance

#new model with cube root transformation
cub.rob <- ts((rob.ts)^(1/3))
```
```{r}
#De-trend the data
robdiff1 <- diff(cub.rob, lag=1)
```
```{r}
#Check the variance to see if it decreased
var(robdiff1)
#the variance did in fact decrease
```
```{r}
#Difference the data once more to check the variance to see if it decrease
robdiff1diff1 <- diff(robdiff1, lag=1)
var(robdiff1diff1)
#the variance did not decrease so only revert to differencing data once
```
```{r}
#Plot of Transformed and Differenced Data
plot(robdiff1, xlab = "Time(months)", ylab = "cuberoot(# Robberies)", main =
expression(nabla[1]~"Transformed Data"))
```
```{r}
#ACF of Transformed and Differenced Data
acf(robdiff1, lag.max=50)
title("ACF of Transformed and Differenced Data", line = 1)

#PACF of Transformed and Differenced Data
pacf(robdiff1, lag.max=50)
title("PACF of Transformed and Differenced Data", line = 1)
```
```{r}
#AIC VALUES
aiccs <- matrix(NA, nr = 6, nc = 6)
dimnames(aiccs) = list(p=0:5, q=0:5)
for(p in 0:5)
{
for(q in 0:5)
{
aiccs[p+1,q+1] = AIC(arima(robdiff1, order = c(p,1,q), method="ML"))
}
````

````
}
aiccs

(aiccs==min(aiccs))
```



```{r}
#BIC VALUES
biccs <- matrix(NA, nr = 6, nc = 6)
dimnames(aiccs) = list(p=0:5, q=0:5)
for(p in 0:5)
{
for(q in 0:5)
{
biccs[p+1,q+1] = BIC(arima(robdiff1, order = c(p,1,q), method="ML"))
}
}
biccs
(biccs==min(biccs))
```
```{r}
#Model Estimation MLE method

#Model 1 - ARIMA(1,1,2)
fit1<- arima(cub.rob, order = c(1,1,2), method = "ML")
fit1

#Model 2 - ARIMA(2,1,2)
fit2<- arima(cub.rob, order = c(2,1,2), method = "ML")
fit2
```
```{r}
source("plot.roots.txt")
par(mfrow = c(1,3))
#Roots of ARIMA(1,1,2)
fit1 <- arima(cub.rob, order = c(1,1,2), method="ML")
fit1
plot.roots(NULL, polyroot(c(1, -0.3718)), main = "Roots for AR Part - ARIMA(1,1,2)")
plot.roots(NULL, polyroot(c(1, 0.0913, -0.302)), main = "Roots for MA Part - ARIMA(1,1,2)")

#Roots for ARIMA(2,1,2)
fit2 <- arima(cub.rob, order = c(2, 1, 2), method = "ML")
fit2
plot.roots(NULL, polyroot(c(1, -0.2991, 0.1702)), main = "Roots for AR - ARIMA(2,1,2)")
plot.roots(NULL, polyroot(c(1, 0.0266, -0.4532)), main = "Roots for MA - ARIMA(2,1,2) ")
```
```{r}
library(astsa)

model.1 <- arima(robdiff1, order = c(1,1,2))
resids1 = model.1$residuals

model.2 <- arima(robdiff1, order = c(2,1,2))
resids2 = model.2$residuals
````

```
## normality checks

## histogram of residuals of ARMA(1,1,2)
hist(resids1, main="Histogram of Residuals of ARMA(1,1,2)", xlab = "Residuals")

## histogram of residuals of ARMA(2,1,2)
hist(resids2, main="Histogram of Residuals of ARMA(2,1,2)", xlab = "Residuals")
```

```{r}
## qq plot of ARMA(1,1,2)
qqnorm(resids1, main = "Nornal Q-Q Plot - ARMA(1,1,2)")
qqline(resids1)

## qq plot of ARMA(2,1,2)
qqnorm(resids2, main = "Nornal Q-Q Plot - ARMA(2,1,2)")
qqline(resids2)
```

```{r}
## Shapiro test for ARMA(1,1,2)
shap1 <- shapiro.test(resids1)
shap1$statistic # W Statistic
shap1$p.value # P-value

## Shapiro test for ARMA(2,1,2)
shap2 <- shapiro.test(resids2)
shap2$statistic # W Statistic
shap2$p.value # P-value
```

```{r}
## Serial correlation check (Check for Independence)

## Box-Pierce ARIMA(1,1,2)
box_pierce <- Box.test(resids1, lag = 11, type = "Box-Pierce", fitdf = 2)$p.value
box_pierce
## Ljung-Box ARIMA(1,1,2)
ljung_box <- Box.test(resids1, lag = 11, type = "Ljung-Box", fitdf = 2)$p.value
ljung_box

## Box-Pierce ARIMA(2,1,2)
box_pierce <- Box.test(resids2, lag = 11, type = "Box-Pierce", fitdf = 2)$p.value
box_pierce
## Ljung-Box ARIMA(2,1,2)
ljung_box <- Box.test(resids2, lag = 11, type = "Ljung-Box", fitdf = 2)$p.value
ljung_box
```

```{r}
## Constant variance Check

## ACF of residuals ARIMA(1,1,2)

acf(resids1, lag.max=500)
title("ACF Plot \n of Residuals - ARIMA(1,1,2)", line = -1)
## PACF of residuals ARIMA(1,1,2)
pacf(resids1, lag.max=500)
title("PACF Plot \n of Residuals - ARIMA(1,1,2)", line = -1)
```

```
## ACF of residuals ARIMA(2,1,2)
acf(resids2, lag.max=500)
title("ACF Plot \n of Residuals - ARIMA(2,1,2)", line = -1)
## PACF of residuals ARIMA(2,1,2)
pacf(resids2, lag.max=500)
title("PACF Plot \n of Residuals - ARIMA(2,1,2)", line = -1)
```

```{r}
#forecasting
fit = arima(cub.rob, order = c(1, 1, 2), method = "ML", xreg=1 : length(cub.rob))
pred.transf <- predict(fit, n.ahead = 12, newxreg=(length(cub.rob)+1) : length((cub.rob)+12))
# upper bound for the C.I. for transformed data
upper.transf= pred.transf$pred + 2*pred.transf$se
# lower bound
lower.transf= pred.transf$pred - 2*pred.transf$se
#plotting cub.rob and forecasting
ts.plot(cub.rob, xlim=c(1,length(cub.rob)+12), ylim = c(0,max(upper.transf)), xlab =
"Time(months)", ylab = "cuberoot(Robberies in Boston)", main = "Forecasting Based on
Transformed Data")
lines(upper.transf, col="blue", lty="dashed")
lines(lower.transf, col="blue", lty="dashed")
points((length(cub.rob)+1):(length(cub.rob)+12), pred.transf$pred, col="red")
```

```{r}
# RETURN TO ORIGINAL DATA
#get predictions and s.e's of transformed time series
rob.ts2 <- ts(rob)
# back-transform to get predictions of original time series
pred.orig <- pred.transf$pred^3
# bounds of the confidence intervals
upper = upper.transf^3
lower = lower.transf^3

# Plot forecasts using the original data
ts.plot(rob.ts2, xlim=c(1,length(rob.ts2)+12), ylim = c(0,max(upper)), main = "Forecasting
Based on Original Data", xlab = "Time (months)", ylab = "Number of Robberies")
lines(upper, col="blue", lty="dashed")
lines(lower, col="blue", lty="dashed")
points((length(rob.ts2)+1):(length(rob.ts2)+12), pred.orig, col="red")

# Zoomed in  plot of  the last 12 values plus forecast:
ts.plot(rob.ts2, xlim=c(length(rob.ts2)-12,length(rob.ts2)+12), ylim = c(0,max(upper)), xlab =
"Time (months)", ylab = "Number of Robberies", main = "Zoomed In Plot of Forecasting Original
Data")
lines((length(rob.ts2)+1):(length(rob.ts2)+12),upper, lty=2, col="blue")
lines((length(rob.ts2)+1):(length(rob.ts2)+12),lower, lty=2, col="blue")
points((length(rob.ts2)+1):(length(rob.ts2)+12),pred.orig, col="red")
```