Modeling Cellular Phone Subscriptions in India

PSTAT 174

June 6, 2018

Group Theta:
Jeremy Berkov
Maximilian Broekhuis
Lauren Wong
Winson Li
Crystal Chau

Contents

Abstract

The purpose of our project was to accurately forecast the number of mobile cellular phone subscriptions in India, based on previous rates. This is of interest in relation to the recent and massive incorporation of a digital identity system named Aadhaar. Aadhaar is a 12-digit unique identity number that can be obtained by residents of India, based on their biometric and demographic data. In our analysis of the time-series, we used various techniques, including Box-Cox transformations and differencing to make the time series stationary and thus allowing us to identify potential models by looking at ACF and PACF plots.

During our analysis, we were able to come up with many candidate models, however, only one of our candidates was most suitable for forecasting: ARIMA(1,1,2) model. The ARIMA(1,1,2) model passed the Box-Ljung and Shapiro-Wilk tests, is viable for forecasting, and best satisfies the principle of parsimony, compared to our other candidates. Through forecasting we were able to plot a potential trajectory with 95% confidence for four years in the future. Despite certain validation points being outside the confidence interval, our forecasted values still remain valid.

Introduction

The Aadhaar Act, passed in 2016, is a money bill from the Parliament of India that is aimed at providing legal backing for the unique digital identification system implemented by Aadhaar. The Aadhaar Act is also known as the Targeted Delivery of Financial and other Subsidies, Benefits and Services Act, which perfectly describes the motives behind its implementation. Though China is often brought up during discussions of overpopulation and population density, India faces many of the same issues and has been actively solving the problems, evidenced by Aadhaar and its great success. The Aadhaar act has seen authentication success rates for government services of up to 96.4% in 2013, though the rates have fallen to around 88% currently. In general, Aadhaar is aimed at increasing financial inclusion and benefit participation for all Indian citizens, and we believe that looking at mobile cellular phone subscriptions is a strong indicator of these factors.

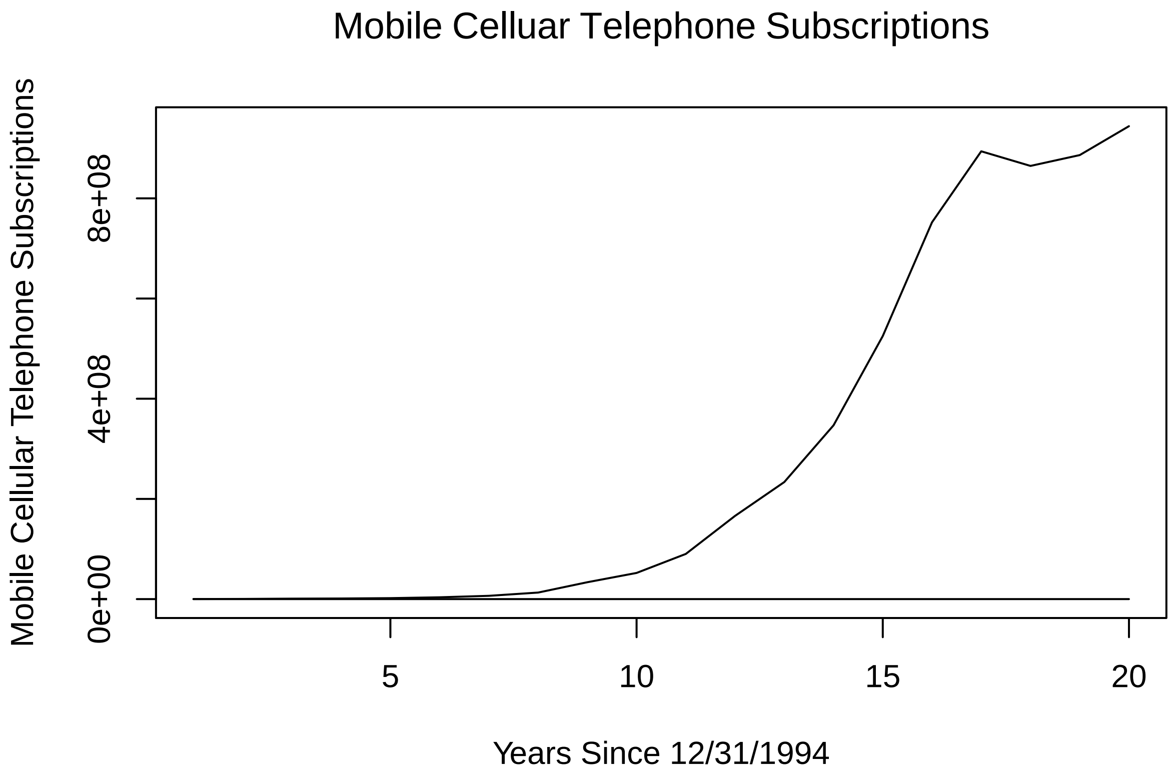
We took our data from the website Quandl.com, which cites United Nations Information and Communication Technology as its source for the data. Though the dataset is defined from 12/31/1960 to 12/31/2014, the cellphone was invented in 1994, and as such the first non-zero values occur at 12/31/1995. From that point until 2014, the number of subscriptions was taken at yearly intervals.

Upon initial observation of the data, we noticed a strong exponential/logarithmic trend and a lack of seasonality. Therefore, before model selection, we decided to difference the data twice, in order to make it stationary, and to apply a Box-Cox transformation, to stabilize the variance. After this, we analyzed ACF and PACF plots of the differenced and transformed data to estimate the best fit model, and from there used R to test the fit of many models by comparing AIC values. When our top choices for possible models were chosen, we applied diagnostic checking techniques such as Shapiro-Wilk (normality of residuals) and Box-Ljung (serial correlation of residuals) to choose our best fit model and used this to make our predictions.

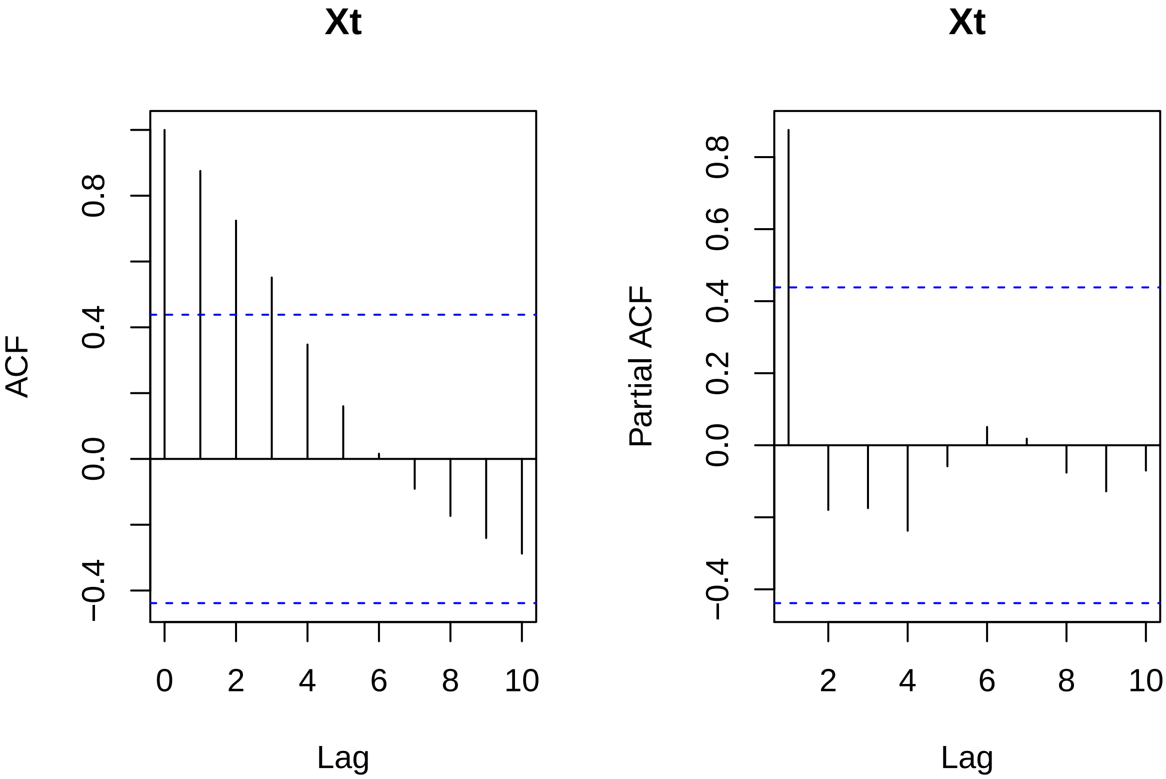
Our final model was an ARIMA(1,1,2), which can be represented as an AR(1) and MA(2), differenced once. We validated and forecasted subscription numbers for the next 5 years using this model.

Initial Time Series Analysis

To start, we plotted the time series of the original data to get an idea of its general form and to identify whether any trend or seasonality is present.

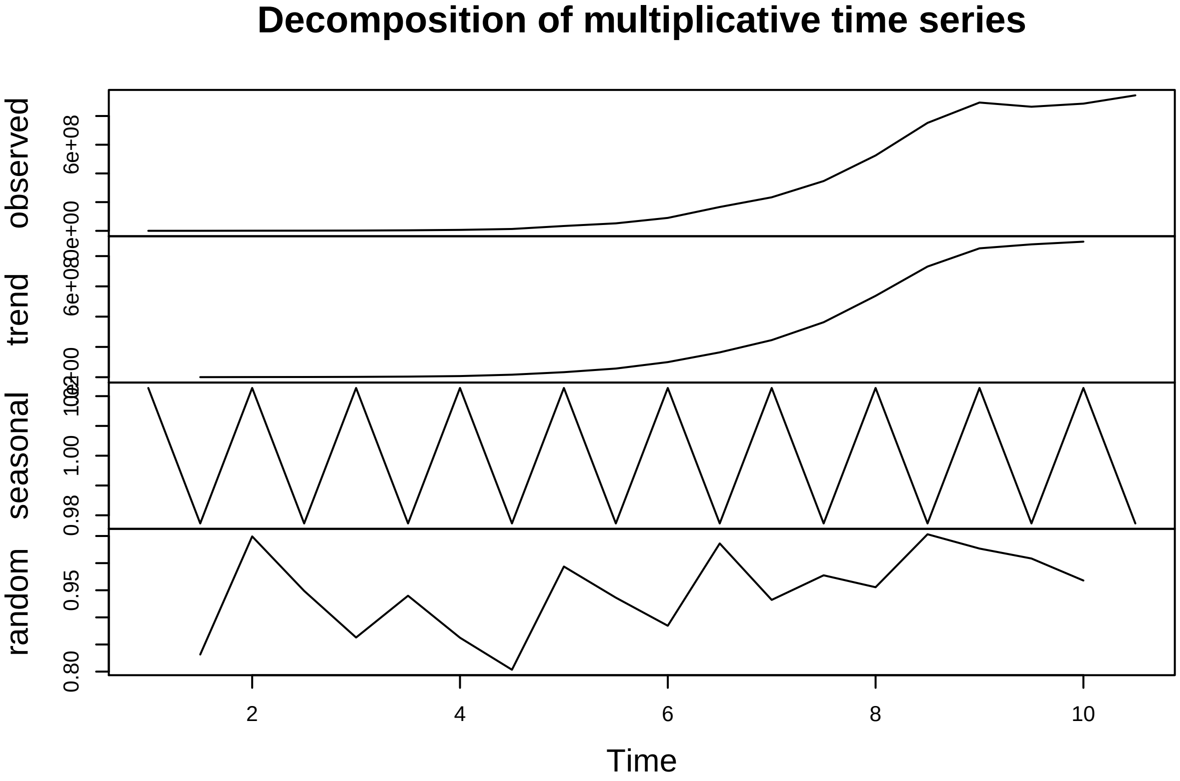


We can see a strong exponential trend from the positive exponential increase of the graph. However, we do seem to see a lack of seasonality.



Although it was already apparent from our initial analysis of the time-series plot, the ACF and PACF plots validate our assumptions that the series is not stationary and has a trend component.

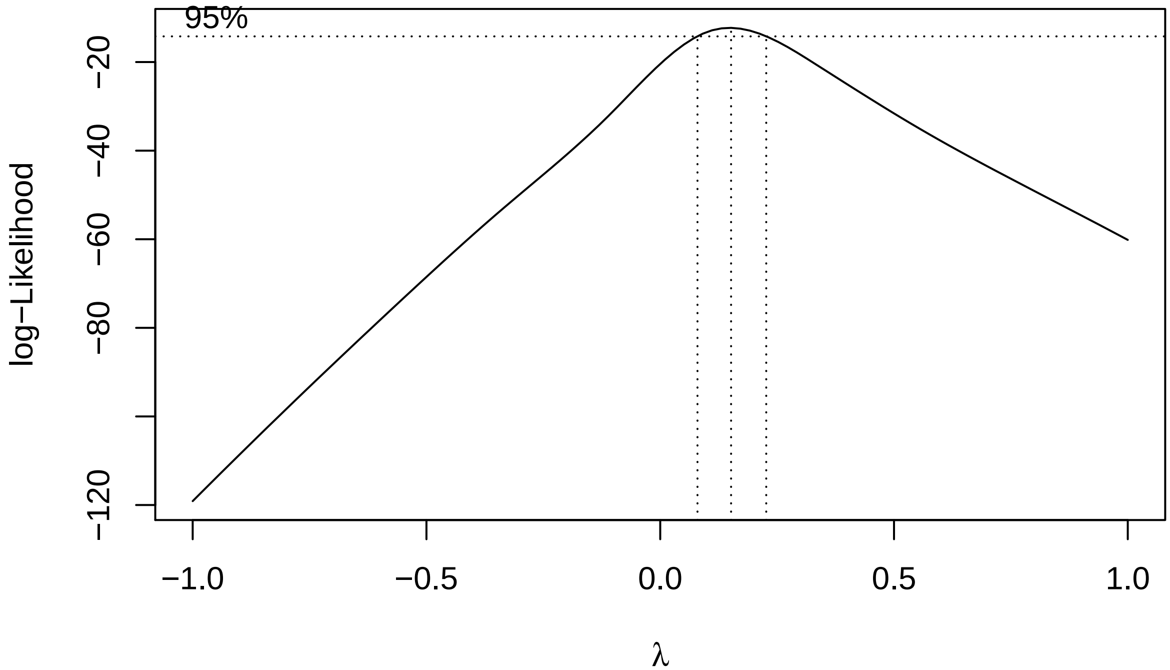
The decompose function can be helpful in visualizing the different components of the model, such as trend, seasonality, and stationarity.

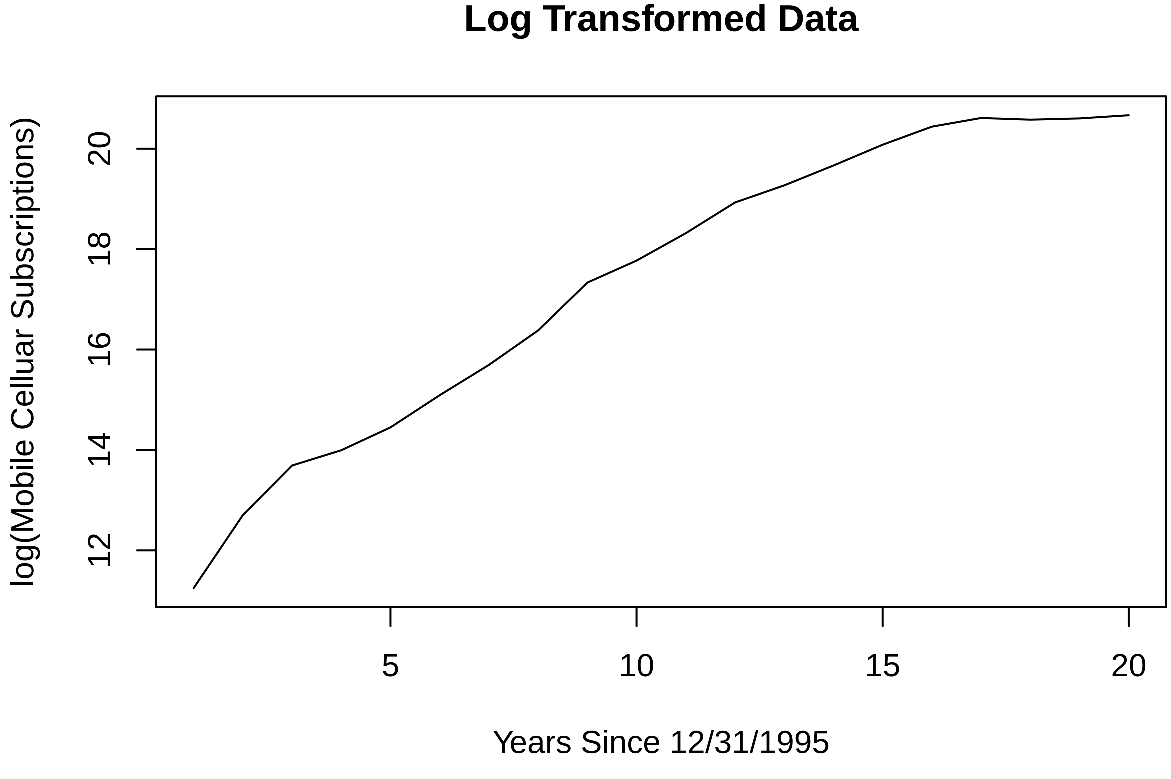


From this chart, we can see that the underlying trend in the graph plays a defining role in its shape. The seasonality is artificial, as we used a frequency = 2 for 20 data points, creating the seasonal effect seen here as well as changing the time interval to 1-10 years.

Box Cox Transformation

Box Cox transformations can help us deal with the problem of heteroscedasticity in our data. However, we must plot a 95% confidence interval to see what value of lambda can maximize our log-likelihood.





```
## [1] 9.103581
```

```
## [1] 1.3622e+17
```

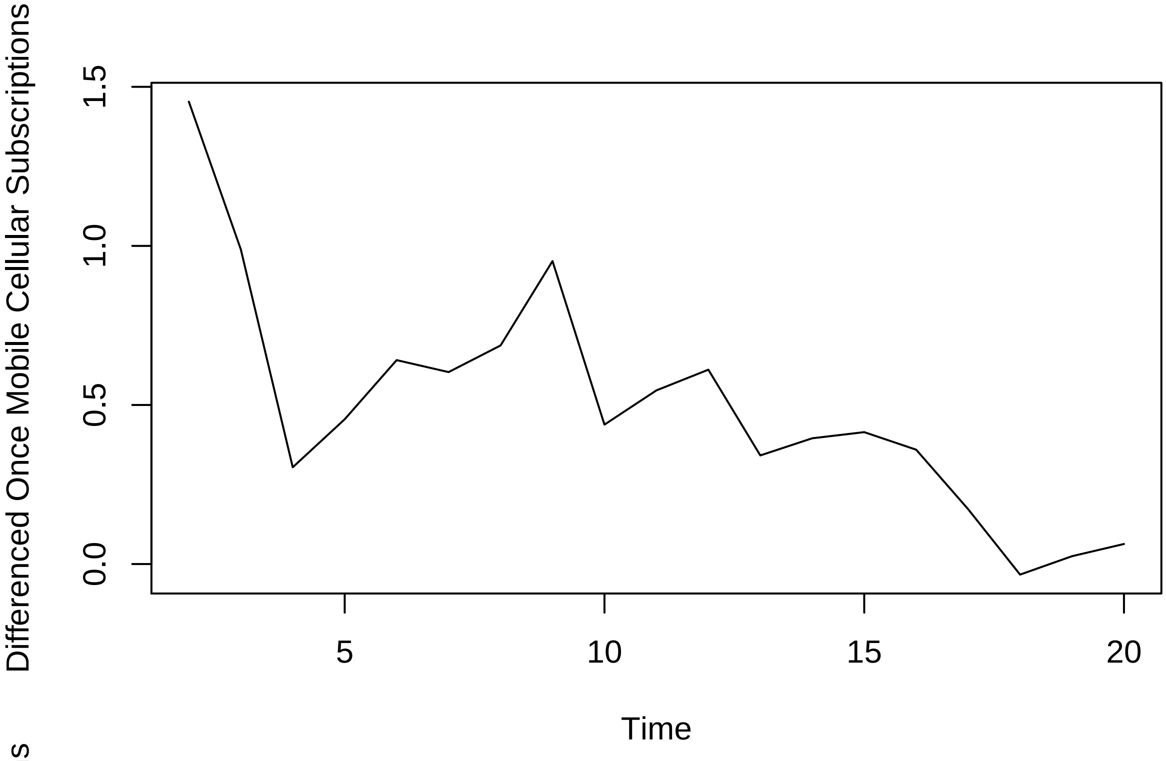
```
## [1] 124605.5
```

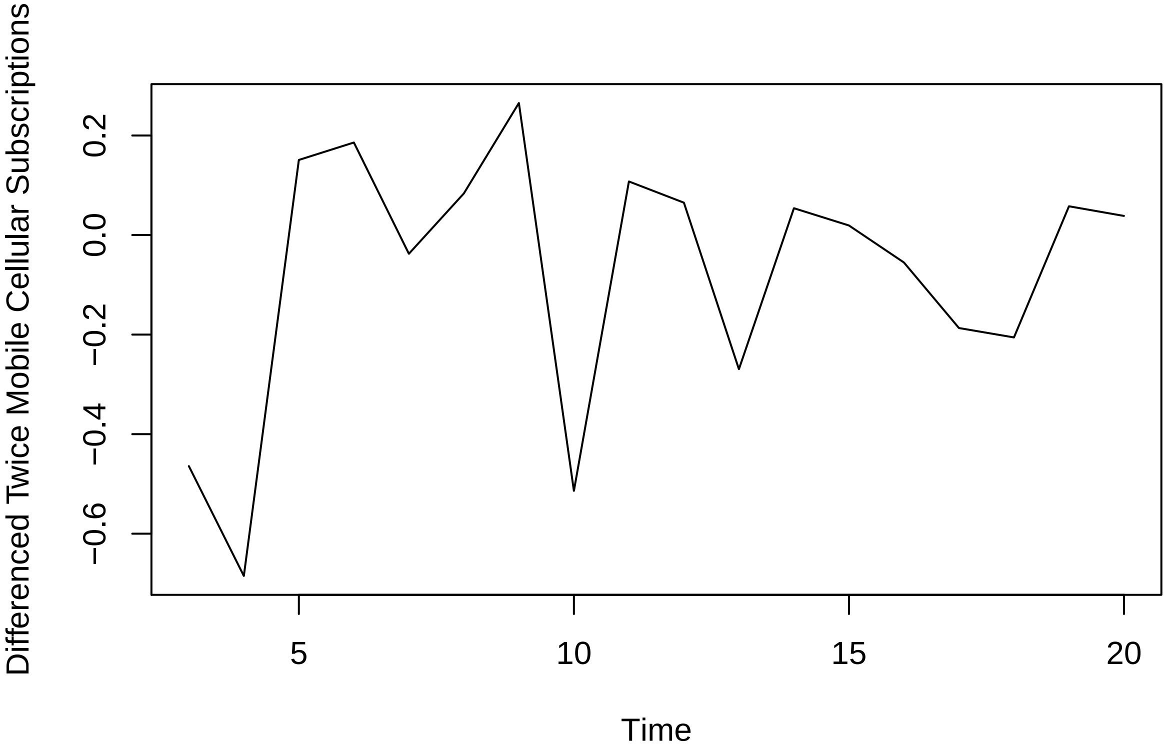
The variance of the log of transformed data is 9.103581.

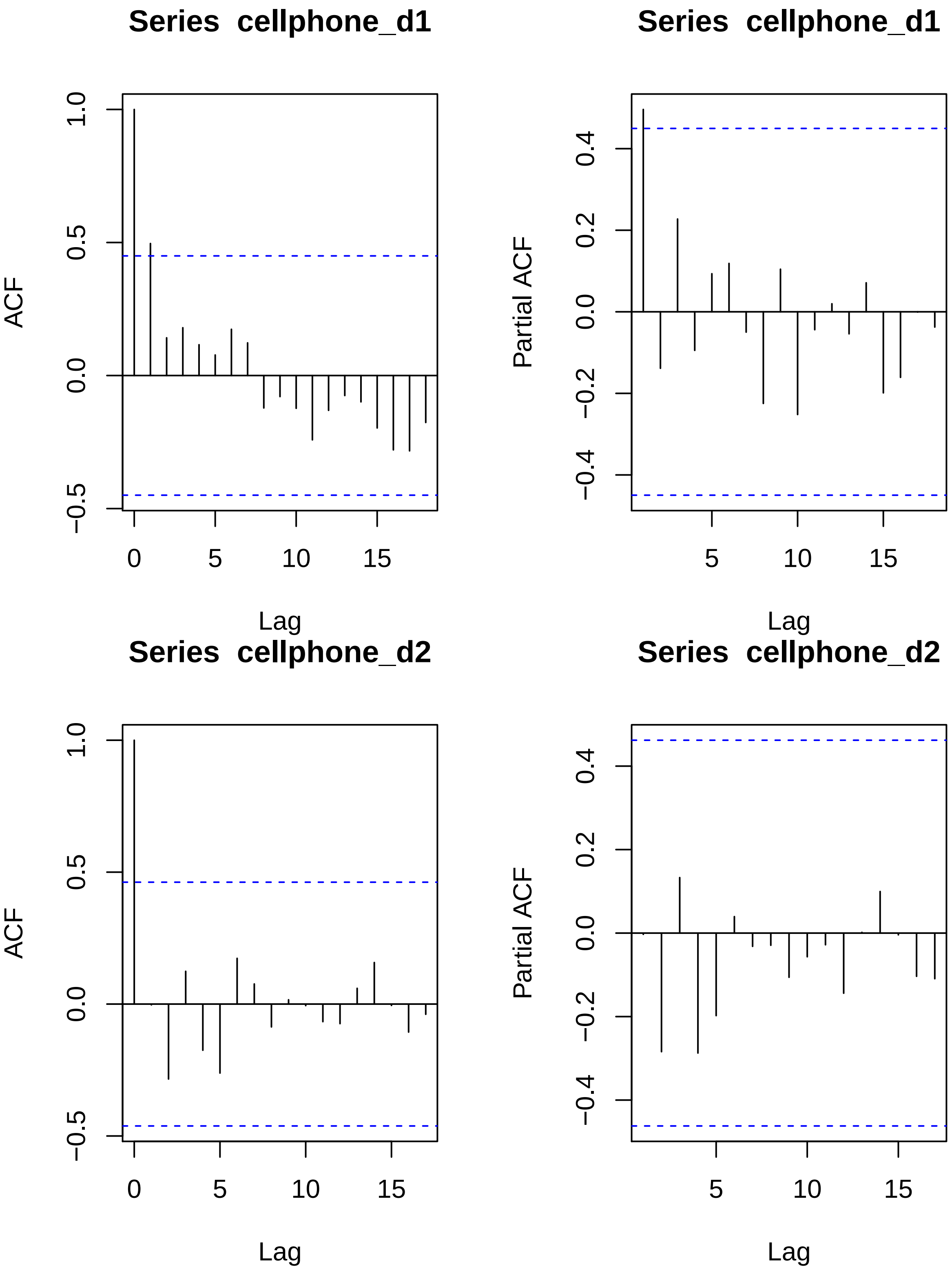
Removing Trend and Seasonality

```
## [1] 0.1304289
```

```
## [1] 0.06789504
```







Upon review of our differenced data plot and their ACF and PACF it appears that our data still displays a significant trend component. The variance of differencing once is: 0.1304289. We difference a second time and re-assess. Differencing a second time further decreases the variance: 0.06789504.

From our twice differenced data plots, ACF, and PACF we can see that the trend is removed. Unfortunately, the twice differenced data creates a unit root in our ARIMA model, the data has been over-differenced.

This gives us significant reason to choose a once differenced model. Our ACF and PACF plots suggest an MA model, but since it is not entirely clear whether there is no AR component we look at AIC values to determine candidate models.

ARMA Models

We will test various combinations of ARIMA models with different p and q parameters using a for loop that returns AIC values up to ARIMA(5,1,5).

```
##    q
## p         0        1        2        3        4        5
##   0 749.4307 733.2720 731.2272 734.1572 735.9515 740.0787
##   1 733.0263 730.0148 732.7141 736.3564 739.6483 745.5028
##   2 734.1342 732.9911 736.4021 739.6944 744.6897 750.8915
##   3 734.7855 736.0861 739.5459 744.6878 750.8018 758.3994
##   4 738.2302 740.4472 746.9700 750.8903 758.4002 768.1925
##   5 742.5449 745.5884 753.0688 757.6202 766.1137 778.3194
```

After running the loop and obtaining the necessary AIC values for several different ARMA models up to ARMA(5,1,5), we concluded that the following models with the smallest AICs are: ARMA(1,1,0), ARMA(0,1,1), ARMA(0,1,2), ARMA(1,1,1), ARMA(2,1,0), ARMA(1,1,2).

Instead of testing all six models, we will conduct diagnostic testing select models from the previous section, indicated by our ACF and PACFs. Our candidate models are: ARMA(1,1,1), ARMA(2,1,0), ARMA(0,1,1), ARMA(1,1,2).

Diagnostics Checking on Test Models

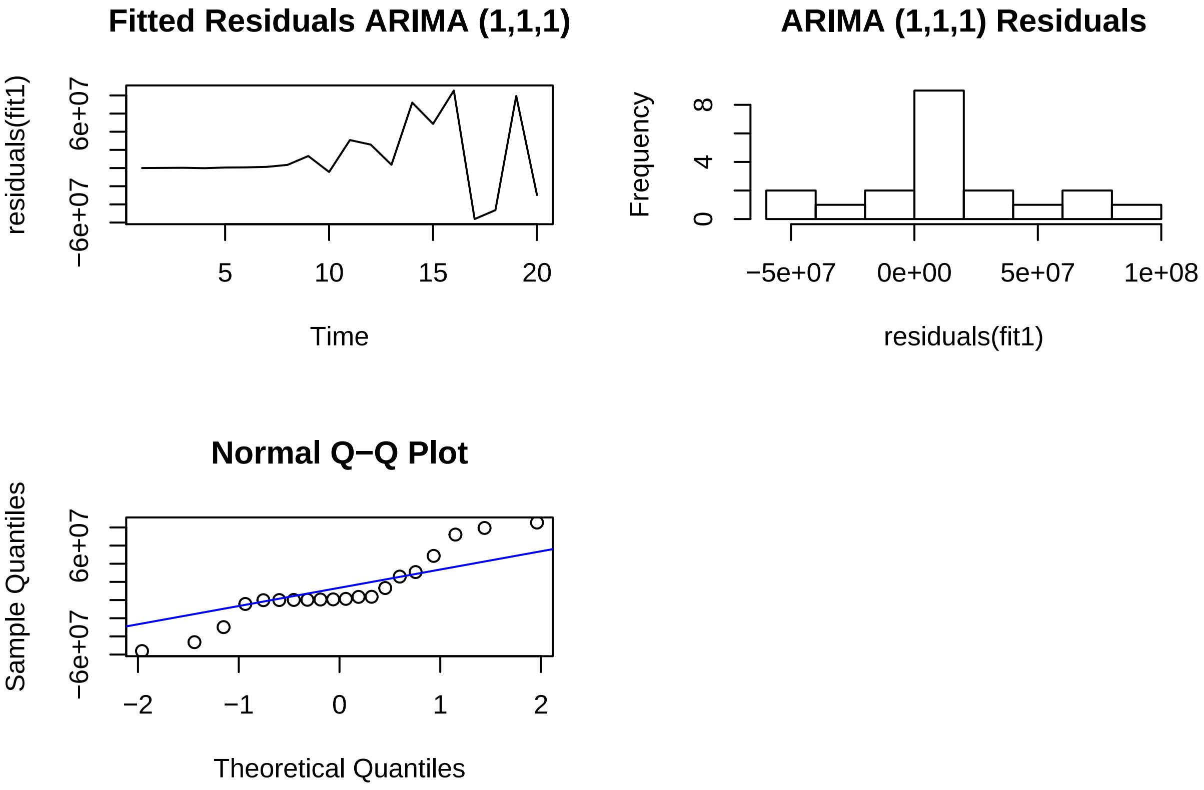
ARIMA(1,1,1)

```
##
## Call:
## arima(x = cellphone.ts, order = c(1, 1, 1), method = "ML")
##
## Coefficients:
##          ar1     ma1
##       0.5159  1.0000
## s.e.  0.1839  0.1741
##
## sigma^2 estimated as 1.536e+15:  log likelihood = -361.21,  aic = 728.41
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit1)
## X-squared = 0.23145, df = 1, p-value = 0.6304
```

```
##
##  Box-Pierce test
##
## data:  residuals(fit1)
## X-squared = 0.19989, df = 1, p-value = 0.6548
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit1)
## W = 0.90318, p-value = 0.04732
```



The p-values for the Box-Ljung and Box-Pierce test are greater than our significance level of 0.05. Our model however doesn't satisfy the normality condition, and there is an ar1 coefficient of 0, so we will not include this model in our forecasting.

ARIMA(2,1,0)

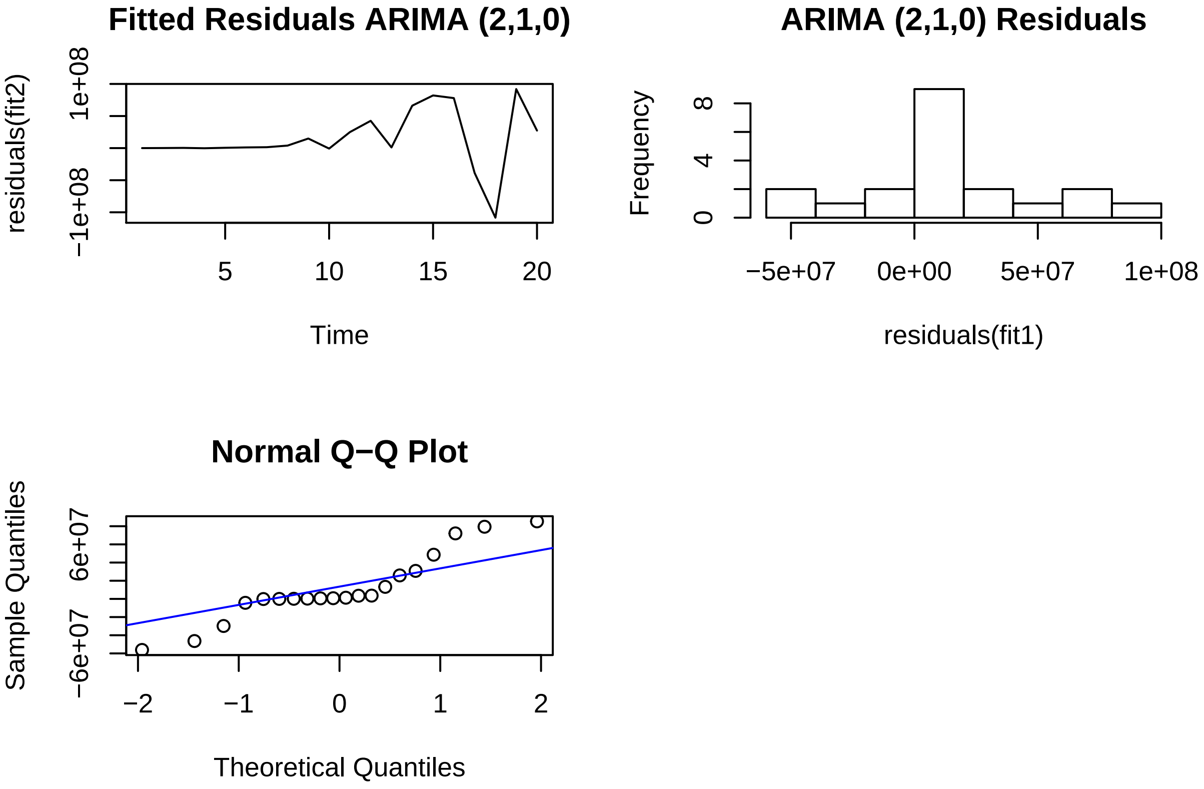
```
##
## Call:
## arima(x = cellphone.ts, order = c(2, 1, 0), method = "ML")
##
## Coefficients:
##          ar1      ar2
##       1.0196  -0.2872
## s.e.  0.2132   0.2101
##
## sigma^2 estimated as 2.228e+15:  log likelihood = -363.27,  aic = 732.53
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit2)
## X-squared = 0.12419, df = 1, p-value = 0.7245
```

```
##
```

```
##  Box-Pierce test
##
## data:  residuals(fit2)
## X-squared = 0.10726, df = 1, p-value = 0.7433
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit2)
## W = 0.86901, p-value = 0.01129
```



The p-values for the Box-Ljung and Box-Pierce test are greater than our significance level of 0.05. Our model however doesn't satisfy the normality condition, so we will not include this model in our forecasting.

ARIMA(0,1,1)

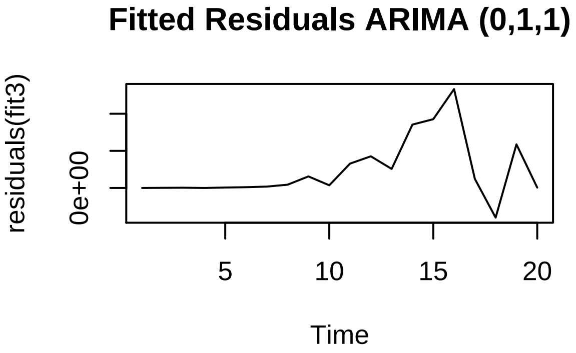
```
##
## Call:
## arima(x = cellphone.ts, order = c(0, 1, 1), method = "ML")
##
## Coefficients:
##          ma1
##       1.0000
## s.e.  0.1367
##
## sigma^2 estimated as 2.246e+15:  log likelihood = -364.26,  aic = 732.52
```

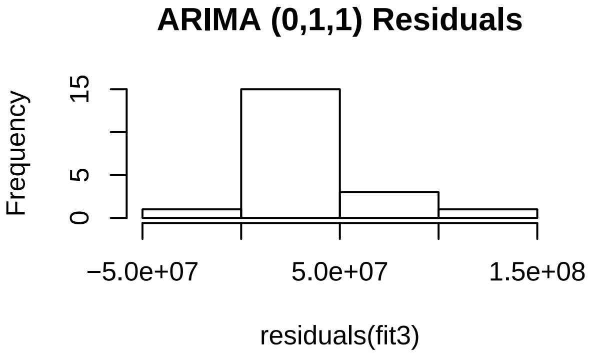
```
##
##  Box-Ljung test
##
```

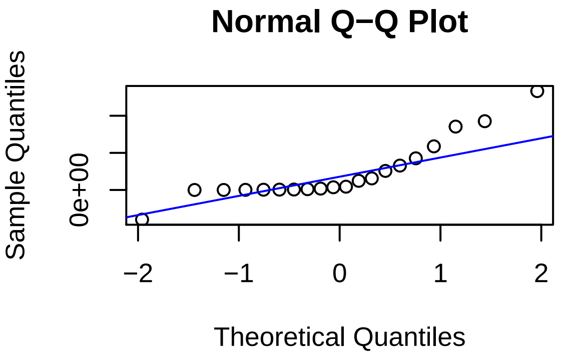
```
## data:  residuals(fit3)
## X-squared = 3.5551, df = 1, p-value = 0.05936
```

```
##
##  Box-Pierce test
##
## data:  residuals(fit3)
## X-squared = 3.0703, df = 1, p-value = 0.07974
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit3)
## W = 0.8251, p-value = 0.002097
```







The p-values for the Box-Ljung and Box-Pierce test are greater than our significance level of 0.05. Our model however doesn't satisfy the normality condition, therefore we will not include this model in our forecasting.

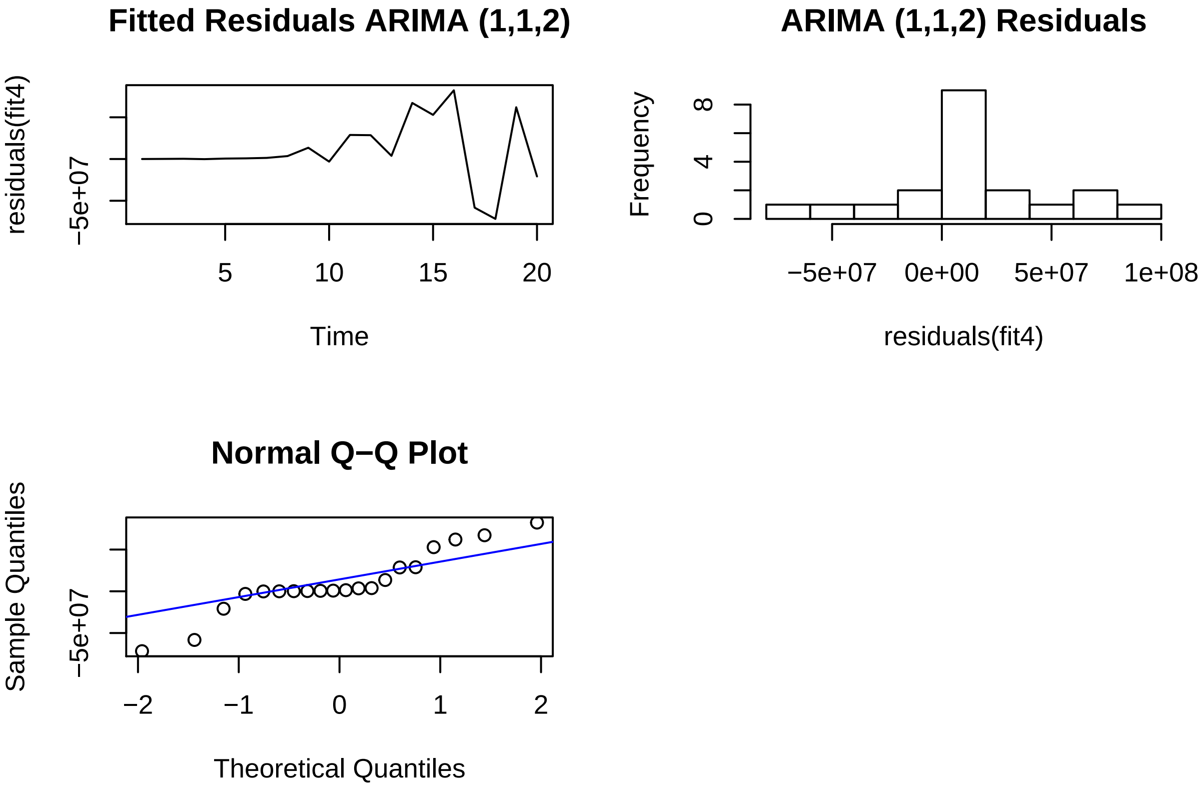
ARIMA(1,1,2)

```
##
## Call:
## arima(x = cellphone.ts, order = c(1, 1, 2), method = "ML")
##
## Coefficients:
##          ar1     ma1      ma2
##       0.7671  0.6173  -0.3826
## s.e.  0.2547  0.4370   0.4110
##
## sigma^2 estimated as 1.514e+15:  log likelihood = -360.93,  aic = 729.86
```

```
##
##  Box-Ljung test
##
## data:  residuals(fit4)
## X-squared = 0.0087968, df = 1, p-value = 0.9253
```

```
##
##  Box-Pierce test
##
## data:  residuals(fit4)
## X-squared = 0.0075972, df = 1, p-value = 0.9305
```

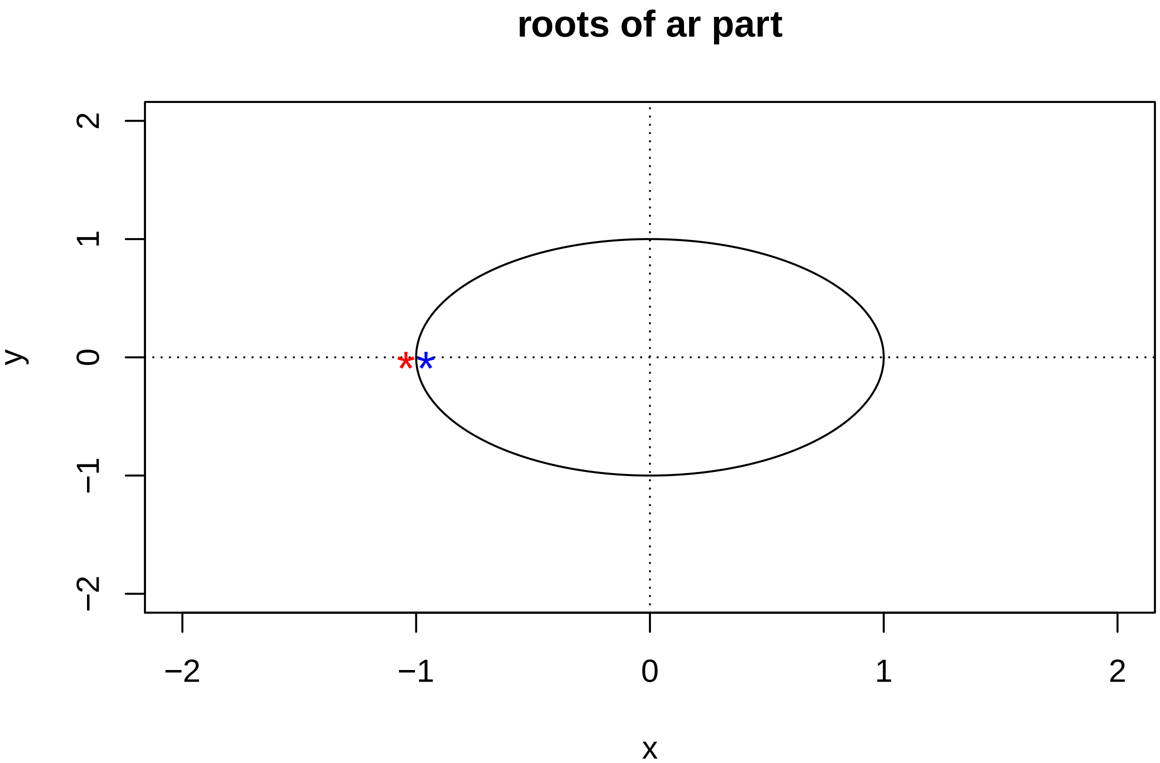
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fit4)
## W = 0.91011, p-value = 0.06406
```

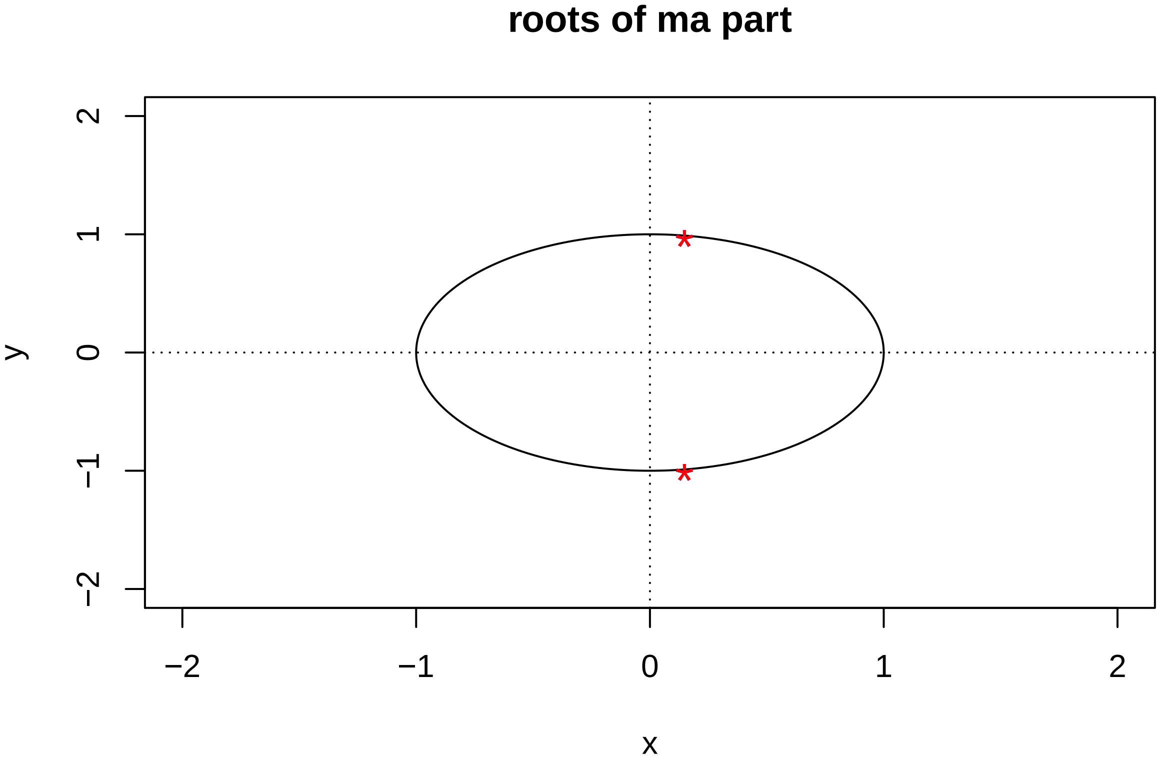


The p-values for the Box-Ljung and Box-Pierce test are greater than our significance level of 0.05 and our model satisfies the normality condition, so we will include this model in our forecasting. Therefore, all of our tests for ARIMA(1,1,2) seem to be significant, allowing it to be a viable model for forecasting after checking for invertibility and stationarity by looking at the unit roots.

Analyzing Roots

The next step is to check the unit roots of ARIMA(1,1,2) for stationarity and invertibility.

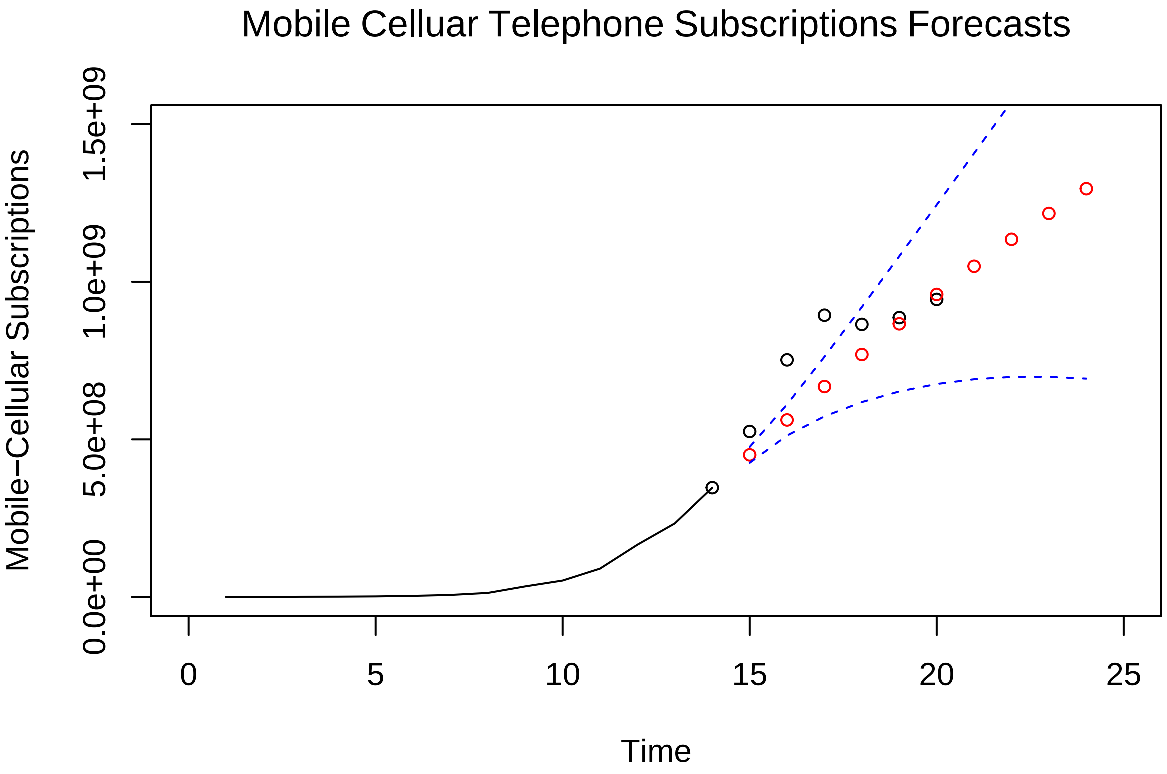




We see that the root of ARIMA(1,1,2) lies outside of the unit circle while the inverse root lies inside, so the model is causal invertible. The roots seem to be very close to the unit circle so we take the square root of the sum of the real and imaginary parts squared to give us an estimate of the distance from the center, giving us 1.000004851 as the distance from the origin for both roots. Since this value is so close to the unit root, this can be a problem for the stationarity condition.

Forecasting

```
## $pred
## Time Series:
## Start = 15
## End = 24
## Frequency = 1
##  [1]  450935836  561613058  667645364  769227690  866546795  959781598
##  [7] 1049103509 1134676745 1216658632 1295199891
##
## $se
## Time Series:
## Start = 15
## End = 24
## Frequency = 1
##  [1]  12853076  25177579  48280354  76946596 109457257 144950741 182863691
##  [8] 222781320 264378868 307392567
```



We removed the last 6 observed values and predicted 10 points, which gave us predictions for 4 years after our last observed point. The observed value is shown as the black circles and our prediction is shown as the red circles. The blue dashed lines gives the 95% confidence interval of our predictions.

Conclusion

Though not all of our validation points are within the 95% CI, the majority of them are and, furthermore, the validation points converge towards our predictions as time goes on. Looking at the original time series, we can see how this might be possible. The period that we are validating exhibits atypical behavior compared to the rest of the graph. There are unusual amounts of growth and then a slight dip. However, with time, we can see that our predictions are indeed valid and accurate, as our prediction for year 20 and our validation point are nearly the same. In conclusion, we believe that cellphone usage is a good indicator of Aadhaar's success, and by modeling the growth in usage, we would expect to see similar rates of growth in Aadhaar and financial inclusion. Furthermore, we have accomplished our goal of understanding what future rates of cellphone subscriptions/usage might look like in India, arriving at our final model of ARIMA(1,1,2).

ARIMA(1,1,2) Model: $Y_t = 0.9580X_{t-1} - (-0.2945)Z_t - Z_{t-2} + Z_t$, where $Y_t \sim WN(0, \sigma^2)$

References

Cellular Phone Use - India, 2014, United Nations Information and Communication Technology, data available on the World Wide Web (Quandl), accessed May 23, 2018, at URL https://www.quandl.com/data/UICT/CELL_IND-Cellular-Phone-Use-India

Appendix

Initial Time Series Analysis

```r
library(MuMIn)
library(MASS)
cellphone = read.csv("/Users/laurenwong/Downloads/UICT-CELL_IND.csv", nrows = 20,
                     colClasses = c(NA,NA,"NULL"))
cellphone = cellphone[nrow(cellphone):1,]
cellphone.ts = ts(cellphone, frequency = 1)
ts.plot(cellphone.ts, ylab = "Mobile Cellular Telephone Subscriptions",
        xlab = "Years Since 12/31/1994")
title(expression(Mobile~Celluar~Telephone~Subscriptions))
par(mfrow=c(1,2))
cellphone = read.csv("/Users/laurenwong/Downloads/UICT-CELL_IND.csv", nrows = 20,
                     colClasses = c("NULL",NA,"NULL"))
cellphone = cellphone[nrow(cellphone):1,]
cellphone.ts = ts(cellphone, frequency = 1)
acf(cellphone.ts, lag.max = 10, main = "Xt")
pacf(cellphone.ts, lag.max = 10, main = "Xt")
cellphone.ts = ts(cellphone, frequency = 2)
decompose_cellphone = decompose(cellphone.ts, type = "multiplicative")
plot(decompose_cellphone)
```

Box Cox Transformation

```r
cellphone = read.csv("/Users/laurenwong/Downloads/UICT-CELL_IND.csv", nrows = 20,
                     colClasses = c(NA,NA,"NULL"))
cellphone = cellphone[nrow(cellphone):1,]
cellphone.ts = ts(cellphone, frequency = 1)
bcTransform = boxcox(cellphone.ts ~ as.numeric(1:length(cellphone.ts)),
                     lambda = seq(-1, 1, length = 10))
#log transformation
cellphone.tr = log(cellphone.ts)
cellphone.test.tr = cellphone.ts^(1/3)
ts.plot(cellphone.tr, xlab = "Years Since 12/31/1995",
        ylab = "log(Mobile Celluar Subscriptions)", main = "Log Transformed Data")
var(cellphone.tr)
var(cellphone.ts)
var(cellphone.test.tr)
```

Removing Trend and Seasonality

```
#differencing
cellphone_d1 = diff(cellphone.tr, differences = 1)
var(cellphone_d1)
cellphone_d2 = diff(cellphone_d1, differences = 1)
var(cellphone_d2)
ts.plot(cellphone_d1, ylab = "Differenced Mobile Cellular Subscriptions")
ts.plot(cellphone_d2, ylab = "Differenced Mobile Cellular Subscriptions")
par(mfrow=c(1,2))
acf(cellphone_d1, lag.max = 60)
pacf(cellphone_d1, lag.max = 60)
acf(cellphone_d2, lag.max = 60)
pacf(cellphone_d2, lag.max = 60)
```

ARMA Models

```
aicc <- matrix(NA, nr = 6, nc = 6)
dimnames(aicc) = list(p= 0:5,q = 0:5)
for (p in 0:5)
{
  for (q in 0:5)
  {
    aicc[p+1,q+1] = AICc(arima(cellphone.ts, order = c(p,1,q), method = "ML"))
  }
}
aicc
```

Diagnostic Checking on Test Models

```
#(1,1,1)
fit1 = arima(cellphone.ts, order=c(1,1,1), method="ML")
fit1
Box.test(residuals(fit1), type="Ljung")
Box.test(residuals(fit1), type ="Box-Pierce")
shapiro.test(residuals(fit1))
op <- par(mfrow=c(2,2))
ts.plot(residuals(fit1),main = "Fitted Residuals ARIMA (1,1,1)")
hist(residuals(fit1),main = "ARIMA (1,1,1) Residuals")
qqnorm(residuals(fit1))
qqline(residuals(fit1),col ="blue")
#(2,1,0)
fit2 = arima(cellphone.ts, order = c(2, 1, 0), method = "ML")
fit2
Box.test(residuals(fit2), type="Ljung")
Box.test(residuals(fit2), type ="Box-Pierce")
```

```
shapiro.test(residuals(fit2))
op <- par(mfrow=c(2,2))
ts.plot(residuals(fit2),main = "Fitted Residuals ARIMA (2,1,0)")
hist(residuals(fit1),main = "ARIMA (2,1,0) Residuals")
qqnorm(residuals(fit1))
qqline(residuals(fit1),col ="blue")
#(0,1,1)
fit3 = arima(cellphone.ts, order = c(0,1,1), method = "ML")
fit3
Box.test(residuals(fit3), type="Ljung")
Box.test(residuals(fit3), type ="Box-Pierce")
shapiro.test(residuals(fit3))
op <- par(mfrow=c(2,2))
ts.plot(residuals(fit3),main = "Fitted Residuals ARIMA (0,1,1)")
hist(residuals(fit3),main = "ARIMA (0,1,1) Residuals")
qqnorm(residuals(fit3))
qqline(residuals(fit3),col ="blue")
#(1,1,2)
fit4 = arima(cellphone.ts, order = c(1,1,2), method = "ML" )
fit4
Box.test(residuals(fit4), type="Ljung")
Box.test(residuals(fit4), type ="Box-Pierce")
shapiro.test(residuals(fit4))
op <- par(mfrow=c(2,2))
ts.plot(residuals(fit4),main = "Fitted Residuals ARIMA (1,1,2)")
hist(residuals(fit4),main = "ARIMA (1,1,2) Residuals")
qqnorm(residuals(fit4))
qqline(residuals(fit4),col ="blue")

#Analyzing roots
plot.roots <- function(ar.roots=NULL, ma.roots=NULL, size=2, angles=FALSE,
                       special=NULL, sqecial=NULL,my.pch=1,first.col="blue",
                       second.col="red",main=NULL)
{xylims <- c(-size,size)
omegas <- seq(0,2*pi,pi/500)
temp <- exp(complex(real=rep(0,length(omegas)),imag=omegas))
plot(Re(temp),Im(temp),typ="l",xlab="x",ylab="y",xlim=xylims,ylim=xylims,main=main)
abline(v=0,lty="dotted")
abline(h=0,lty="dotted")
if(!is.null(ar.roots))
{
  points(Re(1/ar.roots),Im(1/ar.roots),col=first.col,pch=my.pch)
  points(Re(ar.roots),Im(ar.roots),col=second.col,pch=my.pch)
}
if(!is.null(ma.roots))
{
  points(Re(1/ma.roots),Im(1/ma.roots),pch="*",cex=1.5,col=first.col)
  points(Re(ma.roots),Im(ma.roots),pch="*",cex=1.5,col=second.col)
}
if(angles)
{
  if(!is.null(ar.roots))
  {
    abline(a=0,b=Im(ar.roots[1])/Re(ar.roots[1]),lty="dotted")
    abline(a=0,b=Im(ar.roots[2])/Re(ar.roots[2]),lty="dotted")
  }
  if(!is.null(ma.roots))
  {
    sapply(1:length(ma.roots), function(j) abline(a=0,b=Im(ma.roots[j])/Re(ma.roots[j]),
                                                  lty="dotted"))
  }
}
if(!is.null(special))
{
  lines(Re(special),Im(special),lwd=2)
}
if(!is.null(sqecial))
{
  lines(Re(sqecial),Im(sqecial),lwd=2)
}
}
plot.roots(NULL, polyroot(c(1,.9580)), main = 'roots of ar part')
plot.roots(NULL, polyroot(c(1, -0.295, 1)), main = 'roots of ma part')
```

Forecasting

```
cellphone = read.csv("/Users/laurenwong/Downloads/UICT-CELL_IND.csv", nrows = 20,
                     colClasses = c(NA,NA,"NULL"))
cellphone = cellphone[nrow(cellphone):1,]
cellphone.test = cellphone$Mobile.Cellular.Telephone.Subscriptions[14:20]
cellphone.test.ts = ts(cellphone.test, start = c(14,1))
cellphone = cellphone$Mobile.Cellular.Telephone.Subscriptions[1:14]
cellphone.ts = ts(cellphone, frequency = 1)

ARIMA112 = arima(cellphone.ts, order = c(1,1,2), method = 'ML')

mypred1 = predict(ARIMA112, n.ahead = 10)
mypred1
ts.plot(cellphone.ts, xlim = c(0, 25), ylim = c(0, 1500000000),
        ylab = "Mobile-Cellular Subscriptions")
points(cellphone.test.ts, cex=0.8, pch=1, col="black")
points(mypred1$pred, col='red', cex=0.8)
lines(mypred1$pred+1.96*mypred1$se,lty=2,col="blue")
lines(mypred1$pred-1.96*mypred1$se,lty=2,col="blue")
title(expression(Mobile~Celluar~Telephone~Subscriptions~Forecasts))
```