Asylum Applicants Final Project Report: Non-Technical Summary John Attisha and Eva Hare
CSC 425: Time Series Analysis and Forecasting

FORECASTING ASYLUM APPLICANTS FROM SYRIA TO GERMANY

The international news is headlined on a near daily basis with stories of the unprecedented flood of migrants from the Middle East seeking asylum. Their most frequent destinations are member countries of the European Union; so frequent in fact that the "European migrant crisis" has its own Wikipedia page[1].

For our analysis, we extracted data from the United Nations High Commissioner for Refugees (UNHCR) Population Statistics Reference Database on monthly asylum applicants from January 1999 through October 2015[2]. The data lists the number of asylum applicants on a monthly basis by country of origin and country of asylum. Based on this database, we attempted to construct a model to forecast the number of asylum applicants from Syria to Germany for November 2015 through March 2016.

Through our analysis, we determined the monthly number of asylum applicants from Syria to Germany can be modeled through time series analysis. The behavior of the data shows an exponential growth consistent with reports seen in the nightly news and mainstream media.

Our most accurate model indicates that the current number of asylum applicants depends on the shocks to the system for the prior two months, as well as the shocks from one year prior, with the greatest impact coming from the one month prior. The inclusion of the prior two months in the model is in keeping with our real-world expectation that asylum applicants are driven by current events, be they positive, such as a ceasefire, or negative, such as the shelling of a village. Once such an event occurs, it takes approximately two months (as evidenced by our model) for the applicants driven by that particular event to travel from Syria to Germany to have their asylum applications included in the database.

Inclusion of the twelve-month prior term in the model demonstrates what is known as a seasonal effect. In this case, it indicates that people are somewhat more likely to leave their homes in the summer as compared to the winter. This may be due to increased travel challenges in the colder months, or that there are fewer negative events (such as attacks) in the winter, resulting in less impetus to move.

Forecasting using our model indicates a slight decline in the number of asylum applicants from October to November 2015, followed by an upward trend for December 2015 through March 2016. This slight decline is in keeping with the seasonal behavior exhibited in previous years where the number of applicants slightly decreases in the winter, followed by an increase throughout the rest of the year. While the forecast of an increasing trend of asylum applicants from Syria to Germany is unfortunate on a humanitarian basis, it comes as no surprise given that the Syrian Civil War continues unabated, forcing more people to leave their homes and seek asylum in another country.

[1] https://en.wikipedia.org/wiki/European_migrant_crisis

[2] http://popstats.unhcr.org/en/asylum_seekers_monthly. Extracted from UNHCR Statistics Reference Database. Date Extracted 2015-11-26 04:22:38 +01:00.

EXPLORATORY ANALYSIS

We began our analysis by creating a plot of the time series for two subsets of the data: asylum applications from Syria, Afghanistan and Iraq to the European Union (Figure 1) and from Syria to Germany (Figure 2). Both datasets consist of 202 data points corresponding to monthly observations from January 1999 to October 2015. Both plots demonstrate exponential behavior, indicating we should apply a transformation to the data.

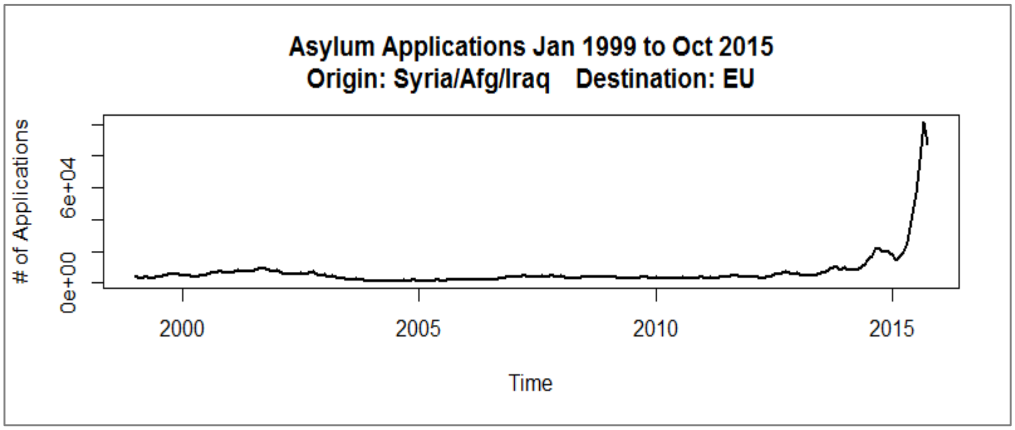

Figure 1

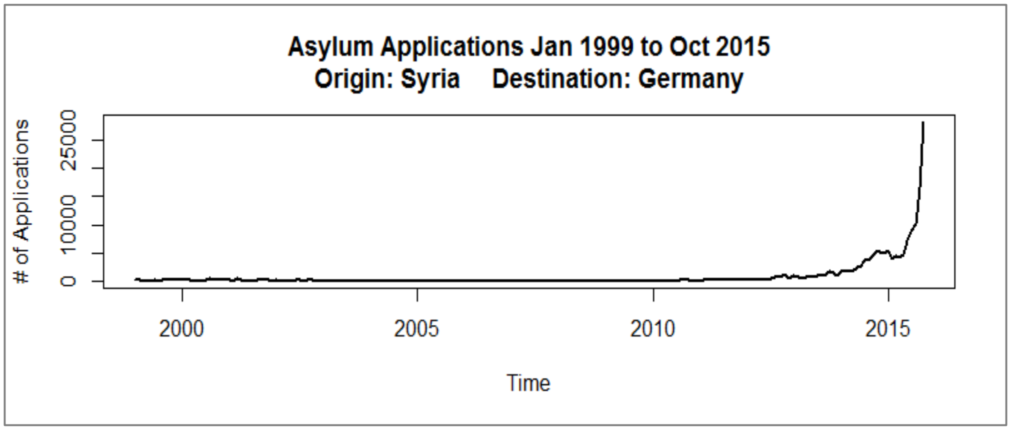

Figure 2

Performing a log transformation reveals that both time series exhibit non-stationary behavior. Figure 3 exhibits clear signs of seasonality during the summer months between 2010 and 2015. As the behavior of asylum applicants from Syria to Germany (Figure 4) is slightly more consistent in terms of an overall trend, we continue our analysis with that particular subset.

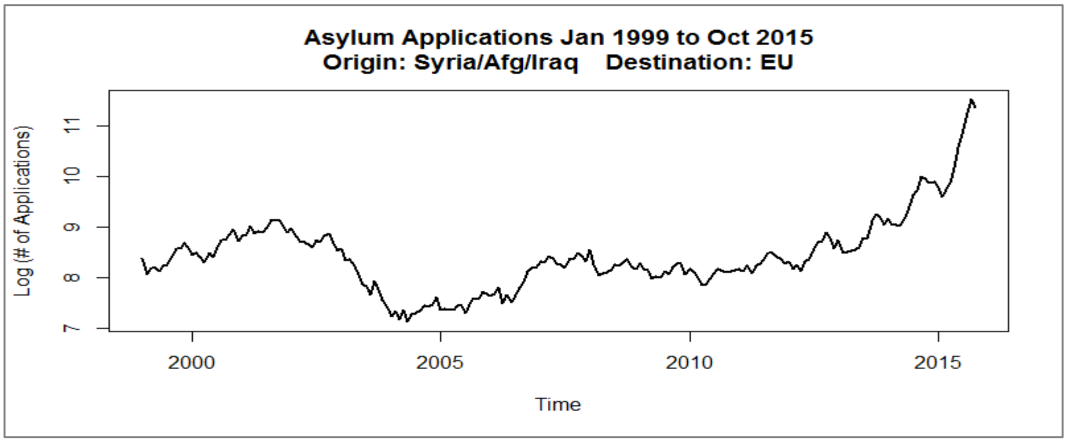

Figure 3

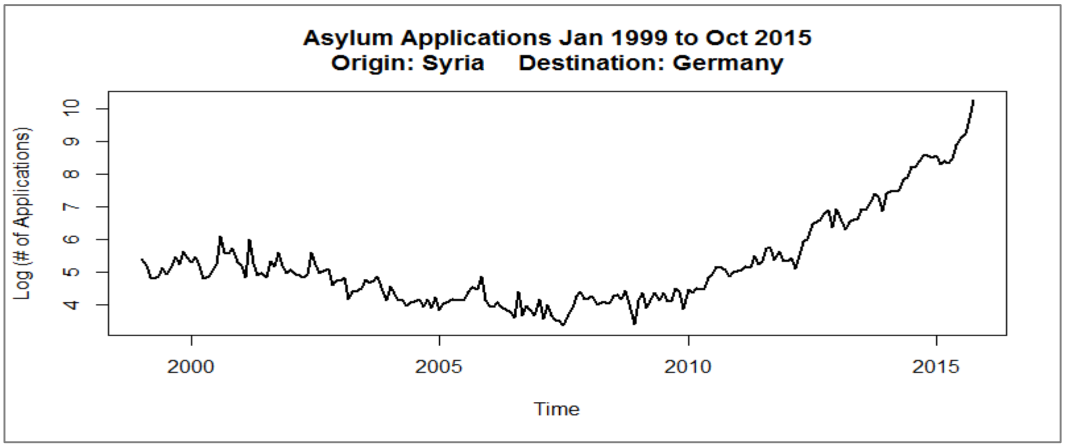

Figure 4

Dataset: Asylum Applications from Syria to Germany

Plotting the autocorrelation function for the log data confirms the time series exhibits non-stationary behavior with a very slow decay[3]. This indicates the series may be unit-root non-stationary. We further confirm this by running the Dickey-Fuller tests for unit-root non-stationary behavior on the log time series, all of which result in p-values near 1.

After computing the first difference of the logged data and constructing both a time plot (Figure 5) and an ACF plot (Figure 6) of this series, we observe a much more stationary behavior. We again run Dickey-Fuller tests for the first-differenced logged time series, which results in p-values less than 0.01. These findings suggest first-

[3] See Appendix I for additional graphs.

order differencing for our model. We also compute and plot the PACF for this time series to give us a sense of the type of model we might expect to construct in order to explain the behavior of this process (Figure 7).

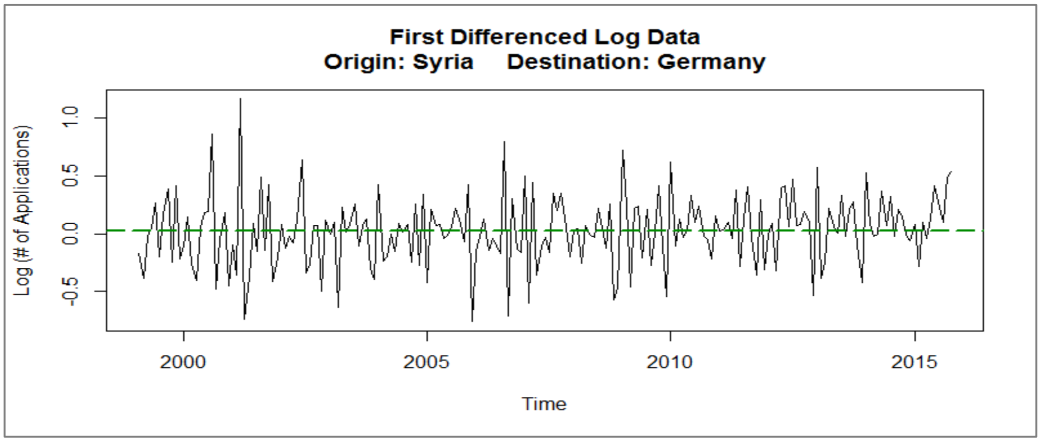

Figure 5

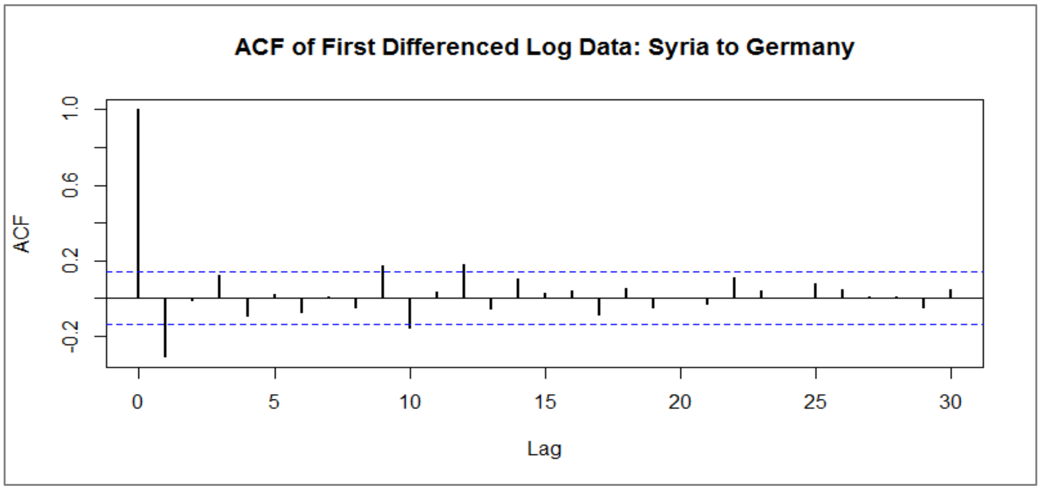

Figure 6

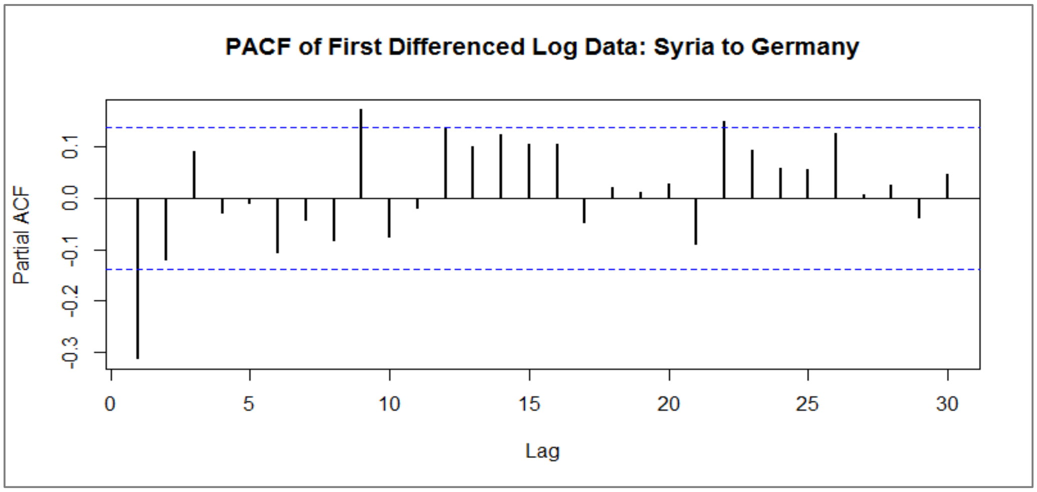

Figure 7

The plots of the ACF and PACF both show evidence of autocorrelation, especially in the first lag. This indicates we may expect both autoregressive and moving average behavior in our model. The ACF also demonstrates a slight recurring pattern at 12 lag intervals which may indicate the presence of seasonality.

Computing basic statistics on the first differenced logged time series reveals a skewness of 0.158 and excess kurtosis of 0.744. The Jarque-Bera normality test results in a p-value of 0.051. These findings indicate the time series is fairly normal with slightly heavy tails, as borne out by the histogram and normal Q-Q plots (Figures 8 and 9, respectively). With normality and stationarity conditions satisfied, we are ready to construct our models.

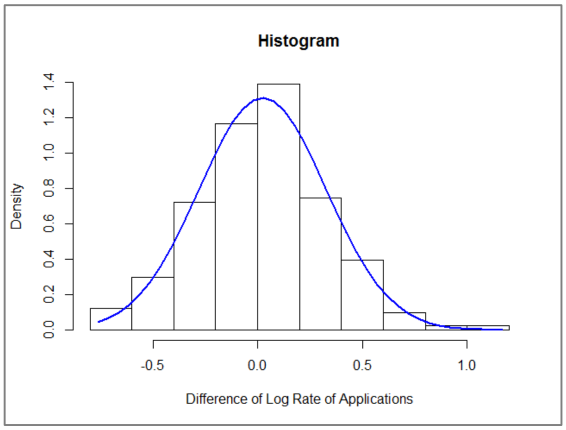

Figure 8

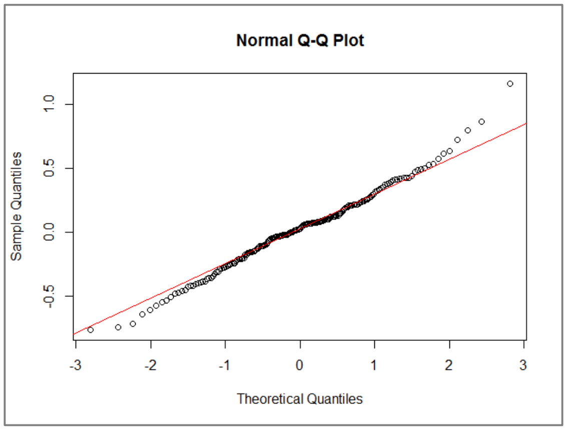

Figure 9

MODEL FITTING

The auto ARIMA function in R recommends an ARIMA(2,2,2)(0,0,1)[12] model – that is, a second differenced model with ARMA(2,2) behavior and 12-month seasonality with MA(1) behavior. The BIC for this model is 91.63. Initial testing of the model's coefficients indicates the AR terms are not statistically significant at the 5% level.

Removing the AR terms results in a new ARIMA(0,2,2)(0,0,1)[12] with BIC of 81.03 where all coefficients are significant at the 1% level[4]. Additionally, the lower BIC tells us Model 2 is likely a better model for our series.

Model 2

$$Y_t = log(X_t)$$

$$(1 - B)^2 Y_t = (1 - 1.383B + 0.401B^2)(1 + 0.193B^{12})a_t$$

Since the software recommended a second differenced model, we ran a check on the summary statistics, stationarity and normality assumptions for the second differenced log time series. The skewness for the time series is -0.291, excess kurtosis is 1.344 and the p-value for the Jarque-Bera test is less than 0.1%. Additionally, the ACF plot exhibits stationary behavior, indicating we can proceed with the second differenced model[5].

We would also like to examine what happens if we restrict ourselves to a first differenced model. The resulting model (Model 3) is an ARIMA(0,1,1)(0,0,1)[12] with a BIC of 76.34. All coefficients are significant at the 1% level.

Model 3

$$Y_t = log(X_t)$$

[4] Hyndman, R.J. "Forecasting: Principles and Practice." Backshift Notation. OTexts. 2013. https://www.otexts.org/fpp/8/2

[5] See Appendix I for additional plots

$$(1-B)Y_t = (1-0.350B)(1+0.220B^{12})a_t$$

RESIDUAL ANALYSIS AND MODEL DIAGNOSTICS

Proceeding with analysis of Model 2, we begin residual testing and model diagnostics. The ACF plot of Model 2's residuals (Figure 10) shows autocorrelations are zero with a possible borderline case at lag 10.

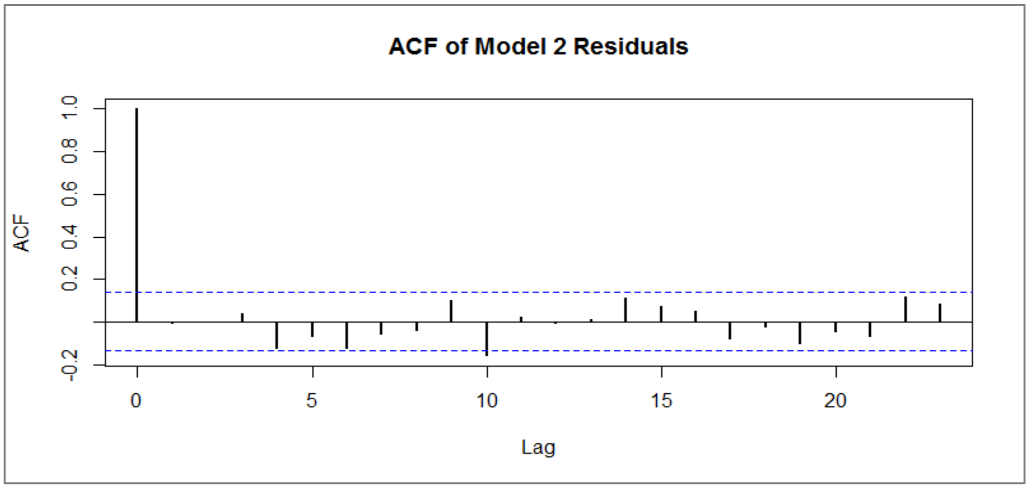

Figure 10

Computing the Ljung Box tests for this model's residuals at 5, 7, 9, 12 and 15 lags with 3 degrees of freedom (due to the 3 calculated parameters) results in p-values of 10%, 7%, 9%, 6% and 6%, respectively. While this fails the test of significance at the 5% level, the p-values are not large enough such that we can comfortably conclude the residuals are white noise. However, this lack of certainty may stem from having somewhat of a small data sample. Examining the histogram (Figure 11) and Q-Q plot (Figure 12) of the residuals, we see that the residuals appear normal with heavy tails.

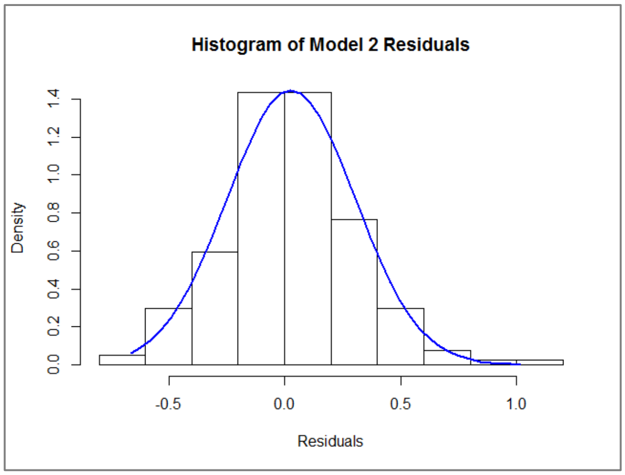

Figure 11

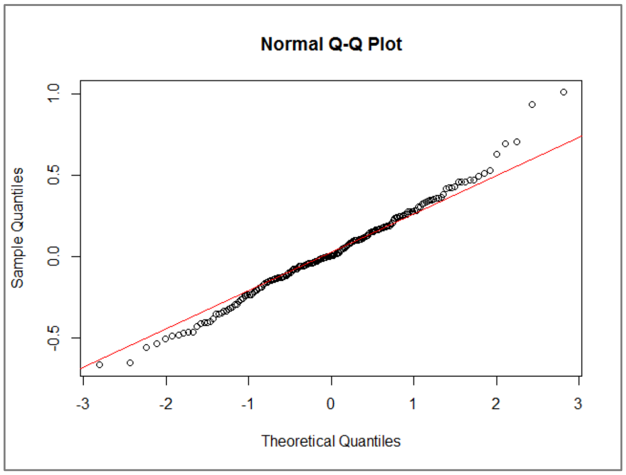

Figure 12

Residual analysis on our first differenced model (Model 3) shows slightly better performance on the residuals, as evidenced by the ACF plot (Figure 13).

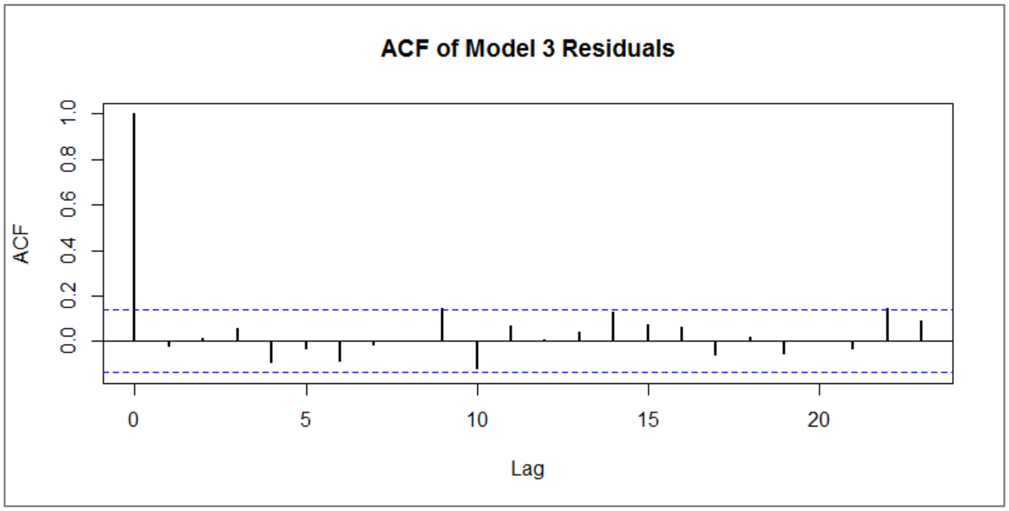

Figure 13

These findings are confirmed by performing Ljung-Box tests on Model 3's residuals at lags 5, 7, 9, 12 and 15, with 2 degrees of freedom (due to the 2 calculated parameters). Resulting p-values of these tests are 41%, 47%, 26%, 22% and 15%, respectively – all of which are large enough that we can conclude the residuals are white noise. Additionally, the histogram (Figure 14) and normal Q-Q plot (Figure 15) of the residuals support the normality assumption.

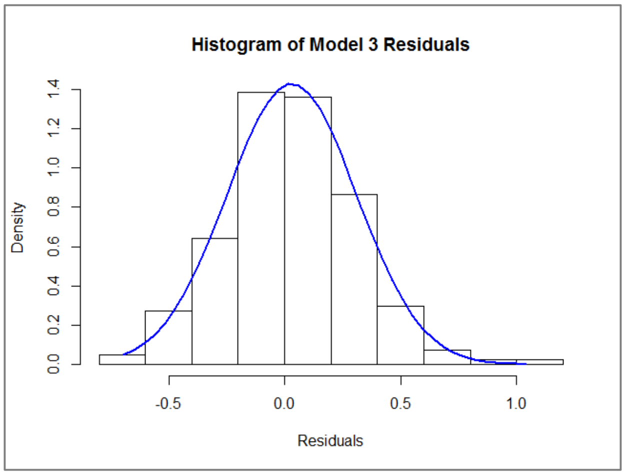

Figure 14

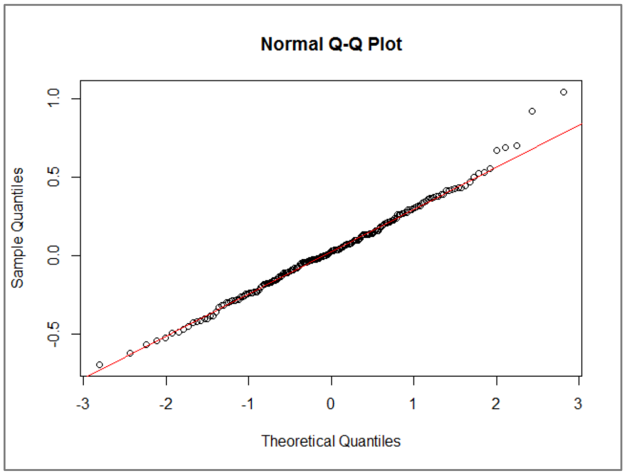

Figure 15

FORECAST ANALYSIS

Forecasting on Model 2 for the next five months results in the following values, along with the given 95% confidence intervals (Table 1).

Date	Forecast	Low 95%	High 95%
OCT 2015	**28,214**		
NOV 2015	**23,815**	13,809	41,072
DEC 2015	**25,820**	13,607	48,994
JAN 2016	**27,970**	13,503	57,936
FEB 2016	**28,510**	12,675	64,128
MAR 2016	**31,412**	12,907	76,449

Table 1

Forecasts appear consistent with the overall dynamic behavior of the system, as further demonstrated by the time plot of the data (Figure 16). The graph shows the last 50 observations of the series. The dashed blue line shows the fitted model, with forecasts in solid blue and confidence intervals in grey.

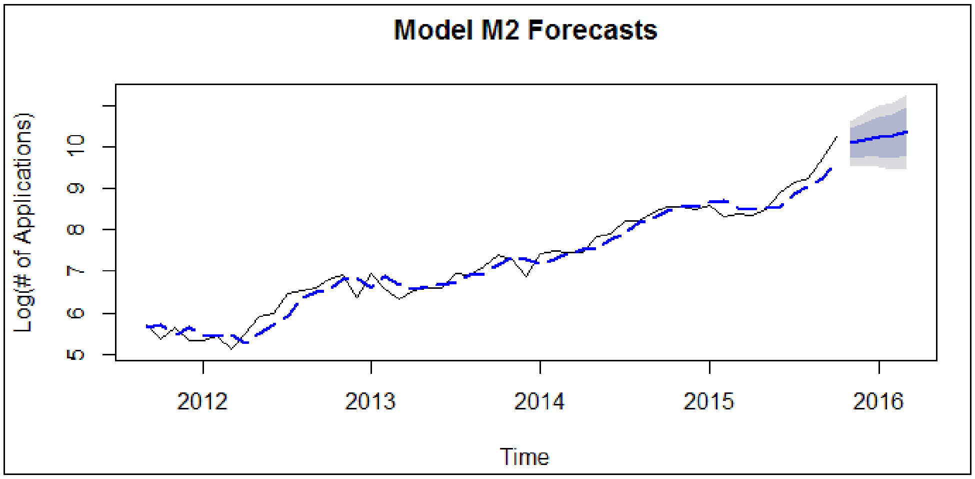

Figure 16

By comparison, Model 3 (Figure 17) also shows a close fit to the known observations (as confirmed by the model diagnostics performed earlier). However, forecasts for the next five months are not consistent with the overall dynamic behavior of the process. This helps confirm the inclusion of second order differencing in our model.

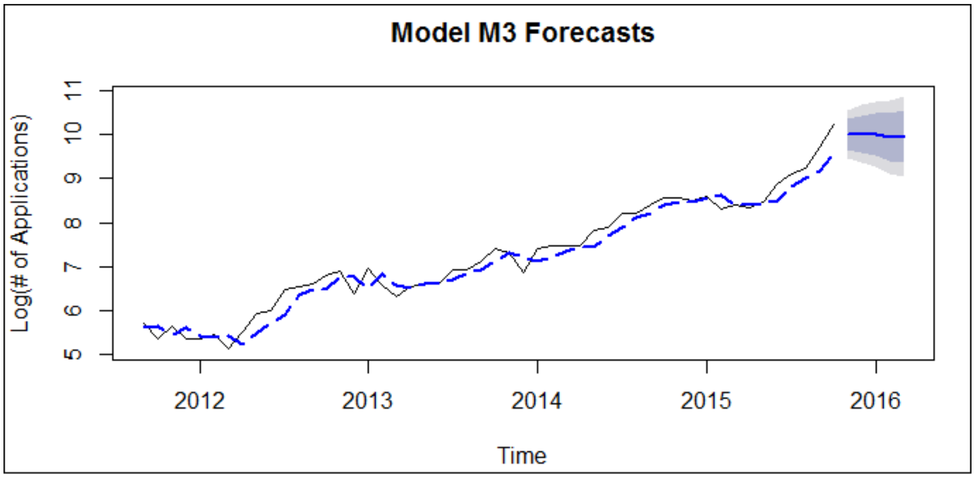

Figure 17

As a final diagnostic, we performed backtesting on Model 2 and Model 3 using 182 observations for training (90% of the dataset). This resulted in a mean absolute percentage error (MAPE) of 1.98% for Model 2, and 2.25% for Model 3. These findings indicate both models are a relatively good fit for forecasting, with Model 2 performing slightly better.

RESULTS

Through our analysis, we determined that the monthly number of asylum applicants from Syria to Germany can be modeled through time series analysis. The behavior of these data show an exponential growth consistent with the reports we see in the nightly news.

The exponential form of the time series required us to begin by taking the log of the series. We then analyzed both a second-differenced model (Model 2) and first-differenced model (Model 3) on the logged time series. While Model 3 had a slightly lower BIC value compared to that of Model 2, its forecasts were not consistent with the overall dynamic behavior of the process. Model 2 performed better in terms of out-of-sample forecasting (MAPE) and the model's forecasts appear to be consistent with the behavior of the time process. As we are most concerned with forecasting future values, we chose to emphasize MAPE and forecast behavior, rather than BIC. Therefore, we determine Model 2 (ARIMA(0,2,2)(0,0,1)[12]) as being our best fit model.

$$Y_t = log(X_t)$$
$$(1-B)^2 Y_t = (1-1.383B+0.401B^2)(1+0.193B^{12})a_t$$

As this model only exhibits MA behavior, it indicates that the time series has a short memory. The current state of the system is affected by the shocks in the system for the previous two months and the shocks in the system 12 months prior due to the seasonal component. This model requires second order differencing to account for the non-stationary curved trend of the logged data series.

Forecasting using this model indicates a slight decline in the number of asylum applicants from October to November, followed by an upward trend for December through March. This slight decline is in keeping with the seasonal behavior exhibited in previous years where the number of applicants decreases in the winter followed by an increase throughout the rest of the year.

The accuracy of our model may be impacted by the small dataset size compared to the relative complexity of our model. Additional observations may increase the accuracy of the model or suggest additional parameters for inclusion. It is also important to note that since our model only includes applicants to a single country, it is highly susceptible to changes in legislation from Germany or other EU member states regarding the acceptance of asylum applications. Significant changes in legislation would invalidate the stationarity condition necessary for accurate modeling and forecasting.

APPENDIX I. Additional Graphs

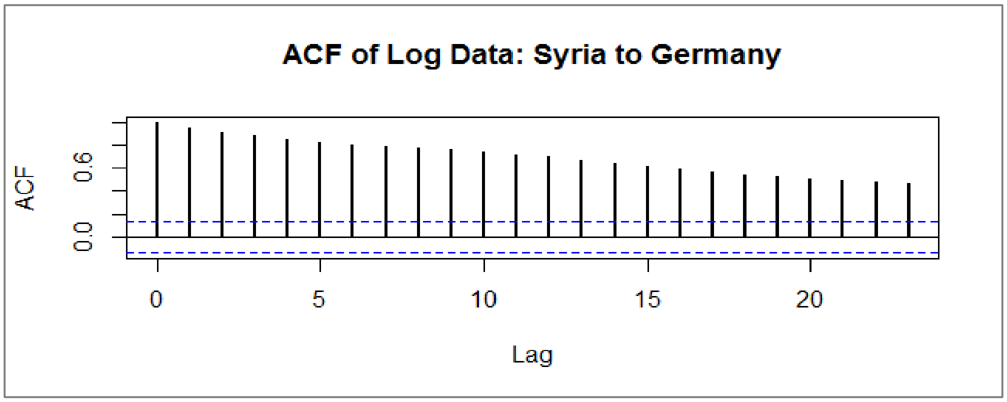

Figure 18

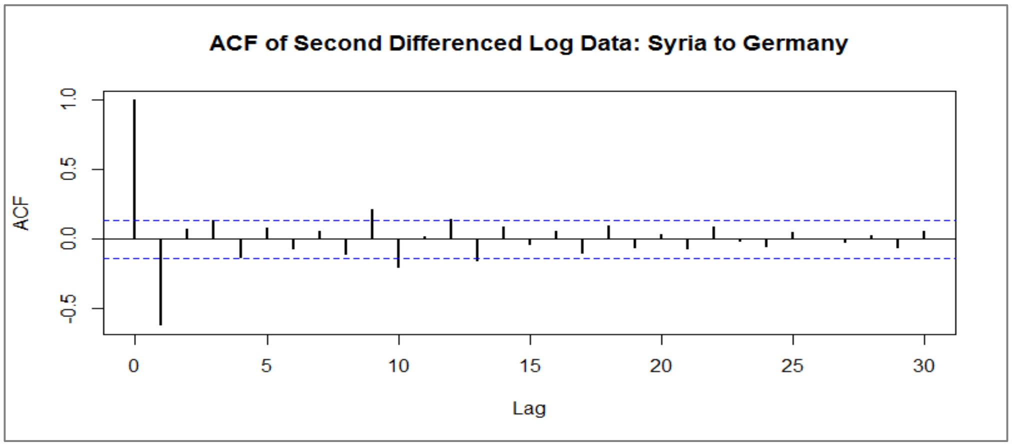

Figure 19

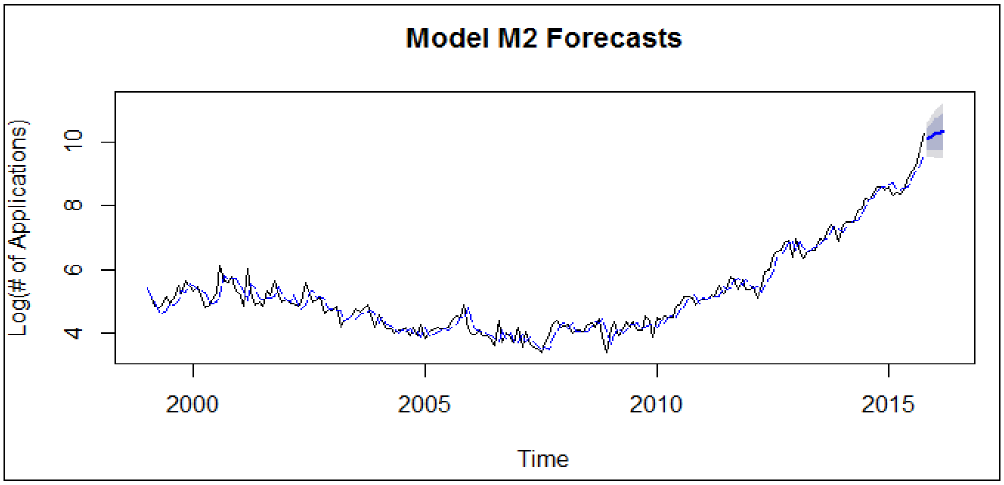

Figure 10

APPENDIX II. R Code

```
########## Asylum Applicants #########

##### Libraries ####

#load libraries
library(tseries)
library(fBasics)
library(forecast)
library(lmtest)
library(fUnitRoots)

##### Data Load and Exploratory Analysis for Syria, Afghanistan, Iraq to EU ####
data=read.table("SyrAfgIrq2EU.csv",header=T, sep=',')
head(data)
rate = data[,1]
head(rate)
ratets = ts(rate, start=c(1999,1), freq=12)
ratets

#time plot of original data
plot(ratets, xlab="Time", ylab="# of Applications", lwd=2,
     main="Asylum Applications Jan 1999 to Oct 2015\nOrigin: Syria/Afg/Iraq
Destination: EU")

#take log
lratets=log(ratets)
plot(lratets, xlab="Time", ylab="Log (# of Applications)", lwd=2,
     main="Asylum Applications Jan 1999 to Oct 2015\nOrigin: Syria/Afg/Iraq
Destination: EU")


##### Data Load and Exploratory Analysis for Syria to Germany ####

myd <- read.table("Syr2Ger.csv",header=T, sep=',')
head(myd)
rate <- myd$Value
head(rate)
ratets <- ts(rate, start=c(1999,1), freq=12)
ratets

#time plot of original data
plot(ratets, type="l", xlab="Time", ylab="# of Applications", lwd=2,
     main="Asylum Applications Jan 1999 to Oct 2015\nOrigin: Syria     Destination:
Germany")

#take the log and plot time series and ACF
lratets = log(ratets)

plot(lratets, xlab="Time", ylab="Log (# of Applications)", lwd=2,
     main="Asylum Applications Jan 1999 to Oct 2015\nOrigin: Syria     Destination:
Germany")
```

```
### Proceed with Syria to Germany data for remainder of code

a <- acf(coredata(lratets), plot = F)
plot(a, main="ACF of Log Data: Syria to Germany", lwd=2)

#dickey fuller test
adfTest(lratets, lags=3, type=c("ct"))
adfTest(lratets, lags=5, type=c("ct"))
adfTest(lratets, lags=7, type=c("ct"))
#all p-values were close to 1. time process is unit-root non-stationary


#let's apply the first difference and then re-run the plotting and
#DF tests to see if we can now reject the H0.
dlratets=diff(lratets)

plot(dlratets, type="l", xlab="Time", ylab="Log (# of Applications)",
     main="First Differenced Log Data\nOrigin: Syria     Destination: Germany")
abline(h=mean(dlratets), col="green4", lty=5, lwd=2)

b <- acf(as.vector(dlratets),lag.max=30, plot=F)
plot(b, main="ACF of First Differenced Log Data: Syria to Germany", lwd=2)

c <- pacf(as.vector(dlratets),lag.max=30, plot=F)
plot(c, main="PACF of First Differenced Log Data: Syria to Germany", lwd=2)

adfTest(dlratets, lags=3, type=c("c"))
adfTest(dlratets, lags=5, type=c("c"))
adfTest(dlratets, lags=7, type=c("c"))
#can reject H0 now.


#basic statistics and normality testing for differenced log series
basicStats(dlratets)

hist(dlratets, xlab="Difference of Log Rate of Applications", prob=TRUE,
main="Histogram")
xfit<-seq(min(dlratets),max(dlratets),length=40)
yfit<-dnorm(xfit,mean=mean(dlratets),sd=sd(dlratets))
lines(xfit, yfit, col="blue", lwd=2)

qqnorm(dlratets)
qqline(dlratets, col = 2)

normalTest(dlratets,method=c("jb"))


##### Model Creation ####

#model m1
m1=auto.arima(lratets, trace=T, ic="bic", allowdrift = T, allowmean = T)
coeftest(m1)
m1
#we get an ARIMA(2,2,2)(0,0,1)[12] model
#both AR coefficients are not significant.
```

```
#try removing both non-significant AR coefficients
#try fitting ARIMA(0,2,2)(0,0,1)[12] model
#model m2
m2=Arima(x=lratets, order=c(0,2,2), seasonal=c(0,0,1), method="ML")
coeftest(m2)
m2
#all coefficients are now significant
#bic value (81.03) was lower than that of model m1 (91.63)

### SIDEBAR ###
#since the function is suggesting a 2nd difference, let's go back and
#check the ACF and plot for the second difference

#take the second difference of the logged rate and plot the ACF values
d2lratets <- diff(dlratets)

plot(d2lratets, type="l", xlab="Time", ylab="Log(# Applications)",
     main="Second Differenced Log Data\nOrigin: Syria     Destination: Germany")
abline(h=mean(d2lratets), col="green4", lty=5, lwd=2)

b2 <- acf(as.vector(d2lratets),lag.max=30, plot=F)
plot(b2, main="ACF of Second Differenced Log Data: Syria to Germany", lwd=2)
#looks stationary

#basic statistics and normality testing for second differenced log series
basicStats(d2lratets)

hist(d2lratets, xlab="Second Difference of Log Rate of Applications", prob=TRUE,
main="Histogram")
xfit<-seq(min(d2lratets),max(d2lratets),length=40)
yfit<-dnorm(xfit,mean=mean(dlratets),sd=sd(d2lratets))
lines(xfit, yfit, col="blue", lwd=2)

qqnorm(d2lratets)
qqline(d2lratets, col = 2)

normalTest(d2lratets,method=c("jb"))

### END SIDEBAR ###


#it was interesting that auto.arima decided to take second difference,
#and not just the first. what if we restrict it to just the first difference?
#(dickey fuller test on first differenced data rejected H0)

#model m3
m3=auto.arima(lratets, d=1, trace=T, ic="bic", allowdrift = T, allowmean = T)
coeftest(m3)
m3
#we get an ARIMA(0,1,1)(0,0,1)[12] model
#both parameters are significant


##### Residual Analysis and Forecasting ####

### Model m1 ###
#diagnostics
```

```
#analyze the residuals
acf(as.vector(m1$resid))
#acf plot shows residual correlations are zero with borderline case at lag 10

Box.test(m1$residuals, 7, "Ljung-Box", fitdf=5)
Box.test(m1$residuals, 9, "Ljung-Box", fitdf=5)
Box.test(m1$residuals, 12, "Ljung-Box", fitdf=5)
Box.test(m1$residuals, 15, "Ljung-Box", fitdf=5)
#p-values for LJB test do not confirm white noise series for all tested lags
#perhaps due to large df and small sample size

hist(m1$residuals)
qqnorm(m1$residuals)
qqline(m1$residuals, col=2)
#residuals appear to be fairly normal

#Forecast for Model m1
f1=forecast.Arima(m1, h=5)
f1exp=exp(f1$mean)
ratets
f1exp
plot(f1)
plot(f1, include=100)
plot(f1, include=50)
lines(ts(tail(f1$fitted,50), frequency=12,start=c(2011,9)), lty=5, lwd=2, 
col="blue")
#forecasts appear to be consistent with the overall dynamic behavior of the process

#backtesting
source("backtest.R")
#there's 202 values in the dataset, so we'll use 182 for training (~90%)
backtest(m1, lratets, h=1, orig=182)
#MAPE is 1.8%. Pretty good




### Model m2 ###
#diagnostics

#analyze the residuals
pResid <- acf(as.vector(m2$resid), plot =F)
plot(pResid, main="ACF of Model 2 Residuals", lwd=2)
#acf plot shows residual correlations are zero with borderline case at lag 10

Box.test(m2$residuals, 5, "Ljung-Box",fitdf=3)
Box.test(m2$residuals, 7, "Ljung-Box", fitdf=3)
Box.test(m2$residuals, 9, "Ljung-Box", fitdf=3)
Box.test(m2$residuals, 12, "Ljung-Box", fitdf=3)
Box.test(m2$residuals, 15, "Ljung-Box", fitdf=3)
#p-values indicate H0 can't be rejected at 5% significance level

hist(m2$residuals, xlab = "Residuals", prob=T, main="Histogram of Model 2 
Residuals")
xfit <-seq(min(m2$residuals), max(m2$residuals), length=40)
yfit <- dnorm(xfit, mean = mean(m2$residuals), sd=sd(m2$residuals))
```

```
lines(xfit, yfit, col="blue", lwd=2)

qqnorm(m2$residuals)
qqline(m2$residuals, col=2)
#residuals appear to be fairly normal


#Forecast for Model m2
f2=forecast.Arima(m2, h=5)
f2exp=exp(f2$mean)
ratets
f2exp

plot(f2, ylab="Log(# of Applications)", xlab="Time", main="Model M2 Forecasts")
lines(ts(f2$fitted, frequency=12,start=c(1999,1)), lty=5, col="blue")

plot(tail(ratets,50), type="l")
plot(f2, include=50, ylab="Log(# of Applications)", xlab="Time", main="Model M2
Forecasts")
lines(ts(tail(f2$fitted,50), frequency=12,start=c(2011,9)), lty=5, lwd=2,
col="blue")
#forecasts appear to be consistent with the overall dynamic behavior of the process


#backtesting
source("backtest.R")
#there's 202 values in the dataset, so we'll use 182 for training (~90%)
backtest(m2, lratets, h=1, orig=182)
#MAPE is 1.9%. Pretty good



### Model m3 ###
#diagnostics
#analyze the residuals
pResid3 <- acf(as.vector(m3$resid), plot =F)
plot(pResid3, main="ACF of Model 3 Residuals", lwd=2)
#acf plot shows residuals are white noise

Box.test(m3$residuals, 5, "Ljung-Box",fitdf=2)
Box.test(m3$residuals, 7, "Ljung-Box", fitdf=2)
Box.test(m3$residuals, 9, "Ljung-Box", fitdf=2)
Box.test(m3$residuals, 12, "Ljung-Box",fitdf=2)
Box.test(m3$residuals, 15, "Ljung-Box",fitdf=2)
#acf of the residuals exhibit white noise

hist(m3$residuals, xlab = "Residuals", prob=T, main="Histogram of Model 3
Residuals")
xfit <-seq(min(m3$residuals), max(m3$residuals), length=40)
yfit <- dnorm(xfit, mean = mean(m3$residuals), sd=sd(m3$residuals))
lines(xfit, yfit, col="blue", lwd=2)

qqnorm(m3$residuals)
qqline(m3$residuals, col=2)
#residuals appear to be fairly normal

#Forecast for Model m3
```

```
f3=forecast.Arima(m3, h=5)
f3exp=exp(f3$mean)
ratets
f3exp

plot(f3, ylab="Log(# of Applications)", xlab="Time", main="Model M3 Forecasts")
lines(ts(f3$fitted, frequency=12,start=c(1999,1)), lty=5, col="blue")

plot(f3, include=50, ylab="Log(# of Applications)", xlab="Time", main="Model M3
Forecasts")
lines(ts(tail(f3$fitted,50), frequency=12,start=c(2011,9)), lty=5, lwd=2,
col="blue")
#forecasts do NOT appear to be consistent with the overall dynamic behavior of the
process



#backtesting
source("backtest.R")
#there's 202 values in the dataset, so we'll use 182 for training (~90%)
backtest(m3, lratets, h=1, orig=182)
#MAPE is 2.2%, still pretty good, but forecasts are not consistent with time
process
```