Time Series Analysis of Proportion of Issued License Plates in Shanghai

Carrie Yan (9497124), Claire Hua (9952425), Ivy Tran (9565979),
Michelle Martin (9658311), Fady Naeim (9737032)

December 5, 2018

Abstract

The purpose of this project is to work with data based off of the monthly Shanghai auction system to sell a limited number of license plates to fossil-fuel car buyers. The data has been constantly collected every month since January of 2002 and continues to be updated to this day. Throughout the project, we use various forms of time series techniques and methods to analyze the features of the data. These methods include ACF, PACF, log transformation, square root transformation, box-cox transformation, differencing, AIC for model comparison, and back transformation. We also use the information to help us forecast the predictions of the license plate proportions up until the year 2020. After making the time series forecast and analysis of the data set, we come to the conclusion that the monthly Shanghai proportion of license plates for fossil-fuel car buyers will remain relatively consistent.

Introduction

For the data we are analyzing, we concentrate on the prediction of monthly auction sales of license plates in Shanghai for fossil-fuel car buyers. Our data begins in January 2002 and is continuously updated each month. We forecast the monthly proportion of licenses issued to the number of applicants up until the year 2020 to determine whether the proportion of license plates issued to number of applicants will increase or decrease as time goes on. The license plate in Shanghai is referred to as "the most expensive piece of metal in the world" and the average price is about $13,000. Due to Shanghai's increasing air pollution problem, this was the government's solution to attempt to combat the problem.

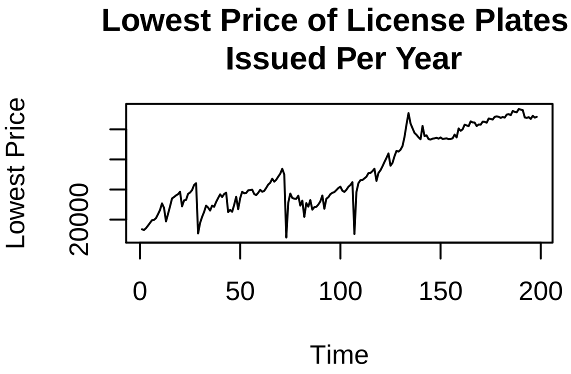
Our data contains the following variables:

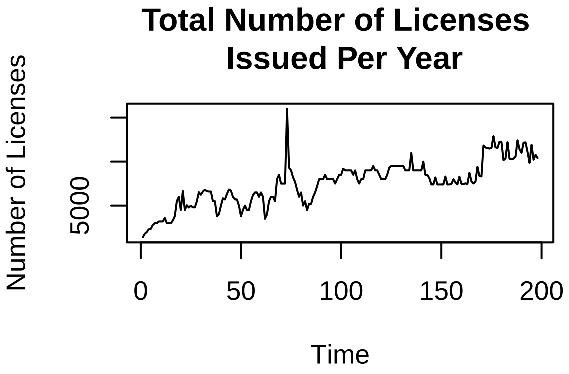
Total Number of Licenses Issued = Number of license plates issued per year
Lowest Price = Price of the lowest auctioned license plate per year
Average Price = Average price of a license plate per year
Total Number of Applicants = Number of people applying for license plates issues per year
Date = Monthly dates starting at January 2002 when the license plates are issued

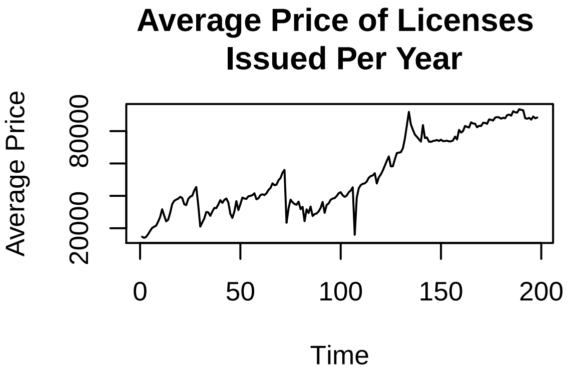
We use time series techniques to predict the coming monthly proportion as well as back-transformation to predict information that has already past.

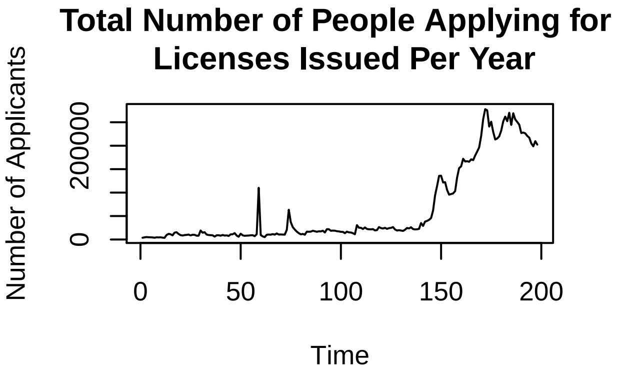
Initial Analysis

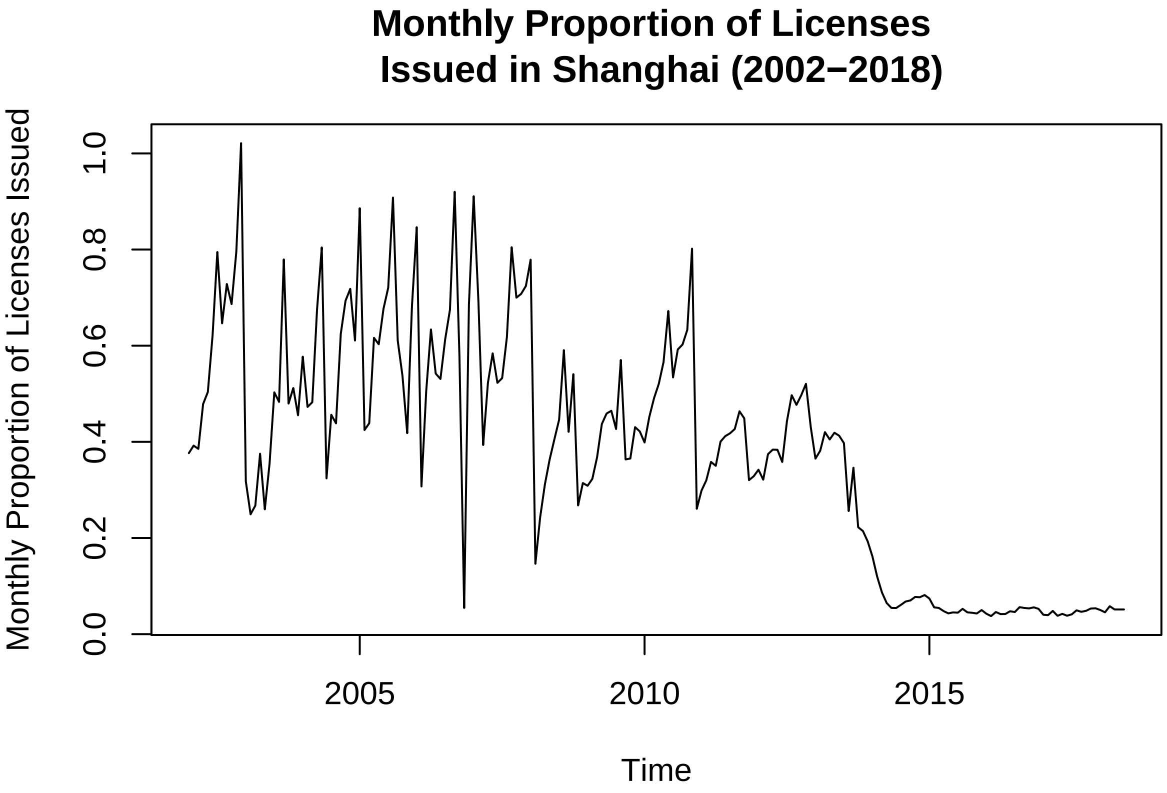
We first convert the data into a time series and plot each of the four variables: lowest price, total number of license plates issued, average price, and total number of applicants. For the plot of lowest price, we see that for about half the plot, the price seems to be slowly increasing with a little fluctuation. Then, there seems to be a sudden spike in which the lowest price increases significantly. For the plot of total number of license plates issued, we can see that the number issued is partially consistent with little increase as time goes on. There are however instances in which the number of license plates issued is dramatically changed, as we can see around 75, and the decrease from approximately 145 to 175. For the plot of average price, we can see that there is an upward trend and for the plot of total number of applicants, we can see that it is a low number up until approximately 150. At this point in time, the number of applicants begins to increase dramatically and then becomes constant at around 250,000, but then seems to begin to drop back down again.

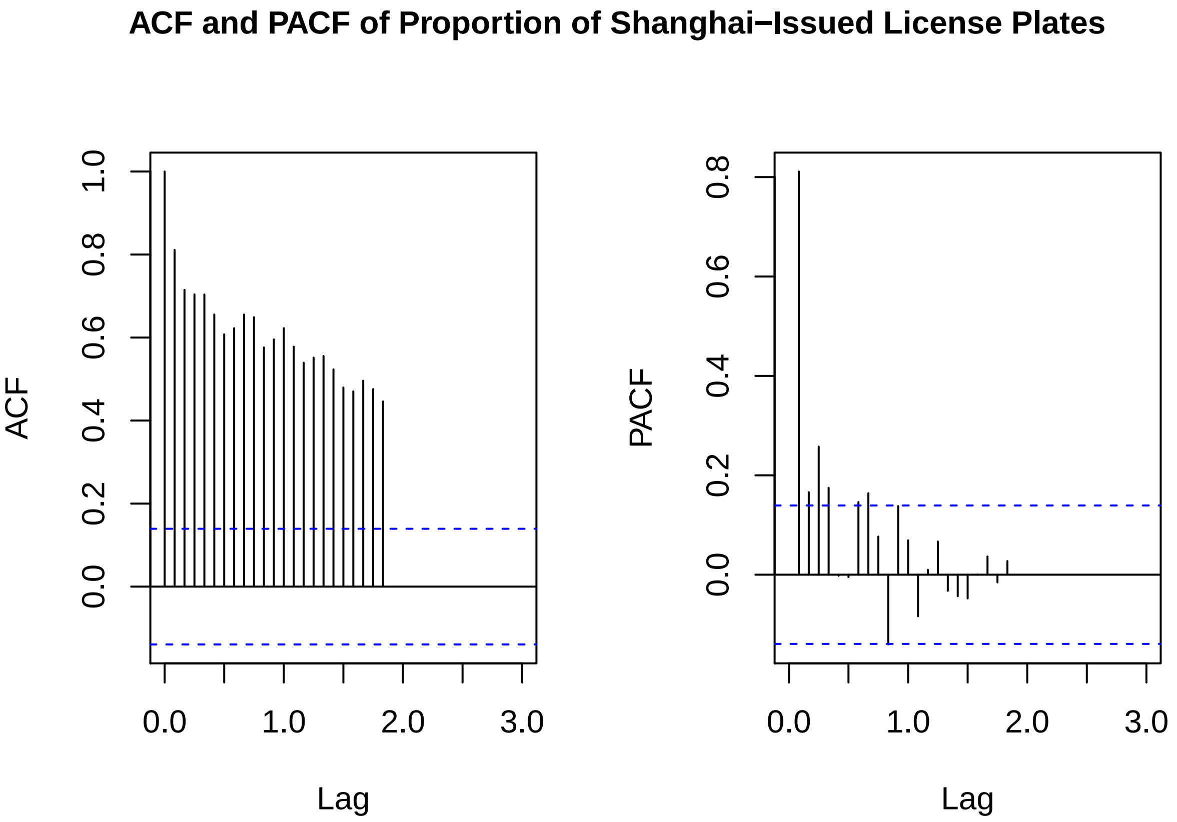

Lowest Price of License Plates Issued Per Year


Total Number of Licenses Issued Per Year


Average Price of Licenses Issued Per Year


Total Number of People Applying for Licenses Issued Per Year


Monthly Proportion of Licenses Issued in Shanghai (2002−2018)



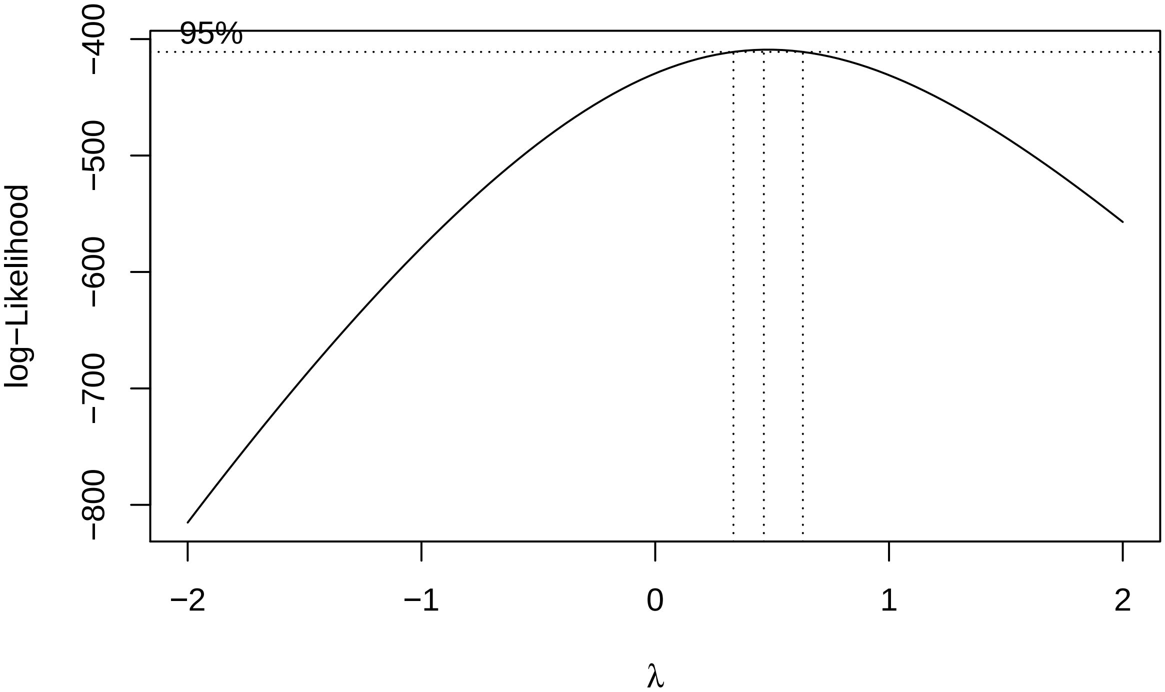
We continue by finding the mean and variance of the proportion of license plates issued to total number of applicants. We get values of 0.3756 for the mean and 0.0609 for the variance. Once we have plotted the time series of the proportion of monthly license plates issued, we see that it is not stationary. We then use ACF and PACF plots to attempt to hypothesize the type of series we are working with. The ACF seems to cut off before lag 2 while the PACF tails off starting near lag 0.7. So with this information, we can hypothesize that the original series is that of an AR model.

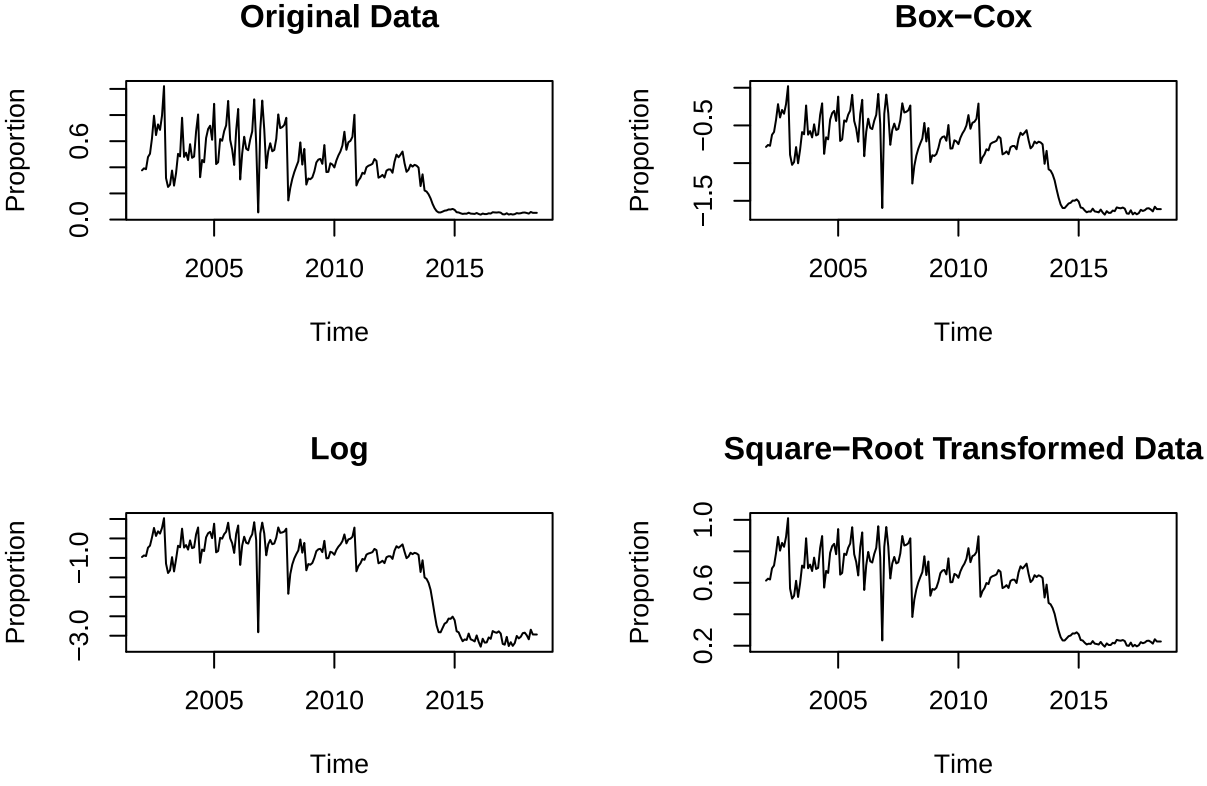
Note: In our ACF and PACF plots, our lags are in increments of years such that lag 1 = 12 months and lag 2 = 24 months.

Transformations

We first begin by testing to see which of the three forms of transformations works best in our situation. We are choosing among Box-cox, Log, and Square root transformations. We plot each of the tranformations and compare them to our original plot.

We applied the transformation because our initial time series was not stationary. Due to heteroscedasticity, our original time series violated our constant error of variance assumption. This is because our variance of error appeared to be changing over time.





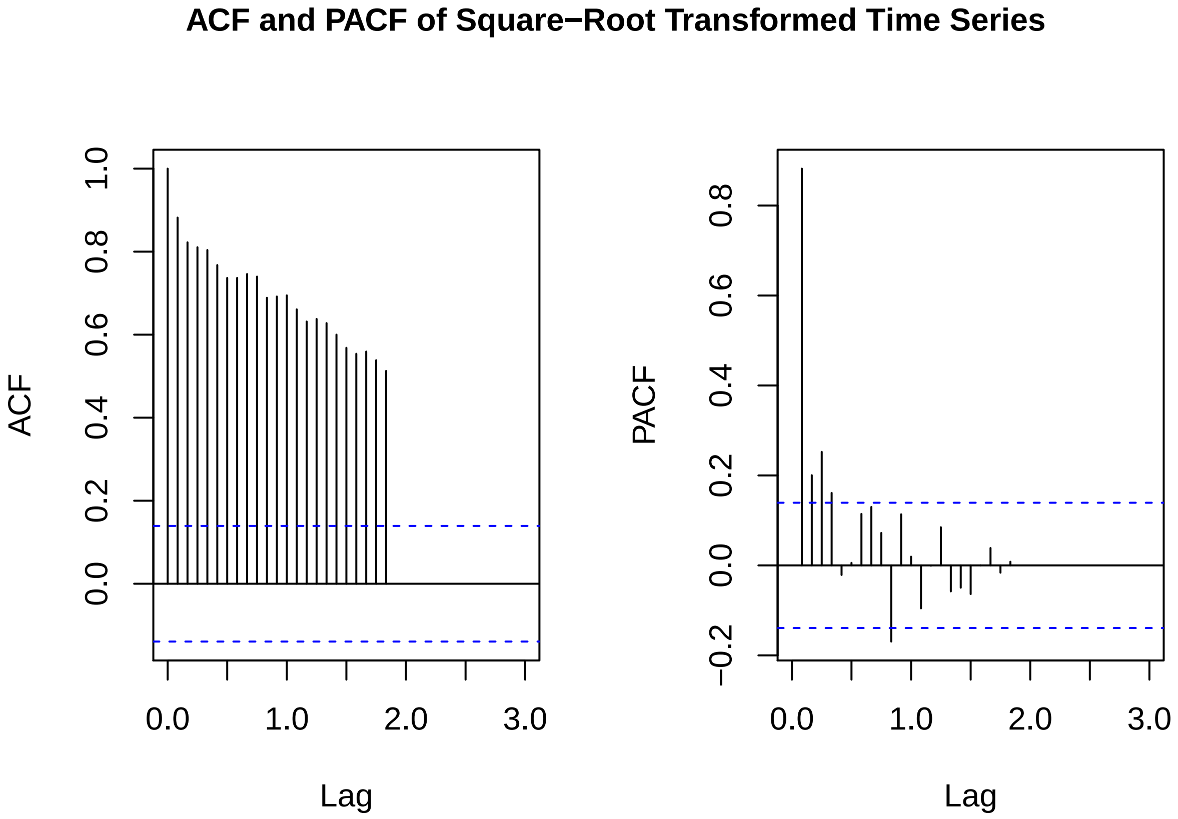
When looking at all the plots above, we quickly realize that the graphs are difficult to interpret, so we find the variances of each to determine which is the best fit for our model. Based off of the results, we can see that the square root transformation gives us the smallest variance value and therefore, we determine that this is the best transformation for our model. Also, the box-cox transformation tells us that lambda is 0.46 which is relatively close to 0.5, which tells us that the square-root transformation performs best.

	Variance

	Variance
Original Time Series	0.0609
Box-Cox Transformation	0.2393
Log Transformation	1.0672
Square Root Transformation	0.0540

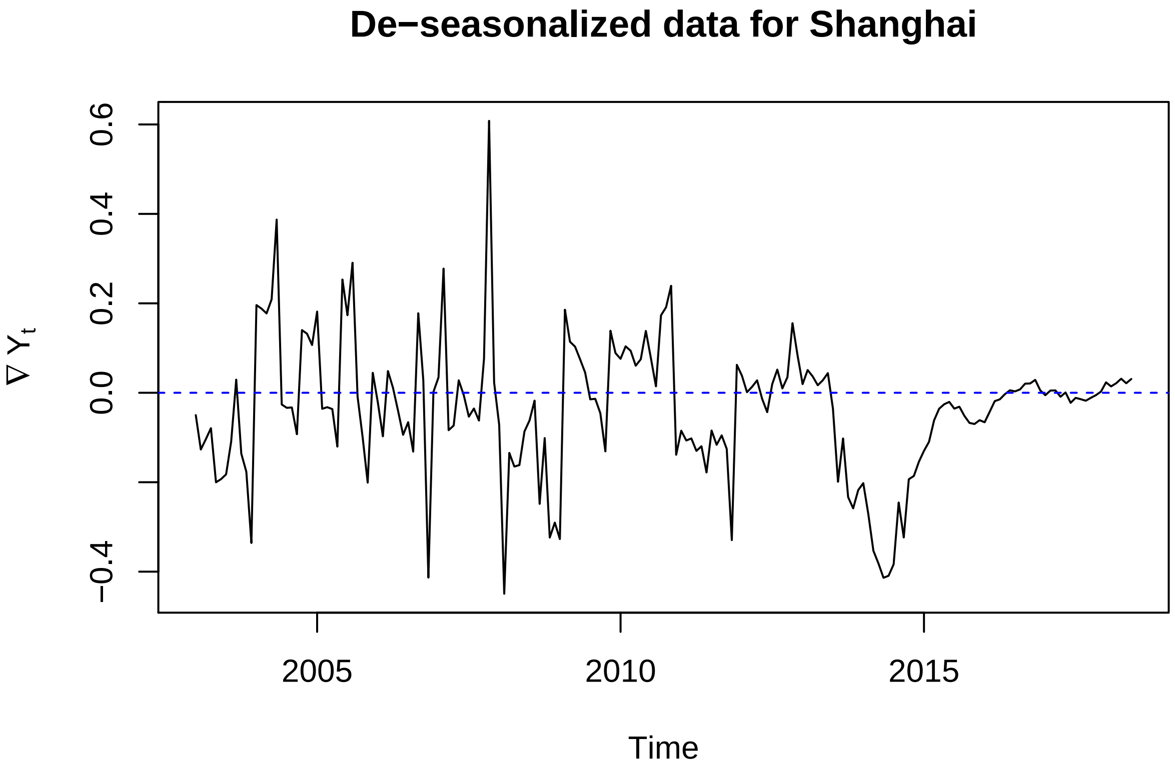
Square Root Transformation

Continuing with our chosen Square Root Transformation, we plot the ACF and PACF time series and see that the ACF is still tailing off while the PACF cuts off at around lag 0.8, which further supports our initial assumption that the series follows an AR(p) model.

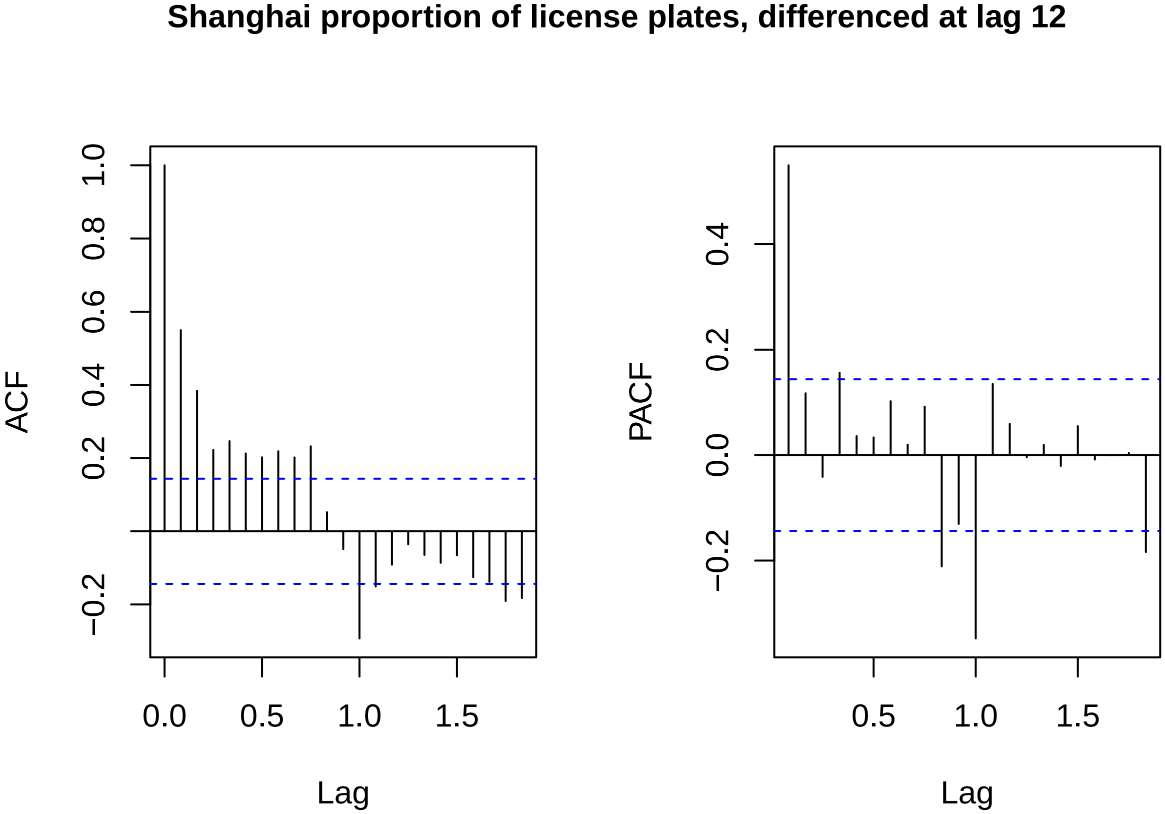


Differencing to Remove Seasonality and Trend

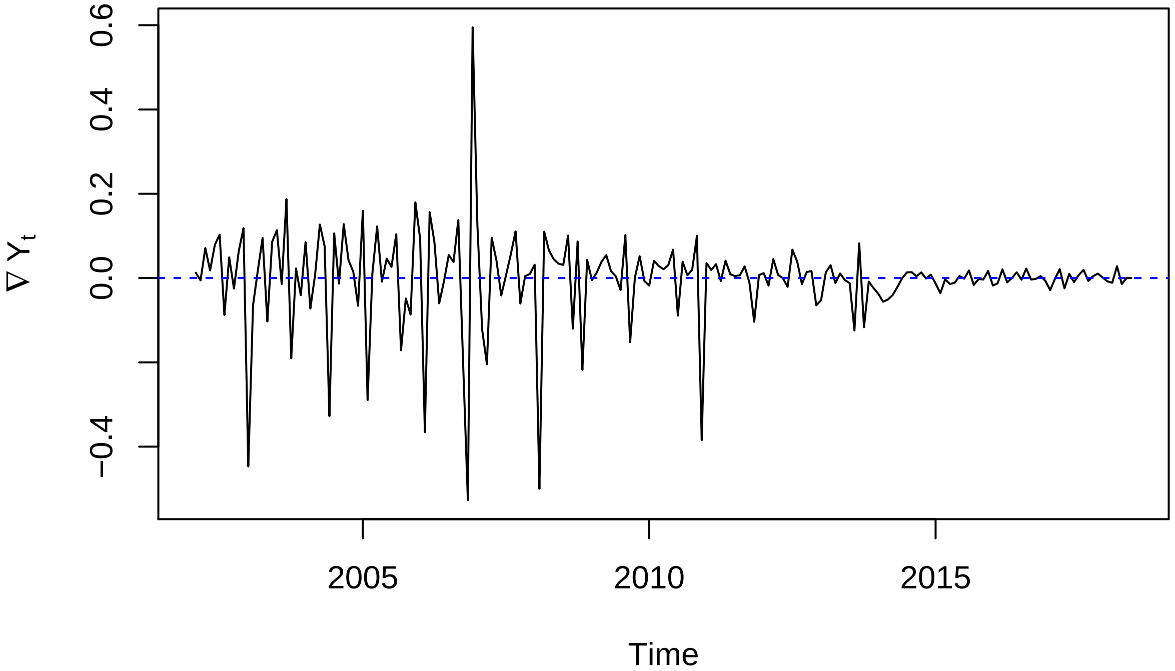
After applying the square root transformation, our data still does not look stationary. Therefore, we will apply differencing to remove trends and seasonality. We difference once at lag 12 to remove the seasonality component so that the de-seasonalized data fluctuates around the mean=0 line. For the ACF, we can see that it begins to slowly decay while the PACF oscillates between the bounds.



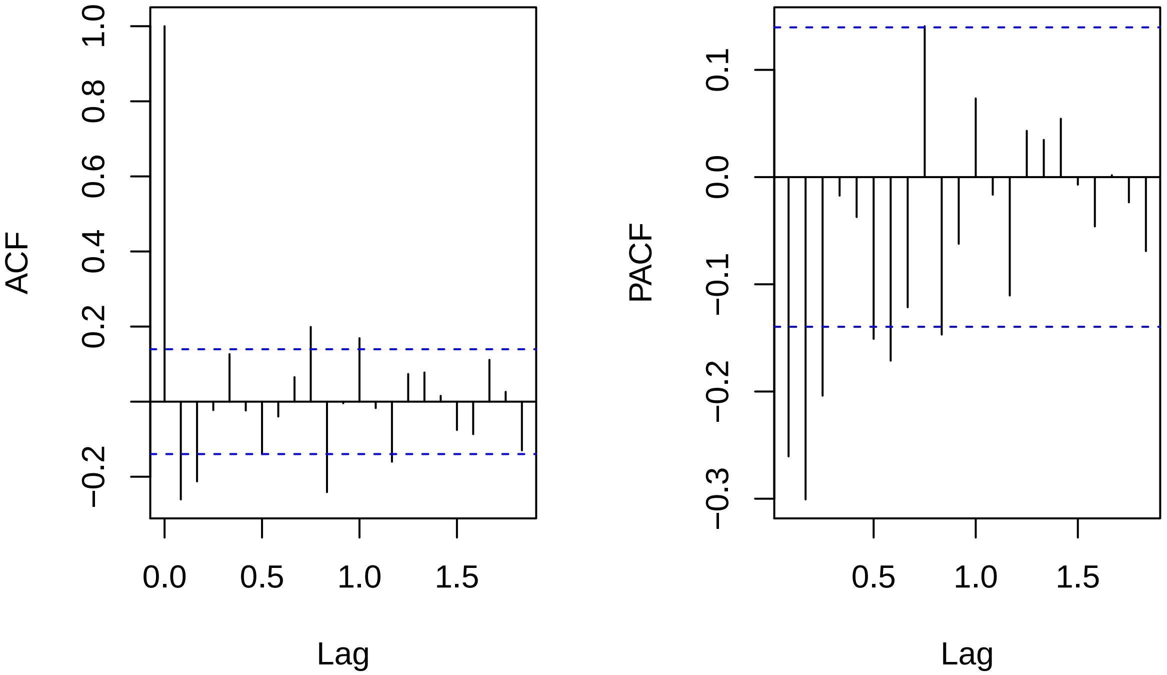
De−seasonalized data for Shanghai



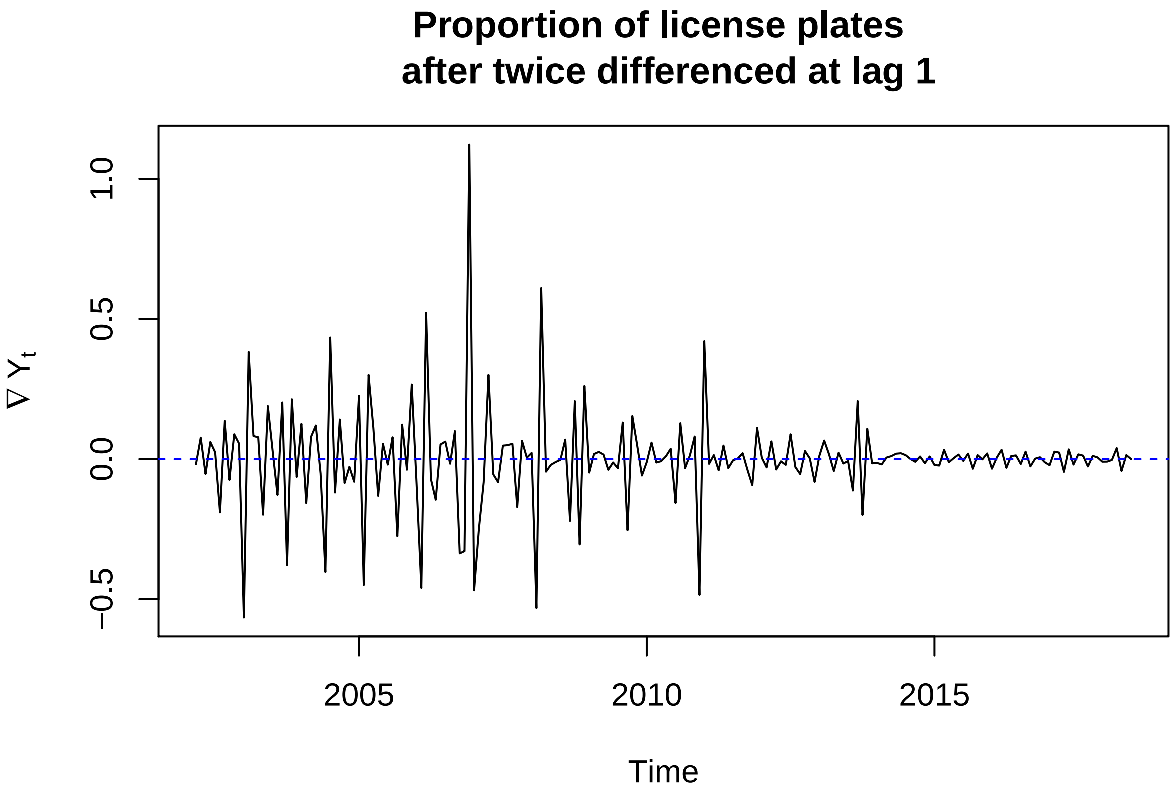
Shanghai proportion of license plates, differenced at lag 12



De−trended and De−seasonalized data for proportion of license plates



ACF and PACF of proportion of license plates differenced at lag 1



We difference again at lag 1 to remove the trend component of the data. This gives us a de-trended and de-seasonalized series to work with. The first time we difference at lag 1, we get a variance value of 0.01221 but when we difference a second time at lag 1, our variances increases to 0.03093 and so this tells us to only difference at lag 1 once. We can see that our de-trended and de-seasonalized data plot is now fluctuating very closely around the mean = 0 line which shows that it is stationary. Our ACF plot oscillates between the bounds while the PACF seems to cut off at lag 0.1.

Parameter Estimation using Yule-Walker

We perform preliminary estimation using Yule-Walker and it gives us an AR model of order 10, so this may be an AR(10) process.

```
## 
## Call:
## ar(x = shanghai_prop.diff1, method = "yule-walker")
## 
## Coefficients:
##       1        2        3        4        5        6        7        8  
## -0.4189  -0.4095  -0.2802  -0.1567  -0.2030  -0.2471  -0.2010  -0.1161  
##       9       10  
##  0.0761  -0.1469  
## 
## Order selected 10  sigma^2 estimated as  0.009353
```

Fitting an ARMA Process

Using the auto.arima() function, we find that the estimated model is a ARIMA (1,0,1) model and so we use the estimated orders of (p,q) to run further AIC tests and find the best model.

```
## Series: shanghai_prop.diff1
## ARIMA(1,0,1) with zero mean
##
## Coefficients:
##          ar1      ma1
##       0.3560  -0.8184
## s.e.  0.0962   0.0570
##
## sigma^2 estimated as 0.009847:  log likelihood=176.33
## AIC=-346.65   AICc=-346.53   BIC=-336.8
```

ARMA Models

Using a for-loop, we test each of the possible ARMA(p,q) parameter values to see which process gives us the smallest value of AIC. Looking at our results, we can see that ARMA(1,1) gives us the lowest AIC value of -217.3153.

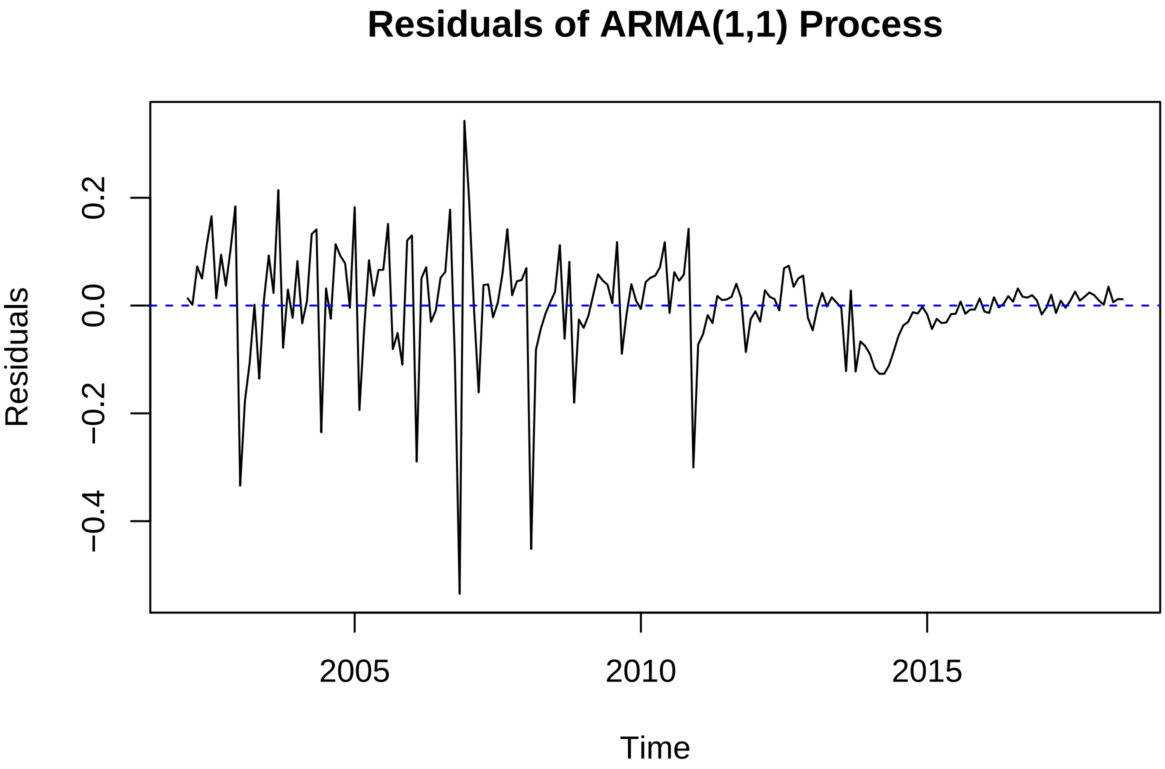
Checking for the best model fit

We do further testing to see if our model can be reduced more. We begin to check each of the three possible models: AR(1), MA(1), and ARMA(1,1) to see which is the best fit. We fit each of the three and then test each individual AIC to see which produces the lowest value. Our results shows us that ARMA(1,1) returns the lowest AIC of -346.0987 while MA(1) gives us -106.3236 and AR(1) gives us -201.9093. Therefore, we can conclude that an ARMA(1,1) model is best for our data.

	AIC
AR(1)	-201.9093
MA(1)	-106.3236
ARMA(1,1)	-346.0987

Plotting Residuals of ARMA(1,1)

After deciding that ARMA(1,1) is the best model, we then plot the residuals. We can see that the residuals seem to oscillate about the line at error 0.



Diagnostic Checking of Residuals

We perform diagnostic checking to check for the normality of errors, if the residuals are serially correlated, and if the residuals are not heteroskedastic and have constant variance.

The Shapiro-Wilk test gives us a p-value of 3.169e-12 which is less than our alpha of 0.05, so we conclude that the ARMA(1,1) does not pass the Shapiro Wilk test.

The Ljung-Box test for constant variance gives us a p-value of 0.7762 which is greater than our alpha of 0.05, so we can accept the assumption of normality and conclude that the residuals are random.

The Box-Pierce test gives us a p-value of p-value = 0.7778, which is very similar to that of the Ljung-Box test, and since that value is greater than out alpha of 0.05, we can conclude that the residuals are serially correlated.

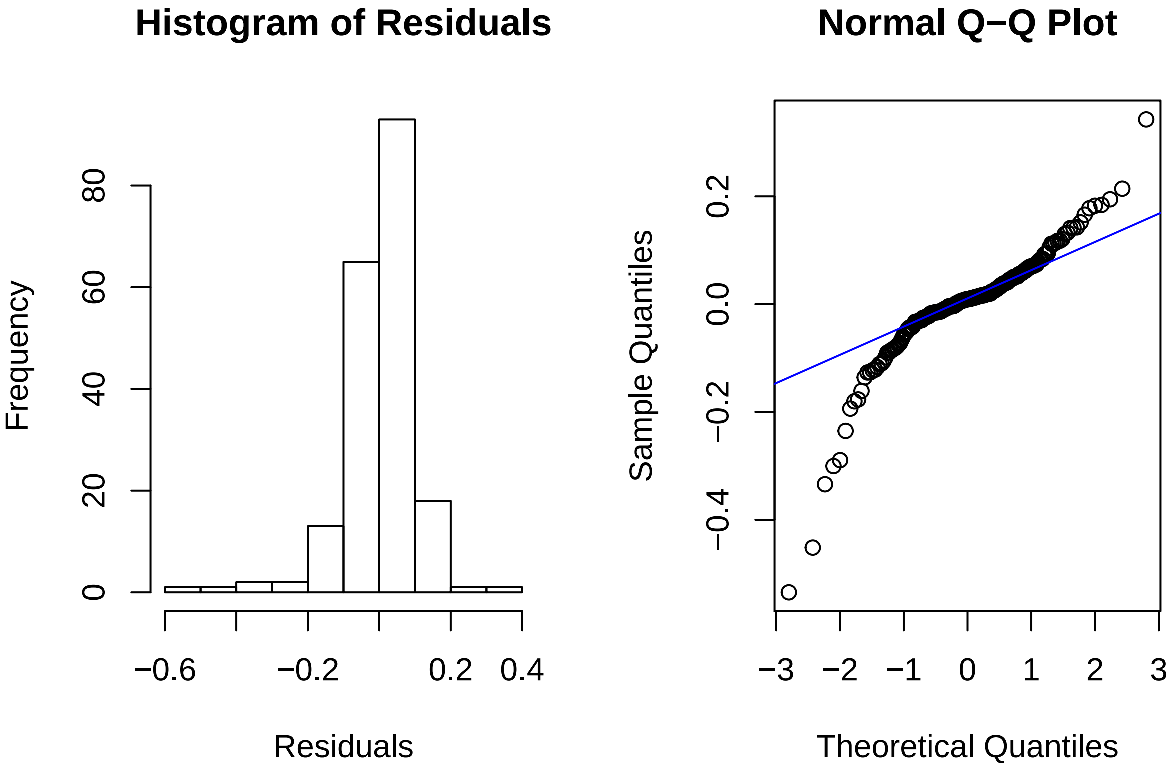
We also plot a QQ-Plot and from that we can see that the errors follow the diagonal line, and so we can assume that the errors are normally distributed. Our histogram shows that our data is normally distributed.

```
##
##  Shapiro-Wilk normality test
##
## data:  err
## W = 0.86508, p-value = 3.169e-12
```

```
##
##  Box-Ljung test
##
## data:  err
## X-squared = 0.080816, df = 1, p-value = 0.7762
```

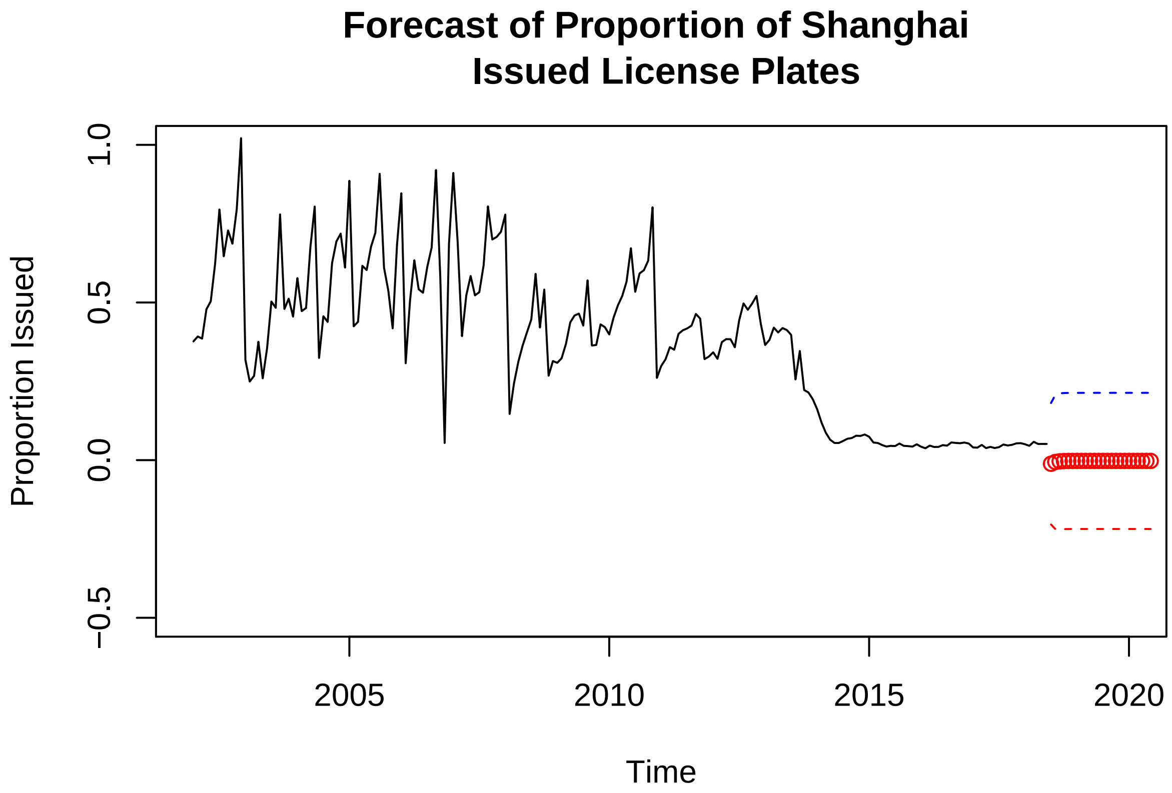
```
##
##  Box-Pierce test
```

```
##
## data:  err
## X-squared = 0.079598, df = 1, p-value = 0.7778
```



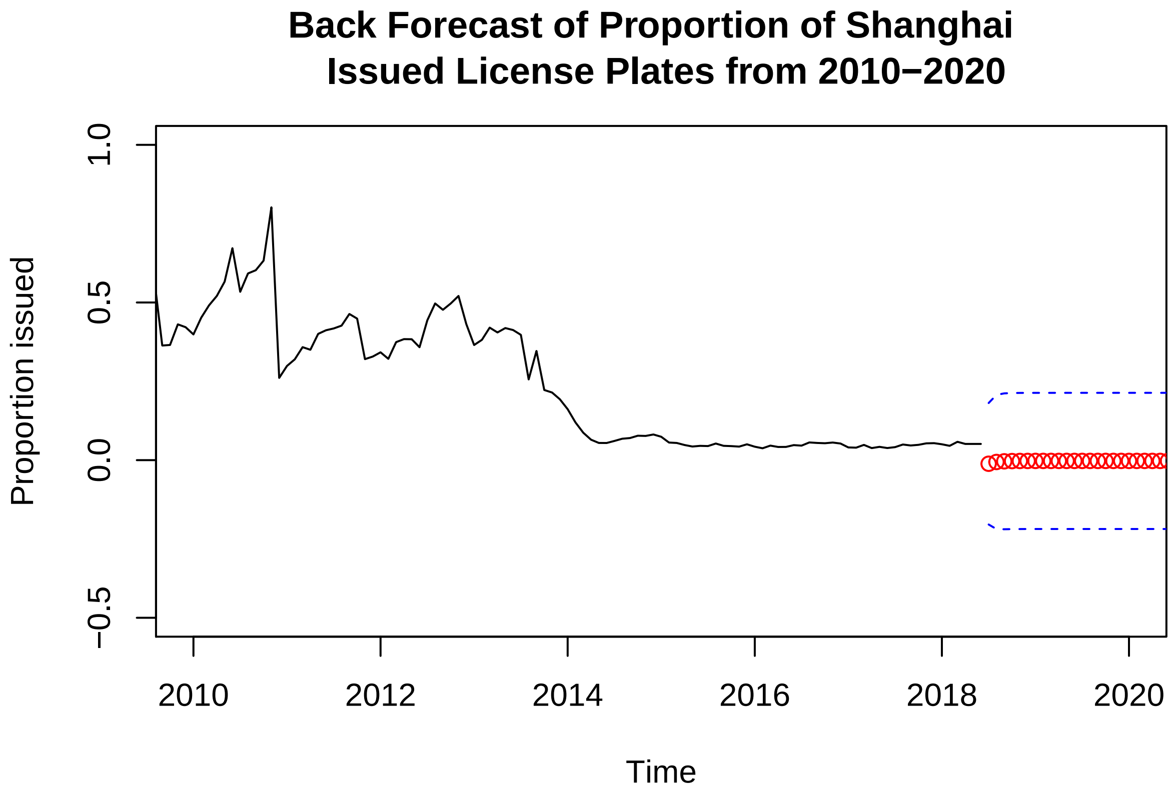
Forecasting

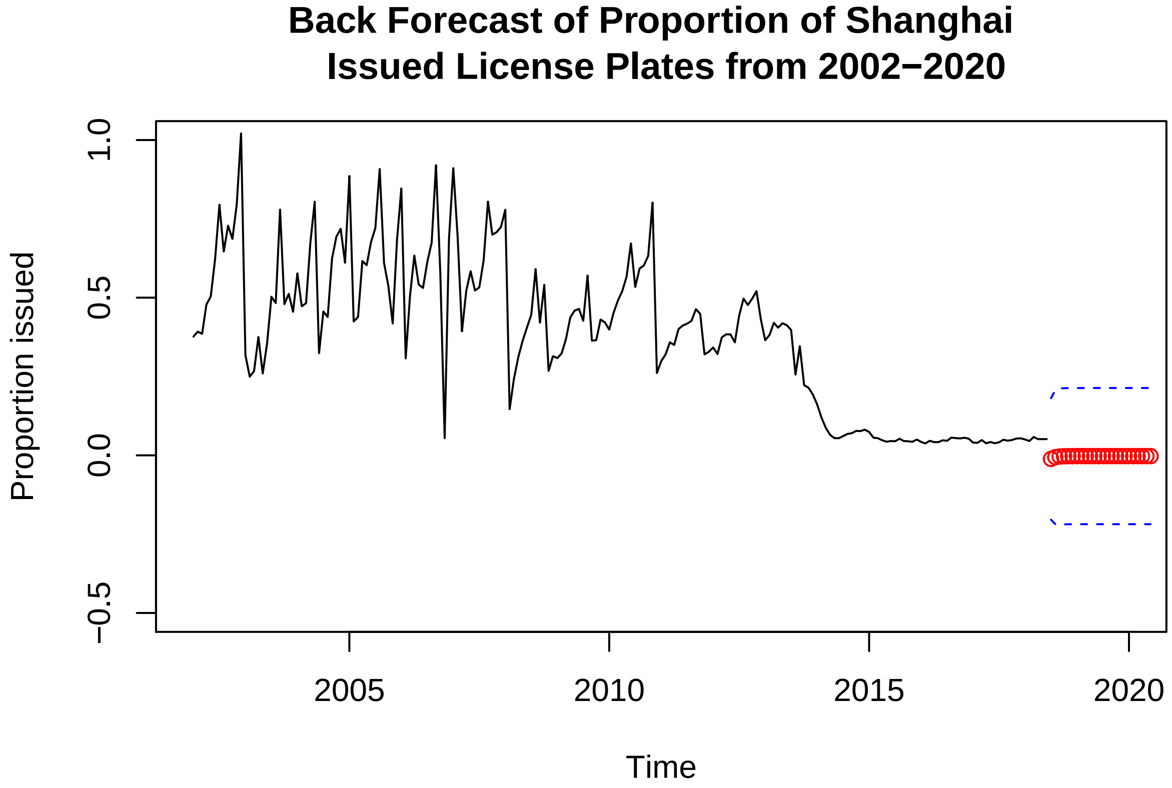
Since we have completed our identification of the proper model, estimated the parameters, and conducted diagnostic checks, we can now move on to forecasting the data. We are going to use forecasting to predict the proportion of license plates issued to number of applicants for the next two years. Since we transformed our data using a Square root transformation, we will need to find the predicted values and then back-transform to forecast our raw data. We used our ARMA(1,1) model and forecasted the next 24 months. We also calculated and plotted an upper and lower confidence interval to calculate a 95% confidence interval for the predicted values.



Back-Transformation

For this portion, we decided to perform back transformation in order to obtain backforecasted values for the proportion of license plates issued for the next 2 years, i.e. 2018-2020. The results show that our predicted values stay consistent for the next two years which means that the proportion of licenses issued in Shanghai to the number of applicants remains relatively constant.


Back Forecast of Proportion of Shanghai
Issued License Plates from 2010−2020
Proportion issued
1.0
0.5
0.0
−0.5
2010
2012
2014
2016
2018
2020
Time


Back Forecast of Proportion of Shanghai
Issued License Plates from 2002−2020
Proportion issued
1.0
0.5
0.0
−0.5
2005
2010
2015
2020
Time

Conclusion

To conclude, we used monthly data to analyze the proportion of Shanghai license plates issued per month to total number of applicants from 2002 to 2018. After transforming and differencing our dataset so that our data is stationary, we used the Yule-Walker method, a for-loop to compare AIC values, and the `auto.arima()` function to conclude that ARMA(1,1) was the best model. After determining our best model, we forecasted values for the next 2 years and found that the predicted values which are closer to 0 show that the proportion will stay relatively consistent as time goes on. There is a consistent trend as number of license plates issued and number of applicants continue to fluctuate as months go on. In other words, if the number of license plates issued increases, the number of applicants will adjust accordingly for the proportion to be stable. This is also true if the number of license plates decreases.

Our results directly relate to the environmental problem at hand where our goal is to contain or reduce pollution. Thus, if a certain number of people apply for a license plate per month, then the Shanghai government attempts to regulate the number of license plates by proportionally reducing the number of license plates available at auction.