STA9701 Time Series Analysis on

TOURIST ARRIVALS IN NEPAL - PART 2

Annapurna Mountain Range[1]

Anustha Shrestha
Baruch College
December 20, 2019

[1] Neupane, Avishek (October 2017)

Introduction

In the previous study of Tourist Arrivals in Nepal[2], monthly tourist arrivals data starting from January 1992 to December 2018 was used to fit a simple ARIMA model. After the analysis, Restricted ARIMA (12,1,1) was proposed as the best model and was used to forecast monthly tourist arrivals for the next 24 periods. Since the real nine-month tourist arrival data was already available, nine-month predictions were used for model comparison. While the restricted ARIMA model seemed to capture the peak and falls of the time series and had a better predictive performance than a basic white noise model, the predictions were found to be more optimistic i.e. the predicted tourist arrivals are slightly higher than the real number of tourist who actually came into the country. The diagnostic plots also show that some autocorrelation still exists in the residuals.

Seasonal patterns were recognized during the first analysis; therefore, this project aims at using Seasonal Multiplicative ARIMA to model tourist arrivals in Nepal. Although tourists visit the country year-round, the popular seasons are during the Spring, between March through June, and in the Fall, between September through November. The Seasonal ARIMA model from this project can be compared with the restricted ARIMA model in terms of predictive performance as well as the diagnostic plots.

It was also noted in the earlier project, that there were some one-time events that might have impacted tourist arrivals; particularly, the earthquakes in 2015 had reduced the number of tourists who entered the country. Furthermore, it is also suspected that the tourist arrivals may have been impacted by the Maoist insurgency and civil riots that occurred between the period from 2001 through 2008. This project will also attempt at analyzing such one-time effects and the impact it has had on tourism. Outlier detection and intervention analysis will be used to understand the impact of these events.

Data Source

Nepal's tourism dataset has been retrieved from Nepal Tourism Statistics 2018 published by the Government of Nepal Ministry of Culture, Tourism and Civil Aviation[3]. The dataset contains monthly tourist arrivals in Nepal starting from January 1992 to December 2018. Altogether there are 324 observations over the years.

Summary

The objective of the project is two-fold: 1) to use monthly tourist arrivals from January 1992 through December 2018 to fit a seasonal multiplicative ARIMA model and compare model performance against those presented in Project 1[2]; 2) analyze the outliers present in the time series and propose a model adjusted for the outlier. The fitted model will be used to forecast for 24 periods, i.e. monthly tourist arrivals for the years 2019 and 2020.

In order to make the time series stationary, log transformation and differencing (d=1) was applied. Since, the ACF of the resulting series still had seasonal non-stationary pattern, seasonal differencing was applied as well (D=1, lag=12). After analyzing the ACF and PACF structures of the seasonal differenced data, a few selections of p and q order were made to test seasonal ARIMA models. Based on AIC and BIC, the best model selected was ARIMA $(1,1,1) \times (0,1,1)_{12}$.

Altogether five outliers were detected in the time series. As suspected the outlier resulting due to the Earthquake in 2015 and the Royal Massacre in 2001 were among the five. Since earthquakes are recurring natural phenomenon in Nepal and it will inevitably impact tourism sector every time it occurs, it is beneficial to study the impact of the disaster. Intervention analysis was conducted and pulse dummy variable was used to specify the intervention effect as the earthquake has an immediate impact that gradually wears off. It is estimated that in the first period the outlier resulted in a decrease of 64.01% in the number of tourists, and by the end of 2018 the percentage had already diminished to 0.07%.

Both the Seasonal ARIMA model as well as the Seasonal ARIMA model adjusted for outlier resulted in more optimistic predictions compared to the real 2019 data available for the first nine months, however, the diagnostic plots look more

[2] Shrestha, Anustha (2019). STA9701 Time Series Analysis on Tourist Arrivals in Nepal.
[3] Government of Nepal Ministry of Culture, Tourism and Civil Aviation. (2018). Nepal Tourism Statistics. Retrieved from: http://tourism.gov.np/statistic

robust for the latter. Based on SSE, the simple Restricted ARIMA model proposed in Project 1[4] is still forecasting better than the models proposed in this project. On a cautionary note, the diagnostic plots of the restricted ARIMA model shows that the model did not satisfy crucial assumptions. While the model may be performing better for the nine-month data available, such restricted model tends to overfit data and may be problematic. The seasonal model seems more robust in terms of the standardized residuals. Future considerations may include analyzing the remaining outliers which was not in the scope of this project.

Exploratory Data Analysis

Data Visualization

The original time series, tourist arrivals in Nepal starting from January 1992 to December 2018 is shown in Figure1. The plot shows an increasing trend, hinting at non-stationarity. The ACF confirms the non-stationarity of the series as it tails off very slowly. It is quite clear that the data has an outlier in 2015, which will be analyzed later using outlier detection and intervention analysis. The big dip in the number of tourists in 2015 is because of the earthquakes that year. Although it doesn't look very clear at first, there is a drop in the tourist arrivals in 2001, which was the year the Royal family was massacred. This resulted in an onset of the Maoist insurgency and a civil unrest in the country, which slowed down tourism.

As noted in the earlier project, the cyclical pattern in the ACF is characteristic of a seasonal model. This project will explore various Seasonal ARIMA models for the time series and forecast 24 periods using the best model available.

Figure1: Monthly Tourist Arrivals in Nepal (1992 - 2018)

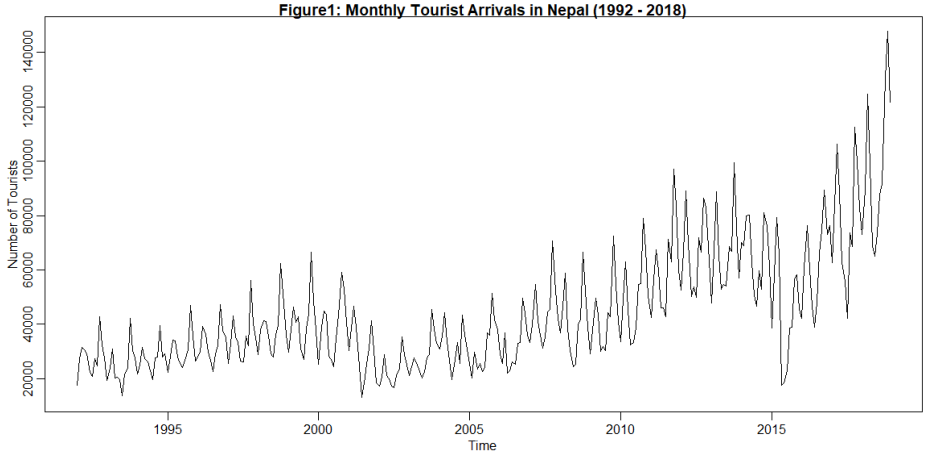
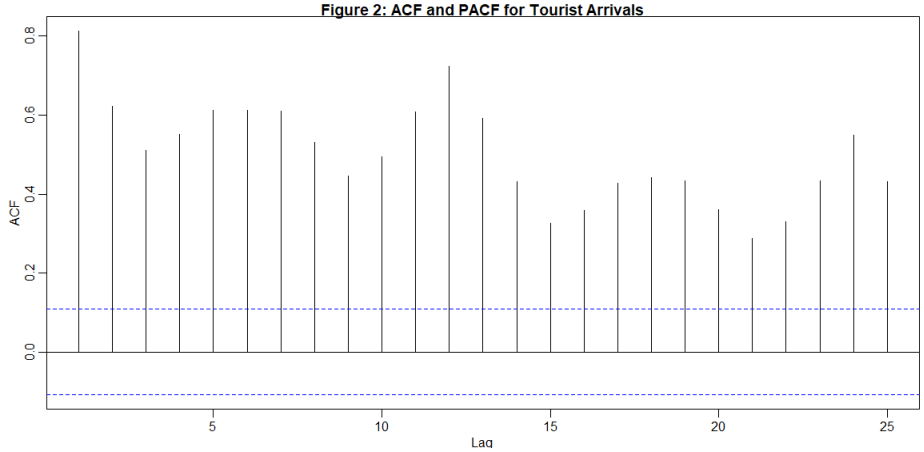


Figure 2: ACF and PACF for Tourist Arrivals



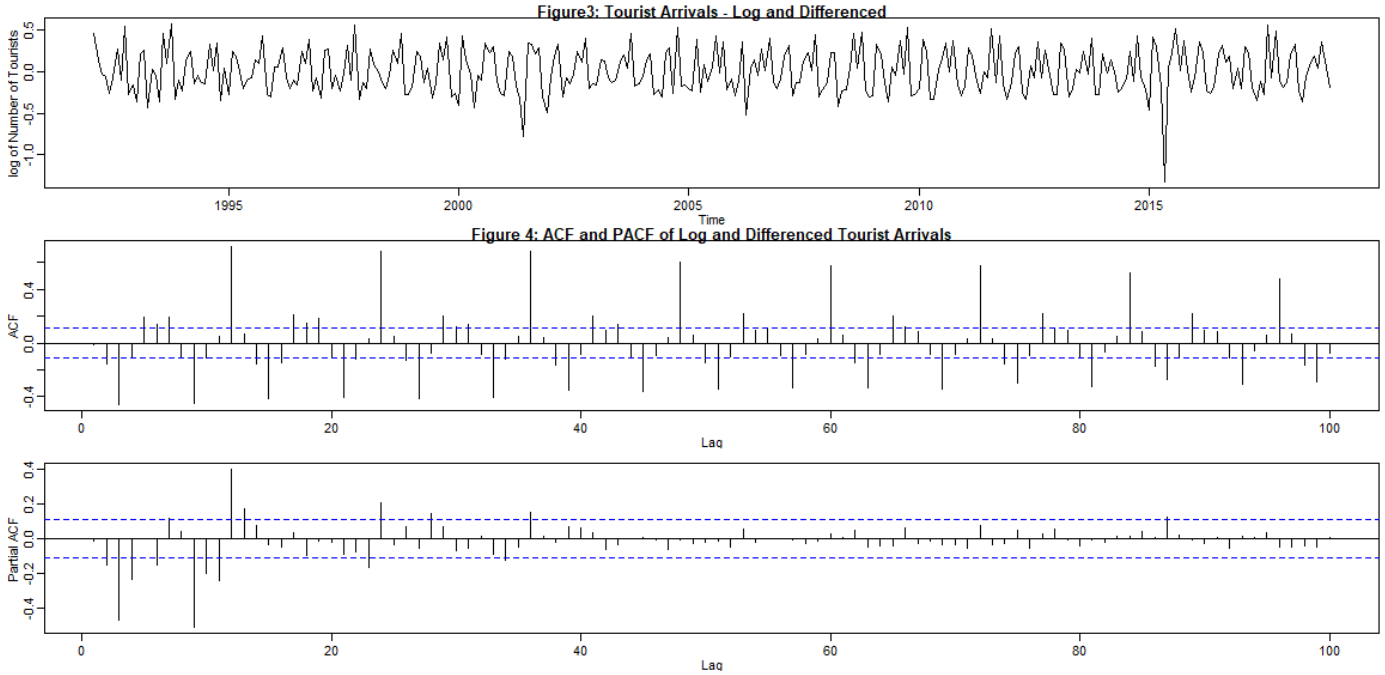
Data Transformation and Differencing

Log transformation and differencing was applied to the data to make the data stationary:

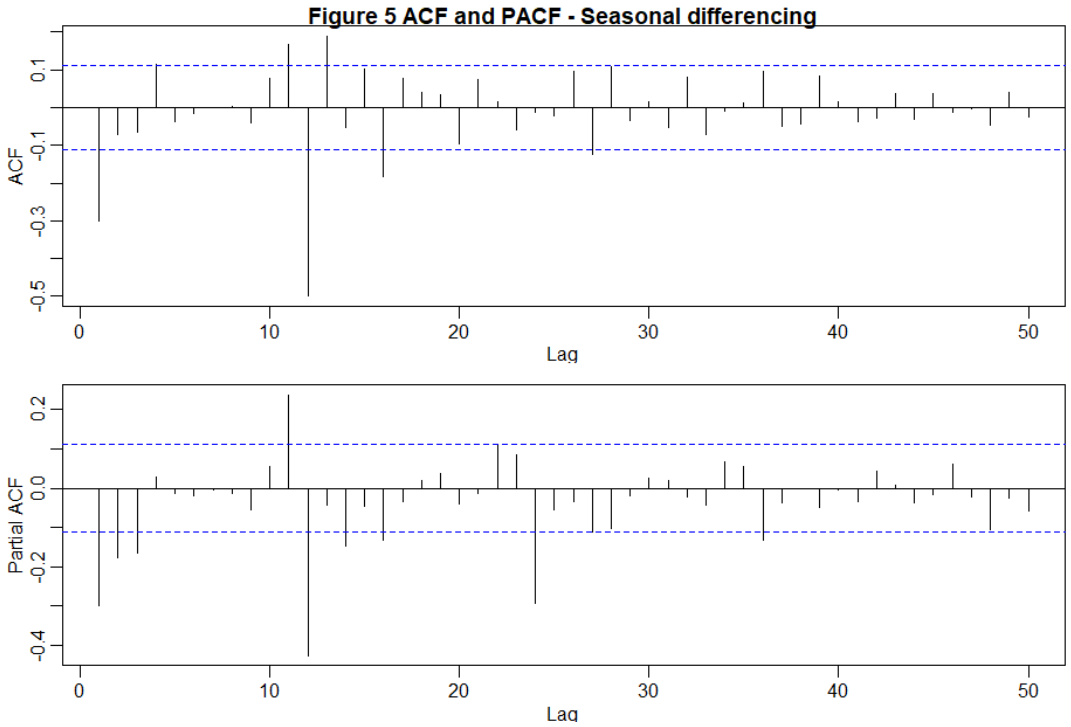
$$\nabla \log(x_t) = (1 - B)\log(x_t) = \log(x_t) - \log(x_{t-1})$$

[4] Shrestha, Anustha (2019). STA9701 Time Series Analysis on Tourist Arrivals in Nepal.

The details behind the transformation has been discussed in the previous project[5]. The resulting series, along with the ACF and PACF can be seen in Figure3 and Figure4.



The ACF plot shows that there is a strong seasonal pattern at lag 12. Furthermore, it can be noted that the ACF at seasonal lags are decreasing, but very slowly, which suggests that seasonal differencing might be appropriate. The ACF and PACF after taking seasonal difference (D=1, lag = 12), is shown in figure5.



[5] Shrestha, Anustha (2019). STA9701 Time Series Analysis on Tourist Arrivals in Nepal.

The seasonal differencing allowed to remove the non-stationary seasonal pattern. Nevertheless, it is important to check for stationarity. Dickey-fuller test (k=0, p-value = 0.01) and augmented Dickey-Fuller test (k=6, p-value = 0.01) both are statistically significant showing that log and differenced time series is stationary.

```
Augmented Dickey-Fuller Test

data:  sdlogtourist
Dickey-Fuller = -7.4961, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

Augmented Dickey-Fuller Test

data:  sdlogtourist
Dickey-Fuller = -23.939, Lag order = 0, p-value = 0.01
alternative hypothesis: stationary
```

Model Selection

Analyzing ACF and PACF of transformed data

We can analyze the ACF and PACF of the resulting series after regular and seasonal differencing (Figure5). PACF is gradually decreasing at every seasonal lag i.e. lag 12, 24, 36 and so on, whereas, ACF is cut off at the first seasonal lag i.e. lag 12 (in lag 24, ACF is on the border). On the other hand, ACF and PACF both seem to be decreasing in the non-seasonal lags. It can be argued that PACF is gradually decreasing and is not significant after the third lag, and that ACF is cut off at the first lag. Since both the ACF and PACF are gradually decreasing in both the seasonal and non-seasonal lags, a variety of Seasonal ARIMA models were selected as candidates and compared in terms of their respective AIC and BIC:

Table1: Model Selection

Candidate Models: AIC and BIC		
Model	**AIC**	**BIC**
ARIMA (3,1,1) × (1,1,1) $_{12}$	-1.0195853	-0.9382855
ARIMA (3,1,1) × (3,1,1) $_{12}$	-1.0169671	-0.9124387
ARIMA (3,1,0) × (3,1,0) $_{12}$	-0.9664184	-0.8851186
ARIMA (3,1,1) × (3,1,2) $_{12}$	-1.0110716	-0.8949289
ARIMA (4,0,1) × (2,0,1) $_{12}$	-1.0128637	-0.896174
ARIMA (1,0,1) × (2,0,2) $_{12}$	-0.9960849	-0.9027332
ARIMA (1,0,0) × (0,0,1) $_{12}$	-0.2709847	-0.2243089
ARIMA (1,1,1) × (0,1,1) $_{12}$	-1.0302574	-0.9838003
ARIMA (1,1,1) × (0,1,0) $_{12}$	-0.560882	-0.5260392
ARIMA (1,1,1) × (1,1,1) $_{12}$	-1.0247187	-0.9666474
ARIMA (1,1,0) × (1,1,0) $_{12}$	-0.7764033	-0.7415605
ARIMA (0,1,1) × (0,1,1) $_{12}$	-1.0163544	-0.9815116
ARIMA (0,1,1) × (0,1,2) $_{12}$	-1.0102947	-0.9638376

Table1 shows the various candidate models with their respective AIC and BIC. The best model, both in terms of AIC and BIC is found to be ARIMA $(1,1,1) \times (0,1,1)_{12}$. We fit the model and check the diagnostic plots.

Model Diagnostics

The final model is given below along with the coefficient estimates. All the coefficients are statistically significant.

ARIMA $(1,1,1) \times (0,1,1)_{12}$:

$$\phi(B)(1-B)(1-B_{12})(X_t) = \theta(B)\Theta(B)w_t$$

where,

$\phi(B) = 1 - \phi_1 B$
$\theta(B) = 1 + \theta_1 B$
$\Theta(B) = 1 + \Theta_1 B^{12}$

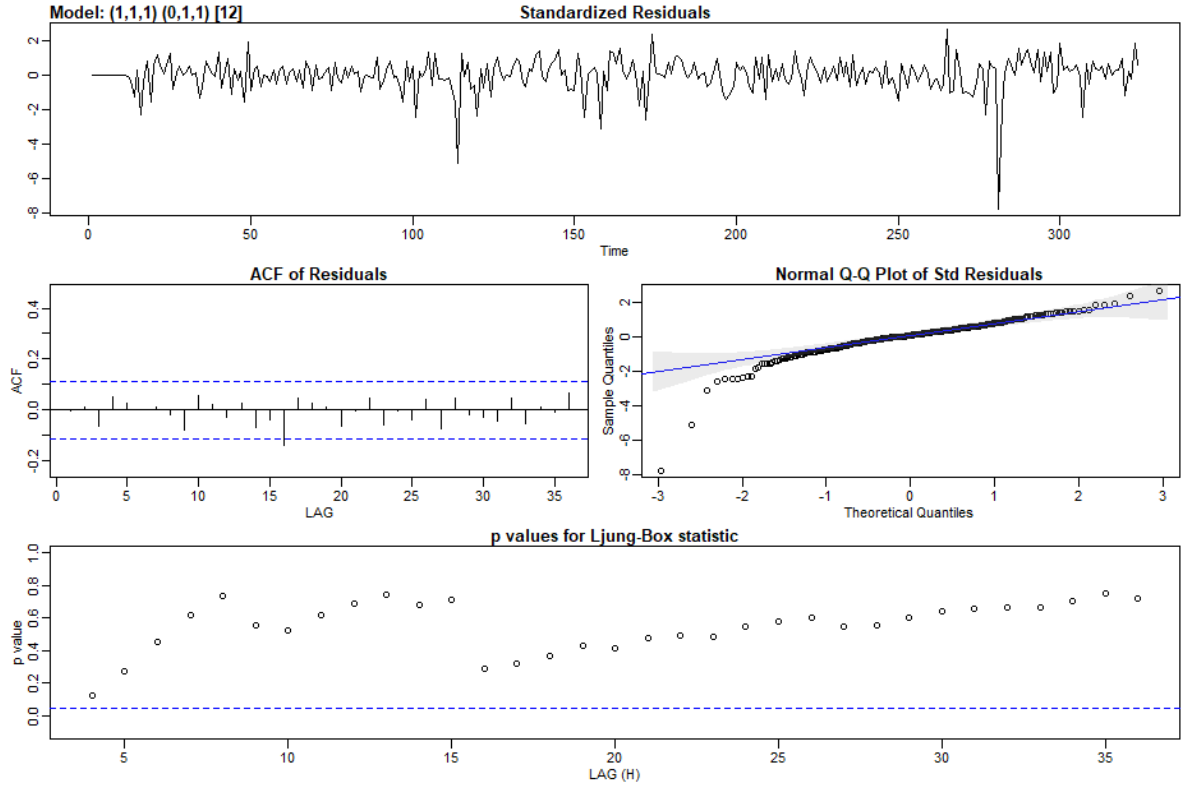
Note: X_t *is the* log of tourist arrivals

Table2: Coefficient Estimates

Coefficients	Estimates	SE	t-value	p-value
$\boldsymbol{\phi_1}$	0.3255	0.1179	2.761	0.0061
$\boldsymbol{\theta_1}$	-0.6788	0.0918	-7.3955	0.0000
$\boldsymbol{\Theta_1}$	-0.8139	0.0403	-20.2018	0.0000

The diagnostic plots of the best seasonal model, ARIMA $(1,1,1) \times (0,1,1)_{12}$ is provided in Figure6. The residuals from the diagnostic plots are mostly spread around 0, except for the outliers we had noted earlier. The Q-Q plot checks for the normality of the residuals. The Q-Q plot looks decent, with a few outliers. ACF of the residuals are not significant and are close to zero, except for lag 16, which is right on the border. The p-value for Ljung-Box statistic are greater than the significance threshold, which shows that the residuals are not correlated.

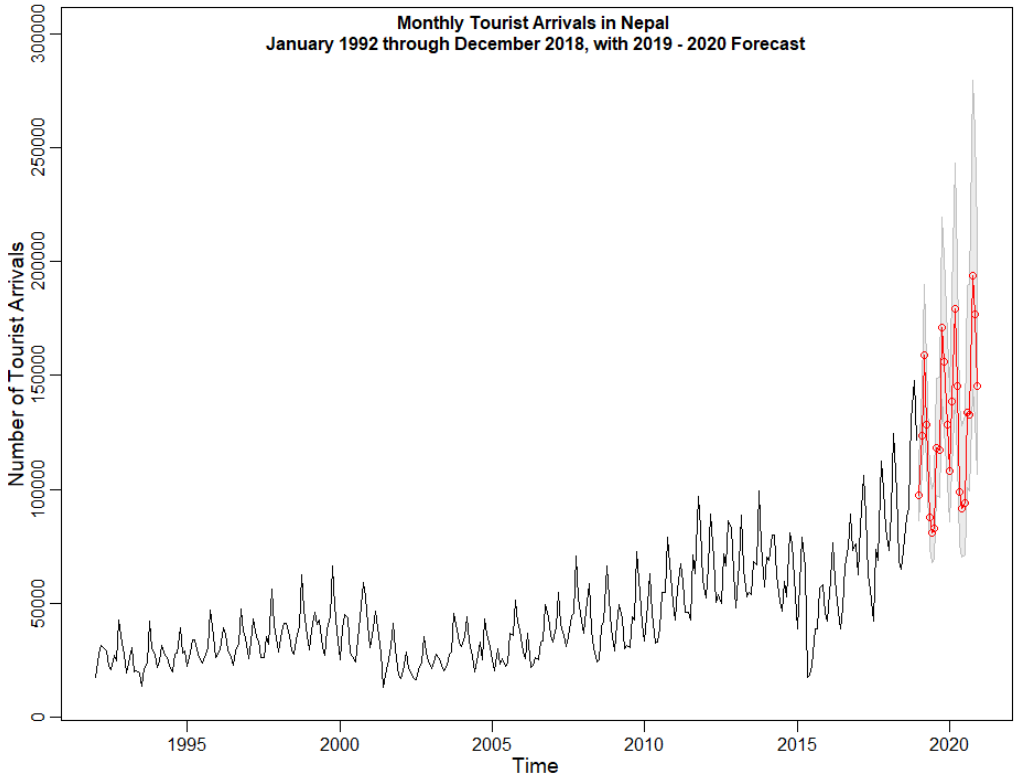
Figure6: Diagnostics - ARIMA $(1,1,1) \times (0,1,1)_{12}$



Forecasting

Future 24 observation were predicted using ARIMA $(1,1,1) \times (0,1,1)_{12}$ model. Figure7 shows the forecast with confidence intervals for 24 periods (2019 and 2020). This model seems to capture the peaks and the falls of the series well, but the confidence interval is very wide.

Figure7: Forecasting for 24 periods- ARIMA $(1,1,1) \times (0,1,1)_{12}$



The final forecast of ARIMA $(1,1,1) \times (0,1,1)_{12}$ model is provided below:

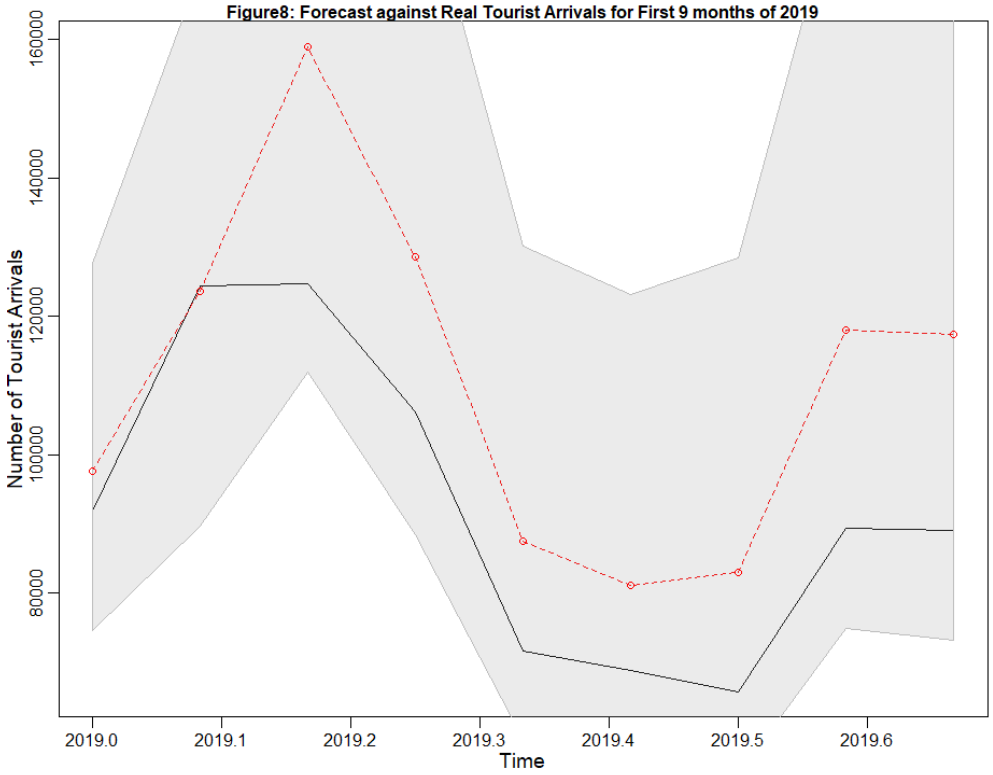
	Jan	Feb	Mar	Apr	May	Jun
2019	97595	123550	158976	128611	87513	81110
2020	107986	138807	179497	145448	99022	91792
	Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2019	83030	118037	117325	171112	156100	128404
2020	93970	133592	132787	193663	176673	145327

Finally, the forecast for 9 periods (January 2019 to September 2019) was compared against the actual tourist arrival data that is available[6]. The forecasted data (red dotted line) against the real 2019 data (black line) is shown in Figure8. The prediction seems a little optimistic, but we notice that it is generally following the pattern of the real data. The SSE for Restricted ARIMA (12,1,1) model was 0.3584, which was found to be better when compared to SSE 0.4030 and 1.4807 of the ARIMA (12,1,1) model and white noise models respectively[7].

The SSE of the Seasonal ARIMA $(1,1,1) \times (0,1,1)_{12}$ model based on the 9-month forecast against the true number of tourist arrival is 0.3746, which is better than the white noise model and simple ARIMA models, but not better than the Restricted ARIMA (12,1,1) model. However, it should be noted that only 9 new data points were used for testing model performance. Increasing the number of test data would allow for a better comparison.

[6] CEIC, (2019). Retrieved from: https://www.ceicdata.com/en/indicator/nepal/visitor-arrivals

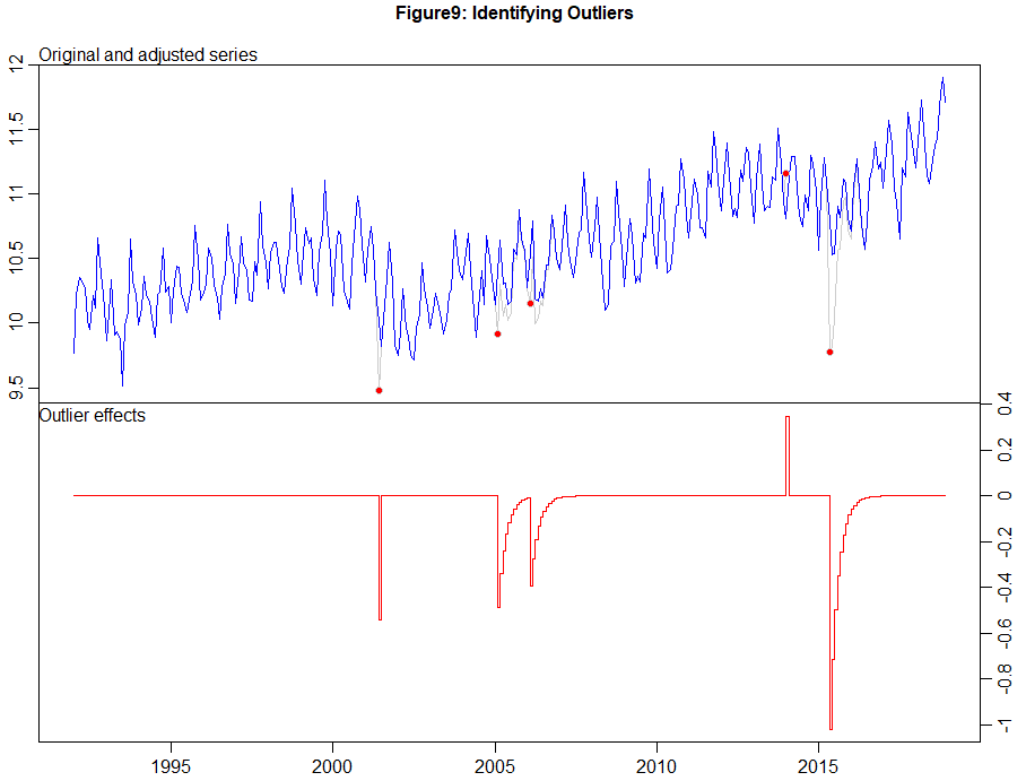
[7] Shrestha, Anustha (2019). STA9701 Time Series Analysis on Tourist Arrivals in Nepal.



Outlier Detection and Intervention Analysis

Identifying Outliers

As mentioned earlier, there are outliers present in the dataset, which impacts the model. In this section, we will attempt to detect outliers and analyze them. Using the tso() functions, the outliers that are detected are shown in Figure9 below:



Altogether there were 5 different outliers detected from the procedure. The outlier that seems to have the most impact in our time series is the Temporary Change (TC) outlier in May 2015. Between April 2015 through May 2015, two major earthquakes of over 7.1 Richter scale devastated the country, which were followed by many aftershocks. It is natural that during this volatile period, the number of tourists travelling into the country diminished dramatically. Such event causes an immediate impact that wears off after some time. Thus, the temporary change outlier makes sense in this case.

The other impactful outlier is in 2001, which is an Additive outlier. In June 2001, a tragedy had struck Nepal – the Royal Family of the then kingdom had been massacred, which created a state of confusion, grief, and national mourning. Therefore, the one time fall in the number of tourists is a reasonable assumption. Similarly, the outliers in 2005 and 2006 were probably related to the insurgency in the country.

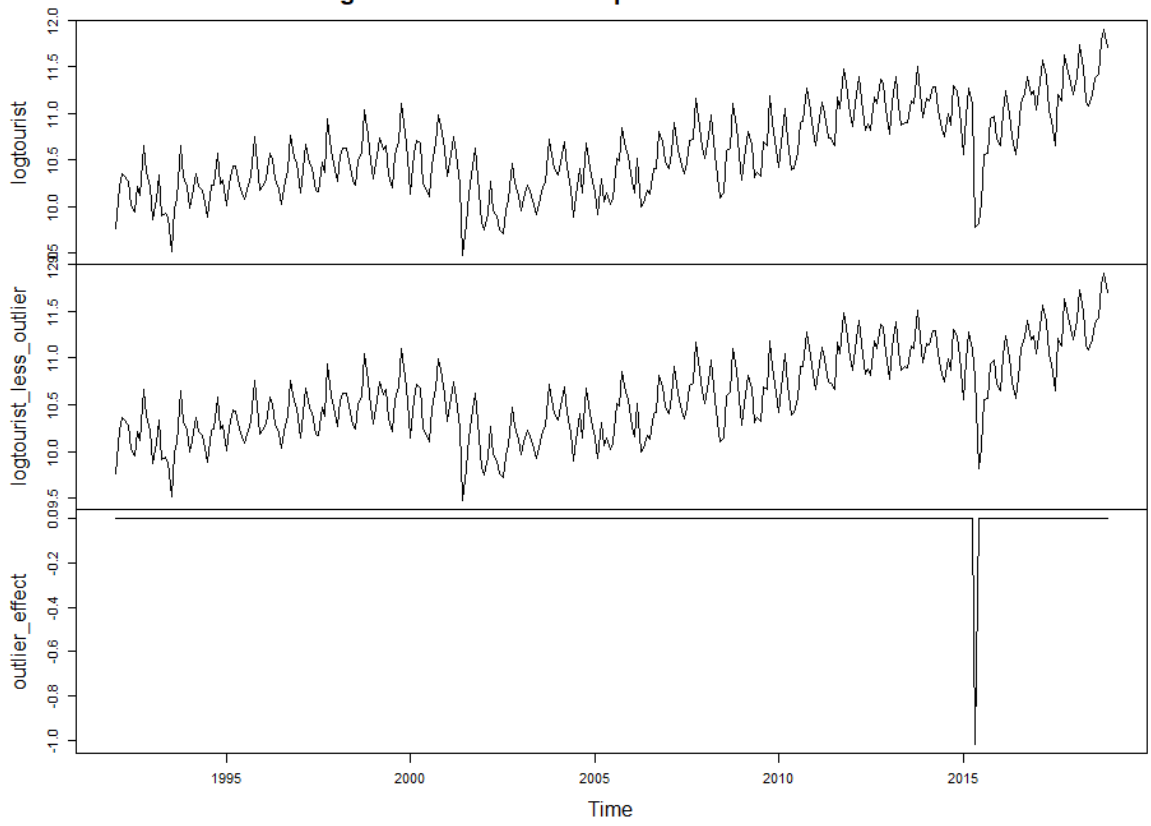
Table3: Outlier Detection

type	ind	time	coefhat	tstat
AO	114	2001:06	-0.5406	-6.245
TC	158	2005:02	-0.4867	-5.125
TC	170	2006:02	-0.3847	-4.032
AO	265	2014:01	0.349	4.028
TC	281	2015:05	-1.0202	-10.693

For my project, I choose to analyze the temporary change outlier in 2015, which resulted from the earthquake. Nepal is in an earthquake-prone zone and occurrence of the natural disaster in the country is inevitable; therefore, it might be beneficial to model the impact earthquakes may have on the economy. First, I will be modeling the impact it has on the number of tourist arrivals for a given period of time; then I will try to forecast for 2019 and 2020 using a new ARIMA $(1,1,1) \times (0,1,1)_{12}$ model adjusted for the outlier.

The plot below shows the original time series, the time series adjusted for outlier and the impact of the outlier on May 2015:

Figure10: Effect of Earthquake on Tourist number



Intervention Analysis

We can consider a simple intervention model:

$$y_t = m_t + x_t$$

where,
x_t is the process model by the Seasonal ARIMA
m_t is the impact of the intervention / change in the mean

Since this is a transient intervention, it can be specified using the pulse function, such that at t = T,

$$m_t = \omega_0 P_t^{(T)} + \frac{\omega_1}{1-\omega_2 B} P_t^{(T)}$$

where,
T is the time when the event occurs and

$$P_t^{(T)} \begin{cases} 1, & t = T \\ 0, & otherwise \end{cases}$$

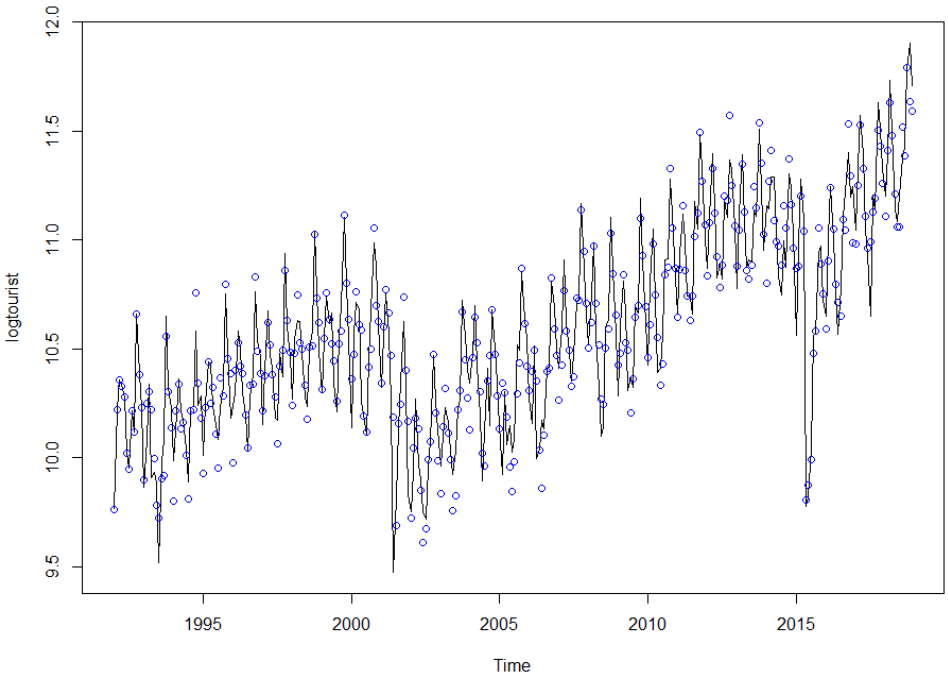
In R, arimax() function allows to model ARIMA adjusting for outliers. The coefficients obtained are as follows:

Table4: Coefficient Estimates with Outlier Effects

Coefficients	Estimates	SE	z	p-value
ϕ_1	0.2735	0.1131	2.417	0.0157
θ_1	-0.6885	0.0871	-7.908	0.0000
Θ_1	-0.8089	0.0471	-17.171	0.0000
ω_0	0.0312	0.1775	0.176	0.8603
ω_2	0.8160	0.0692	11.784	0.0000
ω_1	-1.0532	0.1806	-5.831	0.0000

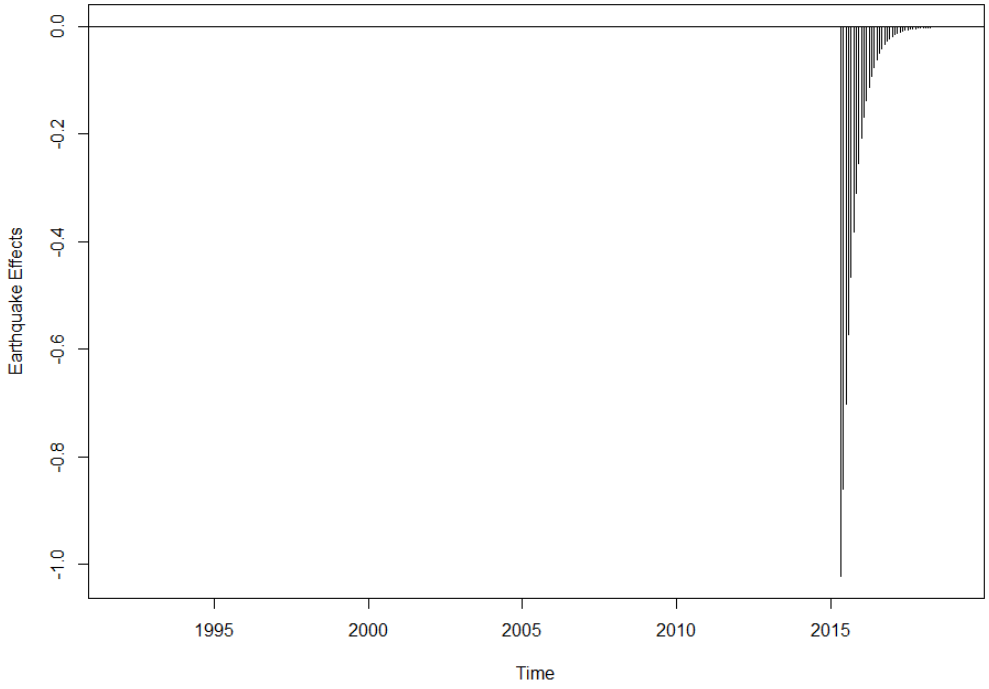
The fitted time series using this ARIMA $(1,1,1) \times (0,1,1)_{12}$ adjusted for the outlier against the original time series is shown in Figure11:

Figure11: Original Time Series against Fitted Time Series



The intervention effect is plotted out in Figure12. In 2015, there is a sudden drop, then the effects gradually wear off after each period.

Figure12: Intervention Effects
Impact of the 2015 Earthquake



In the first year, the model estimates that the earthquake reduced the number of tourists entering Nepal by 64.01%. By the next year, the impact had reduced to 8.76% and by the end of 2018, the effect of the earthquake had already reached to 0.07%, which is an almost non-existent impact.

Adjusted Model and Diagnostic Plots

As noted earlier, the final model can be written as:

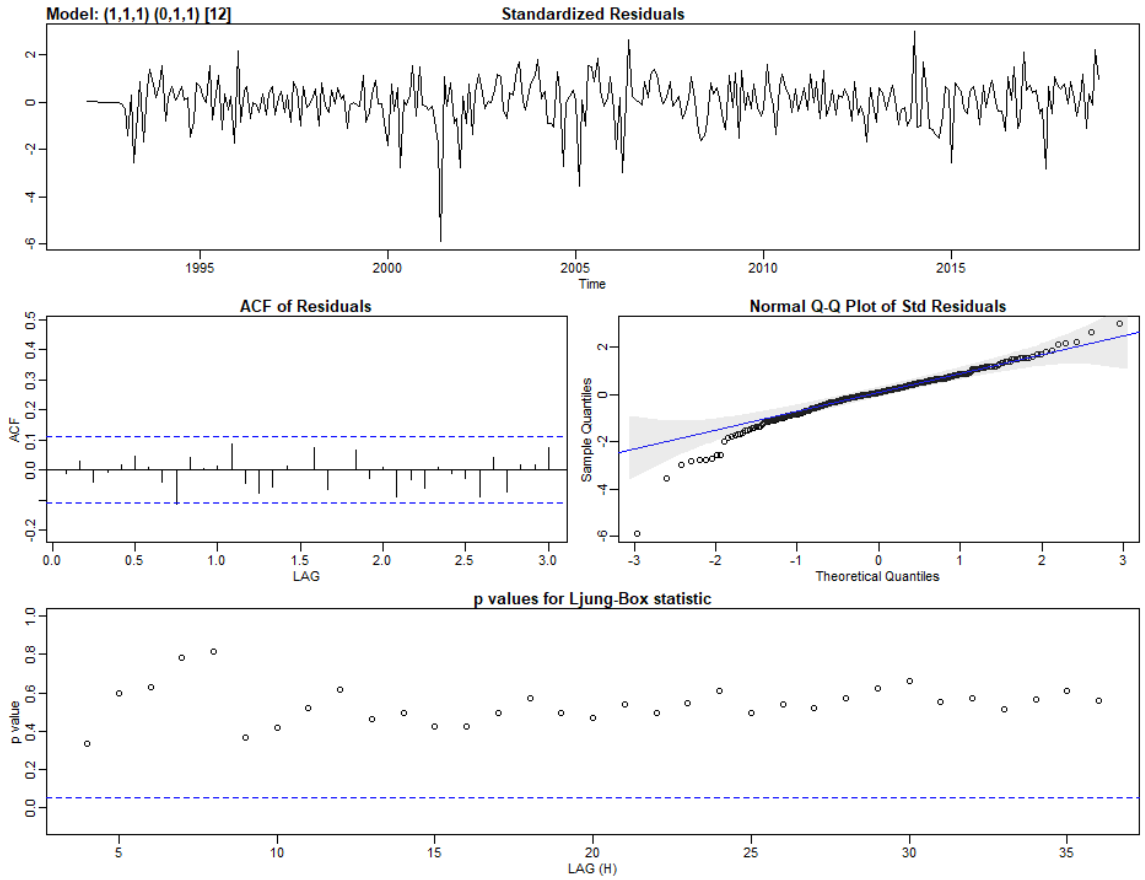
$$y_t = y_{t-1} + y_{t-12} + \epsilon_t + \epsilon_{t-1} + \Theta\epsilon_1 + \omega_0 P_t^{(T)} + \frac{\omega_1}{1-\omega_2 B} P_t^{(T)}$$

Table4: Coefficient Estimates with Outlier Effects

Coefficients	Estimates	SE	z	p-value
$\boldsymbol{\phi_1}$	0.2735	0.1131	2.417	0.0157
$\boldsymbol{\theta_1}$	-0.6885	0.0871	-7.908	0.0000
$\boldsymbol{\Theta_1}$	-0.8089	0.0471	-17.171	0.0000
$\boldsymbol{\omega_0}$	0.0312	0.1775	0.176	0.8603
$\boldsymbol{\omega_2}$	0.8160	0.0692	11.784	0.0000
$\boldsymbol{\omega_1}$	-1.0532	0.1806	-5.831	0.0000

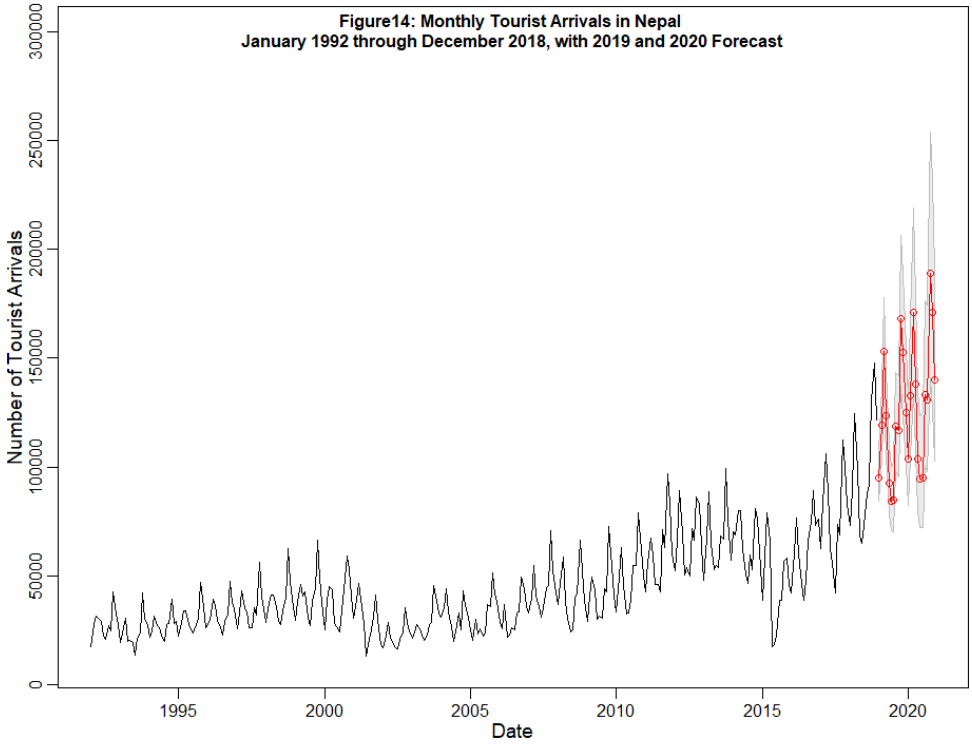
The diagnostic plots for the model are provided in Figure13. We can see that the ACF of the residuals look more like a white noise series. The p-value of the Ljung-Box test also suggests that the residuals are not autocorrelated. However, the Q-Q plot still shows that there are some outliers. In future analysis, the other outliers could be studied in detail and adjusted for accordingly.

Figure13: Diagnostics - ARIMA (1,1,1) × (0,1,1) $_{12}$ adjusted for outlier



Adjusted Model and Forecast

Future 24 observations were predicted using ARIMA (1,1,1) × (0,1,1) $_{12}$, adjusted for the 2015 outlier. Figure14 shows the forecast with confidence intervals for 24 periods (2019 and 2020).

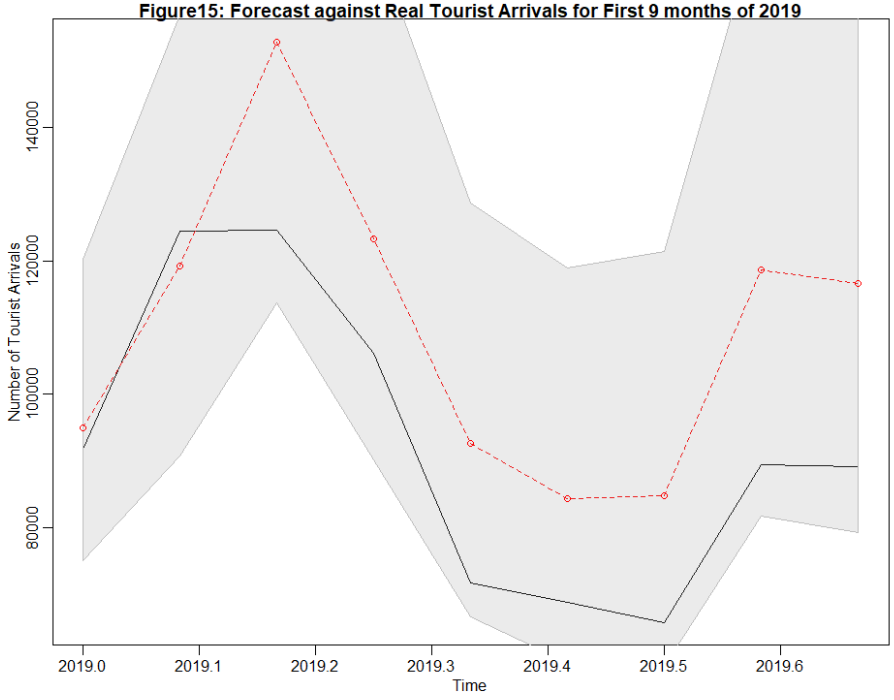


The final forecast of the adjusted ARIMA $(1,1,1) \times (0,1,1)_{12}$ model is provided below:

	Jan	Feb	Mar	Apr	May	Jun
2019	94942	119296	152928	123352	92593	84250
2020	103914	132878	171156	138236	103802	94459
	Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**
2019	84720	118702	116603	168314	152677	124876
2020	94989	133090	130737	188716	171184	140013

Finally, the forecast for 9 periods (January 2019 to September 2019) was compared against the actual tourist arrival data that is available[8]. The forecasted data (in red dotted) against the real 2019 data (black line) is shown in Figure15. The prediction seems optimistic, slightly more than the previous forecast (Seasonal model without outlier adjustment), but we notice that it is generally following the pattern of the real data. The SSE for ARIMA $(1,1,1) \times (0,1,1)_{12}$ adjusted for the temporary change outlier is 0.3914, which is slightly higher than the unadjusted model (0.3746)

Figure15: Forecast against Real Tourist Arrivals for First 9 months of 2019



Conclusion

Although the diagnostic plots for the ARIMA $(1,1,1) \times (0,1,1)_{12}$ model adjusted for outliers are better than the diagnostic plots for the unadjusted Seasonal ARIMA and the Restricted ARIMA models[9], the predictions are a bit optimistic when compared with the actual data for the first 9 months of 2019. On the other hand, using the restricted ARIMA model might result in the problem of overfitting. Given the poor diagnostic plot of the Restricted ARIMA (12,1,1) model, it is better to use a more general model such as the Seasonal ARIMA model.

The outlier in May 2015 which resulted due to the devastating earthquakes in Nepal, was identified as a temporary change outlier; there is an immediate decrease of 64.01% in the number of tourist arrival in the first period, however, the impact gradually decreased to 0.07% by the end of 2018. This model could be used to forecast impact on tourism, given an earthquake event. There were additional outliers identified in the time series, which did not fall in the scope of analysis of this project. The models proposed here could be made more robust by analyzing those outliers and adjusting for them accordingly.

[8] CEIC, (2019). Retrieved from: https://www.ceicdata.com/en/indicator/nepal/visitor-arrivals

[9] Shrestha, Anustha (2019). STA9701 Time Series Analysis on Tourist Arrivals in Nepal.

Appendix

R Code

```
library(forecast)
library(TSA)
library(tseries)
library(astsa)
library(lmtest)
setwd("C:/Users/Anustha Shrestha/Documents/RSTA9701")
tourism = read.csv("NepalTourism.csv")
touristArrivals = tourism[,3]#timeseries data
tourist = ts(touristArrivals, frequency = 12, start = c(1992,1))#data as time series

################################## SIMPLE ARIMA MODEL (PROJECT 1)##############################
#plotting the time series data
plot (tourist, ylab = "Number of Tourists", main= "Figure1: Monthly Tourist Arrivals in Nepal (1992 - 2018)")
acf(touristArrivals, main = "Figure 2: ACF and PACF for Tourist Arrivals")
pacf(touristArrivals)

#Sample acf and pacf of differenced series
dtourist <- diff(touristArrivals, 1)
par(mfrow=c(1,1))
plot.ts(dtourist, ylab = "Number of Tourists: Differenced", main="Tourist Arrivals - Log and Differenced")
acf(dtourist,100, main="ACF : Differenced")
pacf(dtourist,100, main="PACF: Differenced")

# log of tourist arrivals and sample acf and pacf of differenced series
logtourist = log(touristArrivals)
dlogtourist <- diff(logtourist, 1)
par(mfrow=c(1,1))
tsdlogtourist<-ts(dlogtourist, frequency = 12, start=c(1992,2))
par(mfrow=c(3,1))
plot.ts(tsdlogtourist, ylab = "log of Number of Tourists", main="Figure3: Tourist Arrivals - Log and Differenced")
acf(dlogtourist,100, main="Figure 4: ACF and PACF of Log and Differenced Tourist Arrivals")
pacf(dlogtourist,100, main="PACF")

#Model diagnostics
sarima(logtourist,12,1,1)

#Forecasting
#arima (12,1,1)
arima1211<-arima(logtourist, order=c(12,1,1), xreg=1:nobs) #arima(12,1,1)
sarima(logtourist, 12,1,1)
sarima.for(logtourist, 24,12,1,1, newxreg = (nobs+1): (nobs+24))

#Restricted Arima(12,1,1)
```

```
foreR1211 = sarima.for(logtourist, 24,12,1,1, fixed=c(0,0,NA,0,0,NA,0,0,NA,0,0, NA, NA, 0), newxreg = (nobs+1): (nobs+24))
foreWN = sarima.for(logtourist, 24,0,1,0, newxreg = (nobs+1): (nobs+24))

#Model comparison using 9 period prediction against real data
whitenoise <-sarima(logtourist, order=c(0,1,0), xreg=1:nobs)
sarima(logtourist, 0,1,0)
sarima.for(dlogtourist, 24, 0,1,0, newxreg = (nobs+1):(nobs+24))
true <- c(91793, 124421, 124697, 106136, 71640, 68782, 65749, 89382, 89078)
logtrue = log(true)
testforeR1211 = sarima.for(logtourist, 9,12,1,1, fixed=c(0,0,NA,0,0,NA,0,0,NA,0,0, NA, NA, 0), newxreg = (nobs+1): (nobs+9))
testfore1211 = sarima.for(logtourist, 9,12,1,1,newxreg = (nobs+1): (nobs+9))
testforeWN = sarima.for(logtourist, 9,0,1,0, newxreg = (nobs+1): (nobs+9))

RArimaSSE = sum((testforeR1211$pred - logtrue)^2)
ArimaSSE = sum((testfore1211$pred - logtrue)^2)
WNSSE = sum((testforeWN$pred - logtrue)^2)

#Plotting out the predicted vs. 9 period
truets <- ts(true, start = c(2019,2), frequency=12)
Rarima1211P <-exp(testforeR1211$pred)
ts.plot(cbind(truets, ts(Rarima1211P, start=c(2019,2), frequency=12)), gpars = list(lty =c(1:3),col=1:2, ylab="Number of Tourist Arrivals", main="Figure12: Forecast against Real Tourist Arrivals for First 9 months of 2019"))

############################SEASONAL MODEL (Project 2) ######################################
#Differencing
logtourist = log(touristArrivals)
dlogtourist <- diff(logtourist, 1)
par(mfrow=c(1,1))
tsdlogtourist<-ts(dlogtourist, frequency = 12, start=c(1992,2))
par(mfrow=c(3,1))
plot.ts(tsdlogtourist, ylab = "log of Number of Tourists", main="Figure3: Tourist Arrivals - Log and Differenced")
acf(dlogtourist,100, main="Figure 4: ACF and PACF of Log and Differenced Tourist Arrivals")
pacf(dlogtourist,100, main="PACF")

#Seasonal Differencing
sdlogtourist <- diff(dlogtourist, 12)
adf.test(sdlogtourist)
adf.test(sdlogtourist, k=0)
par(mfrow=c(2,1))
acf(sdlogtourist, lag=50, main="Figure 5 ACF and PACF - Seasonal differencing")
pacf(sdlogtourist, lag =50)

#Model Selection
M1<-sarima(logtourist,3,1,1,1,1,1,12)
M2<-sarima(logtourist,3,1,1,3,1,1,12)
M3<-sarima(logtourist,3,1,0,3,1,0,12)
```

```
M4<-sarima(logtourist,3,1,1,3,1,2,12)
M5<-sarima(logtourist,4,0,1,2,0,1,12)
M6<-sarima(logtourist,1,0,1,2,0,2,12)
M7<-sarima(logtourist,1,0,0,0,0,1,12)
M8<-sarima(logtourist,1,1,1,0,1,1,12)##bestest AIC and BIC
M9<-sarima(logtourist,1,1,1,0,1,0,12)
M10<-sarima(logtourist,1,1,1,1,1,1,12)
M11<-sarima(logtourist,1,1,0,1,1,0,12)
M12<-sarima(logtourist,0,1,1,0,1,1,12)
M13<-sarima(logtourist,0,1,1,0,1,2,12)
M14<-sarima(logtourist,0,1,1,0,1,2,12)

AIC_BIC<- data.frame(c1 = c(M1$AIC, M2$AIC, M3$AIC, M4$AIC,
M5$AIC,M6$AIC,M7$AIC,M8$AIC,M9$AIC,M10$AIC,M11$AIC,M12$AIC,M13$AIC,M14$AIC),
                     c2 = c(M1$BIC, M2$BIC, M3$BIC, M4$BIC,
M5$BIC,M6$BIC,M7$BIC,M8$BIC,M9$BIC,M10$BIC,M11$BIC,M12$BIC,M13$BIC,M14$BIC ))

colnames(AIC_BIC) <- c("AIC", "BIC")
rownames(AIC_BIC) <- c( "ARIMA(3,1,1)(1,1,1)[12]", "ARIMA(3,1,1)(3,1,1)[12]",
                        "ARIMA(3,1,0)(3,1,0)[12]", "ARIMA(3,1,1)(3,1,2)[12]",
                        "ARIMA(4,0,1)(2,0,1)[12]", "ARIMA(1,0,1)(2,0,2)[12]",
                        "ARIMA(1,0,0)(0,0,1)[12]", "ARIMA(1,1,1)(0,1,1)[12]",
                        "ARIMA(1,1,1)(0,1,0)[12]", "ARIMA(1,1,1)(1,1,1)[12]",
                        "ARIMA(1,1,0)(1,1,0)[12]", "ARIMA(0,1,1)(0,1,1)[12]",
                        "ARIMA(0,1,1)(0,1,2)[12]", "ARIMA(0,1,1)(0,1,2)2[12]"
                        )
AIC_BIC

################################### Best Seasonal Model ###################################
par(mfrow = c(1,1))
M8<-sarima(logtourist,1,1,1,0,1,1,12)
M8

#24-period forecast and plot
SARIMA_for<-sarima.for(logtourist, 24,1,1,1,0,1,1,12, newxreg = (nobs+1): (nobs+24))
SARIMA_for_ts<-ts(SARIMA_for$pred,frequency=12, start=c(2019,1))
SARIMA_for_se<-ts(SARIMA_for$se,frequency=12, start=c(2019,1))
exp(SARIMA_for_ts) #prediction in original scale
par(cex.axis = 1, cex.lab = 1.2, cex.main = 1)
ts.plot(tourist, exp(SARIMA_for_ts),
      ylab = "Number of Tourist Arrivals",
      xlab = "Time",
      ylim = c(10000,300000))
title("Monthly Tourist Arrivals in Nepal \nJanuary 1992 through December 2018, with 2019 - 2020 Forecast", line=-2)
U1 = exp(SARIMA_for_ts+SARIMA_for_se); L1 = exp(SARIMA_for_ts-SARIMA_for_se)
xx1 = c(time(U1), rev(time(U1))); yy1 = c(L1, rev(U1))
polygon(xx1, yy1, border = 8, col = gray(.6, alpha = .2))
lines(exp(SARIMA_for_ts), type = "p", col = 2)
```

```
lines(exp(SARIMA_for_ts), type = "l", col = 2)


#Compare 9-period forecast against real
true <- c(91793, 124421, 124697, 106136, 71640, 68782, 65749, 89382, 89078)
logtrue = log(true)
SARIMAA<-sarima.for(logtourist, 9,1,1,1,0,1,1,12, newxreg = (nobs+1):(nobs+9))
#SSE
SARIMASSE = sum((SARIMAA$pred - logtrue)^2)
SARIMASSE
#9-period forecast plot
truets <- ts(true, start = c(2019,1), frequency=12)
SARIMAA_ts <-ts(SARIMAA$pred, frequency=12, start=c(2019,1))
SARIMAA_se <-ts(SARIMAA$se, frequency=12, start=c(2019,1))
ts.plot(cbind(truets, exp(SARIMAA_ts)), gpars = list(lty =c(1:3),col=1:2, ylab="Number of Tourist Arrivals", main="Figure8:
Forecast against Real Tourist Arrivals for First 9 months of 2019"))
U2 = exp(SARIMAA_ts+1.96*SARIMAA_se); L2 = exp(SARIMAA_ts-1.96*SARIMAA_se)
xx2 = c(time(U2), rev(time(U2))); yy2 = c(L2, rev(U2))
polygon(xx2, yy2, border = 8, col = gray(.6, alpha = .2))
lines(exp(SARIMAA_ts), type = "p", col = 2)

########################  Outlier Detection ################################################
#outlier Detection on the log series
logtourist<-ts(logtourist, frequency = 12, start=c(1992,1))
outlier<-tso(logtourist)
outlier
plot(outlier)
title("Figure9: Identifying Outliers")

#outlier detection on the log difference series
tsdlogtourist<-ts(tsdlogtourist, frequency = 12, start=c(1992,2))
outlier1<-tso(tsdlogtourist)
outlier1
plot(outlier1)

# As expected there were 2 outliers in the series
# Note that the outliers go away when we difference the series
# I am interested in this
# 1 Additive outlier was in 2001 June <- The royal family massacre
# 2 Temporary change outlier in 2015 May <- Earthquake

# converting to a number #location of TC outlier is 281
tc <- rep(0, length(logtourist))
tc[281]<-1
coefhat <- outlier$outliers["coefhat"]
coefhat <- coefhat[5,]
coefhat <- as.numeric(coefhat)
# obtaining the TC data with same magnitude as determined by the tso() function
```

```
tc_effect <- coefhat*tc

# defining a time series for the temporary change data
outlier_effect <- ts(tc_effect, frequency = frequency(logtourist), start = start(logtourist))

# subtracting the transient change intervention, obtaining a time series without the transient change effect
logtourist_less_outlier <- logtourist - outlier_effect

# plot of the original, without intervention and transient change time series
plot(cbind(logtourist, logtourist_less_outlier, outlier_effect), main="")
title("Figure10: Effect of Earthquake on Tourist number")

# Estimate the coefficients
mod_outlier=arimax(logtourist,order=c(1,1,1),
            seasonal=list(order=c(0,1,1),period=12),
            xtransf=data.frame(EQ=1*(seq(logtourist)==281),
                        EQ=1*(seq(logtourist)==281)),transfer=list(c(0,0),c(1,0)),
            method='ML')
mod_outlier
coeftest(mod_outlier)
w0<-mod_outlier$coef[4]
w1<-mod_outlier$coef[6]
w2<-mod_outlier$coef[5]
w0
w1
w2

#except for EQ.1 MA0 all were significant
plot(logtourist)
points(fitted(mod_outlier), col='blue')
title("Figure11: Original Time Series against Fitted Time Series")

# plot intervention effects
EQ1p=1*(seq(logtourist)==281)
plot(ts(EQ1p*(w0) + filter(EQ1p,filter=w2,method='recursive', side=1)*
      (w1),frequency=12,start=1992),ylab='Earthquake Effects',
   type='h'); abline(h=0)
title("Figure12: Intervention Effects \n Impact of the 2015 Earthquake")

#Estimate intervention effects
Reduced = (1-exp(w0+w1))*100

lowered12 = (1-exp(w1*(w2^12)))*100

lowered24 = (1-exp(w1*(w2^24)))*100

lowered36 = (1-exp(w1*(w2^36)))*100
```

```
lowered42 = (1-exp(w1*(w2^42)))*100

#diagnostics
outlier_eff<- ts(EQ1p*(w0) + filter(EQ1p,filter=w2,method='recursive', side=1)*
       (w1), frequency = frequency(logtourist), start= start(logtourist))
wo_outlier <- logtourist - outlier_eff
SARIMA_out<-sarima(wo_outlier, p=1, d=1, q=1, P=0, D=1, Q=1, S=12)
SARIMA_out

#forecast
par(mfrow = c(1,1))
SARIMA_wo_p<-sarima.for(wo_outlier, 24,1,1,1,0,1,1,12, newxreg = (nobs+1): (nobs+24))
SARIMA_wo_for<-ts(SARIMA_wo_p$pred, frequency=12, start=c(2019,1))
SARIMA_wo_se<-ts(SARIMA_wo_p$se, frequency=12, start=c(2019,1))

par(cex.axis = 1, cex.lab = 1.2, cex.main = 1)
ts.plot(tourist, exp(SARIMA_wo_for),
     ylab = "Number of Tourist Arrivals",
     xlab = "Date",
     ylim = c(10000, 300000))
title("Figure14: Monthly Tourist Arrivals in Nepal \nJanuary 1992 through December 2018, with 2019 and 2020
Forecast", line=-2)
U3 = exp(SARIMA_wo_for+SARIMA_wo_se); L3 = exp(SARIMA_wo_for-SARIMA_wo_se)
xx3 = c(time(U3), rev(time(U3))); yy3 = c(L3, rev(U3))
polygon(xx3, yy3, border = 8, col = gray(.6, alpha = .2))
lines(exp(SARIMA_wo_for), type = "p", col = 2)
lines(exp(SARIMA_wo_for), type = "l", col = 2)

exp(SARIMA_wo_for)#24 period predictions

#9-period predictions
SARIMA_wo<-sarima.for(wo_outlier, 9,1,1,1,0,1,1,12, newxreg = (nobs+1):(nobs+9))
SARIMA_wo_SSE = sum((SARIMA_wo$pred - logtrue)^2)
SARIMA_wo_SSE

SARIMA_wo_pred<- ts(SARIMA_wo$pred, frequency=12, start=c(2019,1))
SARIMA_wo_se1 <-ts(SARIMA_wo$se, frequency=12, start=c(2019,1))

#plot 9-period predictions against true data
truets <- ts(true, start = c(2019,1), frequency=12)
ts.plot(cbind(truets, exp(SARIMA_wo_pred)), gpars = list(lty =c(1:3),col=1:2, ylab="Number of Tourist Arrivals",
                                    main="Figure15: Forecast against Real Tourist Arrivals for First 9 months of 2019"))
U4 = exp(SARIMA_wo_pred+1.96*SARIMA_wo_se1); L4 = exp(SARIMA_wo_pred-1.96*SARIMA_wo_se1)
xx4 = c(time(U4), rev(time(U4))); yy4= c(L4, rev(U4))
polygon(xx4, yy4, border = 8, col = gray(.6, alpha = .2))
lines(exp(SARIMA_wo_pred), type = "p", col = 2)
```

References

Data Sources:

1. Government of Nepal Ministry of Culture, Tourism and Civil Aviation. (2018). Nepal Tourism Statistics. Retrieved from: http://tourism.gov.np/statistic
2. CEIC, (2019). Retrieved from: https://www.ceicdata.com/en/indicator/nepal/visitor-arrivals

Project 1:

1. Shrestha, Anustha (2019). STA9701 Time Series Analysis on Tourist Arrivals in Nepal.

These are some of the materials that I referred to while working on this project

1. Li, Zeda (2019). STA9701 Class Lecture Notes
2. Shumway, R.H., Stoffer D.S. (2016), Time Series Analysis and Its Applications: With R Examples, 4th edition, Springer.
3. Vierse, Anne (2018). Time Series Project 2
4. Outlier Detection and Intervention. Datascience+. Accessed at: https://datascienceplus.com/outliers-detection-and-intervention-analysis/