Time Series Analysis and Forecasting for the Number of Tourists visiting Aruba using SARIMA models

Prepared for

STA4852

University of Central Florida

By

Elaine Ng

April 18, 2019

INTRODUCTION

Analyzing the data set for the monthly number of tourists visiting Aruba from January 1980 to December 2014, four different SARIMA models were estimated. Diagnostic tests were performed for each of the tentative models and then compared to each other, leading to the chosen model ARIMA(3,1,3)x$(1,1,0)_{12}$. With this model, a forecast for the number of tourists visiting Aruba has been derived for the following 24 months, from January 2015 to December 2016.

BACKGROUND

Aruba is a small island in the Caribbean with an estimated population of 106,000. According to the World Travel & Tourism Council, 98.3% of Aruba's total GDP is due to the contribution of tourism. With many of the tourists coming from the United Stated (57%), Venezuela (19%), Canada (4%), The Netherlands (3%), Colombia (2%) and 15% from other parts of the world. With tourism having such a big impact on Aruba's economy it is important to not only collect data related to tourism, but to also analyze that data. This would aid in the business and political decisions made within Aruba such as: signing contracts with U.S. advertising companies to promote tourism, launching island-wide tourism training programs, investing millions of dollars to renovate roads, cruise ship terminals, airports and every other sector related to tourism. Understanding the trend of tourism and forecasting for the upcoming months can help better prepare Aruba for the future.

DATA

The data set that will be used for this time series analysis and forecasting is provided by the Central Bureau of Statistics – Aruba. The data consists of monthly data on the number of tourists that has entered Aruba and stayed for at least one day starting January 1st, 1980 till December 31st, 2014.

Exhibit 1 shows the time series graph of the number of tourists visiting Aruba by month, from January 1980 to December 2014. There appears to be an upward linear trend for most of the years, except at the start of 1980 till the end of 1986, where the plot shows the number of tourists visiting Aruba is nearly constant throughout that time frame.

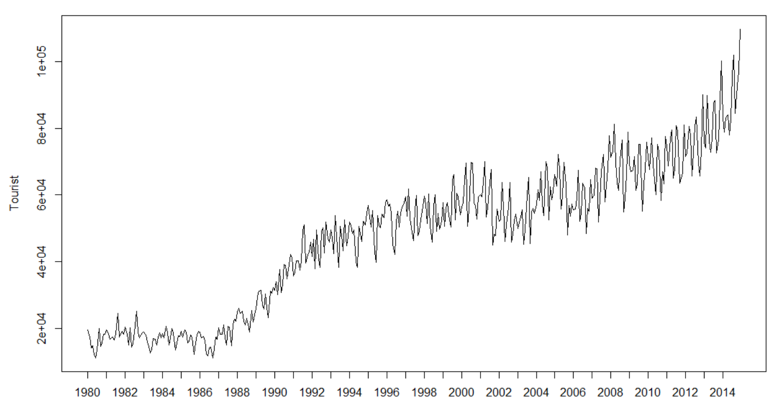

Exhibit 1: Number of Tourists visiting Aruba by month from 1980 to 2014

On January 1st, 1986 Aruba disaffiliated itself with the Netherlands Antilles and for the first time since 1499 Aruba became an autonomous country. The nation was then able to start making changes and decisions for themselves. They were able to start building businesses, sign contracts with airlines and large hotel chains without the need for The Netherlands or its' surrounding island's approval; hence, towards the end of 1986, Aruba's tourism has only gone up since. Considering Aruba before 1986 was under a completely different type of government, the data collected from January 1980 to December 1985 does not seem appropriate for modelling and forecasting the current state of Aruba. Exhibit 2 shows the new time series plot from January 1986 to December 2014. This will be the final data set used for modelling and forecasting.

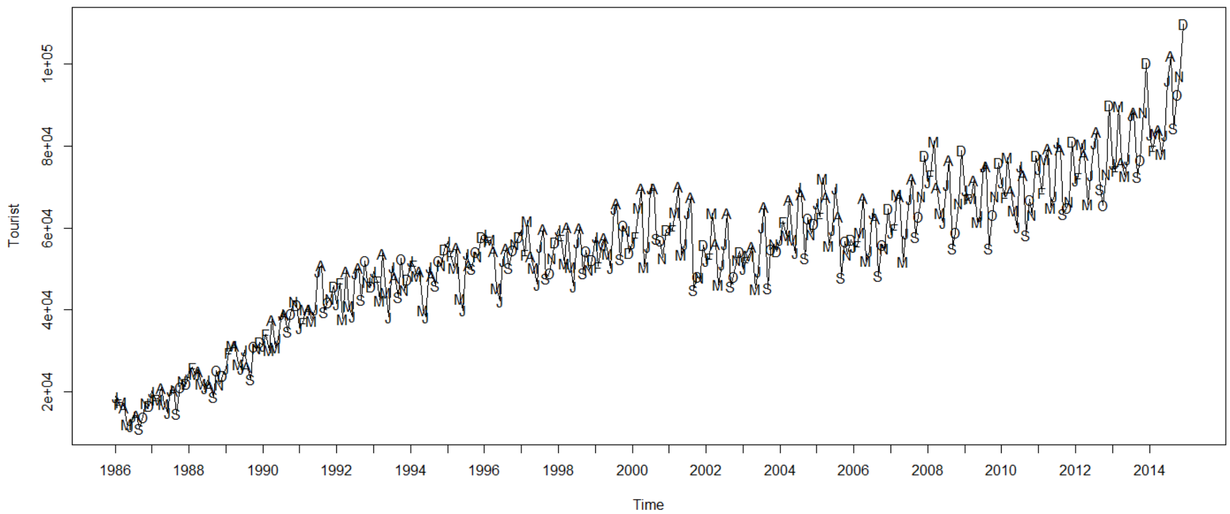

Exhibit 2: Number of Tourists visiting Aruba by month from 1986 to 2014

TREND AND SEASONALITY

As observed earlier, the time series plot for the number of tourists visiting Aruba seems to display an upward linear trend. Also noticeable is the increase of variation with time. A log transformation will be able to stabilize the variation and with a regular differencing it should be able to remove the upward linear trend. Exhibit 3 shows the time series plot after a log transformation has been applied followed by a regular differencing. It appears to be a lot more centered towards the 0 without any visible upward or downward trend, additionally the variation seems to be stable.

A quick glance at Exhibit 2 shows that the months August and December seem to have higher peaks; meanwhile, May and September seem to be the lower points. Most tourists visiting Aruba are from the United States. August being one of the busiest months of the year could be due to summer vacation for U.S. schools and December could be explained by the cold weather for large parts of the United States, leading to tourists traveling to warmer places and spend their holidays there. Conveniently, Aruba is summer all year long. As for the lower peaks of the year, May and September, there is simply no vacations or holidays during those months, especially September being the start of the school year for many U.S. schools.

Looking at its ACF on Exhibit 4, there is definitely a seasonal trend in this time series. That is made clear by the prominent spikes at lag 12, 24 and 36. Suggesting a seasonal differencing might be appropriate.

After taking a seasonal differencing, it is shown that the strong seasonal trend that was present has been reduced and if not, completely removed. The time series plot of the number of visitors to Aruba with a log transformation, a regular differencing and a seasonal differencing is displayed on Exhibit 5. It is slightly difficult to see the removal of the seasonality in the time series plot but with Exhibit 6, the ACF and Exhibit 7, the Partial ACF, it is much more apparent that the seasonal trend has been eliminated.

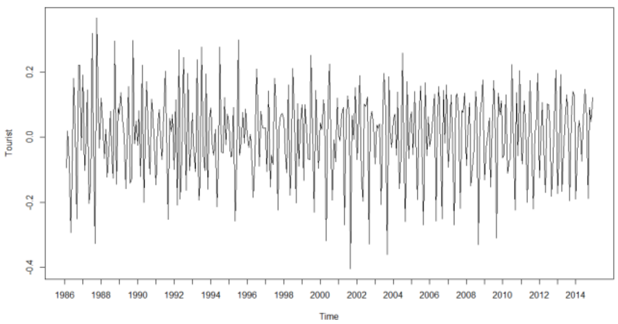

Exhibit 3: Time Series Plot of Log Transformed and First Differenced

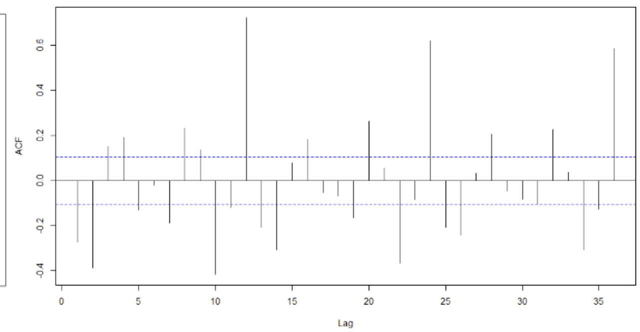

Exhibit 4: Sample ACF of Log Transformed and First Differenced of the Number of Tourists visiting Aruba

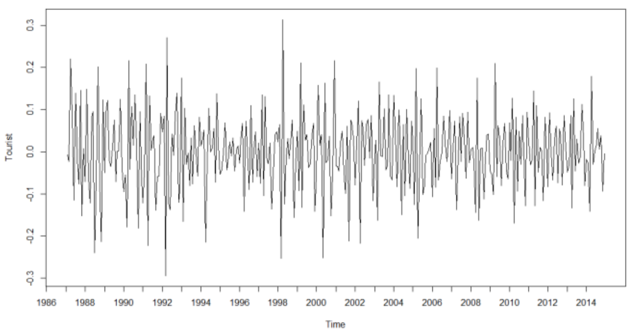

Exhibit 5: Time Series Plot of Log Transformed, First Differenced and Seasonal Differenced Time Series

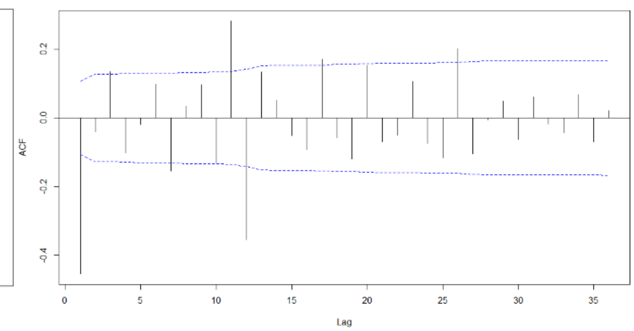

Exhibit 6: Sample ACF of Log Transformed, First Differenced and Seasonal Differenced of the Number of Tourists visiting Aruba

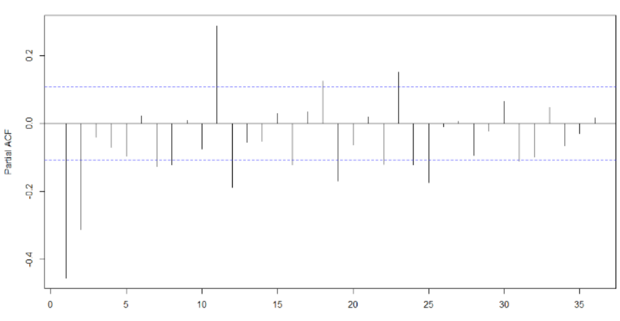

Exhibit 7: Sample Partial ACF of Log Transformed, First Differenced and Seasonal Differenced Time Series

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	o	x	o	o	o	x	o	o	x	x	x	x	o
1	x	x	x	o	o	o	x	o	o	o	o	x	x	x
2	x	x	x	o	o	o	o	o	o	o	o	x	x	o
3	x	o	x	o	o	o	o	o	o	o	o	x	o	o
4	x	o	x	x	x	o	o	o	o	o	o	x	o	o
5	x	x	x	x	x	o	o	o	o	o	o	x	x	o
6	x	x	x	x	o	o	o	o	o	o	o	x	x	o
7	x	x	x	x	x	x	o	x	o	o	o	x	o	o

Exhibit 8: EACF of the Log Transformed, First Differenced and Seasonal Differenced Time Series

SARIMA MODEL FITTING

With the help of Exhibit 6 and Exhibit 7, the ACF and Partial ACF, a tentative seasonal ARIMA model for this particular time series can be estimated. A cut off after a certain lag on the ACF suggests an MA, meanwhile a cut off on the Partial ACF indicates an AR model. Observing the ACF first, lag 1 autocorrelation and lag 3, 7, 11, 12, 17 and 26 are significant. But there appears to be a cut off at lag 1 which could mean an MA(1) model. The rest of them are barely significant with the exception of lag 11 and 12, this could indicate a Seasonal MA(1) or it can also be the effect of an unknown AR term.

The Partial ACF displays significance at lag 1, 2, 7, 8, 11, 12, 16, 18, 19, 22, 23, 24, 25 and 31. Though, many of them are narrowly significant, lag 1, 2, 11 and 12 are very noticeable. There appears to somewhat be a cut off at lag 2, indicating an AR(1) or AR(2), but seeing how they tail off a little this could be caused by an MA term. The spike at lag 11 and 12 could imply a seasonal AR(1).

Another tool that could be used for identifying the order of an ARIMA model is the Extended Autocorrelation Function, EACF. Exhibit 8 shows the EACF for this time series, though locating the theoretical pattern of a triangle of zeroes in this EACF proves to be difficult. The entire column of 11 has been crossed out. Leaving the only triangle of zeroes that can be formed, an ARMA(3,12) or ARMA(4,12). Since simplicity of a model is highly desired and favored, these two tentative models derived from the EACF does not seem very suitable due to their high orders.

The last tool utilized to find a fitting SARIMA model for this time series is the auto.arima() function in the package "forecast". This function will return the best SARIMA fit according to AIC or BIC values. The result given is an ARIMA(1,1,4)x$(2,1,2)_{12}$.

MODEL 1: ARIMA(2,1,0)x$(1,1,1)_{12}$ – FITTING AND DIAGNOSTIC CHECKING

After testing all possible combination between MA(1), Seasonal MA(1), AR(1), AR(2) and Seasonal AR(1), which were derived from the ACF and Partial ACF in Exhibit 6 and Exhibit 7. The Ljung-Box test proves to be a complication, which is set up as,

H_0 : Error terms are uncorrelated

H_a: Error terms are correlated

Test Statistics: $Q_* = n(n+2)(\frac{\hat{r}_1^2}{n-1} + \frac{\hat{r}_2^2}{n-2} + \cdots + \frac{\hat{r}_K^2}{n-K})$

Hence, a p-value > 0.05 is desired, that would indicate that there is not enough evidence to reject H_0, meaning the error terms are completely random. Unfortunately, none of the combination of these models lead to a p-value larger than 0.05 when using the Ljung-Box statistic. Out of all the combinations, the best one would be ARIMA(2,1,0)x$(1,1,1)_{12}$. A summary of this model is shown in Exhibit 9. All 4 coefficients are larger or equal to 2 standard error, proving their significance as calculated below.

ar1: $\frac{-0.5259}{0.0537} = -9.793$ | ar2: $\frac{-0.2692}{0.0530} = -5.079$ | sar1: $\frac{0.3412}{0.0857} = 3.981$ | sma1: $\frac{-0.8492}{0.0598} = -14.201$

Checking this model's diagnostic, displayed in Exhibit 10 is the Standardized Residuals, this appears to be random and the variation decreases with time. Followed by the ACF of the Residuals, there are 3 lags that are significant, specifically at lag 7, 11 and 19. This is quite a lot considering this being based on 95%, meaning it is acceptable for 1 out of 20 to be significant even if it's completely white noise. Though there are 3 significant lags, the lags do not appear to follow a certain pattern. Lastly, the p-values for lag 1 up till lag 6 for the Ljung-Box test is larger than 0.05 but, beginning at lag 7 it starts to decline to 0, meaning the error terms are correlated after lag 7.

```
Call:
arima(x = LogAruba, order = c(2, 1, 0), seasonal = list(order = c(1, 1, 1),
    period = 12))

Coefficients:
          ar1      ar2    sar1     sma1
      -0.5259  -0.2692  0.3412  -0.8492
s.e.   0.0537   0.0530  0.0857   0.0598

sigma^2 estimated as 0.005175:  log likelihood = 401.9,  aic = -795.8
```

Exhibit 9: Maximum Likelihood Estimates and their Standard Errors for ARIMA(2,1,0)x(1,1,1)$_{12}$

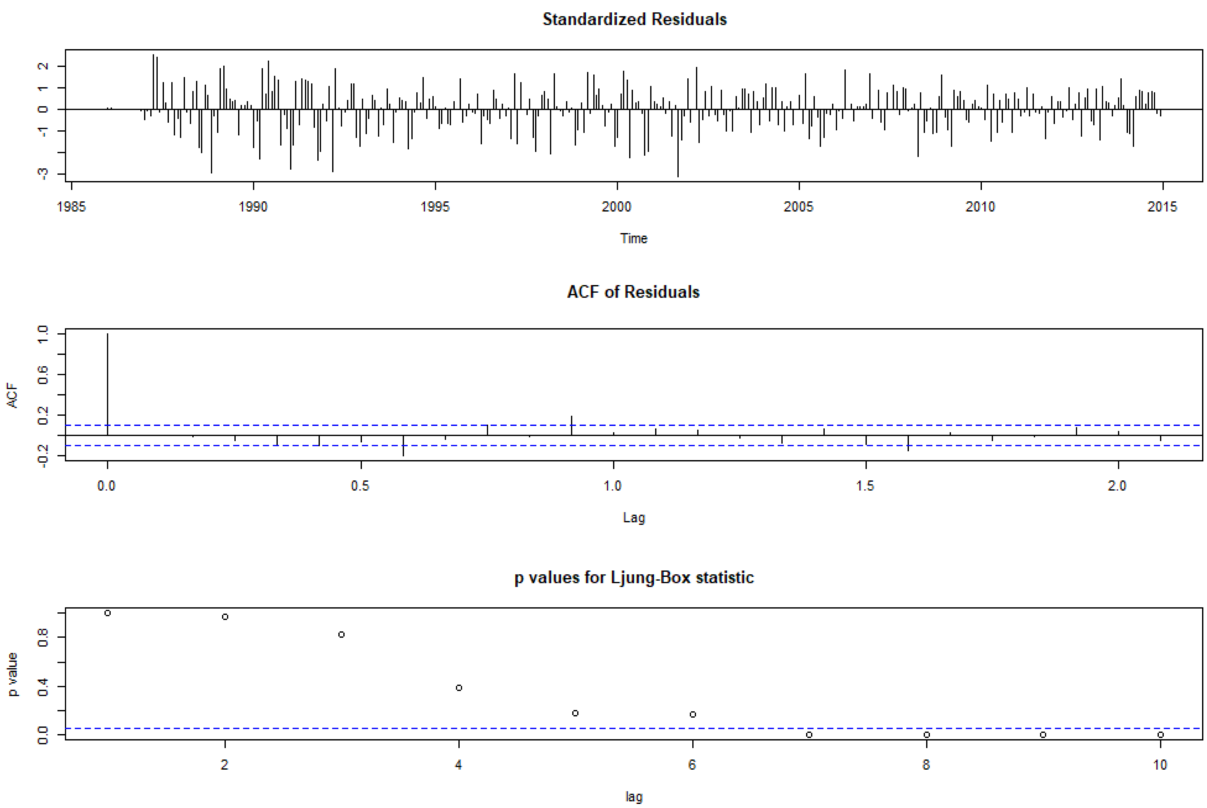

Exhibit 10: Diagnostic Test of ARIMA(2,1,0)x(1,1,1)$_{12}$

MODEL 2: ARIMA(1,1,4)x(2,1,2)$_{12}$ and ARIMA(1,1,4)x(1,1,1)$_{12}$ – FITTING AND DIAGNOSTIC CHECKING

As suggested by the auto.arima() function based on the best AIC and BIC, an appropriate SARIMA model would be ARIMA(1,1,4)x(2,1,2)$_{12}$. Exhibit 11 displays a summary for this model and as shown below, not all coefficients are significant or larger than 2 standard error. For example, ma3 is not significant while ma4 is, and sar1, sar2, sma1 and sma2 are all insignificant as well.

ar1: $\frac{0.5075}{0.1746} = 2.907$	ma1: $\frac{-1.0687}{0.1779} = -6.007$	ma2: $\frac{0.2726}{0.1300} = 2.097$
ma3: $\frac{0.1057}{0.0838} = 1.261$	ma4: $\frac{-0.1687}{0.0563} = -2.996$	sar1: $\frac{-0.3095}{0.3084} = -1.004$
sar2: $\frac{0.1200}{0.1392} = 0.862$	sma1: $\frac{-0.1971}{0.3011} = -0.655$	sma2: $\frac{-0.4669}{0.2517} = -1.855$

```
Call:
arima(x = LogAruba, order = c(1, 1, 4), seasonal = list(order = c(2, 1, 2),
    period = 12))

Coefficients:
         ar1      ma1     ma2     ma3      ma4     sar1    sar2     sma1     sma2
      0.5075  -1.0687  0.2726  0.1057  -0.1687  -0.3095  0.1200  -0.1971  -0.4669
s.e.  0.1746   0.1779  0.1300  0.0838   0.0563   0.3084  0.1392   0.3011   0.2517

sigma^2 estimated as 0.005014:  log likelihood = 407.64,  aic = -797.27
```

Exhibit 11: Maximum Likelihood Estimates and their Standard Errors for ARIMA(1,1,4)x(2,1,2)$_{12}$

Since some of the coefficients are not significant perhaps removing non-significant coefficients, the seasonal AR and MA components and forcing ma3's coefficient to be zero, can lead to a better model. However, after removing the seasonal terms it appears that the Standardized Residual is not white noise but exhibits a wave pattern and the ACF of the Residuals show obvious lag spikes at 12, 24 and 36. All pointing towards the idea that there should be a seasonal term. Testing ARIMA(1,1,4)x(0,1,1)$_{12}$, ARIMA(1,1,4)x(1,1,0)$_{12}$ and ARIMA(1,1,4)x(1,1,1)$_{12}$, Exhibit 12 is the adjusted SARIMA model. This new model ARIMA(1,1,4)x(1,1,1)$_{12}$ has less parameters due to removing some the non-significant parameters. With the exception of ma3, all other coefficients are larger than twice their standard errors.

ar1: $\frac{0.5113}{0.1767} = 2.894$	ma1: $\frac{-1.07060}{0.1801} = -5.944$	ma2: $\frac{0.2673}{0.1304} = 2.050$	ma3: $\frac{0.1171}{0.0828} = -1.414$
ma4: $\frac{-0.1721}{0.0559} = -3.079$	sar1: $\frac{0.2978}{0.0932} = 3.195$	sma1: $\frac{-0.8133}{0.0692} =$ -11.753	

```
Call:
arima(x = LogAruba, order = c(1, 1, 4), seasonal = list(order = c(1, 1, 1),
    period = 12))

Coefficients:
         ar1      ma1     ma2     ma3      ma4    sar1     sma1
      0.5113  -1.0706  0.2673  0.1171  -0.1721  0.2978  -0.8133
s.e.  0.1767   0.1801  0.1304  0.0828   0.0559  0.0932   0.0692

sigma^2 estimated as 0.00502:  log likelihood = 407.34,  aic = -800.68
```

Exhibit 12: Maximum Likelihood Estimates and their Standard Errors for ARIMA(1,1,4)x(1,1,1)$_{12}$

The ARIMA(1,1,4)x(1,1,1)$_{12}$ diagnostics are in Exhibit 13. The Standardized Residuals appears to be random and the variation is decreasing over time. There is a total of 4 lags that are significant within the tested 36 lags, which are lag 7, 11, 19 and 26. There does not seem to be a specific pattern for the ACF of the residuals. Furthermore, the p-values for the Ljung-Box test is 0.1172 which is larger than 0.05. Proving the errors terms are uncorrelated.

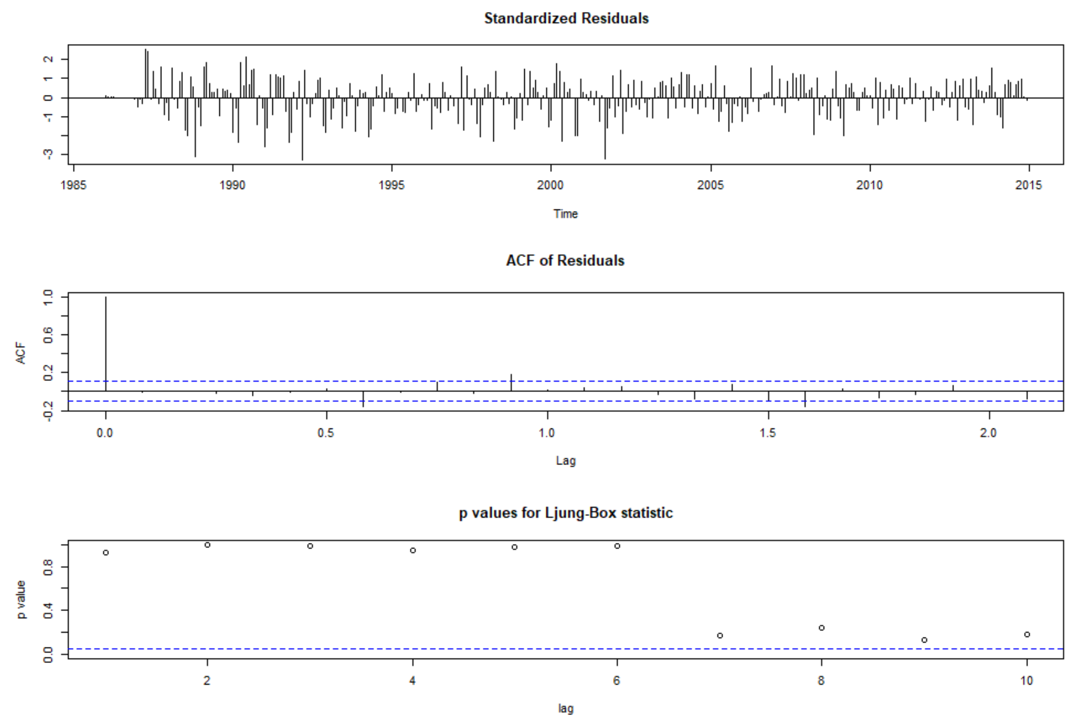

Exhibit 13: Diagnostic Test of ARIMA(1,1,4)x(1,1,1)$_{12}$

MODEL 3: ARIMA(3,1,3)x(1,1,0)$_{12}$ – FITTING AND DIAGNOSTIC CHECKING

Another way to acquire a tentative SARIMA model is by trial and error, without going into very high ordered models which could lead to a better fit but also more complicated. The ARIMA(3,1,3)x(1,1,0)$_{12}$ is one of the models that seems to be appropriate. The model summary is presented in Exhibit 14 and checking whether the coefficients are significant or not below, all besides ma2 have a coefficient more than 2 times their standard error.

ar1: $\frac{-0.8197}{0.1180} = -6.947$	ar2: $\frac{-0.6126}{0.1362} = -4.498$	ar3: $\frac{0.3310}{0.117} = 2.829$	sar1: $\frac{-0.4009}{0.0532} = -7.536$
ma1: $\frac{0.3706}{0.0901} = 4.113$	ma2: $\frac{0.1586}{0.0990} = 1.602$	ma3: $\frac{-0.7505}{0.0888} = -8.452$	

```
Call:
arima(x = LogAruba, order = c(3, 1, 3), seasonal = list(order = c(1, 1, 0)))

Coefficients:
          ar1      ar2     ar3     ma1     ma2      ma3     sar1
      -0.8197  -0.6126  0.3310  0.3706  0.1586  -0.7505  -0.4009
s.e.   0.1180   0.1362  0.1177  0.0901  0.0990   0.0888   0.0532

sigma^2 estimated as 0.004944:  log likelihood = 409.18,  aic = -804.36
```

Exhibit 14: Maximum Likelihood Estimates and their Standard Errors for ARIMA(3,1,3)x(1,1,0)$_{12}$

The diagnostic test for this model is shown in Exhibit 15. The Standardized Residuals seem random and the variation decreases. The ACF of the Residuals have two significant lags, at lag 13 and 24. It does not seem to follow a pattern and 2 significands out of 36 does not seem to be too out of the ordinary considering with white noise there can still be 1 out of 20 that is significant. Additionally, the p-value for the Ljung-Box test is larger than 0.05, indicating that the error terms are uncorrelated.

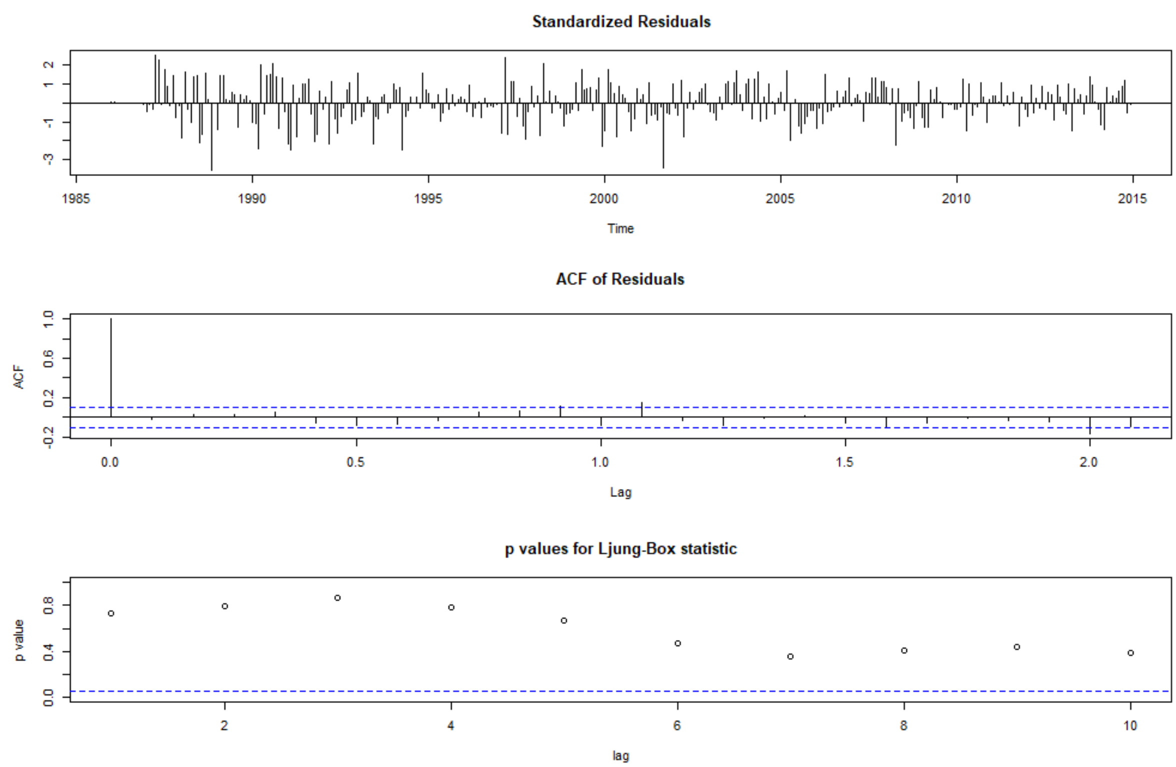

Exhibit 15: Diagnostic test for ARIMA(3,1,3)x(1,1,0)$_{12}$

MODEL COMPARISONS

Below is a table with the 4 tentative models and their respected Log-Likelihood, AIC and BIC.

SARIMA MODEL	LOG LIKELIHOOD	AIC	BIC
ARIMA(2,1,0)X(1,1,1)$_{12}$	401.9	-795.8	-774.7301
ARIMA(1,1,4)X(2,1,2)$_{12}$	407.64	-797.27	-757.1293
ARIMA(1,1,4)X(1,1,1)$_{12}$	407.34	-800.68	-768.1632
ARIMA(3,1,3)X(1,1,0)$_{12}$	409.18	-804.36	-771.8465

Table 1: SARIMA Model Comparison

Besides looking at the diagnostics of each tentative model, a side-by-side comparisons of these 4 different ordered SARIMA model based on their Log-Likelihood, AIC, and BIC can also prove to be helpful in deciding which one to ultimately use for forecasting. The highest log likelihood in the 4 tentative model is ARIMA(3,1,3)x(1,1,0)$_{12}$ and the lowest is ARIMA(2,1,0)x(1,1,1)$_{12}$. Followed by the highest AIC, ARIMA(2,1,0)x(1,1,1)$_{12}$ and the lowest, ARIMA(3,1,3)x(1,1,0)$_{12}$. Lastly the model with the highest BIC is ARIMA(1,1,4)x(2,1,2)$_{12}$ and the lowest BIC, ARIMA(2,1,0)x(1,1,1)$_{12}$. The general idea is the higher the log likelihood the better, and since AIC and BIC are penalized likelihood criteria the lower the better. Hence, with the previous diagnostic checks and these 3 criterions as guidelines, the best model would be ARIMA(3,1,3)x(1,1,0)$_{12}$, with the highest log likelihood, lowest AIC, second lowest BIC, a random standardized residual, only 2 significands out of 36 lags in the ACF of the residuals, and a p-value larger than 0.05 for the Ljung-Box statistics.

COMPARING 2014 FORECASTS WITH ACTUAL VALUES

Removing the data from January 2014 to December 2014 and using the chosen SARIMA model ARIMA(3,1,3)x(1,1,0)$_{12}$, Exhibit 16 shows the forecasted values for January 2014 to December 2014.

Month	Tourists
January	82887
February	78979
March	83131
April	84079
May	78087
June	82708
July	95952
August	102048
September	84509
October	92649
November	97199
December	109854

Table 2: Monthly Tourists 2014

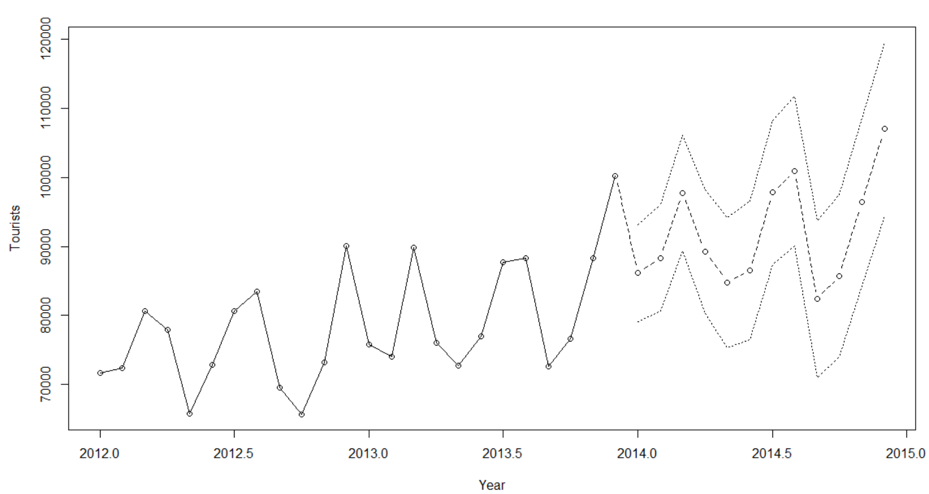

Exhibit 16: Forecast for the number of Tourists visiting Aruba 2014

Comparing the forecasted values in Exhibit 16 with the actual values in Table 2, all the values fall within the 95% confidence band, proving that this model is effective in forecasting for 2014.

FORECASTING

Exhibit 17 shows the forecast and 95% forecast limits for the upcoming 24 months for the ARIMA(3,1,3)x$(1,1,0)_{12}$. As shown before in Exhibit 1, there seems to be an upward trend with seasonality, particularly higher peaks for the months August and December and lower peaks during May and September, both trends of which are visible in our forecast as well.

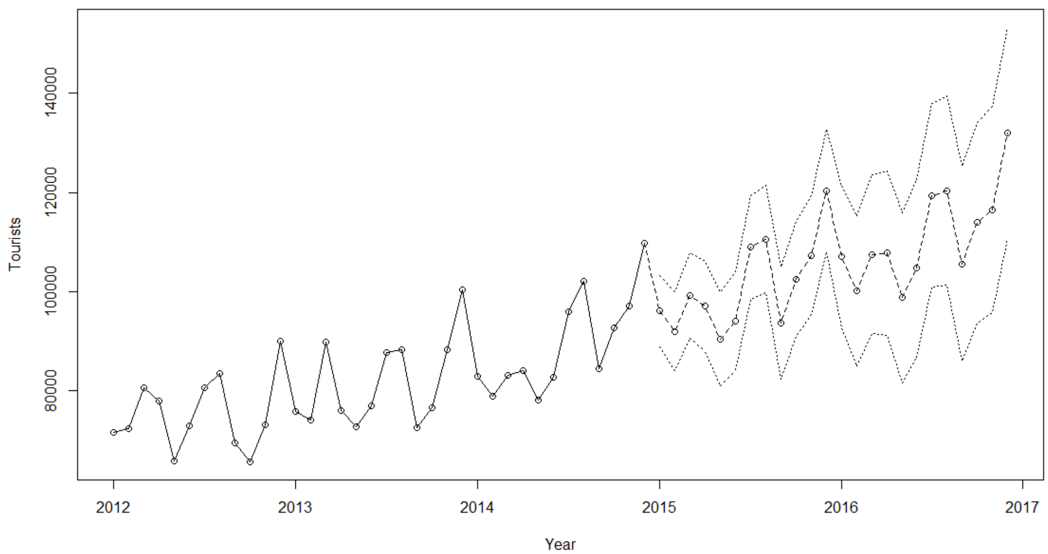

Exhibit 17: Forecast for the number of Tourists visiting Aruba 2015 and 2016

But since the second largest group of visitors to Aruba come from Venezuela and Nicolas Maduro became president of Venezuela in 2013 (which was the start of the crisis in Venezuela), we should expect the actual values to be closer to the lower bound of the forecast, due to less Venezuelans having the opportunity to go on vacation.

CONCLUSION

After applying a log transformation, detrending and removing the seasonality of the original time series of the number of tourists visiting Aruba from January 1980 to December 2014, four different SARIMA models were estimated based on the ACF, Partial ACF, auto.arima() function and by trial and error. Checking their diagnostic tests and comparing their log-likelihood, AIC and BIC, the best model was chosen to be ARIMA(3,1,3)x$(1,1,0)_{12}$. This model was tested by forecasting the removed 2014 data and comparing it with the actual values. Afterwards, a forecast was produced for 2015 and 2016, showing continued growth in the number of tourists visiting Aruba, as well as August and December being the busiest month of the year.

References

Central Bureau of Statistics – Aruba. “Da.1.2. Number of Stayover Visitors”. *Central Bureau of Statistics Aruba*, January 22, 2016, https://cbs.aw/wp/index.php/2016/01/22/da-1-aruba-tourism-authority-2/. Accessed March 31, 2019.

Cryer, J. D. and Chan, K. (2008). *Time Series Analysis with Applications in R,* 2^{nd} ed. New York: Springer.

Encyclopedia Britannica. “Aruba Island, Caribbean Sea”. *Encyclopedia Britannica Inc*., Feb 19, 2019, https://www.britannica.com/place/Aruba. Accessed April 7, 2019.

Hotel Online Special Report. “Aruba Forging Ahead with a $230 million Investment in Tourism — the Island’s Largest Expansion and Renovation Campaign in Years;”. *Eileen Almonte Quinn & Co.,* Mar 2, 2006, https://www.hotelonline.com/News/PR2006_1st/Mar06_ArubaInvestment.html. Accessed April 10, 2019.

World Travel & Tourism Council. “Aruba 2019 Annual Research: Key Highlights”. *World Travel & Tourism Council*, N/A, https://www.wttc.org/economic-impact/country-analysis/countr-data. Accessed April 7, 2019.