SARIMA Modeling and Spectral Analysis of Monthly Airline Traffic in the United States

Yeison Garciaguirre (ygarciaguirre@gmail.com)

June 13, 2024

Abstract

This report focuses on finding a time series model for the monthly total passengers in U.S. airlines between January 2009 and December 2018 and using the fitted model for forecasting the values for the next 12 months. Due to the patterns of the data, different SARIMA models were investigated as potential candidates by evaluating the significance of their coefficients, their residuals, and their AIC, AICc, and BIC values. The model that best fitted the data was a SARIMA$(0,1,1)(0,1,1)_{12}$ process, which was able to forecast the next 12 data points with reasonable accuracy. Finally, spectral analysis was used to identify annual cycles for the data and some other obscure cycles within each year.

1. Introduction

In the airline industry, predicting future demand is an important task to better accommodate their customers and maximize their profitability. In order to do so, airlines need to gather and analyze data about their customers and travel frequency so that they can make accurate forecasts when checking for capacity and setting prices. However, making precise predictions can be challenging since they also need to consider factors, such as seasonality and weather conditions, that might impact their examinations. One way in which airlines can improve their predictions and overcome these challenges is by using time series analysis, which uses historical data to identify patterns, such as trends and seasonality, to fit a time series model and make accurate predictions about future demand.

Due to the importance of forecasting in the airline industry, using time series analysis, this study aims to fit a Seasonal AutoRegressive Integrated Moving Average (SARIMA) model to the U.S. airline traffic data from January 2009 to December 2018 and predict future values for the data that can help make informed decisions. The SARIMA model is suited for the data because it uses both seasonal and non-seasonal pattens to find the best model for the data. Additionally, this study uses spectral analysis, which provides an examination of the frequency domain, to identify cycles in the data and their periodic behavior that might not be explicit in the time domain.

2. Data

Data for the total number of passengers (domestic and international) in U.S. airlines from January 2003 to September 2023 were obtained from Kaggle, and the publisher of this data set retrieved it

from the U.S. Bureau of Transportation Statistics. The website for this U.S. agency was checked, and it was confirmed that the data came from the U.S. Government. The data contains monthly air traffic data measured in millions, and all observations are non-negative. The sample size is of 249 observations, but due to the COVID-19 pandemic and its impact on the whole world and, thus, on the airline industry, the data was truncated to only include the values from January 2009 to December 2018 (training data), reducing the sample size to 120. The observations for 2019 were saved for testing purposes, i.e., to check the forecasting ability of the fitted model.

3. Methodologies

Because the airline industry experiences higher demand during peak seasons (holidays) compared to other times of the year, it is appropriate to fit a SARIMA to the U.S. airline traffic series and use spectral anlysis to identify cyclic and periodic patterns.

3.1 SARIMA Modeling

According to Robert H. Shumway, a multiplicative Seasonal AutoRegressive Integrated Moving Average (SARIMA) model is given by

$$\Phi_p(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \delta + \Theta_Q(B^s)\theta(B)w_t,$$

where w_t is a Gaussian white noise process, the AR and MA components are represented by $\phi_p(B)$ and $\Phi_q(B)$, respectively, and the SAR and SMA components by $\Phi_P(B^s)$ and $\Theta_Q(B^s)$, and the ordinary and seasonal difference components by ∇^d and ∇_s^D.

When fitting a SARIMA model to the U.S. airline traffic series, the following procedure, known as the Box-Jenkins approach, is used:

1. Exploratory data analysis: The time series is plotted and patterns are identified. If the variance of the data seems to be increasing over time, a Box-Cox transformation, given by

$$y_t = \begin{cases} \frac{x_t^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log x_t, & \lambda = 0, \end{cases}$$

might be used to stablize the data. If the data shows a trend over time, we difference the data at lag 1 until we have removed the trend. If the data shows patterns of seasonality, we difference the data until we have removed the trend. The goal is to make the data stationary.

2. Identification of parameters: When deciding the parameters for our SARIMA model, we examine the ACF and PACF plots of the (stationary) transformed data, and the parameters are choosen following *Figure 1*:

3. Estimation of parameters: Once we have chosen the parameter for our possible model(s), we test for the significance of the coefficients, and we compare the AIC, AICc, and BIC values if more we are examining more than one model. We choose the model that has significant coeffients and lowest AIC, AICc, and BIC values.

4. Model diagnostics: We check the residuals from the chosen model by plotting the time plot, ACF, Q-Q plot, and p-values of the residuals and look for patterns of white noise.

5. Model choice: If the residuals look like white noise, we conclude that the model fits the data, and we use the model to calculate forecasts.

	$\text{AR}(P)_s$	$\text{MA}(Q)_s$	$\text{ARMA}(P,Q)_s$
ACF*	Tails off at lags ks, $k = 1, 2, \ldots,$	Cuts off after lag Qs	Tails off at lags ks
PACF*	Cuts off after lag Ps	Tails off at lags ks $k = 1, 2, \ldots,$	Tails off at lags ks

*The values at nonseasonal lags $h \neq ks$, for $k = 1, 2, \ldots,$ are zero

Figure 1: Behavior of the ACF and PACF for SARMA models

3.2 Spectral Analysis

Spectral analysis examines the data by converting the time domain into frequency domain, identifying patterns of cycles and periods in the data. This analysis decomposes the time series into components that ressemble sinusoidal and non-sinusoidal waves and use it to detect periodicities, trends, and noise in the data. One way to do this is by computing the periodogram of the time series, which is a biased estimate of the spectral density function.

4. Results

4.1 SARIMA Modeling and Box-Jenkins Approach

Data Exploration

The plot of the U.S. airline traffic series is given in *Figure 2*.

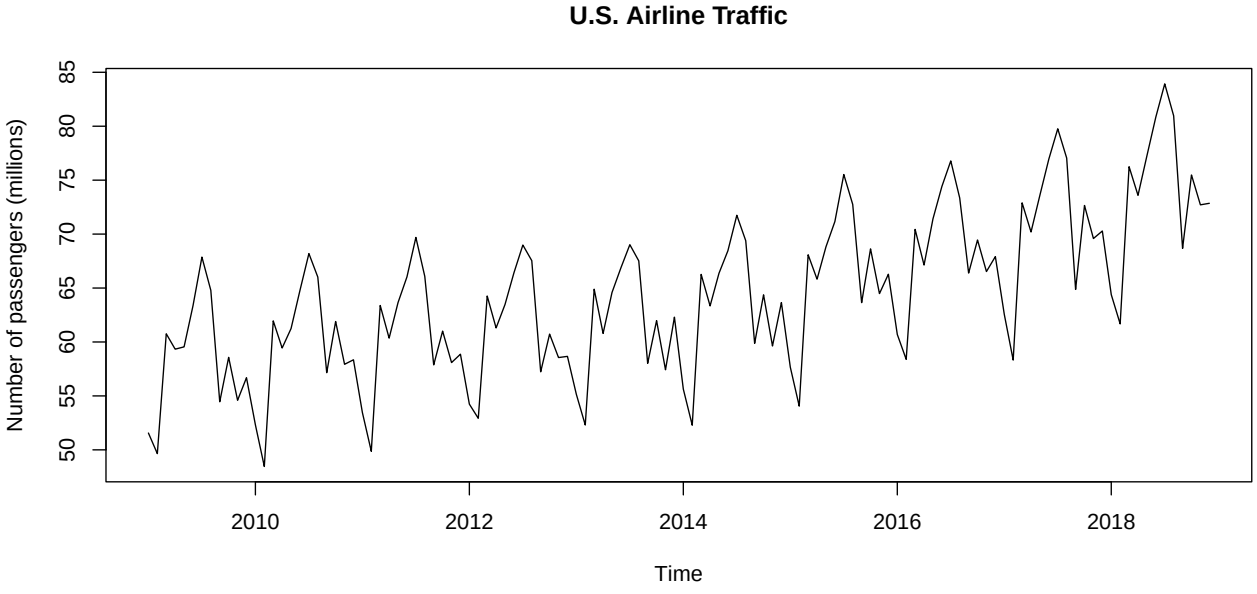

Figure 2: Monthly total of passengers in U.S. airlines (Jan 2009 – Dec 2018)

From *Figure 2*, we can see that there is an increasing trend in U.S. airline traffic over time, and there is also a pattern of seasonality. The variance of the data seems to be increasing slightly over time, and this is more apparent in *Figure 3*, which displays the different components of the series.

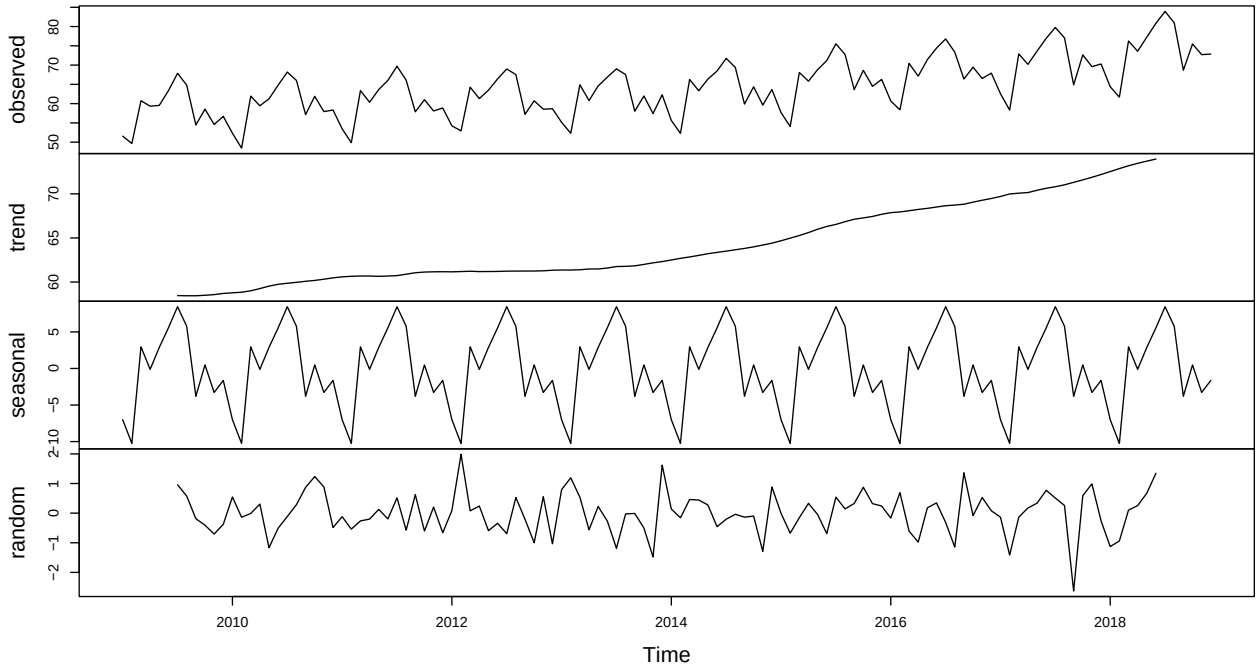

Figure 3: U.S. airline traffic series and its components

In order to assess whether or not a transformation of the series is necessary to stabilize its variance, we solve for the optimal λ, which is found to be approximately 0.87 (*Figure 4*). So, we apply a Box-Cox transformation to the series with this value of λ. However, from *Figure 5*, we can see that the increasing trend and seasonality patterns still persist in the Box-Cox transformed data.

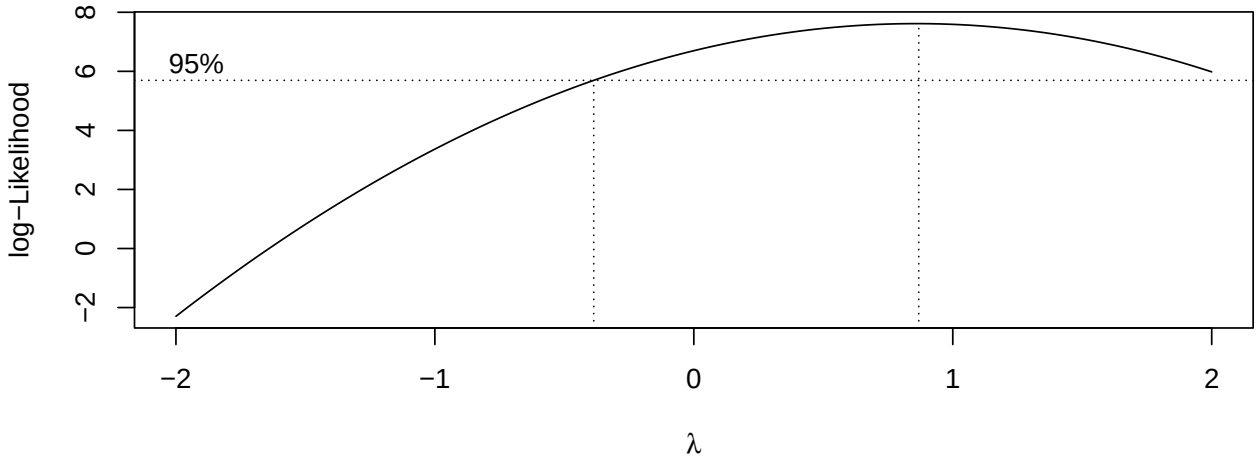

Figure 4: Box-Cox estimate of optimal λ

Thus, we difference the Box-Cox transformed data at lag 1 first to remove the trend, and we difference it again at lag 12 to remove the monthly seasonality of the series. (That is, we perform $y_t = \nabla_{12}\nabla x_t = (1 - B^{12})(1 - B)x_t$.) Clearly, from *Figure 6* we can see that the first-order difference gets rid of the trend, and the twelfth-order difference removes the monthly seasonality. In fact, these transformations further reduce the variance of the data as we can see in *Table 1*.

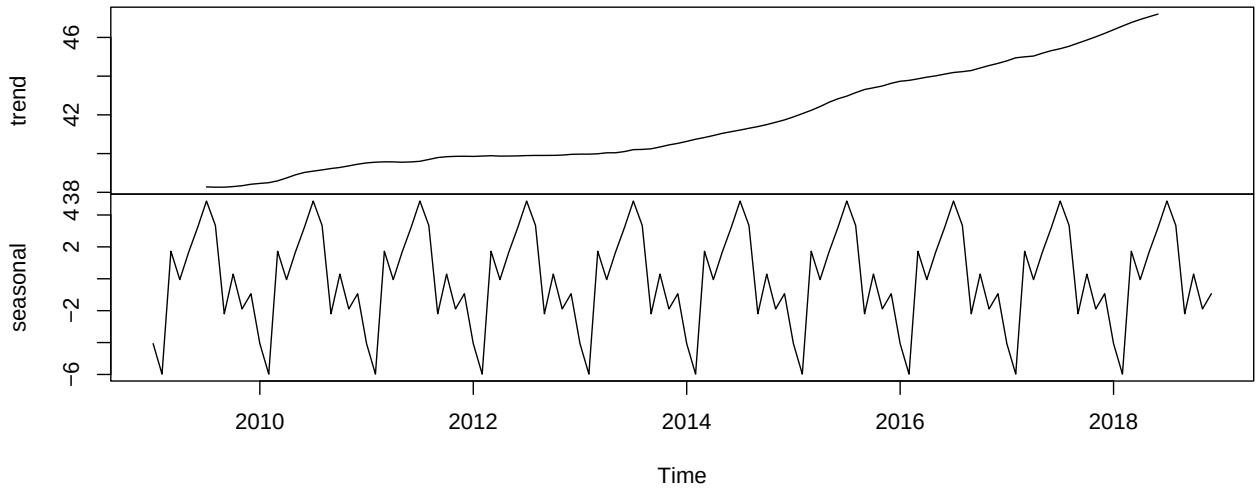

Figure 5: Trend and seasonal plots of the Box-Cox transformed data

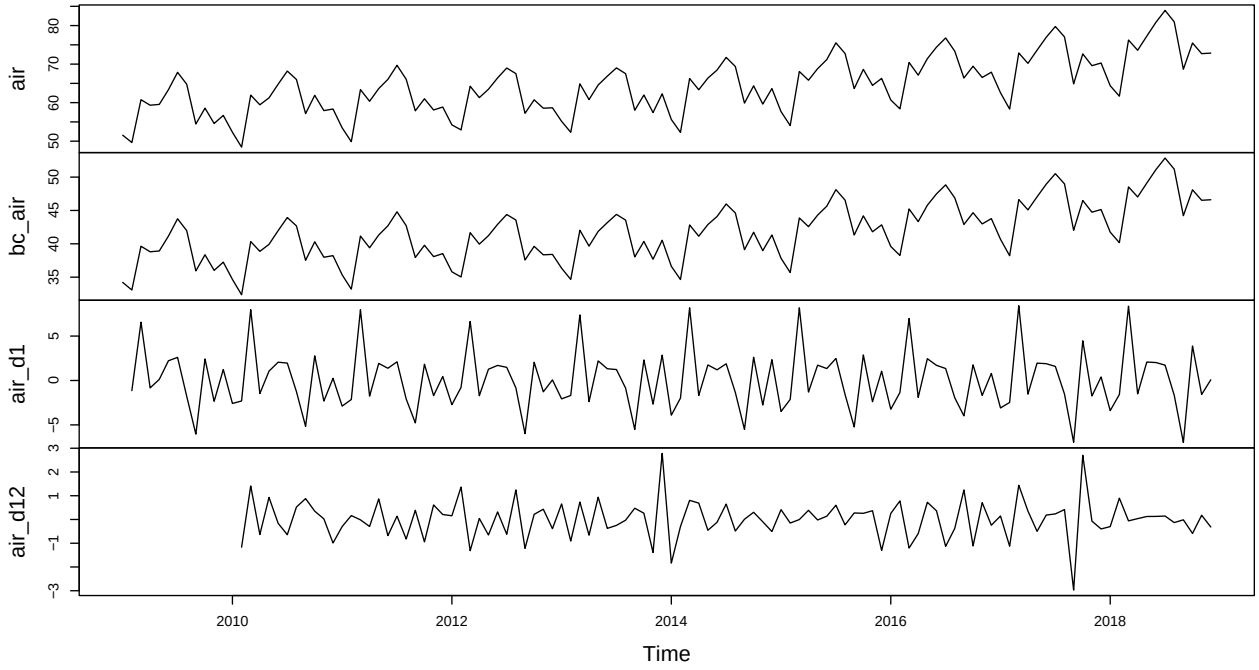

Figure 6: Time plots of the U.S. airline traffic series and its transformations

Table 1: The variances of the U.S. airline traffic series after each transformation

Transformation	Variance
Original	54.669
Box-Cox	18.2872
1st Difference	11.2012
12th Difference	0.6629

The data plot of the twice-differenced data in *Figure 7* shows (weakly) stationarity since it has a constant mean and a stable variance, despite the outliers around 2014 and late 2017. Also, the ACF plot shows that the variance is more stable, and the histogram suggests that the twice-differenced data is normally distributed.

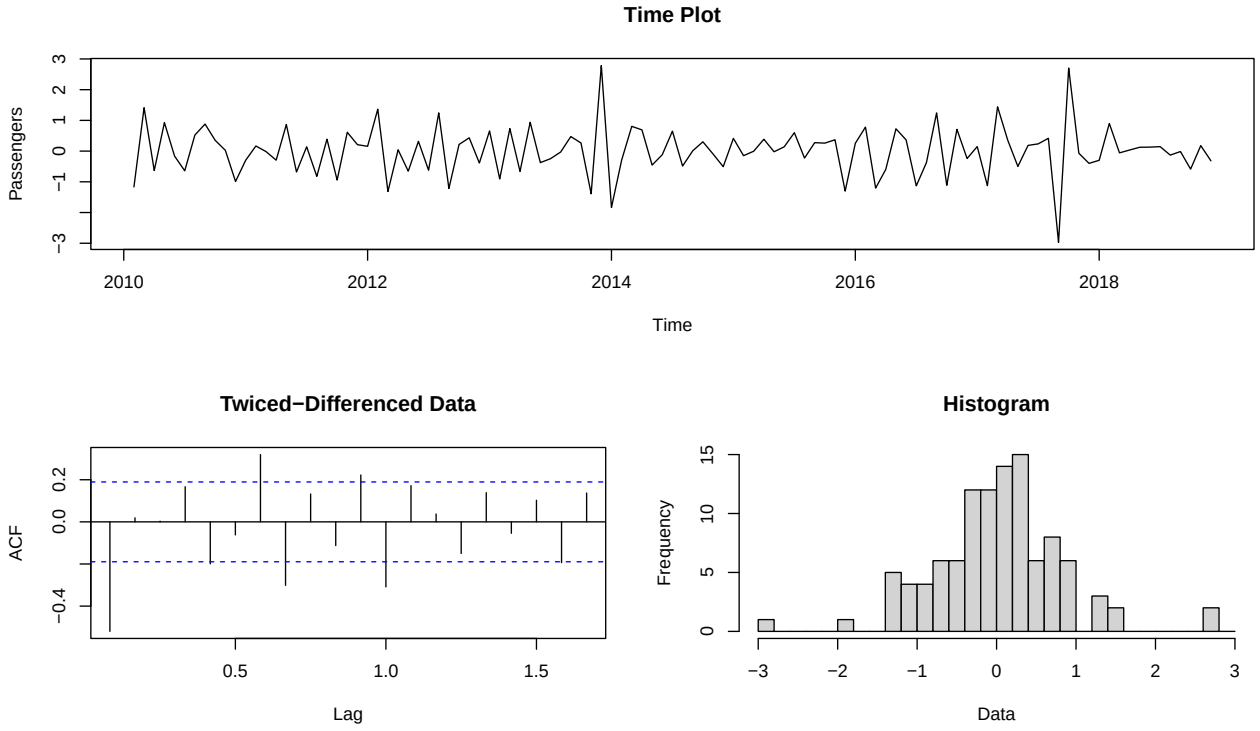

Figure 7: Various plots for the twice-differenced data

Identification of Parameters

The original U.S. airline traffic series shows patterns of an upward trend and seasonality, so we must fit a SARIMA$(p,d,q)(P,D,Q)_s$ model to the data with $s = 12$ because we are working with monthly data. Since we differenced the Box-Cox transformed data at lag 1 to remove the trend and then at lag 12 to get rid of the seasonality, we conclude that $d = 1$ and $D = 1$. It remains to identify the seasonal and non-seasonal parameters for the model.

From *Figure 8*, we notice that the ACF plot cuts off after lag $s = 1$, indicating an SMA component with $Q = 1$. Similarly, we observe that the PACF plot has no significant spikes, implying that our model does not have any SRA component, and thus, $P = 0$. Looking at both plots for the

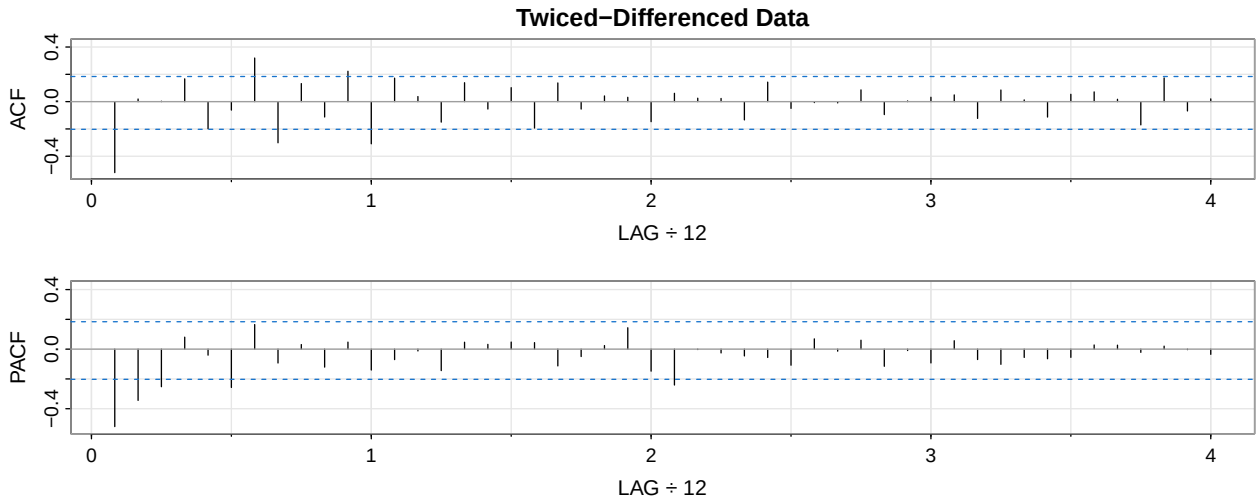

Figure 8: Sample ACF and PACF of the twiced-differenced data. Lag is in terms of years

non-seasonal values, we can see that their patterns are not indicative of a specific model. For instance, we can argue that the ACF cuts off after lag $h = 1$ (the first unnumbered spike) and the PACF is tailing off at lags $h = 1, 2, 3, \ldots$, suggesting that there is an MA component with $q = 1$ but no AR component (so $p = 0$). However, we can also argue that the ACF is tailing off at lags $h = 1, 4, 5, \ldots$ and the PACF cuts off after lag $h = 3$, indicating that there is an AR component with $p = 3$ but no MA component (so $q = 0$).

Hence, we consider both models (a SARIMA$(0, 1, 1)(0, 1, 1)_{12}$ and a SARIMA$(3, 1, 0)(0, 1, 1)_{12}$ model) for comparison, along with similar models but with variations in their parameters. Moreover, the `auto.arima` function points out that a SARIMA$(0, 1, 1)(1, 1, 2)_{12}$ model is a good candidate to fit the data, and so we should also consider it.

Estimation of Parameters

Table 2: Comparison of estimated coefficients based on the MLE method

Model	Significance of Coefficients	AIC	AICc	BIC
SARIMA$(0,1,1)(0,1,1)_{12}$	All coefficients are significant	1.8448	1.8459	1.9198
SARIMA$(3,1,0)(0,1,1)_{12}$	All coefficients are significant	1.8626	1.8663	1.9875
SARIMA$(0,1,1)(1,1,2)_{12}$	SAR1 and SMA1 are not significant	1.8221	1.8258	1.947
SARIMA$(1,1,1)(0,1,1)_{12}$	AR1 is not significant	1.8605	1.8626	1.9604
SARIMA$(0,1,2)(0,1,1)_{12}$	MA2 is not significant	1.8595	1.8617	1.9595
SARIMA$(3,1,1)(0,1,1)_{12}$	AR3 and MA1 are not significant	1.881	1.8866	2.0309
SARIMA$(4,1,0)(0,1,1)_{12}$	AR3 and AR4 are not significant	1.8811	1.8866	2.031
SARIMA$(2,1,0)(0,1,1)_{12}$	All components are significant	1.9038	1.906	2.0037

From *Table 2*, we can see that the first, second, and last models have all significant components based on the p-values from the z-test of their coefficients. Of these three models, the SARIMA$(0, 1, 1)(0, 1, 1)_{12}$ model (Model 1) is the best performing model since it has the lowest

AIC, AICc, and BIC values, followed by the SARIMA$(3,1,0)(0,1,1)_{12}$ model (Model 2). The SARIMA$(2,1,0)(0,1,1)_{12}$ model (Model 3) is the worst performing model out of these three. Consequently, we consider Model 1 as our best candidate for the next step and reserve Model 2 as an alternate in case that Model 1 fails the diagnostics check.

Model Diagnostics

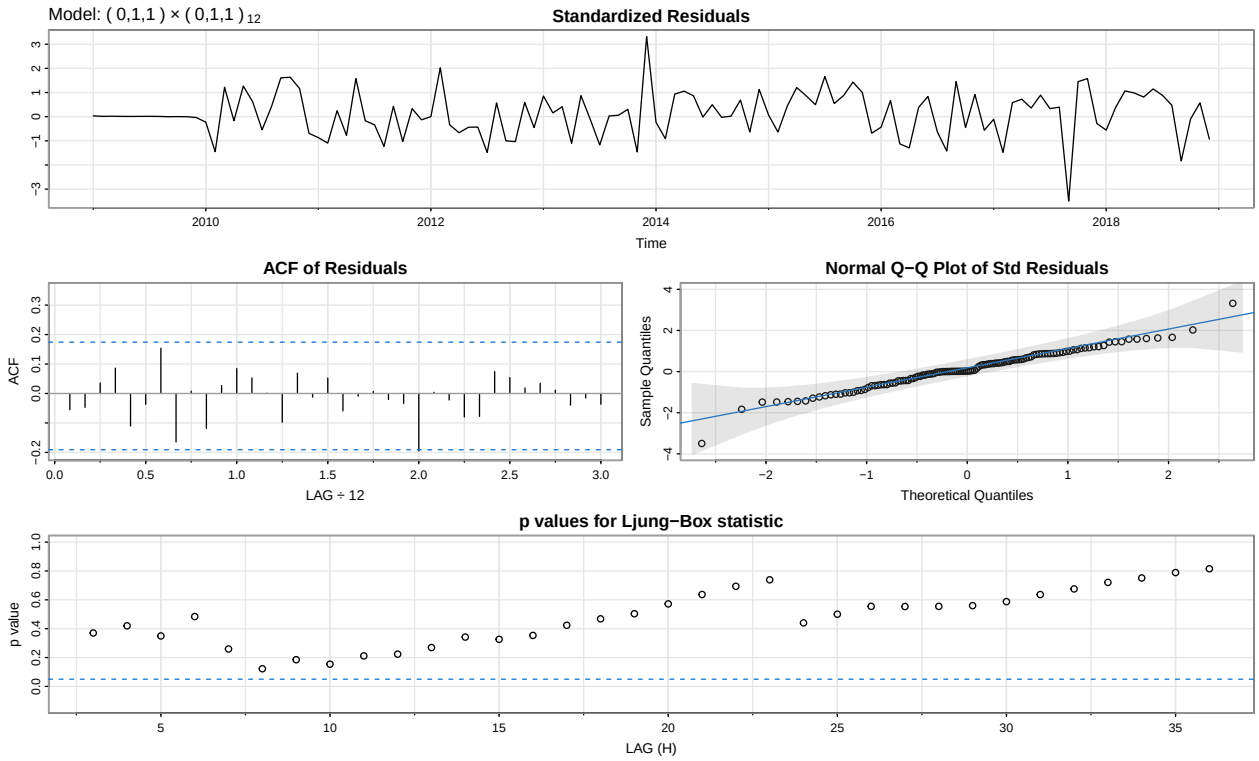

Figure 9: Various plots of the residuals for the SARIMA$(0,1,1)(0,1,1)_{12}$ model fitted to the Box-Cox transformed data

Figure 9 shows various plots for the residuals of Model 1, which we use to check for white noise characteristics. The plot for the residuals shows (weakly) stationarity since it has constant mean and a stable variance, except for the outliers around 2014 and mid 2017 that exceed 3 standard deviations. The ACF plot of the residuals do not exhibit any significant spikes, and so it agrees with our model assumptions. Similarly, since all data points (besides the two outliers) lie on the normal line of the the Q-Q plot, we can assume the residuals are normally distributed. Lastly, because all the p-values for the Q statistic from lag $H = 3$ to lag $H = 36$ exceed 0.05, we can assume that the residuals are white noise. Therefore, we can infer that the residuals of Model 1 have passed the diagnostics check, and since it also has the lowest AIC, AICc, and BIC values, we conclude that Model 1 fits the U.S. airline traffic series well. In fact, we have found that the data can be modeled by

$$(1-B^{12})(1-B)x_t = (1-0.5672_{(0.1192)}B^{12})(1-0.6889_{(0.0663)}B)w_t, \quad w_t \sim \text{wn}(0, 0.3333).$$

Forecasting

Figure 10 shows the forecasts from Model 1 for the next 12 months of the Box-Cox transformed data. We can see that the forecasts have captured the monthly seasonal pattern and the increasing

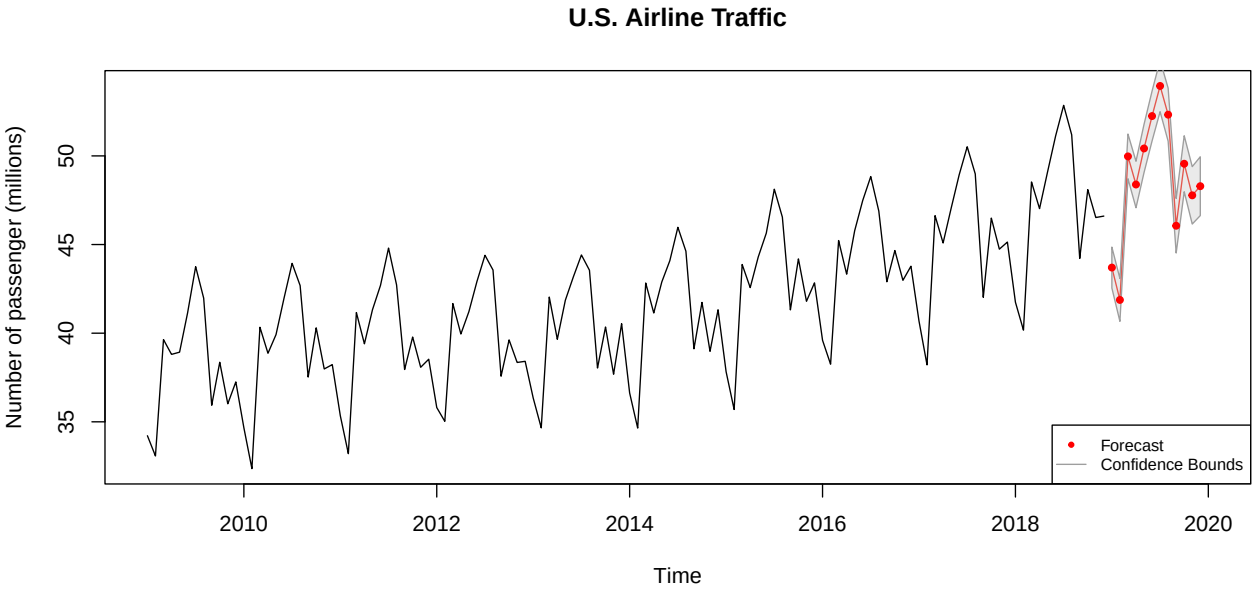

Figure 10: Twelve month forecast using the SARIMA$(0,1,1)(0,1,1)_{12}$ model on the Box-Cox transformed data

trend of the Box-Cox transformed data. We can also notice that the variance is slightly increasing. Moreover, *Figure 11* shows the forecasting ability of Model 1 by comparing the predicted values with the observed values that we saved for testing purposes. The observed values and the forecasted values are close to each other up until October 2019, while the observed values for the last two months are farther from the forecasted values. However, despite the observed value for December 2019, we can see that all observed values are withing the 95 confidence intervals of the forecasted values. Thus, we conclude that our model does a great job at predicting future values.

4.2 Spectral Analysis

The scaled periodogram and the smoothed periodograms for the Box-Cox transformed data are given by *Figure 12*. (Note that the frequency axis for the smoothed periodograms is labeled in multiples of 1/12.) The scaled periodogram has two main peaks at frequencies 0.0833 and 0.0083, corresponding to cycles with periods of $1/0.0833 \approx 12$ months and $1/0.0083 \approx 120$ months, respectively. Hence, we have rediscovered the annual cycle in the U.S. airline traffic series that was found in the SARIMA modeling process, along with a 10-year cycle—but this might not be meaningful since this is the span of the original data. Similarly, both smoothed periodograms show an annual cycle in the data since both have a peak at frequency 1. Also, both smoothed periodograms display peaks at frequencies $2, 3, 4$, and 5 (but the ones in the smoothed log periodogram are more visible), which indicates the presence of harmonics of the annual cycle. That is, the U.S. airline traffic series has some periodic components that are non-sinusoidal besides the annual cycle.

5. Conclusion

This report aimed to identify a SARIMA model of the monthly total of passengers in U.S. airlines in order to forecast future values and carry out its spectral analysis. The original time series was truncated to include only the data between 2009 and 2018, and the observations for 2019 were saved

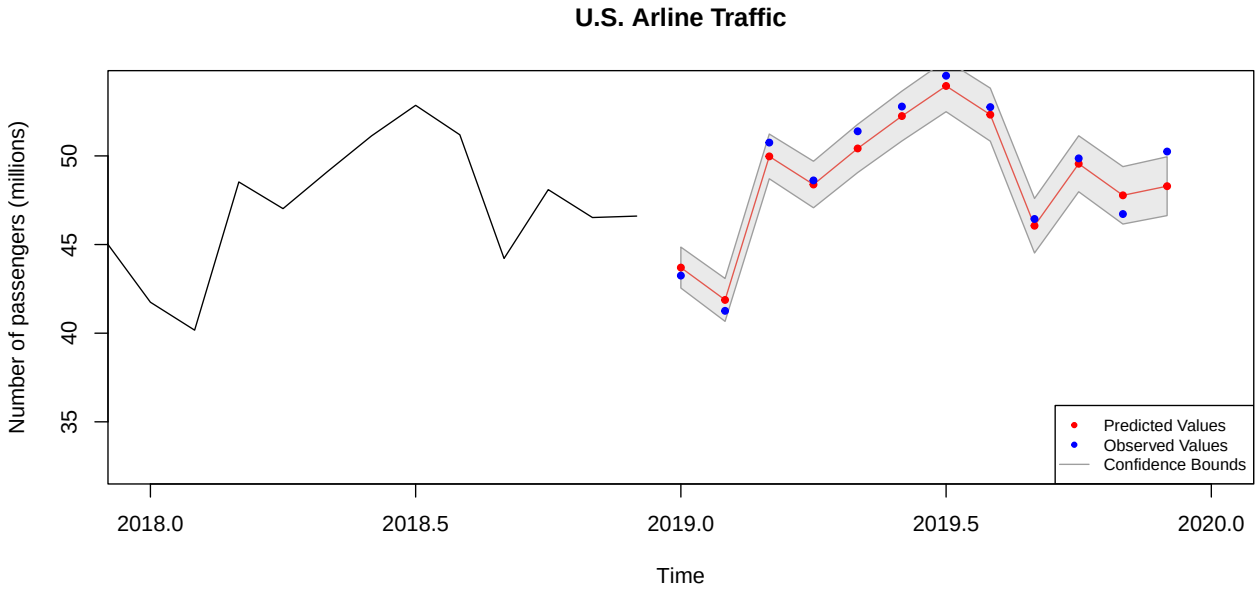

Figure 11: Predicted values versus observed values

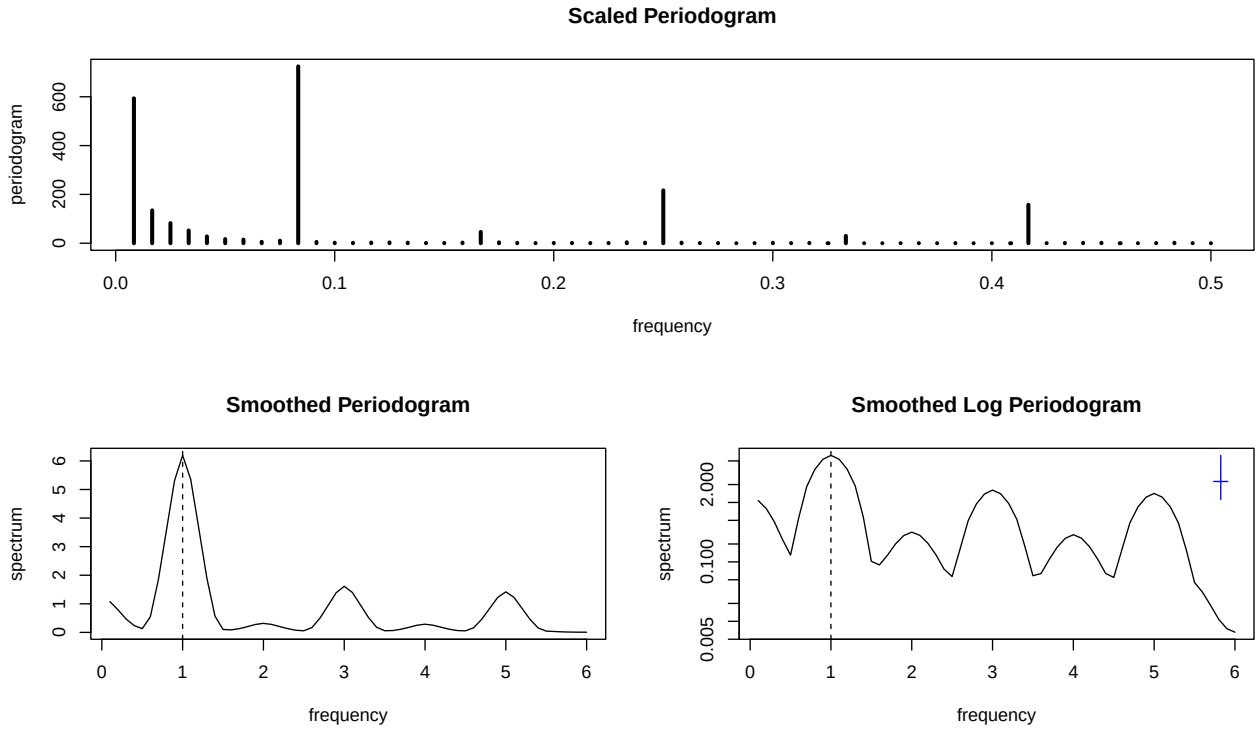

Figure 12: Various plots of the spectral estimates of the Box-Cox transformed data

for testing purposes. The training data showed patterns of an upward trend, monthly seasonality, and an increasing variance over time. So, in order to stabilize the variance and remove the trend and seasonality, a Box-Cox transformation was applied to the series along with a first-order difference and a twelfth-order difference. The analysis of the ACF and PACF plots of the transformed data suggested that a SARIMA$(0,1,1)(0,1,1)_{12}$ model was a good fit for the data, and thus, it was compared with other candidates by evaluating the significance of their coefficients and their AIC, AICc, and BIC values. As originally hypothesized, the SARIMA$(0,1,1)(0,1,1)_{12}$ model was the best fit for the transformed data, and its residuals were checked for patterns of white noise using various plots and the Ljung-Box test. The residuals for the model passed different tests of normality, and we concluded that the data can be modeled by

$$(1-B^{12})(1-B)x_t = (1-0.5672_{(0.1192)}B^{12})(1-0.6889_{(0.0663)}B)w_t, \quad w_t \sim \text{wn}(0, 0.3333).$$

Additionally, the forecasts for the next 12 months were compared with the testing data and showed that the model captured the components (trend, seasonality, and variance) of the transformed data, providing reasonable and accurate values. Lastly, the spectral analysis complemented the SARIMA model by rediscovering the annual cycle of the data and other cycles within the year.

Future research could try to model a SARIMA model on the original (not truncated) time series and explore the impact of COVID-19 on the data. Furthermore, applying more advanced techniques such as machine learning models could help discover more complex patterns of the data or improve the forecast accuracy of the data.

References

Time Series Analysis and Its Applications With R Examples by R. H. Shumway and D. S. Stoffer, Fourth Edition, Springer

Introduction to Time Series and Forecasting by P. J. Brockwell and R. A. Davis, Third Edition, Springer

Introductory Time Series with R by P. S. P. Cowpertwait and A. V. Metcalfe, Springer

U.S. Airline Traffic Data (2003-2023) by YYXIAN, Kaggla

Air Travle – Total by U.S. Department of Transportation

Appendix

```
library(astsa)
library(forecast)
library(MASS)
library(TSA)
library(knitr)
```

Box-Jenkins Approach

Data Exploration

```
# Import data
air_traffic <- read.csv("air traffic.csv")

# Training data
air_traf <- air_traffic$Pax[73:192]/1000000          # Values in millions
air <- ts(air_traf, start=c(2009,1), frequency=12)
air
```

```
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 2009 51.55944 49.64763 60.75897 59.33176 59.54559 63.35820 67.86807 64.78605
## 2010 52.31766 48.45503 61.95539 59.43950 61.24502 64.79035 68.19479 66.01727
## 2011 53.46039 49.86348 63.38290 60.35355 63.65418 66.03533 69.69454 66.08132
## 2012 54.23616 52.92358 64.26006 61.29990 63.47825 66.41246 68.98721 67.53866
## 2013 55.14828 52.30427 64.89192 60.79091 64.58635 66.87443 69.01949 67.51871
## 2014 55.62895 52.28323 66.26507 63.34200 66.36801 68.45878 71.74613 69.38813
## 2015 57.66458 54.05080 68.08174 65.81727 68.81979 71.16352 75.52380 72.75139
## 2016 60.69406 58.38803 70.43915 67.13058 71.41060 74.40686 76.78629 73.34920
## 2017 62.56056 58.32091 72.89789 70.19637 73.62146 76.95736 79.76447 77.04752
## 2018 64.38972 61.67479 76.24706 73.59249 77.26020 80.84706 83.93124 80.96266
##           Sep      Oct      Nov      Dec
## 2009 54.45498 58.57449 54.59309 56.68618
## 2010 57.15216 61.89853 57.93054 58.34781
## 2011 57.87545 61.00877 58.09196 58.86642
## 2012 57.23292 60.73233 58.55454 58.66638
## 2013 58.02114 61.98060 57.42318 62.30011
## 2014 59.87226 64.37110 59.62758 63.65230
## 2015 63.64546 68.63165 64.48298 66.28114
## 2016 66.38692 69.45465 66.52825 67.91402
## 2017 64.87076 72.65534 69.59015 70.28687
## 2018 68.67655 75.48063 72.71218 72.85534
```

```
# Testing data
air_test <- air_traffic$Pax[193:204]/1000000
air_test <- ts(air_test, start=c(2019,1), frequency=12)

# Plot the time series
```

```r
plot.ts(air, xlab="Time", ylab="Number of passengers (millions)",
        main="U.S. Airline Traffic")
```

```r
# Decompose the time series into its components
decomposition <- decompose(air)

observed <- decomposition$x
trend <- decomposition$trend
seasonal <- decomposition$seasonal
random <- decomposition$random

# Plot the time series and its components
plot.ts(cbind(observed,trend,seasonal,random), main="")
```

```r
# Solve for optimal lambda
transform <- boxcox(air~as.numeric(1:length(air)))
lambda <- transform$x[which(transform$y == max(transform$y))]
lambda
```

```
## [1] 0.8686869
```

```r
# Apply Box-Cox transformation to training and testing data
bc_air <- (air^lambda-1)/lambda
bc_air_test <- (air_test^lambda-1)/lambda

# Decompose transformed data into its components
bc_decomposition <- decompose(bc_air)

observed <- bc_decomposition$x
trend <- bc_decomposition$trend
seasonal <- bc_decomposition$seasonal

# Plot the trend and seasonal graphs of the transformed data
plot.ts(cbind(trend,seasonal), main="")
```

```r
# Difference the data to remove trend
air_d1 <- diff(bc_air, 1)

# Difference the data to remove seasonality
air_d12 <- diff(air_d1, 12)

# Plot the time series and each of its transformations
plot.ts(cbind(air,bc_air,air_d1,air_d12), main="")
```

```r
# Get the variances of the data and its transformations
round(c(var(air), var(bc_air), var(air_d1), var(air_d12)), 4)
```

```
## [1] 54.6690 18.2872 11.2012  0.6629
```

```r
# Make a table with the variances
transformations <- c("Original", "Box-Cox", "1st Difference", "12th Difference")
variances <- round(c(var(air), var(bc_air), var(air_d1), var(air_d12)), 4)

kable(cbind(transformations, variances), "pipe", align=c("c", "c"),
      col.names=c("Transformation", "Variance"),
      caption="The variances of the U.S. airline traffic series after each transformation")
```

```r
# Plot the time plot, ACF, and histogram of the differenced data
layout(matrix(c(1,1,2,3), nrow=2, byrow=TRUE))
plot(air_d12, main="Time Plot", ylab="Passengers")
acf(air_d12, main="Twiced-Differenced Data")
hist(as.numeric(air_d12), breaks=seq(from=-3, to=3, by=1/5), main="Histogram",
     xlab="Data")
```

Identification or Parameters

```r
# Plot the ACF and PACF of the differenced data
acf2(air_d12, main="Twiced-Differenced Data")
```

Estimation of Parameters

```r
# Fit the possible models

# SARIMA(0,1,1)(0,1,1)[12]
mod_1 <- sarima(bc_air, p=0, d=1, q=1, P=0, D=1, Q=1, S=12, details=FALSE)
```

```
## <><><><><><><><><><><><><><>
##
## Coefficients:
##      Estimate     SE  t.value p.value
## ma1   -0.6889 0.0663 -10.3937       0
## sma1  -0.5672 0.1192  -4.7586       0
##
## sigma^2 estimated as 0.3332733 on 105 degrees of freedom
##
## AIC = 1.844837  AICc = 1.845915  BIC = 1.919776
##
```

```r
# SARIMA(3,1,0)(0,1,1)[12]
mod_2 <- sarima(bc_air, p=3, d=1, q=0, P=0, D=1, Q=1, S=12, details=FALSE)
```

```
## <><><><><><><><><><><><><><>
##
## Coefficients:
##      Estimate     SE t.value p.value
## ar1   -0.7268 0.0956 -7.6054  0.0000
## ar2   -0.4935 0.1070 -4.6118  0.0000
```

```
## ar3   -0.2445 0.0952 -2.5677  0.0117
## sma1  -0.5321 0.1194 -4.4545  0.0000
##
## sigma^2 estimated as 0.3287642 on 103 degrees of freedom
##
## AIC = 1.86264  AICc = 1.866305  BIC = 1.987538
##
```

```
# SARIMA(0,1,1)(1,1,2)[12]
mod_3 <- sarima(bc_air, p=0, d=1, q=1, P=1, D=1, Q=2, S=12, details=FALSE)
```

```
## <><><><><><><><><><><><><><>
##
## Coefficients:
##      Estimate     SE  t.value p.value
## ma1   -0.6945 0.0656 -10.5803  0.0000
## sar1  -0.5696 0.2990  -1.9053  0.0595
## sma1   0.1152 0.3251   0.3543  0.7238
## sma2  -0.5971 0.1921  -3.1080  0.0024
##
## sigma^2 estimated as 0.3013099 on 103 degrees of freedom
##
## AIC = 1.822127  AICc = 1.825792  BIC = 1.947025
##
```

```
# SARIMA(1,1,1)(0,1,1)[12]
mod_4 <- sarima(bc_air, p=1, d=1, q=1, P=0, D=1, Q=1, S=12, details=FALSE)
```

```
## <><><><><><><><><><><><><><>
##
## Coefficients:
##      Estimate     SE t.value p.value
## ar1   -0.0791 0.1369 -0.5778  0.5647
## ma1   -0.6494 0.1012 -6.4159  0.0000
## sma1  -0.5572 0.1207 -4.6177  0.0000
##
## sigma^2 estimated as 0.3328548 on 104 degrees of freedom
##
## AIC = 1.860452  AICc = 1.86263  BIC = 1.960371
##
```

```
# SARIMA(0,1,2)(0,1,1)[12]
mod_5 <- sarima(bc_air, p=0, d=1, q=2, P=0, D=1, Q=1, S=12, details=FALSE)
```

```
## <><><><><><><><><><><><><><>
##
## Coefficients:
##      Estimate     SE t.value p.value
## ma1   -0.7397 0.1037 -7.1313   0.000
## ma2    0.0714 0.1096  0.6518   0.516
```

```
## sma1  -0.5546 0.1208 -4.5897   0.000
##
## sigma^2 estimated as 0.3327014 on 104 degrees of freedom
##
## AIC = 1.859544  AICc = 1.861721  BIC = 1.959462
##
```

```
# SARIMA(3,1,1)(0,1,1)[12]
mod_6 <- sarima(bc_air, p=3, d=1, q=1, P=0, D=1, Q=1, S=12, details=FALSE)
```

```
## <><><><><><><><><><><><><><>
##
## Coefficients:
##      Estimate     SE t.value p.value
## ar1   -0.7981 0.3812 -2.0934  0.0388
## ar2   -0.5390 0.2572 -2.0954  0.0386
## ar3   -0.2690 0.1536 -1.7505  0.0830
## ma1    0.0750 0.3909  0.1919  0.8482
## sma1  -0.5274 0.1218 -4.3282  0.0000
##
## sigma^2 estimated as 0.3289064 on 102 degrees of freedom
##
## AIC = 1.881031  AICc = 1.886583  BIC = 2.030909
##
```

```
# SARIMA(4,1,0)(0,1,1)[12]
mod_7 <- sarima(bc_air, p=4, d=1, q=0, P=0, D=1, Q=1, S=12, details=FALSE)
```

```
## <><><><><><><><><><><><><><>
##
## Coefficients:
##      Estimate     SE t.value p.value
## ar1   -0.7234 0.0977 -7.4020  0.0000
## ar2   -0.4851 0.1183 -4.1001  0.0001
## ar3   -0.2329 0.1180 -1.9737  0.0511
## ar4    0.0170 0.1020  0.1664  0.8682
## sma1  -0.5276 0.1225 -4.3087  0.0000
##
## sigma^2 estimated as 0.3289097 on 102 degrees of freedom
##
## AIC = 1.881073  AICc = 1.886625  BIC = 2.030951
##
```

```
# SARIMA(2,1,0)(0,1,1)[12]
mod_8 <- sarima(bc_air, p=2, d=1, q=0, P=0, D=1, Q=1, S=12, details=FALSE)
```

```
## <><><><><><><><><><><><><><>
##
## Coefficients:
##      Estimate     SE t.value p.value
```

```
## ar1   -0.6437 0.0926 -6.9539   0e+00
## ar2   -0.3402 0.0918 -3.7064   3e-04
## sma1  -0.5441 0.1206 -4.5119   0e+00
##
## sigma^2 estimated as 0.3489109 on 104 degrees of freedom
##
## AIC = 1.903799  AICc = 1.905977  BIC = 2.003718
##
```

```
# Make a table with the significance of the coefficients and the ICs values
models <- c("SARIMA(0,1,1)(0,1,1)$_{12}$", "SARIMA(3,1,0)(0,1,1)$_{12}$",
            "SARIMA(0,1,1)(1,1,2)$_{12}$", "SARIMA(1,1,1)(0,1,1)$_{12}$",
            "SARIMA(0,1,2)(0,1,1)$_{12}$", "SARIMA(3,1,1)(0,1,1)$_{12}$",
            "SARIMA(4,1,0)(0,1,1)$_{12}$", "SARIMA(2,1,0)(0,1,1)$_{12}$")
coefficients <- c("All coefficients are significant",
                  "All coefficients are significant",
                  "SAR1 and SMA1 are not significant",
                  "AR1 is not significant",
                  "MA2 is not significant",
                  "AR3 and MA1 are not significant",
                  "AR3 and AR4 are not significant",
                  "All components are significant")
AICs <- round(c(mod_1$ICs[1], mod_2$ICs[1], mod_3$ICs[1], mod_4$ICs[1],
          mod_5$ICs[1], mod_6$ICs[1], mod_7$ICs[1], mod_8$ICs[1]), 4)
AICcs <- round(c(mod_1$ICs[2], mod_2$ICs[2], mod_3$ICs[2], mod_4$ICs[2],
          mod_5$ICs[2], mod_6$ICs[2], mod_7$ICs[2], mod_8$ICs[2]), 4)
BICs <- round(c(mod_1$ICs[3], mod_2$ICs[3], mod_3$ICs[3], mod_4$ICs[3],
          mod_5$ICs[3], mod_6$ICs[3], mod_7$ICs[3], mod_8$ICs[3]), 4)

kable(cbind(models, coefficients, AICs, AICcs, BICs), "pipe",
      align=c("c", "c", "c", "c", "c"), row.names = FALSE,
      col.names=c("Model", "Significance of Coefficients", "AIC", "AICc", "BIC"),
      caption="Comparison of estimated coefficients based on the MLE method")
```

Model Diagnostics

```
# Plot the time plot, ACF, Q-Q plot, and p-values of the residuals of model 1
sarima(bc_air, p=0, d=1, q=1, P=0, D=1, Q=1, S=12)
```

Forecasting

```
# Forecast the next 12 data points
forecast <- sarima.for(bc_air, n.ahead=12, p=0, d=1, q=1, P=0, D=1, Q=1, S=12,
                       plot=FALSE)

# Compute the 95% confidence interval
upper_bound <- forecast$pred + 2*forecast$se
lower_bound <- forecast$pred - 2*forecast$se
```

```r
xx <- c(time(upper_bound), rev(time(upper_bound)))
yy <- c(lower_bound, rev(upper_bound))

# Plot the forecast
ts.plot(bc_air, forecast$pred, col=1:2, xlim=c(2009,2020),
        main="U.S. Airline Traffic", ylab="Number of passenger (millions)")
polygon(xx, yy, border=8, col=gray(0.6, alpha=0.2))
lines(forecast$pred, type="p", pch=20, col="red")
legend("bottomright", legend=c("Forecast", "Confidence Bounds"), cex=.75,
       col=c("red", 8), pch=c(20,NA), lty=c(NA,1), )
```

```r
# Zoom in forecast plot
ts.plot(bc_air, forecast$pred, col=1:2, xlim=c(2018,2020),
        main="U.S. Arline Traffic", ylab="Number of passengers (millions)")
polygon(xx, yy, border=8, col=gray(0.6, alpha=0.2))
lines(forecast$pred, type="p", pch=20, col="red")
points(bc_air_test, pch=20, col="blue")
legend(legend=c("Predicted Values", "Observed Values","Confidence Bounds"),
       "bottomright", col=c("red", "blue", 8), pch=c(20,20,NA), lty=c(NA,NA,1),
       cex=.75)
```

Spectral Analysis

```r
# Plot the scaled and smoothed periodograms
layout(matrix(c(1,1,2,3), nrow=2, byrow=TRUE))
periodogram(bc_air, log='no', plot=TRUE, ylab="periodogram", xlab="frequency",
            main="Scaled Periodogram", lwd=3)
spectrum(bc_air, spans=c(5,5), log="no", main="Smoothed Periodogram", sub="")
abline(v=1, lty=2)
spectrum(bc_air, spans=c(5,5), main="Smoothed Log Periodogram", sub="")
abline(v=1, lty=2)
```