Forecasting Air Passengers

Mathew Roberts

15/04/2021

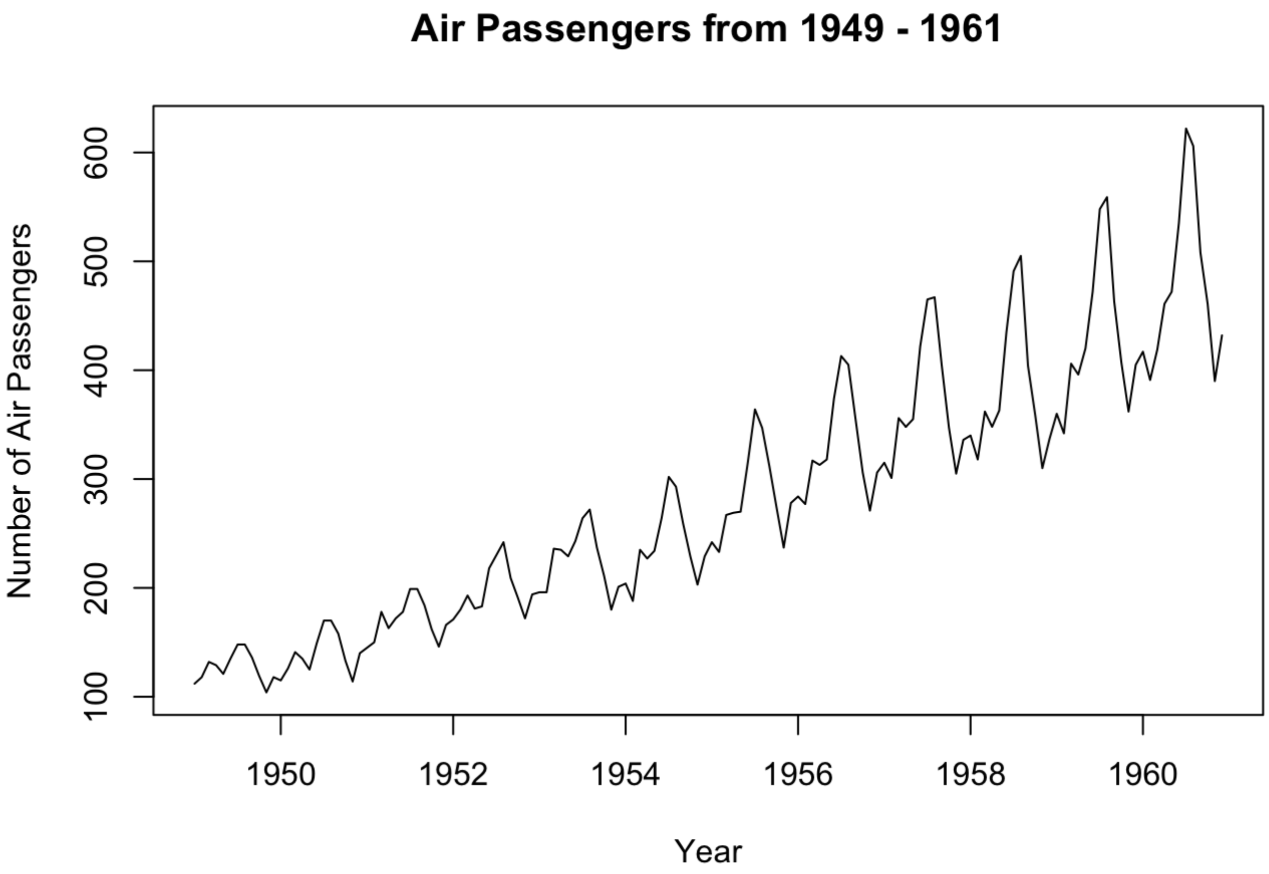
Air Passenger Forecast

The following code aims to provide the optimal forecast for the number of air passengers in the 12 months proceeding the year 1961. We will be using the SARIMA model operating in the Box-Jenkins framework, where we will be assuming that the data is weakly stationary.

Preprocessing

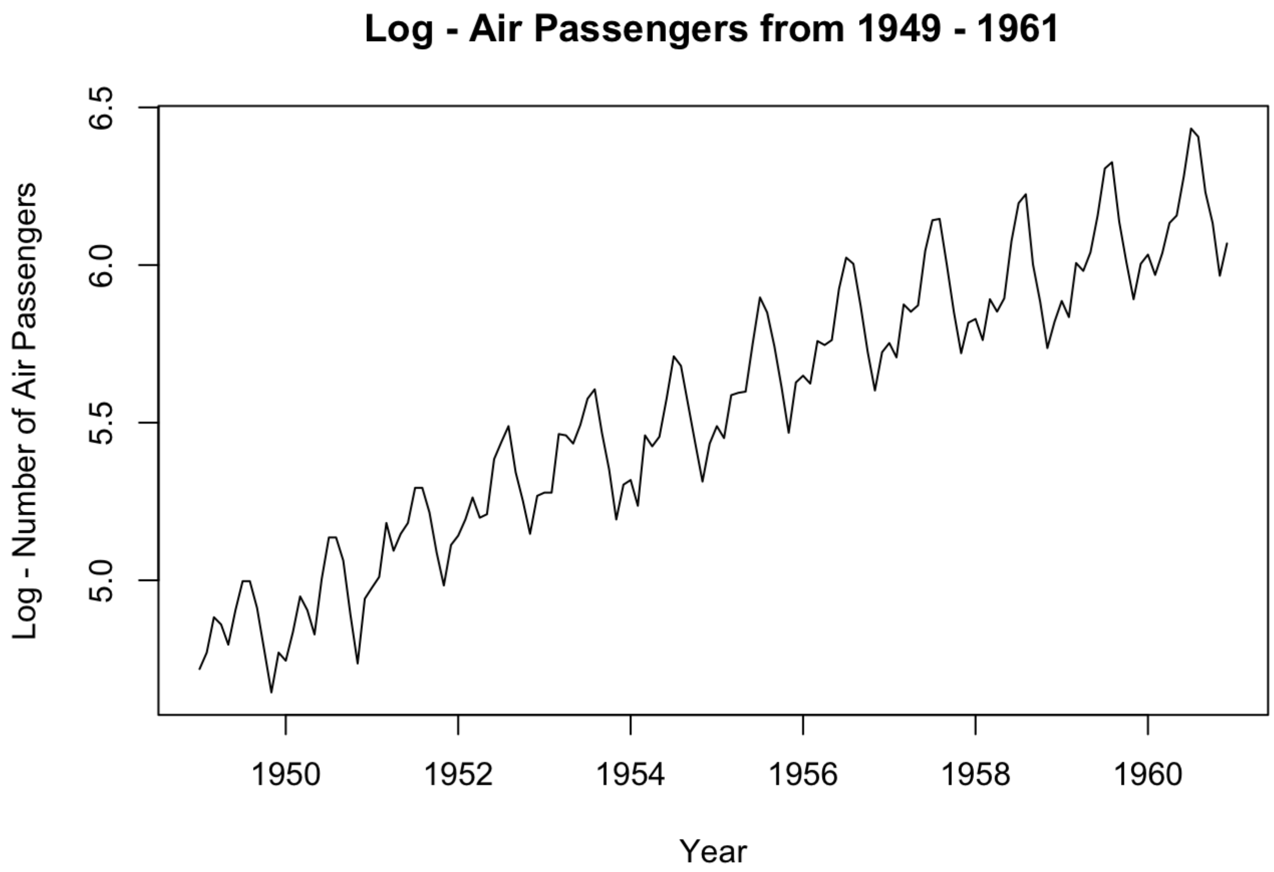
Before we begin fitting the model, we must ensure that our starting data has a homogeneous variance. From the plot below, we can see that the variance increases with time, therefore we must apply a log-transform to the data to stabilise the variance.

```r
library(astsa)
AP <- AirPassengers
plot(AP, xlab = "Year", ylab ="Number of Air Passengers", main = "Air Passengers from 1949 - 1961")
```



We now apply a log-transform and re-plot the data.

```
lAP <- log(AP)
plot(lAP, xlab = "Year", ylab ="Log - Number of Air Passengers", main = "Log - Air Passengers from 1949 - 1961")
```



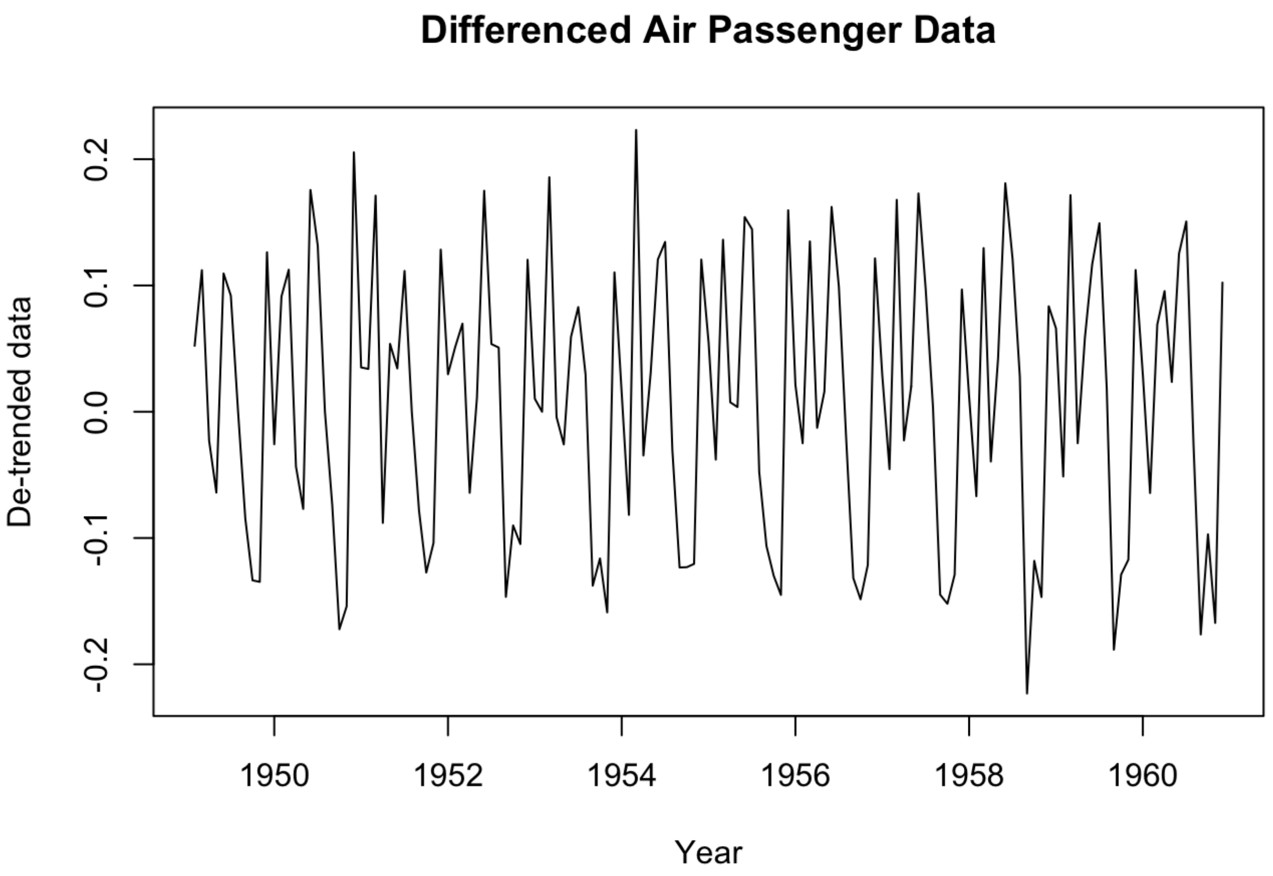
We note that the above plot now has a homogeneous variance, so now we can begin to apply the de-trending and de-seasonality operators.

Constructing the Model

The additive model is given by, $X_t = m_t + S_t + Y_t$, where m_t, S_t and Y_t are the trend, seasonality and noise factors respectively.

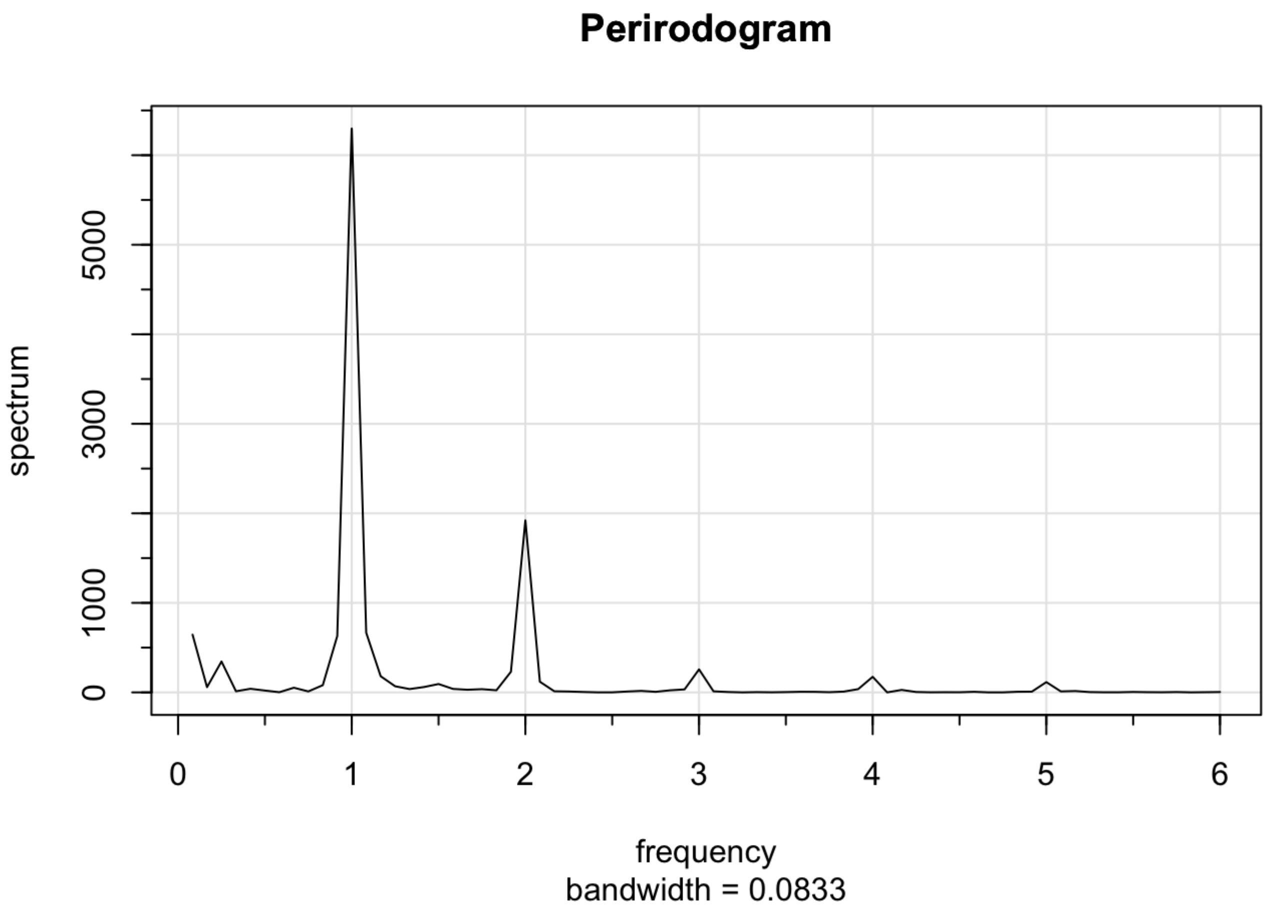
The differencing operations seek to remove the trend and seasonality such that we can focus on modelling the noise, Y_t.

```
dlAP = diff(lAP)
plot(dlAP, xlab = "Year", ylab = "De-trended data", main ="Differenced Air Passenger Data")
```



In the above code we de-trend the data using the differencing operator, however we can still see the seasonality in the data. We must now apply seasonal differencing to this detrended data. But first we must accurately determine the seasonality of the data. This can be done by assessing the frequency of the data in a “Perirodogram” using the $mvspec()$ function.

```
cycle = mvspec(AP, log = "no", main = "Perirodogram")
```

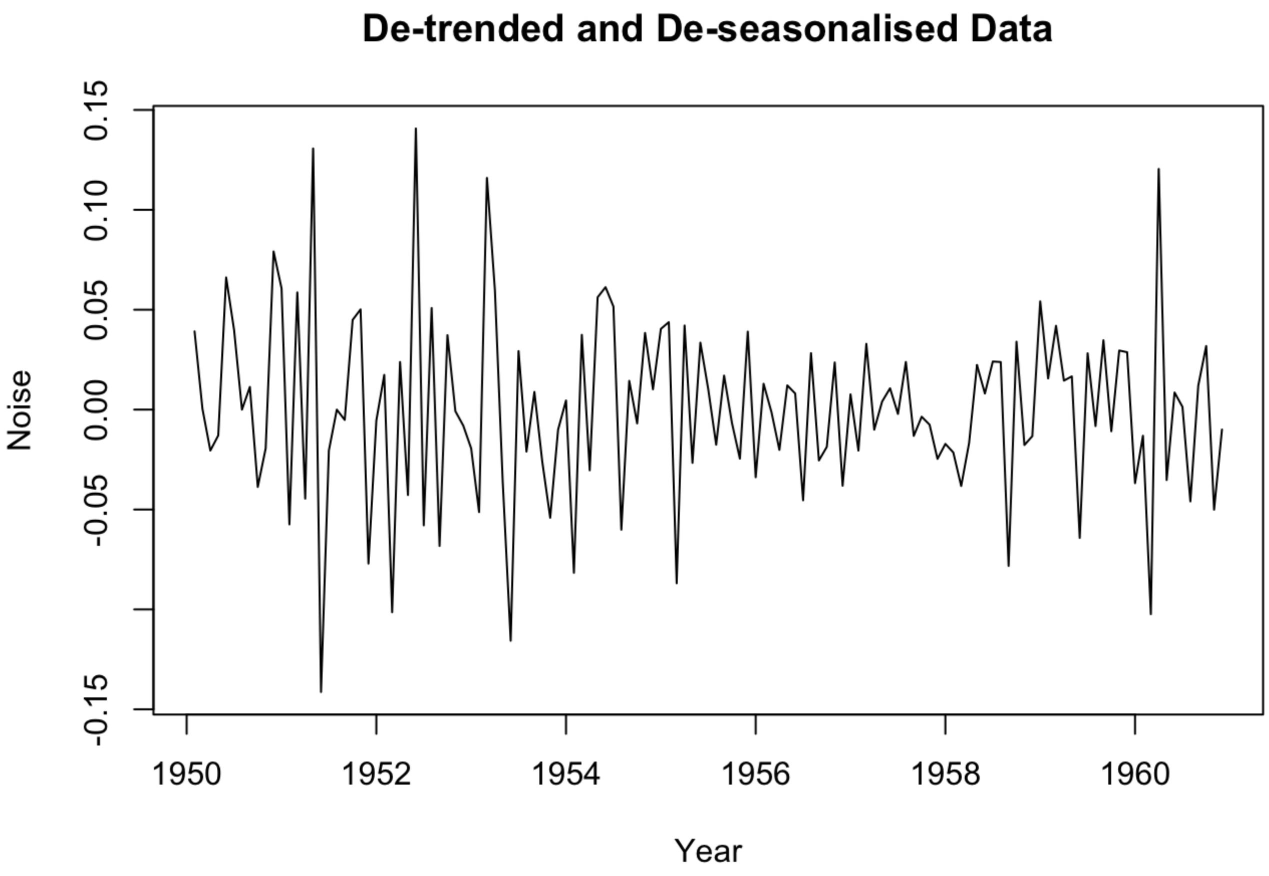


We can see that there is a large spike at frequency equal to 1. Since the data is monthly, each integer frequency represents a time period of 12 months, therefore our seasonality is 12.

We now apply seasonal differencing to the de-trended data, and visually inspect the noise.

```
sdlAP = diff(dlAP, 12)

plot(sdlAP, xlab = "Year", ylab = "Noise", main = "De-trended and De-seasonalised Data")
```



We can see that the data fluctuates around the value 0, with little evidence of structure or seasonality within the data. At a glance, this data looks weakly stationary.

Testing for Weakly Stationarity

We must justify the assertion that this data is weakly stationary by applying tests. These tests are the Augmented Dickey Fuller (ADF) test, and the KPSS test. We seek to reject the ADF test at significance level $\alpha = 0.5$, and fail to reject the KPSS at the same significance. If both tests are passed we can be confident at 95%, that the de-trended and de-seasonalised data is weakly stationary.

```
#We reject adf test - p value smaller than 0.05
adf.test(sdlAP)
```

```
## Warning in adf.test(sdlAP): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  sdlAP
## Dickey-Fuller = -5.1993, Lag order = 5, p-value = 0.01
## alternative hypothesis: stationary
```

```
#We do not reject the kpss test - the p value is greater than significance level 0.05
unitroot_kpss((sdlAP))
```

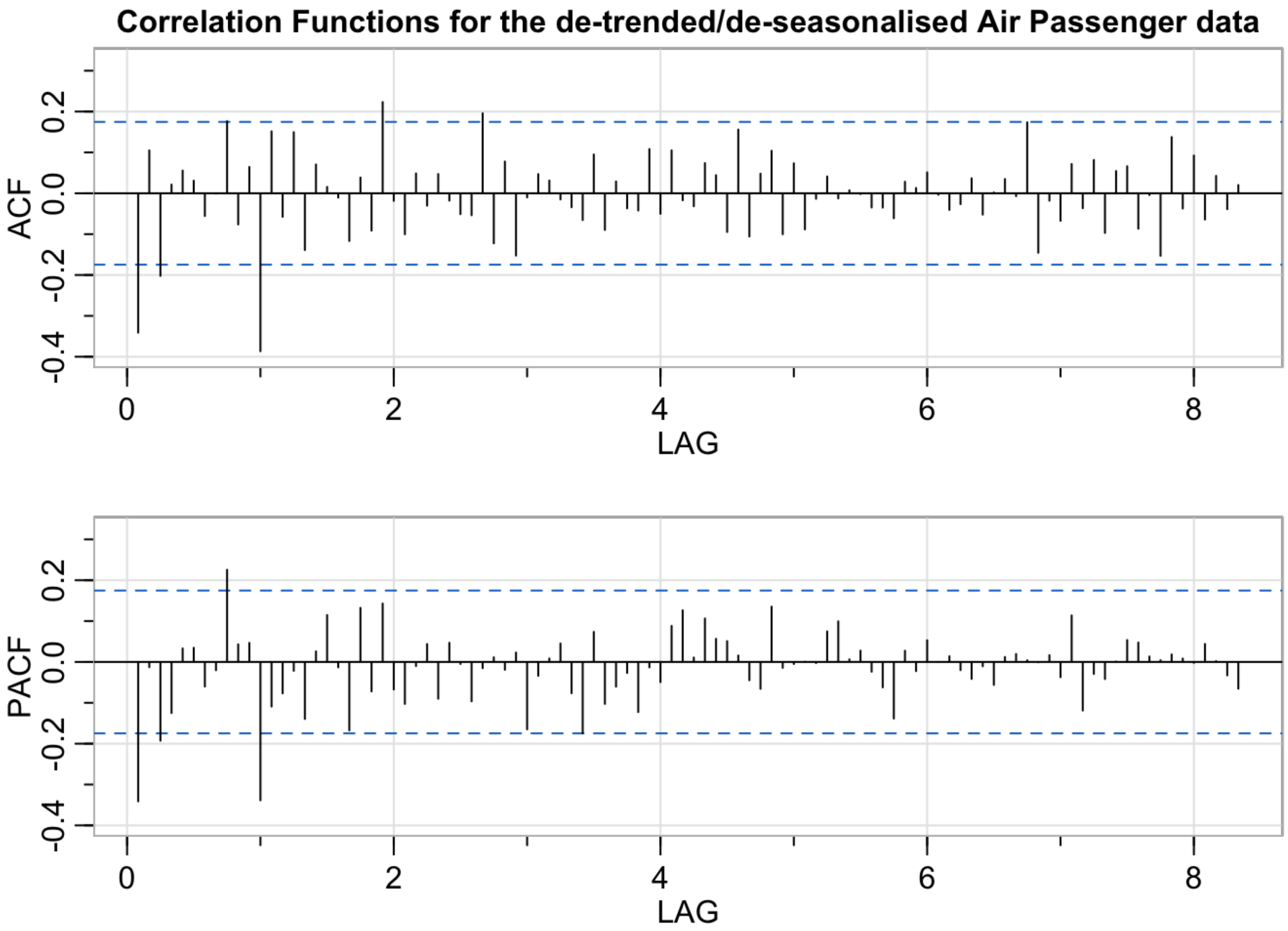
```
##   kpss_stat kpss_pvalue
##  0.08436452  0.10000000
```

We reject the ADF test as the registered p-value is below α, and we do not reject the KPSS test as the p-value is greater than α. Thus we conclude that the data is weakly stationary at a 95% confidence level.

Identifying the Parameters of the SARIMA Model

From the previous tests, we are confident that the de-trended and de-seasonalised data is weakly stationary. We now plot the ACF and PACF, which allow us to estimate the parameters of the SARIMA model.

```
acf2(sdlAP, max.lag = 100, main = "Correlation Functions for the de-trended/de-seasonalised Air Passenger data")
```



```
##       [,1]  [,2]  [,3]  [,4] [,5] [,6]  [,7]  [,8] [,9] [,10] [,11] [,12] [,13]
## ACF  -0.34  0.11 -0.20  0.02 0.06 0.03 -0.06  0.00 0.18 -0.08  0.06 -0.39  0.15
## PACF -0.34 -0.01 -0.19 -0.13 0.03 0.03 -0.06 -0.02 0.23  0.04  0.05 -0.34 -0.11
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
## ACF  -0.06  0.15 -0.14  0.07  0.02 -0.01 -0.12  0.04 -0.09  0.22 -0.02  -0.1
## PACF -0.08 -0.02 -0.14  0.03  0.11 -0.01 -0.17  0.13 -0.07  0.14 -0.07  -0.1
##      [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
## ACF   0.05 -0.03  0.05 -0.02 -0.05 -0.05  0.20 -0.12  0.08 -0.15 -0.01  0.05
## PACF -0.01  0.04 -0.09  0.05  0.00 -0.10 -0.02  0.01 -0.02  0.02 -0.16 -0.03
##      [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46] [,47] [,48] [,49]
## ACF   0.03 -0.02 -0.03 -0.07  0.10 -0.09  0.03 -0.04 -0.04  0.11 -0.05  0.11
## PACF  0.01  0.05 -0.08 -0.17  0.07 -0.10 -0.06 -0.03 -0.12 -0.01 -0.05  0.09
##      [,50] [,51] [,52] [,53] [,54] [,55] [,56] [,57] [,58] [,59] [,60] [,61]
## ACF  -0.02 -0.03  0.07  0.04 -0.09  0.16 -0.11  0.05  0.10 -0.10  0.07 -0.09
## PACF  0.13  0.01  0.11  0.06  0.05  0.02 -0.04 -0.07  0.14 -0.01  0.00  0.00
##      [,62] [,63] [,64] [,65] [,66] [,67] [,68] [,69] [,70] [,71] [,72] [,73]
## ACF  -0.01  0.04 -0.01  0.01  0.00 -0.03 -0.04 -0.06  0.03  0.01  0.05     0
## PACF  0.00  0.07  0.10  0.01  0.03 -0.02 -0.06 -0.14  0.03 -0.02  0.05     0
##      [,74] [,75] [,76] [,77] [,78] [,79] [,80] [,81] [,82] [,83] [,84] [,85]
## ACF  -0.04 -0.03  0.04 -0.05  0.00  0.03 -0.01  0.17 -0.15 -0.02 -0.07  0.07
## PACF  0.01 -0.02 -0.04 -0.01 -0.06  0.01  0.02  0.00  0.00  0.02 -0.04  0.11
##      [,86] [,87] [,88] [,89] [,90] [,91] [,92] [,93] [,94] [,95] [,96] [,97]
## ACF  -0.04  0.08 -0.10  0.05  0.07 -0.09  0.00 -0.15  0.14 -0.04  0.09 -0.06
## PACF -0.12 -0.03 -0.04  0.00  0.05  0.05  0.01  0.00  0.02  0.01  0.00  0.04
##      [,98] [,99] [,100]
## ACF   0.04 -0.04   0.02
## PACF  0.00 -0.03  -0.07
```

The above plots incorporate both seasonal and non-seasonal AR, MA and ARMA models on the same plot. The non-seasonal characteristics are embedded between lag-0 and lag-1, whilst the seasonal characteristics can be determined after lag-1. This is for both ACF and PACF.

For the non-seasonal section of the plots, both the ACF and PACF appear to exponentially decay to 0 as the lag increases to 1. This is indicative of an ARMA(1,1) model for the non-seasonal aspect of the SARIMA model.

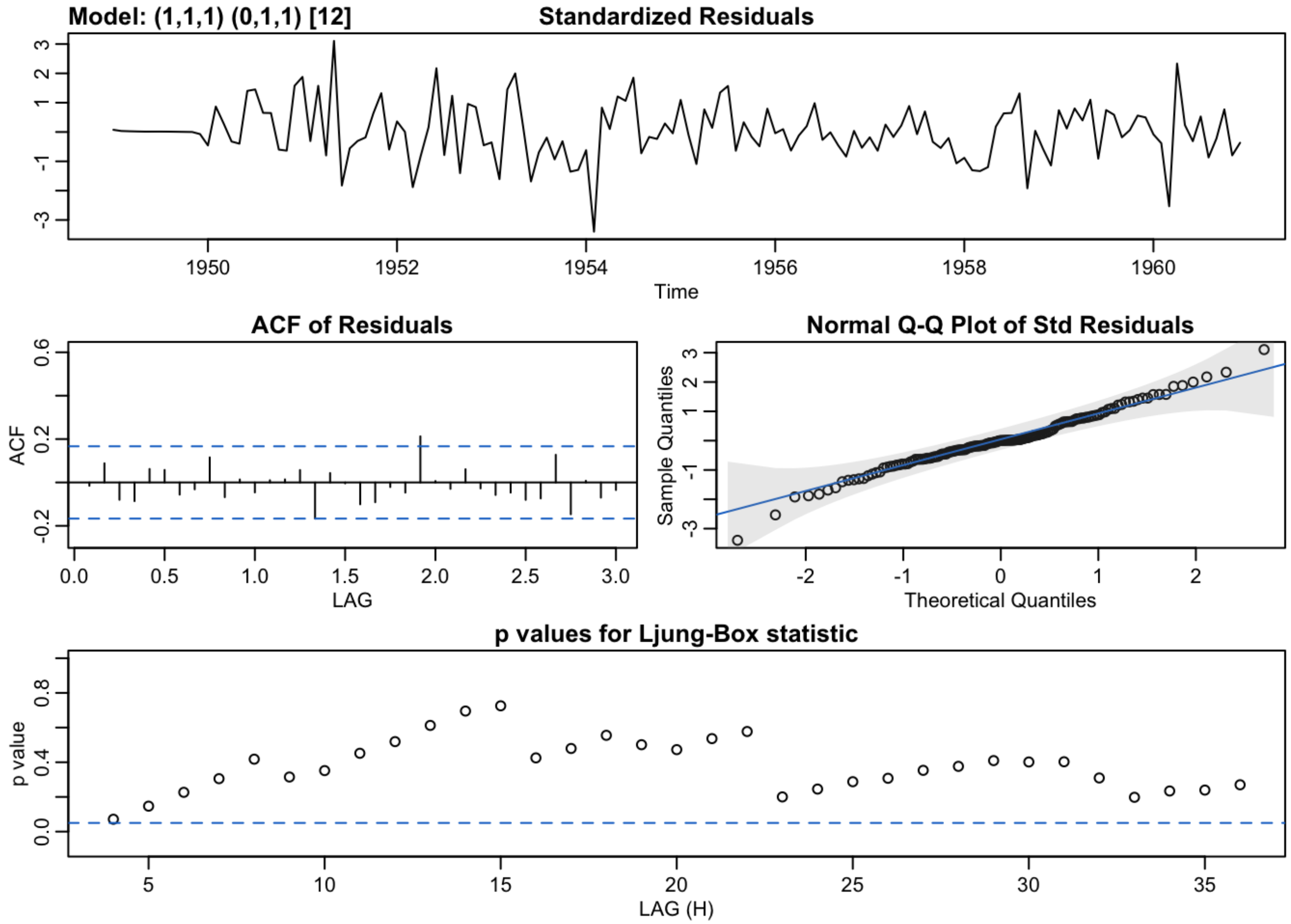
For the seasonal section of the plots, we must look at the behaviour of the graph for each integer lag as each integer represents the full season of 12 months. Through a close inspection, I estimate that the ACF cuts off after lag-1, and the PACF exponentially decays after lag-1 - this observation is hard to see but this has been based on the fact that the difference in size between the spike in the ACF at lag-1 and the following spike is much greater than the equivalent size difference in the PACF plot. Adhering to this model, I suggest that this is characteristic of an SMA(1) model.

Testing the Model

We are now in position to perform diagnostic tests on our SARIMA model. To reiterate, our full SARIMA model includes the initial trend/seasonal differencing as well as the ARMA(1,1) and SMA(1) model inferred from the ACF/PACF plots.

```
model = sarima(lAP, 1,1,1,0,1,1,12)
```

```
## initial  value -3.085211
## iter   2 value -3.225399
## iter   3 value -3.276697
## iter   4 value -3.276902
## iter   5 value -3.282134
## iter   6 value -3.282524
## iter   7 value -3.282990
## iter   8 value -3.286319
## iter   9 value -3.286413
## iter  10 value -3.288141
## iter  11 value -3.288262
## iter  12 value -3.288394
## iter  13 value -3.288768
## iter  14 value -3.288969
## iter  15 value -3.289089
## iter  16 value -3.289094
## iter  17 value -3.289100
## iter  17 value -3.289100
## iter  17 value -3.289100
## final  value -3.289100
## converged
## initial  value -3.288388
## iter   2 value -3.288459
## iter   3 value -3.288530
## iter   4 value -3.288649
## iter   5 value -3.288753
## iter   6 value -3.288781
## iter   7 value -3.288784
## iter   7 value -3.288784
## iter   7 value -3.288784
## final  value -3.288784
## converged
```



The standardised residuals look like an uncorrelated sequence, there are regular oscillations around 0 with not too many spikes. The mean is almost 0 at 0.0529 which is expected for a white noise sequence. The ACF of the residuals also lie within the blue significance lines, which further suggests that the residuals are an i.i.d sequence. The majority of the data points in the Normalised Q-Q plot also lie close to or on the blue line as would be expected of an uncorellated sequence.

The last plot showing the p-values for the Ljung-Box statistic also show that for each lag, the p-value is greater than the significance level of α. The null hypothesis of this statistic states that the residuals are independent, and since the p-value's are greater than the significance level, then we do not reject this null hypothesis.

We include one final test, the Ljung-Box Q test statistic, which is an aggregated version of the above Ljung-Box statistic. It considers all lags and considers the null hypothesis that the data is independently distributed.

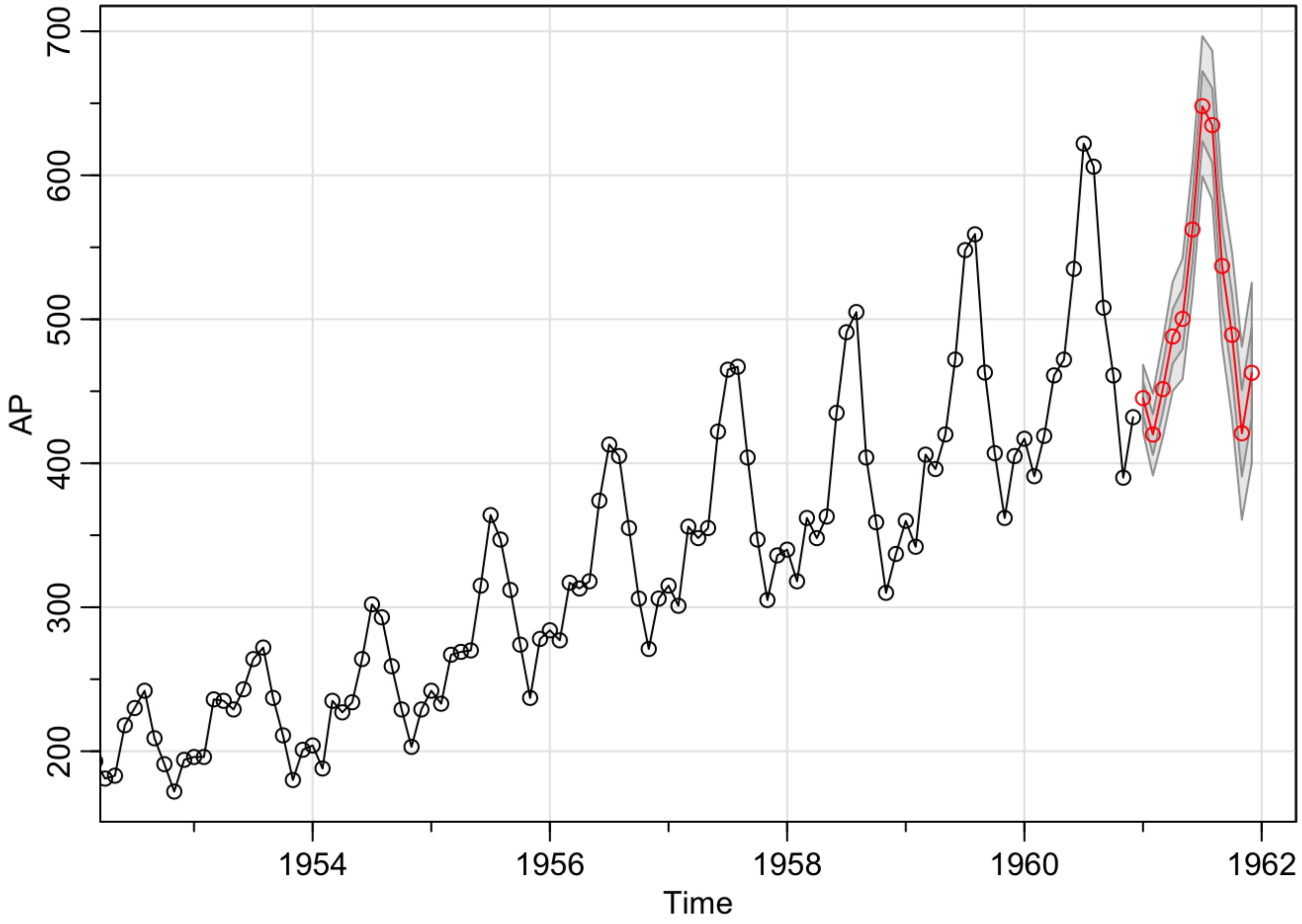
```
Box.test(resid(model$fit), lag = 50, type = "Ljung-Box", fitdf = 3)
```

```
##
##  Box-Ljung test
##
## data:  resid(model$fit)
## X-squared = 52.785, df = 47, p-value = 0.2604
```

Forecast

The model has a satisfactory performance when subject to the above diagnostic tests, therefore I will now use it to forecast 12 months into the future. The lighter grey lines, show the prediction with a variation of 2σ whereas the darker grey lines, show the prediction with a variation of σ

```
sarima.for(AP, 12, 1,1,1,0,1,1,12)
```



```
## $pred
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 1961 445.2039 419.9384 451.4890 488.0462 500.2915 562.3572 647.9834 634.8065
##           Sep      Oct      Nov      Dec
## 1961 537.0047 489.1759 420.8429 462.7791
##
## $se
##           Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
## 1961 11.63320 14.18976 16.79106 18.94353 20.89638 22.67698 24.32867 25.87489
##            Sep      Oct      Nov      Dec
## 1961 27.33385 28.71877 30.03991 31.30534
```

The forecast appears to be visually consistent with the past data. The seasonality is present and whilst also showing the increasing varinace of the Air Passengers as time progresses.