ARIMA forecasting - Spain metro users

Javier Eloy Martinez Ramos

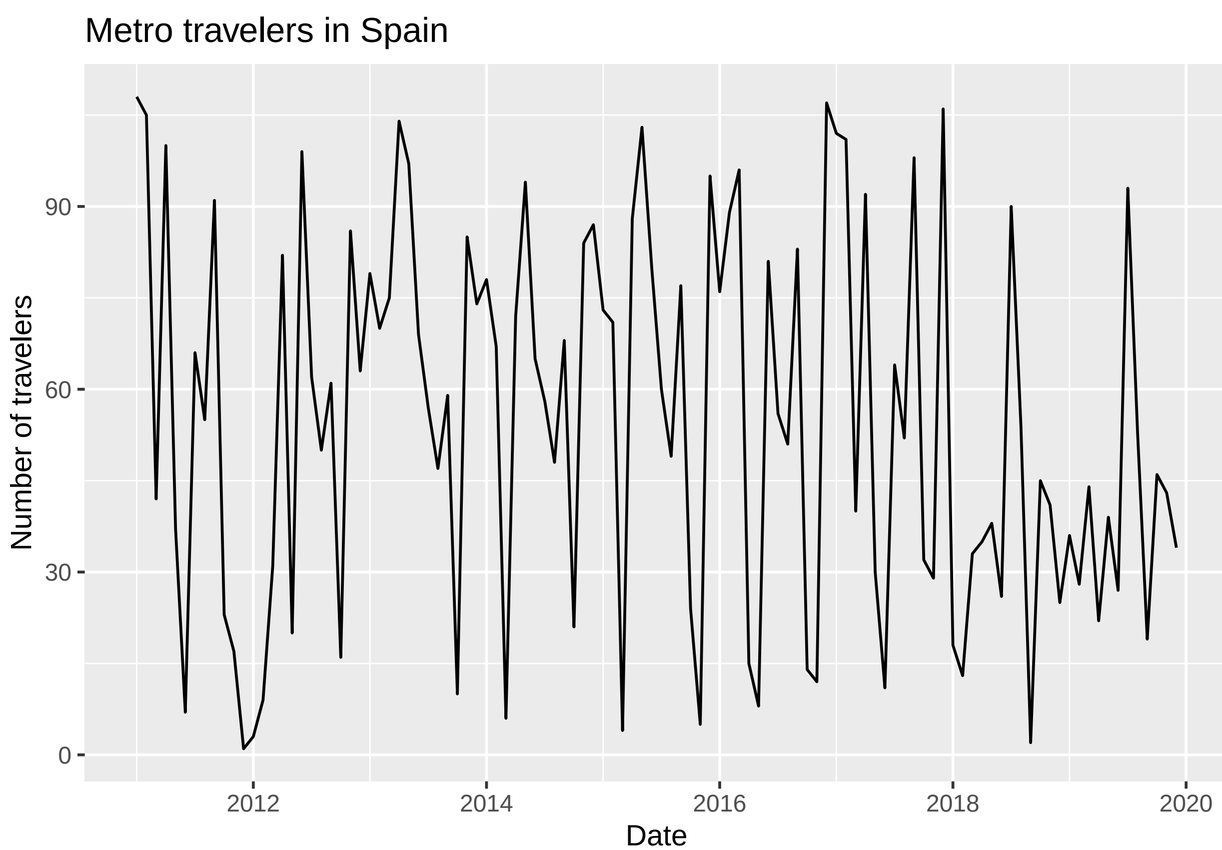
3/3/2022

Introduction

This dataset was downloaded from the spanish National Institute of Statistics (INE) and it contains the quantities of metro users in Spain, from january 2011 to december 2019.

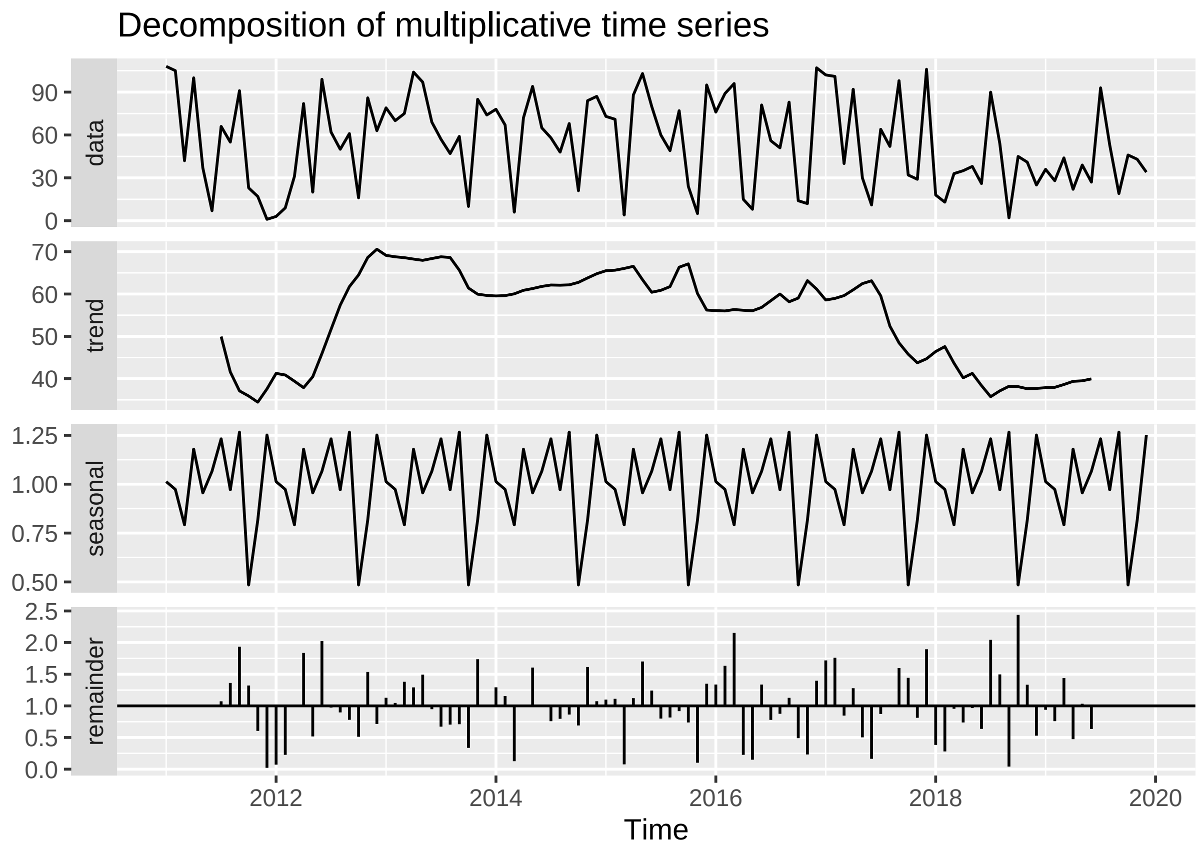
Graphic representation and seasonal decomposition

```
autoplot(viajeros) +
  ggtitle("Metro travelers in Spain") +
  xlab("Date") + ylab("Number of travelers")
```



Here we can see that the data doesn't move too much, this can mean that our serie is stationary.

```r
viajeros_Comp <- decompose(viajeros, type = c("multiplicative"))

autoplot(viajeros_Comp, ts.colour = "blue")
```

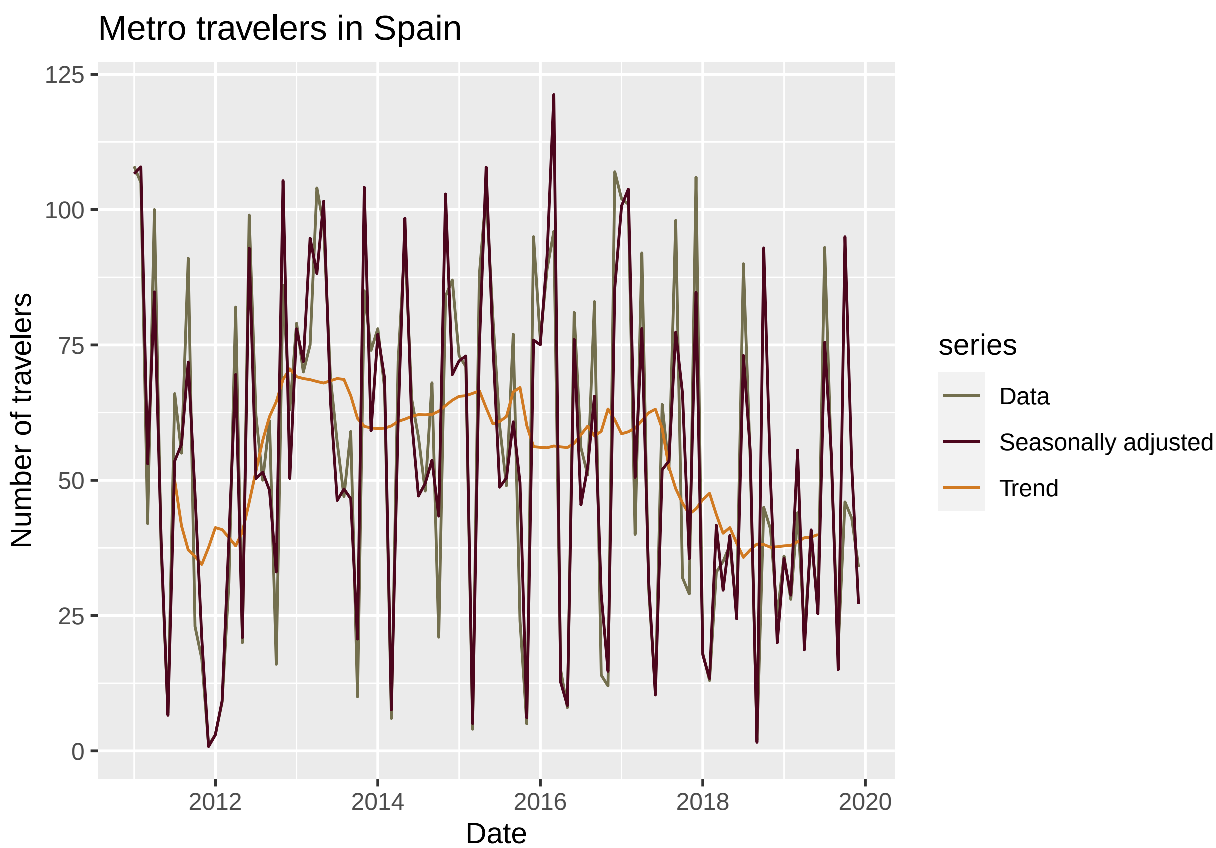


Here we can see that the trend does move, and we clearly have seasonality.

Then we represent the serie with trend and the seasonally adjusted serie:

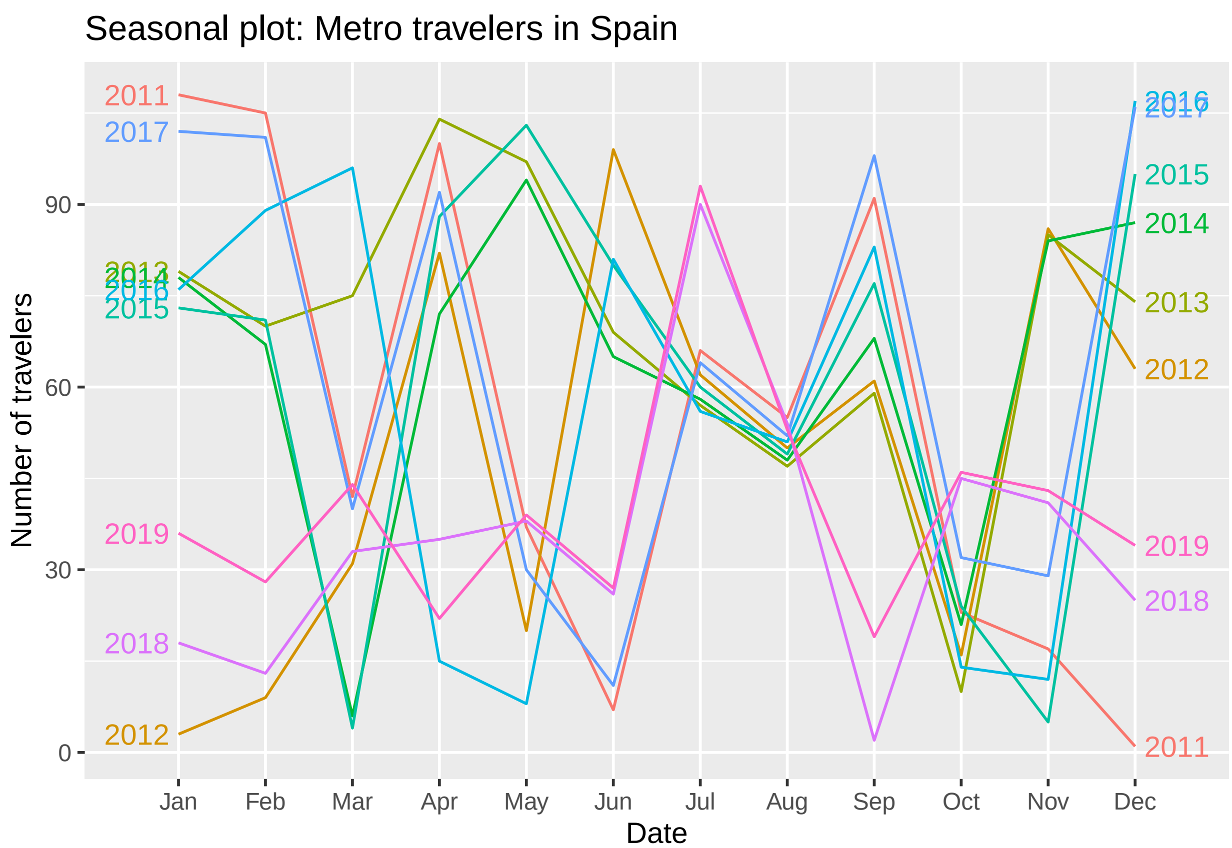
```r
autoplot(viajeros, series = "Data") +
  autolayer(trendcycle(viajeros_Comp), series = "Trend") +
  autolayer(seasadj(viajeros_Comp), series = "Seasonally adjusted") +
  xlab("Date") + ylab("Number of travelers") +
  ggtitle("Metro travelers in Spain") +
  scale_colour_manual(
    values = c("#736F4E", "#4C061D", "#D17A22"),
    breaks = c("Data", "Seasonally adjusted", "Trend")
    )
```

```
## Warning: Removed 12 row(s) containing missing values (geom_path).
```



Then we see the seasonal representation:

```
seasonplot <- ggseasonplot(viajeros, year.labels = TRUE, year.labels.left = TRUE) +
  ylab("Number of travelers") + xlab("Date") +
  ggtitle("Seasonal plot: Metro travelers in Spain")

seasonplot$labels$group <- "Year"
seasonplot$labels$colour <- "Year"

seasonplot
```



There are similarities between years in terms of shape, but not too evident.

Dataset partition

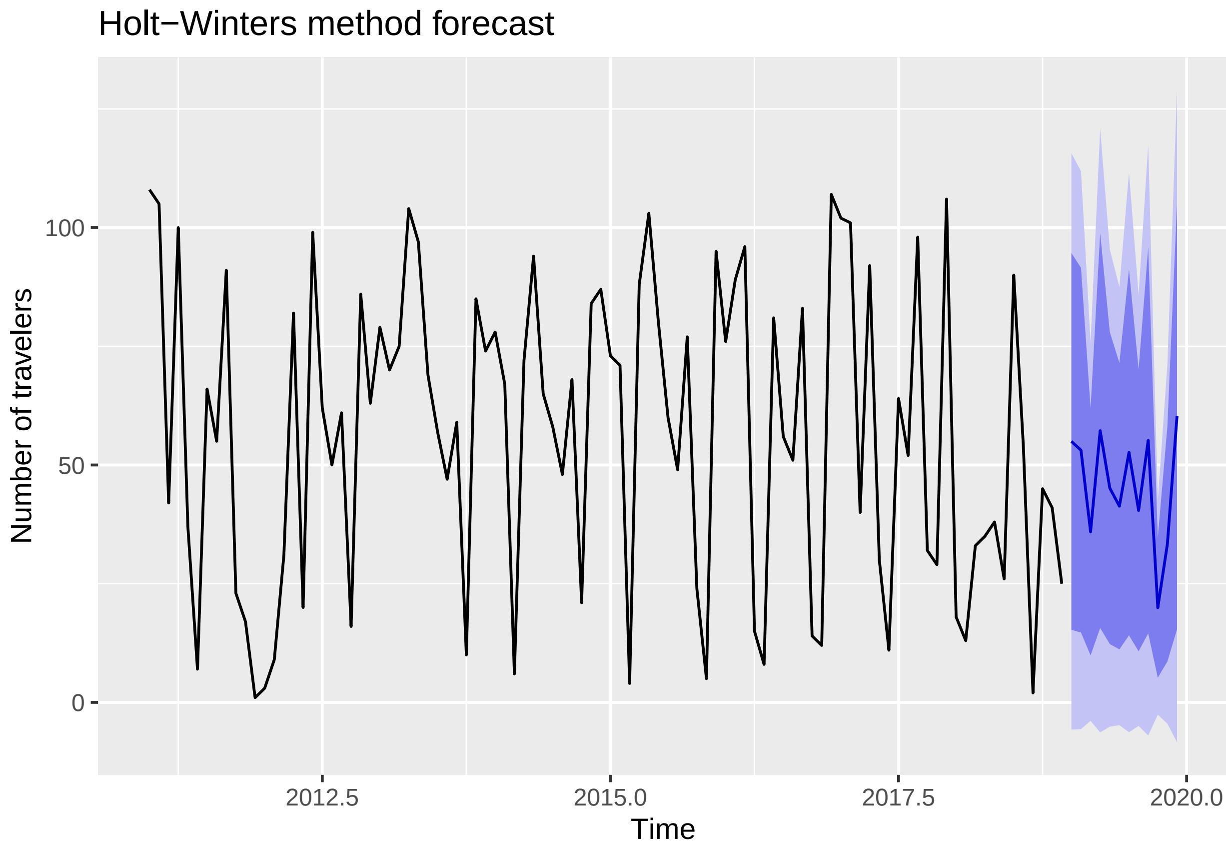
In order to check the accuracy of the forecast methods we are partitioning the dataset to compare forecast and actual events.

```
# We leave out the data corresponding to the last 12 months.

reservados <- 12

viajeros_mod <- viajeros[1:(nrow(viajeros) - reservados),]
viajeros_test <- viajeros[(nrow(viajeros) - reservados):(nrow(viajeros)),]

viajeros_mod <- ts(viajeros_mod, start = c(2011, 1), frequency = 12)
```

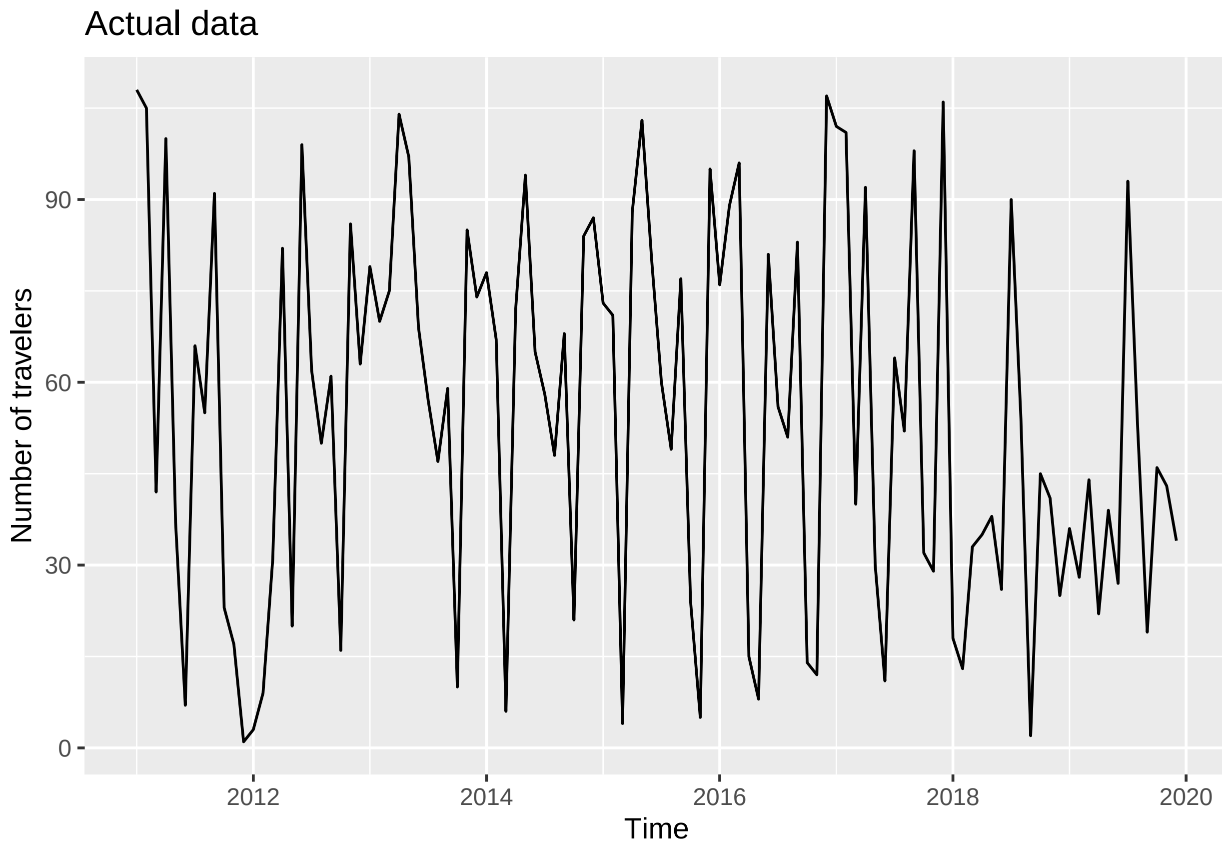
Search for the right time series behaviour model

We chose the Holt-Winters multiplicative model since it's better suited for seasonal series.

```
fit1 <- hw(viajeros_mod, h = reservados, seasonal = "multiplicative", level = c(80, 95))

autoplot(fit1) +
  ggtitle("Holt-Winters method forecast") +
  ylab("Number of travelers") + xlab("Time")
```




```
autoplot(viajeros) +
  ggtitle("Actual data") +
  ylab("Number of travelers") + xlab("Time")
```



Then we are going to try to obtain a better forecast using ARIMA models.

Correlogram representation

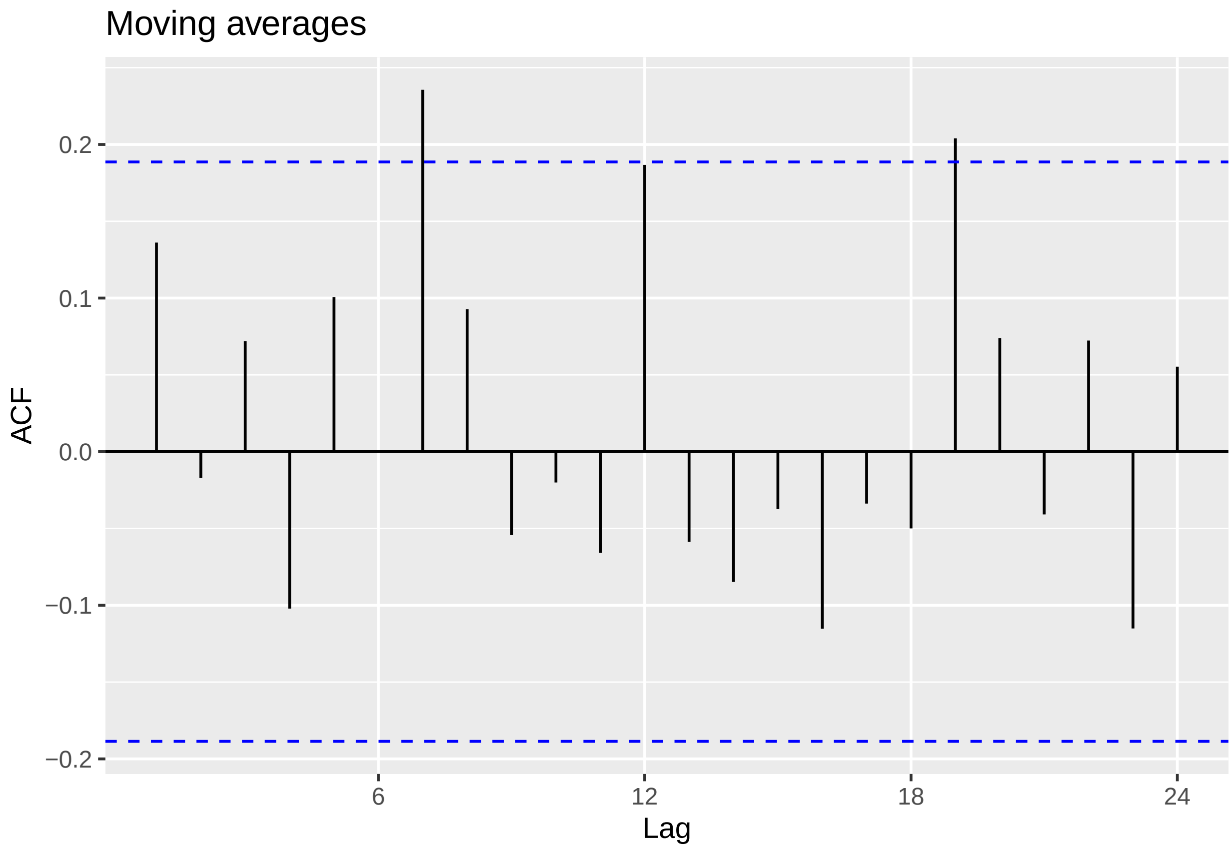
First, we are going to adjust the right model while also checking that its residuals are not correlated.

```
# We make the Dickey-Fuller test to see if we in fact have a seasonal serie:
adf.test(viajeros_mod, alternative = "stationary")
```

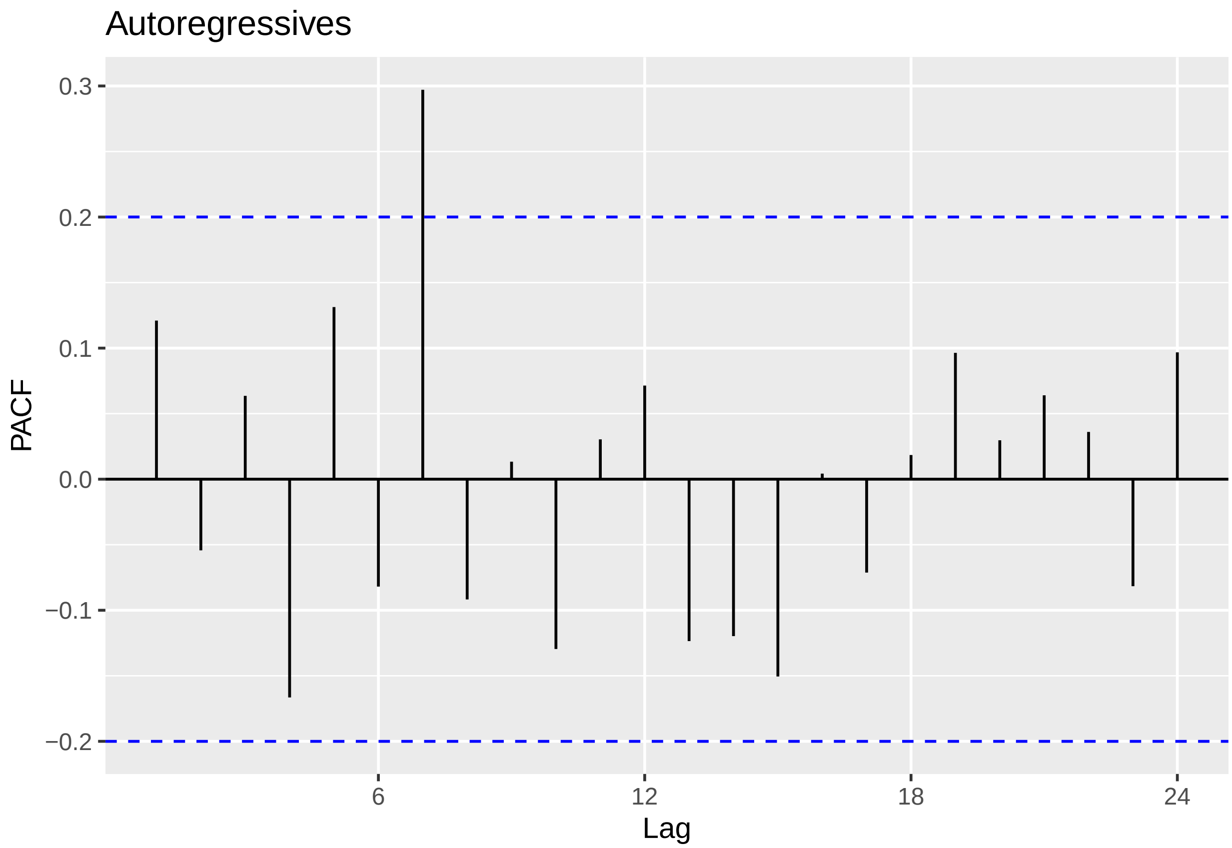
```
##
##  Augmented Dickey-Fuller Test
##
## data:  viajeros_mod
## Dickey-Fuller = -3.8934, Lag order = 4, p-value = 0.01753
## alternative hypothesis: stationary
```

The Dickey Fuller test tells us that the serie we have is seasonal (the P-value is smaller than 0.05), so we don't need to adjust the serie. (0 differences)

```
# ARIMA: (Autoregresivo, diferencias, medias móviles)

# Moving averages:
ggAcf(viajeros) + ggtitle("Moving averages")
```




```
# Autoregressives:
ggPacf(viajeros_mod) + ggtitle("Autoregressives")
```

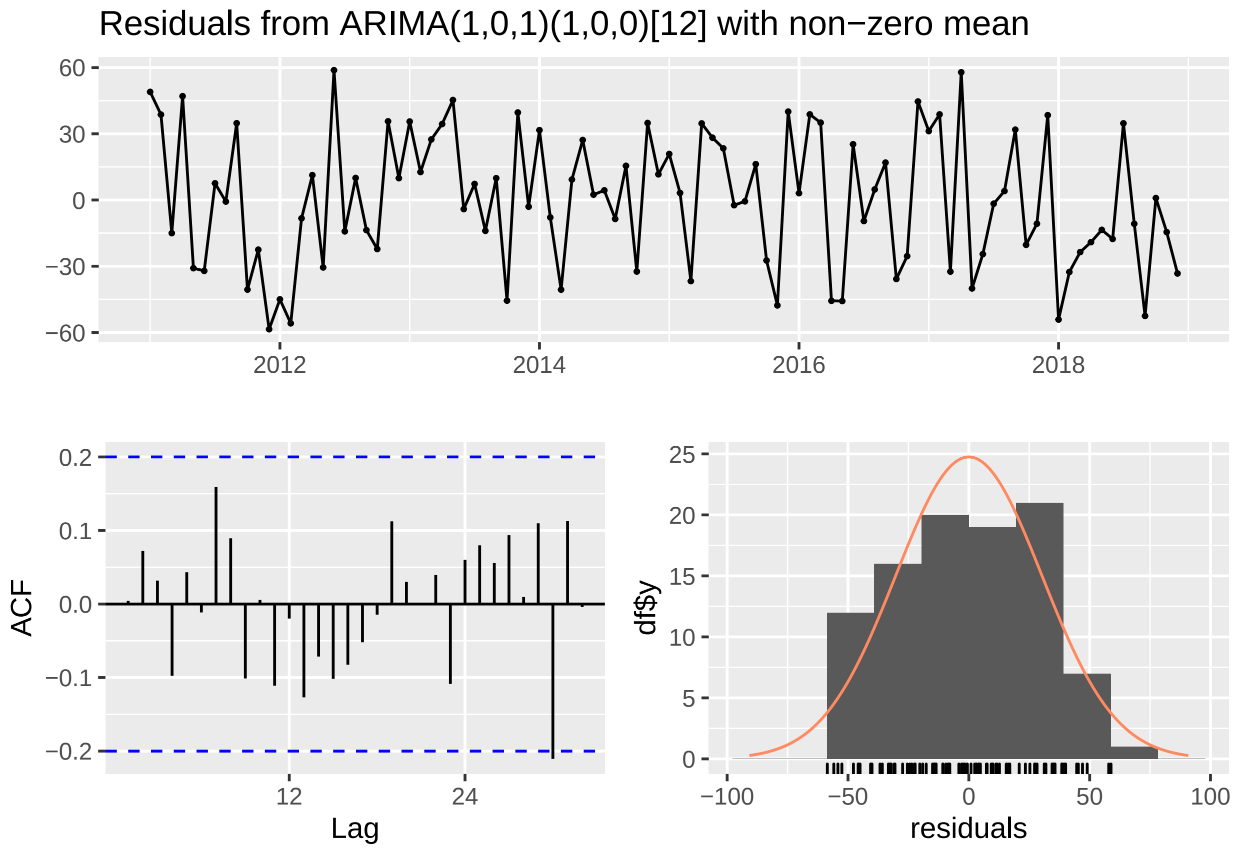


We can see that we have 1 autoregressive and 1 moving average. This means our ARIMA model is (1, 0, 1)(1, 0, 0)[12].

```
modelo <- arima(viajeros_mod, order = c(1, 0, 1), seasonal = c(1, 0, 0))
modelo
```

```
##
## Call:
## arima(x = viajeros_mod, order = c(1, 0, 1), seasonal = c(1, 0, 0))
##
## Coefficients:
##           ar1     ma1    sar1  intercept
##       -0.6440  0.8902  0.1854    55.6409
## s.e.   0.1278  0.0733  0.1101     4.2382
##
## sigma^2 estimated as 906.9:  log likelihood = -463.47,  aic = 936.94
```

```
checkresiduals(modelo)
```



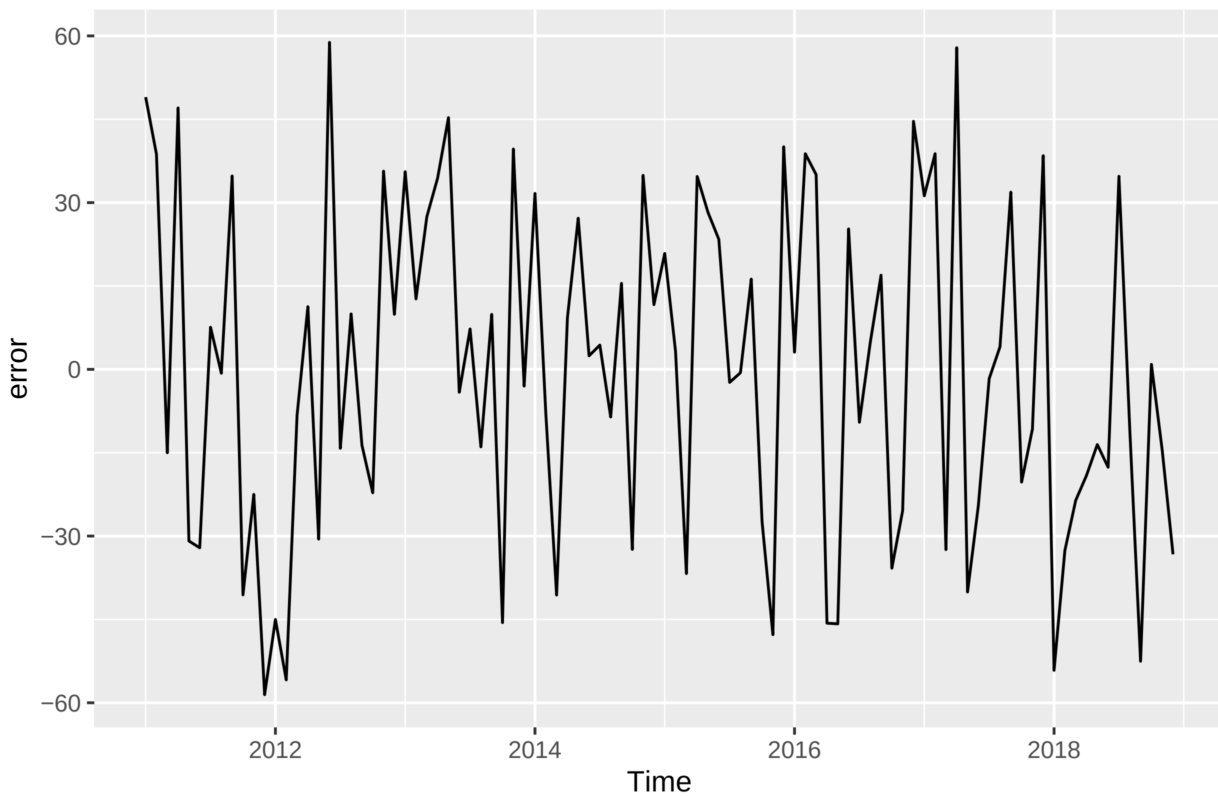

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,1)(1,0,0)[12] with non-zero mean
## Q* = 14.159, df = 15, p-value = 0.5135
##
## Model df: 4.   Total lags used: 19
```

The P-value of the Ljung-Box test is bigger than 0.05, and this means that the model is well-adjusted.

Also, we can see in the residuals graphic that we have a pattern that is similar to white noise, this means we have no correlation between residuals.

Diagnosis:

```
error = residuals(modelo)
# We check that the average of the error is near zero:
autoplot(error)
```



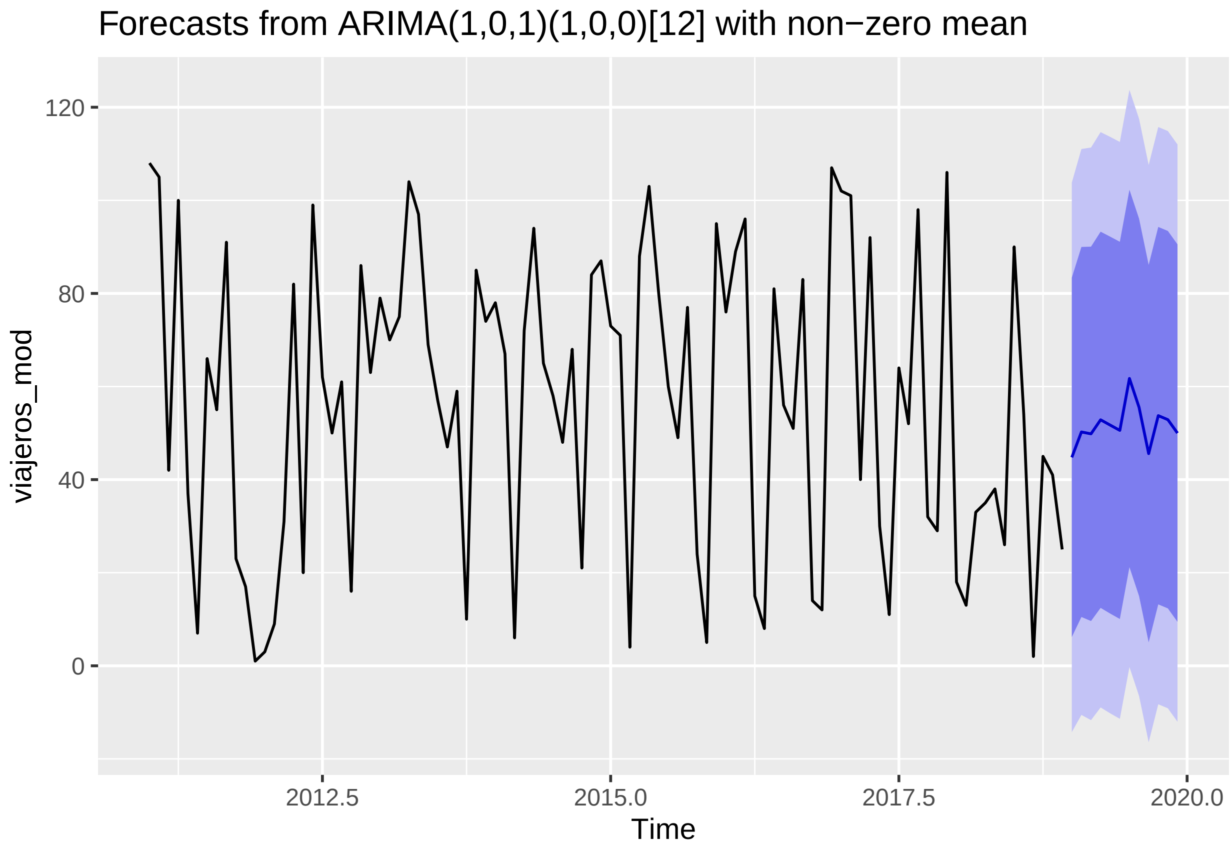
Graphically, the average has the appearance of being zero.

Forecasting with ARIMA:

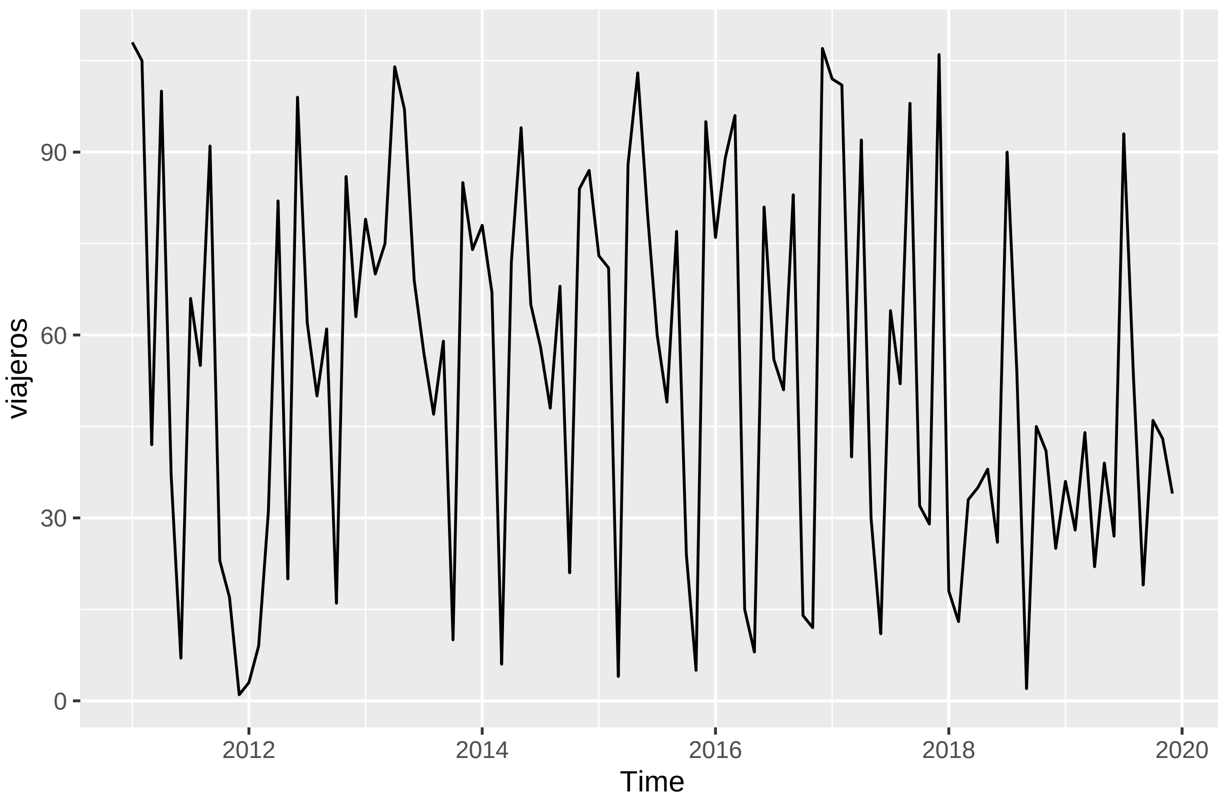
```
pronostico <- forecast(modelo, h = reservados)
pronostico
```

```
##          Point Forecast      Lo 80     Hi 80       Lo 95    Hi 95
## Jan 2019       44.77939  6.185023  83.37375 -14.2455992 103.8044
## Feb 2019       50.23657 10.489641  89.98350 -10.5511139 111.0243
## Mar 2019       49.83281  9.617516  90.04811 -11.6711768 111.3368
## Apr 2019       52.85159 12.443616  93.25956  -8.9470725 114.6503
## May 2019       51.70237 11.214753  92.19000 -10.2180985 113.6228
## Jun 2019       50.57600 10.055389  91.09661 -11.3949268 112.5469
## Jul 2019       61.73370 21.199418 102.26799  -0.2581376 123.7255
## Aug 2019       55.51517 14.975210  96.05513  -6.4853476 117.5157
## Sep 2019       45.58141  5.039102  86.12372 -16.4227000 107.5855
## Oct 2019       53.74220 13.198919  94.28549  -8.2633998 115.7478
## Nov 2019       52.87895 12.335255  93.42263  -9.1272778 114.8852
## Dec 2019       49.99105  9.447197  90.53491 -12.0154253 111.9975
```

```
autoplot(pronostico)
```




```
autoplot(viajeros)
```



Conclusion:

Through the ARIMA model, we have forecasted a similar pattern to the actual data, and even if the amplitude of the variations is not exact, the waveshape and frequency are very similar, and the actual data is between the marked error margins.