Time Series Forecasting for SFO Air Passengers

2023-06-14

# Load Packages

library(fpp3)

## ── Attaching packages ────────────────────────────────────────────── fpp3 0.5 ──

## ✔ tibble 3.2.1 ✔ tsibble 1.1.3  
## ✔ dplyr 1.1.2 ✔ tsibbledata 0.4.1  
## ✔ tidyr 1.3.0 ✔ feasts 0.3.1  
## ✔ lubridate 1.9.2 ✔ fable 0.3.3  
## ✔ ggplot2 3.4.2 ✔ fabletools 0.3.3

## ── Conflicts ───────────────────────────────────────────────── fpp3\_conflicts ──  
## ✖ lubridate::date() masks base::date()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ tsibble::intersect() masks base::intersect()  
## ✖ tsibble::interval() masks lubridate::interval()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ tsibble::setdiff() masks base::setdiff()  
## ✖ tsibble::union() masks base::union()

library(tsibble)  
library(readr)  
library(tseries)

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(knitr)  
library(officedown)

# Load data from SFO air passenger .csv file

## Rows: 50,730  
## Columns: 12  
## $ operating\_airline <chr> "ATA Airlines", "ATA Airlines", "ATA Airli…  
## $ operating\_airline\_iata\_code <chr> "TZ", "TZ", "TZ", "AC", "AC", "CA", "CA", …  
## $ published\_airline <chr> "ATA Airlines", "ATA Airlines", "ATA Airli…  
## $ published\_airline\_iata\_code <chr> "TZ", "TZ", "TZ", "AC", "AC", "CA", "CA", …  
## $ geo\_summary <chr> "Domestic", "Domestic", "Domestic", "Inter…  
## $ geo\_region <chr> "US", "US", "US", "Canada", "Canada", "Asi…  
## $ activity\_type\_code <chr> "Deplaned", "Enplaned", "Thru / Transit", …  
## $ price\_category\_code <chr> "Low Fare", "Low Fare", "Low Fare", "Other…  
## $ terminal <chr> "Terminal 1", "Terminal 1", "Terminal 1", …  
## $ boarding\_area <chr> "B", "B", "B", "B", "B", "G", "G", "A", "A…  
## $ passenger\_count <dbl> 27271, 29131, 5415, 35156, 34090, 6263, 55…  
## $ year <date> 2005-07-01, 2005-07-01, 2005-07-01, 2005-…

# Data Exploration

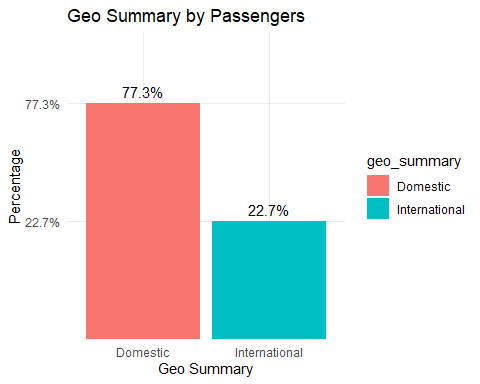
#### Air Passengers Traffic at SFO

#### Air Passengers Traffic at SFO by Activity Type

#### Air Passengers Traffic at SFO by Geography

## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette with 3 different levels  
  
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette with 3 different levels

#### Geo Summary by Passengers



## Selecting by percent

#### Top 10 Operating Airlines by Percentage

**Table** **:** sfo\_top\_passenger table

| operating\_airline | total | percent |
| --- | --- | --- |
| United Airlines | 338.15071 | 30.1% |
| SkyWest Airlines | 95.37713 | 8.5% |
| United Airlines - Pre 07/01/2013 | 89.15942 | 7.9% |
| American Airlines | 80.36718 | 7.2% |
| Delta Air Lines | 79.17331 | 7% |
| Southwest Airlines | 73.45913 | 6.5% |
| Virgin America | 66.76843 | 5.9% |
| Alaska Airlines | 52.80735 | 4.7% |
| JetBlue Airways | 30.02645 | 2.7% |
| US Airways | 18.45011 | 1.6% |

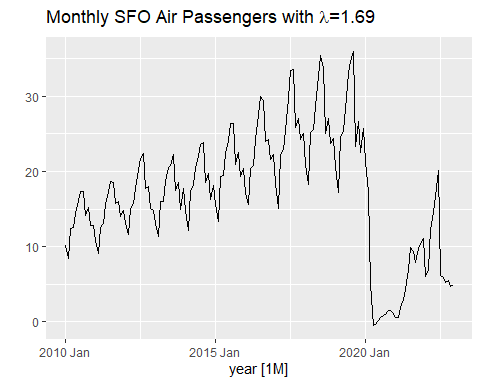
# Data Transformation

#### Convert to tisbble format

**Table** **:** sfo\_passenger table

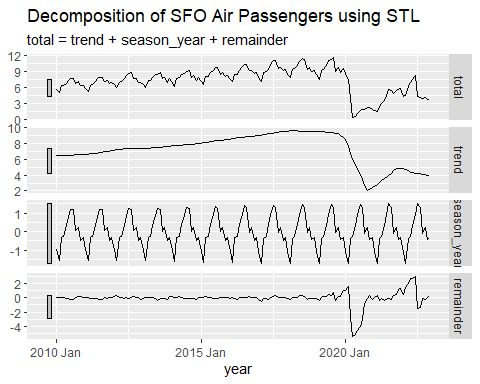
| year | total |
| --- | --- |
| 2010 Jan | 5.570932 |
| 2010 Feb | 5.030722 |
| 2010 Mar | 6.211916 |
| 2010 Apr | 6.278118 |
| 2010 May | 6.760710 |
| 2010 Jun | 7.225772 |

#### Box-Cox transformation for lambda



Note: We tried implementing Box-Cox transformation, but it did not have much impact on the final results, hence while applying final model fitting we used non transformed data.

#### Decomposition using STL



#### KPSS test for differencing

| kpss\_stat | kpss\_pvalue |
| --- | --- |
| 0.6012911 | 0.02251899 |

| ndiffs |
| --- |
| 1 |

p-value (0.02) is significant. Reject null hypothesis. It indicate that the series is not stationary. Therefore a number difference required to make a time series stationary.

#### Dickey-Fuller test

## [1] 0.3541204

#### Split train and test set

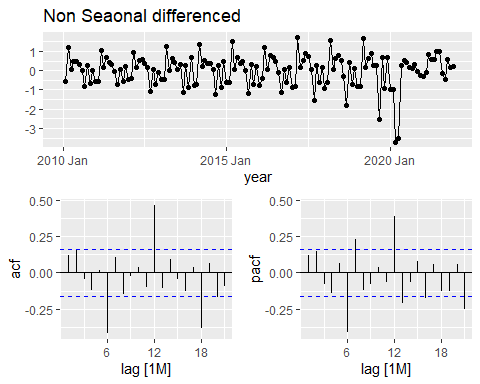
## [1] "Train Min: 2010 Jan Train Max: 2021 Dec"

## [1] "Test Min: 2022 Jan Test Max: 2022 Dec"

#### Examine ACF and PACF plots for the differenced

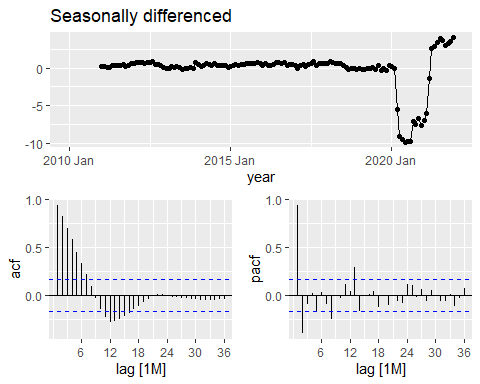
## Warning: Removed 1 row containing missing values (`geom\_line()`).

## Warning: Removed 1 rows containing missing values (`geom\_point()`).



## Warning: Removed 12 rows containing missing values (`geom\_line()`).

## Warning: Removed 12 rows containing missing values (`geom\_point()`).



# Model Fitting

#### Fit the models (ets, arima)

| Model name | Orders |
| --- | --- |
| Ets\_auto | <ETS(A,Ad,A)> |
| Ets\_ses | <ETS(A,A,N)> |
| Ets\_hw\_mul | <ETS(M,A,M)> |
| Ets\_damped\_add | <ETS(A,Ad,A)> |
| Ets\_damped\_mul | <ETS(A,Ad,M)> |
| Arima\_stepwise | <ARIMA(1,0,2)(2,0,0)[12] w/ mean> |
| Arima\_search | <ARIMA(1,0,2)(2,0,0)[12] w/ mean> |
| Arima\_311 | <ARIMA(3,1,1)(2,0,0)[12]> |
| Arima\_410 | <ARIMA(4,1,0)(0,0,2)[12]> |
| Arima\_012 | <ARIMA(0,1,2)(0,0,2)[12]> |
| Arima012011 | <ARIMA(0,1,2)(0,1,1)[12]> |
| Arima210011 | <ARIMA(2,1,0)(0,1,1)[12]> |
| Arima011011 | <ARIMA(0,1,1)(0,1,1)[12]> |
| Arima212011 | <ARIMA(2,1,2)(0,1,1)[12]> |
| Arima210111 | <ARIMA(2,1,0)(1,1,1)[12]> |

# Model Evaluation Metrics

#### Accuracy measures of the forecast

## Warning: The future dataset is incomplete, incomplete out-of-sample data will be treated as missing.   
## 24 observations are missing between 2023 Jan and 2024 Dec

| .model | .type | RMSE | MAE | MAPE |
| --- | --- | --- | --- | --- |
| Ets\_ses | Test | 1.717613 | 1.634137 | 35.92445 |
| Ets\_damped\_mul | Test | 1.795588 | 1.673425 | 37.11414 |
| Arima011011 | Test | 1.800386 | 1.659921 | 35.86206 |
| Arima012011 | Test | 1.813586 | 1.674685 | 36.15545 |
| Arima210011 | Test | 1.820583 | 1.679460 | 36.38118 |
| Arima210111 | Test | 1.923835 | 1.738153 | 38.92653 |
| Arima212011 | Test | 1.985007 | 1.800789 | 40.93028 |
| Ets\_auto | Test | 2.191658 | 1.918942 | 44.33543 |
| Ets\_damped\_add | Test | 2.191658 | 1.918942 | 44.33543 |
| Arima\_012 | Test | 2.637478 | 1.853172 | 29.97824 |
| Arima\_410 | Test | 2.671977 | 1.882842 | 30.48522 |
| Arima\_search | Test | 2.733989 | 1.847116 | 29.07042 |
| Arima\_stepwise | Test | 2.733989 | 1.847116 | 29.07042 |
| Arima\_311 | Test | 3.079463 | 2.311271 | 39.14233 |
| Ets\_hw\_mul | Test | 3.311143 | 2.515613 | 63.20995 |

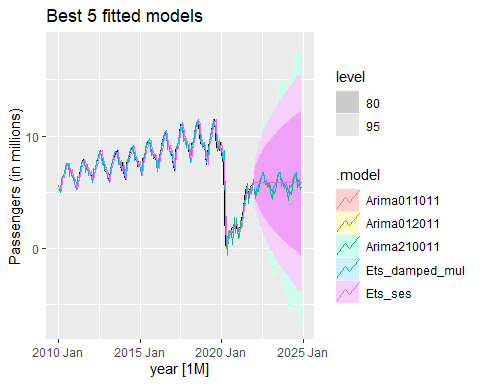
* RMSE (Root Mean Squared Error): measures the average difference between the forecasted and the actual values, taking into account the squared differences.
* The lowest RMSE indicates better accuracy.

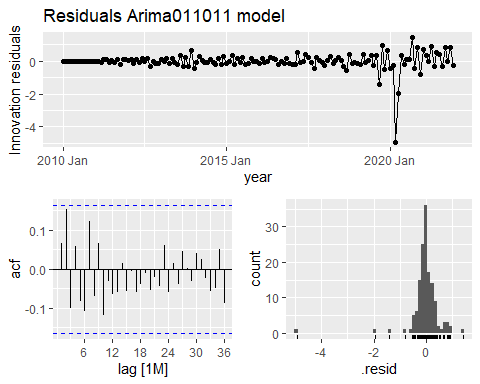
#### Summary report for fitted time series models

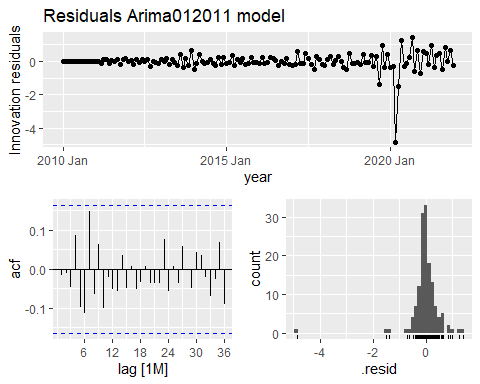
## Warning in report.mdl\_df(selected\_models): Model reporting is only supported  
## for individual models, so a glance will be shown. To see the report for a  
## specific model, use `select()` and `filter()` to identify a single model.

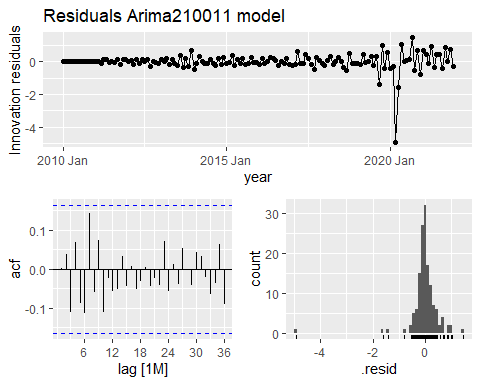
| .model | AIC | AICc | BIC |
| --- | --- | --- | --- |
| Arima012011 | 250.7273 | 251.0448 | 262.2281 |
| Arima210011 | 252.4017 | 252.7191 | 263.9025 |
| Arima011011 | 253.9418 | 254.1308 | 262.5674 |
| Ets\_damped\_mul | 585.9072 | 591.3792 | 639.3638 |
| Ets\_ses | 670.4162 | 670.8510 | 685.2653 |

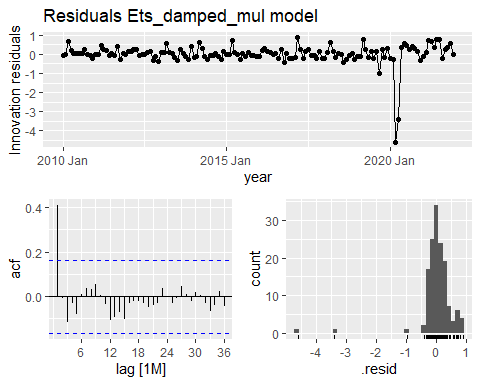
* AICc (Akaike Information Criterion corrected): a lowest AICc value indicates a best-fitting model.

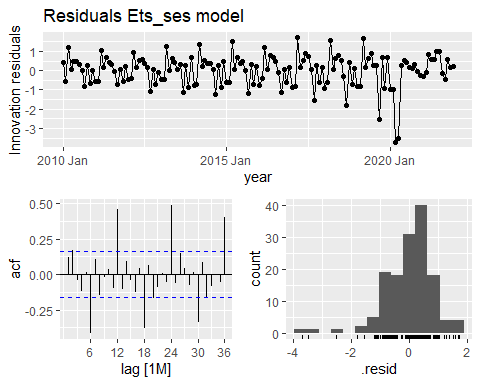












#### Ljung-Box test

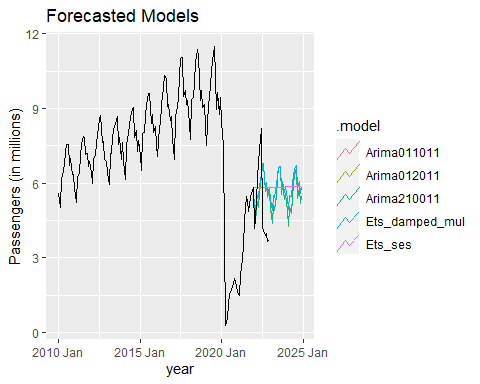
## Series: total   
## Model: ARIMA(0,1,2)(0,1,1)[12]   
##   
## Coefficients:  
## ma1 ma2 sma1  
## 0.4213 0.1963 -0.8873  
## s.e. 0.0927 0.0827 0.1489  
##   
## sigma^2 estimated as 0.3329: log likelihood=-121.36  
## AIC=250.73 AICc=251.04 BIC=262.23

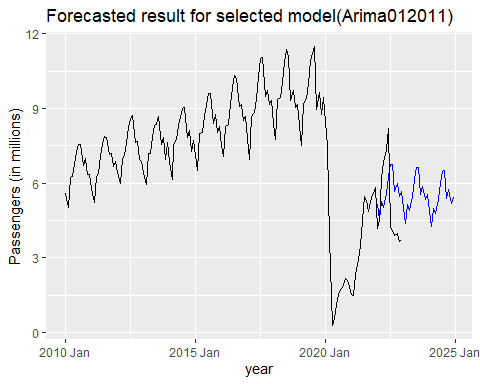
| .model | lb\_stat | lb\_pvalue |
| --- | --- | --- |
| Arima012011 | 11.14494 | 0.1324204 |

For non-seasonal component, ARIMA(0,1,2) has 2 MA terms, for seasonal ARIMA(0,1,1) has 0 MA term estimated. Therefore the degree of freedom should be 5.

Since the p-value > 5%, not reject H0 at a significant level. The residuals could be considered white noise series and the model has adequately captured the underlying data.

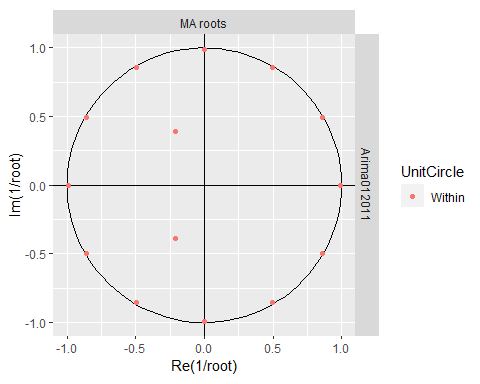
#### Model selection and forecast





## [1] 5.155015 4.517561 5.251238 5.018928 5.594702 6.259479 6.739579 6.765069  
## [9] 5.652866 5.992561 5.458187 5.673563 4.981288 4.387225 5.120902 4.888592  
## [17] 5.464365 6.129142 6.609243 6.634733 5.522530 5.862225 5.327850 5.543226  
## [25] 4.850952 4.256889 4.990566 4.758255 5.334029 5.998806 6.478907 6.504396  
## [33] 5.392194 5.731889 5.197514 5.412890

#### Inverse root test



This inverse toot test implies that the model is stable and a good fit to the data. Also the model has capturing the underlying patterns of the time series.