# Time Series Analysis Report

## Air Traffic in France

**Time Series Analysis Course**

**Dr. Clément Chevalier**

Assistant Ziqing Dong

**Written Report by:**

Isa Castro

Anaïs Zodeougan-Quist

# Contents

# 1 Introduction

Since the 1970's, the air passengers traffic has steadily increased. In France,the air passengers traffic is mostly carried out by the two principal airports : Paris-Charles de Gaulle (CDG) and Paris-Orly (ORY). Together, the airports Paris Charles de Gaulles and Paris Orly represent more than 50% of the air passengers traffic (Union des Aéroports Français, 2017).
However this upward trend has sometimes been disrupted due to some major political, social and economic havocs such as wars, social strikes, and economic crisis that can occur at worldwide level. The French air passengers traffic has not been spared, firstly during the Gulf War in 1991, few years later, with the long-term strikes from Air France flight crew, in 2001 the World Trade Center terrorist attacks and in 2009 the financial crisis.
The purpose of our project is to predict how air passengers traffic in France will behave within the next 5 years , by focusing on the data of the past 20 years from both French principal airports. Thereby, the first section will provide as thoroughly as possible a description of the dataset and then, we will attempt to find the best model, most likely amongst SARIMA and ARIMA processes that will allow us to explain and provide realistic forecasts.

# 2 Descriptive analysis

The present section will detail how the data were provided, their containts and a general statistical description.

## 2.1 Data Collection

The dataset used for our analysis is provided by Groupe ADP - Aéroports de Paris website [1]. It gathers the air traffic monthly data from January 2000 and December 2018. The original dataset contains the monthly data passenger for both Parisian airports (Paris - Charles de Gaulle and Paris - Orly). Which represents a total of 228 data for each airport, and no missing values. As it is referred in the introduction, the air

| Year | Passengers (in millions) | Year | Passengers (in millions) |
|---|---|---|---|
| 2000 | 73642.66 | 2010 | 83370.00 |
| 2001 | 71025.91 | 2011 | 88109.63 |
| 2002 | 71531.34 | 2012 | 88844.20 |
| 2003 | 70677.47 | 2013 | 90327.07 |
| 2004 | 75313.58 | 2014 | 92676.34 |
| 2005 | 78658.84 | 2015 | 95431.98 |
| 2006 | 82471.72 | 2016 | 97171.01 |
| 2007 | 86362.91 | 2017 | 101513.92 |
| 2008 | 87084.38 | 2018 | 105350.41 |
| 2009 | 83014.56 | | |

Table 1: Evolution of the annual air passengers traffic in France, 2000-2018

passengers traffic has increased over the past decades. In France, the latter has gradually expanded passing from 7.3 millions of passengers in 2000 up to 10.5 millions of passengers(cf.Table 1). However, since we are interested in predicting the monthly air traffic from Parisian airports, we have decided to keep only the column which contains the total of passengers and to adjust the scale of measurement in terms of thousand of millions of passengers.

[1] `https://www.parisaeroport.fr/en/group/finance/investor-relations/traffic`

## 2.2 Data transformation and visualisation

In order to start our analysis, we have proceeded to the transformation of the data class into a time series class, by means of the `ts()` function. The plot below, shows the evolution of the monthly air traffic in France beginning from January, 2000 and ending in December, 2018. Later, in our analysis we will perform a log transformation of our data, prior to our attempts in removing the trend and the seasonality. The reason is that it is common practice (Dettling, 2013), while dealing with positive values and a quadratric trend, to log-transform the data; further it tends to diminish the effect of the variance.
It appears that at each periodicity we observe a peak in the number of passengers, followed by a decrease and then a trough.

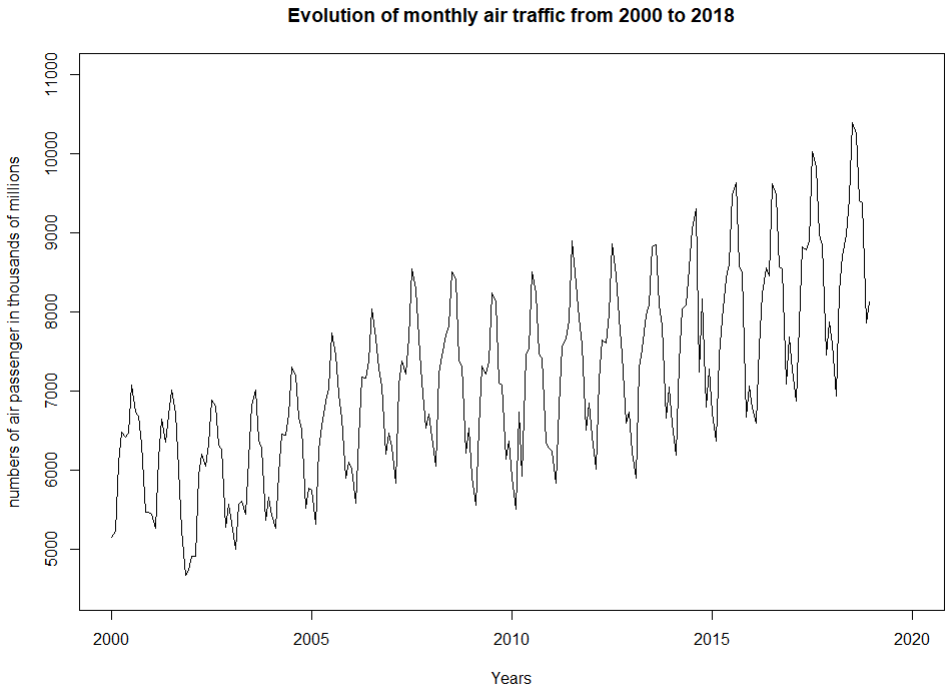


Figure 1: Monthly air passengers traffic in France (Raw data)

Although around the years 2009-2010 we see a sligth collapse, we can observe from year-to-year an overall increase in the number of passengers transported by air. This may denote of an upward trend in the data.

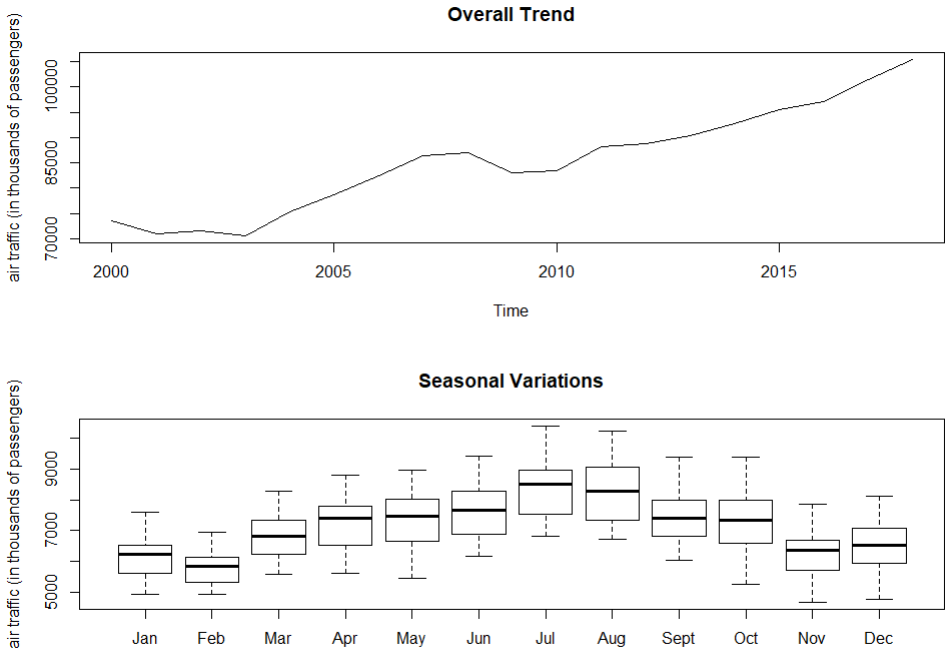


Figure 2: Trend and seasonality

The first part of the Figure 2, emphasizes the overall variations over the years in our dataset. As suggested previously, two noticeable troughs stand out; between the years 2001-2003 and 2008-2010. Regarding the seasonal changes (bottom Figure 2), it confirms the general overview from Figure 1 that some month observes systematically decreases versus the opposite for some other months. Indeed, the months of July and August observe a clear peak of the air traffic, followed by a decrease between the months of September and November. Until a slight recovery from December to January, and so on and so forth. At this point we may be definite concerning the presence of an additive seasonality in our data and a trend (more or less)linear. Other 'informal' procedures to detect the seasonality, is the visualisation of the correlogram (cf. Appendix7).

# 3 Model analysis

In this section we will explore different methods to remove the trend and the seasonality in our data, aiming to obtain a stationary time series for which it will be possible to fit an appropriate model (Himakireeti and Vishnu, 2019). From now on and we will no longer use the log-transformation of the data.

## 3.1 Removing the trend and seasonality

Following the classical model decomposition model (Brockwell et al., 2002; Bourbonnais and Terraza, 2016), we write our seasonal log-additive model as follow, $log(X_t) = log(m_t) + log(S_t) + log(Y_t)$, Where $log(Y_t)$ is the residual time series, $log(m_t)$, the trend and $log(S_t)$ the seasonal component. By trying to remove the trend and seasonality using a linear regression model, the aim is to estimate a polynomial trend and a the seasonal effect of the data (Dettling, 2013). Hence the following equation model,

$$log(X_t) = \beta_0 + \beta_1 t + \alpha_1 + ... + \alpha_{12} + \epsilon_t,$$

where t=1,..,228 and $\alpha$ will be an array of seasonal dummy variables.

```
Call:
lm(formula = log.ts ~ t + as.factor(seasonal.dummy))
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.214934 -0.023525 -0.001848  0.025616  0.087962

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 8.527e+00  1.178e-02 723.755  < 2e-16 ***
t                           1.723e-03  4.667e-05  36.925  < 2e-16 ***
as.factor(seasonal.dummy)2 -5.503e-02  1.503e-02  -3.662 0.000315 ***
as.factor(seasonal.dummy)3  1.048e-01  1.503e-02   6.971  3.8e-11 ***
as.factor(seasonal.dummy)4  1.608e-01  1.503e-02  10.697  < 2e-16 ***
as.factor(seasonal.dummy)5  1.781e-01  1.503e-02  11.848  < 2e-16 ***
as.factor(seasonal.dummy)6  2.103e-01  1.503e-02  13.989  < 2e-16 ***
as.factor(seasonal.dummy)7  3.044e-01  1.503e-02  20.251  < 2e-16 ***
as.factor(seasonal.dummy)8  2.834e-01  1.503e-02  18.851  < 2e-16 ***
as.factor(seasonal.dummy)9  1.879e-01  1.503e-02  12.502  < 2e-16 ***
as.factor(seasonal.dummy)10 1.623e-01  1.503e-02  10.796  < 2e-16 ***
as.factor(seasonal.dummy)11 5.874e-03  1.504e-02   0.391 0.696428
as.factor(seasonal.dummy)12 4.426e-02  1.504e-02   2.944 0.003601 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04632 on 215 degrees of freedom
Multiple R-squared:  0.9252,Adjusted R-squared:  0.921
F-statistic: 221.6 on 12 and 215 DF,  p-value: < 2.2e-16
```

The regression's outputs indicates that all seasonal coefficients are significant except the eleventh one. However the plot of the residuals time series (cf. Appendix6) shows clearly a pattern, which indicates non-stationarity. This is also confirmed by the Box-test and the Augmented Dickey-Fuller (ADF-test) test that have been conducted (see Appendix,6). Therefore to remove the seasonality more efficiently, we performed the Brockwell-Davis method which consists in applying a moving average filter on the data. The algorithm we applied is based on the formula below,

$$\tilde{m}_t = (0.5X_{t-q} + X_{t-q+1} + ... + 0.5X_{t+q})/d,$$

The moving average filter has been adapted to the seasonality of our data as such,

$$\tilde{m}_t = (0.5X_{t-6} + X_{t-5} + ... + 0.5X_{t+6})/12,$$

Then for each month we will compute the centered seasonality coefficients (and replicate them for the whole period), in order to remove those from the entire log time series. Finally the graphic below shows a nice deseasonalized time series. However, it remains few spikes which can be attributed to the remaining error (or remainder in the STL decomposition).

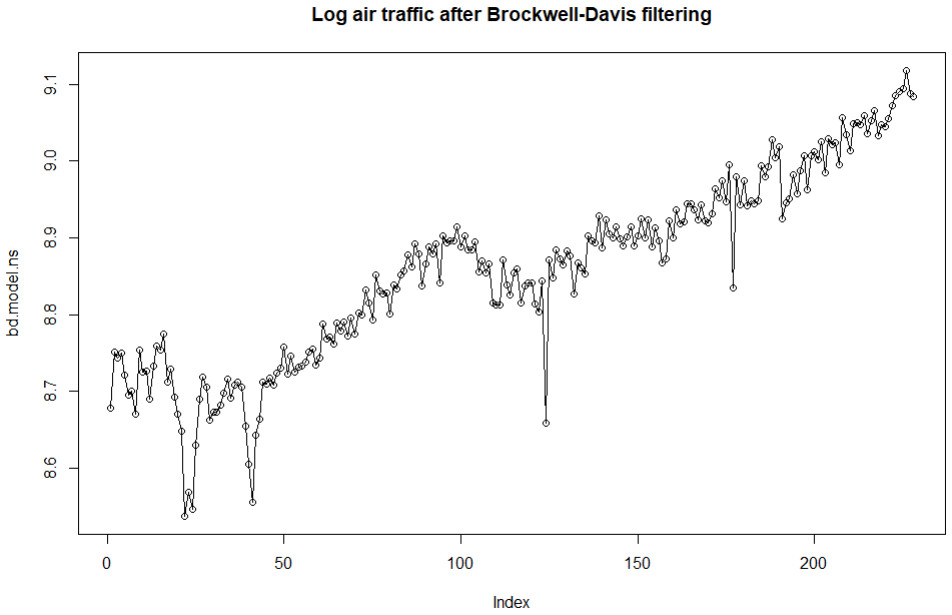


Figure 3: Deseasonalized log air traffic, Brockwell-Davis Method

The so-called STL decomposition (Seasonal and Trend decomposition using Loess) is used to decompose evolving time series (as well as the `decompose()` function)(Dettling, 2013) estimate non-linear models. Comparing the values of the seasonality obtained by the Brockwell-Davis method, to the values obtained from the STL-decomposition allows us to assess the accuracy of the method. Indeed, STL-decomposition is often seen as robust in presence of monthly data.

```
cbind(stl.ts$time.series[1:12, 1:3],w[1:12])
         seasonal    trend    remainder  seasonal.B-Davis
 [1,] -0.13280655 8.717590 -0.036365070  -0.12929347
 [2,] -0.18796764 8.717695  0.032903208  -0.18821807
 [3,] -0.02829052 8.717799  0.025853859  -0.02796463
 [4,]  0.02780393 8.718388  0.031498817   0.02780343
 [5,]  0.04519828 8.718977  0.003205353   0.04670681
 [6,]  0.07773152 8.719963 -0.022830647   0.08032809
 [7,]  0.17221841 8.720949 -0.028357774   0.16410529
 [8,]  0.15156122 8.721660 -0.058617933   0.14409512
 [9,]  0.05650513 8.722371  0.028675261   0.05364449
[10,]  0.03090771 8.723526 -0.001045234   0.02852744
[11,] -0.12547565 8.724681  0.008304226  -0.11828793
[12,] -0.08738585 8.726399 -0.030605866  -0.08144656
```

Here, we see that the difference between the two seasonal coefficients is almost negligeable, and suggest an accurate fit of the Brockwell-Davis method for removing the seasonality. In order to eliminating the trend we have applied the backward shift operator once. It is defined as the first-lag difference applied to our deseasonalized time series $log(\tilde{X}_t)$. In our case, one differenciation has been sufficient. After centering the detrended and deseasonalized data, we obtain a time series with a remains of seasonality (cf. Appendix8). Therefore to be confident regarding the stationarity of our process we have performed several stationarity tests, that will be dicussed in the next section.

## 3.2 Stationarity

In the scope of complex models such as seasonal data, when testing for stationarity, it is not unusual to face tests that deliver ambiguous results (Brockwell et al., 2002). Regarding our analysis, we have performed

4 different tests; Ljung-Box test, and three unit root test, the Augmented Dickey-Fuller-test (ADF), the Phillips-Perron test and the KPSS-test. Except for the Ljung-Box test, all 3 tests state for a the stationarity of the data.
The null hypothesis for the ADF-test and the Phillips-Perron test is that the model has a unit root, which means that there remains a seasonal component. For both these tests, we would want to have a small p-value (¡5%) to decide against the null hypothesis of non-stationarity. Which we obtained.

```
Augmented Dickey-Fuller Test

data:  dYt
Dickey-Fuller = -7.962, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary

Phillips-Perron Unit Root Test

data:  dYt
Dickey-Fuller Z(alpha) = -311.02, Truncation lag parameter = 4, p-value = 0.01
alternative hypothesis: stationary
```

To test the stationarity using KPSS-test, it is recommended to perform an ADF-test because the KPSS test provides a large rate of false positive (type I error), which means that the test will frequently reject the null hypothesis. As we cannot control this parameter without decreasing the test's power.
In the KPSS-test for stationarity, when one specifies the "Level" option, it tests for a leveled stationary of the data. Therefore in this case we would want to have a large p-value (¿5%), to have unsufficient evidence against the null hypothesis, and validate the stationarity of our data.

```
KPSS Test for Level Stationarity

data:  dYt
KPSS Level = 0.022384, Truncation lag parameter = 4, p-value = 0.1
```

### 3.3 Estimates and Forecasts

After the transformations of our original data time series to be stationarized, we will try to fit either a seasonal ARIMA model or an ARIMA, based on the analysis of the shape and spikes of autocorrelation and partial autocorrelation functions shown below.

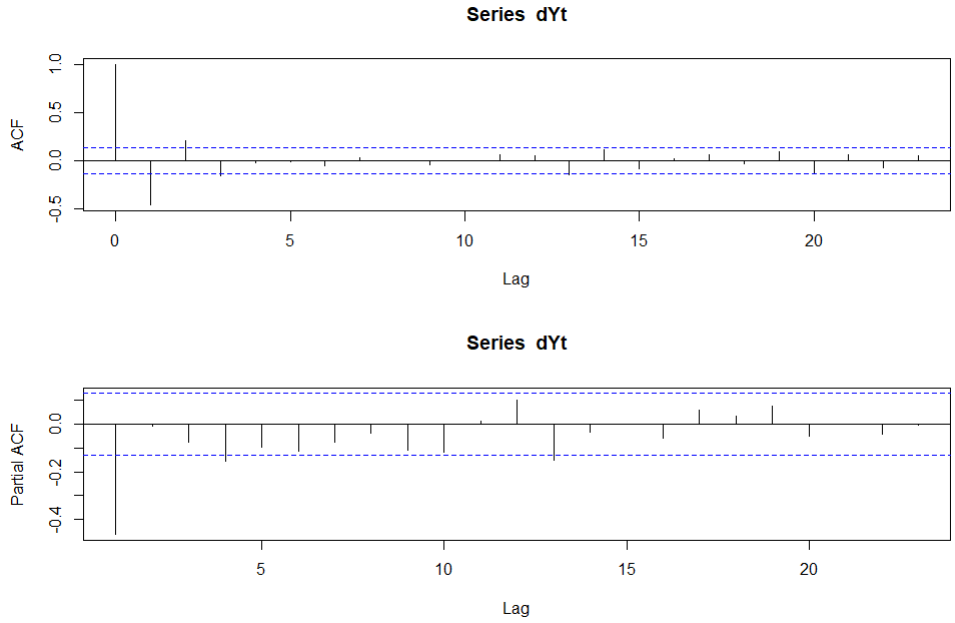


Figure 4: Autocorrelation and Partial Autocorrelation plots of transformed data

Several articles provided an interesting table as a reference to determine the value of p and q for the AR(p) and MA(q) part and for the seasonal model, the rule of thumb will be to focus on the spikes which are multiples of the seasonality of our data (1s, 2s, 3s,...). Given the spikes and the exponential decrease on the

| Model | ACF | PACF |
|---|---|---|
| AR(p) | Spikes decays towards 0 | Spikes cutoff to 0 |
| MA(q) | Spikes cutoff to 0 | Spikes decays towards 0 |
| ARMA(p,q) | Spikes decays towards 0 | Spikes decays towards 0 |

Table 2: Conditions for AR,MA or ARMA

ACF we could guess an AR(3) or AR(4), the PACF suggests that we may have an MA(1). Additionally we did not find any significant spike for the seasonal model. So our model could be an ARIMA(3,0,1) or ARIMA(3,1,1).

```
Best model: ARIMA(4,0,1) with zero mean

Series: dYt
ARIMA(4,0,1) with zero mean

Coefficients:
         ar1     ar2      ar3      ar4      ma1
      0.2798  0.3071  -0.1149  -0.0962  -0.8162
s.e.  0.1048  0.0788   0.0705   0.0739   0.0845

sigma^2 estimated as 0.001036:  log likelihood=460.12
AIC=-908.23   AICc=-907.85   BIC=-887.68
```

The system outputs has eventually selected an ARIMA(4,0,1). Although this differ from our assumptions, the model appeared to be valid and we base our forecastings.

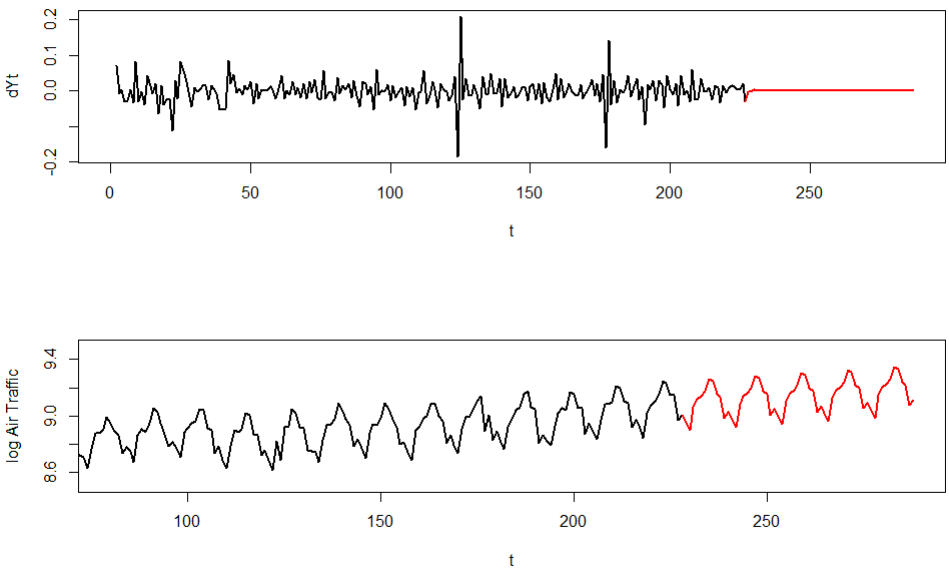


Figure 5: Forecasts of the stationary and the log time series

# 4 Discussion

All the analysis of our data provide some useful outputs at each step that have been necessary to move forward. First the different graphics, gave us information on the trend and the seasonality of our series. Thanks to them, we have seen that the overall trend was not linear increase and there is some dereases between the years 2001-2003 and 2008-2010. These latters correspond exactly to two critical periods: the first one was due to the World Trade Center terrorist attacks in 2001, and the second one to the financial crisis in 2009.
After the WTC attacks, the fear of flying has spread amongst civilians. The financial crisis had disastrous economic consequences such as the increase of the unemployment and in return, the decrease of the 'purchasing power'.Regarding the seasonality, the boxplot indicated clearly that during the months of June to August, the air passengers traffic increases which corresponds to the summer breaks in France.
As for the stationarity, our first attempt to run a linear regression did not provide satisfactory results. Leading to to remaining of the trend and the seasonality. Therefore, we add to explore several methods in order to obtain a valid stationarity. However, for unknown reason to the Ljung-Box test has still provide contrary results. This was one of the aspect that we had not sufficient knowledge to deal with.
Subsquently, visually estimating the parameters p, q, P and Q, was rather difficult (taking into account the appropriate spike was not an obvious task). Therefore we had to rely on the outputs of the auto.arima() function essentially, to select the best model.
Finally, we were satisfied about our forecasts model which seems to follow the tendancy of air passengers traffic.

# 5 Conclusion

Unless major turmoil, overall the air traffic passenger remains steadily increasing over the past decades. Our short analysis has led to an optimistic prediction of the air passengers traffic, showing not much differences with the observed tendency. Indeed, the non-controlled factor of the population development contributes in increasing the travel demand. Furthermore, technological advances, improvements in the air passengers transportation sector as well as the greater availability of aircraft contribute also in the growth of the air traffic.

It is important to remember, that, at the moment, a new societal trend may affect the air traffic in general. Recently, the so called 'flight shame' (or flygskam) and has spread across Western Europe dissuading people from flying unnecessarily. Indeed, mainstream awarness on climate and environmental issues can have a negative impact on the growth of the French air passenger traffic. Yet it is apparent that the air traffic as a whole is very subject to the stability of its macro-environment.
Alongside with the recent series of strikes for the climate, it would therefore be interesting to apprehend to what extend the environmental awareness might affect or not the air traffic over a short period of time in a country such as France. Where societal movements can seriously impact the air traffic.

# Bibliography


Bourbonnais, R. and Terraza, M. (2016). *Analyse des séries temporelles-4e éd.: Cours et exercices corrigés-Applications à l'économie et à la gestion*. Dunod.

Brockwell, P. J., Davis, R. A., and Calder, M. V. (2002). *Introduction to time series and forecasting*, volume 2. Springer.

Chartier, M. and Tounsi, I. (2000). L'évolution du trafic aérien régulier de passagers dans le monde (1987-1997). *L'Information Géographique*, 64(2):148–154.

Dettling, M. (2013). Applied time series analysis. *Applied time series analysis*.

Himakireeti, K. and Vishnu, T. (2019). Air pasengers occupancy prediction using arima model. *International Journal of Applied Engineering Research*, 14.

Okulski, R. R., Heshmati, A., et al. (2010). Time series analysis of global airline passengers transportation industry. *Technology management, economics and policy program, TEMEP discussion paper*, 65.

Union des Aéroports Français, . (2017). Résultats d'activité des aéroports français 2017. statistiques de trafic. Technical report, Union des Aéroports Français.


# Appendix

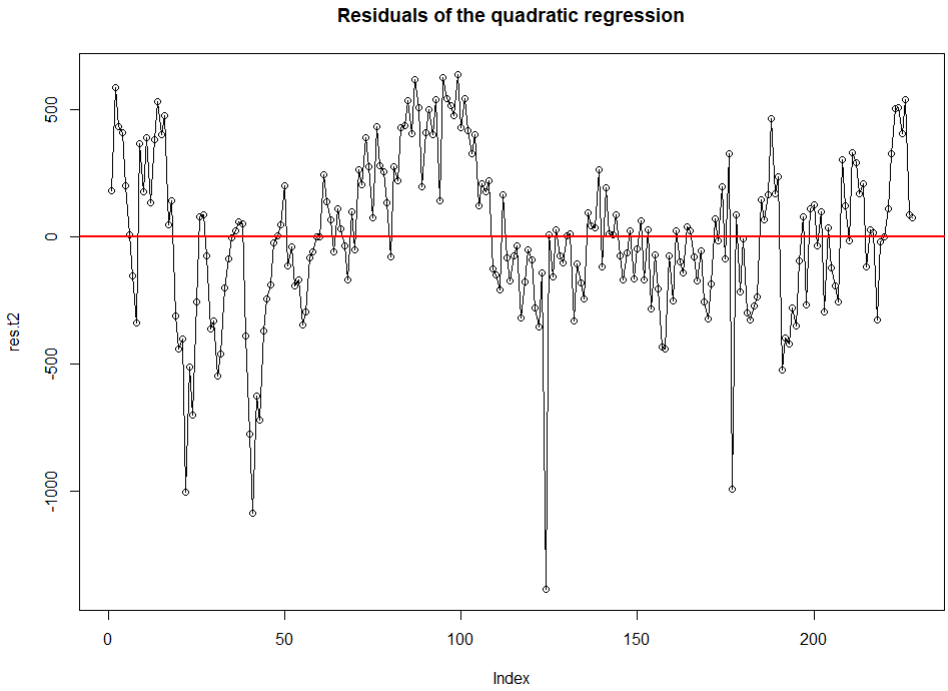


Figure 6: Residuals of the regression on the log of monthly air traffic

```
Box-Ljung test

data:  log.res
X-squared = 477.1, df = 20, p-value < 2.2e-16
```

```
Augmented Dickey-Fuller Test

data:  log.res
Dickey-Fuller = -3.0863, Lag order = 6, p-value = 0.1196
alternative hypothesis: stationary
```

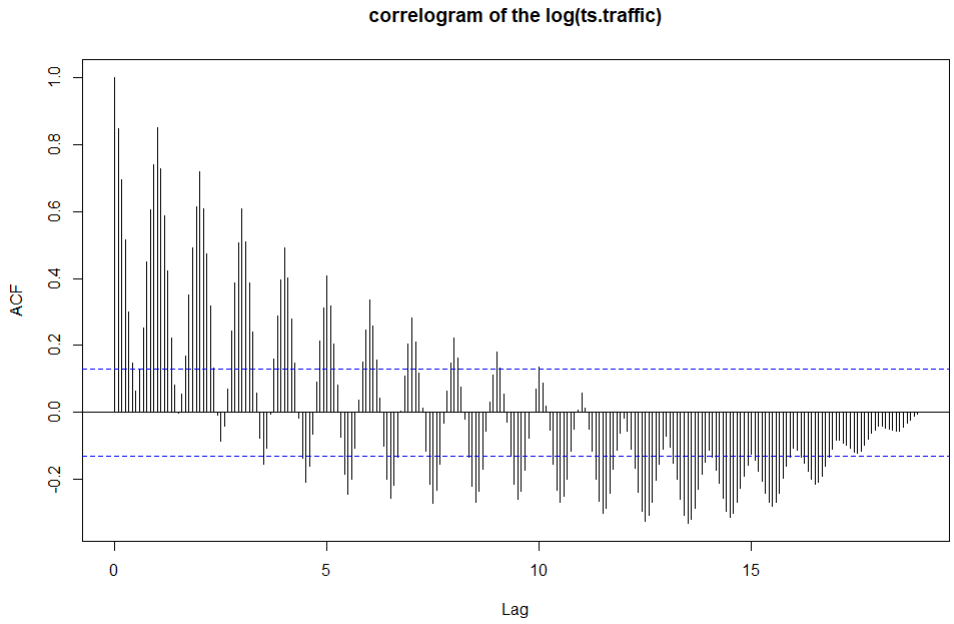


Figure 7: Correlogram of the log of monthly air traffic

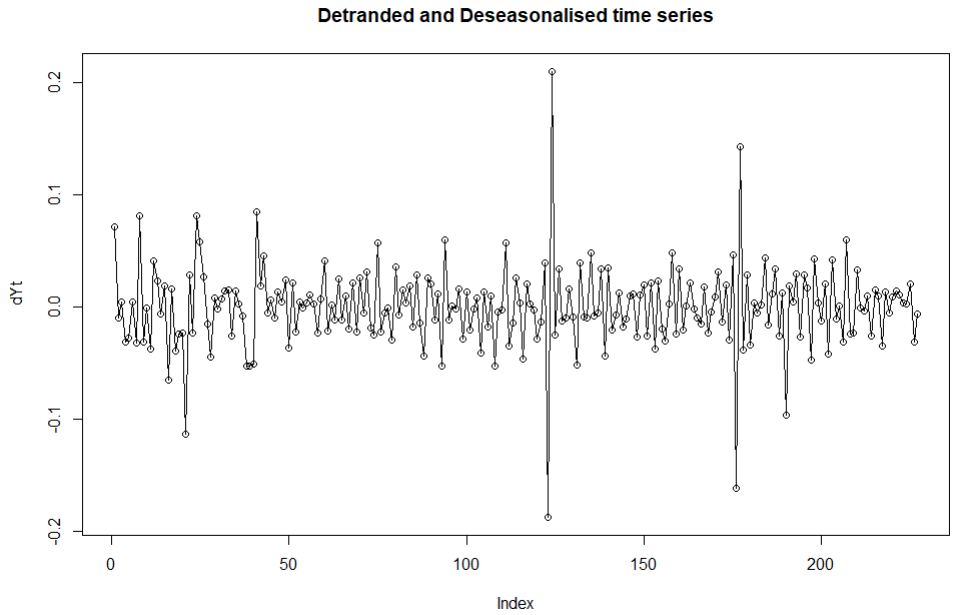


Figure 8: Log aif traffic passenger without trend and seasonality

# List of Figures



# List of Tables