Spatio-Temporal Data Analysis Project

2020-05-02

Patterns in foreign sims connected to OpenWiFi-Milan

Author: Bernardi Riccardo - 864018 Professor: Isadora Antoniano-Villalobos Course: Spatio-Temporal Data Analysis

Contents

Patterns in foreign sims connected to OpenWiFi-Milan 1

1 Statement of the Problem 3

2 Introduction & Motivation 3

3 Data Description 3

4 Exploration of the Data 4

4.1 Preprocessing 4

4.2 Using a boxCox transform 5

4.3 Hist after the transformation 5

4.4 Boxplot after the transformation 6

5 Time serie is built 6

5.1 Week by Week plot 7

6 Peaks Explanation 7

7 Trend recognition 8

7.1 Smoothing 8

7.2 Splitting 9

8 Models 9

8.1 Arima 9

8.2 Searching for multi seasonalities 14

8.3 RUGARCH 14

9 Conclusions 14

1 Statement of the Problem

The project is assigned to every student with a different topic. the topic have to be different between students and should be original to achieve a better score. Students can work in couple and this obviously involves a proportional workload, “more people, more work to be done”. The project should include the analysis of spatial data or temporal data, also both together is possible but the complexity involved is very high. The project should include an introduction, a description of the data, motivation of the choice, a detailed analysis with all the possible tools(interesting tools should be explained and can improve the overall mark).

I Worked by my own, the whole code and report is developed only by me. The code is not plotted on the resulting pdf but is available on the source. I used best practices to write good code such as KISS(keeping the code very very simple) and commenting the code. I used a lot of libraries preferring guaranteed code instead writing unrealiable code. I tried to use realiable libraries since R packages can be developed by all the people and can contain errors. I choosed a very very original problem, this can be checked by simply inputting the name of this project on the internet, no previous works were done until the time i’m writing(or at least they do not appear to me on t hegoogle page).

2 Introduction & Motivation

The dataset that I’ve chosen is about the presence of foreign smartphone’s sims to the OpenWifi of the Municipality of Milan. This data is open and available on the website data.gov.it. The reasons why I would like to go further with this project is that I strongly believe that are present seasonalities that can be interesting to be analysed but also can be more interesting to relate the outliers to some events that happened in the past with a certain mediatic relevance. In practice I would like to both analyse trend and seasonalities to know in which months there are more foreign people and if the trend is increasing in time and both search for outlier peaks to be related to important happenings in the Milan city. Finally I would like to forecast the possible presences in the new year in the city of Milan.

3 Data Description

The dataset comes from the open data provided by all the municipalities of Milan. This repository is available at dati.gov.it. From this repository I selected the data going from June of 2017 to October of the 2019.

Characteristics of the DataSet:

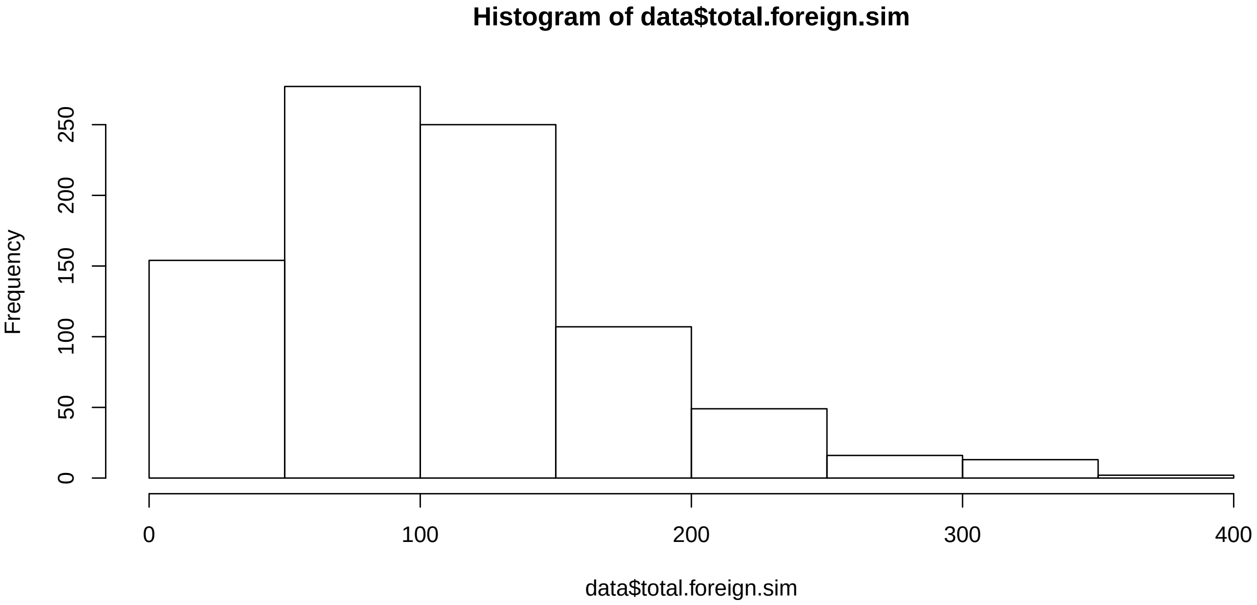
- the dataset contains 2 columns “Date, Number_of_Foreign_Sims”
- has 868 rows
- Dates goes from from 05/06/17 to 30/10/19 (~2 years)
- the datasets have no NA
- no lacking days

- the “Number_of_Foreign_Sims” is a discrete variable about total number of foreign sims in a certain Date connected to the OpenWifi of Milan

4 Exploration of the Data

```
## [1] "minimum, lower-hinge, median, upper-hinge, maximum)"
```

```
## [1]   1.0  61.5 101.0 141.0 378.0
```



4.1 Preprocessing

Checking Nans

```
## [1] 0
```

```
## [1] 0
```

Checking limit values

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

```
## [1] 378
```

```
## [1] 109.9228
```

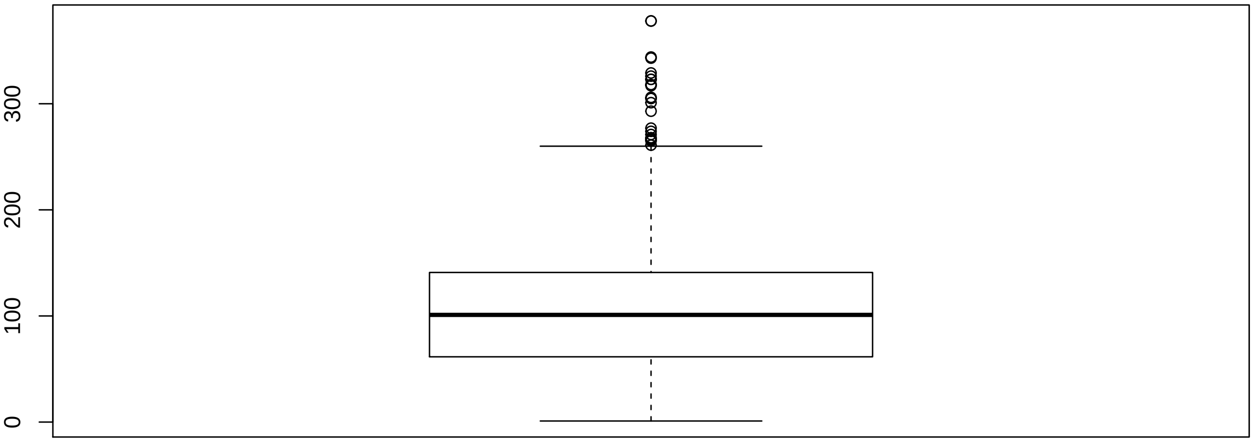
```
## [1] 63.63468
```

Elements that are good in our ts stand between mean±std

```
## [1] 173.5575
```

```
## [1] 46.28813
```

boxplot to check outliers



```
##      0%     25%     50%     75%    100%
##   1.00  61.75 101.00 141.00 378.00
```

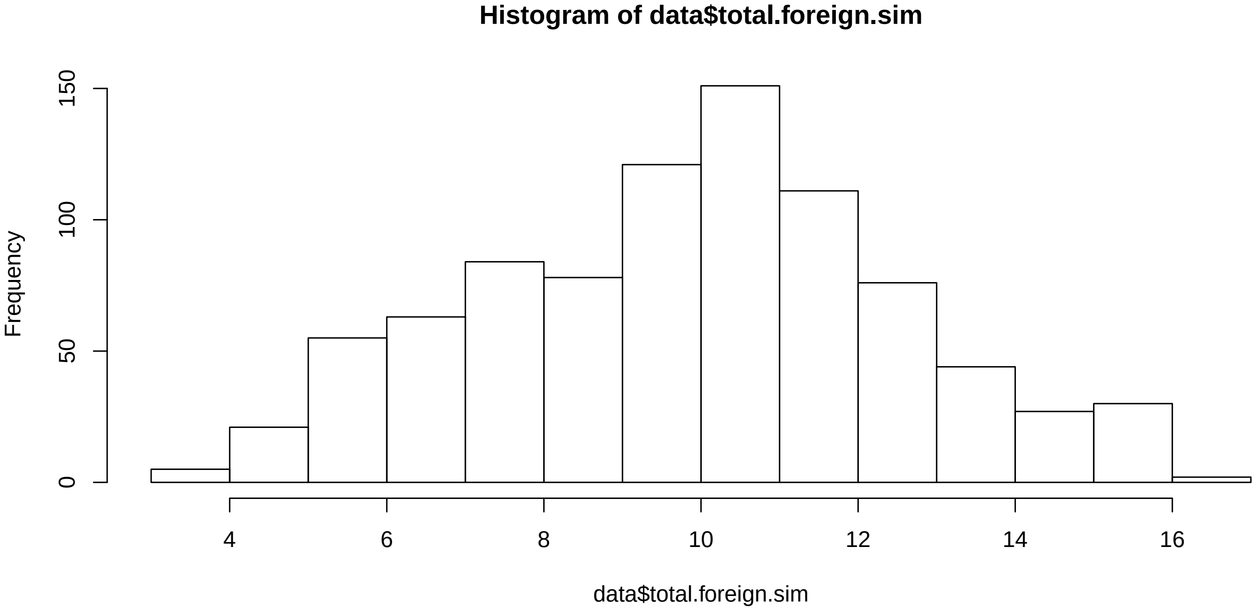
```
##     25%
## -57.125
```

```
##     75%
## 259.875
```

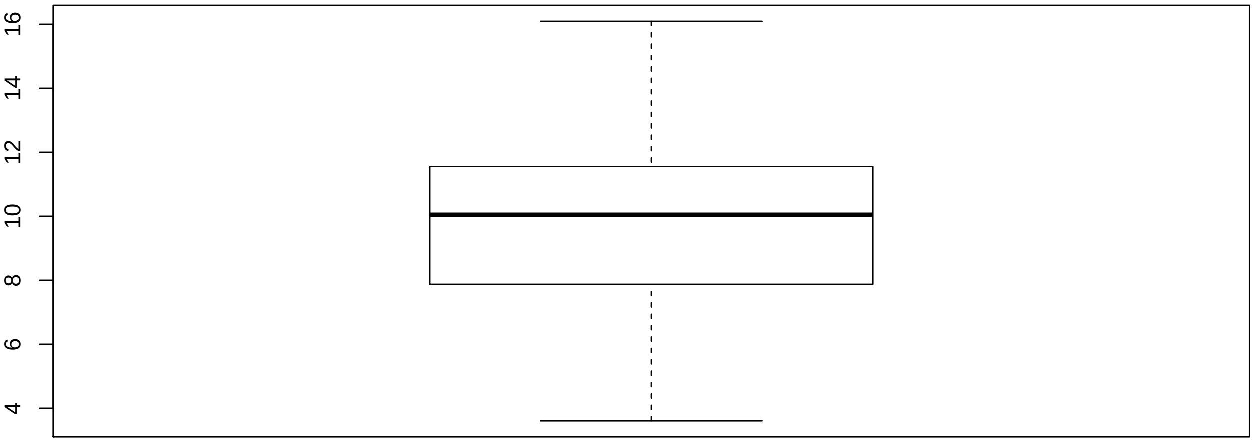
Checking last elements of the serie

4.2 Using a boxCox transform

4.3 Hist after the transformation



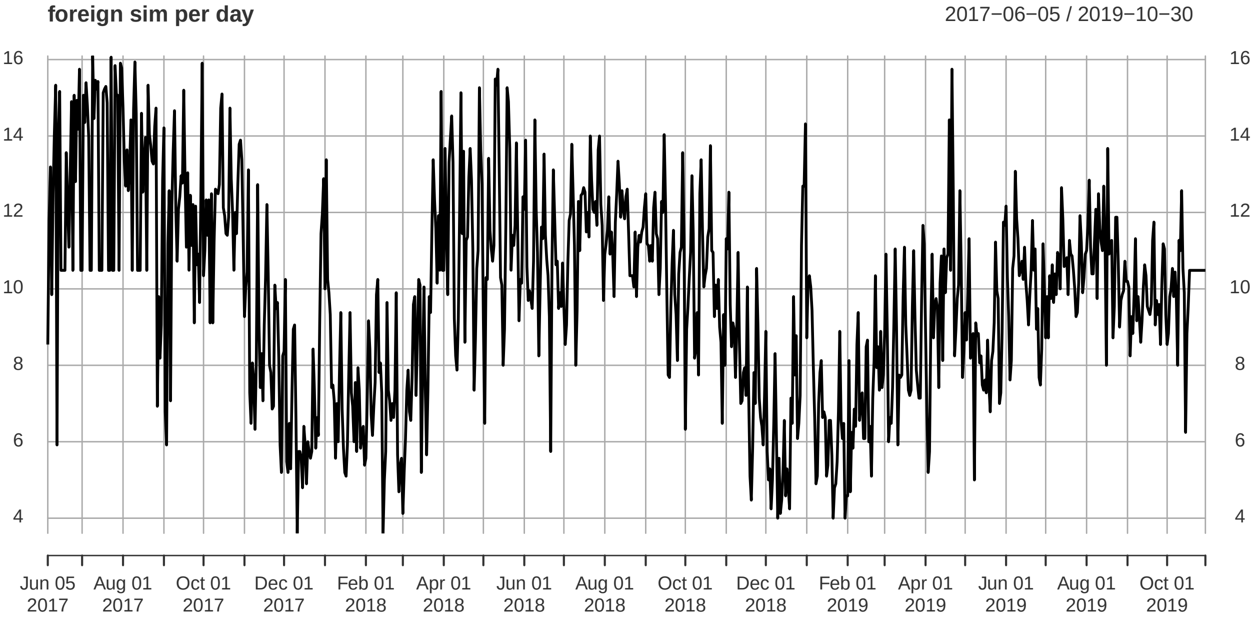
4.4 Boxplot after the transformation



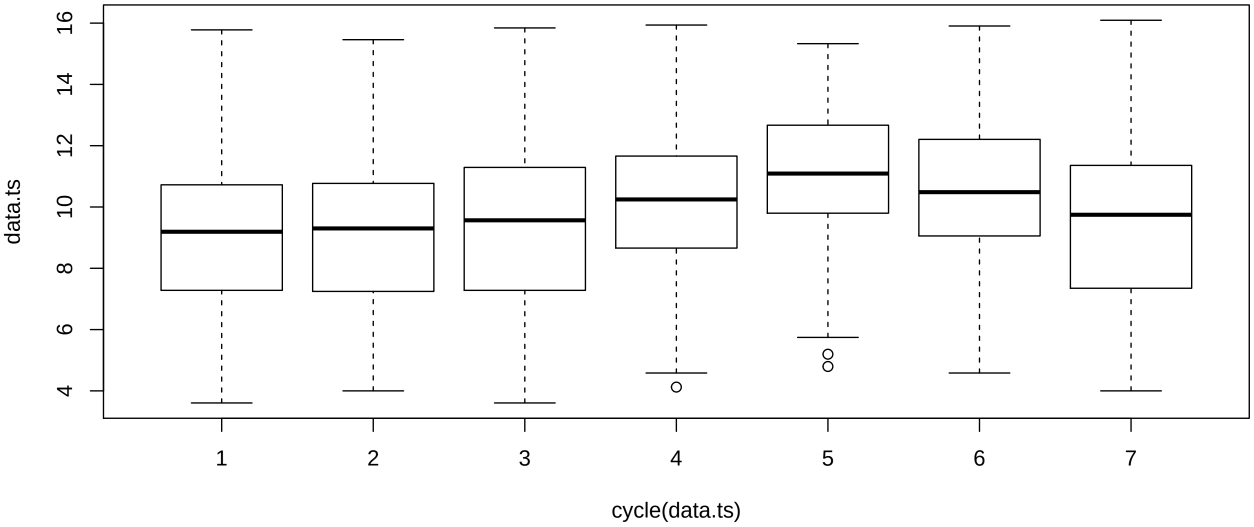
5 Time serie is built

Here the time serie is built

We loaded the dataset from the various datasets aggregating into only one dataset with 655 rows representing 2 years of data gathered. Starting from 05.06.2017 to 30.10.2019. Data is here:



5.1 Week by Week plot



6 Peaks Explanation

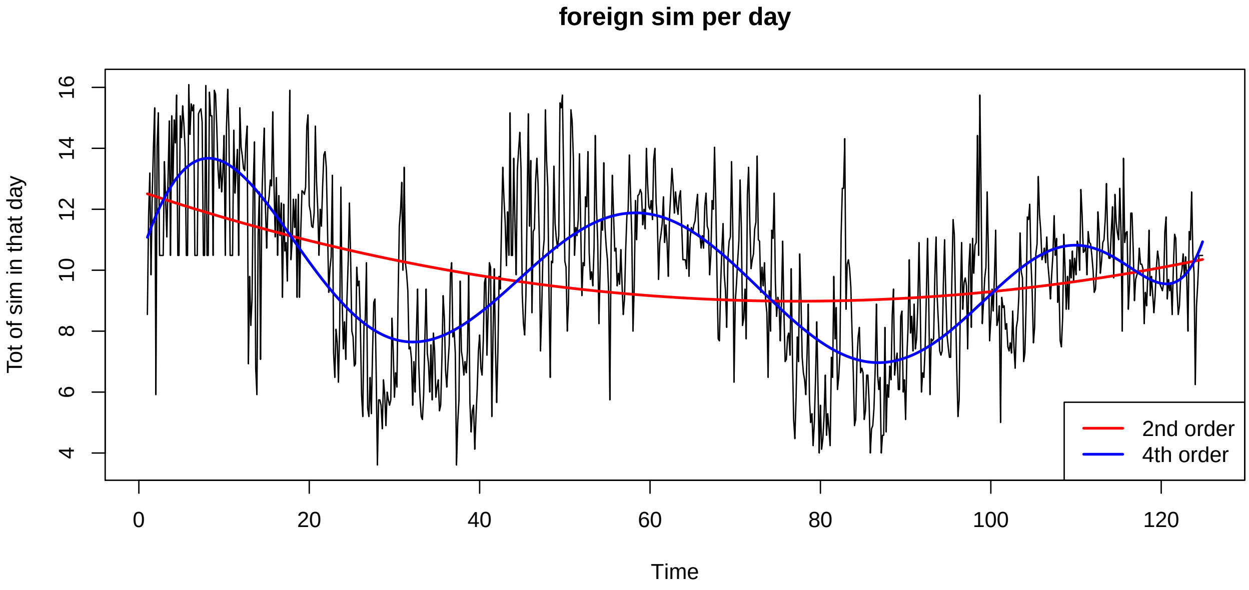
Many peaks are present we would like to exaplin them and to cut them out to be able to predict with a simple arima

- automatic roaming [https://www.mobileworld.it/2017/08/07/roaming-gratis-europa-condizioni-fair-us
- fashion week [https://www.cameramoda.it/it/milano-moda-donna/] february
- fashion week 2017 [https://www.milanoweekend.it/articoli/milano-fashion-week-2017-eventi-programm february

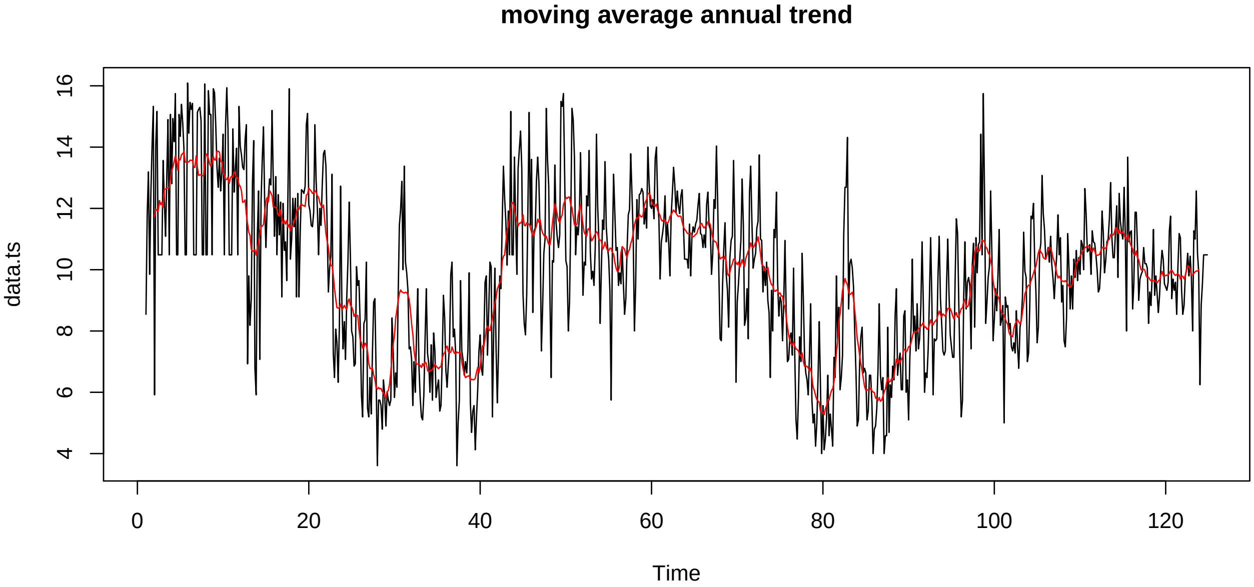
```
## [1] "2017-07-09"
```

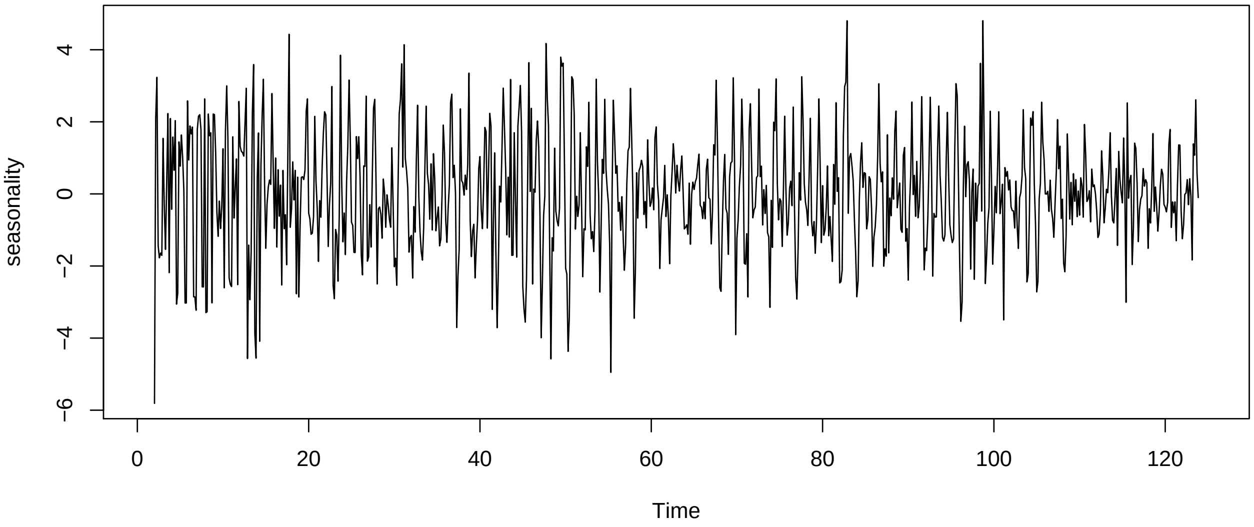
- arch week [https://www.lastampa.it/milano/2017/06/17/news/milano-smart-city-del-futuro-se-ne-par 34584894?refresh_ce]
- it was a saturday!!
- it was the orient festival [https://www.wikieventi.it/milano/index.php?data_selezionata=2017-06-17]
- many mucis events, samsara of papetee and others, folk's festivals, discounts [https://www.wikieventi.it/milano/index.php?data_selezionata=2017-07-22]

7 Trend recognition



7.1 Smoothing





7.2 Splitting

8 Models

We tried many models, the most known is the arma but we tried: - arma - arima - sarima - var - rugarch - fgarch

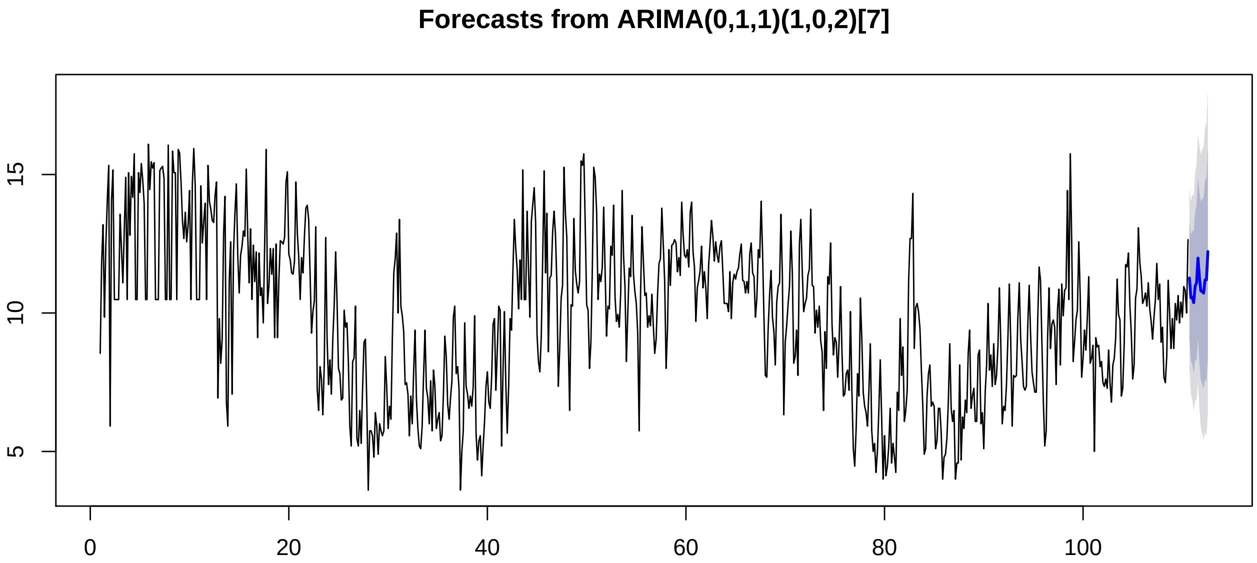
and others, i'm reporting here only the best one and some ideas for the worst performing ones.

8.1 Arima

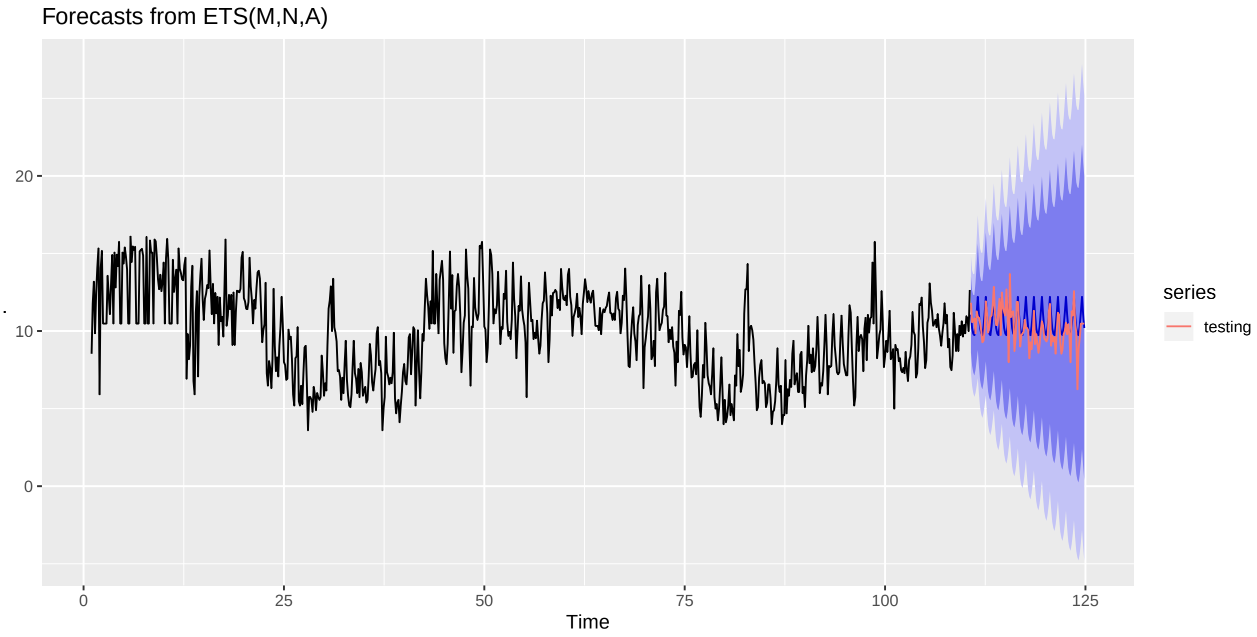
To find the best arima in a such complicate time serie we are going to exploit a grid search algorithm done via the auto arima function provided by the fpp package

```
##
##  ARIMA(2,1,2)(1,0,1)[7] with drift         : Inf
##  ARIMA(0,1,0)           with drift         : 3256.764
##  ARIMA(1,1,0)(1,0,0)[7] with drift         : 3156.221
##  ARIMA(0,1,1)(0,0,1)[7] with drift         : 3125.008
##  ARIMA(0,1,0)                              : 3250.127
##  ARIMA(0,1,1)           with drift         : 3166.984
##  ARIMA(0,1,1)(1,0,1)[7] with drift         : Inf
##  ARIMA(0,1,1)(0,0,2)[7] with drift         : 3088.452
##  ARIMA(0,1,1)(1,0,2)[7] with drift         : 2986.153
##  ARIMA(0,1,1)(2,0,2)[7] with drift         : Inf
##  ARIMA(0,1,1)(2,0,1)[7] with drift         : Inf
##  ARIMA(0,1,0)(1,0,2)[7] with drift         : 3150.281
##  ARIMA(1,1,1)(1,0,2)[7] with drift         : Inf
##  ARIMA(0,1,2)(1,0,2)[7] with drift         : Inf
##  ARIMA(1,1,0)(1,0,2)[7] with drift         : 3062.807
##  ARIMA(1,1,2)(1,0,2)[7] with drift         : Inf
##  ARIMA(0,1,1)(1,0,2)[7]                    : 2979.53
##  ARIMA(0,1,1)(0,0,2)[7]                    : 3081.812
##  ARIMA(0,1,1)(1,0,1)[7]                    : Inf
##  ARIMA(0,1,1)(2,0,2)[7]                    : Inf
##  ARIMA(0,1,1)(0,0,1)[7]                    : 3118.366
##  ARIMA(0,1,1)(2,0,1)[7]                    : Inf
##  ARIMA(0,1,0)(1,0,2)[7]                    : 3143.646
##  ARIMA(1,1,1)(1,0,2)[7]                    : Inf
##  ARIMA(0,1,2)(1,0,2)[7]                    : Inf
##  ARIMA(1,1,0)(1,0,2)[7]                    : 3056.175
##  ARIMA(1,1,2)(1,0,2)[7]                    : Inf
##
##  Best model: ARIMA(0,1,1)(1,0,2)[7]
```

A forecast on the training set looks like this one below:



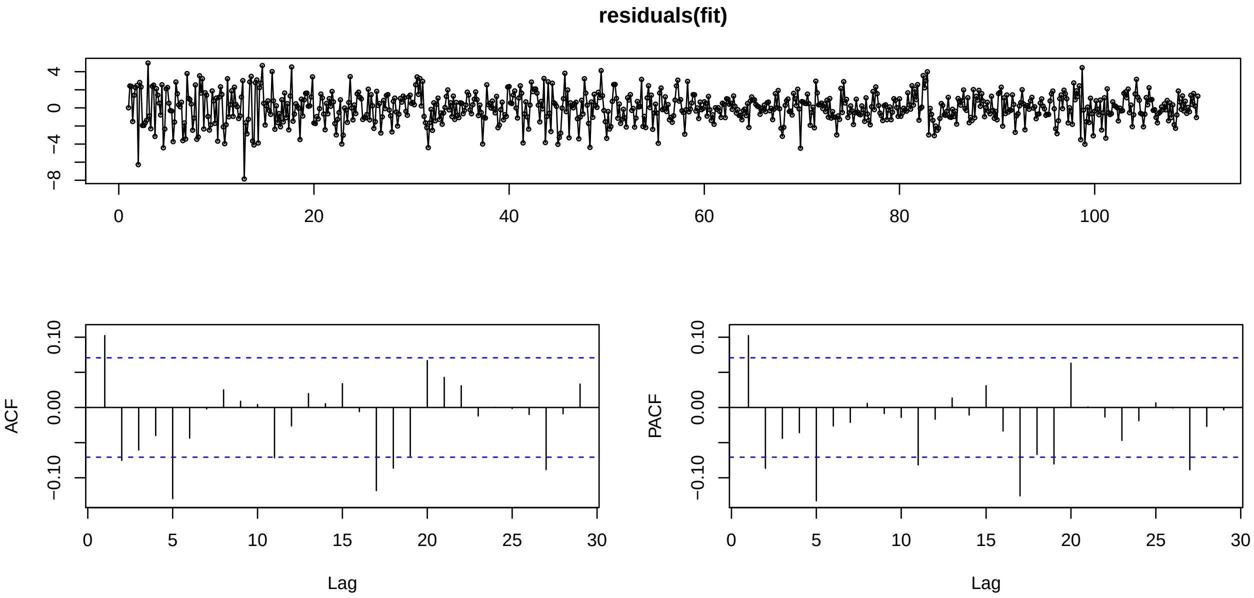
Here we are going to plot a forecast on a part of the data that was never seen by the arima model, the plots looks not so bad.



The accuracy on the test set is only two percentages points lower than on the training set. We are using 100 points.

```
##                        ME     RMSE      MAE        MPE     MAPE      MASE
## Training set  0.003771164 1.645765 1.264390  -2.365973 13.90600 0.7456747
## Test set     -1.824877649 2.313312 1.923329 -19.445302 20.24236 1.1342847
##                   ACF1 Theil's U
## Training set 0.1023087        NA
## Test set     0.6043079  1.676481
```

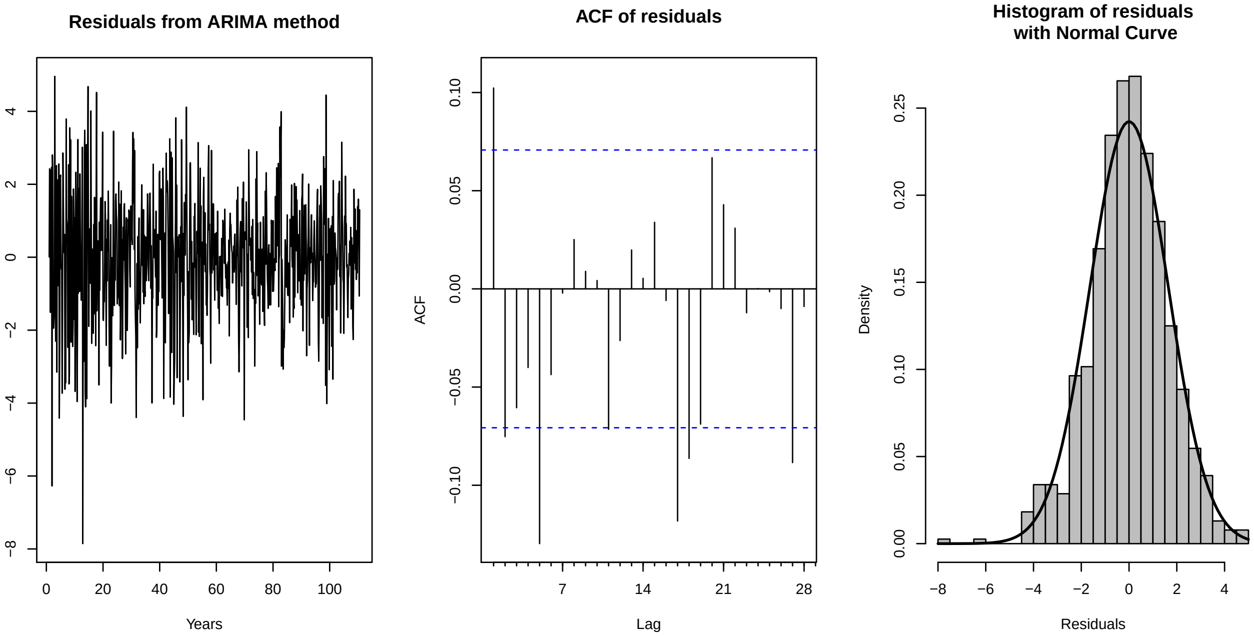
Actually here below we can see that residuals are not pretty good but after many weeks of attempts with many models, all the poossible seasonalities, all the decomposition, all the tricks available i'm pretty sure that this is the best trade-off between complexity of the models used, accuracy and processing of the dataset.



```
##
##  Box-Ljung test
##
## data:  residuals(fit)
## X-squared = 63.891, df = 24, p-value = 1.768e-05
```

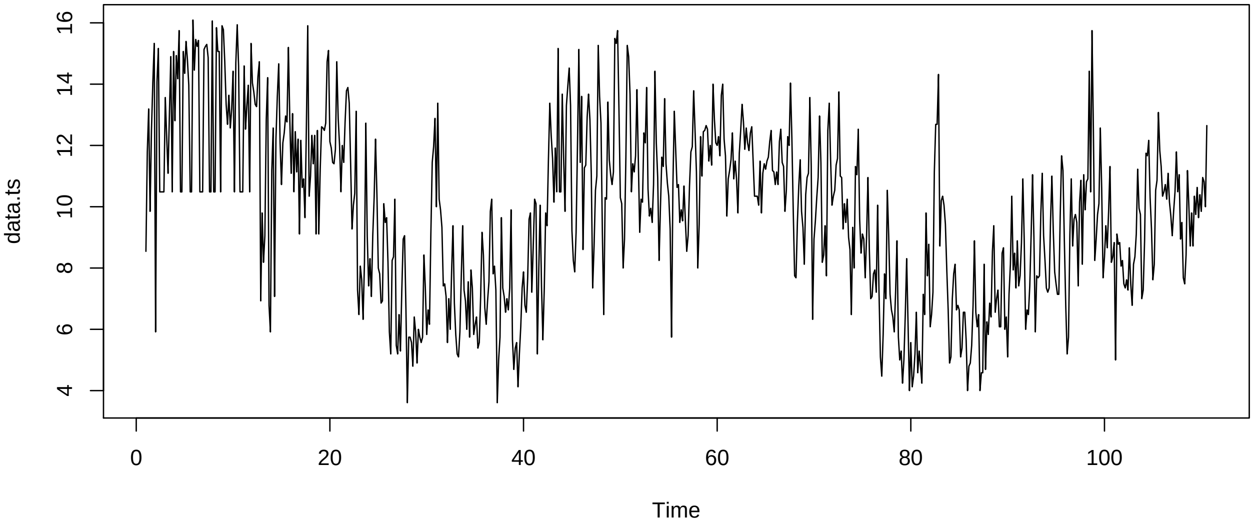
```
##
##  Box-Pierce test
##
## data:  residuals(fit)
## X-squared = 62.865, df = 24, p-value = 2.49e-05
```

The diagnostic here below confirms at least that the auto arima chose the right differentiation parameter.



```
##
##  KPSS Test for Level Stationarity
##
## data:  diff(data.ts)
## KPSS Level = 0.020484, Truncation lag parameter = 6, p-value = 0.1
```

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



The plot is not good but AIC and BIC are very high, we should try with a multi seasonal decomposition

```
## [1] 7
```

8.2 Searching for multi seasonalities

We searched for all the possible multi seasonalities using the msts package but at the end adding many seasonalities did not helped in finding better residuals or clearer trends and seasonalities. Adding more and more just created a complicated useless model that scored the same as the base one.

8.3 RUGARCH

We tried this model but it was not useful for great predictions, residuals were much more scattered than arima.

9 Conclusions

It was a hard work! I tried a lot of methods to fit the data and obtain good forecasting and good residuals. I tried searching for multi seasonalities, difference some times to reach stationarity, detrending with lm and ma, smoothing. I tried by hand many arima models. I tried to decompose the time serie in many ways. I tried all the possible frequencies that can be thought as valid. Eventually it was really interesting! I experimented a lot. The dataset is brand new, never touched by other data scientists for what i know.