Spatio-Temporal Data Analysis Project

2020-04-26

Patterns in foreign sims connected to OpenWiFi-Milan

Author: Bernardi Riccardo - 864018

Contents

1 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.

2 The Data

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 January of 2018 to October of the 2019.

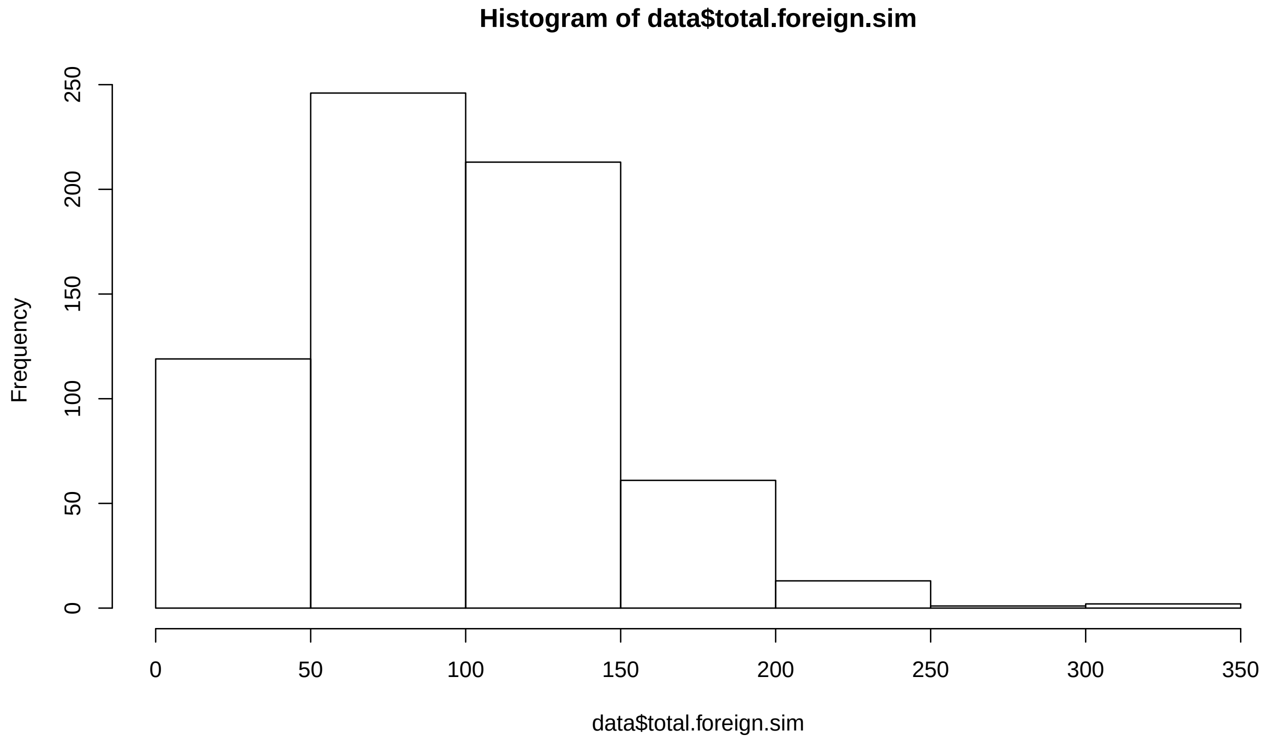
Characteristics of the DataSet:

- the dataset contains 2 columns "Date, Number_of_Foreign_Sims"
- has 658 rows
- Dates goes from from 01/01/18 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

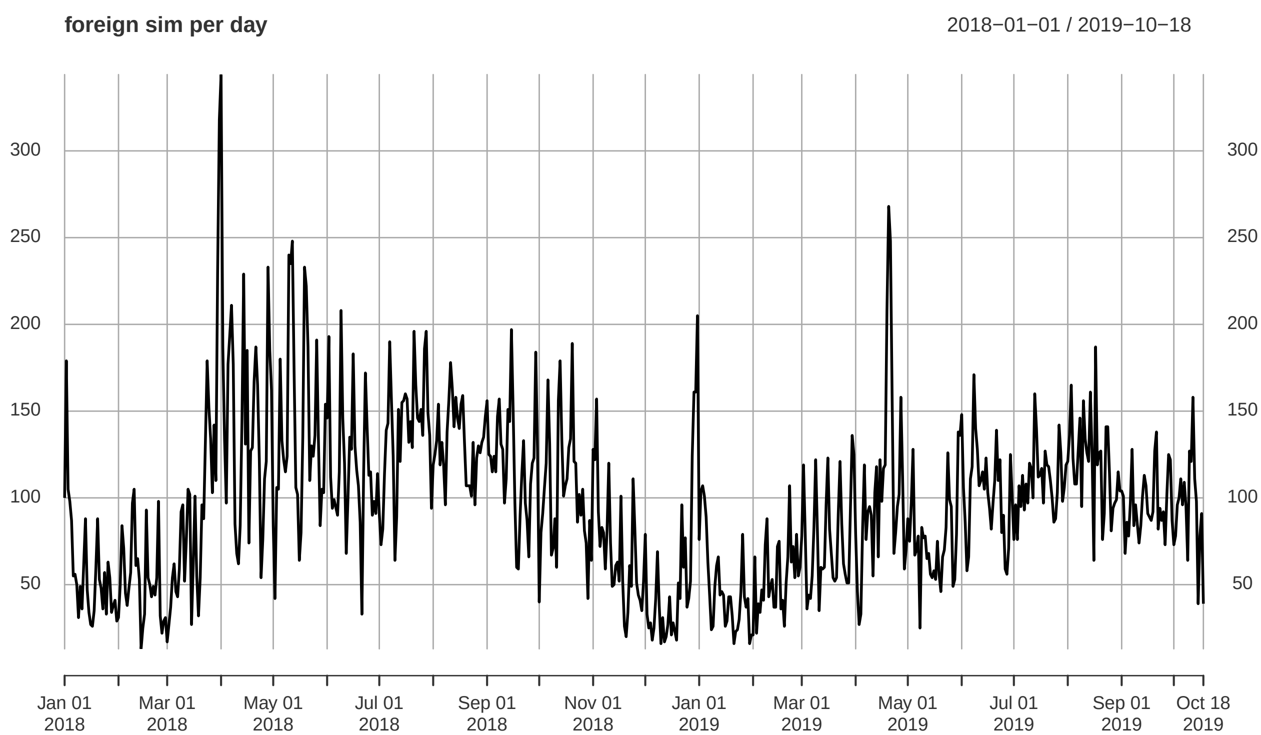
3 Exploration of the Data

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

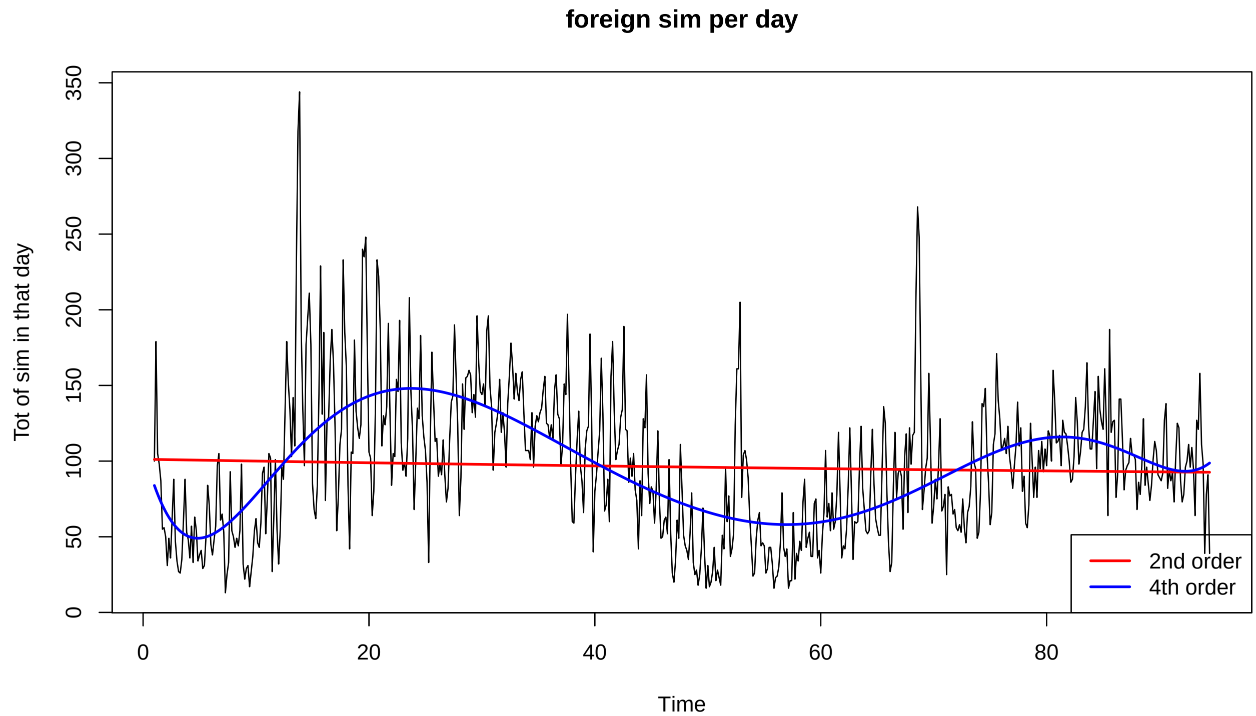
```
## [1]  13  59  95 124 344
```



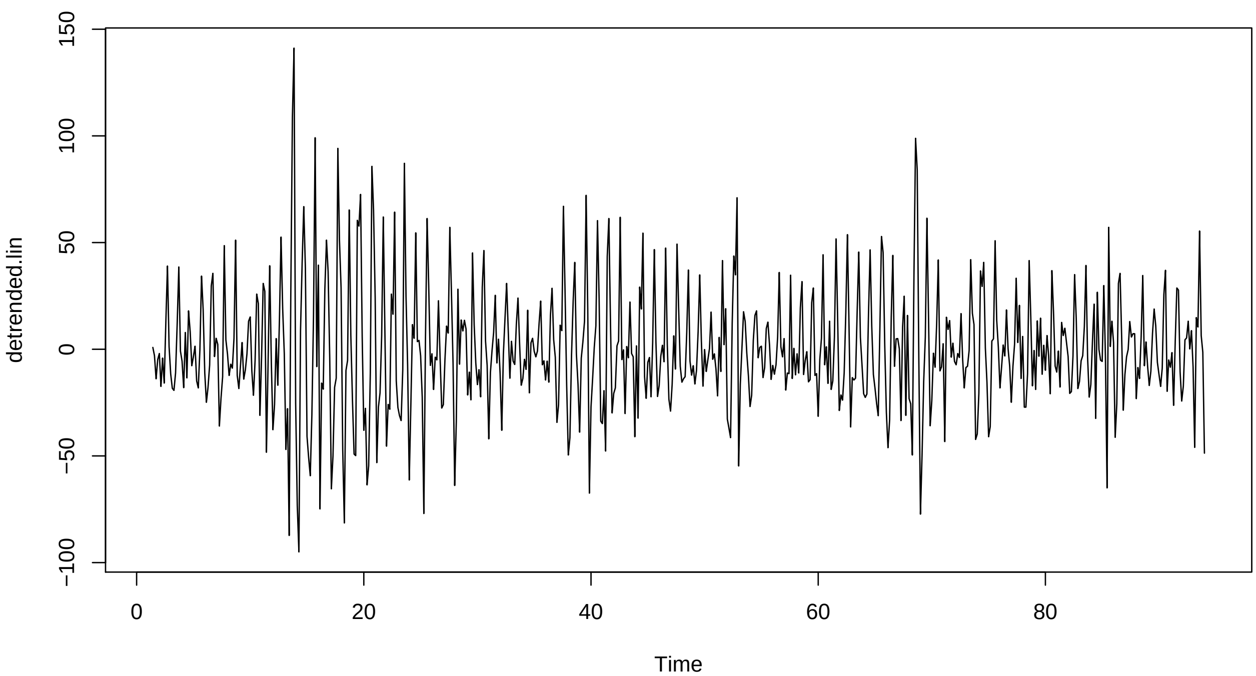
We loaded the dataset from the various datasets aggregating into only one dataset with 655 rows representing 2 years of data gathered. Starting from 01.01.2018 to 30.10.2019. Data is here:



4 Trend recognition

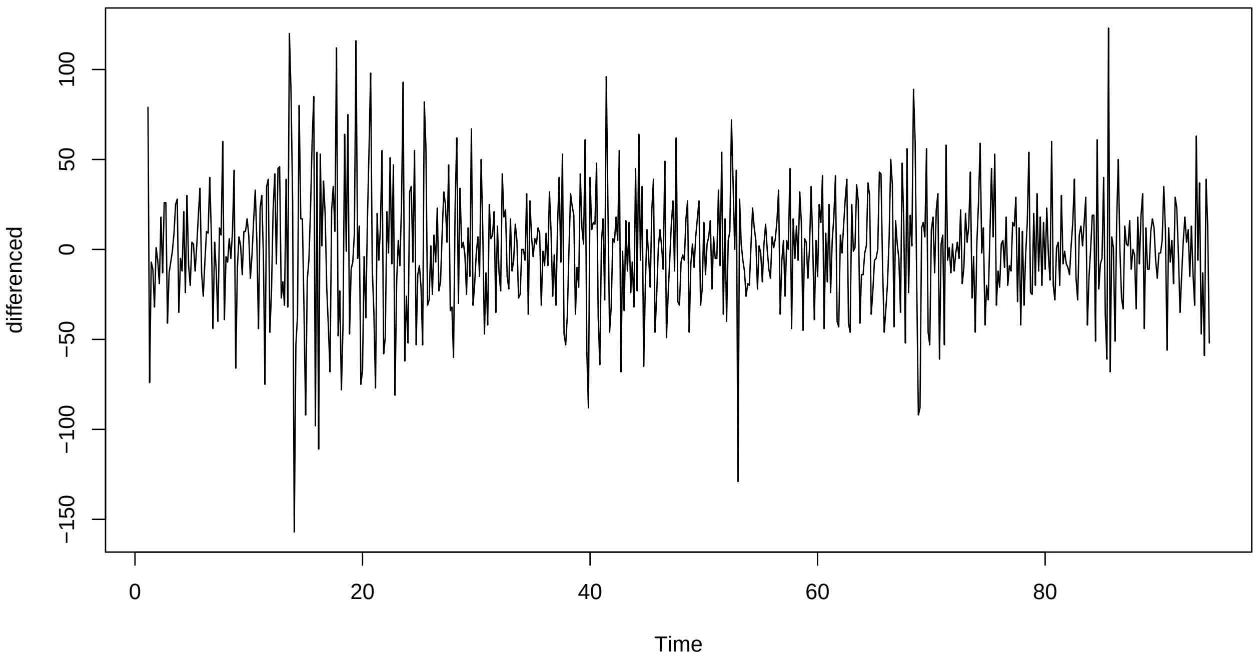


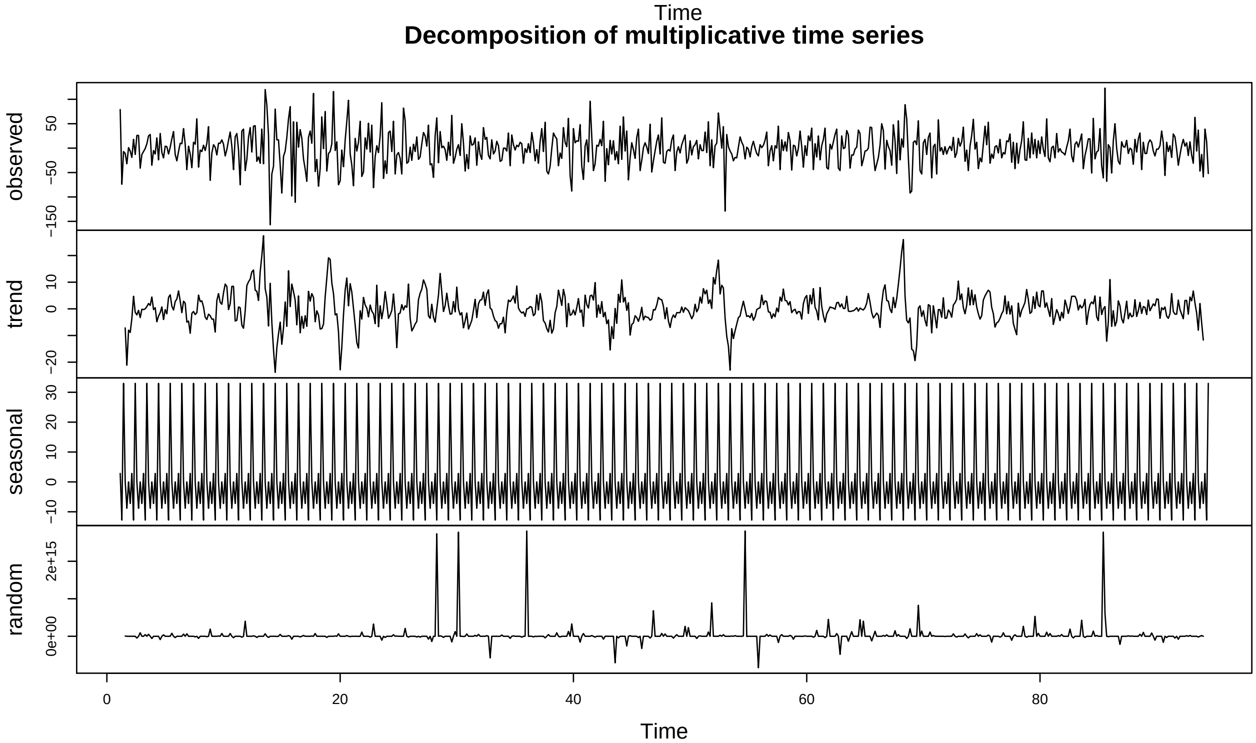
4.1 Detrending using LM

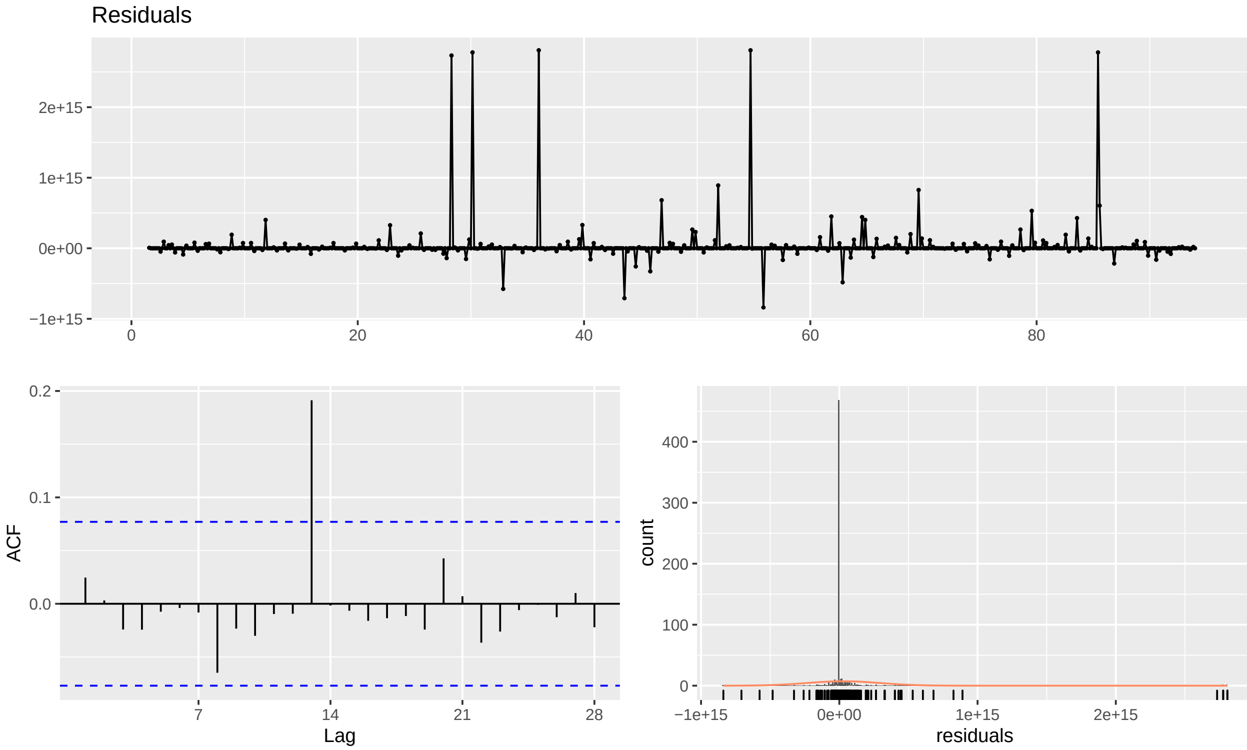


5 Removing seasonality

A good idea is to differenciate before decomposing. With the multiplicative model





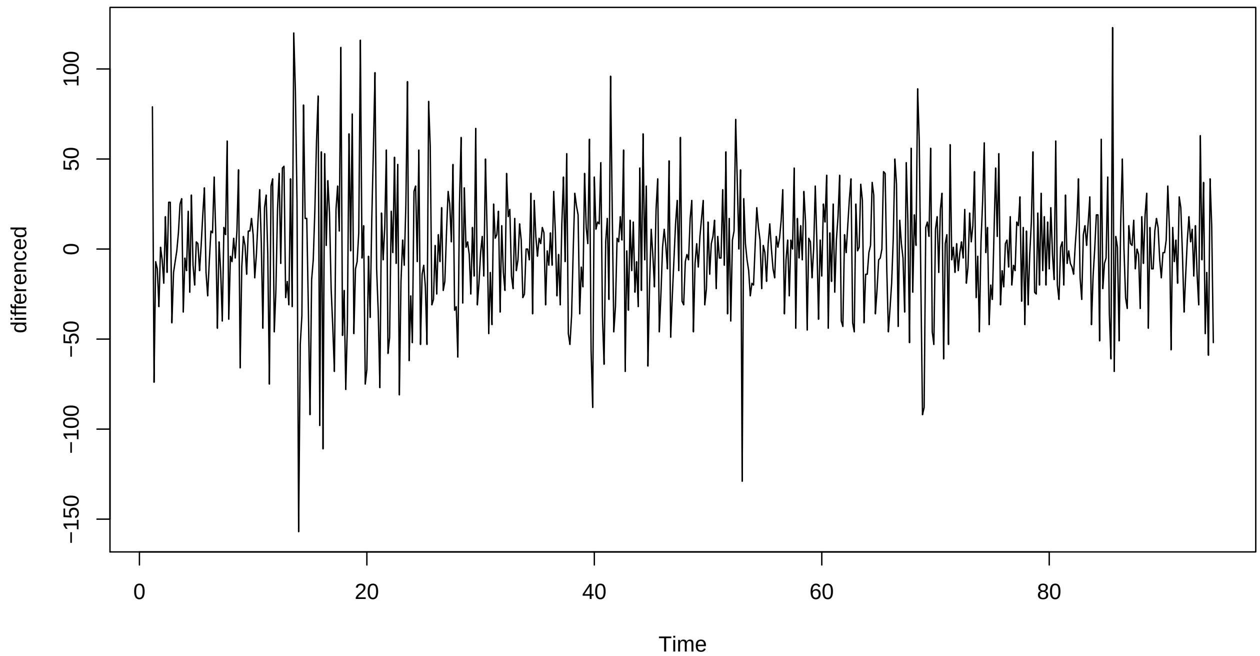


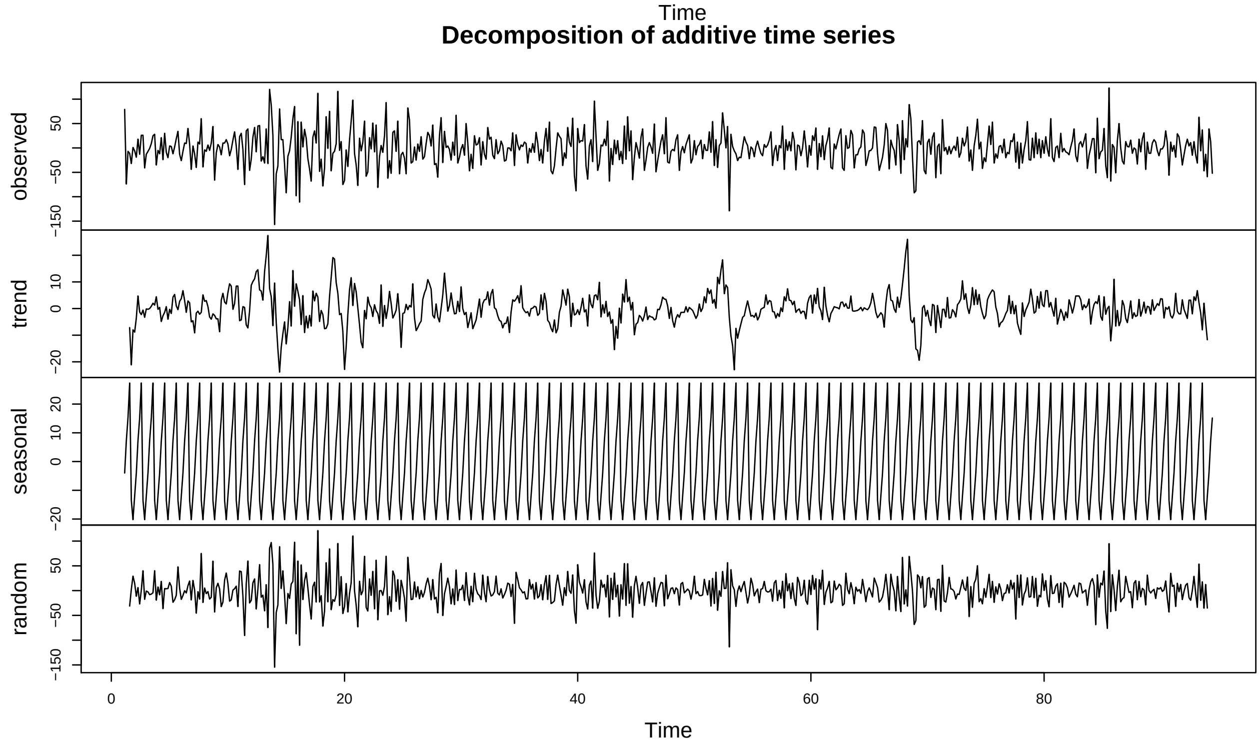

```
##
##  Box-Pierce test
##
## data:  decomposed$random
## X-squared = 1.1981, df = 5, p-value = 0.9451
```

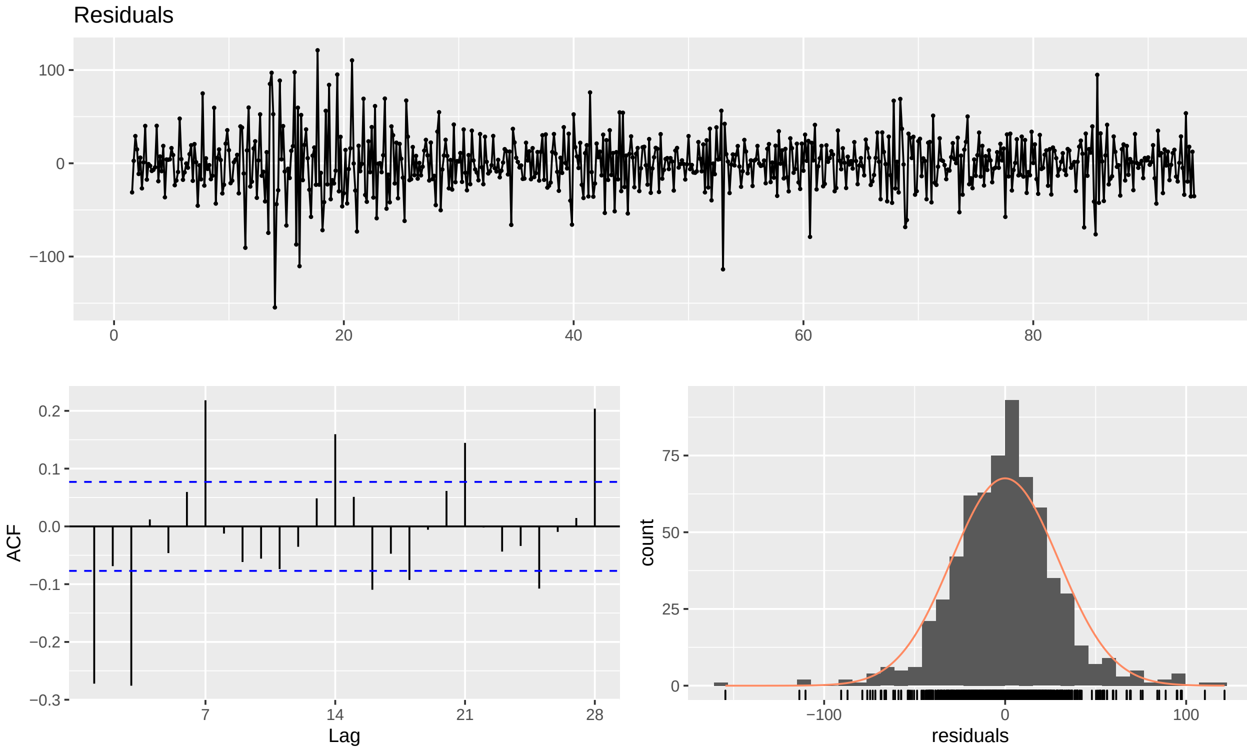
```
##
##  Box-Ljung test
##
## data:  decomposed$random
## X-squared = 1.2068, df = 5, p-value = 0.9442
```

6 The additive model doesn't work for us

With the additive model This model doesn't work at all


differenced
Time
−150
−100
−50
0
50
100
0
20
40
60
80

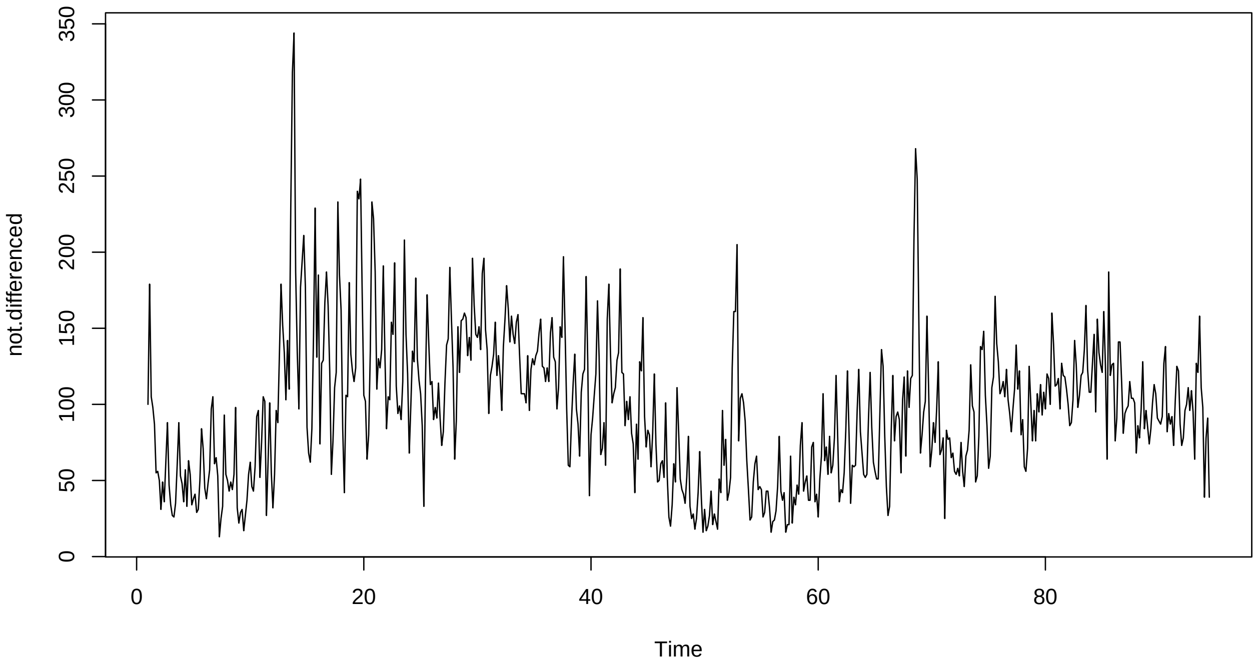

Decomposition of additive time series
observed
trend
seasonal
random
Time
0
20
40
60
80

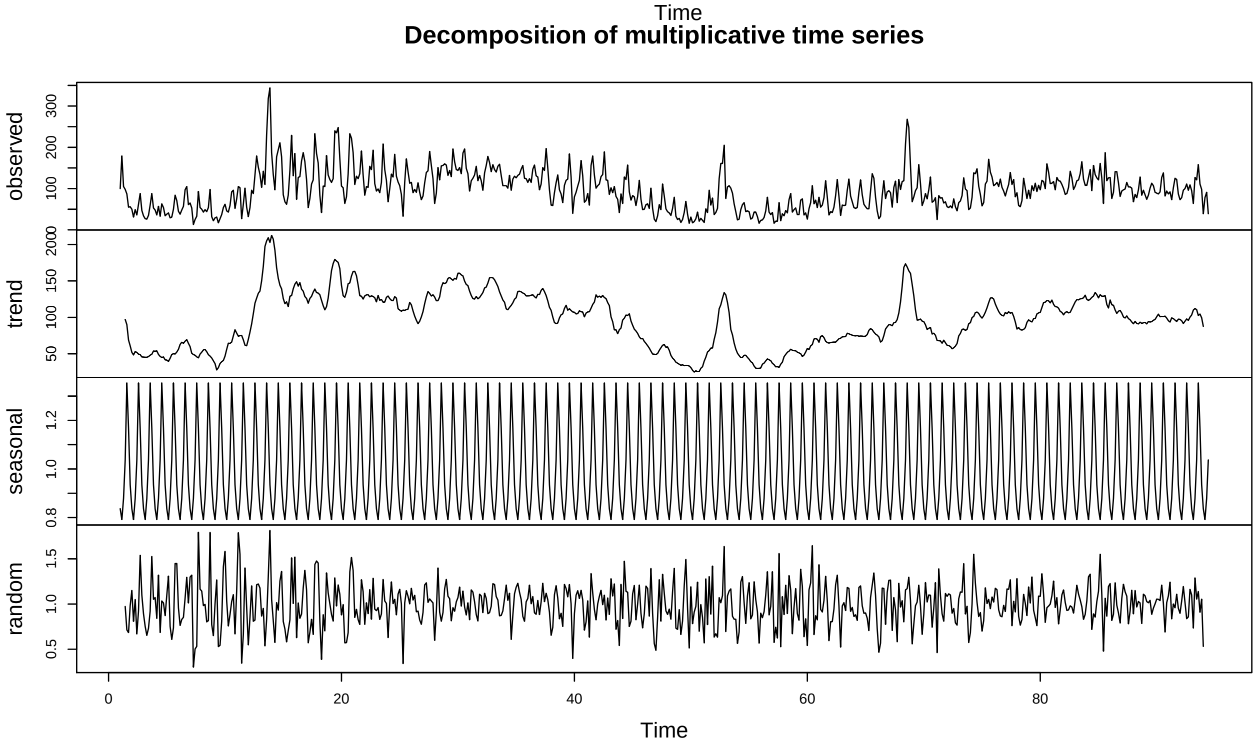


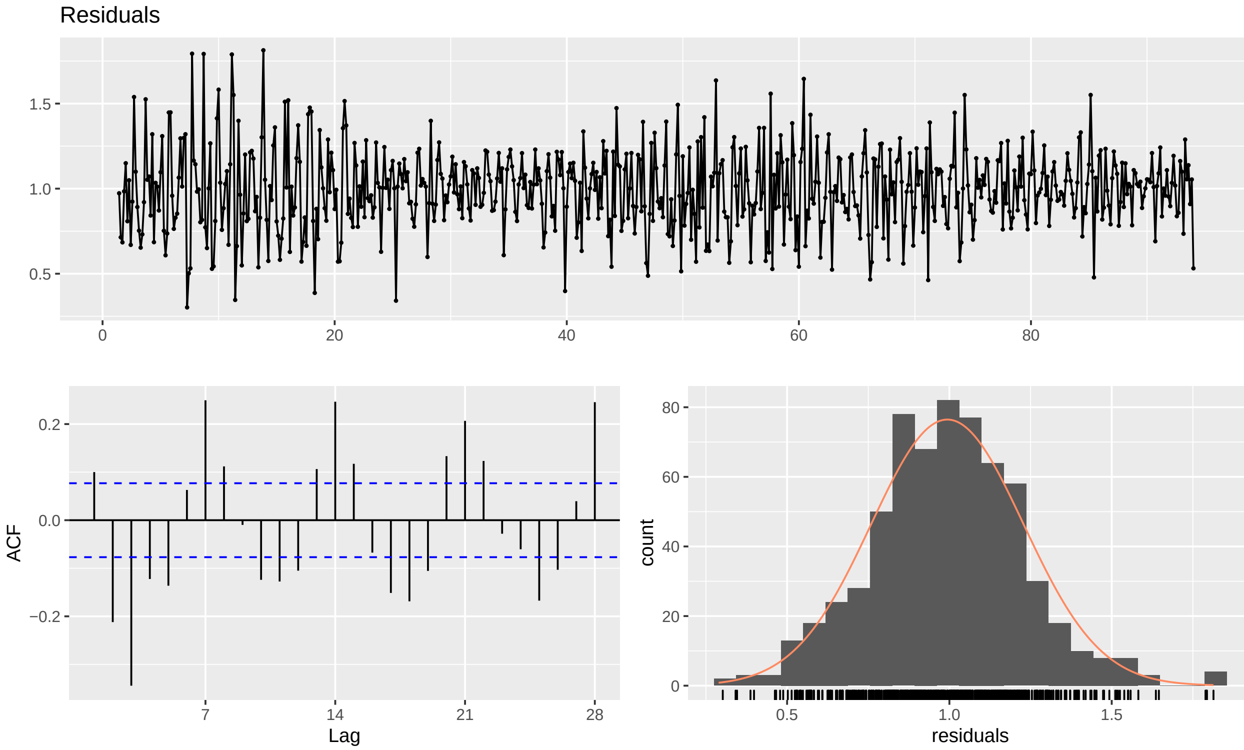

```
##
##  Box-Pierce test
##
## data:  decomposed$random
## X-squared = 101.8, df = 5, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  decomposed$random
## X-squared = 102.44, df = 5, p-value < 2.2e-16
```

Without the first differentiation the result will have been much worse:


not.differenced
0
50
100
150
200
250
300
350
0
20
40
60
80
Time


Decomposition of multiplicative time series
observed
trend
seasonal
random
Time
0
20
40
60
80

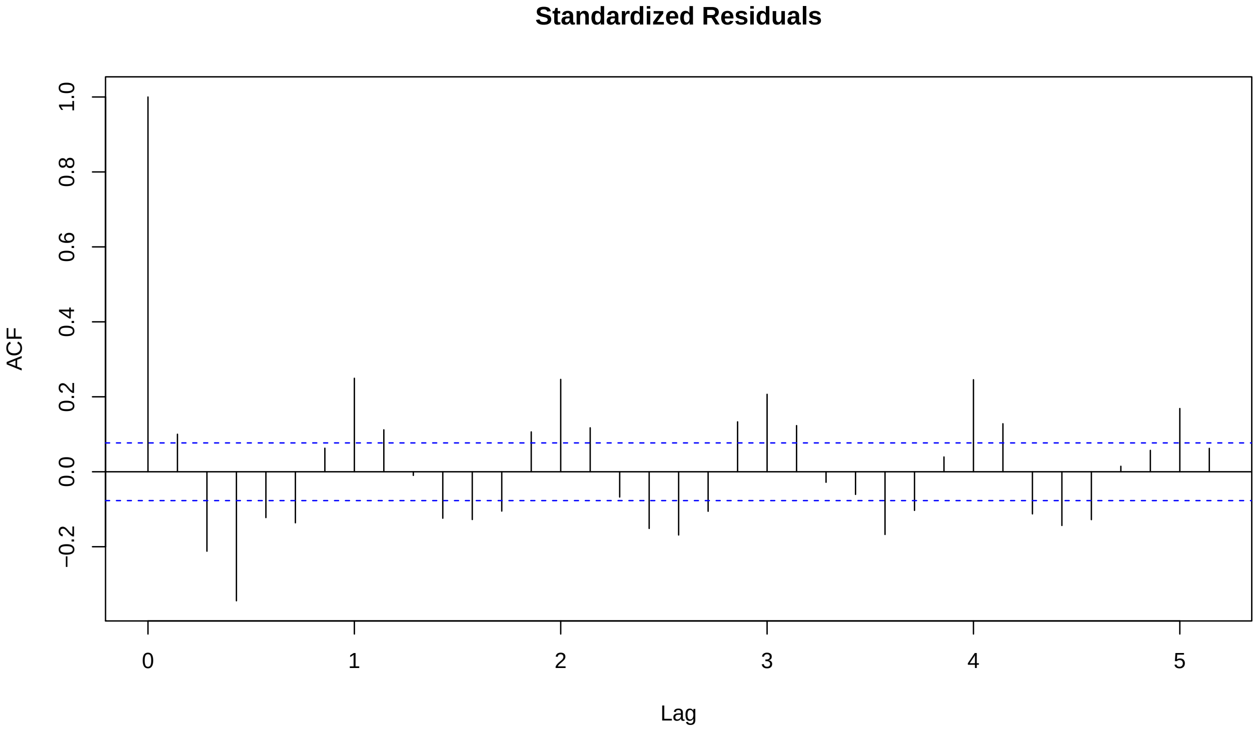


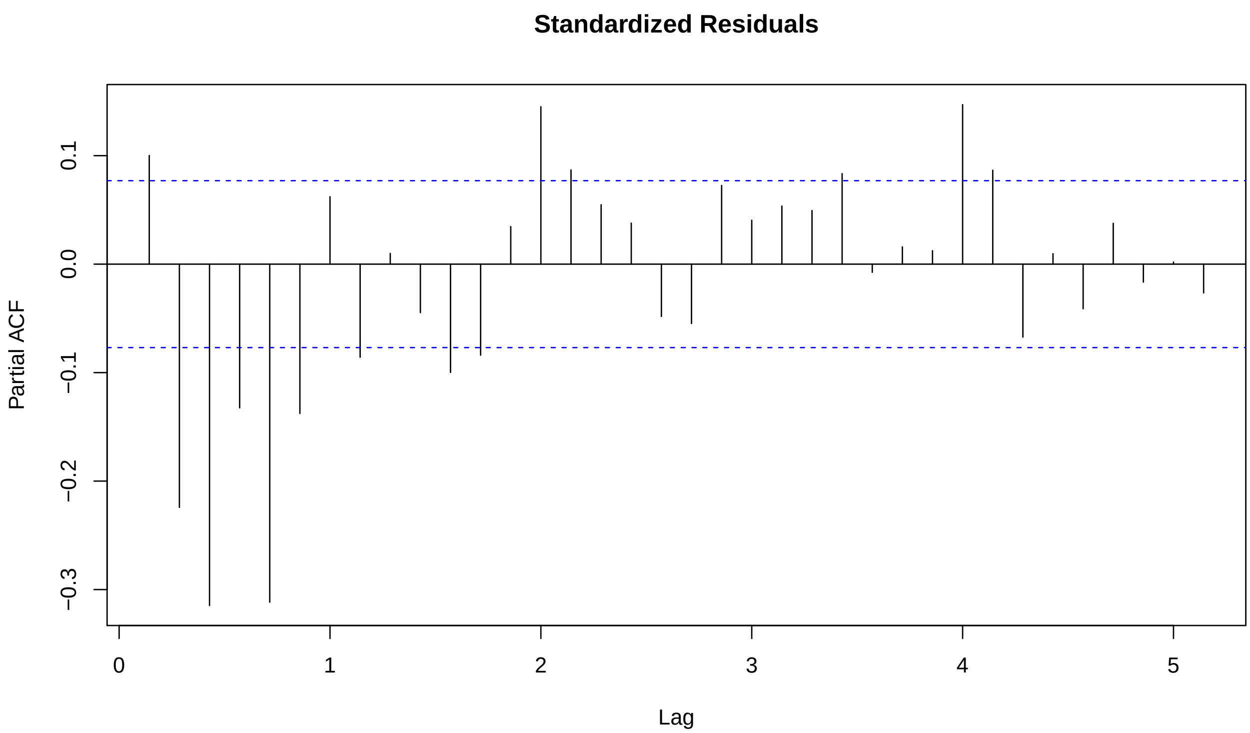

```
##
##  Box-Pierce test
##
## data:  decomposed$random
## X-squared = 134.37, df = 5, p-value < 2.2e-16
```

```
##
##  Box-Ljung test
##
## data:  decomposed$random
## X-squared = 135.4, df = 5, p-value < 2.2e-16
```

Every 7 lags the peak recurs

7 Check Residuals

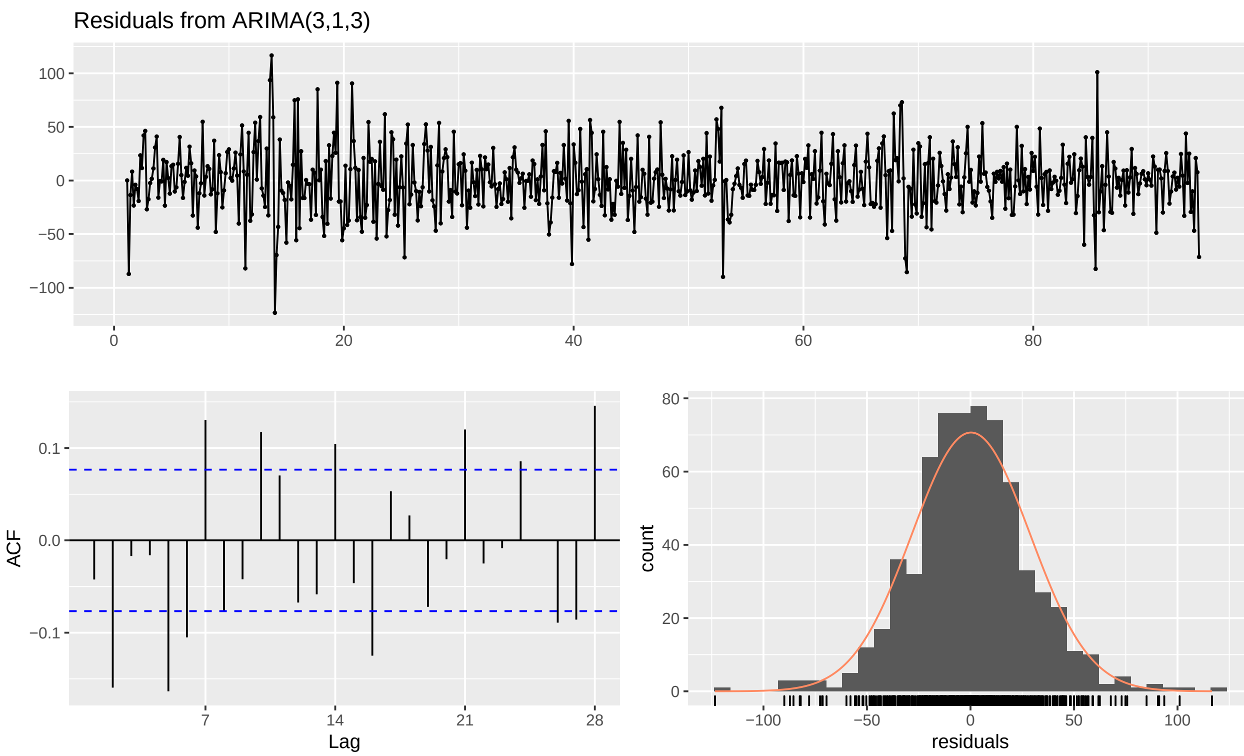




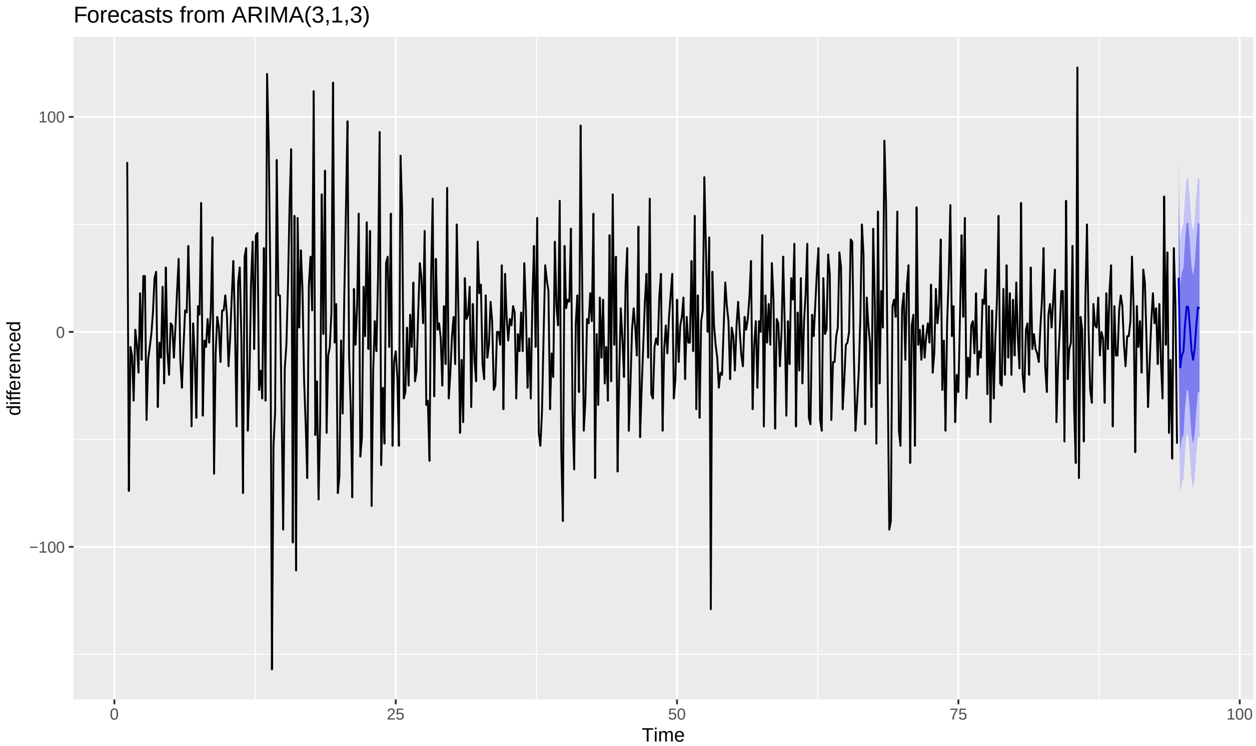
8 Arima

```
## Series: differenced
```

```
## ARIMA(3,1,3)
##
## Coefficients:
##          ar1      ar2      ar3      ma1     ma2      ma3
##       0.8858  -0.5556  -0.3444  -2.1937  2.0793  -0.8856
## s.e.  0.0407   0.0494   0.0389   0.0211  0.0424   0.0267
##
## sigma^2 estimated as 825:  log likelihood=-3120.88
## AIC=6255.75   AICc=6255.93   BIC=6287.12
##
## Training set error measures:
##                     ME     RMSE      MAE MPE MAPE      MASE        ACF1
## Training set 0.2368323 28.56859 21.52922 NaN  Inf 0.7603388 -0.04237868
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(3,1,3)
## Q* = 84.817, df = 8, p-value = 5.218e-15
##
## Model df: 6.   Total lags used: 14
```

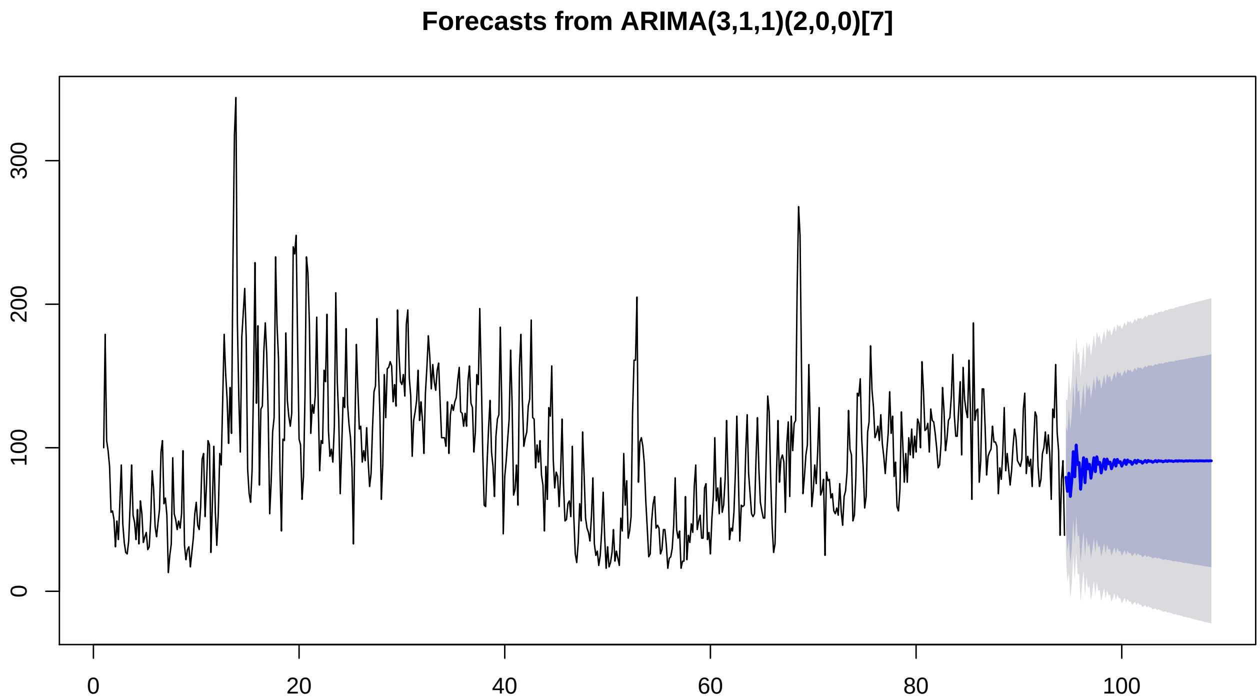


9 Auto Arima

```
##
## Fitting models using approximations to speed things up...
##
## ARIMA(2,1,2)(1,0,1)[7] with drift         : Inf
## ARIMA(0,1,0)           with drift         : 6473.005
## ARIMA(1,1,0)(1,0,0)[7] with drift         : 6346.759
## ARIMA(0,1,1)(0,0,1)[7] with drift         : 6395.033
## ARIMA(0,1,0)                              : 6466.527
## ARIMA(1,1,0)           with drift         : 6469
## ARIMA(1,1,0)(2,0,0)[7] with drift         : 6306.944
## ARIMA(1,1,0)(2,0,1)[7] with drift         : Inf
## ARIMA(1,1,0)(1,0,1)[7] with drift         : Inf
## ARIMA(0,1,0)(2,0,0)[7] with drift         : 6343.017
## ARIMA(2,1,0)(2,0,0)[7] with drift         : 6312.557
## ARIMA(1,1,1)(2,0,0)[7] with drift         : 6244.129
## ARIMA(1,1,1)(1,0,0)[7] with drift         : 6286.522
## ARIMA(1,1,1)(2,0,1)[7] with drift         : Inf
## ARIMA(1,1,1)(1,0,1)[7] with drift         : Inf
## ARIMA(0,1,1)(2,0,0)[7] with drift         : 6298.145
## ARIMA(2,1,1)(2,0,0)[7] with drift         : 6238.386
## ARIMA(2,1,1)(1,0,0)[7] with drift         : 6264.384
## ARIMA(2,1,1)(2,0,1)[7] with drift         : Inf
```

```
##  ARIMA(2,1,1)(1,0,1)[7] with drift         : Inf
##  ARIMA(3,1,1)(2,0,0)[7] with drift         : 6234.382
##  ARIMA(3,1,1)(1,0,0)[7] with drift         : 6258.62
##  ARIMA(3,1,1)(2,0,1)[7] with drift         : Inf
##  ARIMA(3,1,1)(1,0,1)[7] with drift         : Inf
##  ARIMA(3,1,0)(2,0,0)[7] with drift         : 6300.76
##  ARIMA(4,1,1)(2,0,0)[7] with drift         : Inf
##  ARIMA(3,1,2)(2,0,0)[7] with drift         : 6240.819
##  ARIMA(2,1,2)(2,0,0)[7] with drift         : 6243.815
##  ARIMA(4,1,0)(2,0,0)[7] with drift         : 6291.779
##  ARIMA(4,1,2)(2,0,0)[7] with drift         : Inf
##  ARIMA(3,1,1)(2,0,0)[7]                    : 6227.929
##  ARIMA(3,1,1)(1,0,0)[7]                    : 6252.359
##  ARIMA(3,1,1)(2,0,1)[7]                    : Inf
##  ARIMA(3,1,1)(1,0,1)[7]                    : Inf
##  ARIMA(2,1,1)(2,0,0)[7]                    : 6232.056
##  ARIMA(3,1,0)(2,0,0)[7]                    : 6294.277
##  ARIMA(4,1,1)(2,0,0)[7]                    : 6234.823
##  ARIMA(3,1,2)(2,0,0)[7]                    : 6234.373
##  ARIMA(2,1,0)(2,0,0)[7]                    : 6306.074
##  ARIMA(2,1,2)(2,0,0)[7]                    : 6237.558
##  ARIMA(4,1,0)(2,0,0)[7]                    : 6285.297
##  ARIMA(4,1,2)(2,0,0)[7]                    : Inf
##
##  Now re-fitting the best model(s) without approximations...
##
##  ARIMA(3,1,1)(2,0,0)[7]                    : 6244.074
##
##  Best model: ARIMA(3,1,1)(2,0,0)[7]

## Series: data.ts
## ARIMA(3,1,1)(2,0,0)[7]
##
## Coefficients:
##          ar1     ar2      ar3      ma1    sar1    sar2
##       0.5742  0.1332  -0.1068  -0.9754  0.3347  0.2233
## s.e.  0.0404  0.0473   0.0411   0.0128  0.0402  0.0416
##
## sigma^2 estimated as 769.1:  log likelihood=-3099.35
## AIC=6212.69   AICc=6212.87   BIC=6244.07
```

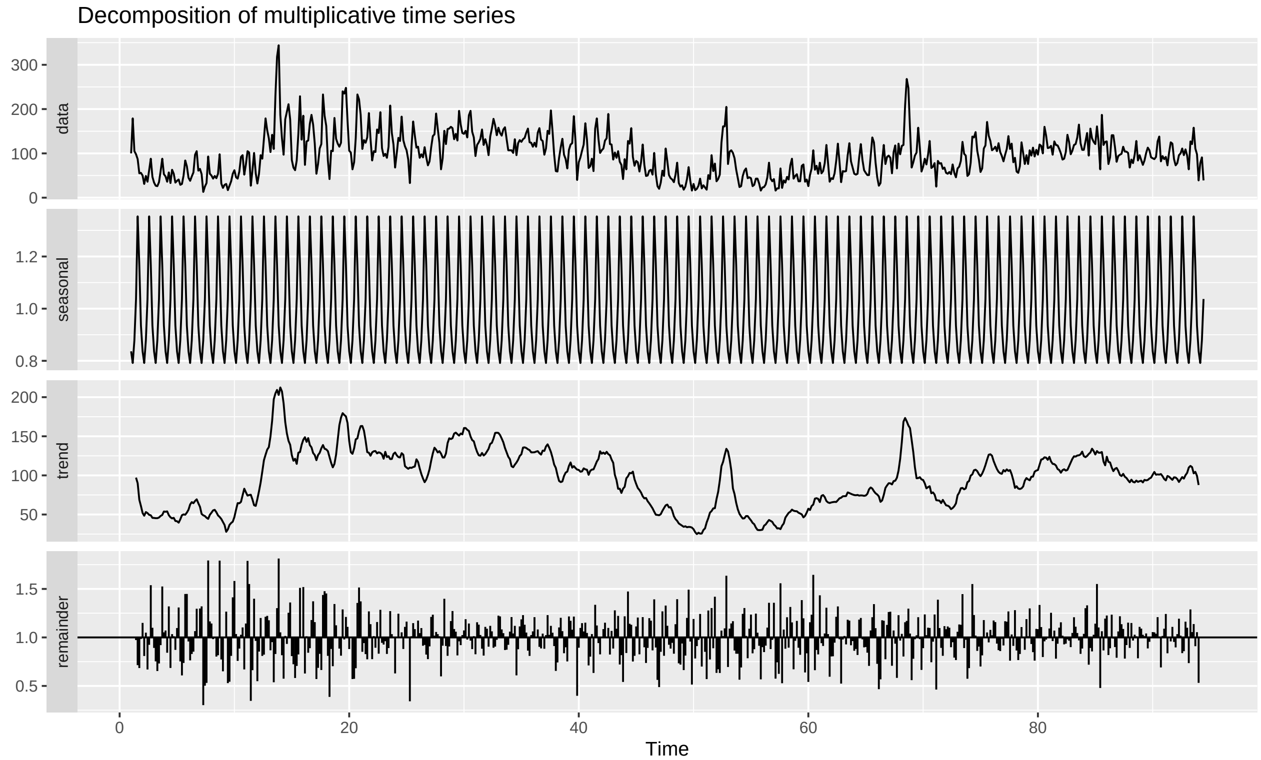


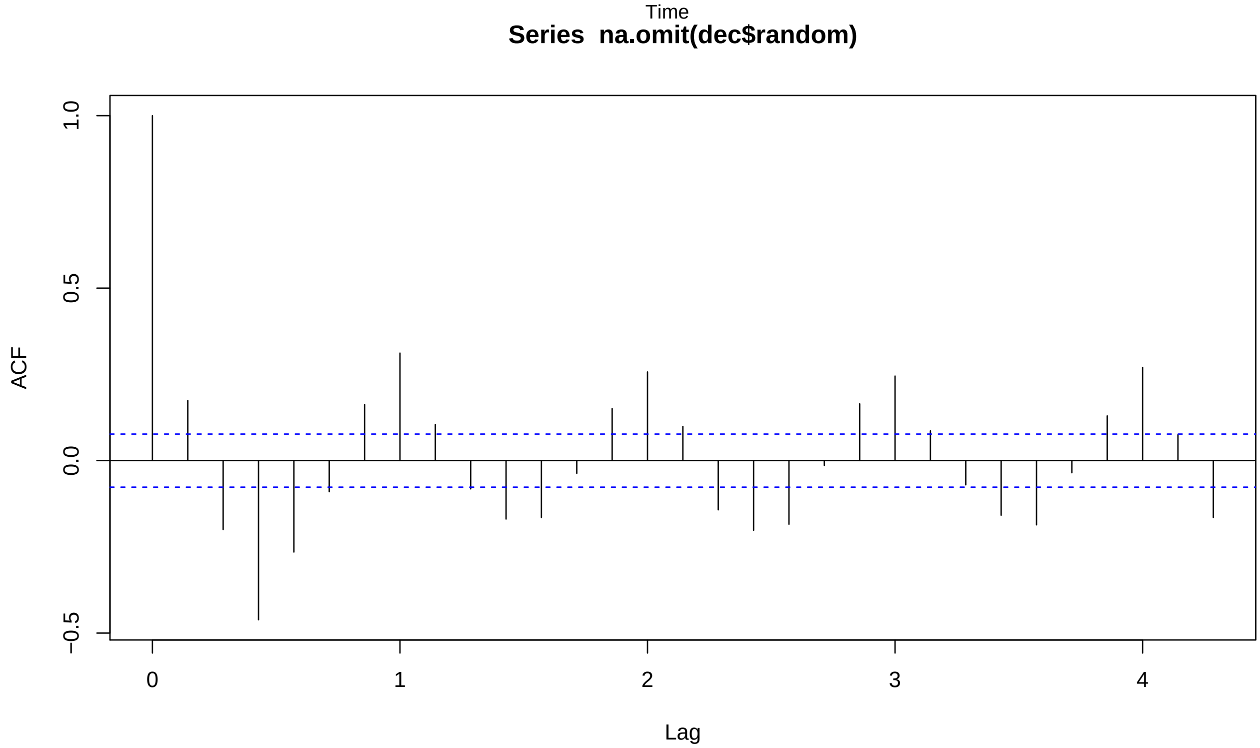
The plot is not good but AIC and BIC are very high, we should try with a multi seasonal decomposition

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

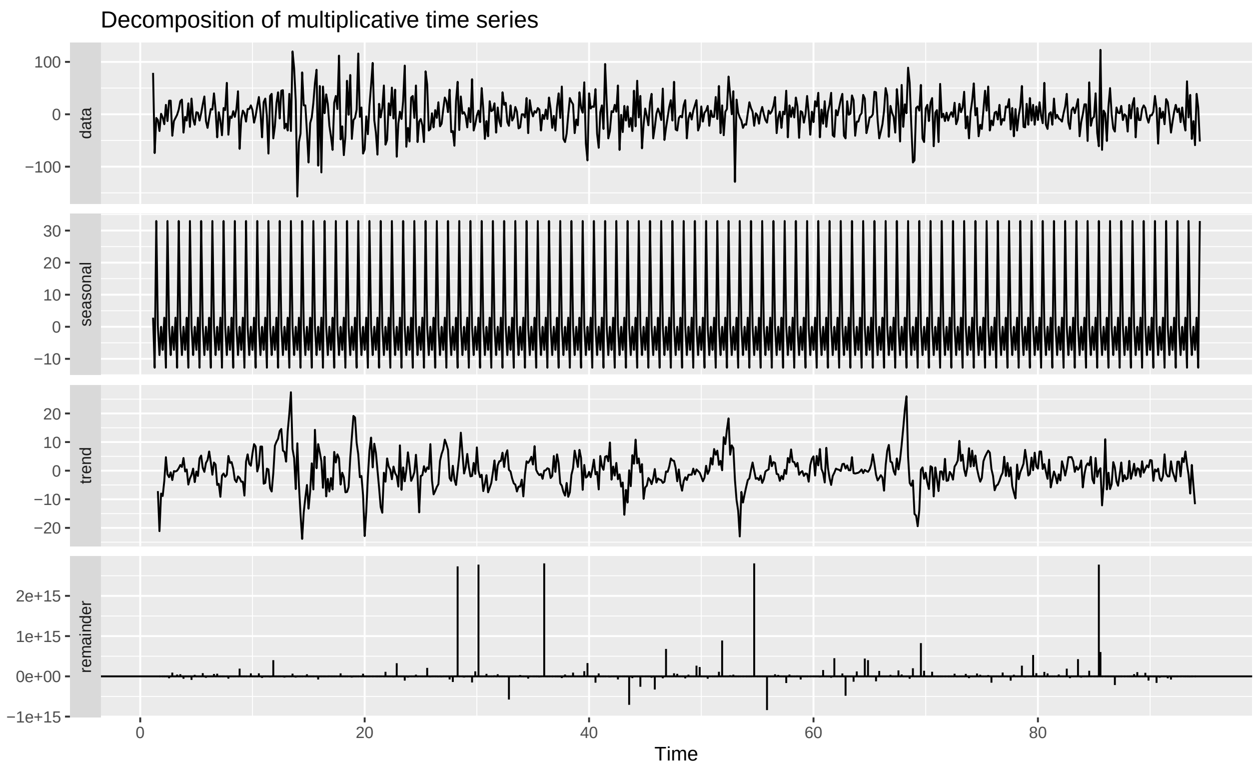
10 Searching for multi seasonalities

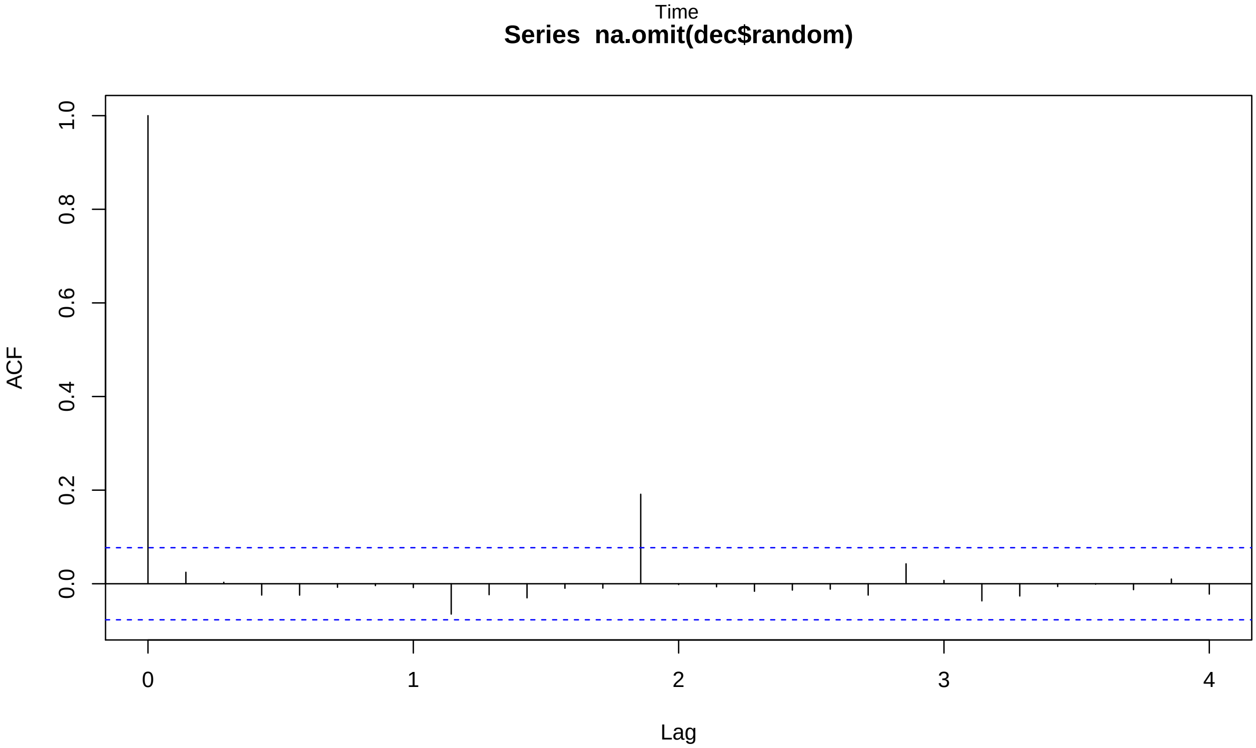
without differentiation residuals looks pretty bad





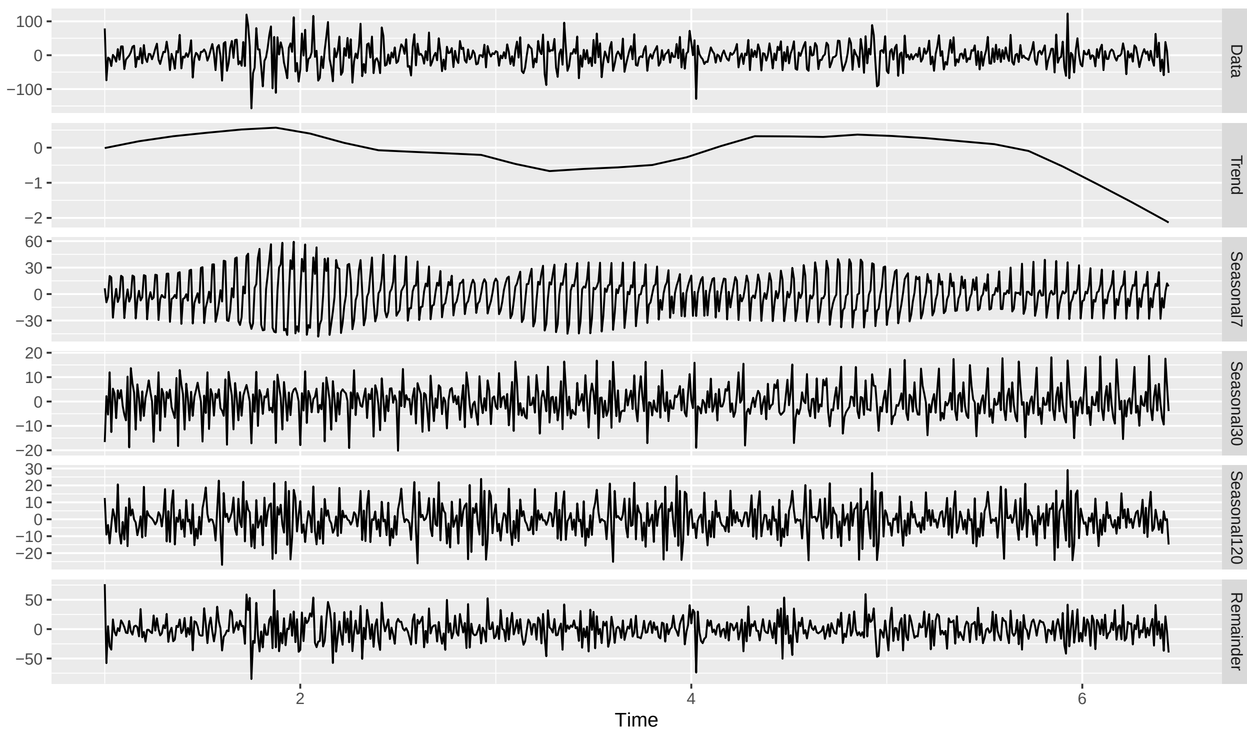
trying with differentiation and a multiplicative model:

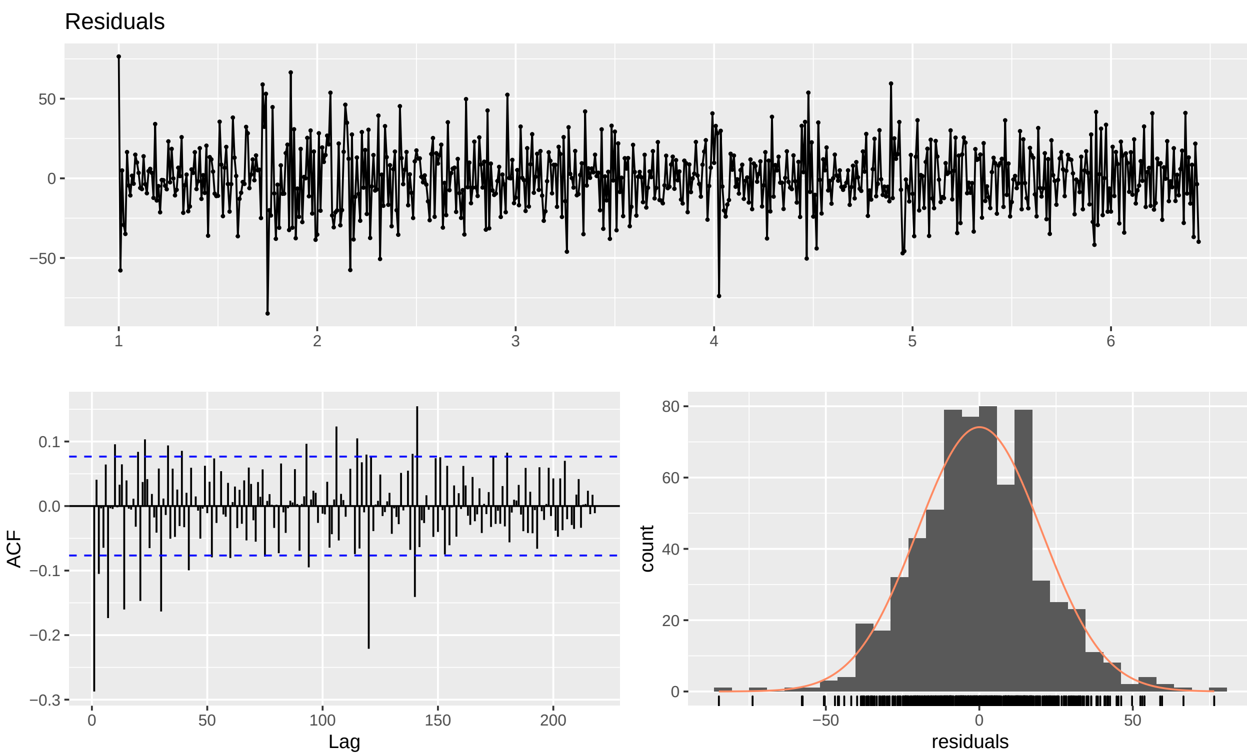




Looks better than before but we can still see every 5(*7) a seasonality/trend left. 5*7 is about a month, probably there is a monthly seasonality

11 Transforming into msts






```
##
##  Box-Pierce test
##
```

```
## data:  remainder(decomposed)
## X-squared = 65.062, df = 5, p-value = 1.088e-12

##
##  Box-Ljung test
##
## data:  remainder(decomposed)
## X-squared = 65.402, df = 5, p-value = 9.248e-13
```

12 Conclusions

It was really interesting!

13 TODO

prima diff, poi prima diff seasonal, check acf pacf, check no trend(trend se con decadono a 0 velocemente) identificare i picchi identificare l estate doppia seasonality una settimanale e una annuale ARCH GARCH VAR<—- stabilizzare con trasformazioni