Modelling daily ozonio mean

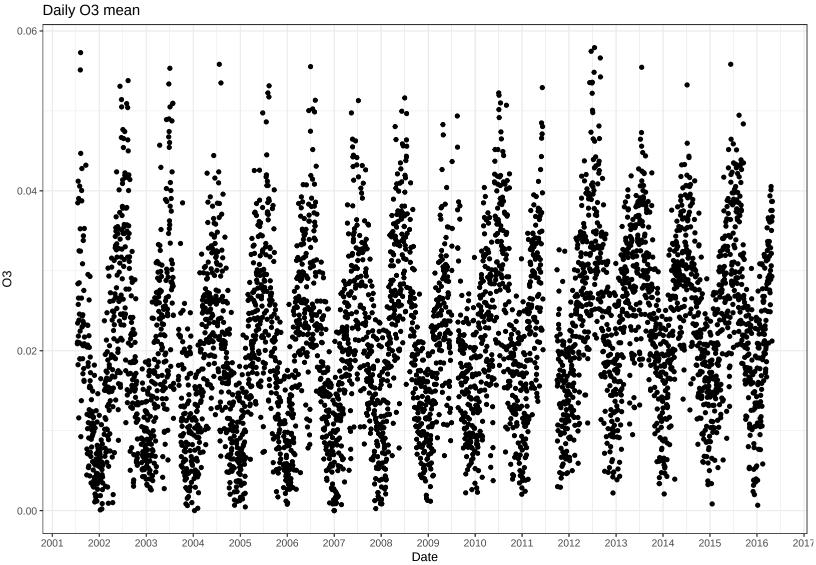
Giovani Valdrighi, Vitória Guardieiro

30/09/2020

Data

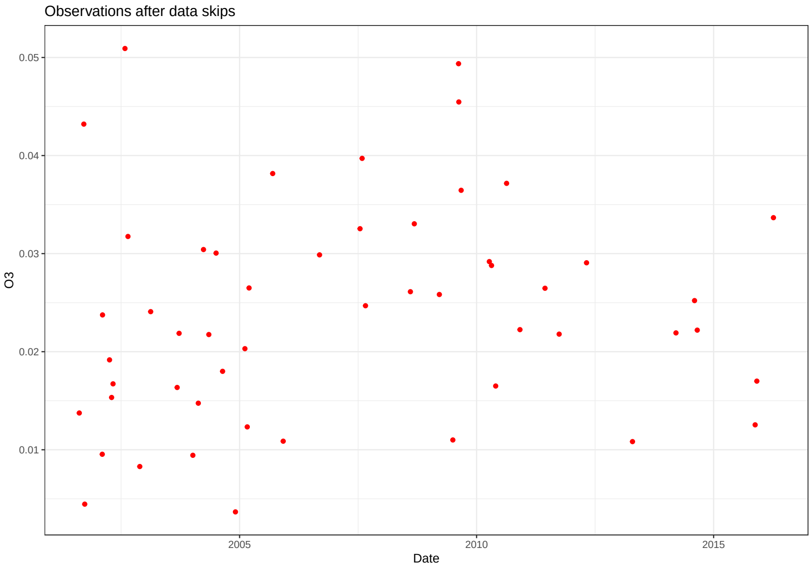
Daily data

- New York data from 15/07/2001 to 30/04/2016.

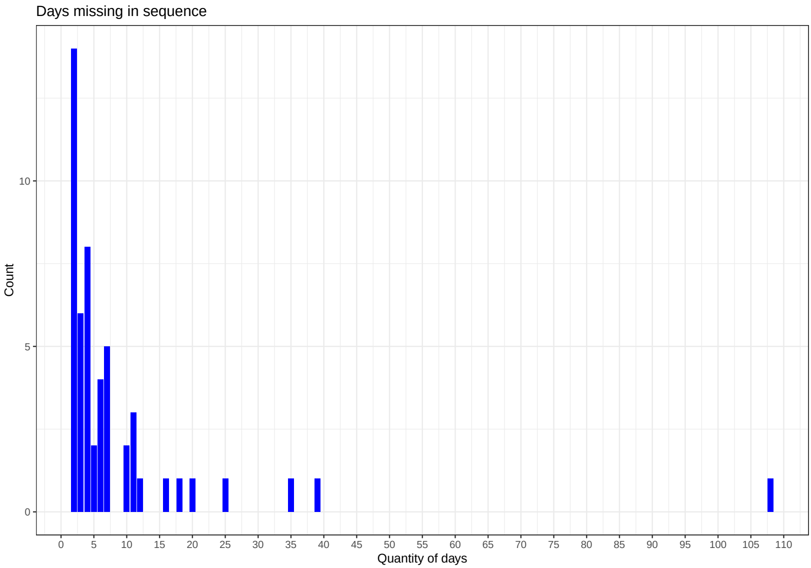


Missing data

- There are 52 time skips in the data, in a total of 473 days.
- The biggest skips is 108 days in 2011.
- The majority of skips are of 1 or 2 days.
- Around 9.5% missing data.
- The missing observations are distributed along the time without a clear pattern.

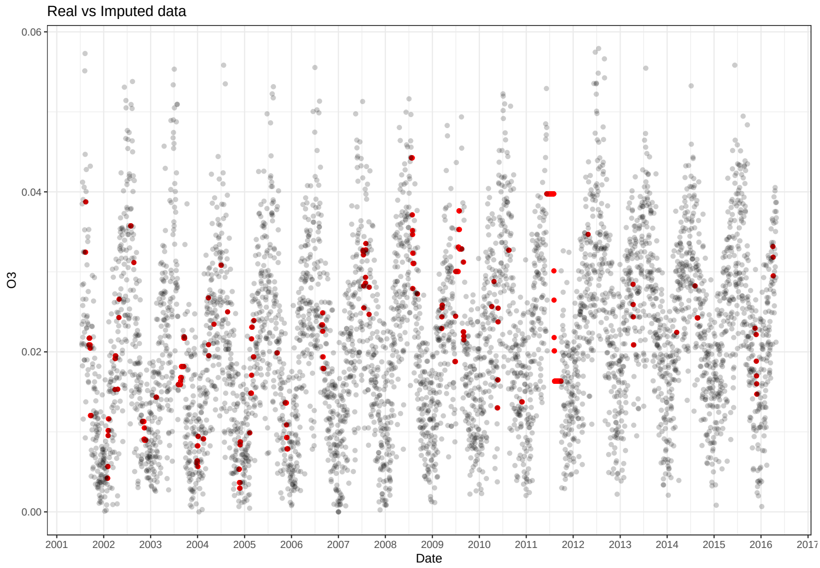


Observations after data skips
O3
0.05
0.04
0.03
0.02
0.01
2005
2010
2015
Date


Days missing in sequence
Count
10
5
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
Quantity of days

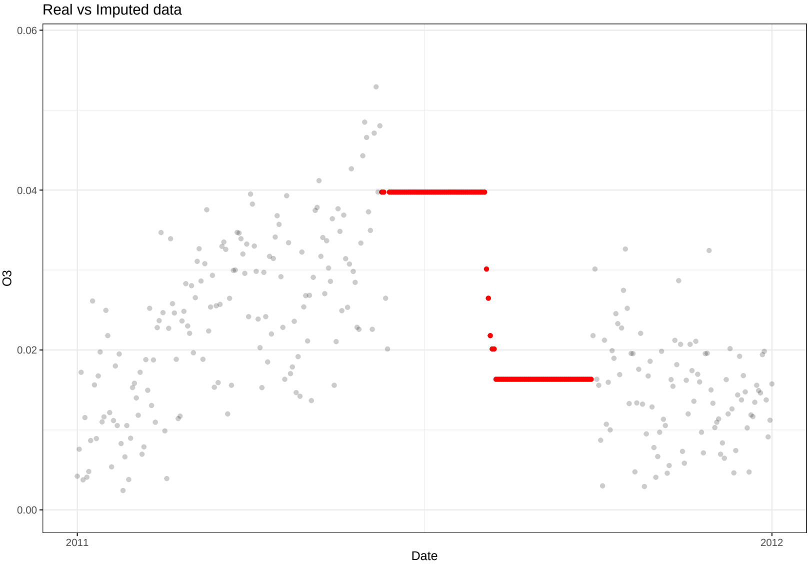
Imputation method

- It was used the kNN method to imputate values on missing observations.
- The kNN method needs the parameter k, the number of closest points considered.
- Starting with $k = 7$.

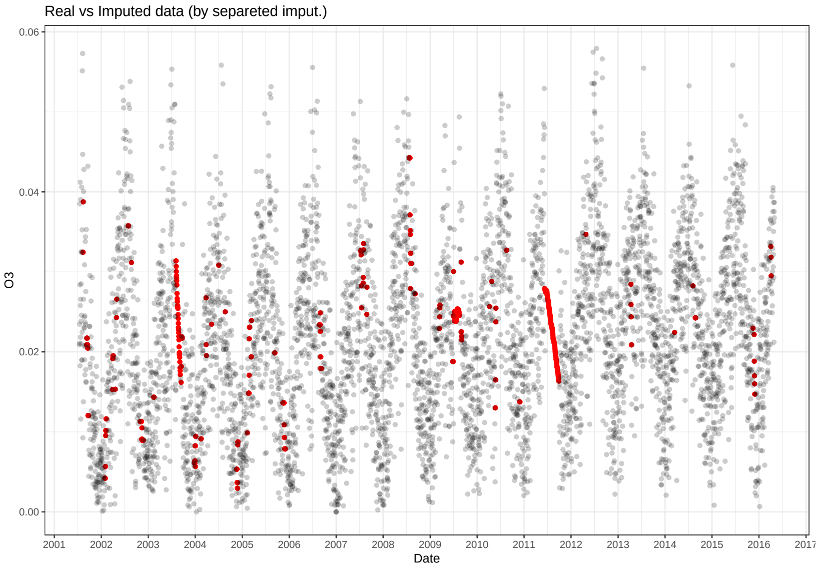


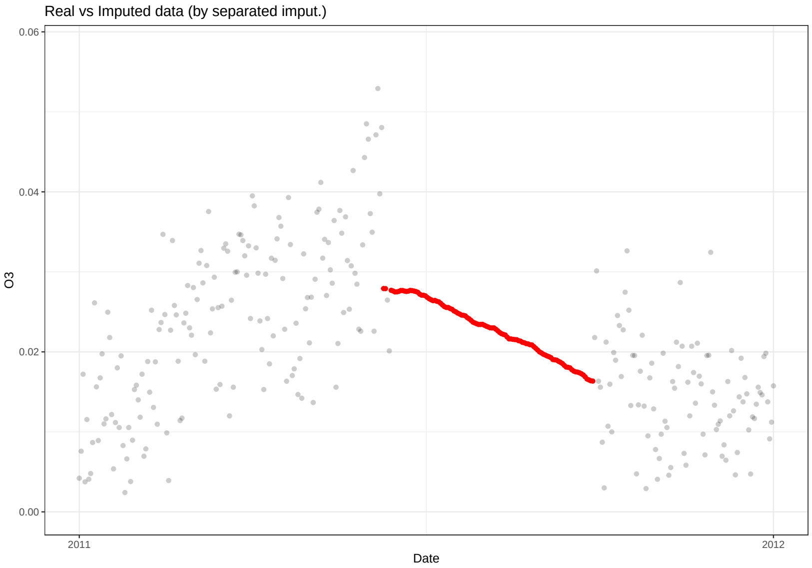
Real vs Imputed data
O3
0.00
0.02
0.04
0.06
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Date

- Method create a bad behavior when the size of the skips is bigger than 7 days.



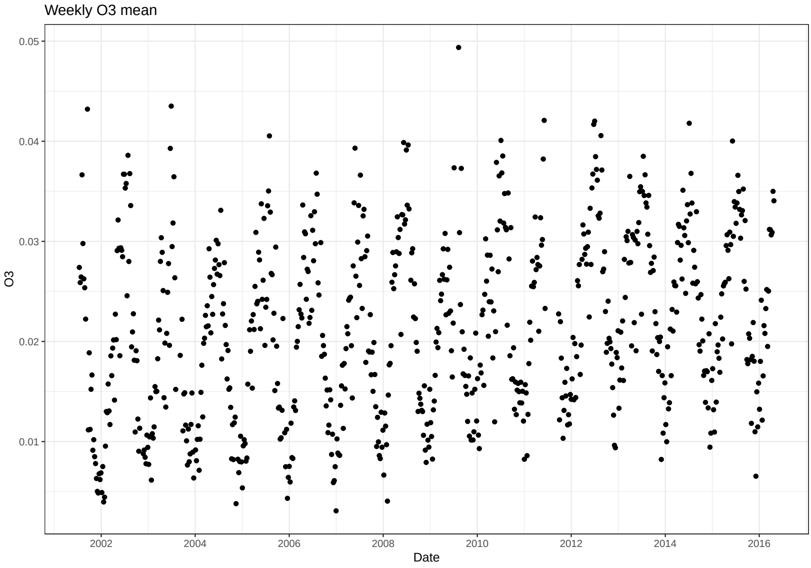
- To deal with this, the parameter k used for imputation will be different if the size of the skip is minor them 30 days, between 30 days and 100 days, or bigger than 100 days.
- $k = 7$, $k = 45$, $k = 120$, respectively.
- We will aggregate closest points by weighted by distance mean.


Real vs Imputed data (by separeted imput.)
O3
0.06
0.04
0.02
0.00
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Date



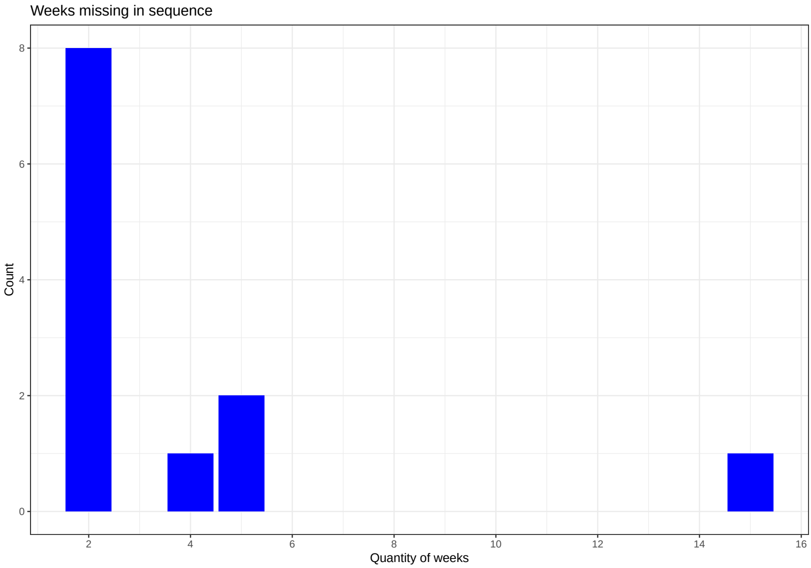
Real vs Imputed data (by separated imput.)
0.06
0.04
0.02
0.00
O3
2011
2012
Date

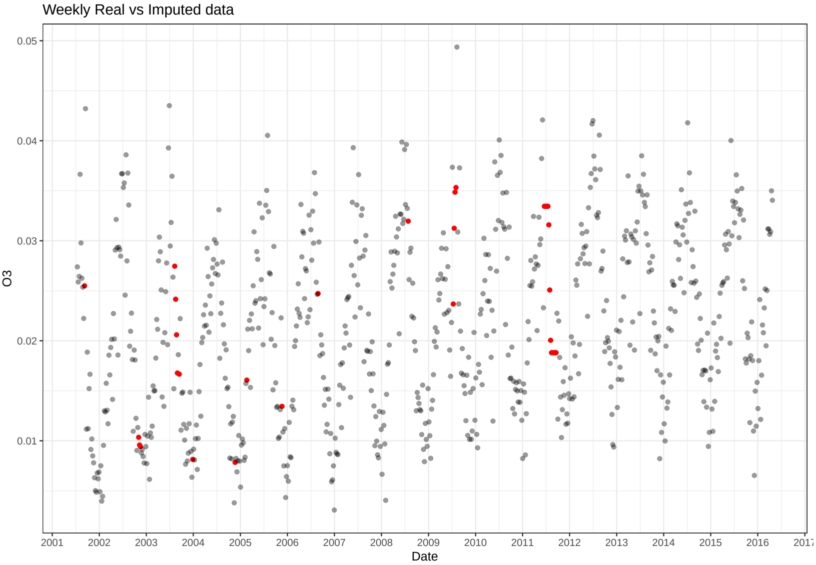
Weekly data



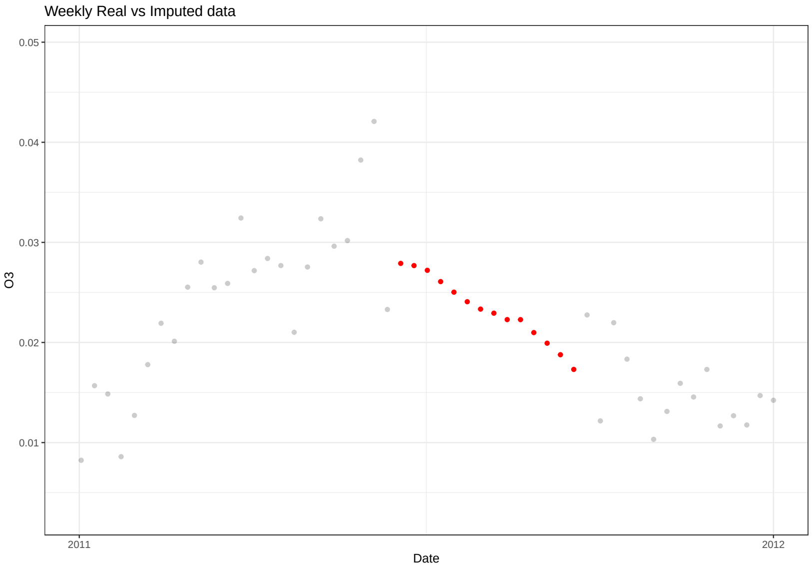
Weekly O3 mean
0.05
0.04
0.03
0.02
0.01
O3
2002
2004
2006
2008
2010
2012
2014
2016
Date

- If the data is grouped by week, ignoring the missing values when aggregating, it'll have 33 missing observations.
- Around 4.3% missing data.





- It has the same problem when the sequence of missing data is to big.



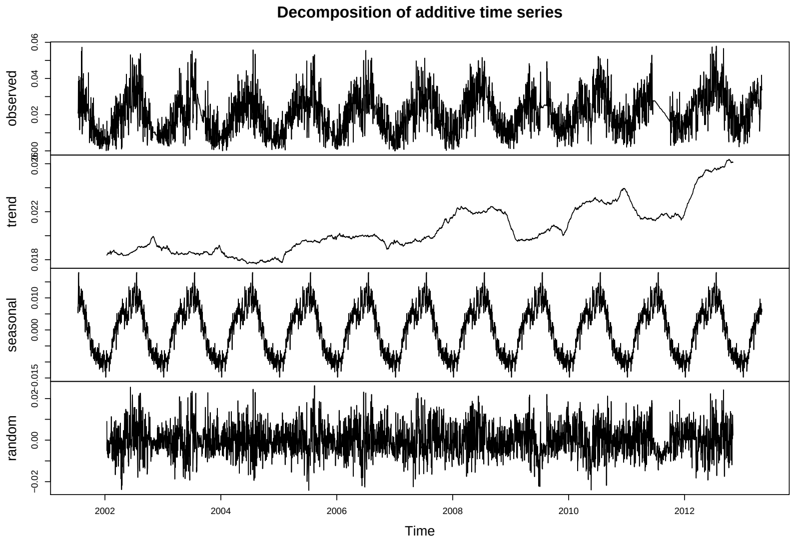
Weekly Real vs Imputed data
0.05
0.04
0.03
0.02
0.01
O3
2011
2012
Date

Daily model

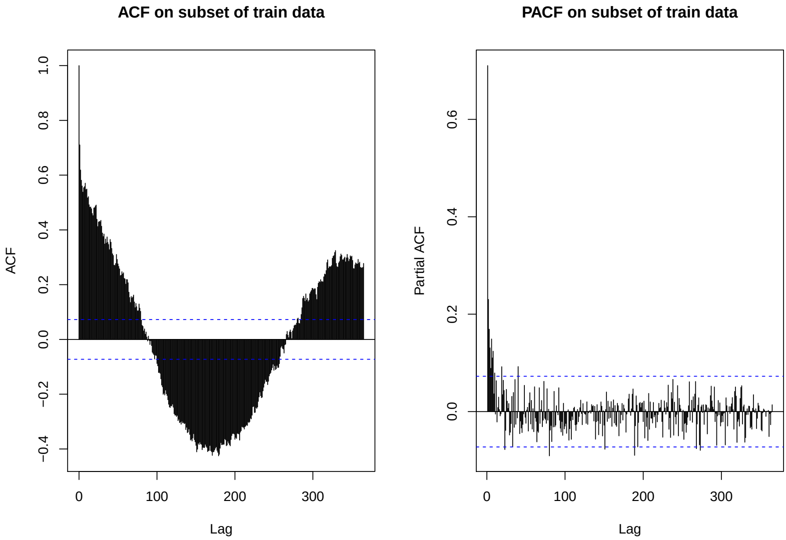
Modelling process

- Metric to be minimized: $\text{MAE} = \frac{1}{n} \sum_n |y_t - \hat{y}_t|$.
- Rolling window of 2 years (730 days).
- Prediction of the next 7 days.
- First: Test if there is tendency with Wald-Wolfowitz runs test.
  - For every 2 years window, the p-value is smaller than $1e-3$.
- Second: Fitting of different models and evaluation of MAE error.

Choice of models - trend



Choice of models - ACF and PACF



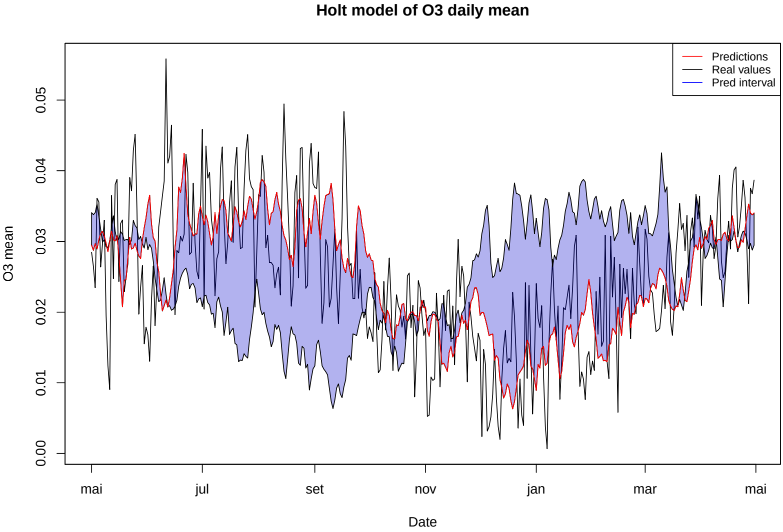
- Naive model: the next 7 days are predict as the mean of the last 4 weeks.
- Exponential smoothing forecast.
- Holt model with trend.
- ARMA(6,0) model.
- Auto ARIMA model.

- Process:
  - 1. For each model, run a 2 years window, for each:
    - Fit model.
    - Generate predictions of next 7 days.
    - Compute mean of residuals for that window.
  - 2. Compute MAE for model as the mean of residuals.

- Results for train data:
  - Auto ARIMA model: 0.005986731
  - Holt model: 0.006142857
  - SES model: 0.00617229
  - ARMA(6,0) model: 0.006279533
  - Naive model: 0.007498889

Evaluating on test data

- MAE: 0.006503142

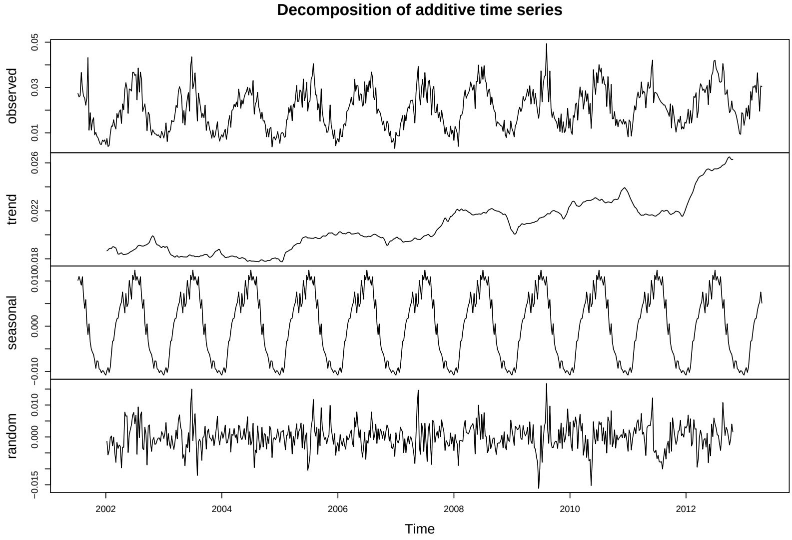


Weekly model

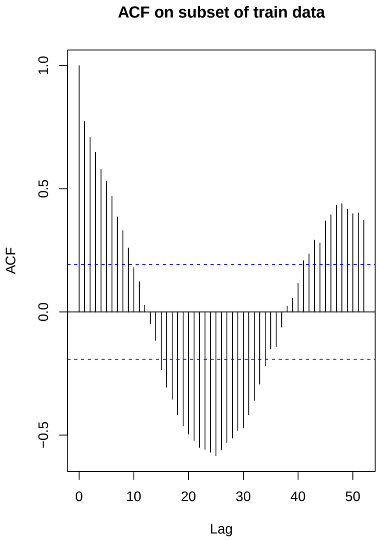
Modelling process

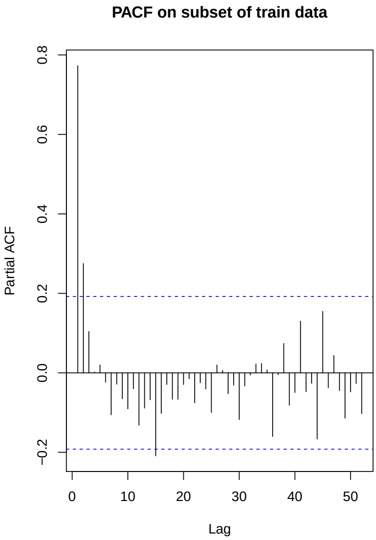
- Metric to be minimized: MAE $= \frac{1}{n} \sum_n |y_t - \hat{y}_t|$.
- Rolling window of 2 years (104 weeks), by skiping 4 weeks.
- Prediction of the next 4 weeks.
- First: Test if there is tendency with Wald-Wolfowitz runs test.
    - For almost every 2 years window, the p-value is bigger than 0.1.
- Second: Fitting of different models and evaluation of MAE error.

Choice of models - trend



Choice of models - ACF and PACF





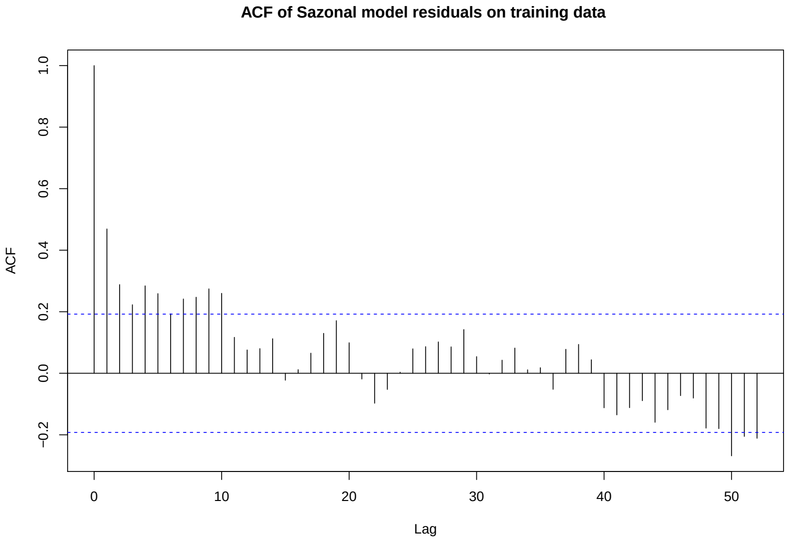
- Baseline model: the next 4 weeks are predict as the mean of the last 4 weeks.
- Seasonal model: linear regression on seasonal dummies variable, each month is a factor.
- Linear model: linear regression on seasonal dummies and time index.
- Poly 2 model: linear regression on seasonal dummies and time index with degree 1 and 2.
- Poly 3 model: linear regression on seasonal dummies and time index with degree 1, 2, and 3.
- Holt Winters model without trend and with seasonality (multiplicative and additive).
- ARMA(1, 0) model.

- Process:
  - 1. For every 2 years window:
    - Fit all the models.
    - Generate predictions of next 4 weeks.
  - 2. With predictions for every week, compute residuals $r_t = y_t - \hat{y}_t$.
  - 3. With residuals, compute MAE.
    - We group the predictions by each window, in this subsets, we compute the MAE of the 4 weeks predicted, than, we compute the mean of the MAE for all windows.
    - We also group the predictions by the numbers of weeks after the last observation, that range from 1 to 4, and compute the MEAN for each of this subset.

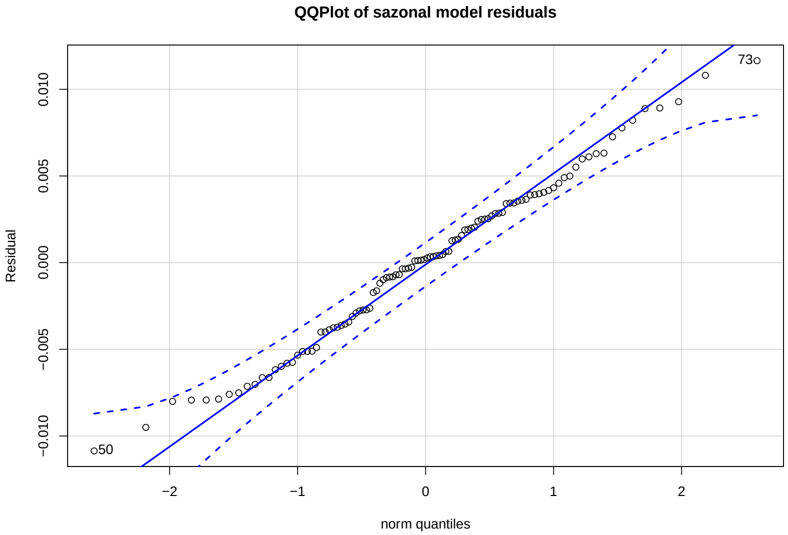
- Results:
  - Seasonal: 0.003887602
  - Linear: 0.004033049
  - Poly 2: 0.004148709
  - Arma(1, 0): 0.004695954
  - Poly 3: 0.004760530
  - HoltWinters additive: 0.005016573
  - HoltWinters multiplicative: 0.005106128
  - Baseline: 0.005218411

```
## # A tibble: 4 x 9
##     day baseline sazonal  linear  poly_2  poly_3  hw_add
##   <int>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1     1  0.00429 0.00378 0.00385 0.00390 0.00421 0.00475
## 2     2  0.00493 0.00386 0.00401 0.00409 0.00466 0.00494
## 3     3  0.00552 0.00393 0.00408 0.00423 0.00487 0.00507
## 4     4  0.00613 0.00398 0.00418 0.00438 0.00531 0.00531
```

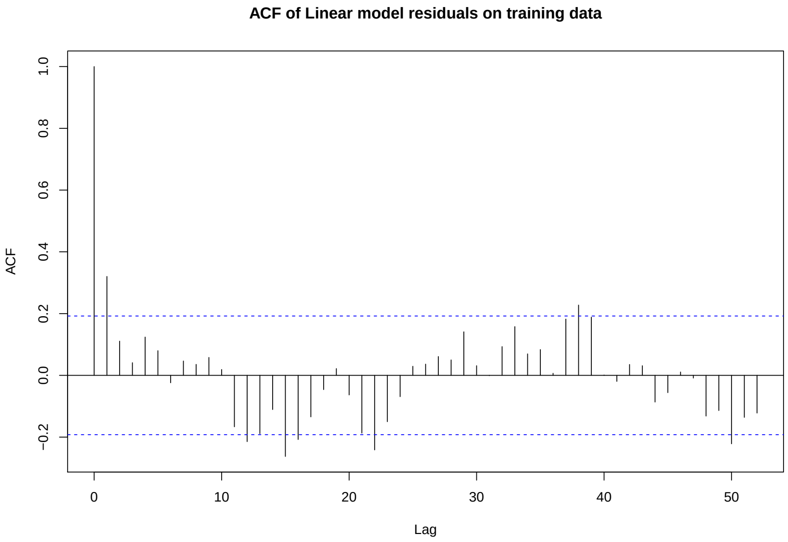
Residuals



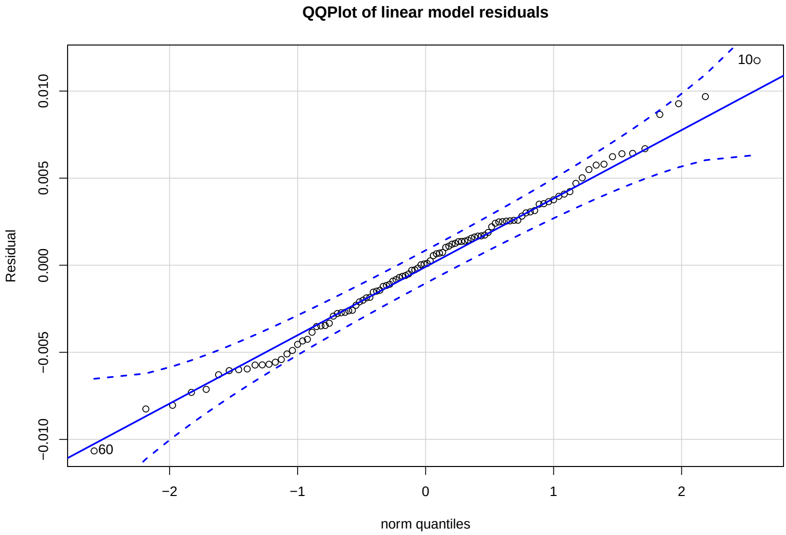
ACF of Sazonal model residuals on training data
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
10
20
30
40
50
Lag


QQPlot of sazonal model residuals
Residual
0.010
0.005
0.000
−0.005
−0.010
−2
−1
0
1
2
norm quantiles
73
50

```
## [1] 73 50
```



ACF of Linear model residuals on training data
ACF
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
10
20
30
40
50
Lag


QQPlot of linear model residuals
Residual
0.010
0.005
0.000
−0.005
−0.010
−2
−1
0
1
2
norm quantiles
10
60

```
## [1] 10 60
```

Evaluating on test data

- MAE: 0.003476891
- MAE by day: 1 - 0.003450016; 2 - 0.003388719; 3 - 0.003502532; 4 - 0.003566297