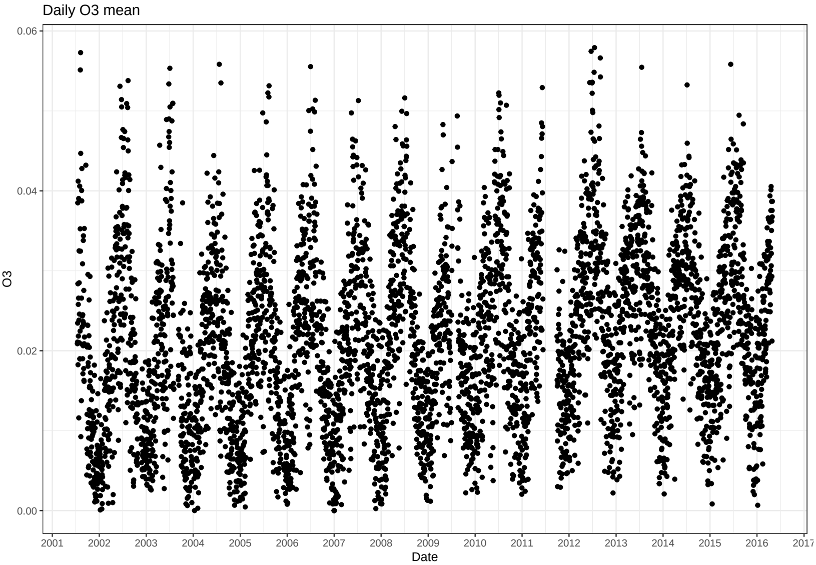
Modelling daily ozonio mean

Giovani Valdrighi, Vitória Guardieiro

24/09/2020

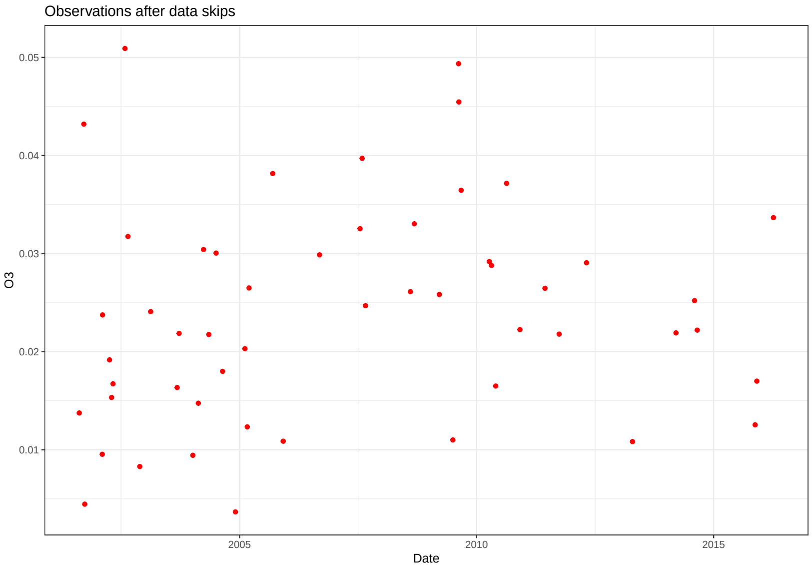
Data

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

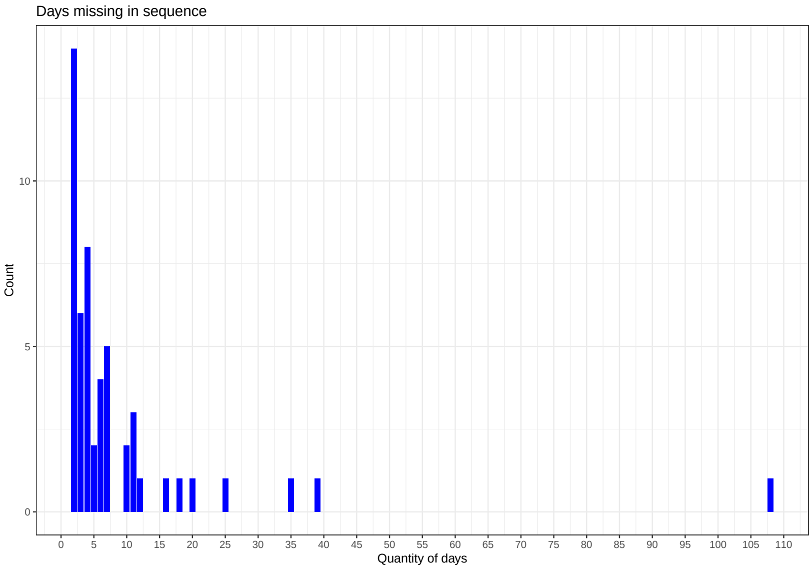


Missing data

- 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.

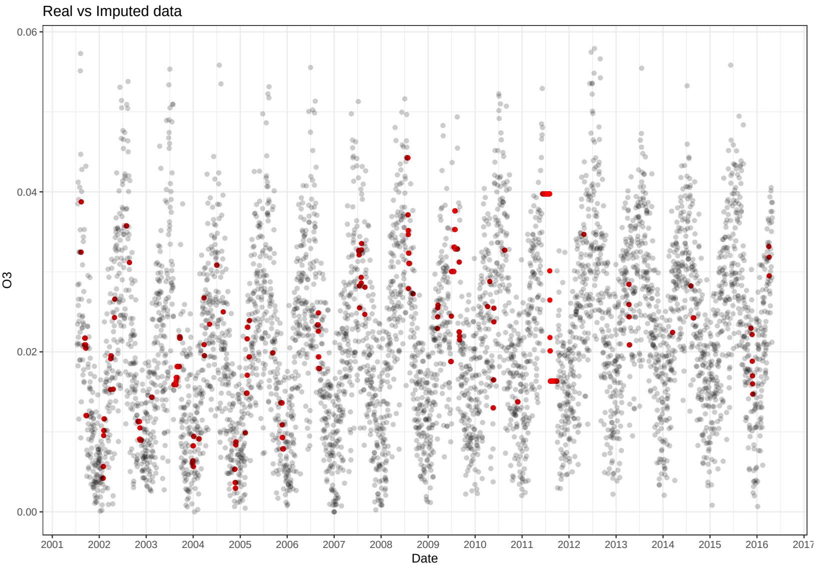


Observations after data skips
0.05
0.04
0.03
0.02
0.01
O3
2005
2010
2015
Date


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

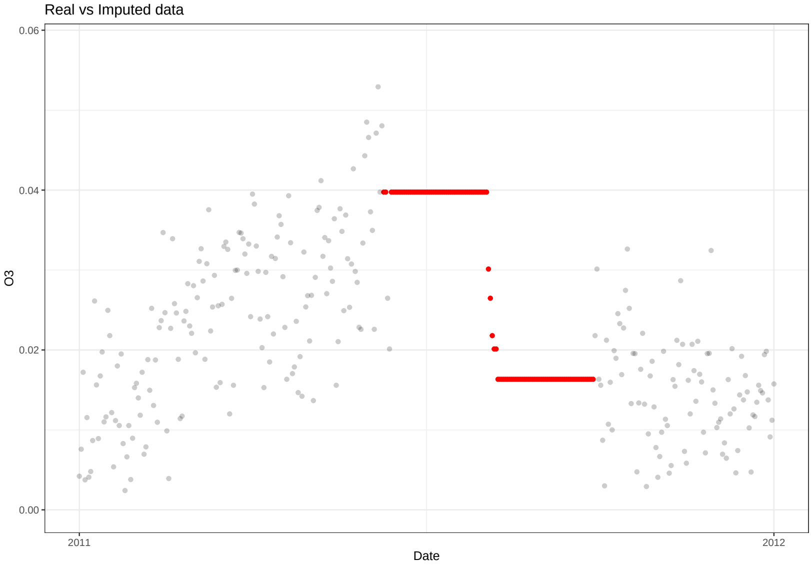
Imputation method

- 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 $k = 7$.

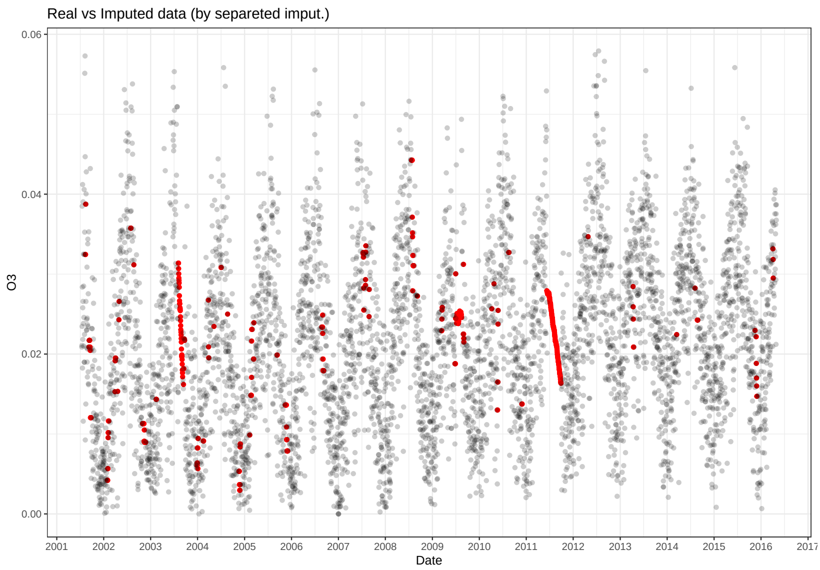


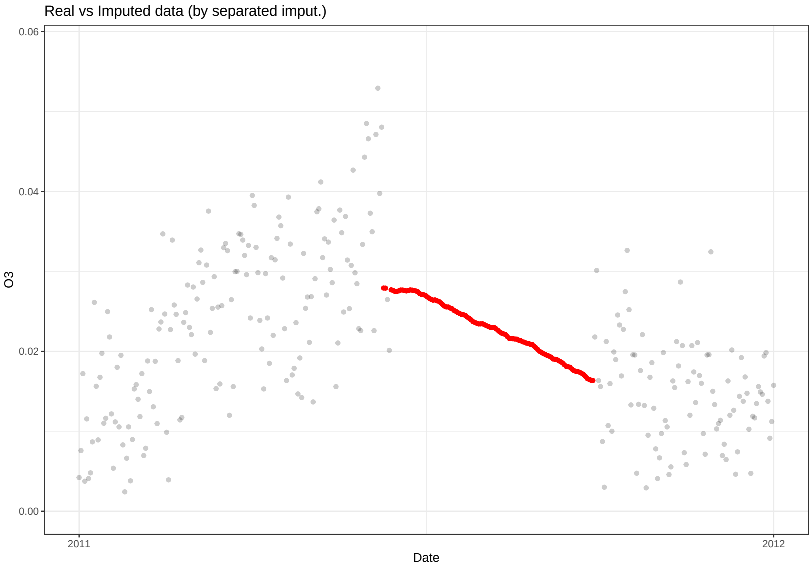
Real vs Imputed data
0.06
0.04
0.02
0.00
O3
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Date

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

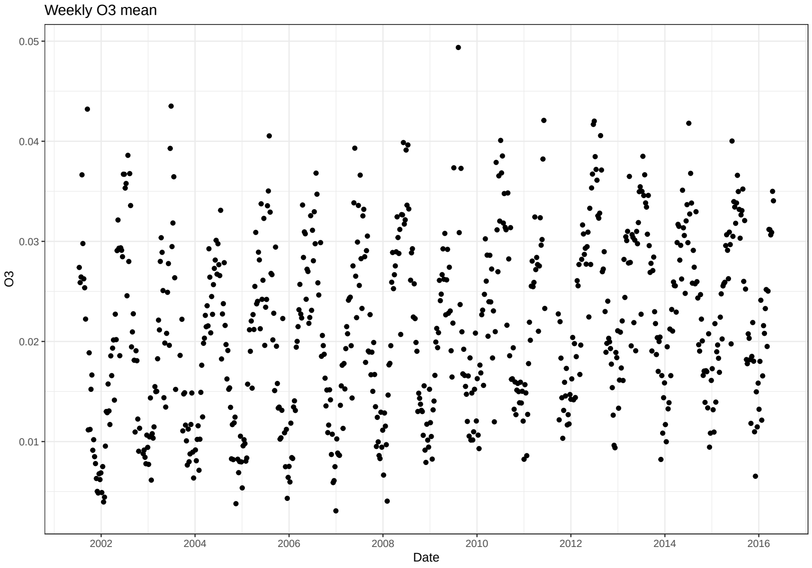


- To deal with this, the parameter k used for imputation will be different if the size of the skip is minor tem 30 days, between 30 days and 100 days, or bigger than 100 days.
- $k = 7$, $k = 45$, $k = 120$, respectively.
- We will aggregate closest points by weighted by distance mean.


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

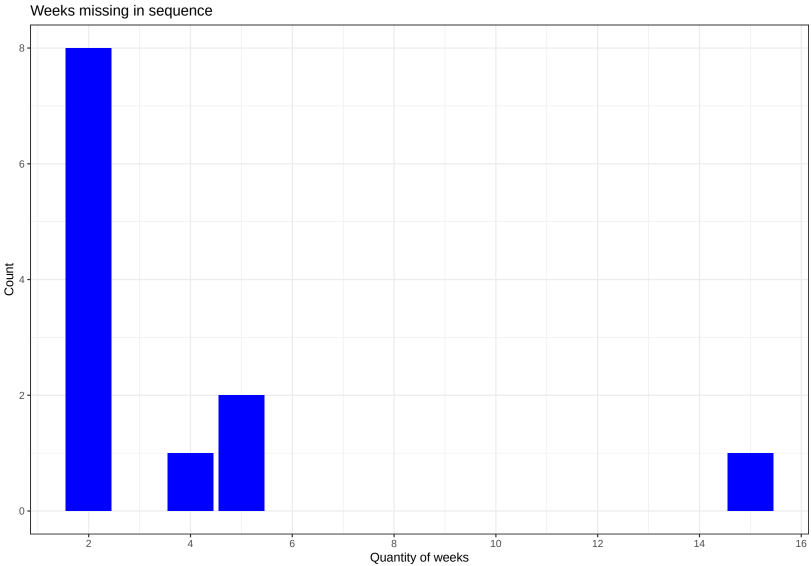

Real vs Imputed data (by separated imput.)
O3
0.06
0.04
0.02
0.00
Date
2011
2012

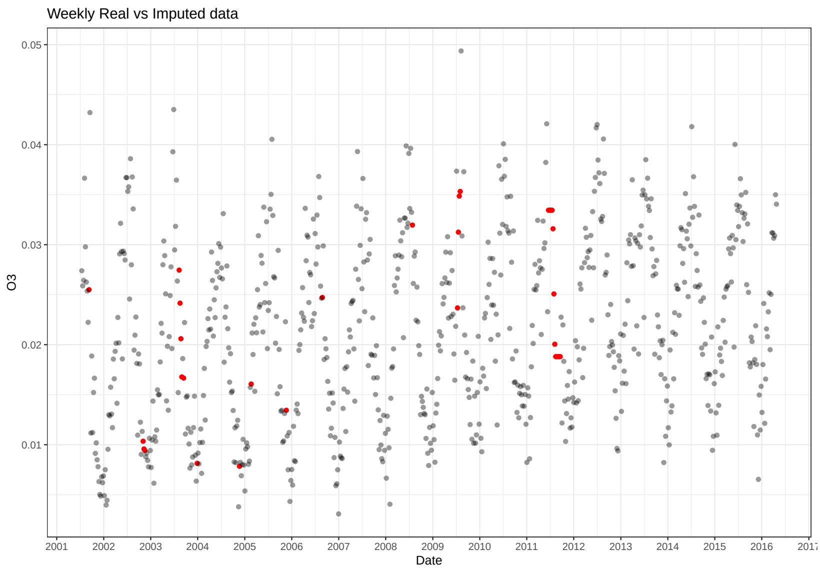
Weekly data



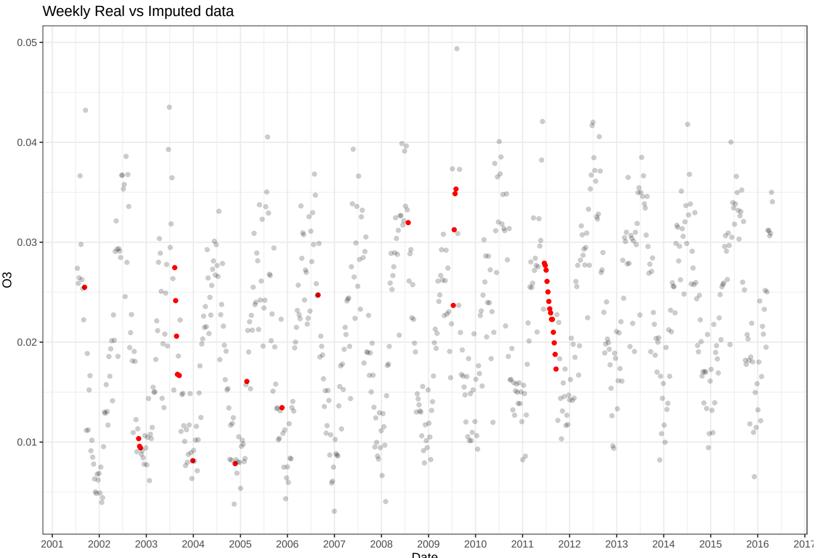
Weekly O3 mean
O3
0.05
0.04
0.03
0.02
0.01
2002
2004
2006
2008
2010
2012
2014
2016
Date

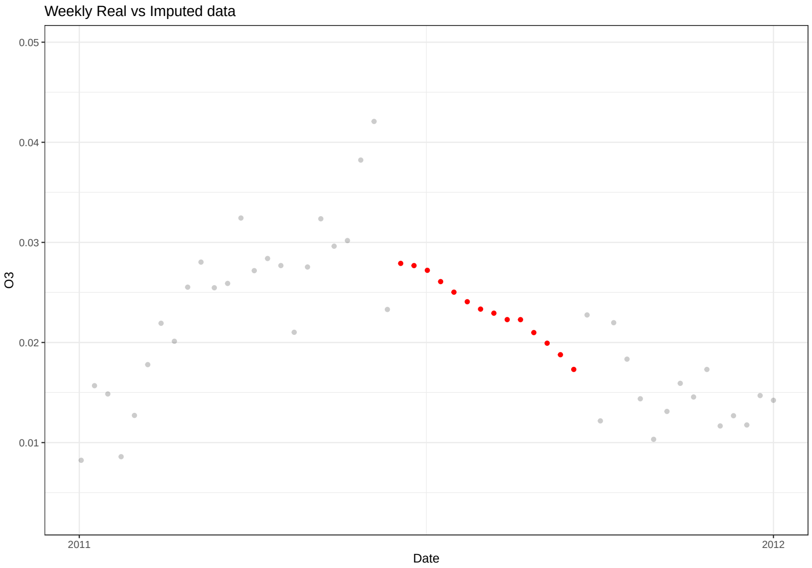
- If the data is grouped by week, ignoring the missing values when aggregating, it'll have 33 missing observations.
- Around 4.3% missing data.




Weekly Real vs Imputed data
O3
0.05
0.04
0.03
0.02
0.01
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Date

- It has the same problem when the sequence of missing data is to big.
- Again, if there is more than 5 missing weeks, it will be used k = 16, if it's less, it'll be k = 4.





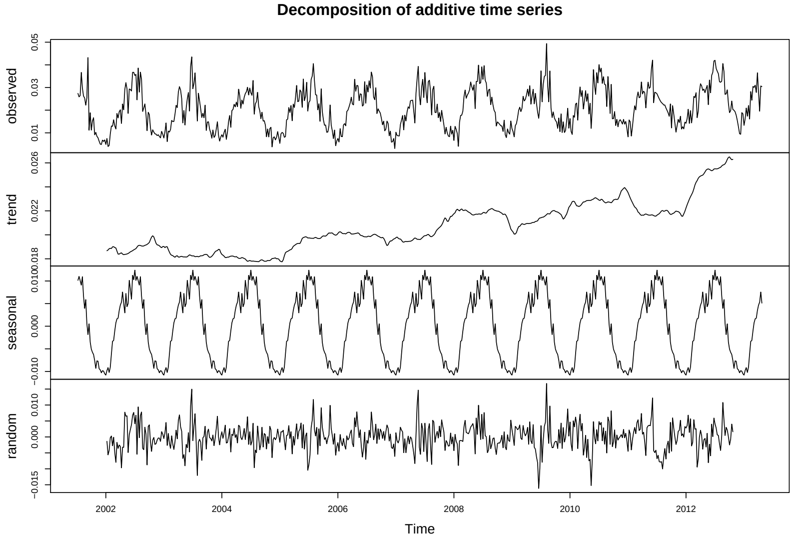
Weekly Real vs Imputed data
0.05
0.04
0.03
0.02
0.01
O3
2011
2012
Date

Weekly model

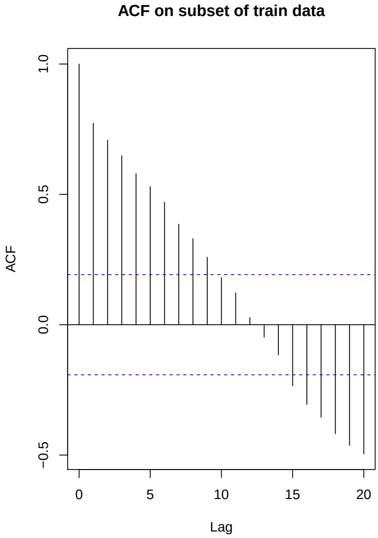
Modelling process

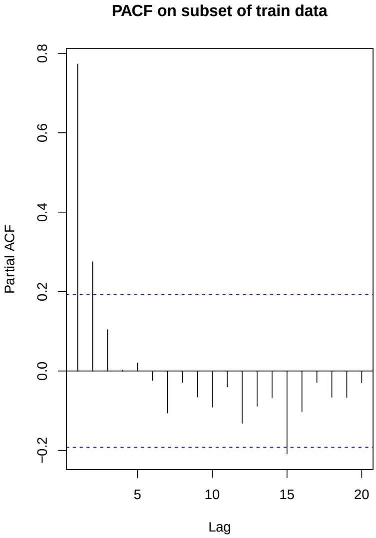
- Metric to be minimized: MAE $= \frac{1}{n} \sum_n |y_t - \hat{y}_t|$.
- Rolling window of 2 years (104 weeks), by skiping 4 weeks.
- Prediction of the next 4 weeks.
- First: Test if there is tendency with Wald-Wolfowitz runs test.
  - For every 2 years window, the p-value is smaller than $1e - 3$.
- Second: Fitting of different models and evaluation of MAE error.

Choice of models - trend



Choice of models - ACF and PACF



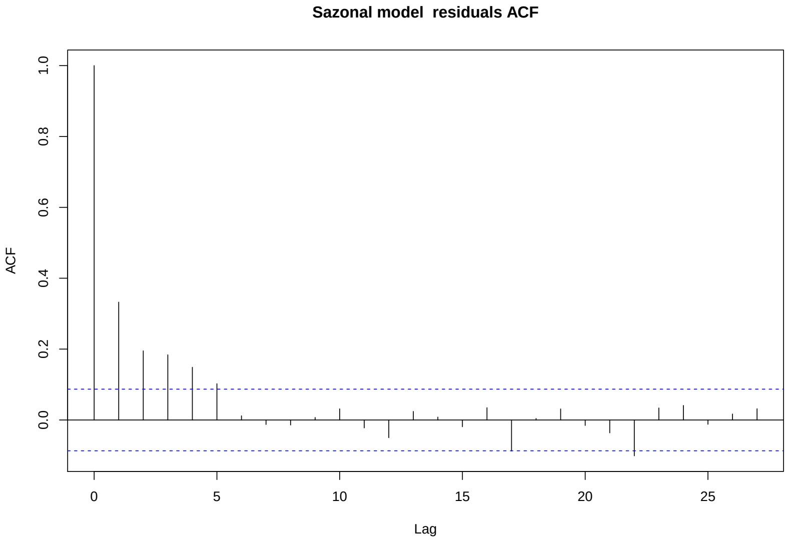


- Naive model: the next 4 weeks are predict as the mean of the last 4 weeks.
- Seasonal model: linear regression on seasonal dummies variable, each month is a factor.
- Linear model: linear regression on seasonal dummies and time index.
- Poly 2 model: linear regression on seasonal dummies and time index with degree 1 and 2.
- Poly 3 model: linear regression on seasonal dummies and time index with degree 1, 2, and 3.
- Holt model with trend.
- Holt Winters model with trend and seasonality (multiplicative and addtive).
- ARMA(1, 0) model.

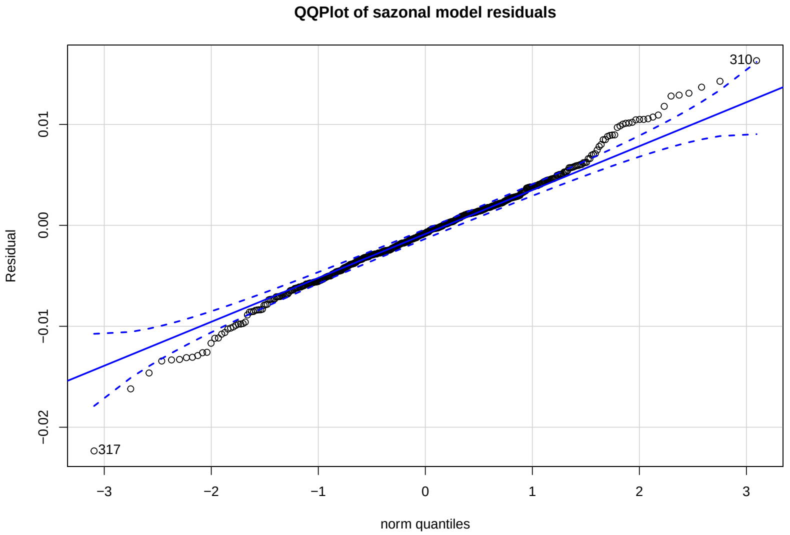
- Process:
  - 1. For every 2 years window:
    - Fit all the models.
    - Generate predictions of next 4 weeks.
  - 2. With predictions for every week, compute residuals $r_t = y_t - \hat{y}_t$.
  - 3. With residuals, compute MAE.

- Results:
  - Sazonal: 0.003903618
  - Linear: 0.004032514
  - Poly 2: 0.004171369
  - Arma(1, 0): 0.004568415
  - Poly 3: 0.004739532
  - Holt: 0.004885386
  - HoltWinters additive: 0.005008383
  - HoltWinters multiplicative: 0.005085043
  - Naive: 0.005122260

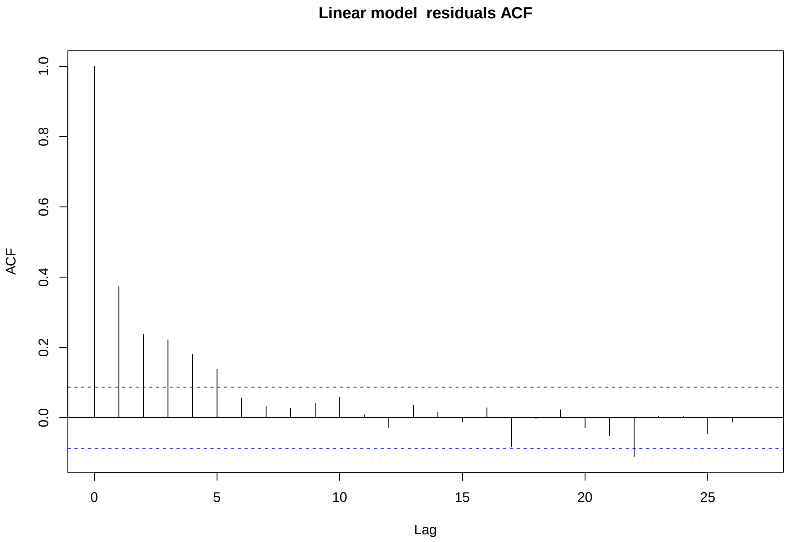
Residuals



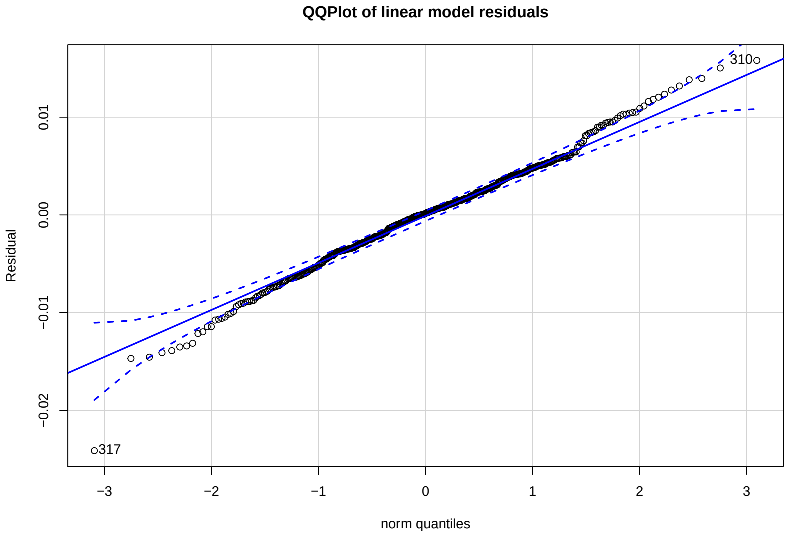
Sazonal model residuals ACF
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
Lag


QQPlot of sazonal model residuals
Residual
norm quantiles
0.01
0.00
−0.01
−0.02
−3
−2
−1
0
1
2
3
310
317

```
## [1] 317 310
```



Linear model residuals ACF
ACF
Lag
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25



QQPlot of linear model residuals
Residual
norm quantiles
0.01
0.00
−0.01
−0.02
−3
−2
−1
0
1
2
3
310
317


```
## [1] 317 310
```

Evaluating on test data

- MAE: 0.003438587