

Modelling daily ozonio mean

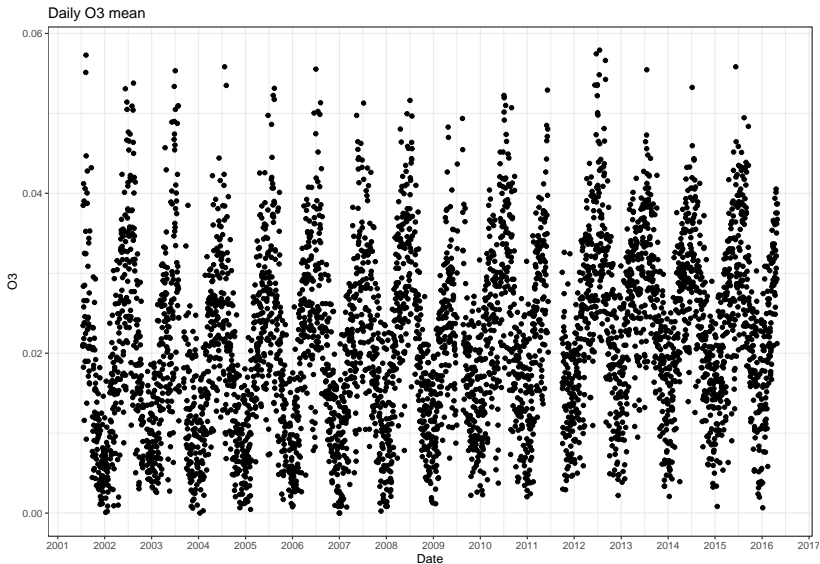
Giovani Valdrighi, Vitória Guardieiro

30/09/2020

Data

Daily data

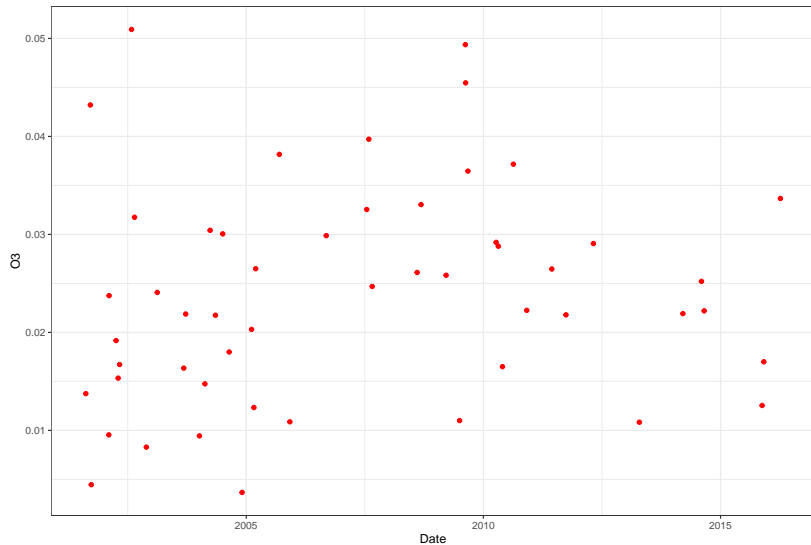
- New York data from 15/07/2001 to 30/04/2016.



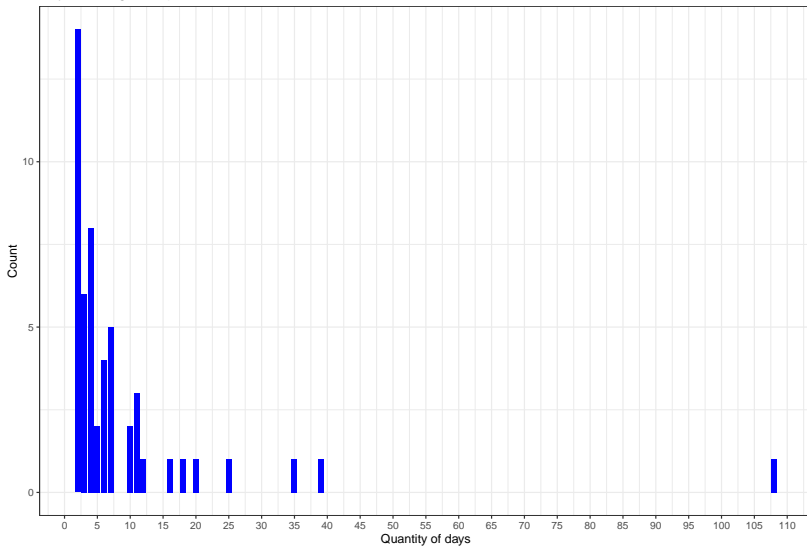
Missing data

- ▶ 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



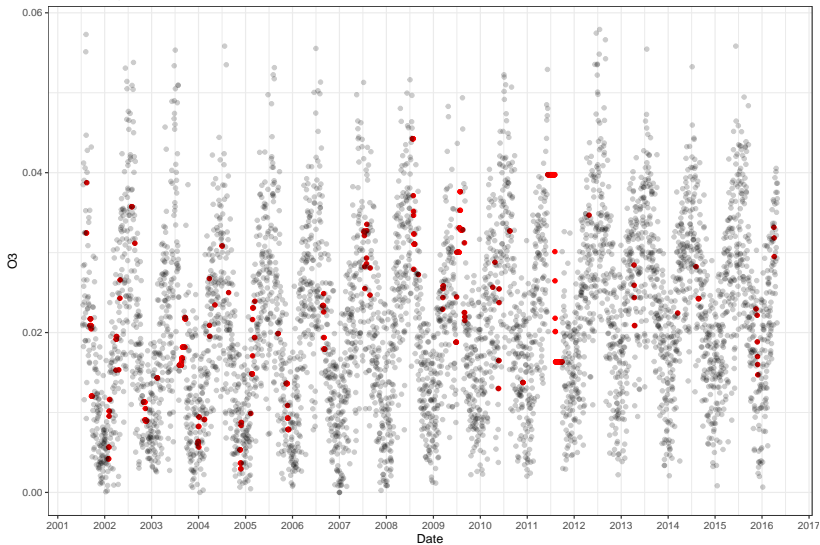
Days missing in sequence



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

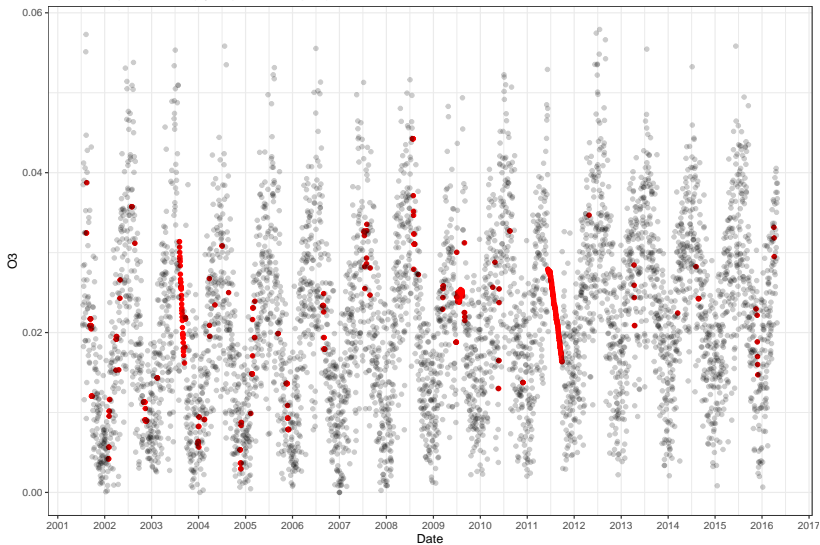


- ▶ Method create a bad behavior when the size of the skips is bigger than 7 days.

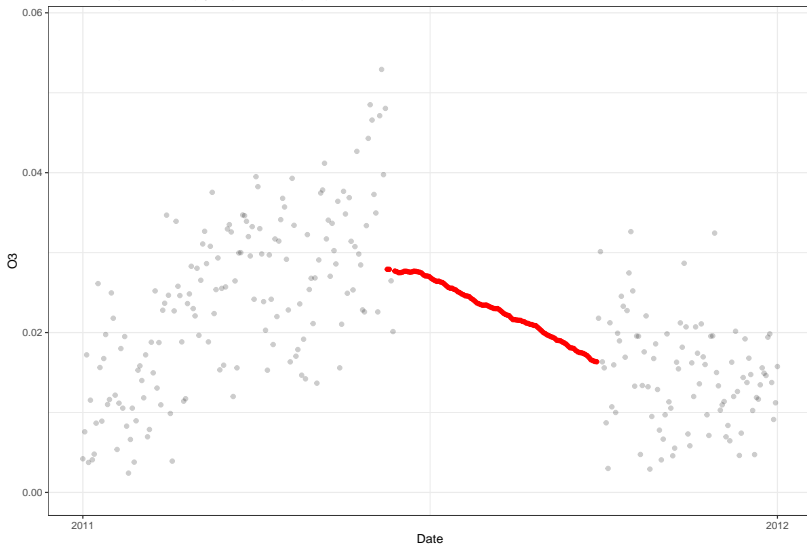


- ▶ To deal with this, the parameter k used for imputation will be different if the size of the skip is minor than 30 days, between 30 days and 100 days, or bigger than 100 days.
- ▶ $k = 7$, $k = 45$, $k = 120$, respectively.
- ▶ We will aggregate closest points by weighted by distance mean.

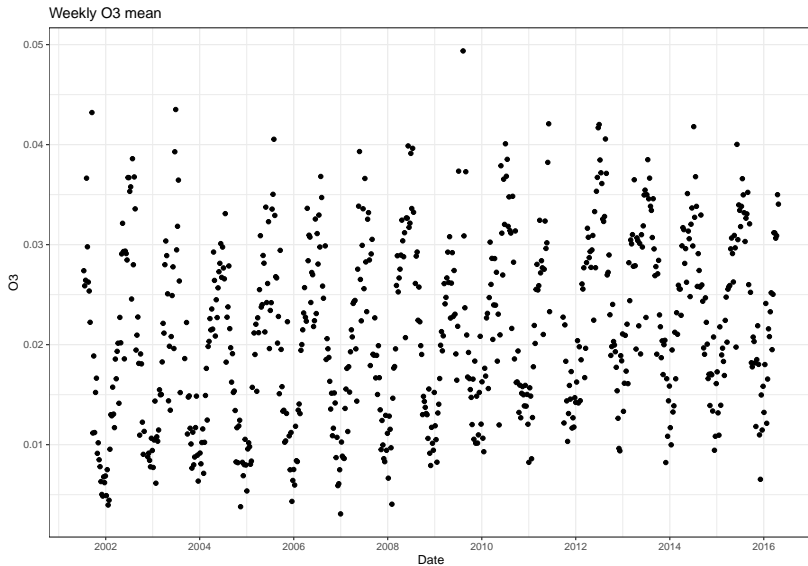
Real vs Imputed data (by separeted input.)



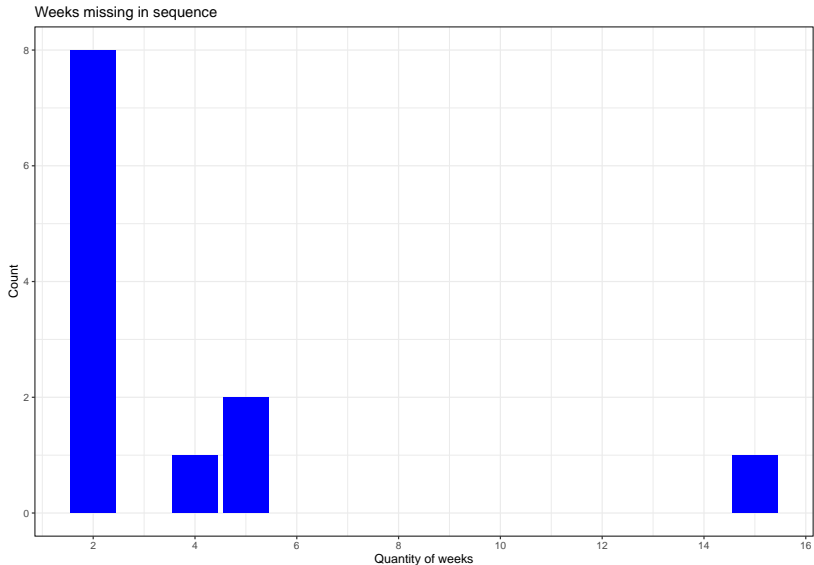
Real vs Imputed data (by separated input.)



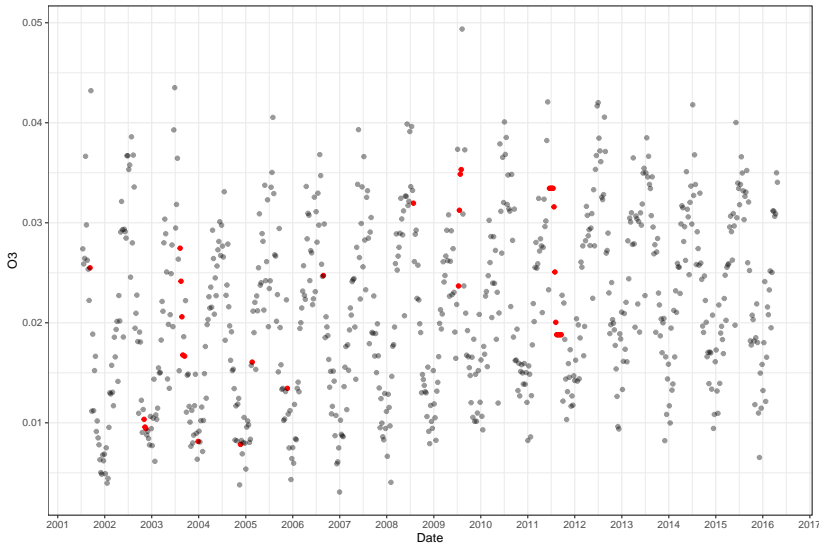
Weekly data



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



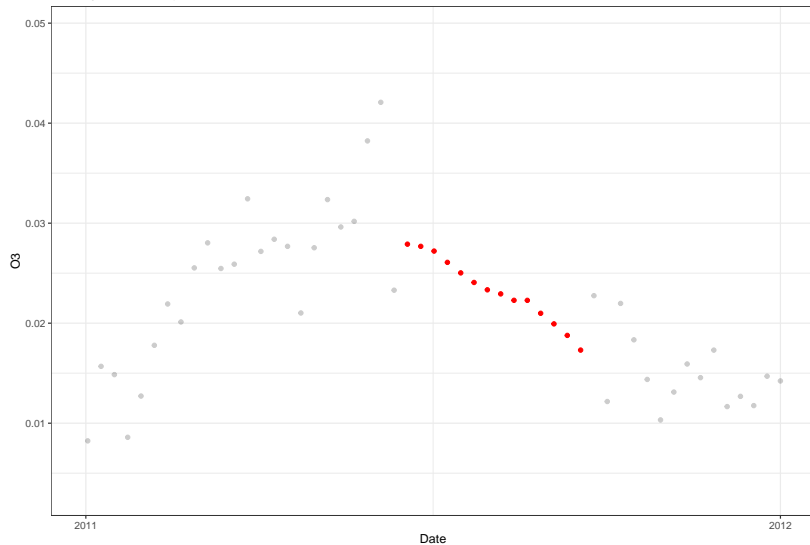
Weekly Real vs Imputed data



— —

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

Weekly Real vs Imputed data

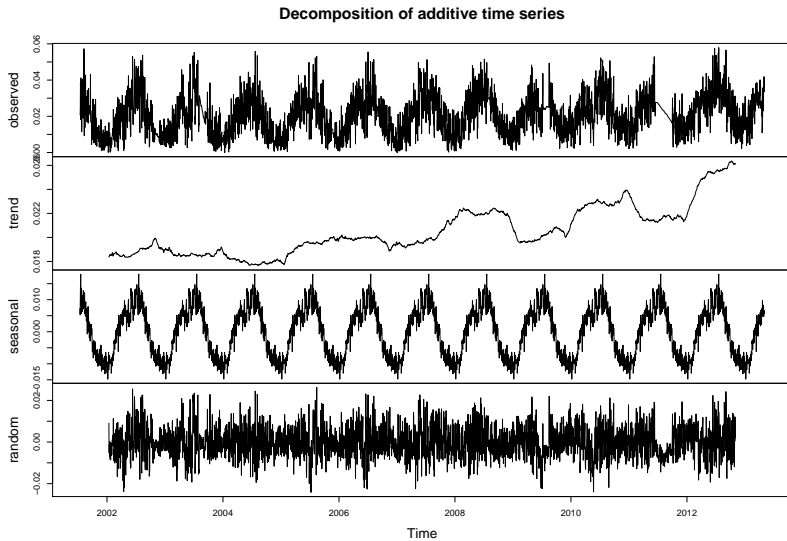


Daily model

Modelling process

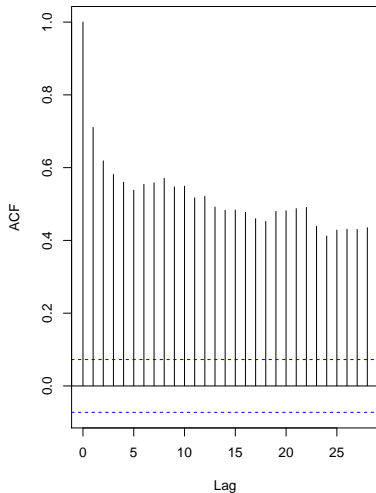
- ▶ Metric to be minimized: $MAE = \frac{1}{n} \sum_n |y_t - \hat{y}_t|$.
- ▶ Rolling window of 2 years (730 days).
- ▶ Prediction of the next 7 days.
- ▶ First: Test if there is tendency with Wald-Wolfowitz runs test.
 - ▶ For every 2 years window, the p-value is smaller than $1e - 3$.
- ▶ Second: Fitting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

Choice of models - trend

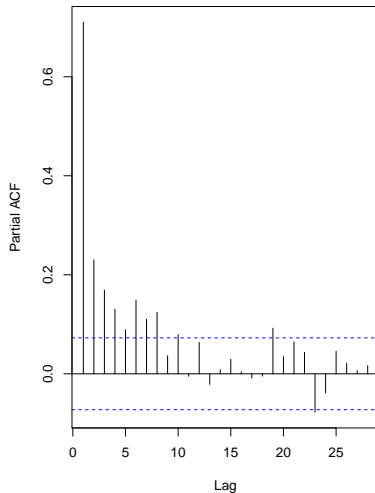


Choice of models - ACF and PACF

ACF on subset of train data



PACF on subset of train data



- ▶ 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.

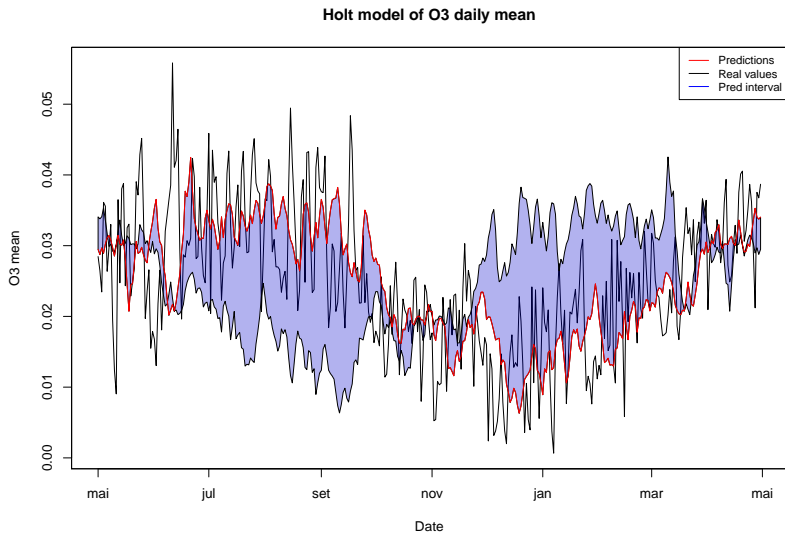
► Process:

- 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

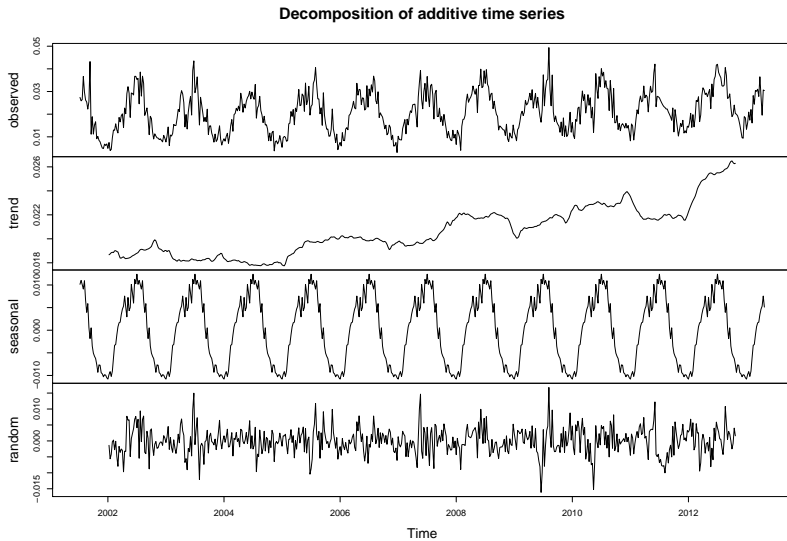


Weekly model

Modelling process

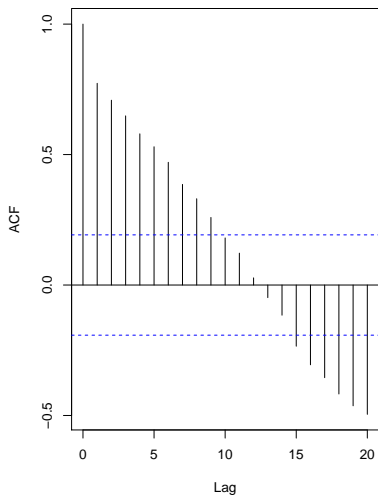
- ▶ Metric to be minimized: $MAE = \frac{1}{n} \sum_n |y_t - \hat{y}_t|$.
- ▶ Rolling window of 2 years (104 weeks), by skipping 4 weeks.
- ▶ Prediction of the next 4 weeks.
- ▶ First: Test if there is tendency with Wald-Wolfowitz runs test.
 - ▶ For every 2 years window, the p-value is smaller than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

Choice of models - trend

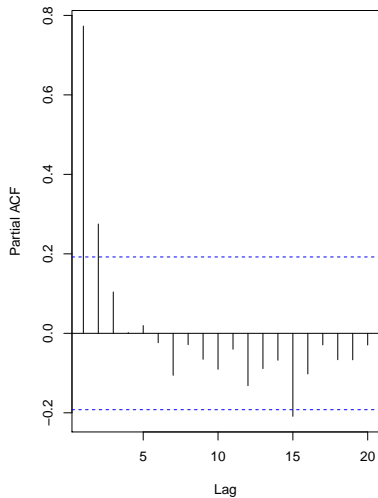


Choice of models - ACF and PACF

ACF on subset of train data



PACF on subset of train data



- ▶ Naive model: the next 4 weeks are predicted 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 additive).
- ▶ ARMA(1, 0) model.

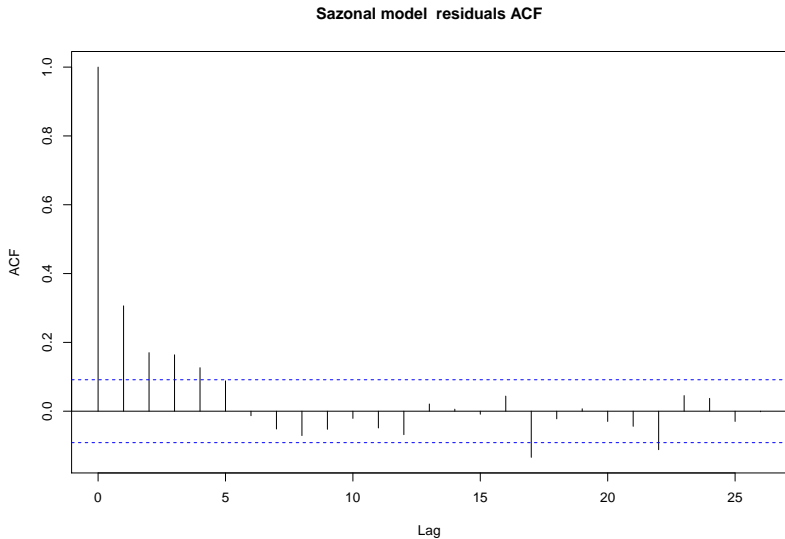
► Process:

- 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.

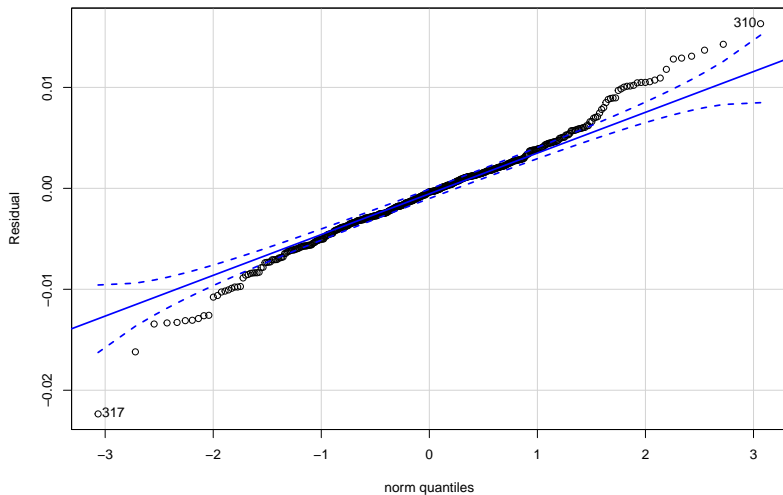
► Results:

- Sazonal: 0.003771788
- Linear: 0.003941300
- Poly 2: 0.004079873
- Poly 3: 0.004562281
- Arma(1, 0): 0.004675624
- Holt: 0.004905350
- HoltWinters additive: 0.005075887
- HoltWinters multiplicative: 0.005126882
- Naive: 0.005160706

Residuals

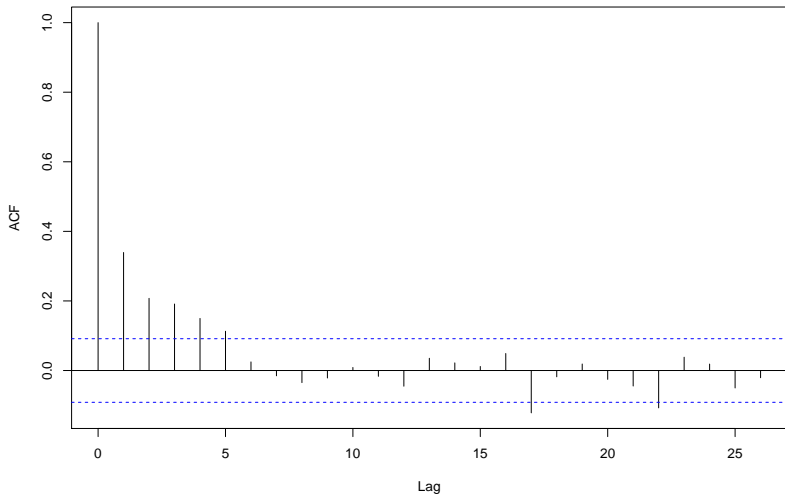


QQPlot of sazonal model residuals

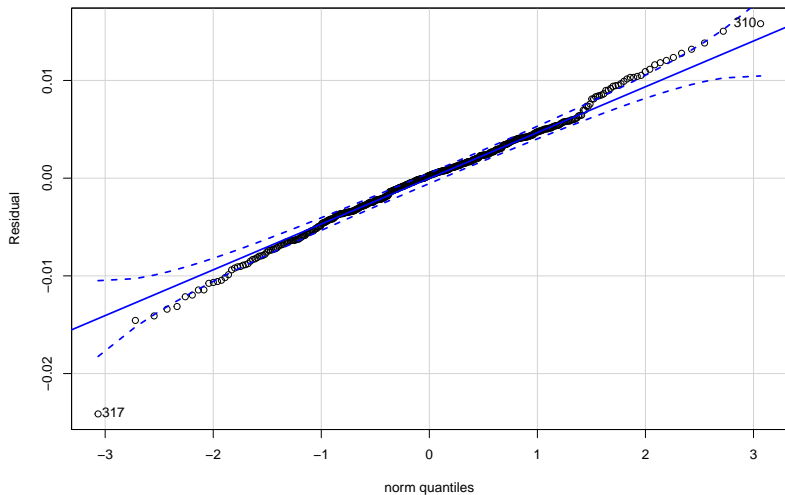


```
## [1] 317 310
```

Linear model residuals ACF



QQPlot of linear model residuals



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

Evaluating on test data

► MAE: 0.003476891

