Ozonio model

Giovani Valdrighi, Vitória Guardieiro

24/09/2020

R Markdown

This is an R Markdown presentation. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document.

Slide with Bullets

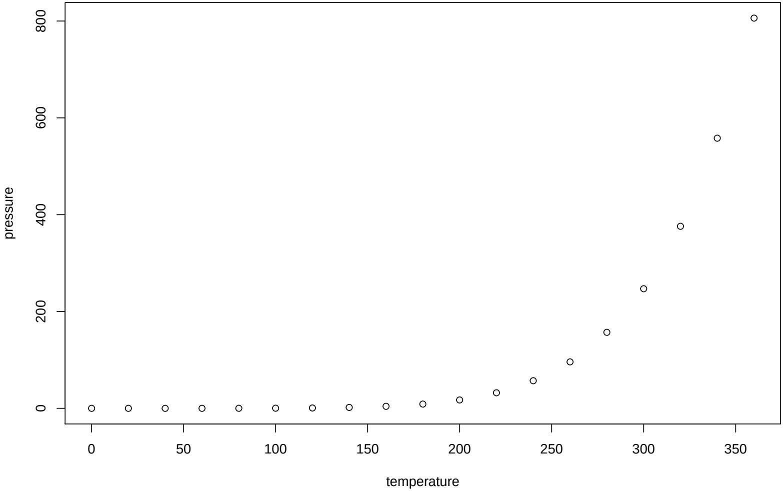
- Bullet 1
- Bullet 2
- Bullet 3

Slide with R Output

```r
summary(cars)
```

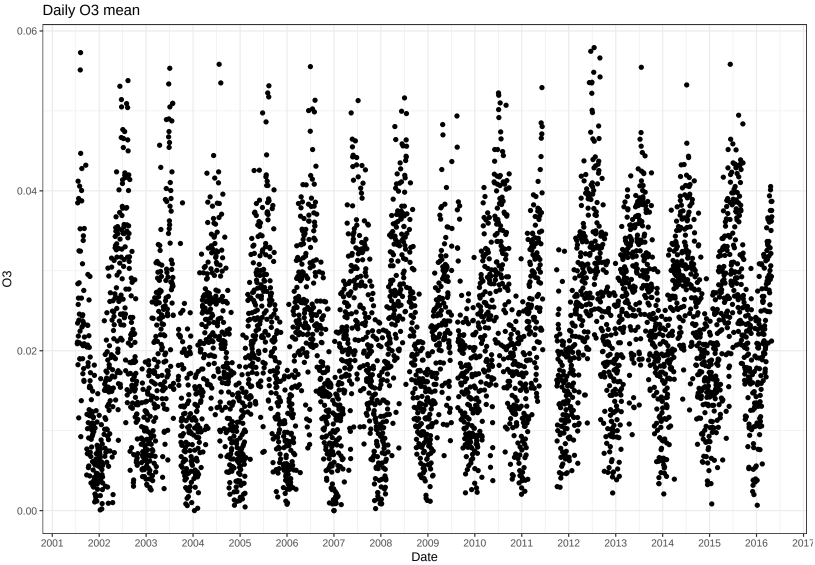
```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Slide with Plot



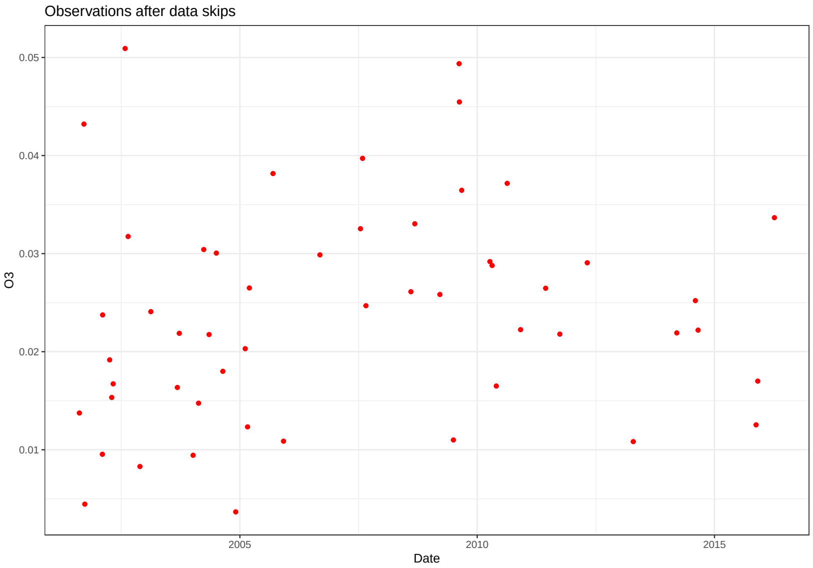
Data

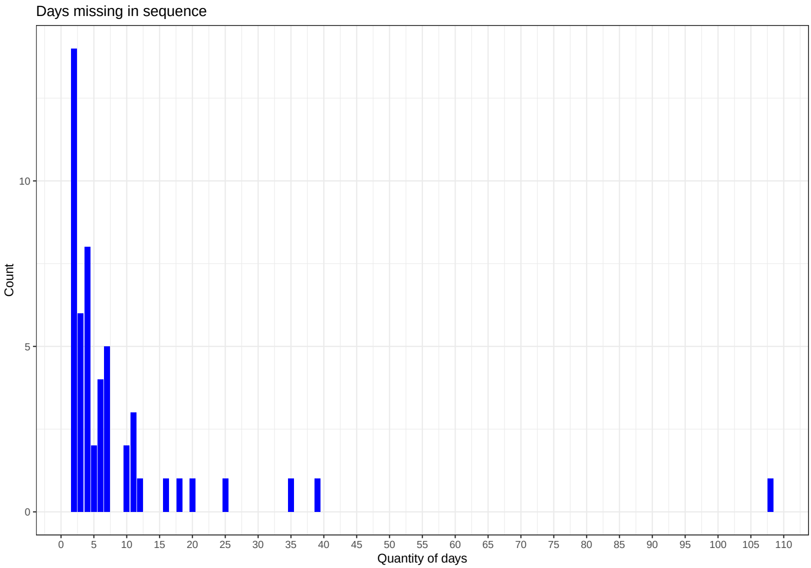
- New York data from 15/07/2001 to 30/04/2016.



Missing data

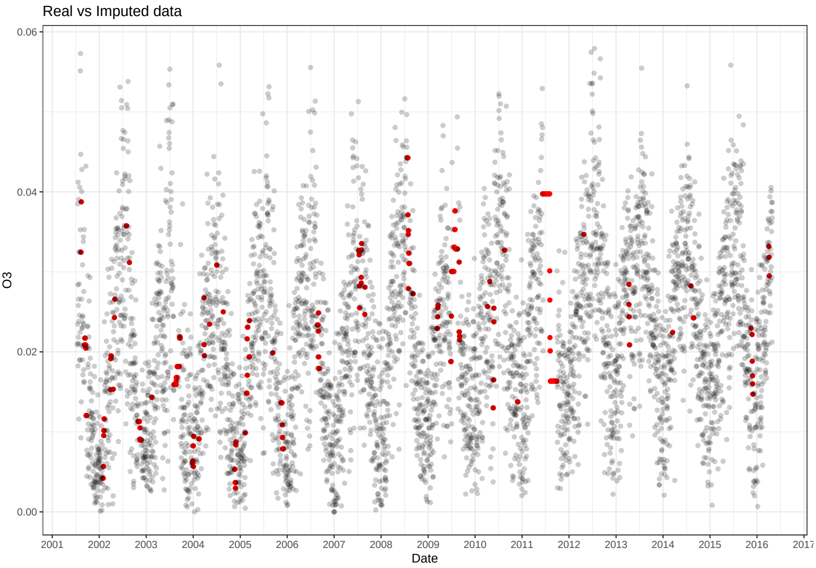
- There are 52 time skips in the data, in a total of 473 days.
- The biggest skips is 108 days in 2011.
- The majority of skips are of 1 or 2 days.
- Around 9.5% missing data.
- The missing observations are distributed along the time without a clear pattern.


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

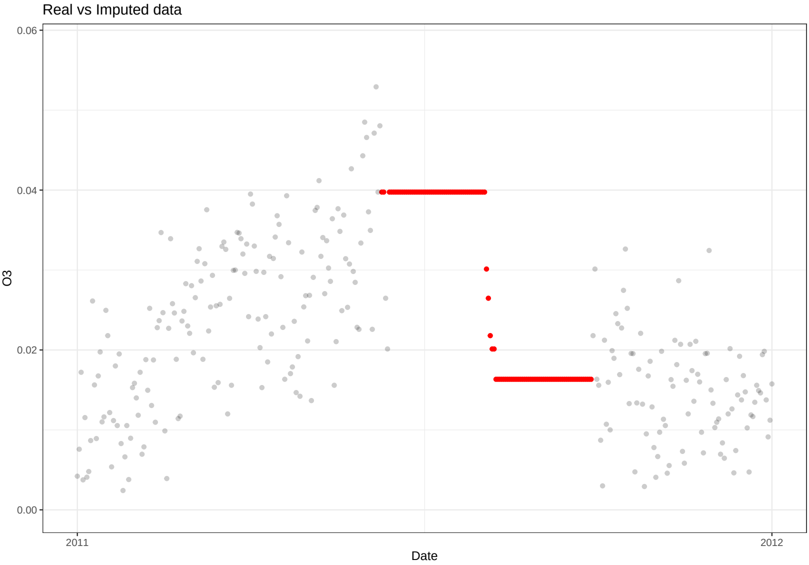

Days missing in sequence
Count
10
5
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
Quantity of days

Imputation method

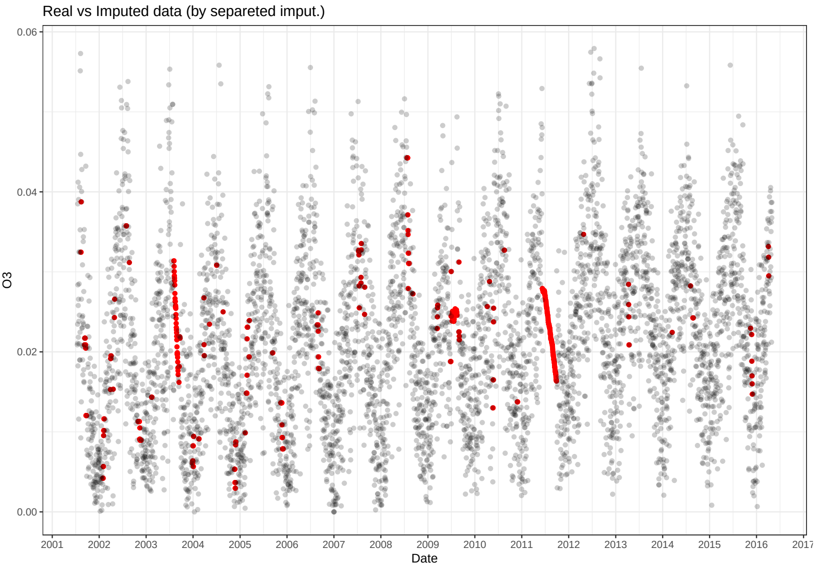
- It was used the kNN method to imputate values on missing observations.
- The kNN method needs the parameter k, the number of closest points considered.
- Starting with $k = 7$.

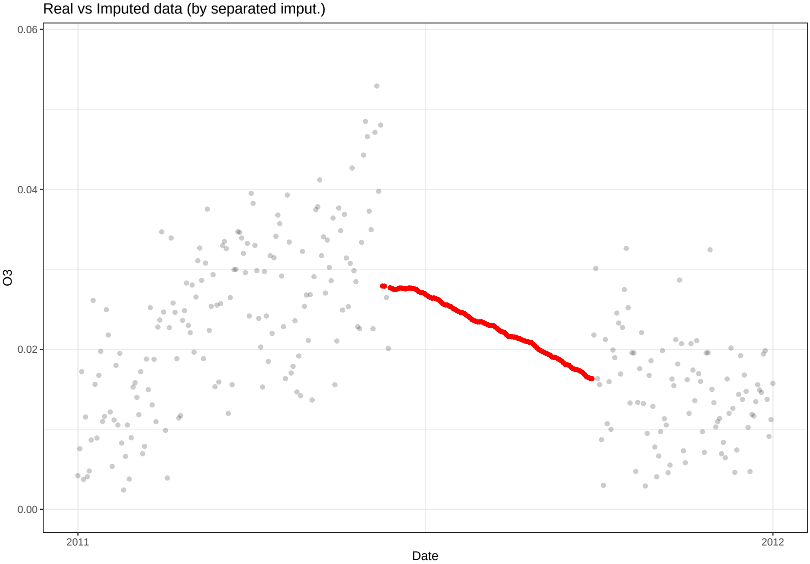

Real vs Imputed data
O3
0.06
0.04
0.02
0.00
Date
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

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

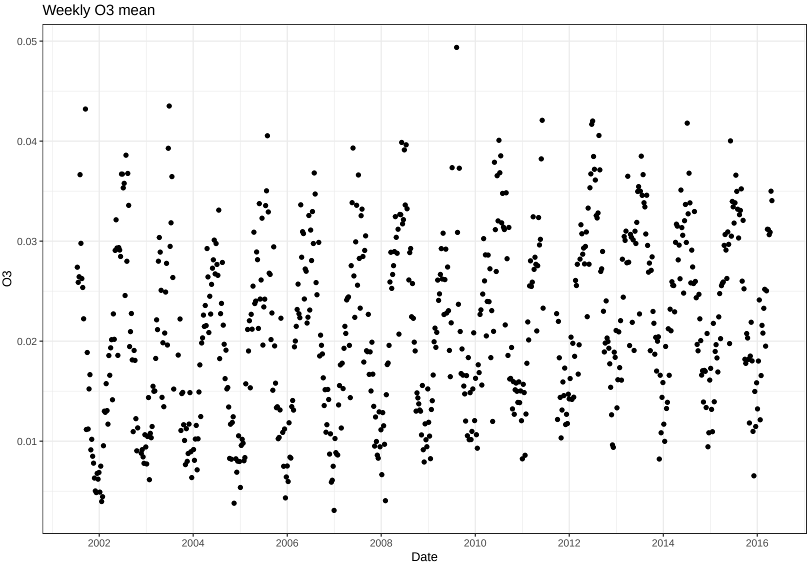


- To deal with this, the parameter k used for imputation will be different if the size of the skip is minor tem 30 days, between 30 days and 100 days, or bigger than 100 days.
- $k = 7$, $k = 45$, $k = 120$, respectively.
- We will aggregate closest points by weighted by distance mean.

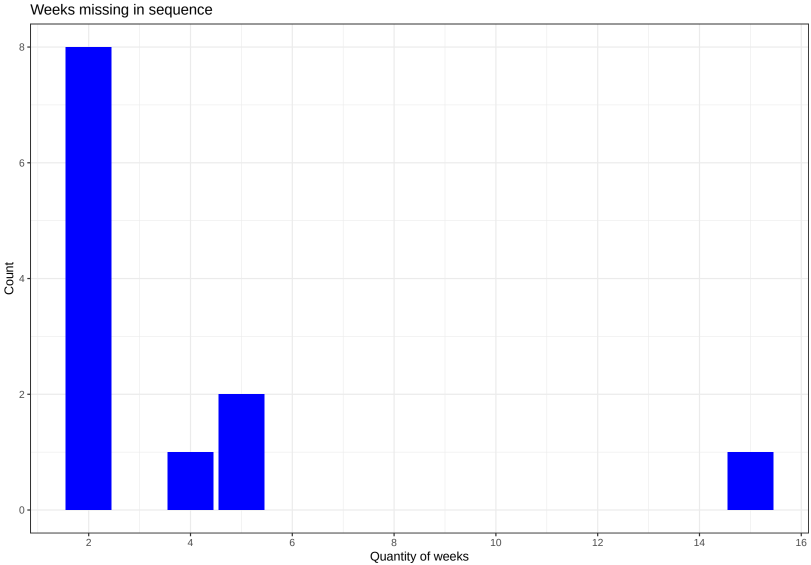

Real vs Imputed data (by separeted imput.)
O3
0.06
0.04
0.02
0.00
Date
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

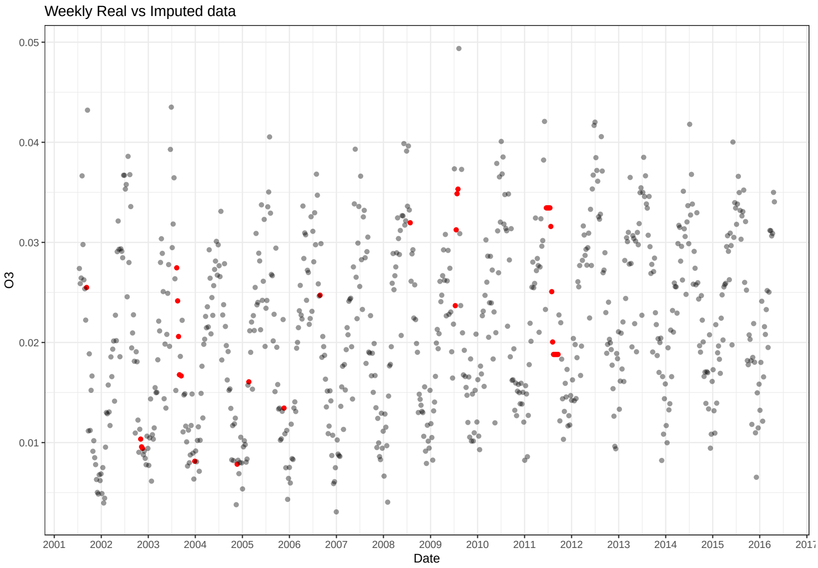

Real vs Imputed data (by separated imput.)
0.06
0.04
0.02
0.00
O3
2011
2012
Date

Weekly data


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

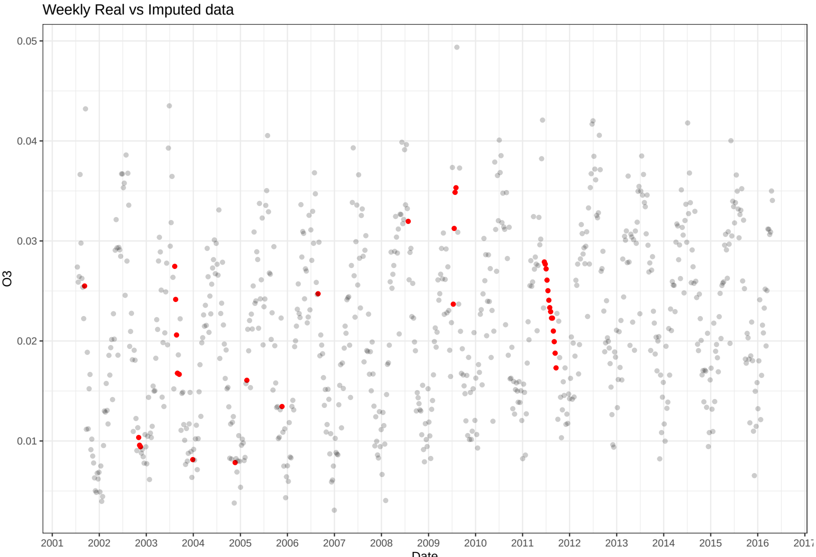
- If the data is grouped by week, ignoring the missing values when aggregating, it'll have 33 missing observations.
- Around 4.3% missing data.



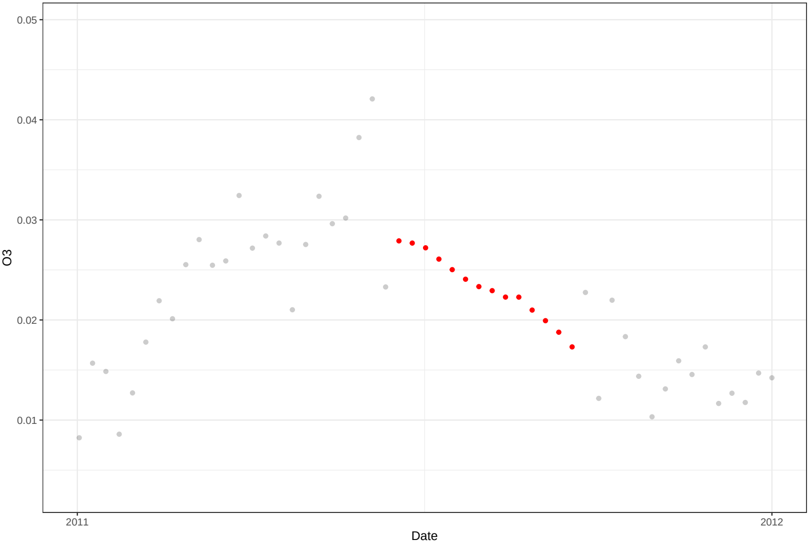


Weekly Real vs Imputed data
O3
0.05
0.04
0.03
0.02
0.01
Date
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

- 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


O3
Date