Experiences from dealing with missing values in sensor time series data

Steffen Moritz, Thomas Bartz-Beielstein

_Institut für Data Science, Engineering, and Analytics, TH Köln_

steffen.moritz@th-koeln.de

useR! 2019, Toulouse
Missing Data a well-known problem

Examples from our own projects:

• Sensor data is prone to missing data

• The reasons are manifold: Measurement, Transmission, Data Storage, Data Processing
Missing Data a well-known problem

Examples from our own projects:

- We have had all kind of unexpected sources for missing data
- Avoiding missing data is (usually) the best solution.

Water reservoir: cell reception problems
imputeTS: Time Series Missing Value Imputation

• imputeTS: Replacing NAs in Time Series

• Lately published version 3.0

• Univariate
  \[ X = \{ x_1, x_2, \ldots, x_n \} \]

• Equi-distant
  \[ |t_1 - t_2| = |t_2 - t_3| = \ldots = |t_{n-1} - t_n| \]

• Numeric
  \[ x_1, \ldots, x_n \in \mathbb{R} \]
Quite a common problem... in time series

Some users of imputeTS:
- Hydrology
- Quantitative Finance
- Meteorology
- Tropical Medicine
- ...

E.g.:
- gauge tide data
- sea-surface temperatures
- rainfall data
## Imputation: Employing Correlations

### Cross Sectional

<table>
<thead>
<tr>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
</tr>
</thead>
<tbody>
<tr>
<td>91</td>
<td>91</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>NA</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>55</td>
<td>55</td>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>32</td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>27</td>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>67</td>
<td>67</td>
<td>67</td>
<td>67</td>
</tr>
</tbody>
</table>

### Time Series Cross Sectional

<table>
<thead>
<tr>
<th>Time</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>13</td>
<td>33</td>
<td>15</td>
</tr>
<tr>
<td>t2</td>
<td>13</td>
<td>34</td>
<td>NA</td>
</tr>
<tr>
<td>t3</td>
<td>13</td>
<td>35</td>
<td>15</td>
</tr>
<tr>
<td>t4</td>
<td>13</td>
<td>36</td>
<td>16</td>
</tr>
<tr>
<td>t5</td>
<td>13</td>
<td>37</td>
<td>16</td>
</tr>
<tr>
<td>t6</td>
<td>14</td>
<td>38</td>
<td>16</td>
</tr>
<tr>
<td>t7</td>
<td>14</td>
<td>39</td>
<td>16</td>
</tr>
<tr>
<td>t8</td>
<td>14</td>
<td>40</td>
<td>17</td>
</tr>
<tr>
<td>t9</td>
<td>14</td>
<td>41</td>
<td>17</td>
</tr>
</tbody>
</table>

### Time Series

<table>
<thead>
<tr>
<th>Time</th>
<th>V1</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>12</td>
</tr>
<tr>
<td>t2</td>
<td>12</td>
</tr>
<tr>
<td>t3</td>
<td>NA</td>
</tr>
<tr>
<td>t4</td>
<td>13</td>
</tr>
<tr>
<td>t5</td>
<td>13</td>
</tr>
<tr>
<td>t6</td>
<td>13</td>
</tr>
<tr>
<td>t7</td>
<td>14</td>
</tr>
<tr>
<td>t8</td>
<td>14</td>
</tr>
<tr>
<td>t9</td>
<td>14</td>
</tr>
</tbody>
</table>

- **Cross Sectional**
  - inter-variable

- **TS Cross Sectional**
  - inter-variable + inter-time

- **Time Series**
  - inter-time
Also TSCS data needs univariate imputation sometimes

<table>
<thead>
<tr>
<th>Time</th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1</td>
<td>13</td>
<td>33</td>
<td>15</td>
</tr>
<tr>
<td>t2</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>t3</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>t4</td>
<td>13</td>
<td>36</td>
<td>16</td>
</tr>
<tr>
<td>t5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>t6</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>t7</td>
<td>14</td>
<td>39</td>
<td>16</td>
</tr>
<tr>
<td>t8</td>
<td>14</td>
<td>40</td>
<td>17</td>
</tr>
<tr>
<td>t9</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

TS Cross Sectional

Problem:
Only whole observations are missing (V1, V2, V3 at one point in time)

This is often common for transmission problems

Thus inter-variable correlation can not be sufficiently employed

--> Pure time series imputation needed
CRAN imputation packages by type

(univariate) Time Series

imputeTS
zoo
forecast
imputePSF
...

TS Cross Sectional

Amelia
mtsdi
...

Cross Sectional

mice
mi
Amelia
VIM
missMDA
missForest
imputeR
simputation
...

Task View Missing Data
https://cran.r-project.org/web/views/MissingData.html

R-miss-tastic
https://rmisstastic.netlify.com/
How to deal with Missing Data in Time Series

• 1. Visualization and statistics of missing data

• 2. Select Approach

  Delete missing data
  Keep missing data
  Replace missing data called imputation, gap filling

• 3. Select Algorithm
Short intro into imputeTS
Our goals:

• Inspired from own sensor data use cases
  Rather big time series. Combination of fast and advanced algorithms.

• Domain experts as users
  Easy and quick access to advanced functions.

• Whole imputation process in one package
  Visualization + Imputation + Result Analysis
Package Scope

• Analysis before NA action
  - 3 Missing Data Plots
  - NA statistics text output

• Analysis after imputation
  - 1 Result Plot

• Imputation functions
  - 5 fast imputation functions
  - 4 more advanced functions
  - NA remove function

• 3 Datasets for testing
Easy to use
# List of algorithms

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>na_locf</td>
<td>Missing Value Imputation by Last Observation Carried Forward</td>
</tr>
<tr>
<td>na_random</td>
<td>Missing Value Imputation by Random Sample</td>
</tr>
<tr>
<td>na_mean</td>
<td>Missing Value Imputation by Mean Value</td>
</tr>
<tr>
<td>na_interpolation</td>
<td>Missing Value Imputation by Interpolation</td>
</tr>
<tr>
<td>na_ma</td>
<td>Missing Value Imputation by Weighted Moving Average</td>
</tr>
<tr>
<td>na_remove</td>
<td>Remove Missing Values</td>
</tr>
<tr>
<td>na_replace</td>
<td>Replace Missing Values by a Defined Value</td>
</tr>
<tr>
<td>na_kalman</td>
<td>Missing Value Imputation by Kalman Smoothing</td>
</tr>
<tr>
<td>na_seadec</td>
<td>Seasonally Decomposed Missing Value Imputation</td>
</tr>
<tr>
<td>na_seasplit</td>
<td>Seasonally Splitted Missing Value Imputation</td>
</tr>
</tbody>
</table>
Easy to use

• `na_`algorithmname``(yourInput, add. param)`
  • Similar syntax also used by other packages like zoo, forecast

• Imputation functions take all kinds of inputs:
  • ts, mts, data.frame, zoo, xts, vector, tibble, tsibble
Example: Pipe and Normal Use

data  %>% na_seadec() %>% further steps

or

imp <- na_seadec(data)
Some other advantage: Speed
Fast: Last observation carried forward

**LOCF**
dendextend
spacetime
zoo
imputeTS

**tsHeating**
Length: 606.837
NAs: 57.391
Algorithms
### Imputation Algorithms to choose from

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>na_locf</td>
<td>Missing Value Imputation by Last Observation Carried Forward</td>
</tr>
<tr>
<td>na_random</td>
<td>Missing Value Imputation by Random Sample</td>
</tr>
<tr>
<td>na_mean</td>
<td>Missing Value Imputation by Mean Value</td>
</tr>
<tr>
<td>na_interpolation</td>
<td>Missing Value Imputation by Interpolation</td>
</tr>
<tr>
<td>na_ma</td>
<td>Missing Value Imputation by Weighted Moving Average</td>
</tr>
<tr>
<td>na_remove</td>
<td>Remove Missing Values</td>
</tr>
<tr>
<td>na_replace</td>
<td>Replace Missing Values by a Defined Value</td>
</tr>
<tr>
<td>na_kalman</td>
<td>Missing Value Imputation by Kalman Smoothing</td>
</tr>
<tr>
<td>na_seadec</td>
<td>Seasonally Decomposed Missing Value Imputation</td>
</tr>
<tr>
<td>na_seasplit</td>
<td>Seasonally Splitted Missing Value Imputation</td>
</tr>
</tbody>
</table>
Algorithm options for Moving Average (na_ma)

- Most of the functions like na_interpolation or na_mean have additional options

- For na_ma e.g. the user can choose between the parameter ‘weighting’

SMA:\[ x_a = \frac{1}{2k} \sum_{i=-k}^{k} x_{a+i} \]

LWMA:\[ x_a = \frac{\sum_{i=-k}^{k} \frac{1}{|i|+1} x_{a+i}}{\sum_{i=k}^{k} \frac{1}{i+1}} \]

EWMA:\[ x_a = \frac{\sum_{i=-k}^{k} \frac{1}{2^{|i|+1}} x_{a+i}}{\sum_{i=k}^{k} \frac{1}{2^i+1}} \]

- \( x_a \) is the position in time series to impute
- \( n \) is the number of observations
- \( k \) width of moving average window in each direction
Imputation Process

Step 1: Visualization
Visualization of NA distribution

plotNA.distribution(yourInput)

Daily morning gold prices from forecast package

Visualization of how the NAs are distributed in the series
Sometimes time series are just too long

plotNA.distribution(tsNH4)

Just too long
Visualization of long time series

plotNA.distributionBar(tsNH4, breaks=20)
Additional Stats

statsNA(tsHeating)

"Length of time series:"
606837
"Number of Missing Values:
57391
"Percentage of Missing Values:
9.46%"

"Stats for Bins"
- Bin 1 (151710 values from 1 to 151710) : 0 NAs (0.003%)
- Bin 2 (151710 values from 151711 to 303420) : 29755 NAs (19.62%)
- Bin 3 (151710 values from 303421 to 455130) : 6153 NAs (4.06%)
- Bin 4 (151707 values from 455131 to 606837) : 21483 NAs (14.2%)

"Longest NA gap (series of consecutive NAs)"
258 in a row

"Most frequent gap size (series of consecutive NA series)"
2 NA in a row (occurring 104 times)

"Gap size accounting for most NAs"
Imputation Process

Step 2: Imputation
Visualization of NA distribution

```
plotNA.distribution(x)
```

AirPassengers from datasets package with manually introduced NAs
Imputation with na_mean

na_mean(x)

plotNA.imputations(x, na_mean(x))

AirPassengers from datasets package with manually introduced NAs
Imputation with na_seasplit

na_seasplit(x)
Imputation Process

Whole Example Workflow
Workflows e.g. with forecast

library("imputeTS")
library("forecast")
gold %>%
na_interpolation() %>% ets() %>%
forecast(h=36) %>%
autoplot()
Outlook & Discussion

• Future plans:
  • Transition plots to ggplot2
  • Additional algorithms (RNN, Pattern based, ...)

• Maybe added in the future
  • Multiple Imputation / accounting for uncertainty
  • Automatic model selection & evaluation / overimputation
Get in contact & download imputeTS

https://github.com/SteffenMoritz/imputeTS

Contributions are welcome.