How to deal with Missing Data in Time Series and the imputeTS package

useR! 2017, Brussels
Steffen Moritz, TH Köln
steffen.moritz10@gmail.com
Talk Overview

• 1. Introduction
• 2. Imputation landscape on CRAN
• 3. Time Series Imputation specifics
• 4. imputeTS Introduction
We are often facing missing data

Examples from our projects:

- Water quality measuring station: sensor problems
- Water reservoir: cell reception problems

- Especially sensor measurements are prone to missing data
- Avoiding missing data should be the prioritized over filling NAs
There are other people with the same problems…

Field of expertise of people asking about imputeTS package:

- Hydrology
- Oceanography
- Quantitive Finance
- Meteorology

This included:

- gauge tide data
- sea-surface temperatures
- rainfall data
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
    - Imputation / gap filling

• 3. Select Algorithm
Simple Map of CRAN imputation packages

(univariate) Time Series
- `imputeTS`
- `zoo`
- `forecast`
- `imputePSF`
- ...

TS Cross Sectional
- `Amelia`
- `mtsdi`
- ...

Cross Sectional
- `mice`
- `mi`
- `Amelia`
- `VIM`
- `missMDA`
- `Hmisc`
- `missForest`
- `imputeR`
- `simputation`
- ...

Input:
- `ts`, `vector`, `zoo`, `xts`, `timeSeries`
- `data.frame`, `mts`, `zoo`, `xts`, `matrix`, `timeSeries`
- `data.frame`, `matrix`
### 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>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>

#### TS 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>12</td>
</tr>
<tr>
<td>t4</td>
<td>NA</td>
</tr>
<tr>
<td>t5</td>
<td>13</td>
</tr>
<tr>
<td>t6</td>
<td>13</td>
</tr>
<tr>
<td>t7</td>
<td>13</td>
</tr>
<tr>
<td>t8</td>
<td>13</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
Time Series Imputation Specifics

• Considering time series characteristics like trend and seasonality is essential

• Although called univariate, time is an additional variable, which is implicitly given

• For MCAR / MAR / MNAR determination time has to be considered as a variable
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>

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

**TS Cross Sectional**

--> Pure time series imputation needed
imputeTS CRAN package

imputeTS: Time Series Missing Value Imputation

The imputeTS package specializes on (univariate) time series imputation. It offers several different imputation algorithm implementations. Beyond the imputation algorithms the package also provides plotting and printing functions of time series missing data statistics. Additionally three time series datasets for imputation experiments are included.

Installation

The imputeTS package can be found on CRAN. For installation execute in R:

```r
install.packages("imputeTS")
```
The idea behind the package

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

• Domain experts as users
  Easy and quick access to advanced functions. No multiple imputation.

• 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
Visualization of NA distribution

plotNA.distribution(yourInput)

Daily morning gold prices from forecast package

Visualization of how the NAs are distributed in the series
Visualization of NA distribution

plotNA.distribution(x)

AirPassengers from datasets package with manually introduced NAs
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%)"
" Bin 2 (151710 values from 151711 to 303420) : 29755 NAs (19.6%)"
" 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"
"NA in a row (leading 232, trailing 0)"
## Imputation Options

<table>
<thead>
<tr>
<th>Function</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>na.interpolation</td>
<td>Missing Value Imputation by Interpolation</td>
</tr>
<tr>
<td>na.kalman</td>
<td>Missing Value Imputation by Kalman Smoothing</td>
</tr>
<tr>
<td>na.locf</td>
<td>Missing Value Imputation by Last Observation Carried Forward</td>
</tr>
<tr>
<td>na.ma</td>
<td>Missing Value Imputation by Weighted Moving Average</td>
</tr>
<tr>
<td>na.mean</td>
<td>Missing Value Imputation by Mean Value</td>
</tr>
<tr>
<td>na.random</td>
<td>Missing Value Imputation by Random Sample</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.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)`
  • Same syntax also used by other packages like zoo, forecast

• Imputation functions take all kinds of inputs:
  • ts, mts, data.frame, matrix, zoo, xts, vector
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)
Fast: Last observation carried forward

**LOCF**
dendextend
spacetime
zoo
imputeTS

tsNH4
Length: 4552
NAs: 883
Fast: Linear Interpolation

**Linear interp.**
imputeTS
zoo

**tsHeating**
Length: 606837
NAs: 57391
Choosing the right algorithm

- Do I need a very fast algorithm?
  - na.seasplit
  - na.locf
  - na.interpolation

- Do I have very short series (< 7 observations)?
  - na.remove
  - na.mean
  - na.random
  - na.replace

- Usually not a good idea!
Choosing the right algorithm

• Does my data have seasonality?
  - na.seasplit
  - na.seadec
  - na.kalman

• Does my data have seasonality + trend?

• A usually very good - but slow choice

Trying and assessing different algorithms is always a good idea.
Get in contact & download the package

steffen.moritz10@gmail.com