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

Technology
Arts Sciences
TH Köln

useR! 2019, Toulouse

12.07.17

Missing Data a well-known problem

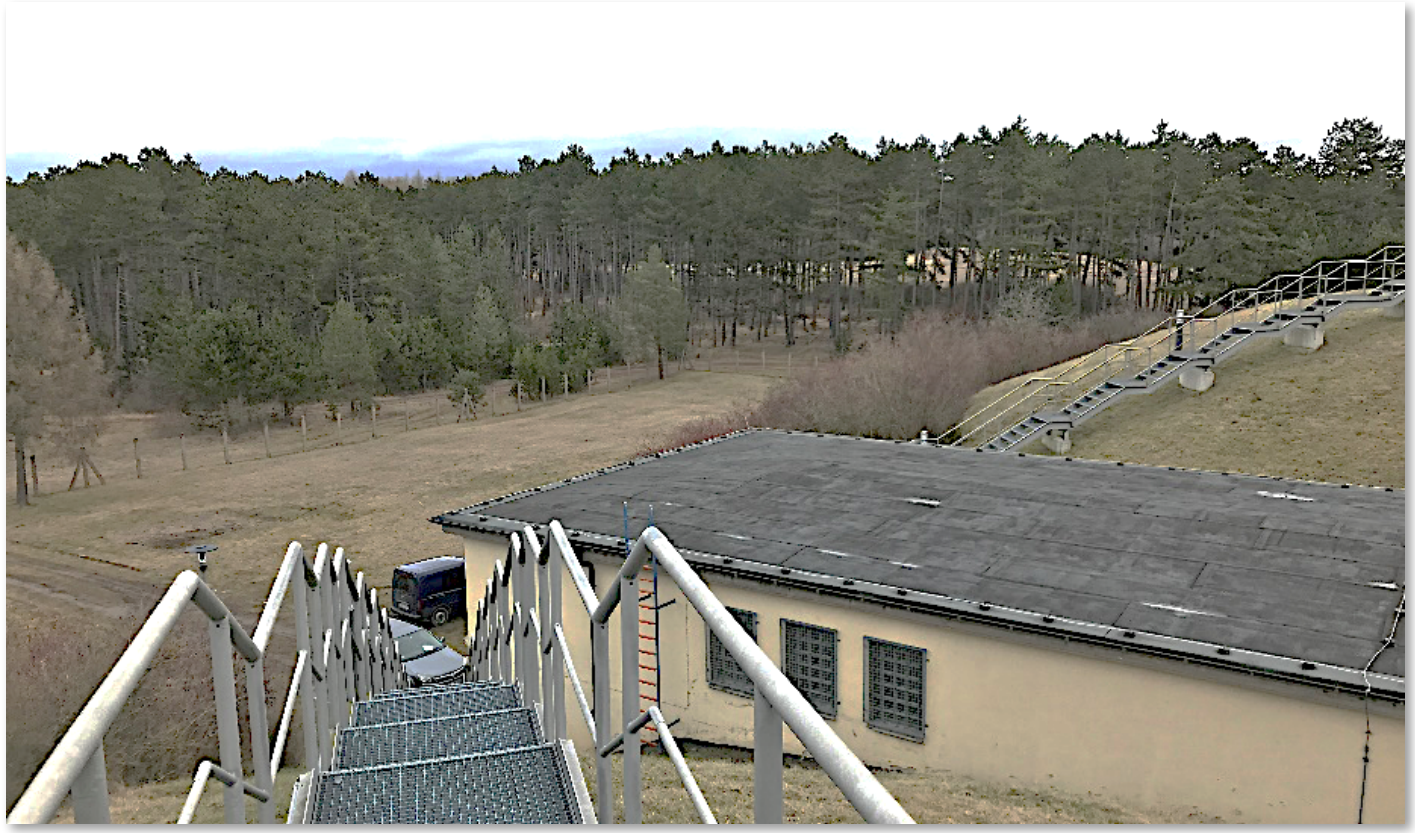
Examples from our own projects:

Water quality measurement panel

- 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:



Water reservoir: cell reception problems

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

imputeTS: Time Series Missing Value Imputation

- imputeTS: Replacing NAs in Time Series

- Lately published version 3.0

- Univariate

$$X = \{x_1, x_2, ..., x_n\}$$

- Equi-distant

$$|t_1 - t_2| = |t_2 - t_3| = ... = |t_{n-1} - t_n|$$

- Numeric

$$x_1, \ldots x_n \in \mathbb{R}$$

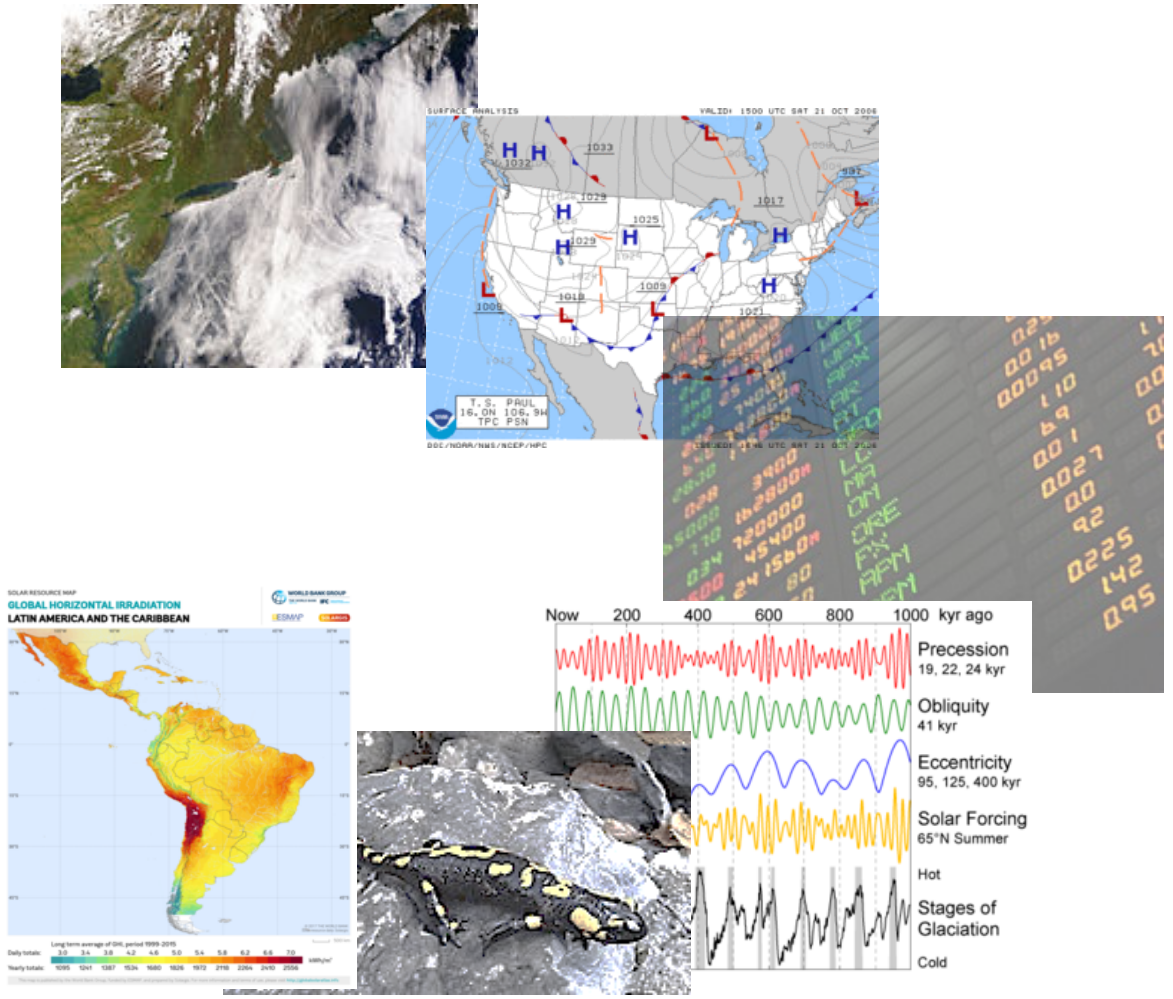
Quite a common problem…in time series

Some users of imputeTS:

- Hydrology
- Quantitive Finance
- Meteorology
- Tropical Medicine
- …

E.g.:

- gauge tide data
- sea-surface temperatures
- rainfall data



Imputation: Employing Correlations

V1	V2	V3	V4
91	91	91	91
NA	13	13	13
14	14	14	14
55	55	55	55
19	19	19	19
32	32	32	32
23	23	23	23
27	27	27	27
67	67	67	67

Cross Sectional

inter-variable

Time	V1	V2	V3
t1	13	33	15
t2	13	34	NA
t3	13	35	15
t4	13	36	16
t5	13	37	16
t6	14	38	16
t7	14	39	16
t8	14	40	17
t9	14	41	17

TS Cross Sectional

inter-variable + inter-time

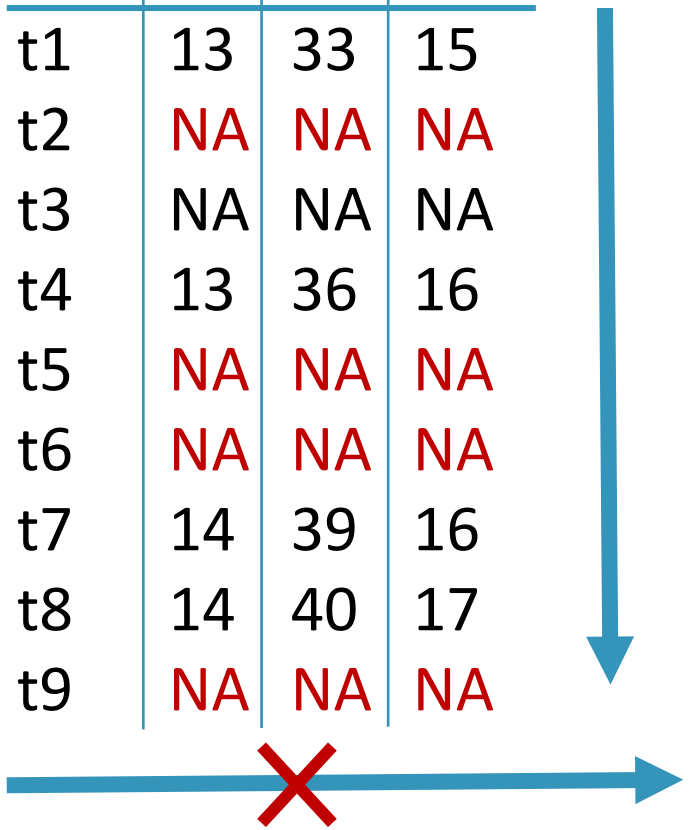
Time	V1
t1	12
t2	12
t3	NA
t4	13
t5	13
t6	13
t7	14
t8	14
t9	14

Time Series

inter-time

Also TSCS data needs univariate imputation sometimes

Time	V1	V2	V3
t1	13	33	15
t2	NA	NA	NA
t3	NA	NA	NA
t4	13	36	16
t5	NA	NA	NA
t6	NA	NA	NA
t7	14	39	16
t8	14	40	17
t9	NA	NA	NA



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

Function	Description
na_locf	Missing Value Imputation by Last Observation Carried Forward
na_random	Missing Value Imputation by Random Sample
na_mean	Missing Value Imputation by Mean Value
na_interpolation	Missing Value Imputation by Interpolation
na_ma	Missing Value Imputation by Weighted Moving Average
na_remove	Remove Missing Values
na_replace	Replace Missing Values by a Defined Value
na_kalman	Missing Value Imputation by Kalman Smoothing
na_seadec	Seasonally Decomposed Missing Value Imputation
na_seasplit	Seasonally Splitted Missing Value Imputation

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

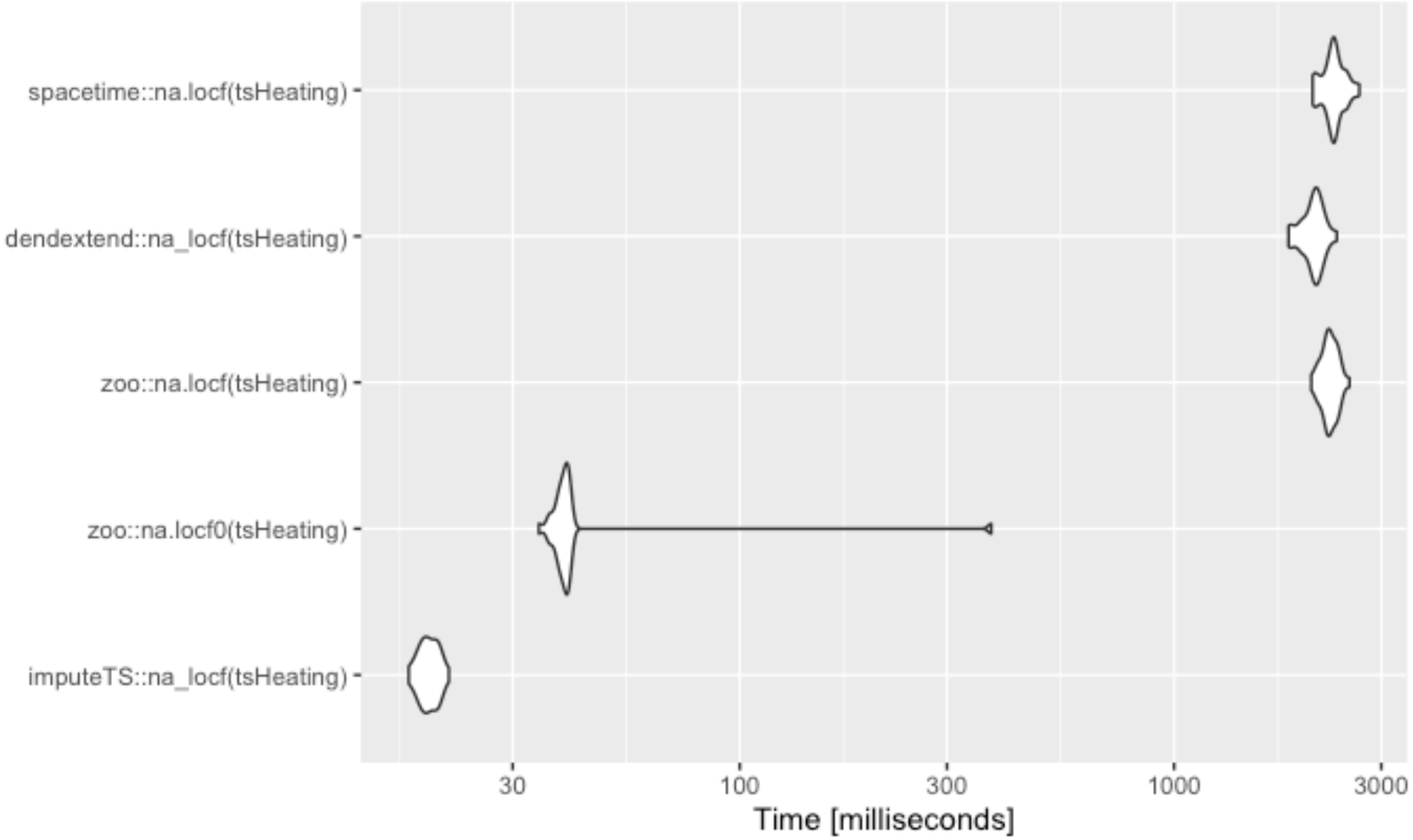
```
imp <– na_seadec(data)
```

Some other advantage: Speed

Fast: Last observation carried forward

LOCF
dendextend
spacetime
zoo
imputeTS

tsHeating
Length: 606.837
NAs: 57.391



Algortihms

Imputation Algorithms to choose from

Function	Description
na_locf	Missing Value Imputation by Last Observation Carried Forward
na_random	Missing Value Imputation by Random Sample
na_mean	Missing Value Imputation by Mean Value
na_interpolation	Missing Value Imputation by Interpolation
na_ma	Missing Value Imputation by Weighted Moving Average
na_remove	Remove Missing Values
na_replace	Replace Missing Values by a Defined Value
na_kalman	Missing Value Imputation by Kalman Smoothing
na_seadec	Seasonally Decomposed Missing Value Imputation
na_seasplit	Seasonally Splitted Missing Value Imputation

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*

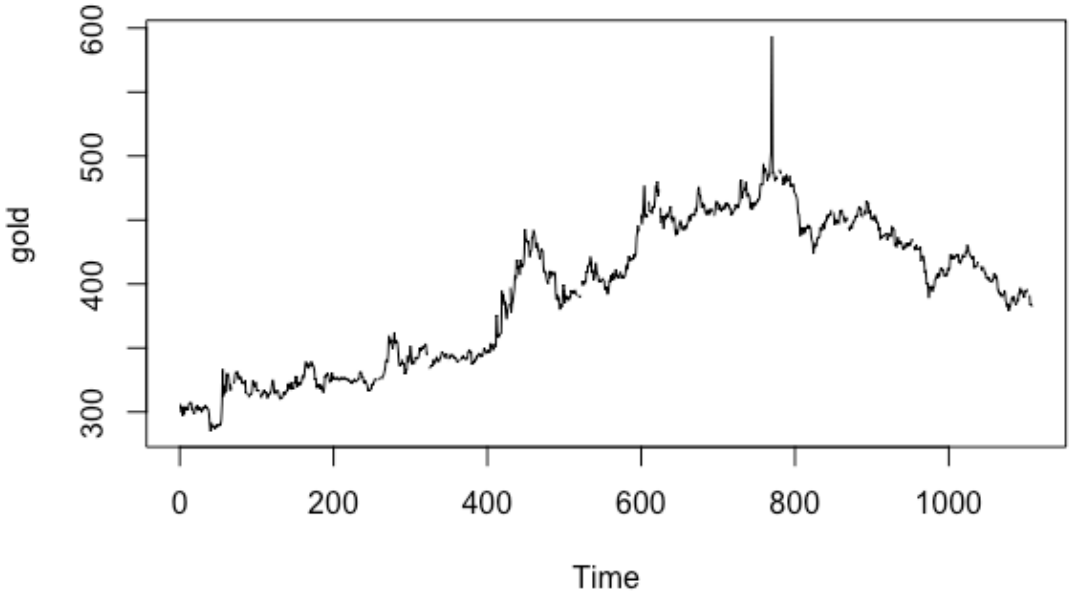
k *width of moving average window in each direction*

Imputation Process

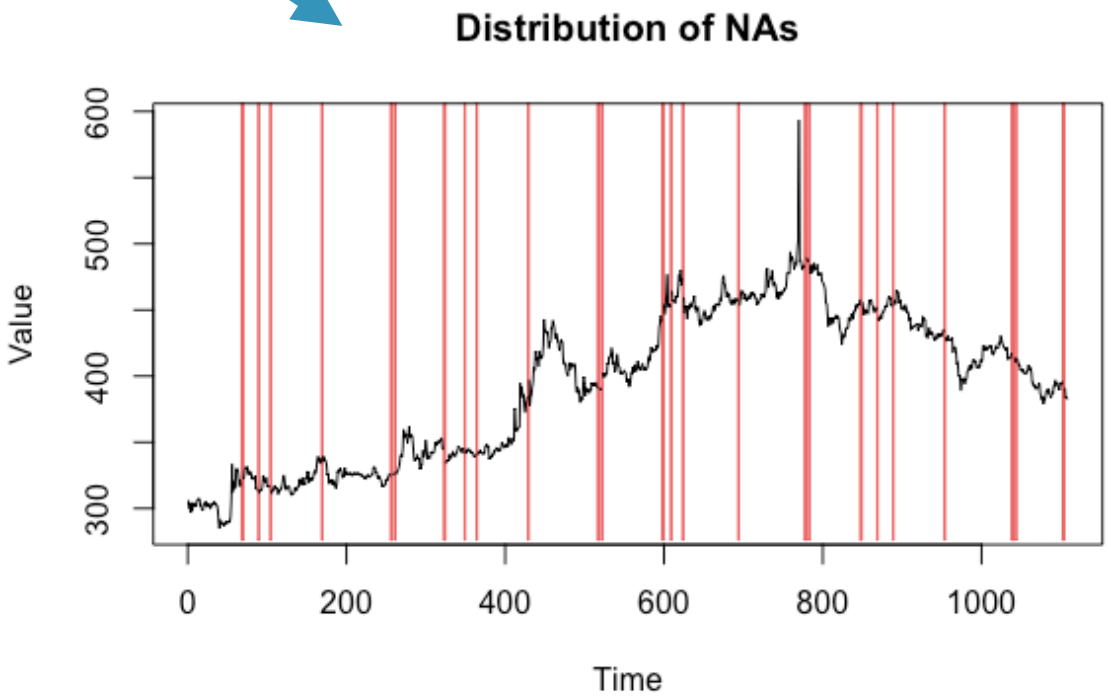
Step 1: Visualization

Visualization of NA distribution

plotNA.distribution(yourInput)



Daily morning gold prices from forecast package

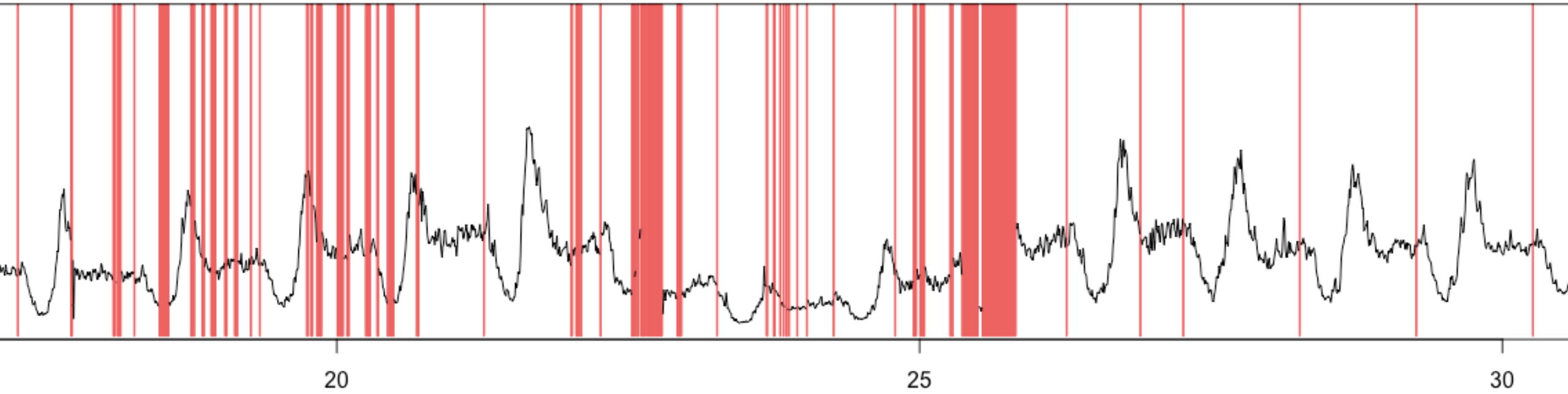


Visualization of how the NAs are distributed in the series

Sometimes time series are just too long

plotNA.distribution(tsNH4)

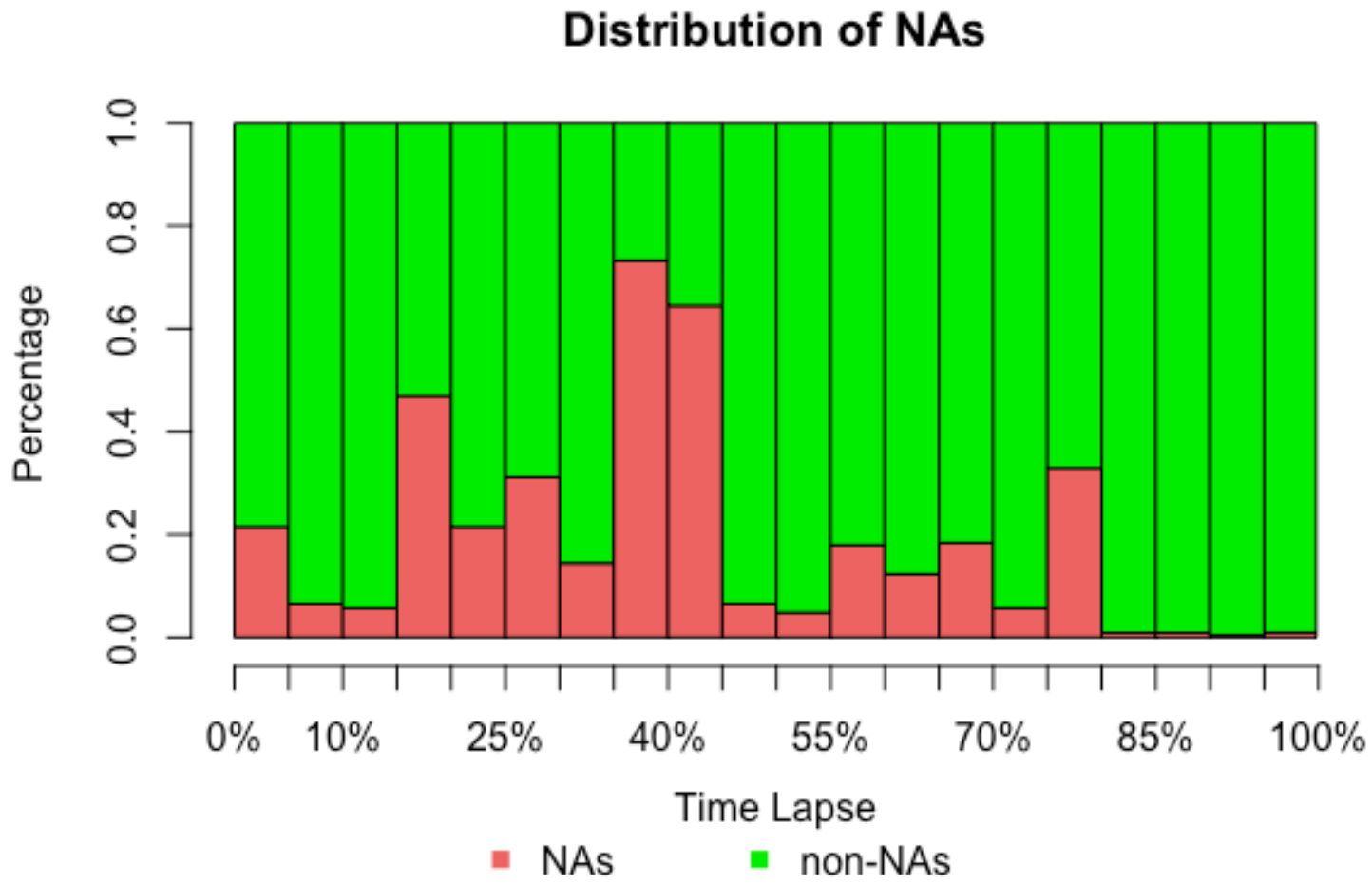
Just too long



tsNH4 from imputeTS package

Visualization of long time series

plotNA.distributionBar(tsNH4, breaks=20)



Additional Stats

statsNA(tsHeating)

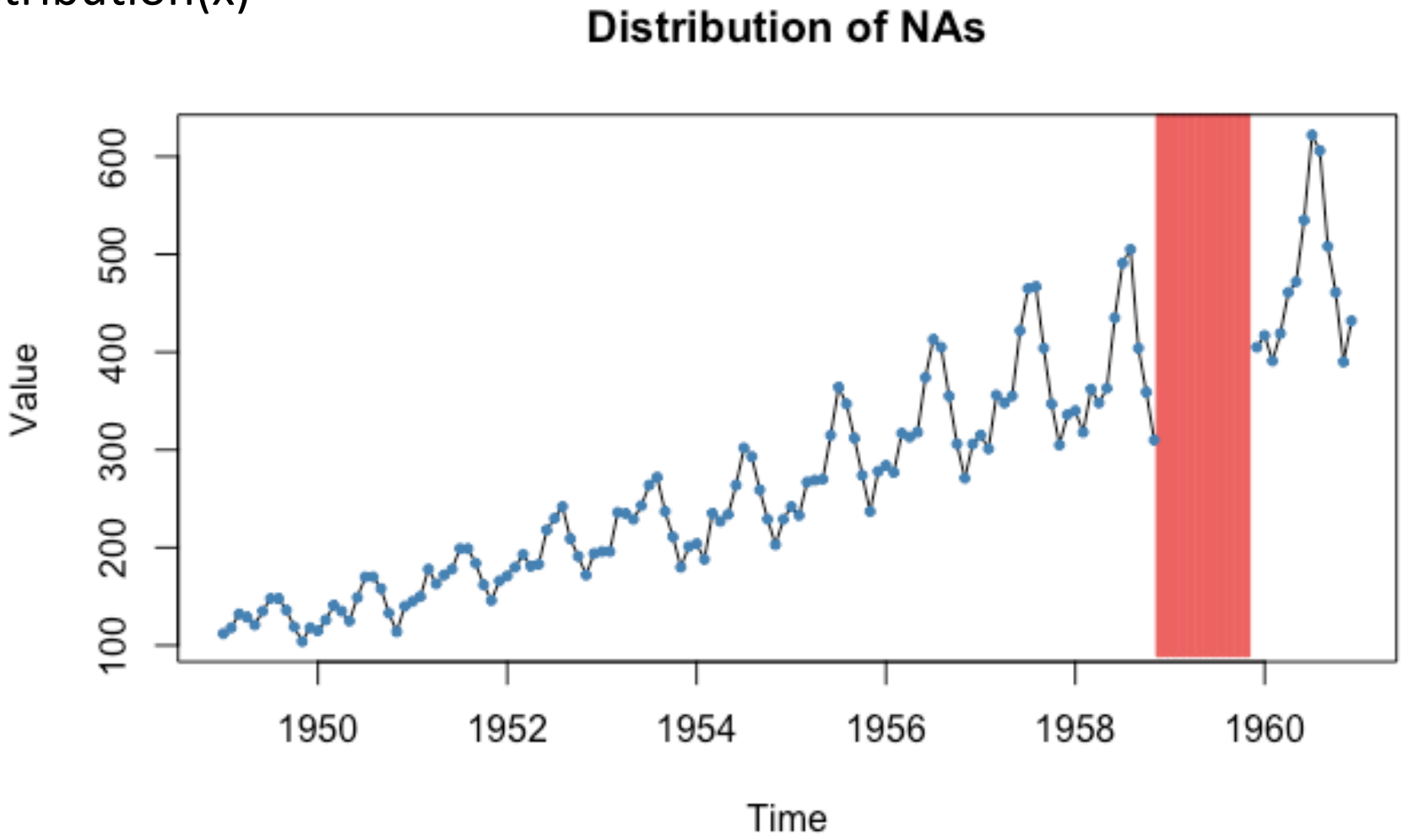
```
"Length of time series:"
606837
"------------------------"
"Number of Missing Values:"
57391
"------------------------"
"Percentage of Missing Values:"
"9.46%"
"------------------------"
"Stats for Bins"
"  Bin 1 (151710 values from 1 to 151710) :      0 NAs (0%)"
"  Bin 2 (151710 values from 151711 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 (occuring 104 times)"
"------------------------"
"Gap size accounting for most NAs"
```

Imputation Process

Step 2: Imputation

Visualization of NA distribution

plotNA.distribution(x)

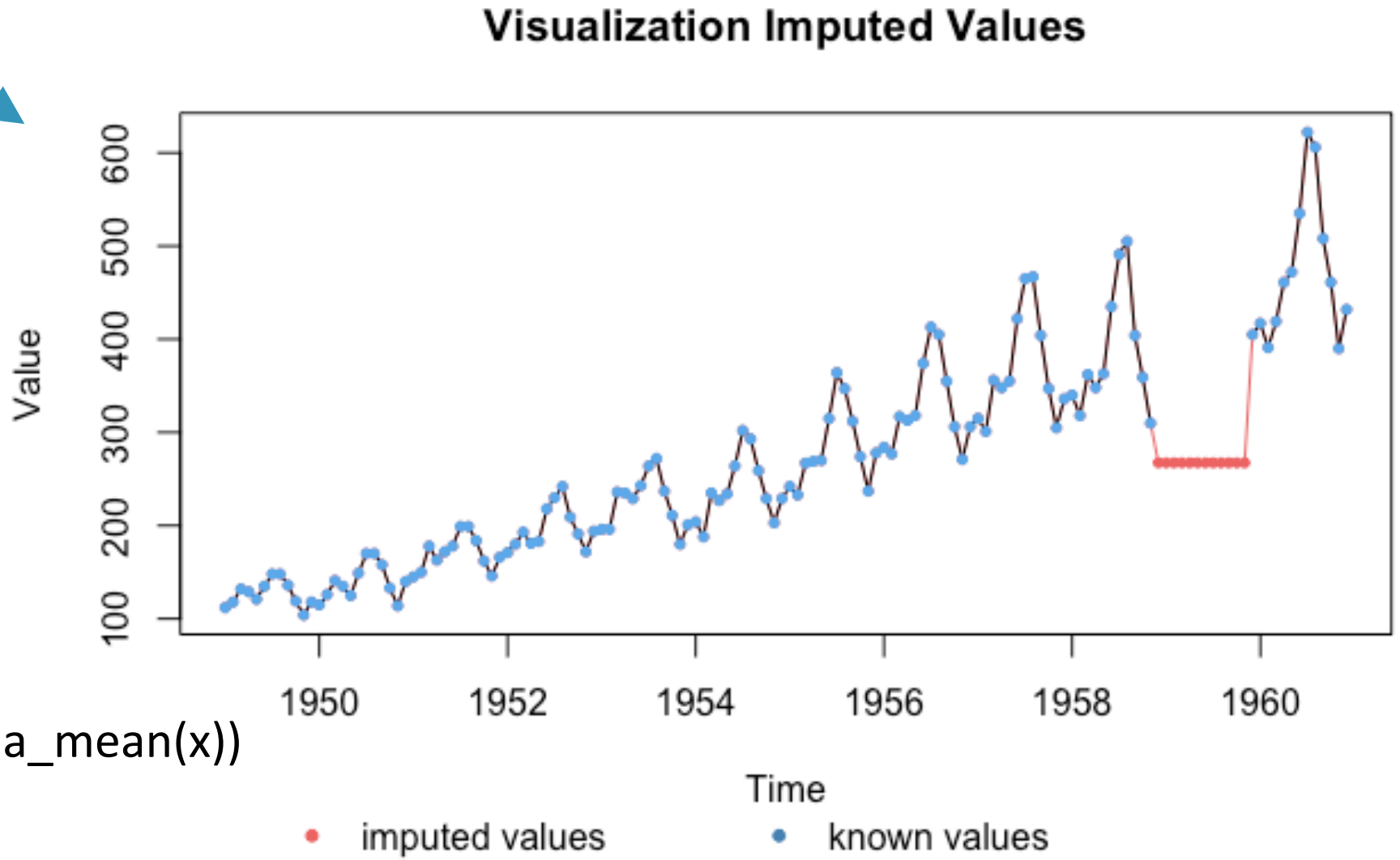


AirPassengers from datasets package with manually introduced NAs

Imputation with na_mean

na_mean(x)

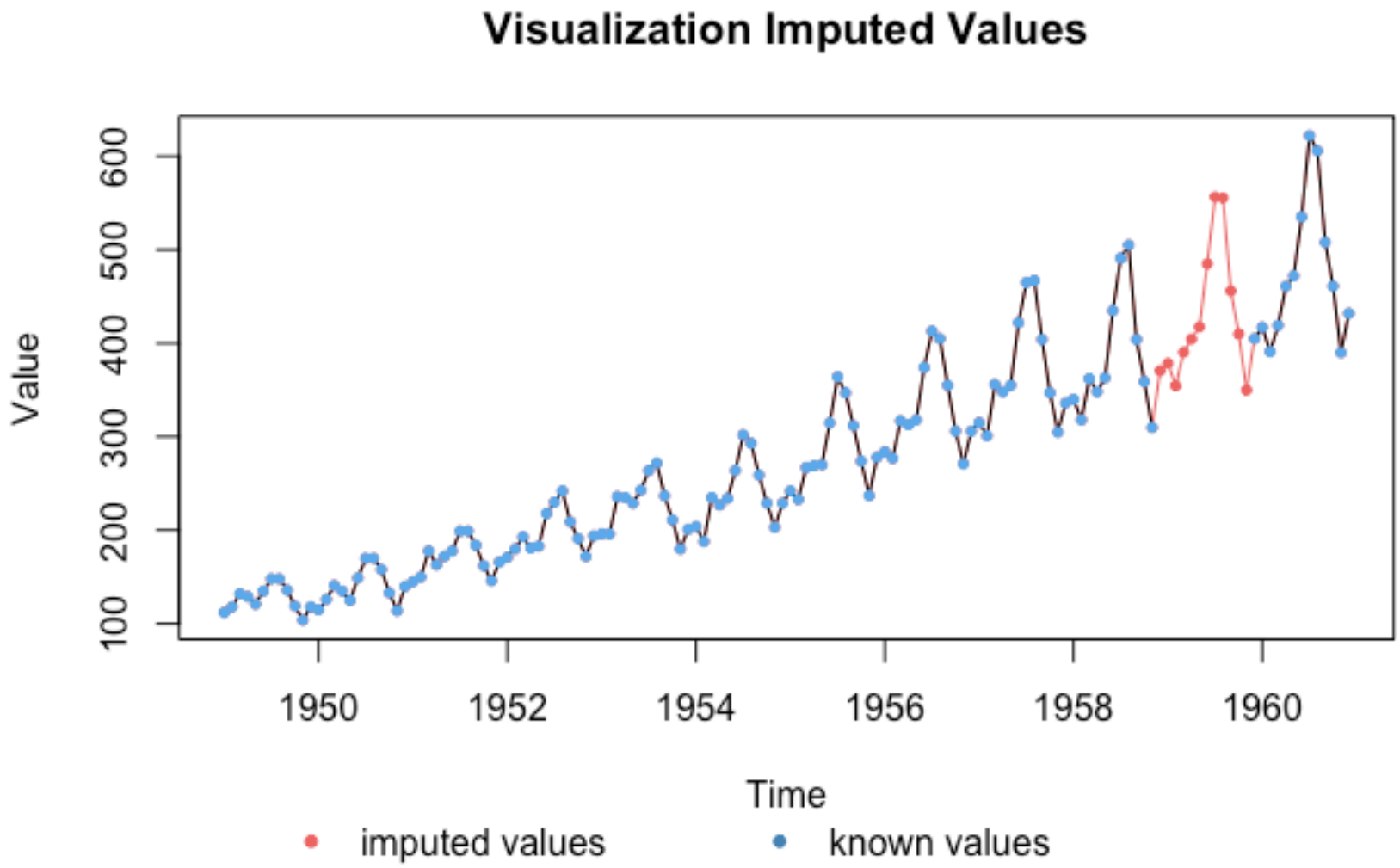
plotNA.imputations(x, na_mean(x))



AirPassengers from datasets package with manually introduced NAs

Imputation with na_seasplit

na_seasplit(x)



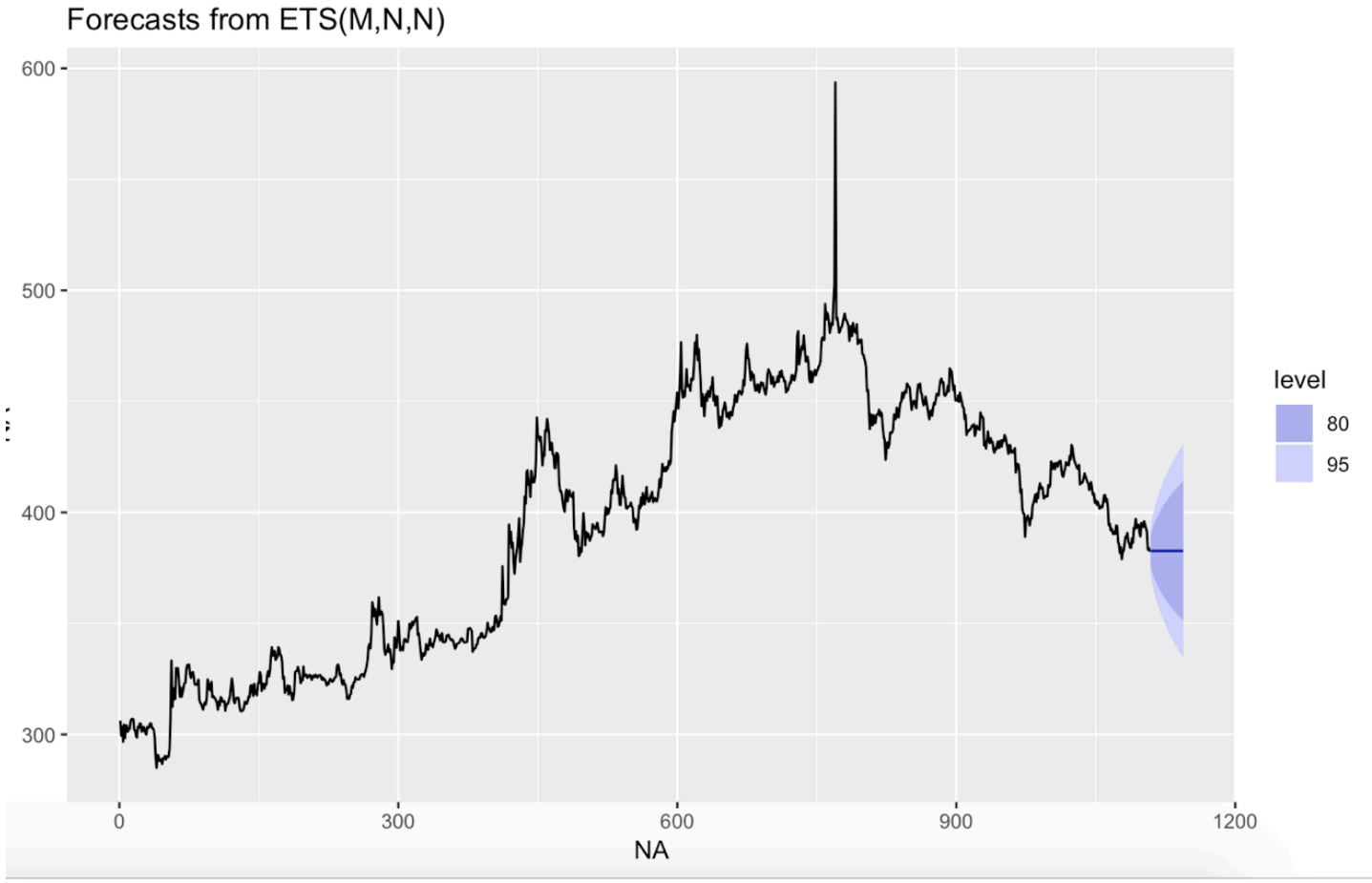
Imputation Process

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.