Regression with time series model errors with interpolation

Yujin Kim | 11-05-2022

Regular time series

- Time series analysis assumes equally spaced time-stamped measurements as well as most tools.
  - Regular time series (with missing values) — filling missing values with imputation method (e.g., ‘traces’ from Python, ‘imputeTS’ from R)

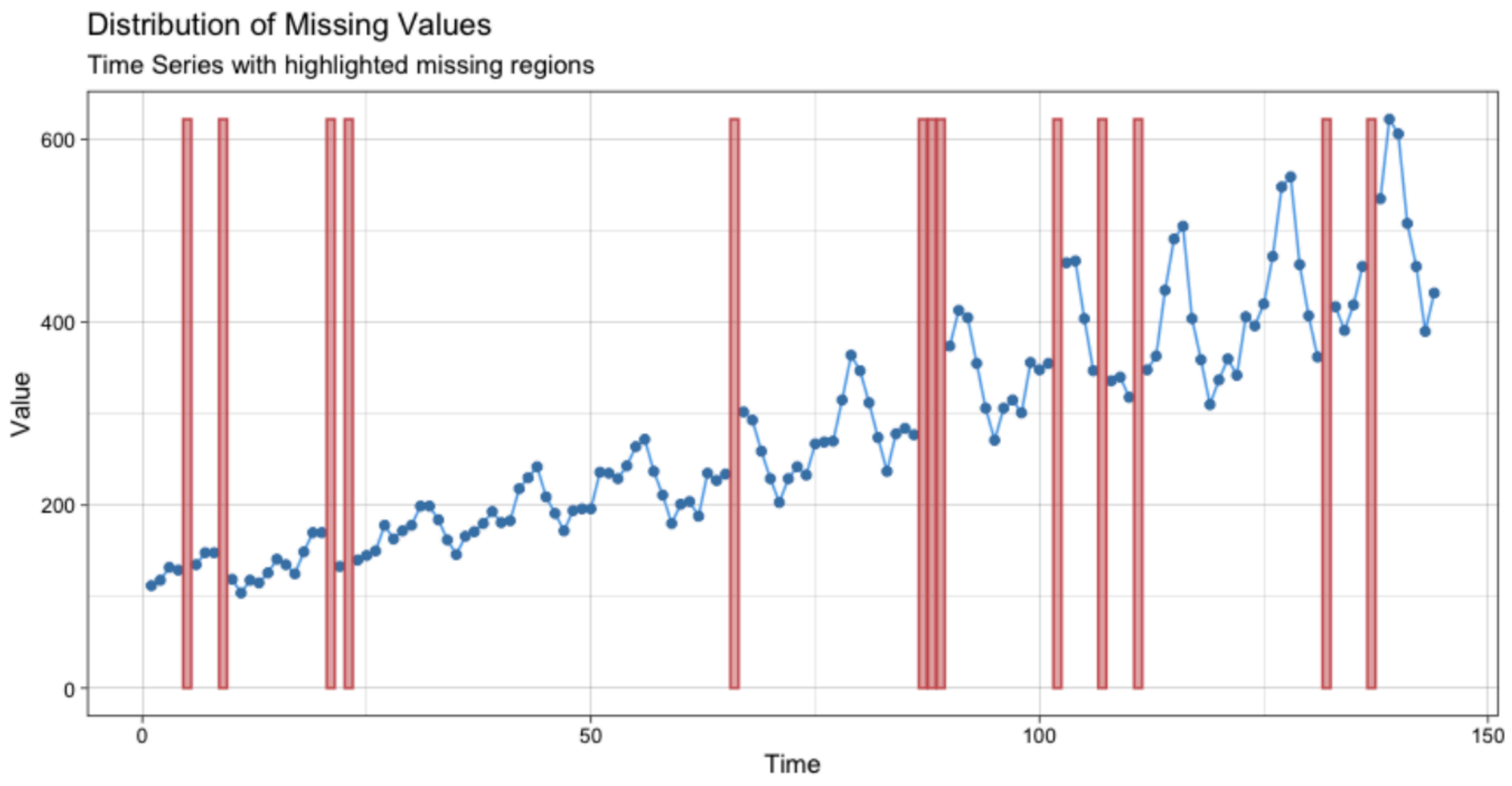
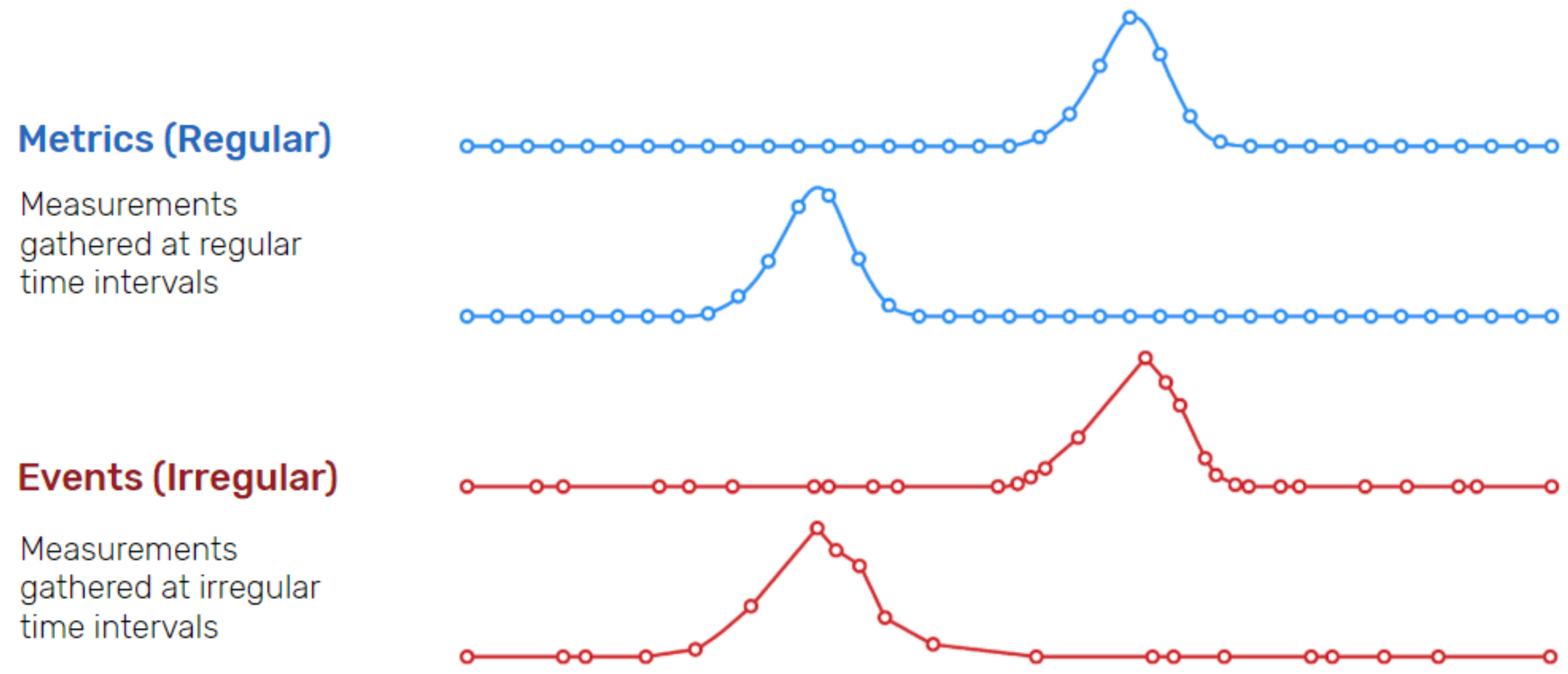


Figure 2: Example for ggplot_na_distribution

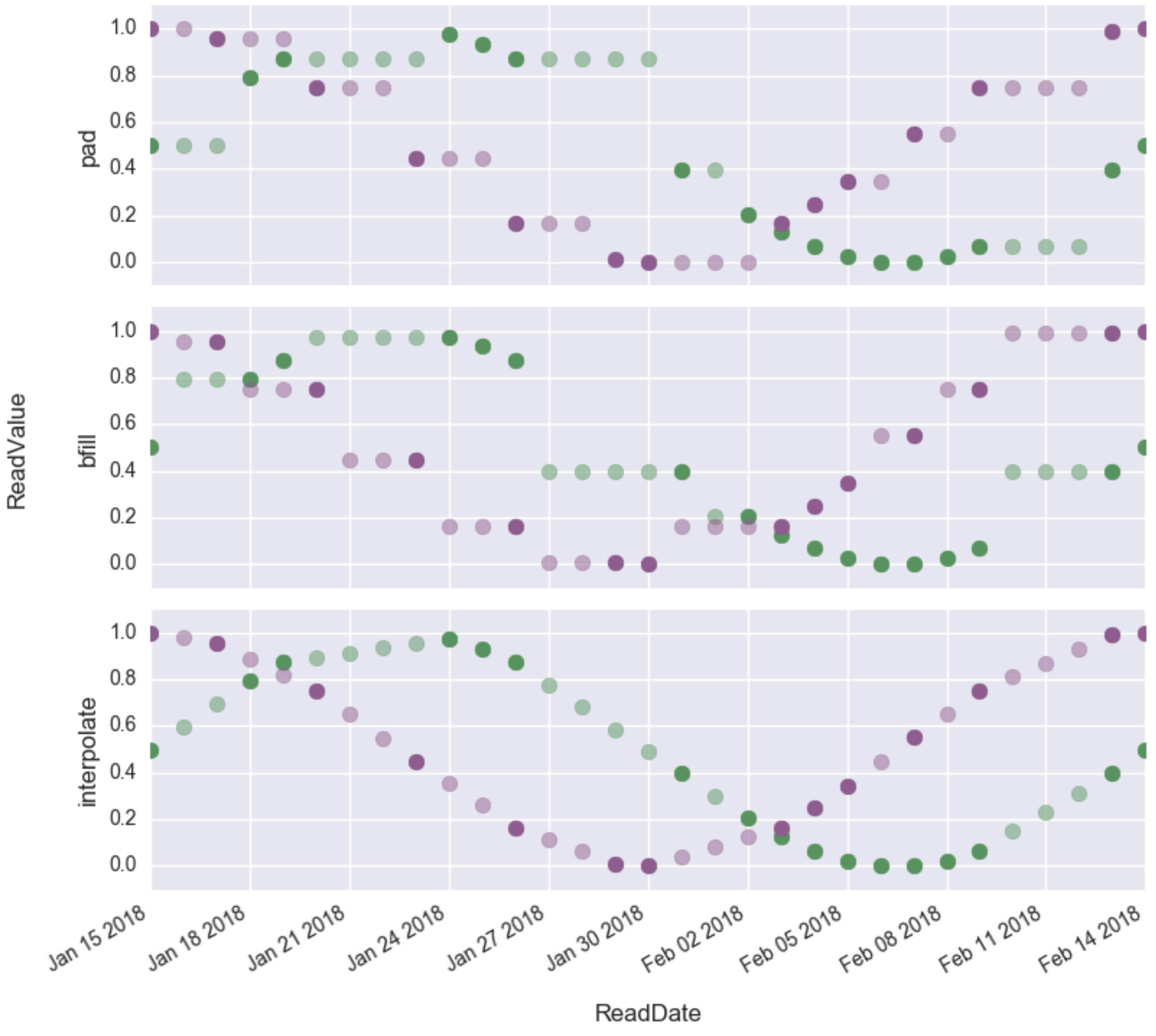
Irregular time series

- Time series with irregular intervals (e.g., events or irregular observations) is unpredictable and cannot be modeled or forecasted (violation of assumption)

- upsampling (e.g., from minutes to seconds) or downsampling (e.g., from days to months) and/or interpolation to make it regular



Interpolation (or smoothing)

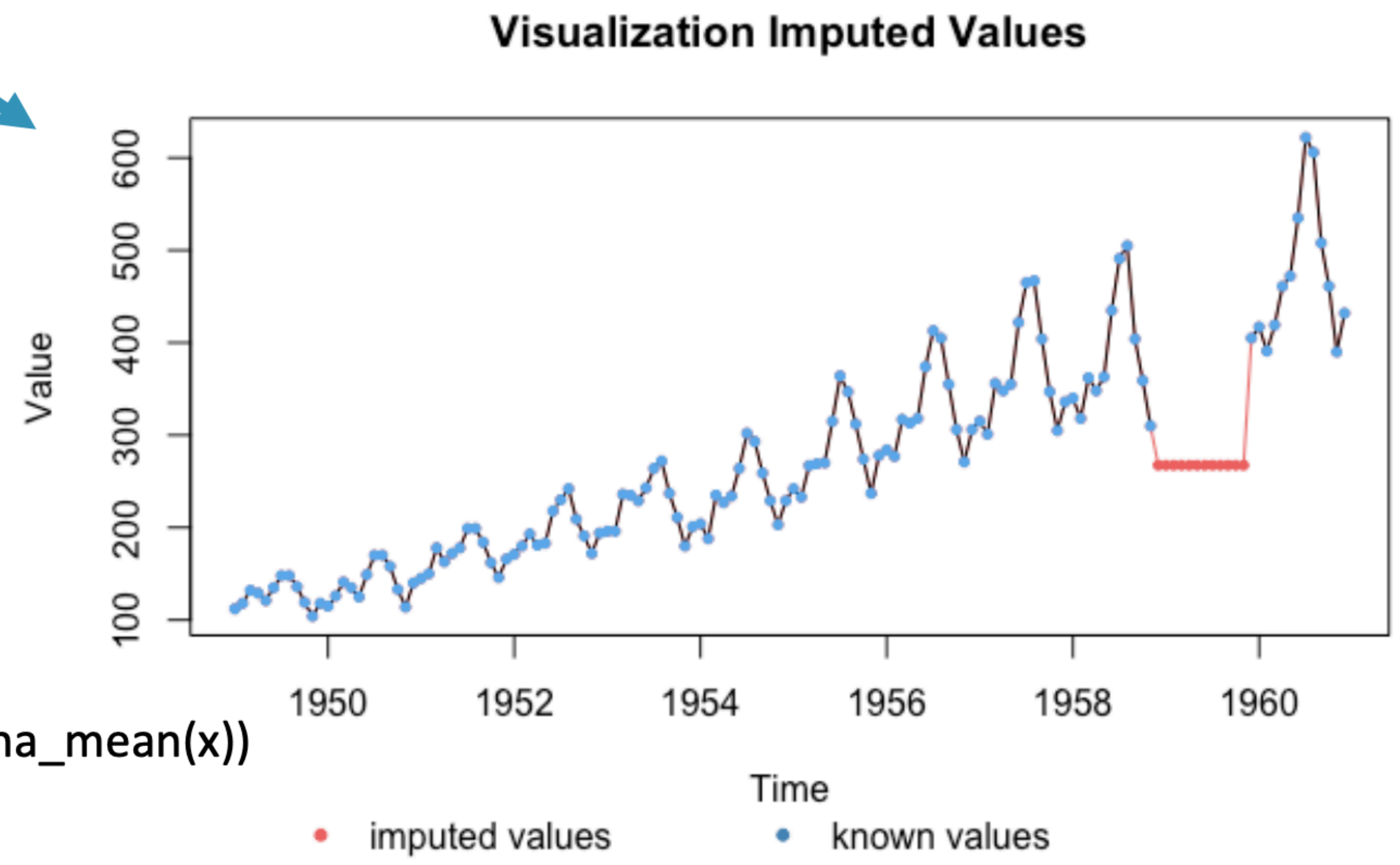


- Concept: predict values that fall within the range of data points taken (caution to use!)
  - Forward-fill (recent) / backward-fill (next) — can use both to interpolate (e.g., a simple spline in the plot, left)
  - Arithmetic mean, median, linear regression, regional weighting, spline interpolation, Stineman interpolation, and Kalman Smoothing imputations etc.
    - For more details see (https://doi.org/10.4136/ambi-agua.2795).
  - Avoid missing data is (usually) the best solution.

Examples

Imputation with na_mean

na_mean(x)

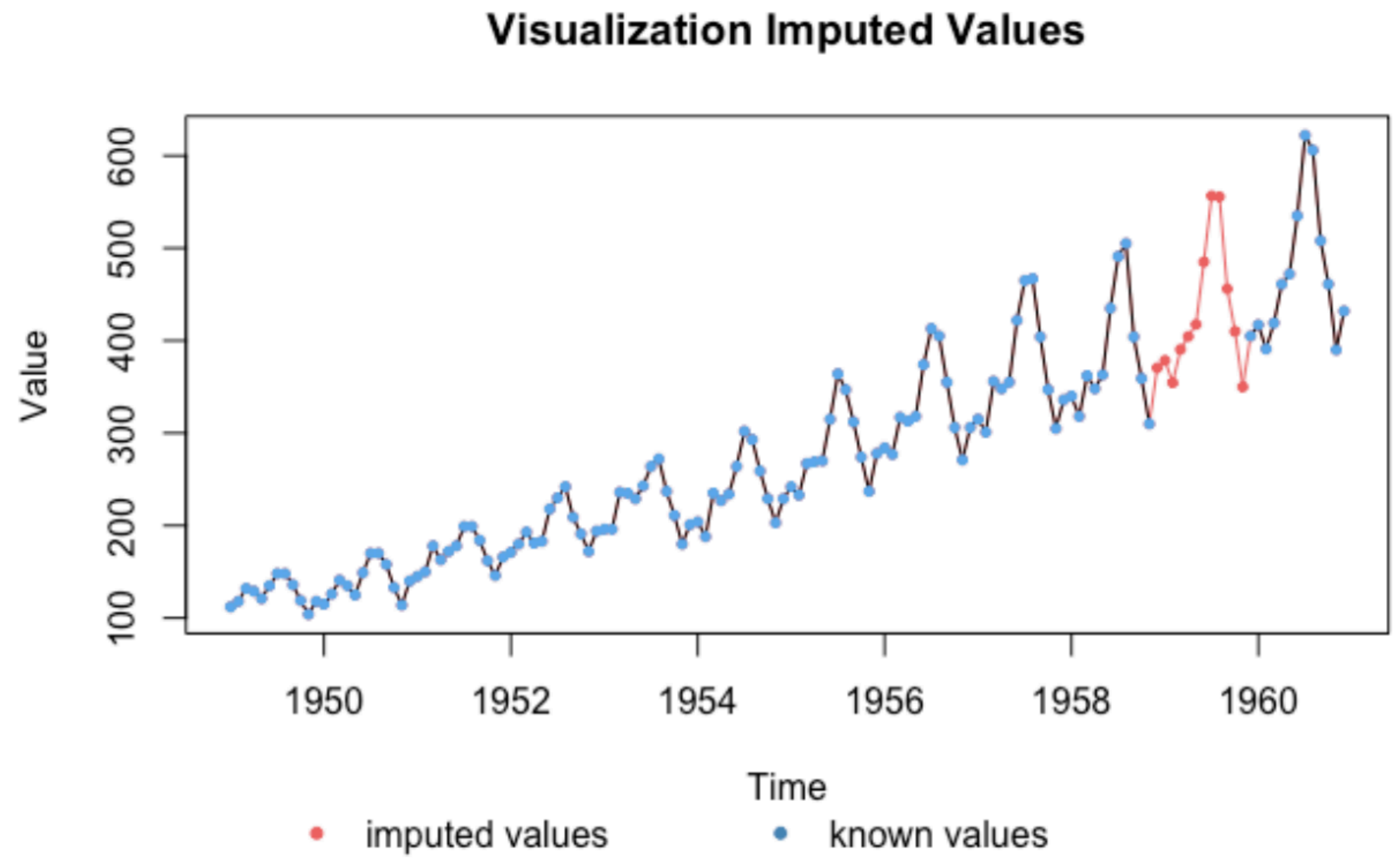


plotNA.imputations(x, na_mean(x))

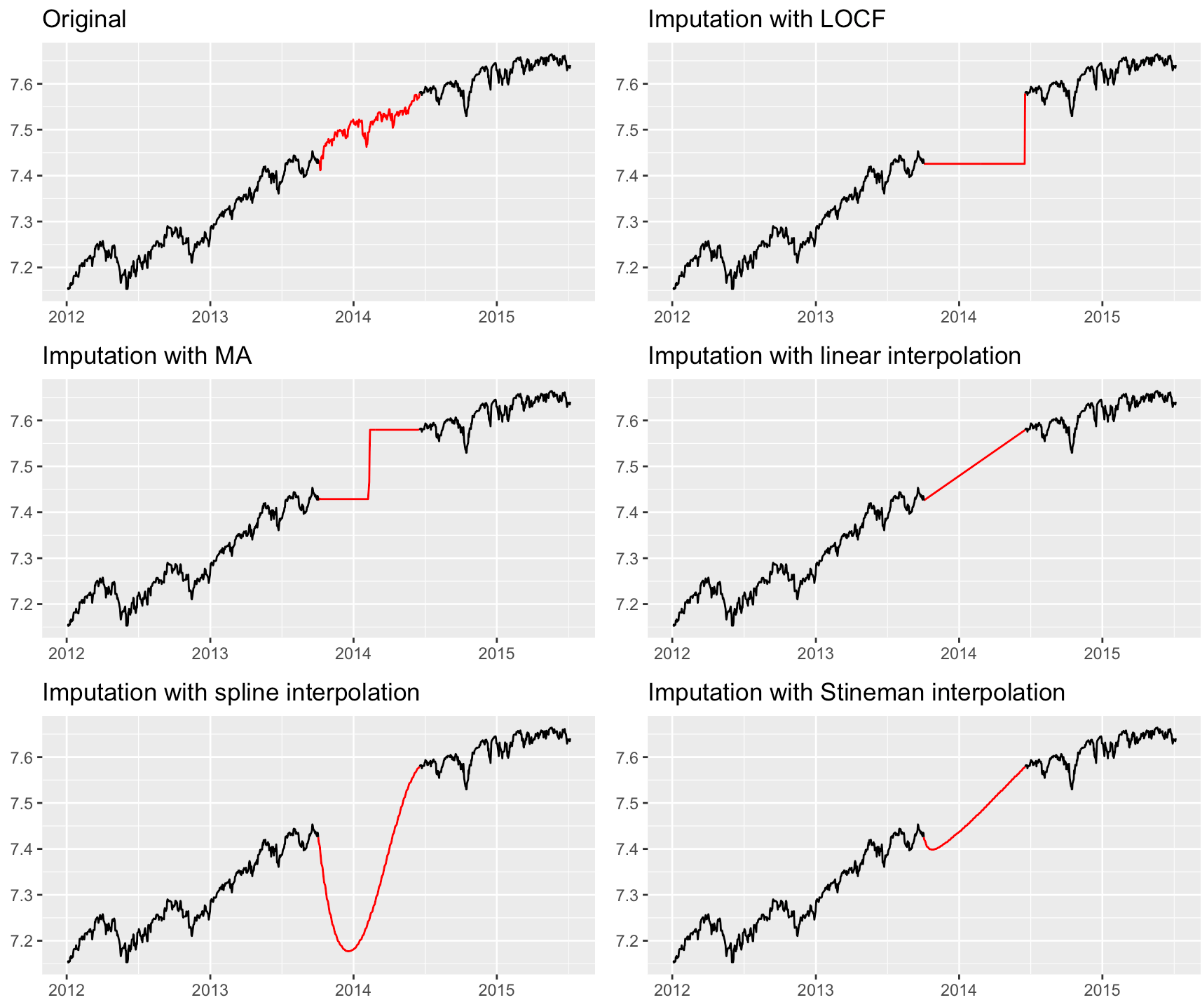
AirPassengers from datasets package with manually introduced NAs

Imputation with na_seasplit

na_seasplit(x)



Examples


Original
Imputation with LOCF
Imputation with MA
Imputation with linear interpolation
Imputation with spline interpolation
Imputation with Stineman interpolation
7.6
7.5
7.4
7.3
7.2
2012
2013
2014
2015

Regression with time series errors

- Let’s assume that we have a regular (or properly interpolated) time series data;
  - A classic regression model with uncorrelated errors

$$\boldsymbol{y} = \widehat{\boldsymbol{\beta}}\boldsymbol{X} + \boldsymbol{\epsilon}$$

$$\widehat{\boldsymbol{\beta}} = fitted\ model's\ regression\ coefficients$$

$$\boldsymbol{\epsilon} = residual\ errors\ of\ regression$$

  - The residual errors of time series are often auto-correlated (i.e., violation of iid)
  - We need the residual errors after controlling for all time series components identified

$$y_i = \widehat{\boldsymbol{\beta}}\boldsymbol{x_i} + \epsilon_i$$

$where\ \epsilon_i\ is\ modeled\ using$
$ARIMA(p,d,q)(P,D,Q)m$

The (S)ARIMA model

- The Auto-Regressive (AR) component is a linear combination of past values of the time series up to some number of lags p.

$$\boxed{y_i = \widehat{\phi_1} y_{i-1} + \widehat{\phi_2} y_{i-2} + \cdots + \widehat{\phi_p} y_{i-p} + \epsilon_i}$$

- The Moving Average (MA) component of SARIMA is a linear combination of the model's past errors up to some number of lags q. The model's past errors are calculated by subtracting the past predictions from past actual values.

$$\boxed{y_i = -\widehat{\theta_1} e_{i-1} - \widehat{\theta_2} e_{i-2} - \cdots - \widehat{\theta_p} e_{i-q} + \epsilon_i}$$

- Order of differencing (d): The ARMA model cannot be used if the time series has a trend (e.g., linear trend, quadratic trend and exponential or logarithmic trends).

- SAR, SMA, D and m: The Seasonal ARIMA or SARIMA model simply extends the above concepts concepts of AR, MA and differencing to the seasonal realm, and the seasonal period (m): ARIMA(p,d,q)(P,D,Q)m.

Error test

Accept/Reject the Null hypothesis of the Ljung-Box test that the residual errors are not auto-correlated.

```
                                      SARIMAX Results
============================================================================================
Dep. Variable:                           PT08_S4_NO2   No. Observations:                 7954
Model:             ARIMA(1, 1, 0)x(0, 1, [1], 24)   Log Likelihood              -48986.687
Date:                               Sun, 06 Sep 2020   AIC                          97983.374
Time:                                       22:30:25   BIC                          98018.265
Sample:                                   03-10-2004   HQIC                         97995.322
                                        - 02-05-2005
Covariance Type:                                 opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
T              5.3000      0.557      9.514      0.000       4.208       6.392
AH           509.7184      8.959     56.895      0.000     492.159     527.278
ar.L1         -0.0615      0.007     -8.385      0.000      -0.076      -0.047
ma.S.L24      -0.9117      0.004   -255.955      0.000      -0.919      -0.905
sigma2      1.353e+04    122.144    110.806      0.000    1.33e+04    1.38e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.72   Jarque-Bera (JB):              6655.59
Prob(Q):                              0.40   Prob(JB):                         0.00
Heteroskedasticity (H):               0.78   Skew:                             0.04
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.49
===================================================================================
```

Key takeaways

- Regression with (Seasonal) ARIMA errors (SARIMAX) is a time series regression model that brings together two powerful regression models namely, Linear Regression, and ARIMA (or Seasonal ARIMA).

- While configuring the time series decompositions, it helps to use a set of well-known rules (combined with personal judgement) for fixing the values of the p,d,q,P,D,Q and m parameters of the model. (in R, auto.arima() fixes)

- Make sure to test the residual errors of regression (e.g., the Ljung-Box test).

- Additionally, you would want the residual errors to be homoscedastic, and (preferably) normally distributed. Experiment with different combinations of p,d,q,P,D,Q until you get a model with the best goodness-of-fit characteristics.