Gap Filling

Time Series Analysis

Replacing missing values in time series: methods, usage, example

English

Wolfgang Boehme

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8 November 2026

The documentation on this website was developed with the support of the AI Claude (Anthropic).

Table of Contents

1 Introduction **2**

2 Loading Data **2**
2.1 Data Format 2
2.2 Example Data 2
2.3 Frequency Detection 2

3 The Six Methods **3**
3.1 Linear Interpolation 3
3.2 Last Known Value (LOCF) 3
3.3 Next Known Value (NOCB) 3
3.4 Value of a Previous Period (Lag) 3
3.5 Mean and Median 4

4 Usage **4**

5 Worked Example: Stock_A **5**

6 Choosing a Method **6**

1 Introduction

Many methods of time series analysis – transformations, filters, model estimation, structural-break tests – require series without gaps. Real data, however, often contain missing values: public holidays, reporting delays, transmission errors or changes in statistical methodology. *Gap Filling* replaces missing values using six common methods, shows the results of all methods side by side and exports the cleaned series together with the original data.

The tool works on the **original data** and independently of the other time series tools. It runs entirely in the browser; only the chart library Chart.js is loaded once from the internet when the page is opened.

2 Loading Data

2.1 Data Format

A CSV file is loaded via "Choose file":

```
Date,Stock_A,Crypto_B,Index_C
2024-01-01,100.5,1000,2500
2024-01-02,102.3,1050,2520
2024-01-03,,980,2510
```

- **Column 1** contains the time (date, year or any label), the value columns follow **from column 2**; each value column is a time series.
- Delimiter: semicolon if the header row contains a semicolon, otherwise comma. In semicolon files the decimal comma is allowed (`102,3`).
- Empty fields, `NaN` and non-numeric entries count as **missing**.
- Rows without an entry in the time column are ignored; quotation marks are removed.

Files from the Time Series Viewer (column `year` first) can also be loaded directly.

2.2 Example Data

"Load example data" loads 20 daily values (1–20 January 2024) of three fictitious series with six gaps in total:

Table 1: Gaps in the example data

Series	Gaps	Date
Stock_A	3	3 Jan, 10 Jan, 17 Jan
Crypto_B	2	5 Jan, 12 Jan
Index_C	1	7 Jan

2.3 Frequency Detection

The tool tries to read the time column as dates (formats `YYYY-MM-DD`, `DD.MM.YYYY` or month names). From the median of the intervals between consecutive points in time (in

days) it determines the frequency:

Table 2: Frequency detection

Median interval m (days)	Frequency
$m \leq 1.5$	daily
$1.5 < m \leq 9$	weekly
$9 < m \leq 45$	monthly
$45 < m \leq 200$	quarterly
$m > 200$	annual

If the time column cannot be read as dates (e. g. plain year numbers), the frequency is "unknown (index)" with step size 1. The detected frequency is only needed for the presets of the lag method.

3 The Six Methods

Let $y_1, \ldots, y_n$ be the series and y_i a missing value.

3.1 Linear Interpolation

If the gap lies between the known values y_s (before) and y_e (after), then

$$\hat{y}_i = y_s + (y_e - y_s)\,\frac{i - s}{e - s}, \qquad s < i < e. \tag{1}$$

Several consecutive gaps are distributed evenly along the connecting line. Gaps at the **beginning** of the series receive the first known value, gaps at the **end** the last known value. Interpolation assumes an even development between the neighbouring values and is usually the best simple choice for trending series.

3.2 Last Known Value (LOCF)

Last Observation Carried Forward: $\hat{y}_i = y_s$ with s = last known point in time before i. Typical for prices and stocks that do not change until the next observation (e. g. weekends). Gaps at the beginning of the series remain empty, as there is no predecessor.

3.3 Next Known Value (NOCB)

Next Observation Carried Backward: $\hat{y}_i = y_e$ with e = next known point in time after i. Gaps at the end of the series remain empty.

3.4 Value of a Previous Period (Lag)

$\hat{y}_i = y_{i-L}$, i. e. the value L points in time earlier. This allows seasonal patterns to be carried over (e. g. previous week for daily data with a weekly pattern, previous year for monthly data). Via the "Period" selection, L is calculated from the detected frequency:

$$L = \max\left(1,\ \text{round}\left(\frac{\text{period length in days}}{m}\right)\right) \tag{2}$$

with the period lengths 1 (previous day), 7 (previous week), 30.44 (previous month) and 365.25 (previous year); "Manual" allows L to be entered freely. If y_{i-L} is itself missing or $i - L$ lies before the start of the series, LOCF and then NOCB are used as a fallback – so the series is always filled completely.

3.5 Mean and Median

All gaps receive the same value: the arithmetic mean or the median of all known values of the series. These methods ignore the time pattern and are only suitable for series without a trend that fluctuate around a fixed level. The median is more robust to outliers.

4 Usage

1. **Select time series** – selection list at the top.
2. **Select method** – for "Value of a previous period" also the period or lag. Below, the detected frequency and the number of gaps are shown.
3. **Preview** – the chart shows the original series with gaps (grey) and the edited series (wine red).
4. **Comparison of all methods** – a second chart marks the values of all six methods at each gap; the selected method is highlighted as a diamond.
5. **Table** – for each gap the values of all methods, the selected method in bold.
6. **"Add as new time series"** – adds the cleaned series as a new column, e. g. "Stock_A (cleaned: interpolated)". The original series is kept; several methods can be added one after another.
7. **"Export as CSV"** – saves the time column, the original series and all added series in a file `gap_filling_export_YYYY-MM-DD.csv` (comma-separated, up to six decimal places, remaining gaps as empty fields). The file has the same format as an upload.

"Start new editing" discards all data and returns to the loading view.

5 Worked Example: Stock_A

The series Stock_A has 20 daily values, of which 17 are known and 3 missing. Detected frequency: daily ($m = 1$).

Table 3: Extract of the series Stock_A

Date (Jan)	2	**3**	4	9	**10**	11	16	**17**	18
Value	102.3	–	105.2	106.2	–	107.9	117.4	–	119.6

Linear interpolation on 3 Jan: $102.3 + (105.2 - 102.3) \cdot \frac{1}{2} = 103.75$; on 10 Jan: $(106.2 + 107.9)/2 = 107.05$; on 17 Jan: $(117.4 + 119.6)/2 = 118.50$.

Mean of the 17 known values: $1869.0/17 = 109.94$; **median**: 108.80 (9th value of the ordered series).

Table 4: Replacement values of all methods for Stock_A

Date	Linear	LOCF	NOCB	Lag (1)	Mean	Median
3 Jan	**103.75**	102.30	105.20	102.30	109.94	108.80
10 Jan	**107.05**	106.20	107.90	106.20	109.94	108.80
17 Jan	**118.50**	117.40	119.60	117.40	109.94	108.80

Lag with previous week: With daily frequency, the preset "Previous week" gives $L = \text{round}(7/1) = 7$. For 3 January the value one week earlier lies before the start of the series; for 10 and 17 January the previous week's value (3 and 10 January) is itself missing. In all three cases LOCF is therefore used, and the result matches the LOCF column – an example showing that the lag method requires sufficiently long series without clustered gaps.

Assessment: Stock_A has a clear upward trend (from 100.5 to 122.8). Mean and median insert values that are much too high at the beginning and too low at the end (on 3 January about 5–8 units above the neighbouring values) and thus distort returns and volatility. LOCF and NOCB each take over a neighbouring value and produce a zero return or a doubled return on one day. Linear interpolation lies between the neighbouring values and distributes the change evenly – the most plausible choice for this series.

6 Choosing a Method

Table 5: Recommendations

Situation	Recommended method
Trend, isolated gaps	linear interpolation
Prices, stocks (value valid until the next report)	LOCF
Values reported later that apply retroactively	NOCB
Pronounced seasonality (week, year)	lag with a suitable period
Stationary series without trend	mean; with outliers median
Long gaps (many periods)	none of the methods is reliable – shorten the period or estimate with a model (e. g. time series modelling)

Important: Replaced values are not observations. They smooth the series (interpolation), create artificial standstills (LOCF/NOCB) or reduce dispersion (mean/median). With many gaps, the results of later analyses should be checked for robustness using several methods; the method comparison in the tool shows how much the replacement values differ from each other.