Заполнение пропусков

Анализ временных рядов

Замена пропущенных значений во временных рядах: методы, работа, пример

Русский

Вольфганг Бёме

wolfgang.boehme@economic-modeling.de

www.economic-modeling.de

8 ноября 2026 г.

Документация на этом сайте разработана при поддержке ИИ Claude (Anthropic).

Содержание

1 Введение **2**

2 Загрузка данных **2**
2.1 Формат данных 2
2.2 Пример данных 2
2.3 Определение частоты 3

3 Шесть методов **3**
3.1 Линейная интерполяция 3
3.2 Последнее известное значение (LOCF) 3
3.3 Следующее известное значение (NOCB) 3
3.4 Значение предыдущего периода (лаг) 4
3.5 Среднее и медиана 4

4 Работа с инструментом **4**

5 Разобранный пример: Акция_A **5**

6 Выбор метода **6**

1 Введение

Многие методы анализа временных рядов — преобразования, фильтры, оценивание моделей, тесты на структурные сдвиги — требуют рядов без пропусков. Реальные данные, однако, часто содержат пропущенные значения: праздничные дни, задержки отчётности, ошибки передачи или изменения статистической методологии. Инструмент *Заполнение пропусков* заменяет пропущенные значения шестью распространёнными методами, показывает результаты всех методов рядом и экспортирует обработанные ряды вместе с исходными данными.

Инструмент работает с **исходными данными** и независимо от других инструментов анализа временных рядов. Он полностью работает в браузере; только библиотека графиков Chart.js один раз загружается из интернета при открытии страницы.

2 Загрузка данных

2.1 Формат данных

CSV-файл загружается кнопкой «Выбрать файл»:

```
Дата,Акция_A,Крипто_B,Индекс_C
2024-01-01,100.5,1000,2500
2024-01-02,102.3,1050,2520
2024-01-03,,980,2510
```

- **Столбец 1** содержит время (дату, год или любое обозначение), **начиная со столбца 2** следуют столбцы значений; каждый столбец значений — временной ряд.
- Разделитель: точка с запятой, если она есть в строке заголовков, иначе запятая. В файлах с точкой с запятой допускается десятичная запятая (`102,3`).
- **Пропущенными** считаются пустые поля, `NaN` и нечисловые записи.
- Строки без записи в столбце времени игнорируются; кавычки удаляются.

Файлы из инструмента «Просмотр временных рядов» (первым идёт столбец `year`) также можно загрузить напрямую.

2.2 Пример данных

Кнопка «Загрузить пример» загружает 20 дневных значений (1–20 января 2024 г.) трёх вымышленных рядов с шестью пропусками:

Таблица 1: Пропуски в примере данных

Ряд	Пропуски	Дата
Акция_A	3	03.01, 10.01, 17.01
Крипто_B	2	05.01, 12.01
Индекс_C	1	07.01

2.3 Определение частоты

Инструмент пытается прочитать столбец времени как даты (форматы `ГГГГ-ММ-ДД`, `ДД.ММ.ГГГГ` или названия месяцев). По медиане интервалов между соседними моментами времени (в днях) он определяет частоту:

Таблица 2: Определение частоты

Медианный интервал m (дни)	Частота
$m \le 1.5$	ежедневная
$1.5 < m \le 9$	еженедельная
$9 < m \le 45$	ежемесячная
$45 < m \le 200$	ежеквартальная
$m > 200$	годовая

Если столбец времени нельзя прочитать как даты (напр. просто номера лет), частота считается «неизвестной (индекс)» с шагом 1. Определённая частота нужна только для предустановок метода лага.

3 Шесть методов

Пусть $y_1, \ldots, y_n$ — ряд, а y_i — пропущенное значение.

3.1 Линейная интерполяция

Если пропуск находится между известными значениями y_s (до) и y_e (после), то

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

Несколько последовательных пропусков равномерно распределяются по соединяющей прямой. Пропуски в **начале** ряда получают первое известное значение, пропуски в **конце** — последнее известное значение. Интерполяция предполагает равномерное изменение между соседними значениями и обычно является лучшим простым выбором для рядов с трендом.

3.2 Последнее известное значение (LOCF)

Last Observation Carried Forward: $\hat{y}_i = y_s$, где s — последний известный момент времени до i. Типично для курсов и запасов, которые не меняются до следующего наблюдения (напр. выходные дни). Пропуски в начале ряда остаются пустыми, так как предшествующего значения нет.

3.3 Следующее известное значение (NOCB)

Next Observation Carried Backward: $\hat{y}_i = y_e$, где e — следующий известный момент времени после i. Пропуски в конце ряда остаются пустыми.

3.4 Значение предыдущего периода (лаг)

$\hat{y}_i = y_{i-L}$, т. е. значение на L моментов времени раньше. Так можно перенести сезонные закономерности (напр. предыдущая неделя для дневных данных с недельной закономерностью, предыдущий год для месячных данных). В поле «Период» величина L вычисляется по определённой частоте:

$$L = \max\left(1,\ \text{round}\left(\frac{\text{длина периода в днях}}{m}\right)\right) \tag{2}$$

с длинами периодов 1 (предыдущий день), 7 (предыдущая неделя), 30.44 (предыдущий месяц) и 365.25 (предыдущий год); вариант «Вручную» позволяет задать L произвольно. Если y_{i-L} само отсутствует или $i - L$ лежит до начала ряда, в качестве замены применяется LOCF, а затем NOCB — поэтому ряд всегда заполняется полностью.

3.5 Среднее и медиана

Все пропуски получают одно и то же значение: среднее арифметическое или медиану всех известных значений ряда. Эти методы игнорируют временную динамику и подходят только для рядов без тренда, колеблющихся вокруг постоянного уровня. Медиана более устойчива к выбросам.

4 Работа с инструментом

1. **Выбрать временной ряд** — список выбора вверху.
2. **Выбрать метод** — для «Значения предыдущего периода» также период или лаг. Ниже показаны определённая частота и число пропусков.
3. **Предпросмотр** — график показывает исходный ряд с пропусками (серый) и обработанный ряд (бордовый).
4. **Сравнение всех методов** — второй график отмечает в каждом пропуске значения всех шести методов; выбранный метод выделен ромбом.
5. **Таблица** — для каждого пропуска значения всех методов, выбранный метод выделен жирным.
6. **«Добавить как новый временной ряд»** — добавляет обработанный ряд как новый столбец, напр. «Акция_А (обработано: интерполяция)». Исходный ряд сохраняется; можно последовательно добавить результаты нескольких методов.
7. **«Экспорт в CSV»** — сохраняет столбец времени, исходные ряды и все добавленные ряды в файл `gap_filling_export_ГГГГ-ММ-ДД.csv` (разделитель — запятая, до шести знаков после запятой, оставшиеся пропуски — пустые поля). Файл имеет тот же формат, что и при загрузке.

Кнопка «Начать новую обработку» удаляет все данные и возвращает к экрану загрузки.

5 Разобранный пример: Акция_A

Ряд Акция_A содержит 20 дневных значений, из них 17 известны и 3 пропущены. Определённая частота: ежедневная ($m = 1$).

Таблица 3: Фрагмент ряда Акция_A

Дата	02.01	**03.01**	04.01	09.01	**10.01**	11.01	16.01	**17.01**	18.01
Значение	102.3	–	105.2	106.2	–	107.9	117.4	–	119.6

Линейная интерполяция 03.01: $102.3 + (105.2 - 102.3) \cdot \frac{1}{2} = 103.75$; 10.01: $(106.2 + 107.9)/2 = 107.05$; 17.01: $(117.4 + 119.6)/2 = 118.50$.

Среднее 17 известных значений: $1869.0/17 = 109.94$; **медиана**: 108.80 (9-е значение упорядоченного ряда).

Таблица 4: Заменяющие значения всех методов для Акция_A

Дата	Линейная	LOCF	NOCB	Лаг (1)	Среднее	Медиана
03.01	**103.75**	102.30	105.20	102.30	109.94	108.80
10.01	**107.05**	106.20	107.90	106.20	109.94	108.80
17.01	**118.50**	117.40	119.60	117.40	109.94	108.80

Лаг «предыдущая неделя»: при ежедневной частоте предустановка «Предыдущая неделя» даёт $L = \text{round}(7/1) = 7$. Для 03.01 значение неделей раньше лежит до начала ряда, для 10.01 и 17.01 значение предыдущей недели (03.01 и 10.01) само пропущено. Поэтому во всех трёх случаях применяется LOCF, и результат совпадает со столбцом LOCF — пример того, что метод лага требует достаточно длинных рядов без скоплений пропусков.

Оценка: у ряда Акция_A выраженный восходящий тренд (со 100.5 до 122.8). Среднее и медиана подставляют в начале ряда слишком высокие, а в конце слишком низкие значения (03.01 примерно на 5–8 единиц выше соседних значений) и тем самым искажают доходности и волатильность. LOCF и NOCB переносят соседнее значение и создают нулевую доходность или удвоенную доходность за один день. Линейная интерполяция лежит между соседними значениями и равномерно распределяет изменение — для этого ряда это наиболее правдоподобный выбор.

6 Выбор метода

Таблица 5: Рекомендации

Ситуация	Рекомендуемый метод
Тренд, отдельные пропуски	линейная интерполяция
Курсы, запасы (значение действует до следующего сообщения)	LOCF
Значения, сообщаемые позднее и действующие задним числом	NOCB
Выраженная сезонность (неделя, год)	лаг с подходящим периодом
Стационарный ряд без тренда	среднее; при выбросах — медиана
Длинные пропуски (много периодов)	ни один метод не надёжен — сократить период или оценить по модели (напр. моделирование временных рядов)

Важно: заменённые значения не являются наблюдениями. Они сглаживают ряд (интерполяция), создают искусственные периоды неизменности (LOCF/NOCB) или уменьшают разброс (среднее/медиана). При большом числе пропусков результаты последующего анализа следует проверять на устойчивость с помощью нескольких методов; сравнение методов в инструменте показывает, насколько сильно различаются заменяющие значения.