РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я
№ 1 1998

УДК 519.246.8

С. Н. Моисеев
(Воронеж)

ЗАПОЛНЕНИЕ ПРОПУСКОВ В СЛУЧАЙНО-ЦЕНЗУРИРОВАННЫХ ВРЕМЕННЫХ РЯДАХ

Предложены алгоритмы заполнения пропусков в случайно-цензурированных снизу стационарных зависимых временных рядах из класса распределений, которые сводятся к гауссову монотонным безынерционным преобразованием. Заполнение не искажает вероятностных свойств временных рядов.

Введение. Построение вероятностной модели изучаемого явления часто приходится вести по выборке временного ряда с пропусками [1]. Существуют два подхода к работе с такими неполными данными: создание алгоритмов обработки, учитывающих пропуски [2, 3], и первоначальное заполнение пропусков с дальнейшей работой по полной выборке [2]. Первый подход требует значительных априорных сведений о вероятностных свойствах ряда, которые обычно не известны на практике. К тому же экспресс-анализ выборки и построение модели предполагают применение большого числа статистических тестов, разработанных, как правило, только для полных выборок.

Второй подход требует заполнения пропусков на самом первом этапе работы с выборкой и в дальнейшем без ограничений использования всего многообразия статистических тестов, разработанных для полных выборок. Он очень удобен для практических целей, чем и объясняется его широкое распространение. Тем не менее в конкретных исследованиях либо работают по участкам выборки, не содержащим пропусков, либо используют простейшие приемы заполнения пропусков подходящей константой, прогнозом и т. п. Это может приводить к существенным ошибкам в статистических выводах.

В настоящей работе предлагаются алгоритмы заполнения пропусков в случайно-цензурированных временных рядах, не приводящие к искажениям их вероятностных свойств при неограниченном возрастании объема выборки.

Постановка задачи. Пусть из n отсчетов выборки временного ряда y_t наблюдаются n_{obs} отсчетов в моменты времени $t \in T_{obs}$, а $n_{mis} = n - n_{obs}$ отсчетов в моменты времени $t \in T_{mis}$ отсутствуют. Обозначим множество наблюдаемых значений ряда через Y_{obs} (observed — наблюдаемый), а множество отсутствующих — через Y_{mis} (missing — отсутствующий). Относительно механизма порождения пропусков известно, что ряд y_t случайно-цензурированный снизу (слева), т. е. отсутствуют отсчеты $y_t < c_t$, где c_t — отсчеты случайной цензуры. Таким образом, механизм порождения пропусков зависит от значений как пропущенных Y_{mis}, так и присутствующих Y_{obs} отсчетов ряда y_t. Такая ситуация является достаточно типичной при сборе данных в условиях слабоконтролируемого эксперимента. Методы же заполнения пропусков, рассмотренные в [2], предполагают выполнение довольно жесткого условия независимости механизма порождения пропусков от значений $y_t \in Y_{mis}$.

Сделаем несколько относительно мягких предположений о вероятностных свойствах рядов y_t и c_t, необходимых для корректного решения задачи. Пусть стационарный ряд y_t принадлежит классу распределений, которые сводятся к гауссову монотонным безынерционным преобразованием. Этот класс распре-

делений довольно широк и часто встречается на практике [1, 4]. Как частный случай он полностью включает в себя ряды с независимыми отсчетами, плотности вероятностей которых не содержат сингулярностей в виде дельта-функций. Положим, что c_t — стационарный в узком смысле независимый от y_t временной ряд с одномерной интегральной функцией распределения $G(x)$. Минимальные возможные значения рядов y_t и c_t обозначим через y_H и c_H. В дальнейшем покажем, как можно еще более смягчить требования к y_t и c_t.

Необходимо заполнить пропуски ряда y_t такими значениями, чтобы при увеличении объема выборки $n \to \infty$ и неизменной доле пропусков была справедлива следующая сходимость по вероятности:

$$\sup_{\mathbf{x}} \left| \hat{F}_p(\mathbf{x}) - F_p(\mathbf{x}) \right| \to 0, \tag{1}$$

где $\hat{F}_p(\mathbf{x})$ и $F_p(\mathbf{x})$, $\mathbf{x} = \{x_1, ..., x_p\}$ — соответственно эмпирическая и истинная p-мерные интегральные функции распределения ряда y_t. Условие (1) означает, что при неограниченном увеличении объема выборки n распределение заполненной выборки будет совпадать с истинным.

Оценка нормализующей функции. На первом этапе найдем оценку монотонной безынерционной функции $f(x)$, которая сводит y_t к гауссову ряду с нулевым средним и единичной дисперсией. Теоретически эта неубывающая функция имеет следующий вид:

$$f(x) = \Phi^{-1}\left[F(x)\right], \tag{2}$$

где $F(x)$ — одномерная интегральная функция распределения ряда y_t; $\Phi^{-1}[\cdot]$ — функция, обратная к интегралу вероятностей:

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left\{-\frac{x^2}{2}\right\} dx.$$

По конечной выборке функцию $f(x)$ можно оценить в конечном числе точек m.

Предположим вначале, что значения цензуры c_t в моменты времени $t \in T_{\text{mis}}$ известны. В этом случае наблюдаем ряд z_t, состоящий из отсчетов $y_t \in Y_{\text{obs}}$ в моменты времени $t \in T_{\text{obs}}$ и отсчетов c_t в моменты времени $t \in T_{\text{mis}}$. При фиксированном t случайная величина z_t будет иметь функцию распределения

$$F_z(x) = F(x)G(x). \tag{3}$$

С помощью соотношения (3) можно найти выборочную оценку функции $F(x)$. Разобьем диапазон возможных значений z_t на $m + 1$ непересекающихся интервалов ненулевой длины. Границы интервалов $x_1, ..., x_m$, $m \geq 2$, $x_m = \max_t z_t$, удобно выбирать так, чтобы количество попавших в интервал отсчетов z_t было одинаковым для всех интервалов. Рекомендации по выбору числа интервалов содержатся в [5]. Согласно им, $m + 1 \approx 1{,}9n^{0{,}4}$. Выборочная оценка функции распределения $F(x)$ в m точках будет иметь вид:

$$F(x_i) = \frac{n_i}{nG(x_i)}, \quad i = \overline{1, m},$$

где n_i — число отсчетов z_t, попавших в интервал $]-\infty, x_i]$, а оценка нормализующей функции $f(x)$ в m точках

$$f_i = f(x_i) = \Phi^{-1}\left[n_i/(nG(x_i))\right], \quad i = \overline{1, m}. \tag{4}$$

Рассмотрим теперь случай, когда значения цензуры c_t не известны. Оценку функции $f(x)$ будем строить по наблюдаемым данным $y_t \in Y_{obs}$. Вероятность того, что значение $y_t \in Y_{obs}$ попадет в интервал dx, равна

$$W_{obs}(x)dx = \frac{1}{q} G(x)W(x)dx, \tag{5}$$

где $W(x)$ — одномерная плотность вероятностей ряда y_t;

$$q = \int_{-\infty}^{\infty} G(x)W(x)dx = P\left[y_t > c_t\right].$$

Откуда

$$F(x) = F(c_H) + q \int_{c_H}^{x} \left[W_{obs}(t)/G(t)\right] dt, \quad x \geq c_H.$$

Учитывая, что выборочные оценки параметра q и плотности вероятностей наблюдаемых отсчетов $W_{obs}(x)$ имеют вид:

$$\hat{q} = n_{obs}/n, \quad \hat{W}_{obs}(x_i) = k_i/n_{obs}, \quad i = \overline{1, m}, \quad m + 1 \approx 1{,}9n_{obs}^{0{,}4}, \tag{6}$$

где k_i — число отсчетов $y_t \in Y_{obs}$, попавших в интервал $]x_{i-1}, x_i]$; $x_m = \max\limits_t y_t \in Y_{obs}$; $x_0 = \min\limits_t y_t \in Y_{obs}$, получаем выборочную оценку функции распределения $F(x)$

$$\hat{F}(x_i) = 1 - \frac{1}{n} \sum_{j=i+1}^{m} k_j / G(x_j), \quad i = \overline{1, m},$$

и оценку нормализующей функции $f(x)$ в m точках

$$f_i = f(x_i) = \Phi^{-1}\left[1 - \frac{1}{n} \sum_{j=i+1}^{m} k_j / G(x_j)\right], \quad i = \overline{1, m}. \tag{7}$$

Между точками x_i, $i = \overline{1, m}$, и вне диапазона их значений можно использовать интерполяцию и экстраполяцию функции $f(x)$. При использовании линейной интерполяции имеем

$$f(x) = \frac{f_i - f_{i-1}}{x_i - x_{i-1}} (x - x_{i-1}) + f_{i-1}, \quad x_{i-1} < x \leq x_i, \quad i = \overline{2, m}. \tag{8}$$

Для значений $x \leq x_1$ берем формулу (8) при $i = 2$, а для $x > x_m$ — при $i = m$ (линейная экстраполяция). Из (8) легко находим функцию, обратную $f(x)$:

$$X(f) = \frac{x_i - x_{i-1}}{f_i - f_{i-1}} (f - f_{i-1}) + x_{i-1}, \quad f_{i-1} < f \leq f_i, \quad i = \overline{2, m}. \tag{9}$$

Для значений $f \leq f_1$ берем (9) при $i = 2$, а для $f > f_m$ — при $i = m$.

В качестве аппроксимации функции $\Phi^{-1}[x]$ можно использовать формулу

$$\Phi^{-1}[x] = \begin{cases} \varphi[\sqrt{-2\ln(1 - x)}], & 0{,}5 \leq x < 1, \\ -\varphi[\sqrt{-2\ln x}], & 0 < x < 0{,}5, \end{cases}$$

где $\varphi(y) = y - \dfrac{a_0 + a_1 y}{1 + b_1 y + b_2 y^2}$, $a_0 = 2{,}30753$, $a_1 = 0{,}27061$, $b_1 = 0{,}99229$, $b_2 = 0{,}04481$.

Заполнение пропусков. При заполнении пропусков необходимо учитывать статистические связи между отсчетами ряда y_t. Перейдем от ряда y_t к ряду

$$u_t = f(y_t). \tag{10}$$

Новый ряд u_t будет гауссовым с нулевым средним и единичной дисперсией.

Во избежание нарушения статистических связей между отсчетами y_t пропуски будем заполнять строго последовательно слева направо. Допустим, что первый слева пропуск наблюдается в момент времени $t = k \in T_{\text{mis}}$. Сначала необходимо по наблюдаемым участкам ряда u_t, $t \in T_{\text{obs}}$, оценить мерность ν функции регрессии (порядок марковости) ряда u_t хорошо известными методами для полных выборок [6, 7] (ν определяется как число первых отличных от нуля частных корреляций). Далее из множества ближайших к пропуску наблюдаемых отсчетов ряда u_t выбираются ν отсчетов слева от пропуска $u_{k-1}, \ldots, u_{k-\nu}$ и $\mu \geq \nu$ отсчетов справа от пропуска $u_{k+l_1}, \ldots, u_{k+l_\mu}$. Число $\mu = \nu$ только для марковских процессов (рядов авторегрессии первого порядка). В общем случае $\mu \geq \nu$, так как разрежение ряда теоретически может увеличивать порядок марковости вследствие отсутствия промежуточной информации. Оценку μ можно получить, вычисляя мерность функции регрессии по наблюдаемым участкам выборки при фиксированной структуре пропусков $u_{t+l_1}, \ldots, u_{t+l_\mu}$. Поскольку при разрежении уменьшается абсолютная корреляция между крайними отсчетами u_k и u_{k+l_μ}, то на практике почти всегда можно положить $\mu = \nu$. Для гауссовых рядов условное среднее отсчета u_k при фиксированных выбранных отсчетах будет линейной функцией

$$\eta_k(u_{k-i},\ u_{k+l_j},\ i = \overline{1,\nu};\ j = \overline{1,\mu}) = \sum_{i=1}^{\nu} \alpha_i u_{k-i} + \sum_{j=1}^{\mu} \alpha_{l_j} u_{k+l_j}, \tag{11}$$

причем $\alpha_i = \alpha_{l_j}$ при $i = l_j$. Оценки параметров α_i, α_{l_j} легко получить из условия минимума дисперсии σ^2 величины $u_t - \eta_t(u_{t-i},\ u_{t+l_j},\ i = \overline{1,\nu};\ j = \overline{1,\mu})$ по наблюдаемым участкам ряда u_t при тех же l_j, $j = \overline{1,\mu}$, что и в (11) [8, с. 75]. Система уравнений для оценок при $\mu = \nu$ имеет вид:

$$\sum_{j=1}^{\nu} (\alpha_j r_{i-j} + \alpha_{l_j} r_{l_j+i}) = r_i; \qquad \sum_{j=1}^{\nu} (\alpha_j r_{l_i+j} + \alpha_{l_j} r_{l_j-l_i}) = r_{l_i}, \quad i = \overline{1,\nu};$$

$$\sigma^2 = 1 - \sum_{i=1}^{\nu} (\alpha_i r_i + \alpha_{l_i} r_{l_i}),$$

где r_i — коэффициент корреляции ряда u_t при задержке i. Например, при $\nu = 1$ (u_t — процесс авторегрессии первого порядка) найдем

$$\eta_k(u_{k-1}, u_{k+l}) = \alpha_1 u_{k-1} + \alpha_l u_{k+l}; \qquad \alpha_i = \frac{r_1^i (1 - r_1^{2l/i})}{1 - r_1^{2l+2}}, \quad i = 1, l;$$

$$\sigma^2 = \frac{(1 - r_1^2)(1 - r_1^{2l})}{1 - r_1^{2l+2}}.$$

Благодаря тому, что ряд u_t гауссов с известными средним и дисперсией, оценки ν, μ, σ^2 и α_i, α_{l_j}, $i = \overline{1,\nu}$; $j = \overline{1,\mu}$, полученные по наблюдаемой части выборки, будут состоятельными.

Пропуск в ряде u_t в момент времени $t = k$ заполняется значением

$$u_k^* = \eta_k(u_{k-i}, u_{k+j},\ i = \overline{1,\nu};\ j = \overline{1,\mu}) + \xi_k \quad \text{при} \quad u_k^* < f(c_k), \tag{12}$$

где ξ_k — моделируемые независимые гауссовы величины с нулевым средним и дисперсией σ^2. Если известных отсчетов слева от пропуска меньше ν или справа от пропуска меньше μ, то в (11) и (12) используется функция $\eta_k(\cdot)$ меньшей размерности. При неизвестном значении цензуры c_k в (12) она моделируется каждый раз в паре с ξ_k из распределения $G(x)$. При наличии зависимости между отсчетами ряда c_t значение c_k моделируется с учетом зависимости от ранее смоделированных отсчетов цензуры. Моделирование ξ_k (или пары ξ_k и c_k) повторяется до тех пор, пока не будет выполнено неравенство в (12). После заполнения пропуска в ряде u_t пропуск в исходном ряде y_t заполняется значением

$$y_k^* = X(u_k^*), \tag{13}$$

где $X(\cdot)$ — функция, обратная $f(x)$ (9). Для известных значений цензуры c_t оценка функции $f(x)$ в (12) и (8) дается выражением (4), для неизвестных цензур — выражением (7). После заполнения пропуска значение u_k считается известным, и заполняется следующий пропуск по правилу (11)—(13).

Заключение. Приведенные алгоритмы заполнения будут удовлетворять критерию (1) при условии $y_H \geq c_H$. Если это условие не выполнено (не могут наблюдаться отсчеты $y_t < c_H$), качество заполнения зависит от точности экстраполяции функции $f(x)$ в области значений $x < c_H$. Например, для гауссовых рядов y_t, даже при $y_H < c_H$, условие (1) будет выполнено, тогда как для рядов с распределениями, отличными от гауссова, может потребоваться более сложная, чем линейная, экстраполяция функции $f(x)$ при $x < c_H$. В любом случае заполнение по правилам (11)—(13) приводит к меньшим ошибкам в статистических выводах, чем при работе только по наблюдаемым отсчетам или заполнении пропусков прогнозом.

Можно несколько смягчить требования, предъявляемые к ряду y_t и цензуре c_t. Например, если отсчеты c_t известны во все моменты времени, то знание функции распределения $G(x)$ необязательно, так как ее можно оценить по выборке c_t. Необязательным является и отсутствие зависимости между рядами c_t и y_t. При наличии такой зависимости увеличится мерность функции регрессии $\eta_k(\cdot)$ за счет добавления аргументов $c_{t-1}, \ldots, c_{t-s}$, (возможно, она потеряет свойство линейности и т. д.). Стационарность рядов y_t и c_t также не является обязательным условием. Однако рассмотрение всех вариантов невозможно провести в общем виде без конкретизации смягченных требований к y_t и c_t.

Для примера рассмотрим задачу заполнения пропусков при полностью известной модели ряда y_t и стационарной независимой цензуре с неизвестной функцией распределения $G(x)$ и неизвестными отсчетами c_t. Используя соотношение (5), находим по данным $y_t \in Y_{\text{obs}}$ выборочную оценку $G(x)$:

$$\hat{G}(x_i) = \frac{k_i}{n[F(x_i) - F(x_{i-1})]}, \quad i = \overline{1,m},$$

где k_i введены в (6). Удовлетворяющее (1) правило заполнения пропуска в момент времени $t = k$ будет следующим. Опираясь на известную модель ряда y_t, можно моделировать значение y_k^* при фиксированных $y_t \in Y_{\text{obs}}$ и значение $c_k^* \in \hat{G}(x)$. Пропуск заполняется значением y_k^*, если $y_k^* < c_k^*$.

Нетрудно заметить, что для заполнения пропусков потребовалось фактически построить полную статистическую модель ряда y_t. Это связано с тем, что заполнение пропусков относится к задачам моделирования, требующим для

своего решения детальной информации о вероятностных свойствах ряда. Иногда [2] задачу заполнения пропусков ставят как задачу оценки или прогноза пропущенного значения. Критерием, оптимизирующим заполнение, в этом случае выступает минимизация отклонения заполняющего значения от истинного. При такой постановке задачи условие (1) не выполняется, что может привести к грубым ошибкам в статистических выводах при работе по заполненной выборке.

Предложенные алгоритмы заполнения случайно-цензурированных временных рядов требуют умеренной априорной информации и могут быть использованы на практике при статистическом анализе неполных данных.

СПИСОК ЛИТЕРАТУРЫ

1. **Моисеев С. Н.** Прогноз огибающей УКВ-сигнала, отраженного от слоя E_s // Геомагнетизм и аэрономия. 1996. **36**, № 4. С. 100.
2. **Литтл Р. Дж., Рубин Д. Б.** Статистический анализ данных с пропусками. М.: Финансы и статистика, 1991.
3. **Скрипник В. М., Назин А. Е., Приходько Ю. Г., Благовещенский Ю. Н.** Анализ надежности технических систем по цензурированным выборкам. М.: Радио и связь, 1988.
4. **Моисеев С. Н.** Прогноз флуктуаций огибающей акустического сигнала в мелком море // Акуст. журн. 1996. **42**, № 5. С. 730.
5. **Новицкий П. В., Зограф И. А.** Оценка погрешностей результатов измерений. Л.: Энергоатомиздат, 1985.
6. **Айвазян С. А., Енюков И. С., Мешалкин Л. Д.** Прикладная статистика: исследование зависимостей. М.: Финансы и статистика, 1985.
7. **Бокс Дж., Дженкинс Г.** Анализ временных рядов, прогноз и управление. М.: Мир, 1974. Вып. 1.
8. **Тихонов В. И.** Статистическая радиотехника. М.: Радио и связь, 1982.

Поступила в редакцию 19 декабря 1996 г.