**УДК** 519.862.6

***Ю.А. Бычков*[1]**

[1] *Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация*

# СОПОСТАВЛЕНИЕ ТОЧНОСТИ ДВУХ МЕТОДОВ ОЦЕНИВАНИЯ ПАРАМЕТРОВ ПРИ ЗАПОЛНЕНИИ ПРОПУСКОВ В ДАННЫХ ГАЗОДОБЫВАЮЩЕГО ПРЕДПРИЯТИЯ

***Аннотация.*** В статье рассматривается проблема заполнения пропусков в массивах данных. Для решения проблемы предлагается использовать алгоритм заполнения пропусков, основанный на инструментарии регрессионного анализа. Восстановление значений элементов, в которых есть пропуски, осуществляется путем построения нескольких линейных регрессионных моделей при помощи метода наименьших модулей и непрерывной формы метода максимальной согласованности. Построение линейных регрессионных моделей организовано с помощью специально разработанного программного обеспечения. Сделан вывод о высокой точности построенных моделей и эффективности используемого в работе алгоритма по заполнению пропусков.

***Ключевые слова:*** регрессионный анализ, метод наименьших модулей, непрерывная форма метода максимальной согласованности, пропуски.

***Yu.A. Bychkov*[1]**

[1] *Irkutsk State Transport University, Irkutsk, Russia*

# COMPARISON OF THE ACCURACY OF TWO METHODS OF ESTIMATION OF PARAMETERS WHEN FILLING IN THE GAPS IN THE DATA OF A GAS PRODUCING ENTERPRISE

***Abstract.*** The article deals with the problem of filling gaps in data arrays. To solve the problem, it is proposed to use a gap filling algorithm based on regression analysis tools. Restoring the values of elements that have gaps is carried out by building several linear regression models using the method of least modules and the continuous form of the method of maximum consistency. The construction of linear regression models is organized using specially developed software. The conclusion is made about the high accuracy of the constructed models and the efficiency of the gap filling algorithm used in the work.

***Keywords:*** regression analysis, least moduli method, continuous form of the maximum consistency method, gaps.

## Введение

Массивам данных, полученных экспериментальным путём, свойственно наличие пропусков и статистических «выбросов», которые могут возникать из-за широкого спектра причин. Типовыми примерами таких причин служат: человеческий фактор, выражающийся в ошибочных записях или трактовках полученных результатов экспериментов или социологических опросов; измерительные ошибки, связанные с неисправностью оборудования; форс-мажорные факторы различной природы.

Проблеме заполнения пропусков в данных посвящено большое количество научных работ. Так, в [1] предложены методы восстановления утраченных значений на основе создания «равномерной» и «неравномерной» временной сетки. В [2] оценивается возможность применения алгоритма MICE для заполнения пропусков в базах данных, содержащих сведения в области здравоохранения. В работе [3] приведён сравнительный анализ и охарактеризованы основные методы восстановления пропущенных значений в массивах данных. Интерес вызывает работа [4], в которой ставится задача по заполнению пропусков в социально-экономических данных на основе нечёткой формализации. В результате предложен метод заполнения пропусков на основе анализа и формализации исходных данных в терминах нечеткой логики взаимосвязи между параметрами аналогичного типа в других объектах сети. В статье [5] для заполнения пропусков во временных рядах экономических показателях применяется интерполяция, основанная на

кубических сплайнах. В качестве примера приводится заполнение пропусков в индексе делового оптимизма в отношении кредитного рынка Беларуси за 2012-2019 г.г. В [6] представлены краткие описания и сравнительный анализ эффективности методов рандомизированной обработки данных, применяемых в целях заполнения пропусков: Монте-Карло, Jackknife и Bootstrap. В [7] описан алгоритм непараметрической оценки кривой регрессии, повышающий точность задач идентификации дискретно-непрерывных многомерных процессов по массивам данных с пропусками. Для оценки качества работы алгоритма проведено два вычислительных эксперимента, по результатам которых сделан вывод об эффективности предлагаемого алгоритма.

Целью данной работы является исследование вопроса о целесообразности применения методов регрессионного моделирования для заполнения пропусков в массивах данных по относительно простому алгоритму.

**Методы оценивания параметров.** Заполнять пропущенные значения в данных будем с помощью метода наименьших модулей (МНМ) и непрерывной формы метода максимальной согласованности (НММС). Применение двух методов обусловлено необходимостью качественного сравнения полученных результатов моделирования.

Приведем краткое описание НММС, предложенного в [8] и применённого в работах [9-12] для решения конкретных практических задач. Рассмотрим линейное регрессионное уравнение (модель) вида:

$$y_k = \sum_{i=1}^{m} \alpha_i x_{ki} + \varepsilon_k, k = \overline{1,n}, \quad (1)$$

где y – зависимая, а $x_i$ – $i$-ая независимая переменные, $\alpha_i$ – $i$-ый подлежащий оцениванию параметр, $\varepsilon_k$ – ошибки аппроксимации, $k$ – номер наблюдения, $n$ – число наблюдений (длина выборки).

Представим линейное регрессионное уравнение (1) в векторной форме:

$$y = X\alpha + \varepsilon,$$

где $y = (y_1, \dots, y_n)^T$, $\alpha = (\alpha_1, \dots, \alpha_m)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$, X – $(n \times m)$ – матрица с компонентами $x_{ki}$.

Линейную модель (1) можно представить в виде:

$$y_k = \hat{y}_k + \varepsilon_k, \quad k = \overline{1,n},$$

где $y_k$ и $\hat{y}_k$ – соответственно фактические (наблюдаемые) и расчетные (вычисленные по модели) значения зависимой переменной y.

Допустим, после построения модели (1) для произвольных номеров наблюдений s и h изучаемой выполняется следующее неравенство [8]:

$$(y_s\text{-}y_h)(\hat{y}_s\text{-}\hat{y}_h) < 0.$$

Оно означает, что на паре номеров наблюдений (s, h) линейная модель (1) неудовлетворительно описывает исследуемый процесс, что не может быть компенсировано малостью величин $|\varepsilon_s|$, и $|\varepsilon_h|$. Подобное обстоятельство негативно влияет на результат исследования, особенно тогда, когда с помощью модели анализируются динамические процессы. Формализовать такие ситуации позволяет критерий согласованности поведения (далее – КСП). Переход от КСП к его непрерывной форме описан в [8]. Применение НКСП заключается в решении задачи оптимизации:

$$L=\sum_{k=1}^{n-1}\sum_{s=k+1}^{n} l_{ks} \rightarrow min, \quad (2)$$

где

$$l_{ks} = \begin{cases} |\hat{y}_k - \hat{y}_s|, (y_k - y_s)(\hat{y}_k - \hat{y}_s) < 0 \\ 0, \quad \text{в противном случае.} \end{cases}$$

Введем в рассмотрение числа $\omega_{ks}$, $k=\overline{1,n-1}$, $s = \overline{k+1,n}$ по правилу:

$$\omega_{ks} = \begin{cases} 1, & y_k - y_s > 0 \\ -1, & y_k - y_s < 0 \\ 0, & y_k - y_s = 0. \end{cases}$$

Задача (2) сводится к следующей задаче линейного программирования (ЛП):

$$r\sum_{k=1}^{n}(u_k + v_k) + (1-r)\sum_{k=1}^{n-1}\sum_{s=k+1}^{n} l_{ks} \rightarrow \min, \quad (3)$$

$$\sum_{i=1}^{m} \alpha_i x_{ki} + u_k - v_k = y_k,\ k = \overline{1,n}, \quad (4)$$

$$\omega_{ks} \sum_{i=1}^{m} \alpha_i (x_{ki} - x_{si}) + l_{ks} \geq 0, k=\overline{1,n-1},\ s = \overline{k+1,n}, \quad (5)$$

$$u_k \geq 0,\ v_k \geq 0, k = \overline{1,n}, l_{ks} \geq 0,\ k=\overline{1,n-1},\ s = \overline{k+1,n} \quad (6)$$

Здесь $r \in (0,1]$ – заранее выбранное число, устанавливающее сравнительный приоритет (компромисс) в целевой функции (3) между функцией потерь M= $\sum_{k=1}^{n}|\varepsilon_k| = \sum_{k=1}^{n}(u_k + v_k)$, соответствующей МНМ и НКСП, $\delta$ – малая положительная константа. Легко видеть, что при $r = 1$ задача ЛП (3) – (6) реализует МНМ, а при $r$, близком к 0, она позволяет определить оценки параметров модели (1), обеспечивающие максимальную согласованность поведения расчетных и фактических значений выходной переменной.

Как показано в [8] реализация МНМ состоит в решение следующей задачи ЛП:

$$M=\sum_{k=1}^{n}|\varepsilon_k| \rightarrow min.$$

**Алгоритм.** Заполнять пропуски в массивах данных будем на основе алгоритма предложенного в [13]. Кратко изложим его суть.

*Первый этап.* Множество номер наблюдений массива данных обозначим через K = {1, 2, …, n}. Исходную выборку (X, y) разобьем на две подвыборки: комплектную ($X_k$, $y_k$), k ∈ $K_1$ и некомплектную ($X_k$, $y_k$), k ∈ $K_2$, где $K_1$ – множество номеров комплектных наблюдений, $K_2$ – множество номеров наблюдений, содержащих пропуски, K = $K_1$ ∪ $K_2$, $K_1$ ∩ $K_2$ = ∅; $X_k$ - k-ая строка матрицы *X*.

*Второй этап.* Для элементов $x_{ki}$ содержащих пропуски строятся регрессионные уравнения (1) на совокупности значений индексного множества $K_1$.

*Третий этап.* На основании ранее построенных регрессионных уравнений рассчитываются значения пропущенных элементов, входящих в индексное множество $K_2$.

*Четвертый этап.* Оценивается точность восстановленных значений по отношению к истинным. Полученные результаты оценивают по показателю относительной погрешности.

В качестве информационной базы, для проведения исследования воспользуемся статистической информацией, представленной в таблице 1. Информация предоставлена региональной газодобывающей компанией.

Сепараторы С101 и С102б являются важными агрегатами установки подготовки газа и непосредственно участвуют в процессе её работы. Они применяются в качестве первой и второй ступеней разделения поступающего от газовых скважин водометанольного раствора, (ВМР) на неочищенный газ и капельную жидкость. В выборке представлено 21 наблюдение. Данное количество наблюдений обусловлено тем, что в оставшиеся дни месяца установка работала в резервном режиме.

**Таблица 1**

Статистические данные

| Дни | Объем добычи газа (общий) | Приход ВМР | Рабочее давление сепаратора С101 | Рабочее давление сепаратора С102б |
|---|---|---|---|---|
| | тыс. м3/сут | т | МПа | МПа |
| № | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
| 1 | 542,919 | 4,8548 | 12,04 | 3,72 |
| 2 | 439,391 | 1,2824 | 12,12 | 3,67 |
| 3 | 438,736 | 1,16332 | 12,67 | 3,68 |
| 4 | 437,829 | 1,16332 | 12,18 | 3,78 |
| 5 | 440,427 | 1,3282 | 12,38 | 3,73 |
| 6 | 440,844 | 1,42896 | 12,37 | 3,68 |
| 7 | 439,334 | 1,11752 | 12,34 | 3,74 |
| 8 | 441,89 | 1,1908 | 12,37 | 3,69 |
| 9 | 441,118 | 1,2824 | 12,38 | 3,74 |
| 10 | 442,114 | 1,1908 | 12,37 | 3,75 |
| 11 | 454,17 | 1,1908 | 12,05 | 3,76 |
| 12 | 454,537 | 1,0992 | 12,18 | 3,77 |

| Дни | Объем добычи газа (общий) | Приход ВМР | Рабочее давление сепаратора С101 | Рабочее давление сепаратора С102б |
|---|---|---|---|---|
| 13 | 455,273 | 1,374 | 12,13 | 3,79 |
| 14 | 452,982 | 1,0076 | 11,99 | 3,85 |
| 15 | 450,391 | 0,8244 | 11,89 | 3,86 |
| 16 | 457,937 | 1,5572 | 12,20 | 3,85 |
| 17 | 451,738 | 1,0992 | 12,23 | 3,8 |
| 18 | 446,666 | 1,0992 | 12,30 | 3,82 |
| 19 | 455,544 | 1,1908 | 12,24 | 3,87 |
| 20 | 454,769 | 1,2824 | 12,31 | 3,86 |
| 21 | 448,521 | 1,0076 | 12,37 | 3,77 |

Предположим, что в наблюдениях с номерами 7, 17 и 19 допущены пропуски, то есть отсутствуют значения $x_{7,1}$, $x_{17,3}$ и $x_{19,4}$. Следовательно, индексные множества $K_1$ и $K_2$ имеют следующий вид:

$$K_1 = \{1,2,3,4,5,6,8,9,10,11,12,13,14,15,16,18,20,21\},$$
$$K_2 = \{7,17,19\}.$$

Уравнение (1) для переменных, вошедших в индексное множества $K_2$, построим с помощью МНМ и НКСП на основе использования программного комплекса [14].

*а) Метод наименьших модулей*

$x_{k1} = 328,63 + 26,53x_{k2} – 18,5x_{k3} + 82,87x_{k4}$,
$x_{k3} = 18,53 – 0,01x_{k1} + 0,21x_{k2} – 0,47x_{k4}$,
$x_{k4} = 4,88 + 0,0019x_{k1} – 0,072x_{k2} – 0,15x_{k3}$.

*б) Непрерывная форма метода максимальной согласованности*

$x_{k1} = 316,08 + 24,75x_{k2} – 9,8x_{k3} + 58,64x_{k4}$,
$x_{k3} = 15,18 – 0,0066x_{k1} + 0,1x_{k2} – 0,013x_{k4}$,
$x_{k4} = 3,27 + 0,0019x_{k1} – 0,057x_{k2} – 0,024x_{k3}$.

Расчетные и истинные значения, а также относительная погрешность восстановленных значений отражены в таблице 2.

**Таблица 2**

Расчетные и истинные значения пропущенных элементов выборки

| Метод | Значения | Относительная погрешность | Средняя относительная погрешность |
|---|---|---|---|
| Метод наименьших модулей | $x^*_{1,7} = 439,92$ | 0,13 % | 1,06 % |
| | $x^*_{17,3} = 12,45$ | 1,86 % | |
| | $x^*_{19,4} = 3,82$ | 1,19 % | |
| Непрерывная форма метода максимальной согласованности | $x^*_{1,7} = 442,12$ | 0,63 % | 1,12 % |
| | $x^*_{17,3} = 12,26$ | 0,24 % | |
| | $x^*_{19,4} = 3,77$ | 2,48 % | |
| Истинные значения | $x_{1,7} = 439,33$ | - | - |
| | $x_{17,3} = 12,23$ | | |
| | $x_{19,4} = 3,87$ | | |

Анализ таблицы 2 позволяет установить, что восстановленные значения весьма близки к истинным. Следовательно, оба метода рационально использовать для решения аналогичных задач. Можно, в частности, в качестве пропущенных значений использовать полусуммы значений работы указанных методов.

**Заключение.** В работе рассмотрена возможность заполнения пропусков на основе использования МНМ и НКСП. Восстановленные числовые значения пропущенных элементов характеризуются высокой точностью.

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Грачев А.В. К восстановлению пропусков в экспериментальных данных // Вестник Нижегородского университета им. Н.И. Лобачевского. Серия: Радиофизика. – 2004. – № 1. – с. 15-23.

2. Аладышкина А.С., Лакшина В.В., Леонова Л.А., Максимов А.Г. Особенности работы с данными, характеризующими здоровье населения: заполнение пропусков в данных // Социальные аспекты здоровья населения. – 2020. – Т. 66. – № 1. – с. 12. – DOI 10.21045/2071-5021-2020-66-1-12.

3. Абраменкова И.В., Круглов В.В. Методы восстановления пропусков в массивах данных // Программные продукты и системы. – 2005. – № 2. – с. 4.

4. Аль-Катабери А.С., Щербаков М.В., Камаев В.А. Методика восстановления пропусков в социально-экономических данных на основе нечеткой формализации // Инженерный вестник Дона. – 2012. – № 1(19). – с. 336-339.

5. Власенко М.Н. Использование интерполяционных кубических сплайнов при восстановлении пропусков во временных рядах данных // Банковский вестник. – 2019. – № 7(672). – с. 31-36.

6. Плотников С.П., Блюмин С.Л. Восстановление пропусков в массивах данных рандомизационными моделями // Современные инструментальные системы, информационные технологии и инновации : сборник научных трудов XII-ой Международной научно-практической конференции: в 4-х томах, Курск, 19–20 марта 2015 года / Ответственный редактор: Горохов А.А.. – Курск: Закрытое акционерное общество «Университетская книга», 2015. – С. 312-314.

7. Осипов П.А., Осипова Я.С., Хоркуш А.В. [и др.]. Заполнение пропусков во входных и выходных данных с помощью алгоритма непараметрической идентификации // Сибирский журнал науки и технологий. – 2018. – Т. 19. – № 4. – с. 589-597. – DOI 10.31772/2587-6066-2018-19-4-589-597.

8. Носков С.И. Применение непрерывного критерия согласованности поведения при построении регрессионных моделей // Известия Тульского государственного университета. Технические науки. – 2021. – № 6. – С. 74-78. – DOI 10.24412/2071-6168-2021-6-74-78.

9. Носков С.И., Бычков Ю.А. Вычислительные эксперименты с непрерывной формой метода максимальной согласованности в регрессионном анализе // Вестник Воронежского государственного технического университета. – 2022. – Т. 18. – № 2. – С. 7-12. – DOI 10.36622/VSTU.2022.18.2.001.

10. Носков С.И., Бычков Ю.А. Модификация непрерывной формы метода максимальной согласованности при построении линейной регрессии // Известия Тульского государственного университета. Технические науки. – 2022. – № 5. – С. 88-94. – DOI 10.24412/2071-6168-2022-5-88-95.

11. Носков С.И., Бычков Ю.А. Применение метода максимальной согласованности для построения многофакторной регрессионной модели ввода жилья на региональном уровне // Инженерно-строительный вестник Прикаспия. – 2022. – № 2(40). – С. 141-145. – DOI 10.52684/2312-3702-2022-39-1-141-145.

12. Носков С.И., Бычков Ю.А. Построение регрессионной модели валового регионального продукта Ставропольского края на основе применения методов наименьших модулей и максимальной согласованности // Электронный сетевой политематический журнал "Научные труды КубГТУ". – 2022. – № 2. – С. 113-120.

13. Носков С.И., Бычков Ю.А. Простой способ заполнения пропусков в данных // Информационные технологии и проблемы математического моделирования сложных систем. – 2017. – № 19. – с. 130-136.

14. Свидетельство о государственной регистрации программы для ЭВМ № 2022618082 Российская Федерация. Программа оптимизации непрерывного критерия согласованности поведения при построении регрессионных моделей: № 2022617381 : заявл. 19.04.2022 :

опубл. 28.04.2022 / Носков С.И., Бычков Ю.А.; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский государственный университет путей сообщения».

**REFERENCES**

1. Grachev A.V. To restore gaps in experimental data // Bulletin of the Nizhny Novgorod University. N.I. Lobachevsky. Series: Radiophysics. - 2004. - No. 1. - p. 15-23.

2. Aladyshkina A.S., Lakshina V.V., Leonova L.A., Maksimov A.G. Peculiarities of working with data characterizing public health: filling gaps in data // Social aspects of public health. - 2020. - T. 66. - No. 1. - p. 12. - DOI 10.21045/2071-5021-2020-66-1-12.

3. Abramenkova I.V., Kruglov V.V. Methods for restoring gaps in data arrays // Software products and systems. - 2005. - No. 2. - p. four.

4. Al-Kataberi A.S., Shcherbakov M.V., Kamaev V.A. A technique for restoring omissions in socio-economic data based on fuzzy formalization // Inzhenerny Bulletin of the Don. - 2012. - No. 1 (19). - With. 336-339.

5. Vlasenko M.N. The use of interpolation cubic splines in restoring gaps in time series data. Bank Vestnik. - 2019. - No. 7 (672). - With. 31-36.

6. Plotnikov S.P., Blyumin S.L. Restoration of gaps in data arrays by randomization models // Modern instrumental systems, information technologies and innovations: a collection of scientific papers of the XII International Scientific and Practical Conference: in 4 volumes, Kursk, March 19–20, 2015 / Managing editor: Gorokhov A .A .. - Kursk: Closed Joint-Stock Company "University Book", 2015. - P. 312-314.

7. Osipov P.A., Osipova Ya.S., Horkush A.V. [and etc.]. Filling gaps in input and output data using the nonparametric identification algorithm // Siberian Journal of Science and Technology. - 2018. - T. 19. - No. 4. - p. 589-597. – DOI 10.31772/2587-6066-2018-19-4-589-597.

8. Noskov S.I. Application of a continuous criterion for the consistency of behavior in the construction of regression models. Izvestia of the Tula State University. Technical science. - 2021. - No. 6. - P. 74-78. – DOI 10.24412/2071-6168-2021-6-74-78.

9. Noskov S.I., Bychkov Yu.A. Computational experiments with the continuous form of the maximum consistency method in regression analysis // Bulletin of the Voronezh State Technical University. - 2022. - T. 18. - No. 2. - S. 7-12. – DOI 10.36622/VSTU.2022.18.2.001.

10. Noskov S.I., Bychkov Yu.A. Modification of the continuous form of the method of maximum consistency in the construction of linear regression // Izvestiya of the Tula State University. Technical science. - 2022. - No. 5. - P. 88-94. – DOI 10.24412/2071-6168-2022-5-88-95.

11. Noskov S.I., Bychkov Yu.A. Application of the maximum consistency method for constructing a multifactorial regression model for housing commissioning at the regional level // Engineering and Construction Bulletin of the Caspian Sea. - 2022. - No. 2 (40). - S. 141-145. – DOI 10.52684/2312-3702-2022-39-1-141-145.

12. Noskov S.I., Bychkov Yu.A. Building a regression model of the gross regional product of the Stavropol Territory based on the application of the least moduli and maximum consistency methods // Electronic network polythematic journal "Scientific Works of KubGTU". - 2022. - No. 2. - P. 113-120.

13. Noskov S.I., Bychkov Yu.A. A simple way to fill gaps in data // Information technologies and problems of mathematical modeling of complex systems. - 2017. - No. 19. - p. 130-136.

14. Certificate of state registration of the computer program No. 2022618082 Russian Federation. The program for optimizing the continuous criterion for the consistency of behavior in the construction of regression models: No. 2022617381 : Appl. 04/19/2022 : publ. 04/28/2022 / Noskov S.I., Bychkov Yu.A.; applicant Federal State Budgetary Educational Institution of Higher Education "Irkutsk State Transport University".

**Информация об авторах**

*Бычков Юрий Александрович* – аспирант кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск e-mail: bychkov_ya@internet.ru.

**Authors**

*Bychkov Yury Aleksandrovich* – post-graduate student of the department of Information systems and information protection, Irkutsk State Transport University, Irkutsk, e-mail: bychkov_ya@internet.ru.

**Для цитирования**

Бычков Ю.А. Сопоставление точности двух методов оценивания параметров при заполнении пропусков в данных газодобывающего предприятия // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2022. – №3(15). – С.7-13– DOI: 10.26731/2658-3704.2022.3(15).7-13 – Режим доступа: http://ismm-irgups.ru/toma/315-2022, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 15.10.2022)

**For citations**

Bychkov. Yu. A. Comparison of the accuracy of two methods of estimation of parameters when filling in the gaps in the data of a gas producing enterprise // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2022. No. 3(15). P. 7-13. DOI: 10.26731/2658-3704.2022.3(15).7-13 [Accessed 15/10/22]