УДК 519.68
Doi: 10.31772/2587-6066-2018-19-4-589-597

Для цитирования: Осипов П. А., Осипова Я. С., Хоркуш А. В., Вдовых П. Е., Верхотурова М. В. 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

For citation: Osipov P. A., Osipova Y. S., Khorkush A. V., Vdovykh P. E., Verkhoturova M. V. [Filling the gaps in the input and output data using the algorithm of nonparametric identification]. *Siberian Journal of Science and Technology*. 2018, Vol. 19, No. 4, P. 589–597 (In Russ.). Doi: 10.31772/2587-6066-2018-19-4-589-597

ЗАПОЛНЕНИЕ ПРОПУСКОВ ВО ВХОДНЫХ И ВЫХОДНЫХ ДАННЫХ С ПОМОЩЬЮ АЛГОРИТМА НЕПАРАМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ

П. А. Осипов*, Я. С. Осипова, А. В. Хоркуш, П. Е. Вдовых, М. В. Верхотурова

Сибирский федеральный университет, Институт космических и информационных технологий
Российская Федерация, 660074, г. Красноярск, ул. Академика Киренского, 26
*E-mail: uoo-ikit@mail.ru

Задача идентификации систем, т. е. определение структуры и параметров систем по наблюдениям, является одной из основных задач современной теории и техники автоматического управления. Точность решения задачи идентификации напрямую зависит от качества исходных данных (выборки наблюдений). Однако данные могут содержать в себе различные недостатки, в частности, пропуски.

Пробелы (пропуски) в данных возникают вследствие множества причин, таких как невозможность наблюдения, отсутствие необходимых инструментов и т. п. Самый простой метод работы с такими данными – исключение из таблицы показателя (столбец) или объекта (строки) с пробелом. При большом количестве пропусков в данных этот подход приводит к уменьшению точности модели из-за сокращения объема выборки. Важно отметить, что в описанном случае сложность решения задачи идентификации повышается, особенно когда плотность пропусков высока, их расположение нерегулярно, а данных недостаточно (крайне мало).

Целью работы является повышение точности решения задачи идентификации дискретно-непрерывных многомерных процессов по выборкам наблюдений с пропусками. Для достижения поставленной цели использовались методы математической статистики, анализа данных, математического моделирования.

Описан алгоритм непараметрической оценки кривой регрессии в дискретно-непрерывном процессе в задаче заполнения пропусков матрицы наблюдений. Также на основе этого алгоритма строится модель. Были проведены два вычислительных эксперимента. Первое исследование проведено в условиях наличия пропусков в выходной переменной матрицы наблюдений. Второй эксперимент проходил при наличии пробелов во входных переменных. Исследования проводились при различных объемах выборки. По итогам работы алгоритма при различных условиях приведены некоторые выводы.

Результаты работы могут быть полезны при создании систем управления многомерными дискретно-непрерывными процессами.

Ключевые слова: непараметрическая идентификация, оценка кривой регрессии, моделирование, анализ данных, пропуски в данных.

FILLING THE GAPS IN THE INPUT AND OUTPUT DATA USING THE ALGORITHM OF NONPARAMETRIC IDENTIFICATION

P. A. Osipov*, Y. S. Osipova, A. V. Khorkush, P. E. Vdovykh, M. V. Verkhoturova

Siberian Federal University, Institute of Space and Information Technology
26, Kirensky Str., Krasnoyarsk, 660074, Russian Federation
*E-mail: uoo-ikit@mail.ru

The task of identifying systems, that is, determining the structure and parameters of systems from observations, is one of the main tasks of a modern theory and technology of automatic control. The accuracy of solving the identification problem directly depends on the quality of the initial data (sample of observations). However, the data may contain various shortcomings, in particular, gaps.

Gaps in the data are due to a variety of reasons, such as inability to observe, lack of necessary tools, and so on. The easiest method of working with such data is to exclude from the table an indicator (column) or an object (line) with a space. With a large number of gaps in the data, this approach leads to a reduction in the accuracy of the model due to a reduction in the sample size. It is important to note that in the described case the complexity of solving

the identification problem increases, especially when the density of passes is high, their location is irregular, and the data is insufficient (very little).

The aim of the paper is to improve the accuracy of solving the problem of identifying discrete-continuous multidimensional processes from samples of observations with gaps. To achieve this goal, methods of mathematical statistics, data analysis, and mathematical modelings were used.

In the article the algorithm of a non-parametric estimation of the regression curve in a discrete-continuous process in the task of filling out the admissions of the observation matrix is described. Moreover, a model is built based on this algorithm. Two computational experiments were carried out. The first experiment was conducted in the presence of gaps in the output variable matrix of observations. The second experiment was conducted with gaps in the input variables. The experiments were conducted at different sample sizes. Based on the results of the algorithm under various conditions, conclusions are given.

The results of the work can be useful in creating control systems for multidimensional discrete-continuous processes.

Keywords: nonparametric identification, regression curve estimation, modeling, data analysis, data gaps.

Введение. В теории автоматического управления принципы построения системы управления разрабатывались на основе заданной модели. Со временем оказалось, что во многих случаях модель, выбранная при проектировании, значительно отличается от реального объекта, что существенно уменьшало эффективность разработанной системы. В связи с этим возникло новое направление в науке, связанное с построением модели на основании наблюдений, полученных в условиях функционирования объекта по его входным и выходным переменным. Это направление известно сегодня как идентификация систем. Теории и методам идентификации посвящено большое количество работ в отечественной и зарубежной литературе, и в этом направлении разработаны свои принципы, подходы и методы [1–6]. Эти подходы нашли широкое применение в различных областях науки и техники, в том числе в биологии, медицине, аэронавтике, экономике, промышленности.

Я. З. Цыпкин отмечает, что задача идентификации систем (определение структуры и параметров систем по наблюдениям) является одной из основных задач современной теории и техники автоматического управления. Эта задача возникает при изучении свойств и особенностей объектов с целью последующего управления ими либо при создании адаптивных систем, в которых на основе идентификации объекта вырабатываются оптимальные управляющие воздействия [5].

В книге Эйкхоффа [6] дано следующее определение: «Задача идентификации формулируется следующим образом: по результатам наблюдений над входными и выходными переменными системы должна быть построена оптимальная в некотором смысле модель, т. е. формализованное представление этой системы».

В исходных данных часто возникают пропуски. Пробелы (пропуски) в данных возникают вследствие множества причин, таких как невозможность наблюдения, отсутствие необходимых инструментов и т. п. Самый простой метод работы с такими данными – исключение из таблицы показателя (столбец) или объекта (строки) с пробелом. При большом количестве пропусков в данных этот подход приводит к уменьшению точности модели из-за сокращения объема выборки. Важно отметить, что в описанном случае сложность решения задачи идентификации повышается, особенно когда плотность пропусков высока, их расположение нерегулярно, а данных недостаточно (крайне мало).

В данной работе реализован один из алгоритмов заполнения пропусков в данных. На сегодняшний день разработано множество методов заполнения пропусков в данных. В работах [2–7] приводятся результаты работы этих методов в различных условиях. Методы заполнения пропусков реализованы в некоторых пакетах прикладных математических программ (например, SPSS Statistic). Задача оценки влияния применения этих методов на точность решения задачи идентификации является актуальной.

Постановка задачи. На рис. 1 представлена общая схема исследуемого процесса, принятая в теории идентификации.

Представленная схема состоит из двух блоков: «Объект», «Модель». На рис. 1 используются следующие обозначения: A – неизвестный оператор объекта; $u(t) = (u_1(t), u_2(t), ..., u_m(t)) \in \Omega(u) \subset R^m$ – векторное входное воздействие объекта размерностью m; $x(t) = (x_1(t), x_2(t), ..., x_n(t)) \in \Omega(x) \subset R^n$ – векторная выходная переменная объекта размерностью n; выполняется условие $m \geq n$; (t) – непрерывное время; Δt – дискретность контроля входных-выходных переменных процесса; $\xi(t)$ – векторная случайная помеха; блоки контроля переменных H^u, H^x подвержены воздействию случайных помех $h^u(t)$ и $h^x(t)$; u_t и x_t – измерения переменных $u(t)$ и $x(t)$ в дискретные моменты времени. Выборка измерений входных-выходных переменных процесса $\{u_i, x_i, i = \overline{1,s}\}$, где s – объем выборки. Измерения входных-выходных переменных объекта поступают на блок «Модель», где на основании заданного алгоритма находятся значения выхода модели x_{st}. Все случайные факторы, действующие в каналах измерения и на процесс, имеют нулевые математические ожидания и ограниченные дисперсии.

Рассматриваемый процесс относится к классу дискретно-непрерывных, т. е. по своей природе процесс является непрерывным, однако входные-выходные переменные процесса контролируются через дискретные моменты времени.

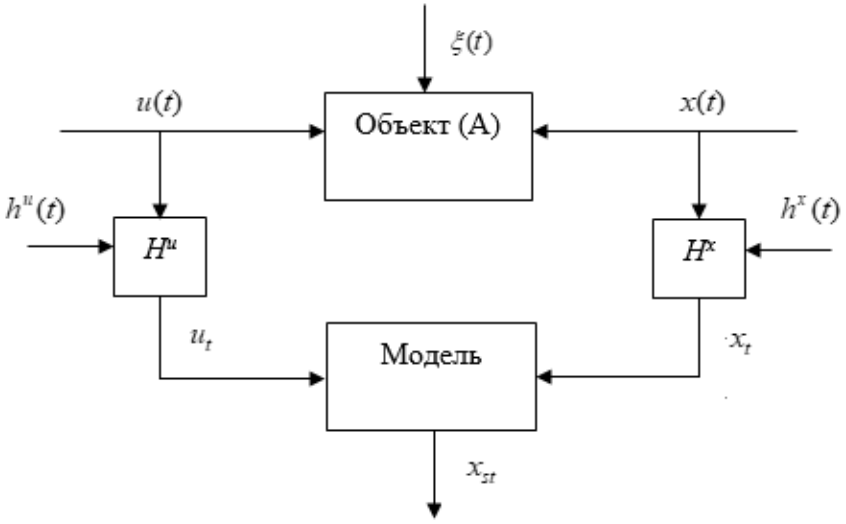

Рис. 1. Общая схема исследуемого процесса

Fig. 1. General scheme of the explored process

При построении модели с помощью методов идентификации используются экспериментальные данные. Ведется регистрация входных и выходных сигналов системы, и модель формируется в результате обработки соответствующих данных.

Формирование модели по наблюдениям включает:

1) данные;

2) множество моделей-кандидатов;

3) правило оценки степени соответствия испытываемой модели данным наблюдений.

Для построения модели важно иметь полные экспериментальные данные (наблюдения), но на практике это встречается крайне редко, т. е. такие данные имеют пропуски. Точность решения задачи идентификации зависит от качества данных. Если удалять строки из в матрицы наблюдений, которые имеют пропуски, очевидно, что число строк будет меньше, соответственно, точность модели снизится. Поэтому существует задача заполнения таких пробелов в матрице наблюдений.

Непараметрический алгоритм оценки кривой регрессии. Существуют параметрические и непараметрические методы идентификации. Методы параметрической идентификации требуют большого объема априорной информации. Часто возникают случаи, когда априорная информация об объекте очень бедна, поэтому структуру объекта нельзя определить с требуемой точностью. Непараметрические методы не ориентированы на указанные параметрические семейства, имеют более универсальную структуру и более широкую область применения [7; 8]. В условиях малой априорной информации целесообразно использовать методы непараметрической идентификации [9; 10].

С другой стороны, необходимо решать большое количество дополнительных задач: выбор структуры системы, задание класса моделей, оценивание степени и формы влияния входных переменных на выходные и др. [11]. Один из вариантов построения модели в условиях непараметрической неопределенности – это применение непараметрической оценки кривой регрессии. Вид такой оценки в многомерном случае имеет вид [12]

$$\hat{x_s}(u) = \frac{\sum_{i=1}^{s} x_i \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{C_s}\right)}{\sum_{i=1}^{s} \prod_{j=1}^{m} \Phi\left(\frac{u^j - u_i^j}{C_s}\right)}, \quad (1)$$

где $u = (u_1, u_2, ..., u_3)$ – m-мерный вектор входных воздействий объекта; x – выходная величина; $\Phi\left(C_s^{-1}(u^j - u_i^j)\right)$ – ядерная колоколообразная функция; C_s – коэффициент размытости ядра. Ядерная функция $\Phi(.)$ и коэффициент размытости ядра C_s удовлетворяют следующим условиям сходимости:

$$\begin{gathered} C_s > 0; \Rightarrow \Phi\left(C_s^{-1}(u^j - u_i^j)\right) < \infty; \\ \lim_{s\to\infty} C_s = 0; \Rightarrow C_s^{-1} \int_{\Omega(u)} \Phi\left(C_s^{-1}(u^j - u_i^j)\right) dx = 1; \\ \lim_{s\to\infty} s C_s^m = \infty; \Rightarrow \lim_{s\to\infty} C_s^{-1} \Phi\left(C_s^{-1}(u^j - u_i^j)\right) = \delta(u^j - u_i^j), \end{gathered} \quad (2)$$

где $\delta(u^j - u_i^j)$ – дельта-функция Дирака [13]. Ядерная функция имеет различные формы: треугольное ядро, параболическое ядро, кубическое ядро и др. Важно отметить, точность восстановления функции регрессии по наблюдениям с пропусками несущественно зависит от формы ядра и определяется практическими соображениями исследователя. В данных вычислительных экспериментах используется треугольное ядро, которое имеет вид

$$\Phi\left(\frac{x - x_i}{C_s}\right) = \begin{cases} 1 - \left|C_s^{-1}(x - x_i)\right|, \left|C_s^{-1}(x - x_i) \le 1\right| \\ 0, \left|C_s^{-1}(x - x_i) > 1\right| \end{cases}. \quad (3)$$

Для вычислительных экспериментов был выбран объект, описываемый следующей структурой:

$$x_i = 0,5 \cdot u_1 + 2 \cdot u_2^2 + u_3^{\frac{2}{3}}. \quad (4)$$

Следует отметить, что данный вид зависимости (4) известен только в рамках данных вычислительных экспериментов. Сразу отметим, что ошибка моделирования W считается по формуле

$$W = \frac{1}{s}\left(\sum_{i=1}^{s}\left|x_i - \hat{x_i}\right|\right). \quad (5)$$

Проведенные эксперименты по заполнению пропусков в данных разделены на три этапа.

Этап 1. На данном этапе имеем полностью заполненную матрицу наблюдений с входными переменными u_1, u_2, u_3 и выходом x_i. С помощью формулы (1) находим значения оценки $\hat{x_i}$ и настраиваем коэффициент размытости C_s, т. е. значение, при котором ошибка W минимальна. То есть параметр размытости C_s определяется путем решения задачи минимизации квадратичного показателя соответствия выхода объекта и выхода модели, основанного на методе «скользящего экзамена», когда в модели (1) исключается i-я переменная, предъявляемая для экзамена:

$$R(c_s) = \sum_{k=1}^{s}\left(x_k - x_s\left(u_k, c_s\right)\right)^2 = \min_{c_s}, k \neq i. \quad (6)$$

Важно отметить, что C_s был взят в промежутке от 0,1 до 5 с шагом 0,1.

Этап II. Добавляем пропуски по определенному правилу. Удаляем из матрицы наблюдений строчки с пробелами. Далее выполняем то же самое, что на первом этапе.

Этап III. На данном этапе восстанавливаем значения пропущенных данных в первом эксперименте по формуле (1), а во втором – по формуле

$$u_m = \frac{\sum_{\substack{i=1\\k\neq i}}^{s} u_{1i}\Phi\left(\frac{x_k - x_i}{C_s}\right)\prod_{\substack{j=1\\k\neq i}}^{m}\Phi\left(\frac{u_{jk} - u_{ji}}{C_s}\right)}{\sum_{\substack{i=1\\k\neq i}}^{s}\Phi\left(\frac{x_k - x_i}{C_s}\right)\prod_{\substack{j=1\\k\neq i}}^{m}\Phi\left(\frac{u_{jk} - u_{ji}}{C_s}\right)}. \quad (7)$$

Значения пробелов заполнены, матрица наблюдений снова становится полной. Далее выполняем то же самое, что на первом этапе.

Рассмотренные этапы были реализованы на выборках S = 300; 600; 900; 1200; 1500; 1800; 2100.

Далее рассмотрим результаты вычислительных экспериментов.

Первый вычислительный эксперимент. В рассматриваемом эксперименте пропуски находятся в выходных данных (во всех значениях x_i, кроме каждого третьего). Дискретность таких пропусков объясняется особенностью контроля выхода. Это означает, что одни переменные процесса измеряются в один промежуток времени, другие – в другой. На практике часто встречаются данные с различной дискретностью контроля выхода, например в процессах сжигания угля в котлоагрегате энергоблока, кислородно-конвертерной плавки стали.

Значения входных переменных сгенерированы случайным образом в промежутке от 0 до 3 с точностью 0,00001.

Этап I. Найдены значения оценки $\hat{x_i}$ при заполненной матрице наблюдений.

Этап II. Удалены строчки с пробелами из матрицы наблюдений. В результате выборка уменьшилась в 3 раза. Снова находим значения оценки $\hat{x_i}$.

Этап III. Восстановлены пропущенные значения в матрице наблюдений.

Результаты данного вычислительного эксперимента представлены в табл. 1.

После изучения табл. 1 следуют следующие выводы:

1. При увеличении выборки S уменьшается ошибка моделирования (5). Это говорит о том, что чем больше текущая информация, тем точнее модель.

2. Результаты 2 этапа во всех случаях хуже итогов 1 и 3 этапов. Очевидно, что выборка на 2 этапе меньше остальных, поэтому ошибка (5) больше ($W_{1s} < W_{2s}$, $W_{3s} < W_{2s}$; $C_{s1} < C_{2s}$, $C_{s3} < C_{s2}$).

3. Результаты на 3 этапе во всех случаях лучше итогов 1 этапа ($W_{3s} < W_{1s} < W_{2s}$). В большинстве случаев $C_{s1} = C_{s3}$, в остальных случаях – отличие в одну десятую, это говорит о правильности реализации алгоритма.

На рис. 2 представлен график зависимости ошибки W от объема выборки S.

Важно отметить, что с увеличением выборки уменьшается ошибка моделирования (5). Поэтому чем больше данных, тем точнее будет результат.

Вышеописанные выводы доказывают эффективность применения непараметрического метода для заполнения пропусков и построения модели при малой априорной информации.

Второй вычислительный эксперимент. В отличие от первого эксперимента, теперь пропуск находится во входных данных, а именно, в каждой 5-й строке в случайной переменной u_1 или u_2, или u_3.

Рассматриваемый эксперимент будет проведен с разными входными данными. В первом случае между входами есть некоторые зависимости:

$$u_2 = 0,5 \cdot u_1 + 0,1 \cdot g, \quad (8)$$

$$u_3 = 0,5 \cdot u_1 + 0,3 \cdot u_2 + 0,05 \cdot g, \quad (9)$$

где g – случайным образом сгенерированное число в промежутке от 0 до 3 с точностью 0,001.

Коэффициент корреляции на всех выборках составляет примерно 0,97. Во втором случае данные сгенерированы случайным образом в промежутке от 0 до 3.

Этап I. Найдены значения оценки $\hat{x_i}$ при заполненной матрице наблюдений.

Этап II. Удалены строчки с пробелами из матрицы наблюдений. Тем самым выборка сокращается на 20 %. Снова находим значения оценки $\hat{x_i}$.

Этап III. Восстановлены пропущенные значения в матрице наблюдений.

Результаты данного вычислительного эксперимента представлены в табл. 2.

Подведем некоторые выводы по табл. 2. В данном эксперименте от общего объема выборки пропусков не так много, поэтому процедура удаления строк с пробелами не дает существенного отличия от результатов 1-го этапа (в пределах пяти сотых). Интересно следующее: по результатам 3-го этапа ошибка (5) увеличивается даже в сравнении со 2-м этапом.

Таблица 1

Результаты вычислительного эксперимента

Структура модели: $x_1 = a_1u_1 + a_2u_2^2 + a_3u_3^{\frac{2}{3}}$						
S	1 этап		2 этап		3 этап	
	W_1	C_{s1}	W_2	C_{s2}	W_3	C_{s3}
300	0,5249346	0,6	0,86835486	0,8	0,34579805	0,6
600	0,40851948	0,5	0,6613773	0,6	0,2906845	0,5
900	0,32048658	0,4	0,52412224	0,6	0,25944132	0,5
1200	0,27469125	0,4	0,43487462	0,5	0,20380524	0,4
1500	0,25636882	0,4	0,40102375	0,5	0,19891156	0,4
1800	0,2437991	0,4	0,37384215	0,5	0,19284262	0,4
2100	0,22616474	0,3	0,3535294	0,5	0,18599847	0,4

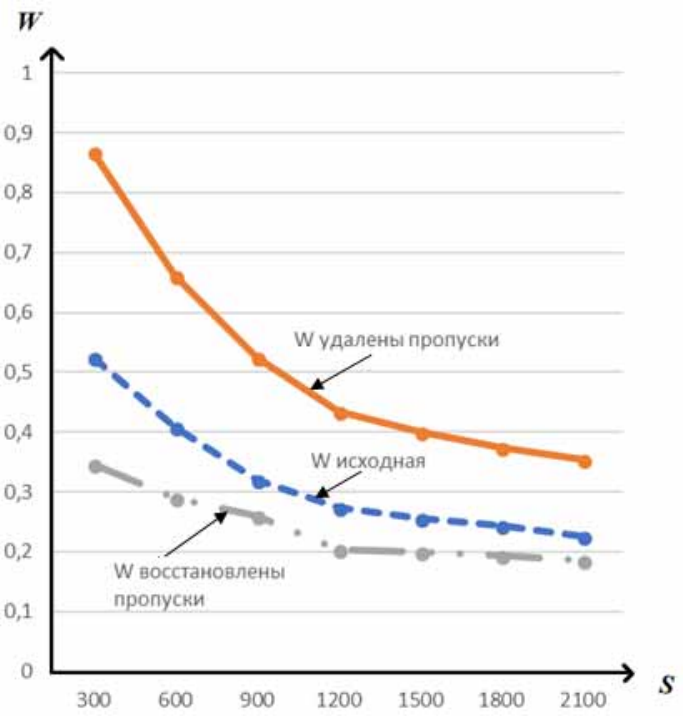

Рис. 2. График зависимости относительной ошибки (5) от объема выборки на разных этапах экспериментов

Fig. 2. The graph of dependence of the relative error (5) on the sample size at different stages of the experiments

Таблица 2

Результаты вычислительного эксперимента с зависимыми между собой входами с пропусками в каждой 5-й строке матрицы наблюдений

Структура модели: $x_1 = a_1u_1 + a_2u_2^2 + a_3u_3^{\frac{2}{3}}$						
S	1 этап		2 этап		3 этап	
	W_1	C_{s1}	W_2	C_{s2}	W_3	C_{s3}
300	0,10365894	0,2	0,107344694	0,2	0,17787787	0,3
600	0,09081579	0,2	0,093733266	0,2	0,1875585	0,3
900	0,045822605	0,1	0,04874478	0,1	0,12278164	0,2
1200	0,03565259	0,1	0,040589087	0,1	0,12119869	0,2

Таблица 3

Результаты вычислительного эксперимента 3-го и 4-го этапа с пропуском во входных переменных в каждой пятой строке

Структура модели: $x_1 = a_1u_1 + a_2u_2^2 + a_3u_3^{\frac{2}{3}}$				
S	3 этап		4 этап	
	W_1	C_{s1}	W_2	C_{s2}
300	0,17787787	0,3	1,3138921	0,9
600	0,1875585	0,3	1,2184308	0,9
900	0,12278164	0,2	1,1859775	0,85
1200	0,12119869	0,2	1,0735321	0,8

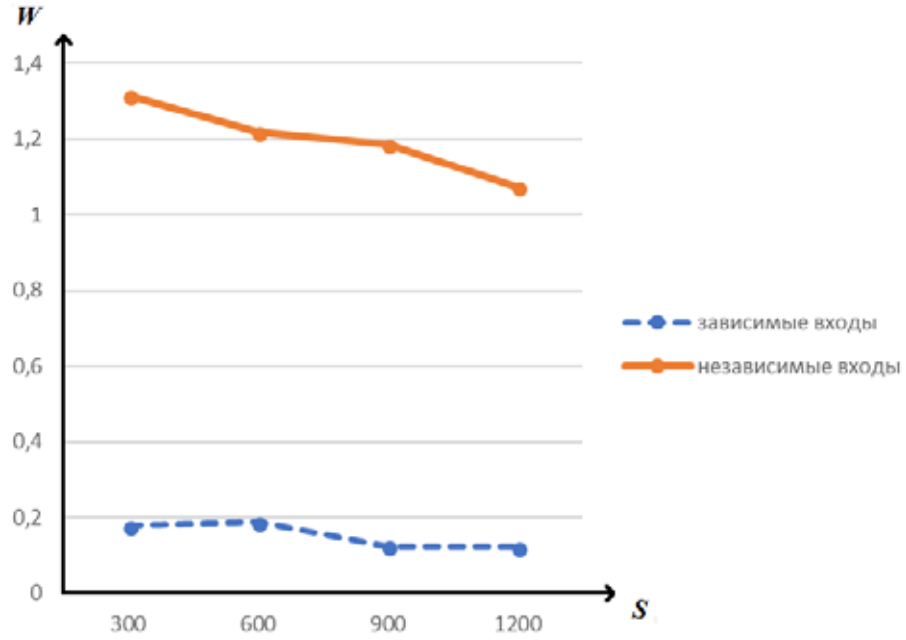

Рис. 3. График зависимости W от S в модели (4) при наличии пропуска в каждой пятой строке матрицы наблюдений на входе

Fig. 3. The graph of the dependence of W on S in the model (4) with the presence of a gap in each 5 line of the observation matrix at the input

Также был проведен этот же эксперимент, но с двумя входными переменными. Картина сохранилась, по результатам 3-го этапа ошибка (5) увеличивается даже в сравнении со 2-м этапом.

Данные результаты говорят о том, что эффективность применения этого алгоритма к данным, содержащим пропуски по входным переменным, значительно ниже, чем к данным с пропусками по выходам.

Далее эксперимент был повторен с независимыми между собой входными данными. Пусть это будет 4-й этап. Результаты 3-го и 4-го этапа представлены в табл. 3.

На рис. 3 представлен график зависимости W от S в модели (4), визуализирующий данные из табл. 3.

По вышеописанным данным следует вывод, что при независимых между собой входах результаты в несколько раз хуже, чем при зависимых. Это говорит о том, что использование непараметрического алгоритма для восстановления пропусков на входах при независимых между собой значениях в матрице наблюдений нецелесообразно использовать (или использовать, но в задачах, не требующих большой точности). Предполагается, что в строках и столбцах имеется избыточность, т. е. между свойствами могут быть зависимости, а объекты могут быть похожи между собой. Если избыточность не наблюдается, то все строки и столбцы имеют одинаковый вес при прогнозировании и смысл локальности алгоритма теряется, что и происходит при независимых данных [14].

Далее проведем исследование, аналогичное прошлому эксперименту. Но теперь будет пропуск в каждой 3-й строке в случайной переменной u_1 или u_2, или u_3.

Результаты проведенного исследования представлены в табл. 4.

Результат на 3-м этапе хуже, чем на первом и втором (почти всегда), но разрыв между результатами относительно небольшой.

На рис. 4 представлен график зависимости ошибки (5) от объема выборки текущего и прошлого эксперимента с зависимыми между собой данными в матрице наблюдений для модели (4).

В данном эксперименте по сравнению с прошлым пробелов в матрице наблюдений больше. По графикам очевидно, что ошибка (5) меньше в текущем эксперименте, т. е. получается так, что при малой априорной информации непараметрические алгоритмы имеют высокую эффективность.

Далее эксперимент был повторен с независимыми между собой входными данными. Пусть это будет 4-й этап. Результаты 3-го и 4-го этапа представлены в табл. 5.

Визуализация результатов табл. 5 представлена на рис. 5.

Таблица 4

Результаты вычислительного эксперимента с зависимыми между собой входами с пропусками в каждой 3-й строке матрицы наблюдений

Структура модели: $x_1 = a_1u_1 + a_2u_2^2 + a_3u_3^{\frac{2}{3}}$						
S	1 этап		2 этап		3 этап	
	W_1	C_{s1}	W_2	C_{s2}	W_3	C_{s3}
300	0,10365894	0,2	0,11760777	0,2	0,10815763	0,2
600	0,09081579	0,2	0,09333572	0,2	0,09623865	0,2
900	0,045822605	0,1	0,056772906	0,1	0,08182337	0,2
1200	0,03565259	0,1	0,047689456	0,1	0,07308948	0,2

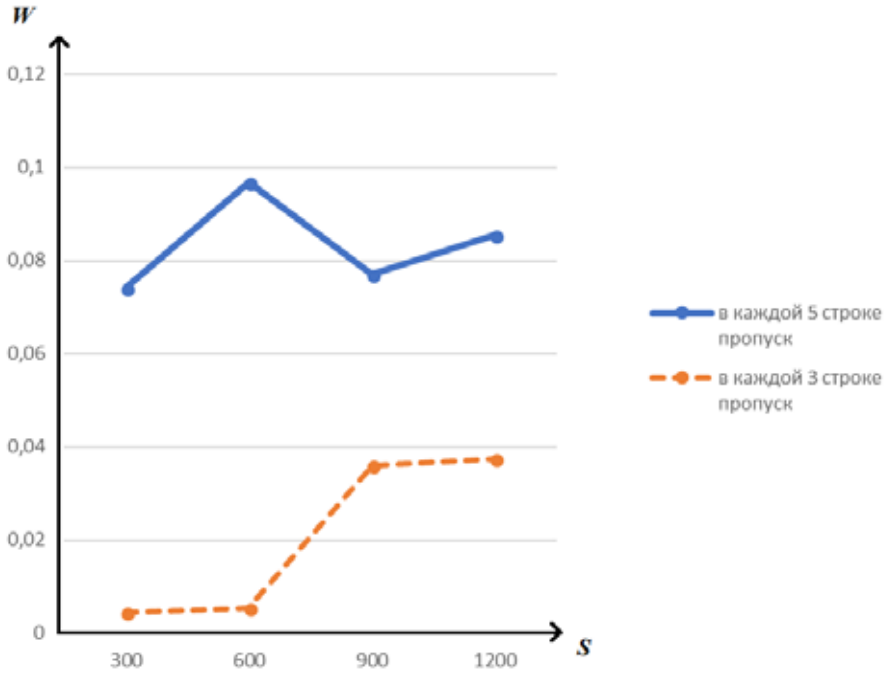

Рис. 4. График зависимости W от S проведенных экспериментов в модели (4)

Fig. 4. The graph of the dependence of W on S of the experiments carried out in model (4)

Таблица 5

Результаты вычислительного эксперимента 3-го и 4-го этапа с пропуском во входных переменных в каждой третьей строке

Структура модели: $x_1 = a_1u_1 + a_2u_2^2 + a_3u_3^{\frac{2}{3}}$				
S	3 этап		4 этап	
	W_1	C_{s1}	W_2	C_{s2}
300	0,10815763	0,2	0,6380898	0,9
600	0,09623865	0,2	0,4356285	0,6
900	0,08182337	0,2	0,3983642	0,6
1200	0,07308948	0,2	0,3711715	0,6

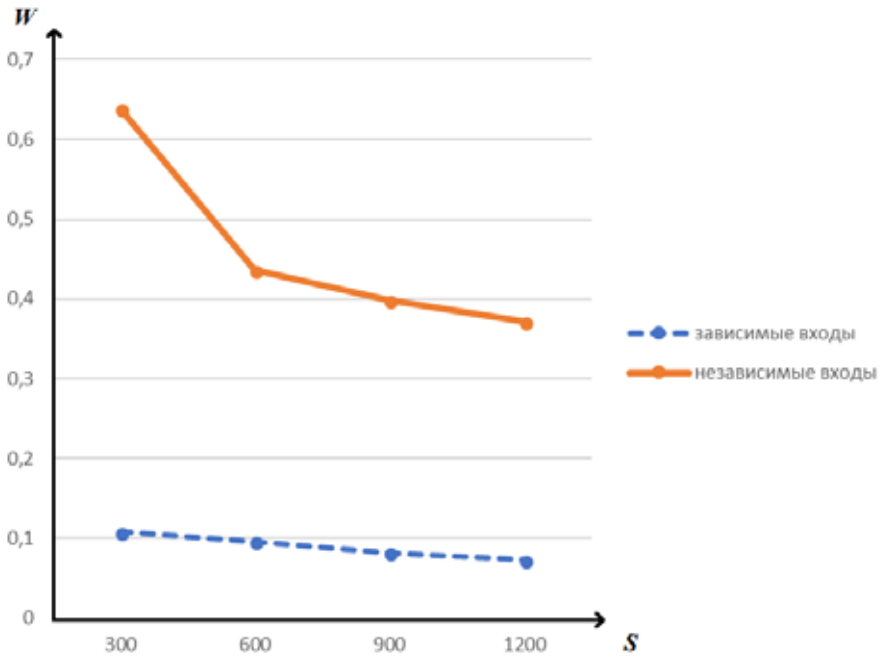

Рис. 5. График зависимости ошибки (5) от объема выборки в модели (4)

Fig. 5. The graph of the dependence of the error (5) on the sample size in the model (4)

Анализируя графики и табл. 5, еще раз можно убедиться в том, что при независимых между собой входах результаты в несколько раз хуже, чем при зависимых. Использование непараметрического алгоритма для восстановления пропусков при случайных входных значениях в матрице наблюдений нецелесообразно.

Заключение. Вышеописанные выводы доказывают эффективность применения непараметрического алгоритма для заполнения пропусков и построения модели при малой априорной информации. Ошибка моделирования (5) по заполненной матрице наблюдений с помощью рассматриваемого алгоритма оказалась меньше, чем по исходной.

Эффективность применения непараметрического алгоритма к данным, содержащим пропуски по входным переменным, значительно ниже, чем к данным с пропусками по выходам. Также важно отметить, что при зависимых входных данных результат работы алгоритма будет намного точнее, что описано в гипотезе избыточности [15].

В дальнейшем планируется исследование Zet-алгоритма, а именно, его применение в задаче заполнения пропусков в данных. Также будет проведено сравнение результатов работы Zet-алгоритма с алгоритмом непараметрической оценки кривой регрессии.

Библиографические ссылки

1. Карлов И. А. Методы восстановления пропущенных значений с использованием инструментария DataMining // Вестник СибГАУ. 2011. № 7 (40). С. 29–33.

2. Льюнг Л. Идентификация систем. М. : Наука, 1991. 423 с.

3. Райбман Н. С. Что такое идентификация. М. : Наука, 1970. 119 с.

4. Цыпкин Я. З. Адаптация и обучение в автоматических системах. М. : Наука, 1968. 400 с.

5. Эйкхофф П. Основы идентификации систем управления. М. : Мир, 1975. 681 с.

6. Кесман К. Дж. Идентификация системы. Введение. Лондон : Спрингер, 2011. 351 с.

7. Рубан А. И. Методы анализа данных. Красноярск : ИПЦ КГТУ, 2004. 319 с.

8. Шуленин В. П. Математическая статистика : учебник. Ч. 2. Непараметрическая статистика. Томск : Изд-во НТЛ, 2012. 388 с.

9. Корнеева А. А., Медведев А. В. К анализу данных в задаче идентификации // Кибернетика и высокие технологии XXI века : тр. XIII Междунар. науч.-техн. конф. 2012. Т. 1. С. 52–62.

10. Семенов А. Д., Артамонов Д. В., Брюхачев А. В. Идентификация объектов управления. Пенза : Изд-во Пенз. гос. ун-та, 2003. 211 с.

11. Медведев А. В. Анализ данных в задаче идентификации // Компьютерный анализ данных моделирования. 1995. Т. 2. С. 201–206.

12. Хардле В. Прикладная непараметрическая регрессия. М. : Мир, 1993. 349 с.

13. Надарая Э. А. Непараметрическое оценивание плотности вероятностей и кривой регрессии. Тбилиси : Изд-во Тбилис. ун-та, 1983. 194 с.

14. Гассер Т. Ядровая оценка функции регрессии. Гейдельберг : Спрингер, 1979. С. 23–68.

15. Загоруйко Н. Г. Методы распознавания и их применение. М. : Советское радио, 1972.

References

1. Karlov I. A. [Methods for restoring missing values using the DataMining toolkit]. *Vestnik SibGAU*. 2011, Vol. 161, No. 7 (40), P. 29–33 (In Russ.).

2. L'yung L. *Identifikatsiya sistem* [Identification of systems]. Moscow, Nauka Publ., 1991, 423 p.

3. Raybman N. S. *Chto takoe identifikatsiya* [What is identification]. Moscow, Nauka Publ., 1970, 119 p.

4. Tsypkin Ya. Z. *Adaptatsiya i obuchenie v avtomaticheskikh sistemakh* [Adaptation and training in automatic systems]. Moscow, Nauka Publ., 1968, 400 p.

5. Eykkhoff P. *Osnovy identifikatsii sistem upravleniya* [Basics of Identification of Management Systems] Moscow, Mir Publ., 1975, 681 p.

6. Keesman Karel J. *Sistema identifikatsii. Vvedenie* [System identification. An introduction]. London, Springer, 2011, 351 p.

7. Ruban A. I. *Metody analiza dannykh* [Methods of data analysis: a tutorial]. Krasnoyarsk, IPTs KGTU Publ., 2004, 319 p.

8. Shulenin V. P. *Matematicheskaya statistika. Ch. 2. Neparametricheskaya statistika* [Math statistics. Part 2. Nonparametric statistics]. Tomsk, NTL Publ., 2012, 388 p.

9. Korneyeva A. A., Medvedev A. V. [To the analysis of data in the identification problem] *Kibernetika i vysokie tekhnologii XXI veka: trudy XIII mezhdunarodnoy nauchno-tekhnicheskoy konferentsii* [Cybernetics and high technologies of the XXI century: proceedings of the XIII international scientific and technical conference]. Voronezh, 2012, P. 52–62 (In Russ.).

10. Semenov A. D., Artamonov D. V., Bryukhachev A. V. *Identifikatsiya ob''ektov upravleniya* [Identification of management objects: a tutorial]. Penza, publishing house of the Penza state university, 2003, 211 p.

11. Medvedev A. V. [Analysis of data in the identification problem]. *Komp'yuternyy analiz dannykh modelirovaniya*. 1995, Vol. 2, P. 201–206 (In Russ.).

12. Khardle V. *Prikladnaya neparametricheskaya regressiya* [Applied nonparametric regression]. Moscow, Mir Publ., 1993, 349 p.

13. Nadaraya E. A. *Neparametricheskoe otsenivanie plotnosti veroyatnostey i krivoy regressii* [Nonparametric estimation of probability density and regression curve]. Tbilisi, Izdatel'stvo Tbilisskogo universiteta Publ., 1983, 194 p.

14. Gasser T. *Yadrovaya otsenka funktsii regressii* [Kernel estimation of regression function]. Heidelberg, Springer, 1979, P. 23–68.

15. Zagoruyko N. G. *Metody raspoznavaniya i ikh primenenie* [Methods of recognition and their application]. Moscow, Sovetskoe Radio Publ., 1972.