Инженерный вестник Дона, №6 (2023)
ivdon.ru/ru/magazine/archive/n6y2023/8466

Импутация данных методами статистического моделирования

О.Н. Яркова

Санкт-Петербургский государственный архитектурно-строительный университет

Аннотация: Одной из задач предварительной обработки данных является задача устранения пропусков в данных, т.е. задача импутирования. В работе предложены алгоритмы заполнения пропусков в данных на основе метода статистического имитационного моделирования. Предлагаемые алгоритмы заполнения пропусков включают этапы кластеризации данных по набору признаков, классификации объекта с пропуском, построения функции распределения для признака, имеющего пропуски по каждому кластеру, восстановления пропущенных значений методом обратной функции. Проведены вычислительные эксперименты на основе статистических данных социально-экономических показателей по субъектам РФ за 2022 год. Проведен анализ свойств предлагаемых алгоритмов импутирования в сравнении с известными методами. Показана эффективность предлагаемых алгоритмов.
Ключевые слова: алгоритм импутации, пропуски в данных, статистическое моделирование, метод обратной функции, имитационное моделирование данных.

Введение

В современном мире практически ни одна сфера деятельности не обходится без решения задач исследования и анализа данных. Данные накапливаются в потоковом режиме, анализируются «на лету», создаются витрины данных, позволяющие аналитикам в режиме реального времени исследовать данные, строить модели, визуализировать информацию, готовить данные для лиц, принимающих решения. В таких условиях особенно актуальны задачи предварительной обработки, которые должны быть решены быстро и гарантировать качество подготовленных к анализу данных. Одной из задач предварительной обработки данных является задача устранения пропусков, т.е. задача импутирования. Способы решения задачи импутирования зависят от механизма формирования пропусков: MCAR (Missing Completely At Random) — механизм формирования пропусков, при котором вероятность пропуска для каждой записи набора одинакова; MAR (Missing At Random) —данные пропущены не случайно, а ввиду некоторых закономерностей; MNAR (Missing Not At Random) — механизм

формирования пропусков, при котором данные отсутствуют в зависимости от неизвестных факторов [1, 2]. Вопросам разработки и исследования методов предварительной обработки данных посвящено множество работ. Так, в работе Сидоровой Д.Н. [3] отмечается важность задачи предварительной обработки данных и показаны этапы подготовки данных для использования алгоритмов кластеризации на примере журналов информационной безопасности. В научных работах, к примеру, Абраменковой И.В. [4], Зингаевой И.К. [5], Пимонова А.Г. [6] и др., описывается и анализируется множество методов импутирования и инструментария для их реализации. Аналитикам доступны простые методы импутирования: заполнение средними/медианой, Hot Deck (ближайшего соседа), регрессионное моделирование и др., а также сложные методы: Zet и его модификации [7], Бартлета, множественное импутирование [8], EM-оценивание [5], Resampling и др. [5, 9], подходы на основе методов машинного обучения [10] и нечеткой логики [11], гибридные методы [2]. Простые методы достаточно просто реализуются, могут быть автоматизированы, позволяют обрабатывать потоковые данные, но обладают рядом недостатков, к примеру, метод заполнения средним, модой, нулями искажают статистические свойства выборки - закон распределения. Аналогично ведут себя методы типа LOCF (Last observation carried forward) — повторение результата последнего наблюдения. Метод Complete-case Analysis (удаление данных с пропусками), который используется во многих инструментальных средствах обработки данных в качестве метода по умолчанию, приводит к тому, что, в случае MAR и MNAR используется не вся доступная информация, средние квадратические отклонения возрастают, полученные результаты становятся менее репрезентативными. Метод Available-case analysis может приводить к некорректным значениям статистических показателей, таких, как, например, коэффициент корреляции. Кроме того, исключение объектов наблюдения в

принципе может быть неприемлемо. В работе Скрипкиной Т.Б. [12] предлагается для данных муниципальной статистики метод k-ближайших соседей. Однако, методы по типу ближайшего соседа, Hot Deck, практически не применимы при большом числе пропусков и требуют исследований о характере связи между переменными в каждом конкретном исследовании, что делает их мало приемлемыми для потоковой обработки данных. То же касается и сложных методов импутирования. Так, например, Маниковский А.С. и Мухопад А.Ю. [13] исследовали методы восстановления пропущенных значений во временных рядах с помощью линейного сплайна и методов прогнозирования, ими была показана эффективность методов построения двустороннего прогноза временного ряда на основе нейросетей. В работе Аль-Катабери А.С. [11] предлагается подход импутирования на основе нечеткой логики, в работе Фоминой Е.Е. [10] метод на основе нейросетей, но при этом отмечаются значительные временные затраты на его реализацию. Таким образом, несмотря на большое количество исследования в указанной сфере, задача разработки и исследования методов импутирования остается актуальной и недостаточно проработанной.

Цель и задачи исследования

Цель исследования: разработка и исследование свойств метода импутирования на основе метода статистического имитационного моделирования.

Задачи исследования:

– разработка вычислительно-простого алгоритма, позволяющего решать задачу импутирования в автоматизированном режиме;

– исследование свойств предложенного алгоритма импутирования путем проведения серии вычислительных экспериментов на реальном наборе данных;

– разработка рекомендаций по применению алгоритма импутирования, основанного на использовании метода статистического моделирования.

Методы исследования

Рассмотрим задачу: пусть необходимо заполнить пропуски в ряду данных, обладающем свойством стационарности, при условии, что не доступна информации по другим показателям, связанным с исследуемым. В этом случае предлагается для устранения пропуска использовать метод статистического имитационного моделирования обратной функции, где функция распределения строится на основе имеющихся статистических данных, в том числе, распределенных по произвольному закону распределения, не совпадающему со стандартными распределениями. Показатель может быть представлен как непрерывной, так и дискретной случайной величиной.

Алгоритм А

1. Дан ряд $X=\{x_i\}$, $i=1,2..n$ имеющий пропуски в данных. Пусть K – множество индексов, значения которых характеризуют пропущенные данные в ряду X.

2. Удаляем из ряда пропущенные значения $X^* = \{x_i\} : i \notin K$.

3. По ряду $X^* = \{x_i\} : i \notin K$ строится вариационный ряд (дискретный / непрерывный) относительных частот и аппроксимируется функция распределения $F_\xi = P(\xi < X^*)$, к примеру линейным сплайном.

4. Далее в цикле по i от $k+1$ до n:

4.1 С помощью датчика случайных чисел генерируется число p равномерно распределенное на отрезке [0,1];

4.2 Численно решается уравнение $p = F_\xi(x_i)$ относительно x_i, т.е. $x_i = F_\xi^{-1}(p)$;

Конец цикла по i.

Конец алгоритма.

Для генерации случайных чисел, равномерно распределенных на отрезке [0,1] можно воспользоваться алгоритмом Вихрь Мерсена [14]. Генерируемая последовательность по алгоритму Вихрь Мерсена обладает хорошими статистическими свойствами и подходит для применения в методах имитационного моделирования.

Алгоритм A позволяет заполнять пропущенные значения в рядах данных, не изменяя свойств имеющейся последовательности. Будем использовать его, как базовый для решения более сложных вопросов импутирования.

Рассмотрим следующую постановку задачи: пусть необходимо заполнить пропуски в ряду данных для некоторого показателя, при условии что доступна информация по другим показателям связанным с исследуемым, т.е. данные заданы в виде $X=\{x_{i,j}\}$, где $i=1,2..n$ – номер объекта в таблице данных; $j=1,2..m$ – номер признака; $x_{i,j}$ – значение признака j для i-го объекта наблюдения.

Пусть в значениях одного из признаков, для простоты первого $x_{1,*}$, имеются пропущенные данные. Для решения задачи импутирования в этом случае предлагаются следующие алгоритмы.

Алгоритм B.1

1. Пропуски заполняются средними значениями: пусть K – множество номеров объектов, содержащих пропущенные данные в признаке 1: $K=\{k_l\}: x_{k_l,1}=0$, тогда положим $\forall k \in K \;\; x_{k,1}=M[X_1^*]$, $X_1^*=\{x_{j,1}\}_{j\notin K}$.

2. На основе всех представленных в таблице данных, полученных после выполнения этапа 1, $X=\{x_{i,j}\}$, $i=1,2..n$; $j=1,2..m$ осуществляется кластеризация объектов Q_i, $i=1,2..n$. Допустим, на этом этапе получены классы V_l, $l=1,2..v$.

3. Уточняются значения методом обратной функции, с помощью алгоритма A, в котором для заполнения элемента $x_{k,1} : k \in K$ строится функция распределения F_{Xl*}=P$(\xi<X_l)$ (дискретная / непрерывная) на основе значений $x_{j,1} : j \notin K,\ j,k \in V_i$ исследуемого показателя для класса V_l в который попал объект Q_k со значением признака с пропуском $x_{k,1}$.

Алгоритм В.2

Данные заданы в виде $X=\{x_{i,j}\}$, где i=1,2..n – номер объекта в таблице данных; j=1,2..m – номер признака; $x_{i,j}$ – значение признака j для i-го объекта наблюдения. K – множество номеров объектов, содержащих пропущенные данные в признаке 1: $K = \{k_l\} : x_{k_l,1} = 0$,

1. На основе таблицы данных $X = \{x_{i,j}\} : i \notin K, j = 1,2..m$ т.е. по объектам без пропусков осуществляется кластеризация объектов $Q_i : i \notin K$ по всем признакам таблицы данных.

2. Классифицируются объекты с пропусками Q_i, $i \in K$ по полученным на этапе 1 классам V_l, l=1,2..v методом k ближайших соседей по признакам не содержащим пропуски $X = \{x_{i,j}\} : i \in K, j = 2,3..m$.

4. Заполняются пропущенные значения методом обратной функции, с помощью алгоритма A, в котором для заполнения элемента $x_{1,k} : k \in K$ строится функция распределения F_{Xl*}=P$(\xi<X_l)$ (дискретная / непрерывная) на основе значений $x_{j,1} : j \notin K,\ j,k \in V_i$ исследуемого показателя для класса V_l в который попал объект Q_k со значением признака с пропуском $x_{k,1}$.

Предлагаемые алгоритмы можно использовать в случае пропусков типа MCAR. При наличии качественных признаков необходимо использовать алгоритмы кластеризации, допускающие качественные признаки в данных.

Вычислительные эксперименты

Исследование свойств предлагаемых алгоритмов будем проводить на примере данных статистического ежегодника за 2022 год [15].

В анализе участвуют следующие показатели по субъектам РФ (85 объектов): $X1$ - численность населения, тыс. человек; $X2$ - уровень занятости, процентов; $X3$ - среднедушевые денежные доходы населения, тыс. руб.; $X4$ - реальные денежные доходы населения, в процентах к предыдущему году; $X5$ - среднемесячная номинальная начисленная заработная плата работников организаций, тыс. руб.; $X6$ - доля численности населения с денежными доходами ниже границы бедности; $X7$- число предприятий и организаций: сельское, лесное хозяйство, охота, рыболовство, рыбоводство; $X8$ - число предприятий и организаций: добыча полезных ископаемых; $X9$ - число предприятий и организаций: обрабатывающие производства; $X10$ - число предприятий и организаций: строительство; $X11$ - число предприятий и организаций: торговля оптовая и розничная ремонт автотранспортных средств и мотоциклов; $X13$ - удельный вес убыточных организаций в общем числе организаций, процентов; $X14$ – категориальный показатель, характеризующий принадлежность к округу (1 - Центральный федеральный округ; 2 - Северо-Западный федеральный округ; 3 - Южный федеральный округ; 4 - Северо-Кавказский федеральный округ; 5 - Приволжский федеральный округ; 6 - Уральский федеральный округ; 7 - Сибирский федеральный округ; 8 - Дальневосточный федеральный округ).

Анализ позволил выявить наличие линейной регрессионной зависимости доли численности населения, с денежными доходами ниже границы бедности ($X6 \sim Y$) от показателей $X2$, $X3$, $X4$, $X5$, $X9$, $X11$, $X12$:

$$\hat{Y} = -\underset{0,1173}{0.3869}\cdot X2 - \underset{0.0654}{0,3902}\cdot X3 + \underset{0.0596}{0.3268}\cdot X4 + \underset{0.0487}{0.2462}\cdot X5 - \underset{0.0002}{0.0007}\cdot X9 + \\ + \underset{4.8\cdot10^{-5}}{0.0001}\cdot X11 + \underset{0.0750}{0.1652}\cdot X12, \quad R^2 = 0.96. \qquad (1)$$

Остатки модели (1) нормально распределены, не автокоррелированы. Коэффициенты модели значимы.

Будем использовать выявленную зависимость для анализа методов импутирования.

Предварительно проведена кластеризация субъектов РФ на основе исходной (без пропусков) таблицы данных. Метод разбиения - K-средних, выбранная мера – Манхеттен, обусловлена отсутствием нормального распределения признаков. Наилучшее значение функционала качества классификации достигнуто при 8 классах. Распределение по классам достаточно логичное, к примеру, Московская область, г. Москва, г. Санкт-Петербург попали в один класс, как наиболее продвинутые регионы по исследуемым показателям. Сделаем оговорку: в виду последующего применения алгоритмов кластеризации, очевидно, что в случае, если в таблице данных пропущены значения во всех трех из указанных регионов, приемлемые значения исследуемого показателя будет восстановить сложно. Указанное количество классов использовалось, как ориентир при проведении кластеризации в вычислительных экспериментах. Далее в экспериментах для кластеризации использовался метод K-средних, для данных, представленных только количественными показателями, применялась мера Манхэттен, для данных, включающих в том числе качественные признаки – мера сходства Миркина [16], которая отличается от других тем, что получена не просто как содержательная экспликация понятия близости, а как результат определенных теоретических предпосылок процесса классификации [17]. Для решения задачи кластеризации использовался разработанный автором инструментарий [18].

Проведена серия экспериментов. В качестве признака, содержащего пропуски, использовался признак $X6$, для которого удалось на исходном наборе данных построить регрессионную зависимость. Объекты, в которых

моделировались пропуски, выбирались случайным образом, с помощью датчика случайных чисел. Вычислительные эксперименты проводились: при 6 пропусках, что соответствует 5% пропущенных данных, 12 пропусках - 10%, 18 пропусках – 15%, 24 пропусках – 20%; 30 пропусках – 25% пропущенных значений.

В анализе участвовали методы импутирования: метод 1 – заполнение нулями; метод 2 – заполнение оценкой математического ожидания; 3 – регрессионный метод; 4 – алгоритм *B*.1, где кластеризация проводилась без учета принадлежности к округу с использованием меры Манхэттен; 5 – алгоритм *B*.1, где кластеризация проводилась с учетом принадлежности к округу методом K-средних с использованием меры сходства Миркина [16]; 6 – алгоритм *A*, кластеризация не проводилась, в качестве классов использовалась географическая принадлежность к округу; 7 – алгоритм *B*.2, где кластеризация проводилась без учета принадлежности к округу с использованием меры Манхэттен.

В качестве показателей эффективности работы алгоритмов импутирования будем использовать оценки статистических показателей выборочной совокупности: математическое ожидание (далее $M[Y]$), среднее квадратическое отклонение ($\sigma[Y]$), коэффициент асимметрии ($A[Y]$), эксцесс ($E[Y]$), коэффициент вариации ($V[Y]$) и среднее значение вектора относительных погрешностей восстановления пропусков:

$$\varepsilon = \frac{1}{n}\left|\frac{y_i - y_i^*}{y_i}\right|. \quad (2)$$

где $Y=\{y_i\},i=1,2..n$ исходный ряд данных по исследуемому показателю, $Y^*=\{y_i^*\},i=1,2..n$ - ряд с восстановленными пропусками.

Лучшим по критерию (2) является метод с наименьшим средним относительных погрешностей: min ε. Показатели эффективности работы

алгоритмов импутирования, полученные в результате вычислительных экспериментов представлены в таблице 1.

Таблица № 1

Показатели эффективности работы алгоритмов импутирования

Кол-во / процент пропущенных данных	Показатель качества метода	Метод							Исходный ряд данных
		1	2	3	4	5	6	7	
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Эксперимент 1: 6 пропусков / 5%	$M[Y]$	13,14	13,14	13,07	13,12	13,01	12,88	12,99	13,08
	$\sigma[Y]$	5,69	4,57	4,69	4,65	4,63	4,72	4,68	4,72
	$A[Y]$	0,14	1,03	0,98	0,96	1,06	1,02	0,99	0,97
	$E[Y]$	1,21	2,21	1,80	1,94	2,10	1,92	2,00	1,72
	$V[Y]$	0,43	0,35	0,36	0,35	0,36	0,37	0,36	0,36
	ε	1,00	0,36	0,05	0,31	0,32	0,28	0,32	-
Эксперимент 2: 12 пропусков / 10%	$M[Y]$	13,07	13,07	-	12,90	12,92	13,08	13,06	13,08
	$\sigma[Y]$	6,11	4,05	-	4,35	4,23	4,33	4,29	4,72
	$A[Y]$	-0,23	0,85	-	0,79	0,75	0,74	0,79	0,97
	$E[Y]$	0,21	2,08	-	1,31	1,53	1,15	1,29	1,72
	$V[Y]$	0,47	0,31	-	0,34	0,33	0,33	0,33	0,36
	ε	1,00	0,45	-	0,33	0,41	0,32	0,28	-
Эксперимент 3: 18 пропусков / 15%	$M[Y]$	11,36	12,96	-	13,19	13,16	13,28	13,33	13,08
	$\sigma[Y]$	7,01	4,42	-	4,66	4,79	4,54	4,60	4,72
	$A[Y]$	0,03	1,28	-	0,99	0,85	0,95	0,99	0,97
	$E[Y]$	-0,22	2,80	-	1,77	1,33	1,94	1,78	1,72
	$V[Y]$	0,62	0,34	-	0,35	0,36	0,34	0,35	0,36
	ε	1,00	0,34	-	0,17	0,39	0,35	0,25	-
Эксперимент 4: 24 пропуска / 20%	$M[Y]$	10,25	12,42	-	13,05	12,95	13,27	13,21	13,08
	$\sigma[Y]$	7,17	4,14	-	4,86	4,13	4,49	4,22	4,72
	$A[Y]$	0,12	1,71	-	1,30	1,30	1,26	1,16	0,97
	$E[Y]$	-0,36	4,63	-	2,69	3,98	2,79	3,14	1,72
	$V[Y]$	0,70	0,33	-	0,37	0,32	0,34	0,32	0,36
	ε	1,00	0,41	-	0,27	0,25	0,38	0,24	-

1	2	3	4	5	6	7	8	9	10
Экспери-мент 5: 30 пропусков / 25%	$M[Y]$	9,41	12,06	-	12,48	12,87	13,39	13,21	13,08
	$\sigma[Y]$	7,69	4,57	-	4,82	4,45	5,20	4,90	4,72
	$A[Y]$	0,35	1,62	-	1,23	1,23	0,93	1,12	0,97
	$E[Y]$	-0,61	3,15	-	1,85	2,72	0,94	1,69	1,72
	$V[Y]$	0,82	0,38	-	0,39	0,35	0,39	0,37	0,36
	ε	1,00	0,32	-	0,23	0,27	0,42	0,22	-

Получено автором в результате вычислительного эксперимента

Модель регрессии для эксперимента 1: 6 пропусков (5%):

$$\begin{aligned} y = &-\underset{0.0005}{0.0003}\cdot X1-\underset{0,1314}{0.3796}\cdot X2-\underset{0.0787}{0,3822}\cdot X3+\underset{0.0675}{0.3280}\cdot X4+\underset{0.0614}{0.2367}\cdot X5-\underset{0.0003}{0.0006}\cdot X9+ \\ &+\underset{5.6\cdot10^{-5}}{0.0001}\cdot X11+\underset{0.0816}{0.1590}\cdot X12+\underset{0.7401}{9.3965}\cdot f_1+\underset{0.1.5300}{9.0752}\cdot f_2+\underset{0.4490}{12.7479}\cdot f_3+ \\ &+\underset{1.2001}{15.8397}\cdot f_4+\underset{0.9027}{18.3144}\cdot f_5+\underset{1.61546}{7.72659}\cdot f_6,\ R^2=0.96, \end{aligned} \quad (3)$$

где f_i – фиктивные переменные, принимающие значение f_i=-1, если пропущено i-е значение объекта для признака Y, f_i=0 в противном случае.

Остатки модели (3) нормально распределены, не автокоррелированы, коэффициенты при фиктивных переменных модели значимы. Коэффициенты при фиктивных переменных соответствуют восстановленным по регрессионной модели значениям пропусков в ряду данных.

Для экспериментов 2-5: (10%-25% пропусков) построить регрессионную зависимость для восстановления пропусков не удалось, так как не выполнялись условия Гаусса-Маркова и незначимыми оказались, в том числе, коэффициенты при фиктивных переменных.

Дополнительно проводился тест по критерию Пирсона на однородность выборок для исходного ряда данных и ряда с заполненными пропусками. Гипотеза $H0$ – выборки однородны, $H1$ – нет оснований считать выборки однородными. В результате:

– метод 1: 5% пропусков – принимается гипотеза $H0$, 10 % и более – $H1$;

– метод 2: 5%-15% пропусков – принимается гипотеза *H*0, 20 % и более – *H*1;

– метод 3: 5% пропусков – принимается гипотеза *H*0;

– методы 3-7: для всех проведенных экспериментов принимается гипотеза *H*0.

Обсуждение и выводы

При 5% пропусков наилучшие результаты дает метод регрессии, однако стоит отметить, что данные для моделирования подбирались, обладающие свойством линейной регрессионной зависимости, в общем же случае подобного рода зависимости построить не удается и метод, зачастую, оказывается неприменим.

Метод заполнения пропусков нулями показывает наихудшие результаты уже при 5% пропусков в данных, и далее, при увеличении количества пропусков, не соответствует статистическим свойствам исходного ряда, метод занижает значение математического ожидания, увеличивает среднее квадратическое отклонение.

Метод заполнения средними значениями имеет лучшие показатели по сравнению с заполнением нулями, при этом он занижает математическое ожидание ряда и существенно ухудшает статистические свойства ряда при росте количества пропусков более 15 %, аналогичные выводы приведены, к примеру, в работах Абраменковой И.В [4], Chhabra G. [2].

Алгоритмы *B*.1 и *B*.2 показывают примерно одинаковые характеристики качества заполнения пропусков во всех экспериментах, и соответствует качеству заполнения пропусков методом заполнения математическим ожиданием (метод 2) при 5% пропусков. При увеличении количества пропусков показатели качества сохраняются, в отличие от метода 2, показатели которого ухудшаются.

Метод 6, в реализации которого кластеризация не проводилась, а использовалась естественная географическая принадлежность к округу, показал наихудшие результаты среди методов на основе статистического моделирования методом обратной функции, следовательно, не рекомендуется исключать этап кластеризации, и при необходимости географическую принадлежность можно учесть при реализации методов кластеризации, допускающих качественные показатели (например, на основе меры Миркина).

Отметим, что показатель средней ошибки методов на основе алгоритмов *B*.1, *B*.2 показывает разброс значений в одном диапазоне, что обусловлено случайностью генерируемых значений, и метод чувствителен к этапам кластеризации данных и классификации объектов с пропусками. В некоторых случаях, по итогу этапа 2 алгоритма *B*.1, в классе с пропуском оказывались все объекты с пропущенными значениями, в этом случае заполнить пропуски можно либо оценкой математического ожидания исходного ряда, либо с использованием базового алгоритма *A* по всем доступным (заполненным) данным ряда, алгоритм *B*.2 лишен подобного недостатка, так как кластеризация проводится по объектам, содержащим полную информацию.

Предлагаемые алгоритмы могут быть применимы при количестве пропусков – более 10%. При меньшем количестве пропусков лучше использовать более простой в реализации метод – метод 2. Если данные допускают использование метода регрессии, то он является преимущественным при небольшом количестве пропусков (не более 10%). При значениях более 10%, метод регрессии существенно искажает коэффициенты модели регрессии, что показывает потеря значимости коэффициентов уже при 5% пропусков.

Заключение

В работе предложены алгоритмы заполнения пропусков в данных на основе метода статистического моделирования обратной функции, где функция распределения строится на основе имеющейся статистической информации для исследуемого показателя в рамках класса, которому принадлежит объект с пропущенным значением. Предложены 3 алгоритма: базовый алгоритм *A* позволяет восстановить пропуски в случае отсутствия информации по другим показателям для исследуемого объекта, а также для восстановления пропусков в данных для объектов в рамках одного класса; алгоритм *B*.1 предусматривающий этап кластеризации объектов по признакам без пропусков; алгоритм *B*.2, включающий этап кластеризации объектов без пропусков и классификации объектов с пропусками. Вычислительные эксперименты проведены на основе статистических данных социально-экономических показателей по субъектам РФ за 2022 год. Анализ свойств предлагаемых алгоритмов импутирования проводился в сравнении с известными методами: заполнение нулями, заполнение оценкой математического ожидания, регрессионным методом. Показана эффективность предлагаемых алгоритмов при количестве пропусков более 10%.

Предлагаемые алгоритмы можно расширить на случай метода множественного импутирования, для наборов данных, допускающих увеличение количества объектов наблюдения.

Литература

1. Зангиева И.К. Проблема пропусков в социологических данных: смысл и подходы к решению // Социология: методология, методы, математическое моделирование. 2011. №33. С. 28-56.

2. Chhabra G., Vashisht V., Ranjan J. A Review on Missing Data Value Estimation Using Imputation Algorithm // Journal of Advanced Research in Dynamical and Control Systems. 2019. no.11. pp. 312-318.

3. Сидорова Д.Н. Подготовка данных для кластеризации событий в журналах информационной безопасности // Инженерный вестник Дона, 2022. №6. URL: ivdon.ru/ru/magazine/archive/n6y2022/7736.

4. Абраменкова И.В., Круглов В.В. Методы восстановления пропусков в массивах данных // Программные продукты и системы. 2005. № 2. С. 18-22.

5. Зангиева И.К., Тимонина Е.С. Сравнение эффективности алгоритмов заполнения пропусков в данных в зависимости от используемого метода анализа // Мониторинг общественного мнения. 2014. №1(119). С. 42- 55.

6. Пимонов А.Г., Глебова Е.А., Сарапулова Т.В., Глебов В.В. Методы, алгоритмы и программные средства для восстановления пропущенных данных в массивах экономической статистики // Экономика и управление инновациями. 2017. №3. С. 52-66. DOI: 10.26730/2587-5574-2017-3-52-65

7. Калинин А.В., Ченцов С.В. Алгоритм восстановления пропусков на поле «плохих» данных // Сибирский аэрокосмический журнал. 2008. №2(19). с. 91-95.

8. Horton N. J; Lipsitz S.R. Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables. // The American Statistician. 2001. Vol.55. No.3. pp. 244-254. URL: jstor.org/stable/2685809

9. Шамрик Д.Л. Базовые методы восстановления пропусков в массивах данных / V Всероссийская молодежная научно-техническая конференция «Информационные технологии в науке и производстве». Омск, 2018. С. 73-83.

10. Фомина Е.Е. Сравнительный анализ методов импутации категориальных переменных в массивах с результатами социологических

опросов // Вестник ПНИПУ. Социально-экономические науки. 2021. №1. С. 83-96. DOI: 10.15593/2224-9354/2021.1.7

11. 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. URL: ivdon.ru/ru/magazine/archive/n1y2012/681.

12. Скрипкина Т.Б. Импутация данных муниципальной статистики // Вестник НГУЭУ. 2020. №3. С. 277-286. DOI: 10.34020/2073-6495-2020-3-277-286

13. Маниковский А.С., Мухопад А.Ю. Методы восстановления пропущенных значений во временных рядах в системе прогнозирования электропотребления // Инженерный вестник Дона, 2022. №7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7824.

14. Matsumoto M., Nishimura T. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom number generator // ACM Transactions on Modeling and Computer Simulation. 1998. vol.8. no.1. pp.3-30. DOI: doi.org/10.1145/272991.272995

15. Российский статистический ежегодник. 2022: Стат.сб. Росстат. М., 2022. 691 с.

16. Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.

17. Чудинова О.С., Ротова А.М., Храмова К.В. Использование методов кластеризации для анализа благосостояния населения // Международная конференция «Наука. Исследования. Практика», СПб:. ГНИИ «НАЦРАЗВИТИЕ», 2019. С. 226-229.

18. Яркова О.Н., Ротова А.М., Храмова К.В. Кластеризация данных. Патент на изобретение № 2021614882. Бюллетень. 2021. № 4. URL: elibrary.ru/item.asp? id=45820952.

References

1. Zangieva I.K. Sotsiologiya: metodologiya, metody, matematicheskoe modelirovanie. 2011. №33. pp. 28-56.

2. Chhabra G., Vashisht V., Ranjan J. Journal of Advanced Research in Dynamical and Control Systems. 2019. no.11. pp. 312-318.

3. Sidorova D.N. Inzhenernyj vestnik Dona, 2022. №6 URL: ivdon.ru/ru/magazine/archive/n6y2022/7736.

4. Abramenkova I.V., Kruglov V.V. Programmnye produkty i sistemy. 2005. № 2. pp. 18-22.

5. 2Zangieva I.K., Timonina E.S. Monitoring obshchestvennogo mneniya. 2014. №1(119). pp. 42- 55.

6. Pimonov A.G., Glebova E.A., Sarapulova T.V., Glebov V.V. Ekonomika i upravlenie innovatsiyami. 2017. №3. pp. 52-66. DOI: 10.26730/2587-5574-2017-3-52-65

7. Kalinin A.V., Chentsov S.V. Sibirskiy aerokosmicheskiy zhurnal. 2008. №2 (19). pp. 91-95.

8. Horton N. J; Lipsitz S.R. The American Statistician. 2001. Vol.55. No.3. pp. 244-254. URL: jstor.org/stable/2685809

9. Shamrik D.L. V Vserossiyskaya molodezhnaya nauchno-tekhnicheskaya konferentsiya «Informatsionnye tekhnologii v nauke i proizvodstve». Omsk, 2018. pp. 73-83.

10. Fomina E.E. Vestnik PNIPU. Sotsial'no-ekonomicheskie nauki. 2021. №1.p. 83-96. DOI: 10.15593/2224-9354/2021.1.7

11. Al'-Kataberi A.S., Shcherbakov M.V., Kamaev V.A. Inzhenernyj vestnik Dona, 2012. №1. URL:ivdon.ru/ru/magazine/archive/n1y2012/681.

12. Skripkina T.B. Vestnik NGUEU. 2020. №3. pp. 277-286. DOI: 10.34020/2073-6495-2020-3-277-286

13. Manikovskiy A.S., Mukhopad A.Yu. Inzhenernyj vestnik Dona, 2022. №7. URL: ivdon.ru/ru/magazine/archive/n7y2022/7824.

14. Matsumoto M., Nishimura T. ACM Transactions on Modeling and Computer Simulation, 1998. vol.8. no.1. pp. 3-30. DOI: doi.org/10.1145/272991.272995.

15. Russian Statistical Yearbook 2022: Stat .book/Rosstat. M., 2022. 691 p.

16. Mandel' I.D. Klasternyy analiz[Cluster analisis] M.: Finansy i statistika, 1988. 176 p.

17. Chudinova O.S., Rotova A.M., Khramova K.V. // International scientific conference «Science. Research. Practice». SPb.: GNII «NATsRAZVITIE», 2019. pp. 226-229.

18. Yarkova O.N., Rotova A.M., Khramova K.V. Patent na izobretenie № 2021614882. Byulleten'. 2021. № 4. URL: elibrary.ru/item.asp?id=45820952