DOI: 10.34020/2073-6495-2020-3-277-286

УДК 311

ИМПУТАЦИЯ ДАННЫХ МУНИЦИПАЛЬНОЙ СТАТИСТИКИ

Скрипкина Т.Б.
Новосибирский государственный университет
экономики и управления «НИНХ»
E-mail: t.b.skripkina@nsuem.ru

Рассмотрена ключевая проблема проведения статистического анализа на массиве данных муниципальной статистики: наличие существенного количества пропущенных данных. Приведено понятие и виды импутации данных. Обосновано применение метода *K* ближайших соседей для проведения импутации на этапе использования данных официальной статистики, содержащихся в Базе данных показателей муниципальных образований. Предложен авторский алгоритм восстановления пропусков в массиве данных показателей муниципальной статистики с использованием системы STATISTICA. Верификация результатов импутации проведена путем сравнения распределений исходных и восстановленных данных на основе критерия согласия Пирсона χ^2.

Ключевые слова: импутация статистических данных, восстановление пропусков данных, муниципальная статистика, метод *K* ближайших соседей, критерий согласия Пирсона χ^2.

IMPUTATION OF MUNICIPAL STATISTICS DATA

Skripkina T.B.
Novosibirsk State University of Economics and Management
E-mail: t.b.skripkina@nsuem.ru

The article considers the key problem of conducting statistical analysis on an array of municipal statistics: the presence of a significant amount of missing data. The concept and types of data imputation are given. The application of the *K* nearest neighbor method for imputation at the stage of using official statistics data contained in the database of municipal indicators is justified. The author's algorithm for restoring omissions in the data set of municipal statistics indicators using the STATISTICA system is proposed. Verification of the imputation results was performed by comparing the distributions of the original and restored data based on the Pearson's consent criterion χ^2.

Key words: imputation of statistical data, restoration of data omissions, municipal statistics, *K* nearest neighbor method, Pearson's consent criterion χ^2.

Введение. Муниципальная статистика – одно из самых молодых направлений в статистике, не насчитывающее и полутора десятков лет своего существования. Основным источником статистической информации является База данных показателей муниципальных образований, открывающая широкие возможности для статистического анализа. При этом одним из наиболее существенных недостатков является большое количество пропусков данных.

Пропуски данных по показателям муниципальной статистики могут возникать по разным причинам: отсутствие явления в данном муниципальном образовании, соблюдение принципа конфиденциальности респондентов (малый размер явления, позволяющий однозначно идентифицировать респондента), недисциплинированность органов местного самоуправления, предоставляющих отчеты в органы государственной статистики, несовершенства информационной системы и др. Массив статистических данных, имеющий пропуски, не позволяет применять большинство методов статистического анализа, автоматизированных в специальных статистических программах. В связи с этим возникает необходимость применять методы импутации данных на этапе статистического исследования.

Понятие импутации данных официально закреплено в Методологических положениях по организации процессов производства официальной статистической информации, утвержденных приказом Росстата от 07.12.2018 N 732 [6]. Согласно Методологическим положениям «процесс импутации – это замещение ошибочных, противоречивых и отсутствующих ответов в процессе редактирования данных другими ответами – значениями показателей». Импутация данных является третьим этапом процесса редактирования данных и может выполняться автоматизированным, ручным способом, а также с помощью комбинации этих двух способов. В документе также сказано о том, что «редактирование и импутация могут осуществляться: в интерактивном режиме по отдельной единице сбора данных; методами пакетной обработки в ходе специальных редакторских «прогонов» данных с использованием специально разработанного программного обеспечения; с использованием комбинации вышеперечисленных методов». В данной работе рассматривается подход к импутации данных автоматизированным способом с использованием специального программного обеспечения.

Целью импутации данных является уменьшение смещения в оценках показателей, обусловленное отсутствием данных или некорректными данными.

Классификация методов импутации данных по версии Межгосударственного статистического комитета Содружества Независимых государств [9] представлена на рис. 1.

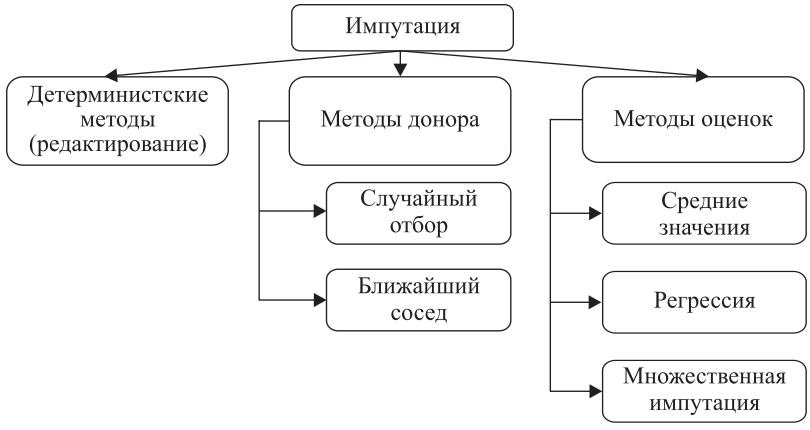

Рис. 1. Классификация методов импутации данных

Детерминистские методы подразумевают возможность восстановления пропущенных значений исходя из арифметических или логических соотношений. Методы донора используют значения признаков объекта-донора для импутации данных объекта-реципиента. При этом донор может быть выбран случайным способом или методом ближайшего соседа (наиболее близкого по характеристикам объекта), который может быть определен различными метриками. Методы оценок предполагают замещение пропущенного значения некоторой его оценкой, которая может быть получена несколькими способами: вычислением среднего значения признака, построением уравнения регрессии, историческими данными и др.

На практике Федеральной службой государственной статистики для импутации данных используется разработка Статистической службы Канады CANCEIS (Canadian Census Edit and Imputation System – канадская система редактирования переписей и условного исчисления). Импутация данных выполняется методом донора с выбором ближайшего соседа и минимизацией вносимых изменений. Данные импутируются по домашнему хозяйству и членам домохозяйства [4].

Алгоритм импутации данных муниципальной статистики. Необходимость разработки авторского алгоритма импутации данных муниципальной статистики обусловлена следующими причинами:

1. Многомерный массив данных муниципальной статистики содержит существенное количество пропусков, которые не позволяют корректно применять многомерные статистические методы анализа. Так, массив данных об инфраструктуре муниципальных районов Российской Федерации за 2018 г. (53 признака) содержит всего 6 полностью заполненных строк из 1750.

2. По большинству признаков причиной пропусков данных не является отсутствие явления в данном муниципальном образовании, следовательно, бóльшая часть пропусков подлежит восстановлению.

3. Официальной статистикой описаны методы импутации данных, применимые на этапе сбора и подготовки статистического материала к публикации. Алгоритмы, позволяющие проводить импутацию данных пользователям (а не производителям) статистических данных о муниципальных образованиях, в настоящее время не описаны.

На входе алгоритма имеется исходный массив $n \times m$, сформированный на основе данных муниципальной статистики исходя из целей исследования, где n – число объектов наблюдения (муниципальных образований), m – количество изучаемых признаков.

Первый подготовительный этап предполагает исключение из массива данных мало наполненных признаков. В данном алгоритме мало наполненными признаками считаются признаки с долей валидных наблюдений менее 50 %. Другими словами, если данные по признаку m_i имеются менее, чем у половины наблюдений, такой признак исключается из анализа. На выходе данного этапа получаем массив данных $n \times (m - k)$, где k – количество не валидных признаков.

На втором этапе проводится собственно восстановление пропусков. Выбор метода импутации (рис. 2) обусловлен следующими рассуждениями. Применение детерминистских методов наиболее эффективно на этапе

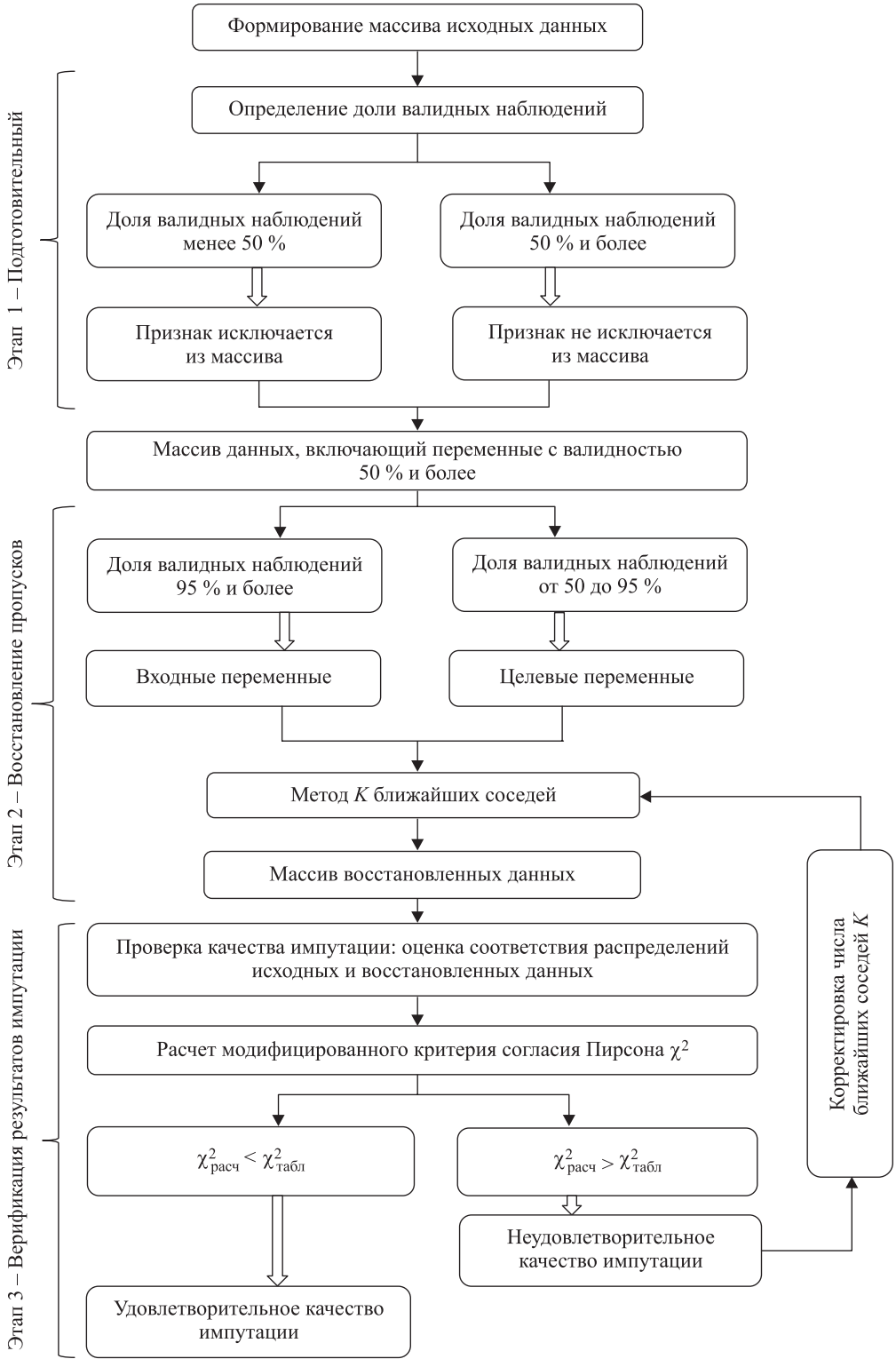

Рис. 2. Авторский алгоритм восстановления пропусков в массиве данных показателей муниципальной статистики с использованием системы STATISTICA

сбора статистических данных по каждому объекту наблюдения. Методы оценок требуют достаточно однородной совокупности и могут найти свое применение уже после типологии муниципальных образований – внутри типов. Среди донорских методов случайный метод импутации также требует предварительной классификации объектов. По указанным причинам наиболее подходящим является метод ближайшего соседа, а точнее K ближайших соседей, предполагающий поиск подобных объектов в количестве K штук, находящихся на наименьшем расстоянии от объекта-реципиента. Затем вычисляется среднее значение признака m_i по K объектам-донорам, которое импутируется в качестве искомого пропущенного значения объекту-реципиенту.

Так как выбранные объекты-доноры являются самыми похожими на объект-реципиент, подразумевается, что значения признака m_i у данных объектов достаточно близки. Поэтому импутация данным способом дает наиболее правдоподобные результаты. Средством реализации данного метода выступает система STATISTICA.

В модуле MD Imputation системы STATISTICA заложен метод K ближайших соседей, предполагающий оценку близости объектов на основе стандартизированных евклидовых расстояний. В качестве входных переменных (input variables) используются переменные, у которых доля валидных наблюдений выше 95 %. По данным переменным вычисляется матрица стандартизованных расстояний Евклида и определяются ближайшие соседи – доноры. Данные восстанавливаются в целевых переменных (target variables), где валидность наблюдений находится в пределах 50–95 %.

На выходе второго этапа получаем массив восстановленных данных размерностью $n \times (m - k)$, в котором значения целевых переменных заполнены по всем объектам, а входных переменных – минимум у 95 % объектов.

Третьим этапом алгоритма импутации данных является верификация полученных результатов. В научной литературе предложены различные способы верификации. Так, в качестве критерия качества импутации Е.Е. Фомина предлагает использовать модифицированную формулу средней ошибки аппроксимации, с помощью которой исследуется стандартизованные отклонения между фактическими и предсказанными значениями [8]. Данный критерий может быть использован только в случае, когда известны фактические значения и предсказанные разными методами значения переменных. Наименьшая величина средней ошибки аппроксимации характеризует наиболее подходящий метод импутации. Другой коллектив авторов [3] для определения качества импутации использует критерий Стьюдента для проверки равенства средних на уровне значимости $p = 0{,}05$. Однако при равенстве средних значений исходная и восстановленная совокупность могут существенно отличаться другими параметрами (например, дисперсией). В рамках модель-ориентированного подхода М.С. Фабрикант утверждает, что если «…импутация не привела к искажению данных, то полученные после импутации коэффициенты не должны существенно отличаться от первоначальных…» [7]. В данном случае речь идет о коэффициентах уравнений регрессии, построенных до и после импутации. При этом не понятен критерий существенности отличий в коэффициентах.

Таблица 1

Соотношение объема совокупности и числа интервалов рядов распределения по рекомендации ВНИИ Метрологии

n	*k*
40–100	7–9
100–500	8–12
500–1000	10–16
1000–10000	12–22

В настоящем алгоритме для оценки качества импутации предложено использовать распределение частот. В случае, если форма распределения после восстановления пропущенных данных существенно не изменилась, качество импутации можно признать хорошим, поскольку в данном случае смещение параметров не происходит. В качестве критерия оценки качества используется модифицированный критерий χ^2, позволяющий оценить степень согласованности двух распределений – до восстановления данных и после.

Для оценки качества импутации формируются ряды распределения с равными интервалами, в связи с чем встает вопрос о количестве интервалов. Как отмечается в публикации «О выборе числа интервалов в критериях согласия типа χ^2» (авторы Б.Ю. Лемешко, Е.В. Чимитова), «при больших объемах выборок *n* разброс значений *k*, задаваемых различными формулами, достаточно велик. Поэтому на практике при выборе числа интервалов больше руководствуются разумными соображениями, выбирая число интервалов так, чтобы в интервалы попадало число наблюдений не менее 5–10». Авторы ссылаются на рекомендации ВНИИ Метрологии [1], в которых в зависимости от *n* предлагаются следующие величины *k* (табл. 1) [5].

Наиболее часто применяемая формула Стерджесса дает заниженные результаты на большом объеме совокупности. Так, при количестве наблюдений $N = 1750$ формула Стерджесса определяет 11 интервалов, когда в соответствии с рекомендациями их должно быть не менее 12. В связи с этим вопрос о количестве интервалов предлагается решить с помощью формулы Брукса и Каррузера [2]:

$$k = 5\lg N, \tag{1}$$

где *k* – число интервалов; *N* – объем совокупности.

При том же объеме совокупности формула дает 16 интервалов, что соответствует рекомендациям ВНИИ Метрологии.

После построения двух рядов распределения по исходным и по восстановленным данным вычисляется модифицированный критерий согласия Пирсона χ^2 по формуле

$$\chi^2_{\text{расч}} = \sum_{i=1}^{k} \frac{(f_{\text{восст}} - f_{\text{исх}})^2}{f_{\text{исх}}}, \tag{2}$$

где $f_{\text{восст}}$ – частость в ряду распределения восстановленных данных, $f_{\text{исх}}$ – частость в ряду распределения исходных данных.

Расчетное значение критерия сравнивается с теоретическим (табличным). В случае, если $\chi^2_{\text{расч}} < \chi^2_{\text{табл}}$, то различия между распределениями исходных и восстановленных данных признаются не существенными, а качество импутации данных удовлетворительным. В противном случае необходимо вернуться на второй этап алгоритма и изменить количество ближайших соседей. *K* должно быть достаточно большим, чтобы минимизировать ве-

роятность неверной импутации, и достаточно малым, чтобы *K* ближайших точек были достаточно близки к точке запроса. Слишком малое *K* ведет к увеличению дисперсии распределения. Слишком большое *K* – к смещению параметров распределения. Таким образом, существует оптимальное *K*, которое обеспечивает равновесие между смещением и дисперсией распределения.

Апробация. С целью статистического исследования инфраструктуры муниципальных районов всех субъектов Российской Федерации на основе Базы данных показателей муниципальных образований сформирован массив статистических данных за 2018 г. размерностью 53×1750. На первом этапе алгоритма в результате анализа валидности шесть показателей исключены из анализа, так как данными заполнены менее половины всех наблюдений. Оставшийся массив признаков разделился на 15 входных и 32 целевых переменных. С помощью модуля MD Imputation системы STATISTICA проведена процедура импутации данных. С целью верификации ее результатов построены ряды распределения по исходным и восстановленным данным, вычислен модифицированный критерий χ^2 по каждому признаку (табл. 2).

Таблица 2

Верификация результатов импутации статистических данных об инфраструктуре муниципальных районов Российской Федерации в 2018 г.

№ п/п	Признак	Валидность, %	Расчетный критерий согласия Пирсона ($\chi^2_{расч}$)
1	2	3	4
1	Протяженность автодорог общего пользования местного значения с усовершенствованным покрытием, находящихся в собственности муниципального образования (на конец года), км	92,3	0,001
2	Количество автозаправочных станций (АЗС), расположенных на автомобильных дорогах общего пользования местного значения, единиц	64,9	0,019
3	Общая протяженность улиц, проездов, набережных на конец года, км	93,9	0,003
4	Общая протяженность освещенных частей улиц, проездов, набережных на конец года, км	93,6	0,004
5	Одиночное протяжение уличной газовой сети, м	71,9	0,096
6	Количество негазифицированных населенных пунктов, ед.	94,4	0,001
7	Число источников теплоснабжения мощностью до 3 Гкал/ч	92,3	0,005
8	Протяженность тепловых и паровых сетей в двухтрубном исчислении, нуждающихся в замене, м	85,3	0,001
9	Протяженность тепловых и паровых сетей, которые были заменены и отремонтированы за отчетный год, м	68,7	0,001
10	Одиночное протяжение уличной водопроводной сети, нуждающейся в замене, м	93,5	0,001
11	Одиночное протяжение уличной водопроводной сети, которая заменена и отремонтирована за отчетный год, м	77,8	0,001

Окончание табл. 2

1	2	3	4
12	Количество населенных пунктов, не имеющих водопроводов (отдельных водопроводных сетей), ед.	87,0	0,000
13	Одиночное протяжение уличной канализационной сети, м	80,5	0,006
14	Одиночное протяжение уличной канализационной сети, нуждающейся в замене, м	72,8	0,012
15	Количество населенных пунктов, не имеющих канализаций (отдельных канализационных сетей), ед.	93,0	0,007
16	Число лечебно-профилактических организаций – всего, ед.	94,7	0,000
17	Численность воспитанников, посещающих организации, осуществляющие образовательную деятельность по образовательным программам дошкольного образования, присмотр и уход за детьми, ед.	94,4	0,009
18	Число детско-юношеских спортивных школ (включая филиалы), ед.	81,7	0,003
19	Число самостоятельных детско-юношеских спортивных школ, ед.	67,6	0,011
20	Численность занимающихся в детско-юношеских спортивных школах, ед.	82,3	0,009
21	Текущие (эксплуатационные) затраты на охрану окружающей среды, включая оплату услуг природоохранного назначения, тыс. руб.	67,7	0,002
22	Вывезено за год твердых коммунальных отходов, тыс. куб. м	84,2	0,000
23	Введено в действие жилых домов на территории муниципального образования, квадратный метр общей площади	94,1	0,000
24	Введено в действие индивидуальных жилых домов на территории муниципального образования, квадратный метр общей площади	94,2	0,000
25	Оборот розничной торговли (без субъектов малого предпринимательства), тыс. руб.	62,4	0,021
26	Общий объем всех продовольственных товаров, реализованных в границах муниципального района, в денежном выражении за финансовый год, тыс. руб.	94,5	0,007
27	Число ярмарок, ед.	60,4	0,506
28	Число мест в объектах общественного питания, ед.	70,7	0,006
29	Число торговых мест на ярмарках, ед.	57,8	0,379
30	Число объектов бытового обслуживания населения, оказывающих услуги, ед.	71,7	0,170
31	Число коллективных средств размещения, ед.	84,1	0,002
32	Число мест в коллективных средствах размещения, ед.	84,1	0,001

Теоретическое значение критерия согласия Пирсона при уровне значимости $\alpha = 0{,}05$ и $df = 16 - 1 = 15$ составляет 7,26. Расчетные значения критерия согласия Пирсона по всем признакам, подвергнутым импутации, меньше теоретического, что свидетельствует о хорошем качестве импутации данных.

Заключение. Предложенный алгоритм позволяет восстановить пропуски в массиве данных при отсутствии признаков со стопроцентной валидностью. Показал удовлетворительные результаты при апробации на реальном массиве данных. Собранные из официальных источников, а именно Базы данных муниципальных образований данные подготовлены к дальнейшему статистическому анализу, в том числе применению многомерных статистических методов.

Литература

1. *Бурдун Г.Д., Марков Б.Н.* Основы метрологии. М.: Изд-во стандартов, 1985. 120 с.
2. *Зайков К.А.* К вопросу оценки уровня инновационного потенциала субъектов Российской Федерации // Вестник НГУЭУ. 2019. № 1. С. 134–151.
3. *Бых А.И., Высоцкая Е.В., Рак Л.И., Порван А.П., Болибок Е.Е., Сватенко О.А.* Выбор метода восстановления пропущенных данных для оценки сердечно-сосудистой деятельности подростков // Восточно-Европейский журнал передовых технологий. 2010. № 3/4 (45). С. 4–7. [Электронный ресурс]. URL: https://cyberleninka.ru/article/n/vybor-metoda-vosstanovleniya-propuschennyh-dannyh-dlya-otsenki-serdechno-sosudistoy-deyatelnosti-podrostkov/viewer
4. *Ковтун А.* Развитие и совершенствование процессов обработки данных выборочных обследований. [Электронный ресурс]. URL: https://www.gks.ru/free_doc/new_site/rosstat/smi/prezent23-2013/kovtun.pdf
5. *Лемешко Б.Ю., Чимитова Е.В.* О выборе числа интервалов в критериях согласия типа χ^2 // Заводская лаборатория. Диагностика материалов. 2003. Т. 69. С. 61–67. URL: https://www.researchgate.net/publication/315333672_O_vybore_cisla_intervalov_v_kriteriah_soglasia_tipa_X2
6. Приказ Росстата от 07.12.2018 N 732 «Об утверждении Методологических положений по организации процессов производства официальной статистической информации». [Электронный ресурс]. URL: http://www.consultant.ru/document/cons_doc_LAW_313411/ (дата обращения: 13.03.2020).
7. *Фабрикант М.С.* Модель-ориентированный подход к отсутствующим значениям: множественная импутация в многоуровневой регрессии посредством R (на примере анализа опросных данных) // Социология: методология, методы, математическое моделирование. 2015. № 41. С. 7–29. [Электронный ресурс]. URL: https://www.elibrary.ru/download/elibrary_26539204_98997170.pdf
8. *Фомина Е.Е.* Обзор методов и программного обеспечения для восстановления пропущенных значений в массивах социологических данных // Гуманитарный вестник. 2019. № 4. С. 1–12. [Электронный ресурс]. URL: http://www.hmbul.ru/articles/611/611.pdf
9. Хрестоматия практико-ориентированного комплекса учебно-методических материалов по курсу «Организация выборочных обследований». Межгосударственный статистический комитет Содружества Независимых государств. С. 42. URL: http://www.cisstat.com/Trainig_courses/CIS_training_course_Organization_of_sample_surveys/CIS_training_course_Organization_of_sample_surveys_07%20-%20reader.pdf

Bibliography

1. *Burdun G.D., Markov B.N.* Osnovy metrologii. M.: Izd-vo standartov, 1985. 120 p.
2. *Zajkov K.A.* K voprosu ocenki urovnja innovacionnogo potenciala sub#ektov Rossijskoj Federacii // Vestnik NGUJeU. 2019. № 1. P. 134–151.
3. *Byh A.I., Vysockaja E.V., Rak L.I., Porvan A.P., Bolibok E.E., Svatenko O.A.* Vybor metoda vosstanovlenija propushhennyh dannyh dlja ocenki serdechno-sosudistoj dejatel'nosti podrostkov // Vostochno-Evropejskij zhurnal peredovyh tehnologij. 2010.

№ 3/4 (45). P. 4–7. [Jelektronnyj resurs]. URL: https://cyberleninka.ru/article/n/vybor-metoda-vosstanovleniya-propuschennyh-dannyh-dlya-otsenki-serdechno-sosudistoy-deyatelnosti-podrostkov/viewer

4. *Kovtun A*. Razvitie i sovershenstvovanie processov obrabotki dannyh vyborochnyh obsledovanij. [Jelektronnyj resurs]. URL: https://www.gks.ru/free_doc/new_site/rosstat/smi/prezent23-2013/kovtun.pdf
5. *Lemeshko B.Ju., Chimitova E.V*. O vybore chisla intervalov v kriterijah soglasija tipa χ^2 // Zavodskaja laboratorija. Diagnostika materialov. 2003. Vol. 69. P. 61–67. URL: https://www.researchgate.net/publication/315333672_O_vybore_cisla_intervalov_v_kriteriah_soglasia_tipa_X2
6. Prikaz Rosstata ot 07.12.2018 N 732 «Ob utverzhdenii Metodologicheskih polozhenij po organizacii processov proizvodstva oficial'noj statisticheskoj informacii». [Jelektronnyj resurs]. URL: http://www.consultant.ru/document/cons_doc_LAW_313411/ (data obrashhenija: 13.03.2020).
7. *Fabrikant M.S*. Model'-orientirovannyj podhod k otsutstvujushhim znachenijam: mnozhestvennaja imputacija v mnogourovnevoj regressii posredstvom R (na primere analiza oprosnyh dannyh) // Sociologija: metodologija, metody, matematicheskoe modelirovanie. 2015. № 41. P. 7–29. [Jelektronnyj resurs]. URL: https://www.elibrary.ru/download/elibrary_26539204_98997170.pdf
8. *Fomina E.E*. Obzor metodov i programmnogo obespechenija dlja vosstanovlenija propushhennyh znachenij v massivah sociologicheskih dannyh // Gumanitarnyj vestnik. 2019. № 4. P. 1–12. [Jelektronnyj resurs]. URL: http://www.hmbul.ru/articles/611/611.pdf
9. Hrestomatija praktiko-orientirovannogo kompleksa uchebno-metodicheskih materialov po kursu «Organizacija vyborochnyh obsledovanij». Mezhgosudarstvennyj statisticheskij komitet Sodruzhestva Nezavisimyh gosudarstv. P. 42. URL: http://www.cisstat.com/Trainig_courses/CIS_training_course_Organization_of_sample_surveys/CIS_training_course_Organization_of_sample_surveys_07%20-%20reader.pdf