УДК 316:004.9 DOI: 10.18698/2306-8477-2019-4-611

# Обзор методов и программного обеспечения для восстановления пропущенных значений в массивах социологических данных

© Е.Е. Фомина

Тверской государственный технический университет, Тверь, 170026, Россия

*Анализ социологических данных сопряжен с исследованием больших массивов переменных, которые могут содержать пропущенные значения. Наличие значительного числа некомплектных записей приводит к искажению результатов статистического анализа, неверной интерпретации результатов моделирования. В статье представлен обзор методов и программного обеспечения, предназначенных для импутации отсутствующих значений при проведении социологических исследований. Рассмотрены математическая сущность, преимущества и недостатки наиболее распространенных методов восстановления пропусков, используемых при решении практических задач. Приведен обзор современного программного обеспечения, используемого для решения подобных задач. Предложена методика выбора наиболее эффективного алгоритма импутации.*

***Ключевые слова:*** *импутация данных, восстановление пропущенных значений, некомплектные наблюдения*

Анализ данных, полученных в результате проведения опросов, измерения значений социально-экономических показателей, описывающих изучаемый объект или явление, предполагает обработку больших массивов, включающих набор некоторого количества переменных.

Часто такие переменные содержат пропущенные значения. Природа пропусков бывает различной. Они могут быть связаны как с нежеланием респондента отвечать на тот или иной вопрос, так и с отсутствием значения какого-либо показателя у некоторых объектов.

Д. Рубин и Р. Литтл предложили следующую классификацию механизмов формирования пропусков [1].

1. MCAR (Missing Completely At Random, полностью случайные пропуски) — это механизм формирования пропусков, при котором для каждой записи набора данных вероятность пропуска одинакова.

Если рассмотреть набор переменных $X_1$, …, $X_N$, $X_p$, то условная вероятность того, что значение переменной $X_p$ пропущено, не зависит ни от $X_p$, ни от других переменных $X_1$, …, $X_N$ ($P = \{X_p = \text{пропуск} \mid X_1, \ldots, X_N, X_P\} = P\{X_p = \text{пропуск}\}$).

2. MAR (Missing At Random, случайные пропуски) — это механизм формирования пропусков, при котором данные пропущены не

случайно, а ввиду некоторых факторов. Пропуск относят к типу MAR, если его вероятность может быть вычислена на основе другой имеющейся в наборе данных информации.

Условная вероятность того, что значение переменной $X_p$ пропущено, не зависит от $X_p$, но может зависеть от других переменных $X_1$, …, $X_N$ ($P = \{X_p = \text{пропуск} \mid X_1, \ldots, X_N, X_P\} = P\{X_p = \text{пропуск} \mid X_1, \ldots, X_N\}$).

3. MNAR (Missing Not At Random, неслучайные пропуски) — механизм формирования пропусков, при котором отсутствие значений связано с неизвестными факторами. Как следствие, вероятность пропуска невозможно выразить на основе информации, содержащейся в наборе данных.

Наличие в массиве пропусков, относящихся к типу MNAR, является для исследователя сигналом о том, что необходимо совершенствовать инструментарий или способ сбора информации.

Понимание механизма формирования пропуска существенно, так как оно дает представление о степени важности пропущенных значений [2]. Поэтому его выявление — первый этап в решении задачи импутации.

В работе [3] рассмотрены критерии оценивания типа пропуска.

Следует отметить, что методы восстановления пропущенных значений, или импутации, применимы только в том случае, когда пропуски данных соответствуют типу MCAR или MAR, в противном случае отсутствие значимых переменных приведет к построению некачественной модели.

Сегодня разработано большое количество алгоритмов для решения задачи импутации. Однако нужно отметить, что отсутствует методология решения задач подобного рода. Нет методических рекомендаций, согласно которым можно было бы определить наиболее эффективный алгоритм для восстановления пропущенных значений в массиве данных различных типов и структуры.

Цель настоящей статьи — провести обзор математических методов и современного программного обеспечения, которые могут быть применены для решения задачи импутации в социологических исследованиях, а также предложить алгоритм выбора наиболее эффективного метода заполнения пропусков при решении практических задач.

**Постановка задачи.** Обозначим $\mathbf{A_{n\times m}}$ матрицу исходных данных, где $n$ — количество строк матрицы, каждая из которых соответствует тому или иному наблюдению (объекту); $m$ — количеством столбцов (переменных), каждый из которых содержит значения некоторого признака, измеренного у каждого объекта; $a_{ij}$ — элементы матрицы (могут быть как непрерывными, так и категориальными переменными).

Некоторые элементы $a_{ij}$ матрицы $\mathbf{A_{n\times m}}$ могут содержать пропущенные значения, которые требуется восстановить.

**Обзор методов.** На практике используются рассмотренные ниже подходы при работе с пропущенными значениями.

1. Удаление некомплектных наблюдений из базы данных.

Наблюдения, содержащие пропуски, будут удалены из анализа. Этот метод является методом по умолчанию в большинстве статистических пакетов обработки данных. Недостаток его очевиден, так как удаление наблюдений ведет к сокращению объема выборки, потере информации и смещению результатов анализа.

Так, например, при кластерном анализе исключение наблюдений с пропусками приведет к тому, что ряд объектов не будет отнесен ни к одному из кластеров. В свою очередь, включение этих объектов может дать совершенно иной результат разбиения на группы.

Данный метод без существенных последствий можно применять в том случае, если пропуски соответствуют категории MCAR и процент неполных наблюдений достаточно мал (менее 5 %) [4].

2. Взвешивание наблюдений.

Метод позволяет сохранить требуемый объем выборки, удалив при этом неполные наблюдения. Комплектным наблюдениям будет присвоен весовой коэффициент, определяемый той переменной, для которой необходимо сохранить структуру. Или же веса будут увеличены для случайно отобранных наблюдений. Присвоение веса, большего единицы, позволит дополнить выборку до требуемого объема [4].

Недостаток метода заключается в том, что присвоение веса может привести к существенному смещению в оценке параметров.

3. Анализ доступных наблюдений.

Данный метод предполагает использование в исследовании тех наблюдений, которые содержат значения анализируемой в настоящий момент переменной (переменных). Остальные переменные могут содержать пропуски.

Недостаток данного метода заключается в невозможности сравнения значений итоговых показателей, так как они могут быть рассчитаны по разным подвыборкам исходной совокупности.

4. Восстановление (импутация) пропущенного значения.

Р. Литтл предложил систему классификации методов импутации [1], согласно которой методы делятся на простые и сложные. Сложные, в свою очередь, подразделяются на локальные, которые используют для заполнения пропусков не весь массив данных, а только «близкие» наблюдения, и глобальные, которые для восстановления пропущенных значений опираются на весь массив данных.

Будем придерживаться этой классификации и приведем обзор методов каждой группы.

Напомним, что применение описанных ниже методов возможно, если механизм формирования пропусков соответствует MCAR или MAR.

***Простые методы импутации.*** При *замене мерой средней тенденции* восстановление пропущенных данных происходит путем подстановки вместо пропущенных значений в столбце матрицы $\mathbf{A_{n\times m}}$ среднего арифметического, рассчитанного по этому столбцу, в случае если данные являются интервальными, моды, если данные являются номинальными, и медианы, если данные являются порядковыми.

Этот метод может применяться только для нормально распределенных данных, в противном случае (при наличии выраженной асимметрии или эксцесса) основная масса значений может быть далека от среднего; таким образом, процедура замены приведет к искажению структуры данных, недооценке ее неоднородности.

На практике данный метод, как правило, дает высокую ошибку предсказания [5].

*Метод HotDeck* восстанавливает пропущенные значения признака того или иного объекта (строки матрицы $\mathbf{A_{n\times m}}$), используя значения сходных комплектных объектов из представленного набора данных.

В основе метода лежит предположение о том, что если объекты схожи между собой по значениям $(m-1)$-й переменной, то они схожи и по значениям $m$-й переменной.

Заполнение пропуска в значении переменной у некомплектного объекта происходит путем подстановки значения той же переменной ближайшего комплектного объекта. Для определения меры близости вычисляются расстояния между объектами, которые выбираются исходя из типов представленных данных.

В качестве меры сходства может быть использован коэффициент сходства Гауэра, допускающий применение переменных, представленных в различных шкалах [6].

Недостаток метода заключается в вычислительных затратах при поиске ближайшего объекта.

Метод HotDeck может применяться в сочетании с кластерным анализом. На первом этапе комплектные объекты будут разбиты на кластеры. Далее пропущенные значения переменных некомплектного объекта заполняются соответствующими значениями центроида того кластера, который является ближайшим.

Сущность метода *регрессионного анализа* заключается в том, что строится модель множественной линейной регрессии, в которой зависимая переменная — столбец, содержащий пропуски ($P$), независимые переменные — столбцы, значения которых не содержат пропусков ($X_1, ..., X_m$): $P = a_0 + a_1X_1 + a_2X_2 + ... + a_mX_m$.

Для поиска коэффициентов уравнения используется метод наименьших квадратов. Неизвестное пропущенное значение рассчитывается путем подстановки в уравнение регрессии значения других признаков, соответствующих обрабатываемой записи.

Недостаток способа заключается в том, что в набор независимых переменных должны входить только те переменные, которые имеют высокую корреляцию с зависимой. На практике возможна ситуация, когда переменные, имеющие высокую корреляцию с зависимой, сами содержат пропуски.

Для восстановления значений дихотомической переменной используется метод логистической регрессии [6, 7].

***Сложные методы импутации. Локальные алгоритмы.*** В основе алгоритма ZET лежат следующие предположения, являющиеся одновременно ограничениями для его применения:

- избыточности. Предположение заключается в том, что в исследуемой совокупности имеются наблюдения (строки матрицы $\mathbf{A_{n\times m}}$), схожие между собой и зависящие друг от друга признаки (столбцы матрицы $\mathbf{A_{n\times m}}$);
- локальной компактности. Предположение заключается в том, что для предсказания пропущенных значений используется не вся матрица, а ее компетентная часть, состоящая из элементов близких строк и похожих столбцов. Компетентная часть не должна иметь пропусков;
- линейной зависимости. Предположение заключается в том, что зависимости между строками и столбцами носят линейный характер.

Алгоритм ZET включает в себя следующие этапы: отбор компетентной части для заполнения пропуска; расчет коэффициентов уравнения, используемого для прогнозирования пропущенного значения; вычисление прогнозируемого значения [8].

Алгоритм ZET разработан сотрудниками ИМ СО РАН Н.Г. Загоруйко, В.Н. Елкиной и В.С. Темиркаевым [9].

Данный алгоритм показывает более высокое качество восстановления пропусков по сравнению с регрессионным методом, так как он учитывает закономерности исследуемого набора данных [5].

***Глобальные алгоритмы.*** *Алгоритм Бартлета* основан на регрессионном методе и включает три итерации:

1) пропуски в некомплектных записях заполняются начальными значениями. В качестве начального приближения можно использовать среднее арифметическое значение по соответствующей переменной;

2) строится модель множественной регрессии, где зависимая переменная соответствует столбцу с пропусками, остальные переменные — независимые;

3) построенное уравнение регрессии используется для прогнозирования пропущенных значений [10].

Недостаток метода Бартлета такой же, как у метода регрессионного анализа: он связан с предположением о линейной зависимости между переменными, что не всегда наблюдается на практике.

Особенность *EM-алгоритма* заключается в построении модели порождения пропусков с получением выводов на основании функции максимального правдоподобия. Каждая итерация алгоритма состоит из двух шагов. На Е-шаге (expectation) вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. На М-шаге (maximization) вычисляется оценка максимального правдоподобия, таким образом, увеличивается ожидаемое правдоподобие, вычисляемое на Е-шаге. Затем это значение используется для Е-шага на следующей итерации. Алгоритм выполняется до сходимости [8].

Недостаток метода заключается в построении модели порождения пропусков [5].

*Алгоритм resampling* — итерационный метод, который возможен в двух модификациях. В первой модификации пропущенные значения некомплектных наблюдений случайным образом заменяются на соответствующие значения комплектных наблюдений из исходного массива данных, далее строится уравнение регрессии. Во втором варианте уравнение регрессии получают из комплектной подматрицы.

Случайные значения переменной в первом случае и малая мощность набора комплектных переменных во втором могут послужить причиной получения неверных результатов [5].

*Множественная импутация данных* — наиболее распространенный сегодня метод заполнения пропусков в социологической практике. Сущность метода заключается в том, что на место каждого пропуска подставляют несколько значений, т. е. формируются $k$ наборов данных, или $k$ матриц $\left\{A^{j}_{n\times m}\right\}_{j=\overline{1,k}}$ (как правило, $k = 5$). Далее на место отсутствующего значения подставляют среднее значение, рассчитанное по всем построенным моделям [4].

Каждый из наборов получается с использованием одной из следующих моделей — предиктивной, степени предрасположенности или дискриминантной [3].

К недостаткам метода следует отнести большие временные и вычислительные затраты по сравнению с любым из рассмотренных выше методов.

*Нейронные сети* используются для решения широкого класса задач, к которым относятся задачи кластеризации, распознавания образов, оптимизации и другие. В том числе нейронные сети могут применяться для решения задачи прогнозирования пропущенных значений.

Для реализации этого метода требуется специальное программное обеспечение, содержащее модуль работы с нейронными сетями.

Наиболее эффективен для решения задачи импутации значений интервальных переменных многослойный персептрон с обратным распространением ошибки [6, 11, 12].

**Программное обеспечение.** Существующие статистические пакеты обработки данных не позволяют проводить процедуру импутации всеми описанными выше способами. В них в основном реализованы такие методы, как удаление некомплектных записей, замена пропущенных значений средним арифметическим, восстановление с использованием регрессионного метода и метода множественной импутации.

Среди программного обеспечения, которое может применяться в социологических исследованиях для решения задачи восстановления пропущенных значений, можно отметить MS Excel, SPSS Statistics, среду статистического анализа R , ряд специализированных пакетов, не имеющих широкого распространения, речь о которых пойдет ниже.

Рассмотрим каждый из представленных выше пакетов подробнее.

Табличный процессор MS Excel содержит набор статистических функций и надстройку «анализ данных» для реализации простых способов импутации.

SPSS Statistics включает два модуля для работы с некомплектными записями [13–16]:

- модуль Missing Value Analysis (MVA, анализ пропущенных значений) позволяет осуществлять проверку данных и обнаруживать закономерности в распределении пропущенных значений;
- модуль Multiple Imputation (множественная импутация) позволяет восстанавливать пропущенные значения методом множественной импутации. Работа модуля возможна в двух режимах: в автоматическом (автоматически выбирает метод импутации на основе анализа сканирования данных) и в пользовательском.

Импутация в пользовательском режиме предполагает применение итерационного метода Монте-Карло с использованием Цепей Маркова (как для монотонной, так и для немонотонной структуры данных) и неитерационного метода, который можно использовать только при наличии монотонной структуры пропущенных значений. Для каждой переменной в монотонном порядке метод строит одномерную модель (линейной регрессии или логистической регрессии в зависимости от типа переменных), используя все предыдущие переменные как независимые, далее на основе построенной модели происходит импутация пропущенного значения.

Среда статистического анализа R включает в себя такие библиотеки (пакеты), как Amelia, MICE и MI, содержащие функции для вос-

становления пропущенных значений как простыми методами, так и методом множественной импутации [4].

К специализированным пакетам, которые не имеют широкого распространения, но могут эффективно использоваться для решения задачи импутации, относятся STEPS и AGGITS (США), GEIS (Канада), SOLAS (Ирландия) [17].

**Методика выбора наиболее эффективного алгоритма импутации.** Выбирая тот или иной математический алгоритм для решения практической задачи, необходимо опираться не только на его преимущества и недостатки, но и на структуру выборки, взаимосвязи между описывающими ее переменными, дальнейший инструментарий исследования, который будет применяться после восстановления пропусков.

Эффективность того или иного метода устанавливается экспериментально в такой последовательности:

1) формируется массив комплектных записей. Для этого неполные наблюдения исключаются из рассмотрения;

2) «искусственно» создаются пропущенные значения, т. е. в таблице удаляются некоторые элементы $a_1, ..., a_n$;

3) пропущенные значения поочередно предсказываются с использованием разных методов $M_1, ..., M_k$;

4) рассчитываются суммарные относительные погрешности для каждого метода:

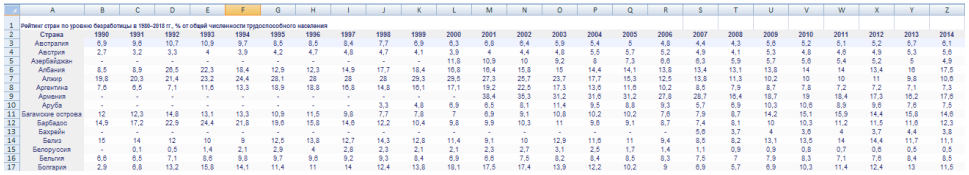
$$\Delta_{M_j} = \sum_{i=1}^{n} \frac{|a_i - \overline{a}_i|}{a_i} \cdot 100\,\%,$$

где $j$ — номер метода, $j = 1, ..., k$; $a_i$ — пропущенное значение; $\overline{a}_i$ — предсказанное значение.

Метод, для которого суммарная относительная погрешность будет минимальной, наиболее эффективный.

Поскольку набор методов импутации достаточно велик, то тестировать можно те методы, которые реализованы в пакете обработки данных, используемом исследователем, поскольку проверка эффективности алгоритмов данных без применения программного обеспечения — трудоемкая задача.

**Пример использования методики выбора наиболее эффективного алгоритма импутации в социологическом исследовании.** Используем описанный выше алгоритм для импутации пропущенных значений в массиве данных, содержащем информацию по уровню безработицы в странах мира (см. рисунок). Конечной целью исследования являлась кластеризация стран по уровню безработицы, поэтому для корректного разбиения необходимо восстановить все пропуски.



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Рейтинг стран по уровню безработицы в 1980–2018 гг., % от общей численности трудоспособного населения | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | Страна | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 |
| 3 | Австралия | 6,9 | 9,6 | 10,7 | 10,9 | 9,7 | 8,5 | 8,5 | 8,4 | 7,7 | 6,9 | 6,3 | 6,8 | 6,4 | 5,9 | 5,4 | 5 | 4,8 | 4,4 | 4,3 | 5,6 | 5,2 | 5,1 | 5,2 | 5,7 | 6,1 |
| 4 | Австрия | 2,7 | 3,2 | 3,3 | 4 | 3,9 | 4,2 | 4,7 | 4,8 | 4,7 | 4,1 | 3,9 | 4 | 4,4 | 4,9 | 5,5 | 5,7 | 5,2 | 4,9 | 4,1 | 5,3 | 4,8 | 4,6 | 4,9 | 5,3 | 5,6 |
| 5 | Азербайджан | - | - | - | - | - | - | - | - | - | - | 11,8 | 10,9 | 10 | 9,2 | 9 | 7,3 | 6,6 | 6,3 | 5,9 | 5,7 | 5,6 | 5,4 | 5,2 | 5 | 4,9 |
| 6 | Албания | 8,5 | 8,9 | 20,5 | 22,3 | 18,4 | 12,9 | 12,3 | 14,9 | 17,7 | 18,4 | 16,8 | 16,4 | 15,8 | 15 | 14,4 | 14,1 | 13,8 | 13,4 | 13,1 | 13,8 | 14 | 14 | 13,4 | 16 | 17,5 |
| 7 | Алжир | 19,8 | 20,3 | 21,4 | 23,2 | 24,4 | 28,1 | 28 | 28 | 28 | 29,3 | 29,5 | 27,3 | 25,7 | 23,7 | 17,7 | 15,3 | 12,5 | 13,8 | 11,3 | 10,2 | 10 | 10 | 11 | 9,8 | 10,6 |
| 8 | Аргентина | 7,6 | 6,5 | 7,1 | 11,6 | 13,3 | 18,9 | 18,8 | 16,8 | 14,8 | 16,1 | 17,1 | 19,2 | 22,5 | 17,3 | 13,6 | 11,6 | 10,2 | 8,5 | 7,9 | 8,7 | 7,8 | 7,2 | 7,2 | 7,1 | 7,3 |
| 9 | Армения | - | - | - | - | - | - | - | - | - | - | - | 38,4 | 35,3 | 31,2 | 31,6 | 31,2 | 27,8 | 28,7 | 16,4 | 18,7 | 19 | 18,4 | 17,3 | 16,2 | 17,6 |
| 10 | Аруба | - | - | - | - | - | - | - | - | 3,3 | 4,8 | 6,9 | 6,5 | 8,1 | 11,4 | 9,5 | 8,8 | 9,3 | 5,7 | 6,9 | 10,3 | 10,6 | 8,9 | 9,6 | 7,6 | 7,5 |
| 11 | Багамские острова | 12 | 12,3 | 14,8 | 13,1 | 13,3 | 10,9 | 11,5 | 9,8 | 7,7 | 7,8 | 7 | 6,9 | 9,1 | 10,8 | 10,2 | 10,2 | 7,8 | 7,9 | 8,7 | 14,2 | 15,1 | 15,9 | 14,4 | 15,8 | 14,8 |
| 12 | Барбадос | 14,9 | 17,2 | 22,9 | 24,4 | 21,8 | 19,6 | 15,8 | 14,6 | 12,2 | 10,4 | 9,8 | 9,9 | 10,3 | 11 | 9,6 | 9,1 | 8,7 | 7,4 | 8,1 | 10 | 10,3 | 11,2 | 11,5 | 11,6 | 12,3 |
| 13 | Бахрейн | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 5,6 | 3,7 | 4 | 3,6 | 4 | 3,7 | 4,4 | 3,8 |
| 14 | Белиз | 15 | 14 | 12 | 10 | 9 | 12,5 | 13,8 | 12,7 | 14,3 | 12,8 | 11,4 | 9,1 | 10 | 12,9 | 11,6 | 11 | 9,4 | 8,5 | 8,2 | 13,1 | 13,5 | 14 | 14,4 | 11,7 | 11,1 |
| 15 | Белоруссия | - | 0,1 | 0,5 | 1,4 | 2,1 | 2,9 | 4 | 2,8 | 2,3 | 2,1 | 2,1 | 2,3 | 2,7 | 3,1 | 2,5 | 1,7 | 1,4 | 1,1 | 0,9 | 0,9 | 0,8 | 0,7 | 0,6 | 0,5 | 0,5 |
| 16 | Бельгия | 6,6 | 6,5 | 7,1 | 8,6 | 9,8 | 9,7 | 9,6 | 9,2 | 9,3 | 8,4 | 6,9 | 6,6 | 7,5 | 8,2 | 8,4 | 8,5 | 8,3 | 7,5 | 7 | 7,9 | 8,3 | 7,1 | 7,5 | 8,4 | 8,5 |
| 17 | Болгария | 2,9 | 6,8 | 13,2 | 15,8 | 14,1 | 11,4 | 11 | 14 | 12,4 | 13,8 | 18,1 | 17,5 | 17,4 | 13,9 | 12,2 | 10,2 | 9 | 6,9 | 5,7 | 6,9 | 10,3 | 11,4 | 12,4 | 13 | 11,5 |

Фрагмент таблицы с исходными данными

Для поиска наиболее эффективного метода из таблицы были исключены страны, для которых отсутствовала информация по уровню безработицы за тот или иной год. В массиве комплектных данных случайным образом были сгенерированы пропуски. Для генерации номеров пропущенных значений в каждом столбце таблицы использовался инструмент «Анализ данных» в MS Excel.

Пропущенные значения поочередно предсказывались с использованием следующих методов: замены средним, множественной импутации, регрессионного анализа, алгоритма ZET. В качестве программного обеспечения использовался статистический пакет SPSS.

Суммарные относительные погрешности для каждого метода, %:

- замена средним — 115 %;
- множественная импутация — 89 %;
- регрессионный анализ — 32 %;
- алгоритм ZET — 42 %.

Таким образом, наиболее эффективным методом восстановления пропусков для данной задачи оказался метод регрессионного анализа, который в дальнейшем был использован для импутации фактически пропущенных значений.

## ЛИТЕРАТУРА


[1] Литтл Р.Дж.А., Рубин Д.Б. *Статистический анализ данных с пропусками*. Москва, Финансы и статистика, 1990, 336 с.

[2] Злоба Е., Яцкие И. Статистические методы восстановления пропущенных данных. *Computer Modelling & New Technologies*, 2002, vol. 6, no. 1, pp. 51–61.

[3] Зангиева И.К. Проблема пропусков в социологических данных: смысл и подходы к решению. *Социология: методология, методы, математическое моделирование*, 2011, № 33, с. 28–56.

[4] Фабрикан М.С. Практики сбора и анализа формализованных данных. *Социология: методология, методы, математическое моделирование*, 2015, № 41, с. 7–29.

[5] Абраменкова И.В., Круглов В.В. Методы восстановления пропусков в массивах данных. *Программные продукты и системы*, 2005, № 2, с. 4.

[6] Silva-Ramírez E.-L., et al. Missing value imputation on missing completely at random data using multilayer perceptrons. *Neural Networks*, 2011, vol. 24, iss. 1, pp. 121–129.

[7] Орлова И.В., ред. *Многомерный статистический анализ в экономических задачах: компьютерное моделирование в SPSS.* Москва, Вузовский учебник, 2009, 309 с.

[8] 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.

[9] Загоруйко Н.Г., Елкина В.Н., Тимеркаев В.С. Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм ZET). *Эмпирическое предсказание и распознавание образов*, 1975, вып. 61: Вычислительные системы, с. 3–27.

[10] Снитюк В.Е. Эволюционный метод восстановления пропусков в данных. *Сборник трудов VI Международной конференции «Интеллектуальный анализ информации». Киев, НТУУ «КПИ», 16–19 мая 2006 г.* Киев, 2006, с. 262–271.

[11] Silva-Ramírez E.-L., Pino-Mejías R., Lopez-Coello M. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, 2015, no. 29, pp. 128–132.

[12] Yoon S., Lee S. Training algorithm with incomplete data for feed-forward neural networks. *Neural Processing Letters*, 1999, no. 10 (3), pp. 171–179.

[13] Бююль А., Цефель П. *SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей.* Санкт-Петербург, ДиаСофтЮП, 2005, 608 с.

[14] Таганов Д.Н. *Статистический анализ в маркетинговых исследованиях*. Санкт-Петербург, Питер, 2005, 192 с.

[15] Наследов А.Д. *SPSS — Компьютерный анализ данных в психологии и социальных науках.* Санкт-Петербург, Питер, 2005, 416 с.

[16] Крыштановский А.О. *Анализ социологических данных с помощью пакета SPSS*. Москва, Изд. дом ГУ ВШЭ, 2006, 281 с.

[17] *Организация массовых данных и алгоритмы выявления выбросов.* URL: https://studme.org/120986/matematika_himiya_fizik/organizatsiya_massovyh_dannyh_algoritmy_vyyavleniya_vybrosov#aftercont (дата обращения 06.05.2019).

Статья поступила в редакцию 12.08.2019

Ссылку на эту статью просим оформлять следующим образом:

Фомина Е.Е. Обзор методов и программного обеспечения для восстановления пропущенных значений в массивах социологических данных. *Гуманитарный вестник*, 2019, вып. 4. http://dx.doi.org/10.18698/2306-8477-2019-4-611

**Фомина Елена Евгеньевна** — канд. техн. наук, доцент кафедры «Информатика и прикладная математика» Тверского государственного технического университета. e-mail: f-elena2008@yandex.ru

# Review of software and methods for recovering missing values in sociological data sets

© E.E. Fomina

Tver State Technical University, Tver, 170026, Russia

*Analysis of sociological data involves the study of large arrays of variables that may contain missing values. The presence of a significant number of incomplete records leads to distortion of the statistical analysis results, misinterpretation of modeling results. The article surveys software and mathematical methods designed to fill in the gaps in the data sets during sociological research. The mathematical nature, advantages and disadvantages of the most common methods of restoration of omissions used in solving practical problems are considered. An overview of modern software used to solve such problems is presented. A method for selecting the most efficient imputation algorithm was proposed.*

***Keywords:*** *imputations of data, missing data restoration, incomplete observations*

## REFERENCES

[1] Little R.J.A., Rubin D.B. *Statistical analysis with missing data*. New York, Wiley J. and Sons Publ., 1987, 278 p. [In Russ.: Little R.J.A., Rubin D.B. Statisticheskiy analis dannykh s propuskami. Moscow, Finansy i statistika Publ., 1990, 336 p.].

[2] Zloba E., Yatskie I. *Computer Modeling & New Technologies*, 2002, vol. 6, no. 1, pp. 51–61.

[3] Zangieva I.K. *Sotsiologiya: metodologiya, metody, matematicheskoe modelirovanie — Sociology: methodology, methods, mathematical modeling*, 2011, no. 33, pp. 28–56.

[4] Fabrikan M.S. *Sotsiologiya: metodologiya, metody, matematicheskoe modelirovanie — Sociology: methodology, methods, mathematical modeling*, 2015, no. 41, pp. 7–29.

[5] Abramenkova I.V., Kruglov V.V. *Programmnye produkty i sistemy — Software and systems*, 2005, no. 2, p. 4.

[6] Silva-Ramírez E.-L. et al. *Neural Networks*, 2011, vol. 24, no. 1, pp. 121–129.

[7] Orlova I.V., ed. *Mnogomernyy statisticheskiy analiz v ekonomicheskikh zadachakh: komputernoe modelirovanie v SPSS* [Multivariate statistical analysis in economic problems: computer modeling in SPSS]. Moscow, Vuzovskiy uchebnik Publ., 2009, 309 p.

[8] Pimonov A.G., Glebova E.A., Sarapulova T.V., Glebov V.V. *Ekonomika i upravlenie innovatsiyami — Economics and innovation management*, 2017, no. 3, pp. 52–66.

[9] Zagoruyko N.G., Elkina V.N., Timerkaev V.S. Algoritm zapolneniya propuskov v empiricheskikh tablitsakh (algoritm ZET) [Algorithm for filling gaps in empirical tables (algorithm ZET)]. In: *Empiricheskoe predskazanie i raspoznavanie obrazov. Sbornik trudov* [Empirical prediction and pattern recognition. Collected works], Novosibirsk, 1975, no. 61: Vychislitelnye sistemy [Computer systems], pp. 3–27.

[10] Snituk V.E. Evolutsionnyy metod vosstanovleniya propuskov dannykh [Evolutionary method of restoration of omissions in data]. *Sbornik trudov*

*VI Mezhdunarodnoy konferentsii "Intellektualnyy analiz informatsii, Kiev, Nationalnyy tekhnicheskiy universitet Ukrainy "Kievskiy politekhnicheskiy institute", 16–19 maya 2006 g.* [Proceedings of the VI-th International Conference «Intellectual analysis of information», Kyiv, National technical University of Ukraine "Kyiv Polytechnic Institute", May 16–19, 2006]. Kiev, 2006, pp. 262–271.

[11] Silva-Ramírez E.-L., Pino-Mejías R., Lopez-Coello M. *Applied Soft Computing*, 2015, no. 29, pp. 128–132.

[12] Yoon S., Lee S. *Neural Processing Letters*, 1999, no. 10 (3), pp. 171–179.

[13] Büyül A., Zefeldt P. *SPSS: iskusstvo obrabotki informatsii. Analiz ststisticheskikh dannykh i vosstanovlenie skrytykh zakonomernostey* [SPSS: the art of information processing. The analysis of statistical data and restoring hidden patterns]. St. Petersburg, DiaSoftYup Publ., 2005, 608 p. [In Russ.]

[14] Taganov D.N. *Statisticheskiy analiz v marketingovykh issledovaniyakh* [Statistical analysis in marketing research]. St. Petersburg, Peter Publ., 2005, 192 p.

[15] Nasledov A.D. *SPSS — komputernyy analiz dannykh v psikhologii i sotsialnykh naukakh* [SPSS — Computer analysis of data in psychology and social sciences]. St. Petersburg, Peter Publ., 2005, 416 p.

[16] Kryshtanovsky A. O. *Analiz sotsiologicheskikh dannykh s pomoshchyu paketa SPSS* [Analysis of sociological data using SPSS package]. Moscow, GU VShE Publ., 2006, 281 p.

[17] *Organizatsiya massovykh dannykh i algoritmy vyyavleniya vybrosov* [Organization of mass data and algorithms for detecting outliers]. Available at: https://studme.org/120986/matematika_himiya_fizik/organizatsiya_massovyh_dannyh_algoritmy_vyyavleniya_vybrosov#aftercont (accessed May 6, 2019).

**Fomina E.E.**, Cand. Sc. (Eng.), Assoc. Professor, Department of Informatics and Applied Mathematics, Tver State Technical University. e-mail: f-elena2008@yandex.ru