УДК 303.4.025

И. К. Зангиева, Е. С. Тимонина

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ ЗАПОЛНЕНИЯ ПРОПУСКОВ В ДАННЫХ В ЗАВИСИМОСТИ ОТ ИСПОЛЬЗУЕМОГО МЕТОДА АНАЛИЗА

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ АЛГОРИТМОВ ЗАПОЛНЕНИЯ ПРОПУСКОВ В ДАННЫХ В ЗАВИСИМОСТИ ОТ ИСПОЛЬЗУЕМОГО МЕТОДА АНАЛИЗА

ЗАНГИЕВА Ирина Казбековна – кандидат социологических наук, старший преподаватель и заместитель заведующего кафедрой методов сбора и анализа социологической информации НИУ-ВШЭ. E-mail: zangieva.irina@gmail.com
ТИМОНИНА Елизавета Сергеевна – выпускница бакалавриата факультета социологии НИУ-ВШЭ (2013), студентка магистратуры факультета социологии НИУ-ВШЭ. E-mail: eliza.tim.92@gmail.com

Аннотация. Статья посвящена описанию проведенного исследования, направленного на выявление наиболее эффективного алгоритма заполнения пропусков в данных для таких методов анализа, как регрессионное моделирование, факторный анализ, описательная статистика, расчет коэффициента корреляции. Ставится проблема неоднозначности выбора в каждой конкретной исследовательской ситуации алгоритма заполнения пропусков ввиду недостатка в современной науке обоснованных рекомендаций по их выбору.
Авторы утверждают, что алгоритм заполнения пропусков должен быть отобран исходя из последующего метода анализа заполненных от пропусков данных. Другими словами, авторы полагают, что для каждого метода анализа данных эффективность одного и того же метода заполнения пропусков будет различна. Оценить эффективность нескольких алгоритмов заполнения для каждого метода анализа данных удалось с помощью статистического эксперимента.
Суть статистического эксперимента заключалась в сравнении результатов применения каждого метода на эталонном (без пропусков) массиве с

COMPARING IMPUTATION ALGORITHMS EFFICIENCY RESPECTIVE TO THE DATA ANALYSIS METHODS

ZANGIEVA Irina Kazbekovna - PhD in Sociology, Senior Lecturer and Deputy Head at the Department of Sociological Research Methods, National Research University «Higher School of Economics»; e-mail: zangieva.irina@gmail.com
TIMONINA Elizaveta Sergeevna – Bachelor at the Faculty of Sociology, National Research University «Higher School of Economics» (2013); Master`s Degree student at the Faculty of Sociology, National Research University «Higher School of Economics»; e-mail: eliza.tim.92@gmail.com

Abstract. The paper describes a recent study aimed at investigating the most efficient data imputation algorithm for several methods of data analysis such as regression modeling, factor analysis, descriptive statistics, and correlation analysis. The lack of recommendations when choosing the data imputation algorithm poses the problem of choice ambiguity in each situation.
The authors consider that the data imputation algorithm should be selected according to the method employed after data improvement. In other words, it is believed that for each data analysis method the efficiency of the same data imputation algorithm is different. The statistical experiment was used to evaluate the efficiency of several data imputation algorithms for each method of data analysis.
The core idea of statistical experiment was to compare the results of each method application used in the etalon data set (without missing values) with the results obtained on a large number of artificial subsamples generated from the original data set where missing values were filed with comparable data imputation algorithms. Generation of subsamples was carried out via the bootstrap procedure, which allowed to undertake

результатами, полученными на большом количестве искусственно созданных из исходного массива подвыборок, пропуски в которых были заполнены несколькими алгоритмами. Для статистической оценки генерация подвыборок была проведена с помощью процедуры bootstrap, что позволило оценить доверительные интервалы для каждого показателя до и после заполнения пропусков.
В результате эксперимента удалось оценить эффективность таких алгоритмов заполнения пропусков, как заполнение мерами средней тенденции, EM-алгоритм, заполнение с помощью регрессии и Hot Deck алгоритм для уже указанных методов анализа данных.

Ключевые слова: пропуски в данных, неответ, алгоритмы заполнения пропусков, статистический эксперимент, bootstrap, бутстреп.

statistical evaluation and to build confidence intervals for each parameter before and after the data imputation.
Through this experiment the authors managed to evaluate the efficiency of such data imputation algorithms as imputation with the average trend measures, the EM algorithm, the imputation via regression model and Hot Deck algorithm for the mentioned methods of data analysis.

Keywords: missing values; item nonresponse; data imputation algorithms; statistical experiment; bootstrap.

В информации, собранной в ходе количественных социологических исследований, зачастую наблюдаются пропуски данных. Причиной пропусков может быть как недостижимость респондентов, так и неответы на вопросы анкеты. Если с первым типом пропусков исследователь после сбора данных не может сделать ничего, кроме попытки повторно связаться с респондентом, то со вторым типом пропусков можно работать и после завершения полевого этапа.

Американские статистики Р. Литтл и Д. Рубин выделили следующие типы пропусков. Полностью случайные (missing completely at random; MCAR): вероятность появления полностью случайных пропусков в данных не зависит ни от значений самого измеряемого признака, ни от значений других признаков. Частично случайные (missing at random; MAR) – вероятность появления частично случайных пропусков в данных не зависит от значений самого измеряемого признака, но обусловлена значением другого признака. Например, неответ на вопрос о возрасте (исходный признак) зависит от пола (другой признак). Неслучайные (not missing at random) – вероятность появления неслучайных пропусков в данных зависит от значения самого признака. Например, на вопрос о возрасте не отвечают респонденты более старших возрастов [1].

По мнению авторов, заполнение неслучайных пропусков возможно лишь при глубоком понимании характера влияния признака на неответ, в противном случае заполнение может привести к значительным смещениям результатов.

Самая простая стратегия работы с данными, содержащими пропуски, – это удаление пропусков или их игнорирование и анализ только полных наблюдений, иными словами, это всегда потеря определенного количества данных. Другой способ корректировки предлагает взвешивание доступных данных для искусственного достижения запланированного объема выборки, когда каждому полному наблюдению присваивается определенный весовой коэффициент. В большинстве подобных случаев конечные результаты расчетов существенно смещены. Наиболее распространенная сегодня среди социологов стратегия работы – заполнение пропусков различными методами. Преимущество этого способа неоспоримо: он позволяет сохранить исходное количество данных и получить вместо пропусков значения, близкие к реальности.

При заполнении пропусков перед исследователем возникают вопросы: какой из нескольких десятков существующих алгоритмов выбрать? будет ли это заполнение пропусков модальным, средним или медианным значением признака? Помимо указанных простых алгоритмов заполнения мерами средней тенденции, разработаны алгоритмы расчета пропущенного значения через построение регрессионной модели по другой переменной или посредством подбора значения на место пропуска на основе максимального сходства с имеющимися и др.

Обзор публикаций по этой теме показал, что в настоящее время большинство работ, посвященных сравнению эффективности различных алгоритмов заполнения пропусков в данных, основываются на *теоретическом* сравнении математической базы используемых алгоритмов [1, 7, 10]. Мы же полагаем, что основным критерием выбора конкретного алгоритма заполнения пропусков должен быть последующий метод анализа: скорее всего, для одного метода анализа более эффективным окажется один алгоритм, для другого – другой. Однако при современном уровне развития методов анализа данных и разработки алгоритмов заполнения пропусков теоретически соотносить их математический аппарат удается далеко не всегда. При этом исследований, направленных на *эмпирическое* сравнение эффективности алгоритмов заполнения пропусков в данных, мало. В тех же работах, где осуществлялось подобное сравнение, не учитывался используемый метод анализа данных, а также не проводилась статистическая верификация выводов. На наш взгляд, выбрать конкретный алгоритм можно на базе результатов эмпирического сравнения результатов анализа, полученных на заполненных различными способами данных, с результатами, полученными на исходном полном массиве – эталоне, в ходе статистического эксперимента. Однако подобное однократное сравнение даст обоснованные рекомендации только применительно к конкретной исследовательской ситуации, а нашей целью является выбор способа заполнения пропусков, оправданного с определенной статистической вероятностью. Мы полагаем, что выбрать наиболее эффективный алгоритм заполнения пропусков для определенного метода анализа можно посредством статистического эксперимента. В частности, в нашем исследовании эффективность алгоритмов заполнения пропусков сравнивалась при использовании наиболее распространенных методов анализа данных: расчета коэффициента корреляции, регрессионных моделей, факторного анализа, показателей описательной статистики.

Обозначим этапы проведенного эксперимента:

1 Создание эталонного массива данных (без пропусков по содержательным переменным).
2 Решение содержательных задач на эталонном массиве данных с помощью описанных методов анализа данных – получение эталонного результата и построение эталонного доверительного интервала для коэффициента корреляции, регрессионных коэффициентов, среднего и дисперсии собственного значения полученных факторов.
3 Внесение в эталонный массив разного количества полностью случайных пропусков по содержательным переменным – создание массивов с пропусками в данных.
4 Заполнение внесенных на шаге 3 пропусков отобранными алгоритмами заполнения пропусков.
5 Для статистической верификации полученных результатов решение содержательной задачи на большом числе выборок, сгенерированных с

помощью бутстрепа из массивов с заполненными пропусками; получение доверительных интервалов для значения коэффициента корреляции, регрессионных коэффициентов, среднего и дисперсии собственных значений факторов на *уже заполненных от пропусков данных*.

6 Сравнение полученных на шаге 5 результатов с эталонным результатом (шаг 2).

7 На основании введенных критериев выбор наиболее эффективного алгоритма заполнения пропусков для каждого метода анализа данных.

Процедура бутстрепа – наименее трудоемкая процедура генерации случайных событий. Ее суть заключается в создании некоторого количества выборок из массива данных той же полноты, что и исходный массив, при этом в каждую выборку может несколько раз попасть один и тот же человек. Таким образом, бутстреп – это генерация выборок с возвращением.

Бутстреп является одним из методов непараметрической статистической оценки и используется в тех случаях, когда априори неизвестно, какое теоретическое распределение выбирать для статистической оценки какого-либо параметра. Генерация выборок с помощью бутстрепа позволяет построить доверительный интервал изменения какого-либо параметра с определенной вероятностью

С помощью бутстрепа из исходного массива создается различное, заданное исследователем количество случайных выборок *с возвращением*, совпадающих по объему с исходным массивом: предполагается, что в одну выборку может несколько раз попасть один и тот же респондент. Процедура бутстрепа была предложена Б. Эфроном в качестве альтернативы методике Монте-Карло [6], ее суть заключается в многократном извлечении случайных выборок из исходной выборки.

Реализуется процедура бутстрепа в такой последовательности:

1 Из исходной выборки создаются наборы случайных выборок с возвращением исходной полноты: «берется конечная совокупность из n членов исходной выборки x_1, x_2... x_{n-1}, x_n , откуда на каждом шаге из n последовательных итераций с помощью датчика случайных чисел, равномерно распределенных на интервале [1, n], вытягивается произвольный элемент x_k, который снова возвращается в исходную выборку (т.е. может быть извлечен повторно)» [2, с. 13].

2 Для каждой выборки рассчитываются такие значения анализируемой характеристики, как медиана, среднее, стандартная ошибка, дисперсия, коэффициент корреляции, регрессионные коэффициенты. На основе полученных расчетов строится гистограмма распределения значения признака и доверительный интервал изменения этого признака.

На основе разброса значений параметра переменной, полученного в процессе генерации выборок, бутстреп позволяет рассчитать стандартную ошибку изменения указанного параметра и построить доверительный интервал на основе эмпирического распределения полученных значений [9, p. 577].

Как уже отмечалось, процедура бутстрепа – это создание выборок с возвращением, что обычно считается недостатком метода. Однако бутстреп является наименее трудоемким методом рандомизации случайных событий по сравнению с методикой Монте-Карло, а его

недостаток пытаются элиминировать созданием большого количества выборок, чтобы добиться их бо́льшей разнородности.

Анализ литературы, в которой различные способы заполнения пропусков сравнивались посредством создания некоторого количества выборок (не всегда с помощью процедуры бутстрепа), обнаружил, что это количество изменяется в пределах от 100 до 100 000 [3, 5, 9, 11, 12]. Мы остановились на создании 100 000-ных выборок. Наше предположение заключалось в том, что с увеличением количества выборок границы доверительного интервала измеряемых показателей (в рамках исследования – значения коэффициента корреляции, регрессионных коэффициентов, среднего и дисперсии значений факторов) будут более близки к значению этих показателей в генеральной совокупности, а значит, более правдивы.

Построение доверительного интервала указанных величин с помощью методики бутстреп позволило статистически оценить выбор того или иного алгоритма заполнения пропусков в качестве наиболее эффективного для различных методов анализа данных.

Для выбора наиболее эффективного алгоритма заполнения пропуска на основании результатов бутстрепа были введены два критерия. Первый критерий был назван **степенью отклонения доверительных интервалов (Δ)**.

$$\Delta = \frac{|x_e - x_n| + |y_e - y_n|}{y_e - x_e} \times 100\% \quad (1)$$

,

где xe – нижняя граница эталонного доверительного интервала;
xn – нижняя граница доверительного интервала, полученная в ходе заполнения пропусков;
ye – верхняя граница эталонного доверительного интервала;
ye – верхняя граница доверительного интервала, полученная в ходе заполнения пропусков.

В числителе мы получали абсолютное значение отклонения доверительного интервала после заполнения от эталонного доверительного интервала, модуль использовался для предотвращения обращения числителя в 0 в том случае, когда отклонение нижних границ равно отклонению верхних границ, но с противоположным знаком. В знаменателе – длина эталонного доверительного интервала. Соответственно, введенный критерий изменяется в пределах от 0 до бесконечности, принимает значение 0, когда границы доверительного интервала после заполнения полностью совпадают с эталонными, и исчисляется в %. Таким образом, чем ближе этот показатель к 0, тем меньше степень отклонения интервалов, тем ближе результат после заполнения к эталонному.

Второй критерий – устойчивость степени отклонения доверительных интервалов на разном количестве созданных с помощью бутстрепа выборок. С одной стороны, чем больше количество выборок, тем более надежен результат (что следует из закона больших чисел), а с другой – бо́льшее количество выборок предполагает их бо́льшую однородность, так как бутстреп – это техника создания выборок с возвращением. Было решено рассматривать также устойчивость степени отклонения доверительных интервалов на разном количестве выборок, предполагается, чем она устойчивее, тем более надежна. Устойчивость в данном случае будет проявляться в получении одинаковых результатов на разном количестве выборок, созданных с помощью бутстрепа.

Наиболее простые методы единичного (single imputation) заполнения пропусков связаны с заполнением пропусков мерами средней тенденции, такими как выборочная мода, медиана, среднее значение. Обычно придерживаются заполнения модой и медианой, когда данные с пропусками измерены на номинальном или порядковом уровнях, а среднее значение используют также для заполнения пропусков в интервальных переменных.

К более сложным методам единичного заполнения пропусков можно отнести заполнение с помощью алгоритма Hot Deck. Его суть состоит в подборе значения на место пропущенного исходя из распределения ответов. Другими словами, алгоритм заполняет пропущенное значение (реципиент) конкретного респондента значением того респондента (донор), чьи ответы наиболее похожи на ответы респондента с пропущенным значением [4, p. 643]. Предполагается, что результаты опроса вводятся в массив в определенном порядке и в некоторых исследованиях на основе подобного временного эффекта может быть вычислено пропущенное значение [14]. С математической точки зрения, алгоритм высчитывает расстояние от пропущенного значения до каждого полного наблюдения, после чего заполняет пропуск тем значением, расстояние до которого минимально. Меру расстояния определяет сам исследователь исходя из конкретной задачи исследования и характера связи между переменными (чаще всего используется Евклидово расстояние и его квадрат). Алгоритм Hot Deck реализован в пакете SOLAS и позволяет заполнять пропуски в данных, измеренных на разных уровнях.

Заполнение пропусков с помощью модели регрессии реализуется в два этапа: на первом строится регрессионное уравнение, зависимой переменной в котором выступает переменная с пропущенным значением, и анализируется полученная модель; на втором в полученное регрессионное уравнение подставляются значения по полным независимым переменным, на основании чего пропуск по зависимой переменной заполняется предсказанным по уравнению значением [13, p. 533]. Обычно для заполнения пропусков используют линейную регрессию, которая предполагает интервальный уровень измерения как зависимой переменной, так и независимых. Заполнение с помощью регрессии может быть реализовано только в случае интервального уровня измерения переменной с пропущенным значением.

EM-алгоритм заполняет пропуски только в интервальных данных и представляет итеративную процедуру, которая, как и регрессионное моделирование, реализуется в два этапа [13, p. 534]:

1 На первом этапе E (expectation) на полных наблюдениях происходит расчет ожидаемого значения переменной с пропуском, оцениваются ковариационные матрицы, меры средней тенденции и взаимной корреляции переменной с пропусками и других переменных в массиве, в соответствии с чем пропуски заполняются.

2 На этапе M (maximization) происходит итеративное оценивание полученных на шаге E-матриц ковариаций и вектора средних значений до тех пор, пока степень соответствия между ожидаемыми и подставляемыми данными не будет максимальна. Итерационный процесс продолжается до тех пор, пока разница между ковариационными матрицами, рассчитанными последовательно на шаге M, будет минимальна.

Активно развиваются методы множественного заполнения пропусков. Их суть состоит в заполнении пропуска сразу несколькими рассчитанными значениями, наличие которых говорит о неопределенности причин возникновения пропусков. Множественное заполнение обычно реализуется в три этапа [8, p. 244]:

1 На первом этапе исследователь получает заполненные несколькими значениями пропуски, которые могут быть сохранены как отдельные наборы в разных массивах либо в одном массиве размера $k \times n$, где k – количество заполнений пропущенных значений, а n – исходное количество данных в массиве, который на втором этапе анализируется с весовым коэффициентом $1/k$.

2 На втором этапе полученные наборы/массив анализируются разными методами анализа как полные наблюдения.

3 На третьем шаге полученные на каждом шаге результаты анализируются в совокупности, что позволяет принять во внимание неопределенность природы пропусков.

Итак, выбор того или иного алгоритма заполнения пропусков для каждого метода анализа данных зависел от уровня измерения содержательных переменных, на которых эти методы были применены. Для коэффициента корреляции и описательной статистики, рассчитываемых на интервальных данных, были отобраны 6 алгоритмов: традиционные методы заполнения пропусков мерами средней тенденции (мода, медиана, среднее), заполнение с помощью регрессионного моделирования и EM-алгоритм, а также Hot Deck алгоритм. Для регрессии было отобрано только заполнение модой и с помощью Hot Deck алгоритма, поскольку независимые переменные были измерены на номинальном и порядковом уровнях.

На схеме 1 представлены выбранные алгоритмы заполнения пропусков для каждого используемого метода анализа данных.

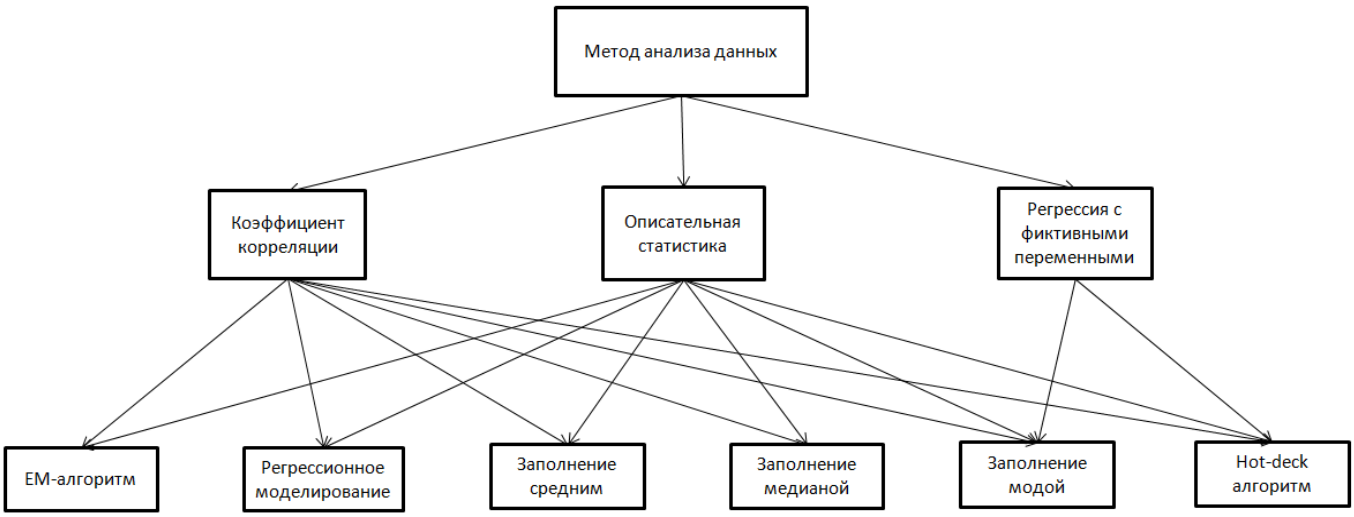

Рисунок 1 – Выбранные алгоритмы заполнения пропусков в данных для каждого метода анализа

Создание эталонного массива и получение эталонных результатов

Чтобы эмпирически сравнить результаты анализа, полученные на заполненных различными способами данных, с результатами, полученными на исходном полном массиве – эталоне, в ходе статистического эксперимента с помощью указанных методов на массиве данных 5-й волны Европейского социального исследования (European Social Survey)[1] были решены содержательные задачи, посвященные трудовым ценностям россиян. В свою очередь, это потребовало решения следующих задач:

1 Определение типа *потенциальной* работы, который стоит за важностью тех или иных ценностей работы при трудоустройстве россиян (факторный анализ).

– Определение влияния текущего типа работы россиян на выраженность того или иного типа *потенциальной* работы:
– определение типа текущей работы (факторный анализ).

2 Поиск связи между типом текущей работы и выраженностью того или иного типа работы потенциальной (коэффициент корреляции Пирсона).

3 Определение влияния социально-демографических параметров на выраженность типа потенциальной работы: поиск связи между полом и образованием и выраженностью того или иного типа потенциальной работы (регрессия с фиктивными переменными).

Таким образом, получение эталонных результатов в ходе решения перечисленных задач потребовало применения следующих методов анализа данных: факторный анализ; поиск связи через коэффициент корреляции Пирсона для интервальных данных; регрессия с фиктивными переменными.

Анализ проводили на эталонном массиве – без пропусков по содержательным переменным, поэтому в рамках решения содержательных задач исходным массивом выступала не вся база данных ESS, а подвыборка работавших на момент опроса респондентов – таких в массиве 1357 человек.

После удаления из массива неполных наблюдений в исходном массиве осталось 1013 наблюдений без пропущенных значений. Иными словами, для получения эталонного массива из исходного пришлось удалить 25% наблюдений. Данное обстоятельство не является причиной для недоверия полученным на эталонном массиве содержательным результатам.

В итоге с помощью факторного анализа были выявлены типы текущей работы россиян, а также типы потенциальной работы – той, которую они бы предпочли при трудоустройстве (схема 2).

[1] http://www.europeansocialsurvey.org.

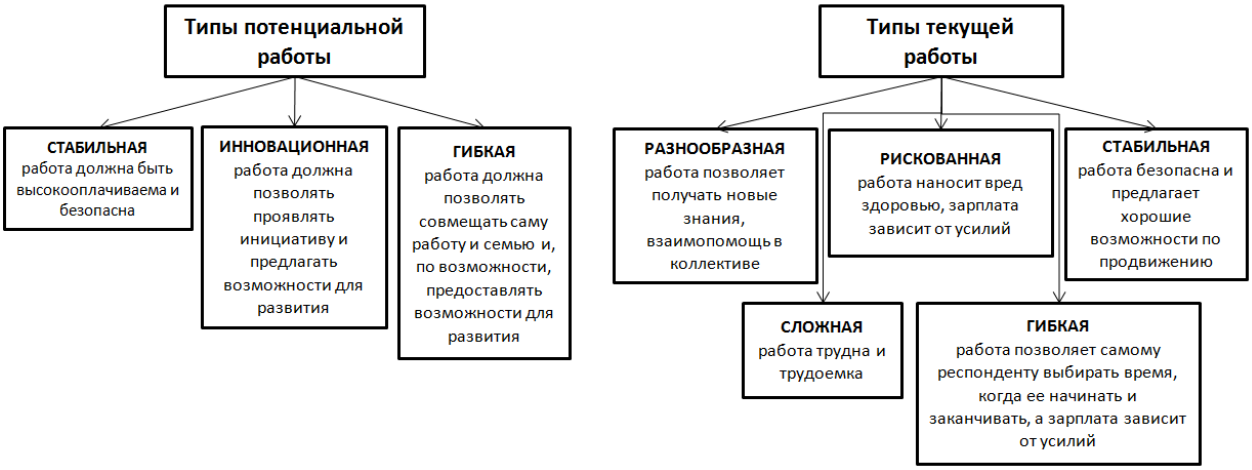

Рисунок 2 – Выявленные в ходе факторного анализа типы потенциальной и текущей работы россиян

Коэффициент корреляции Пирсона позволил выявить связь между полученными типами текущей и потенциальной работы: так, наиболее выраженная значимая связь (коэффициент корреляции 0,372) была выявлена между разнообразным типом текущей работы и инновационным типом потенциальной работы. Можно предположить, если текущая работа респондента разнообразна или требует определенного творческого подхода, вероятно, его требования при выборе места работы будут включать эти характеристики.

Регрессия с фиктивными переменными позволила определить влияние социально-демографических параметров респондентов на выявленные типы потенциальной работы. Так, пол и образование значимо влияют только на выраженность инновационного и гибкого типа потенциальной работы. Инновационный тип более выражен у мужчин и респондентов с высшим образованием, а гибкий тип – у женщин. Однако были получены достаточно низкие коэффициенты детерминации: скорее всего, выраженность того или иного типа потенциальной работы опосредована влиянием иных характеристик, нежели социально-демографических параметров, не учтенных в построенных моделях. Несмотря на это, вывод о связи между этими переменными может быть сделан.

Организация статистического эксперимента потребовала реализации следующих шагов:

1 Был сформирован эталонный массив данных путем удаления из исходного массива ESS всех наблюдений с пропусками по изучаемым переменным.

2 С помощью методов анализа (факторный анализ, регрессия с фиктивными переменными, коэффициент корреляции, расчет разброса и среднего показателя) на эталонных данных были получены эталонный результат и эталонные доверительные интервалы: выявлены типы текущей и потенциальной работы россиян (факторный анализ); установлено, как социально-демографические параметры влияют на выраженность выявленных типов потенциальной работы (регрессионный анализ); определена взаимосвязь между выявленными типами текущей работы и потенциальной (коэффициент корреляции Пирсона); были изучены меры средней тенденции и

разброса факторных нагрузок выявленных типов потенциальной и текущей работы (описательная статистика).

3 В рамках проведения статистического эксперимента в эталонный массив было внесено некоторое количество искусственных, полностью случайных пропусков, в результате чего создано *n* массивов с разным набором пропусков.

4 С помощью выбранных алгоритмов искусственно созданные пропуски были заполнены в каждом из *n* массивов.

5 На каждом из *n* массивов данных, где пропуски были заполнены каждым алгоритмом, воспроизведены расчеты с помощью процедуры бутстрепа на разном количестве выборок, посчитаны доверительные интервалы после заполнения пропусков отобранными алгоритмами.

6 Полученные в пункте 5 результаты сравнивались с эталонным результатом. В итоге был выбран наиболее эффективный способ заполнения пропусков для того или иного метода анализа на основании описанных ранее критериев.

Было решено остановиться на трех вариантах количества пропусков: 10% пропусков по каждой переменной, участвующей в анализе; 50% пропусков по каждой переменной, участвующей в анализе; количество пропусков, повторяющее структуру пропусков исходного массива данных.

В итоге были созданы 9 массивов, по 3 для каждого метода анализа: в первом 10% пропусков по каждой содержательной переменной, во втором – 50% пропусков, в третьем была воссоздана исходная структура пропусков. Кроме того, массивы содержали ряд вспомогательных переменных, необходимых для последующего заполнения пропусков такими методами, как EM-алгоритм, Hot Deck алгоритм и регрессионное моделирование.

Затем в полученных массивах пропуски были заполнены отобранными алгоритмами: мерами средней тенденции, с помощью модели регрессии, EM-алгоритмом, Hot Deck алгоритмом.

После заполнения созданных массивов с различным количеством пропусков описанными выше алгоритмами было получено 42 массива: для расчета коэффициента корреляции – 6 массивов, заполненных 6 разными методами (модой, средним, медианой, регрессионным моделирование, EM-алгоритмом и Hot Deck алгоритмом) для 10% пропусков, 6 массивов для 50% пропусков и 6 массивов для исходной структуры пропусков; для расчета описательной статистики также получено по 6 массивов для каждого количества пропусков; для регрессии с фиктивными переменными получено по 2 массива для каждого количества пропусков, заполненных с помощью моды и Hot Deck алгоритма.

Была решена содержательная задача, расчеты которой полностью повторяли расчеты, полученные на эталоне, но при этом была задействована процедура бутстрепа.

Для повышения надежности полученных результатов было решено реализовать для каждого из созданных 42 массивов 4 процедуры бутстрепа: с созданием 1000, 10 000 и 100 000 выборок. Таким образом, для статистической проверки эффективности различных алгоритмов заполнения пропусков было запущено 126 процедур бутстрепа, для каждого из 42 заполненных разными методами массивов по 42 процедуры бутстрепа для построения 1000 выборок, 10 000 и 100 000, полностью повторяющих расчеты коэффициента корреляции, описательной статистики и регрессии с фиктивными переменными, проведенные на эталонном массиве.

После запуска бутстрепа результаты, полученные после заполнения пропусков, сравнивали с результатами, полученными на эталонном массиве, и на основании введенных ранее критериев был выбран наиболее эффективный алгоритм заполнения пропусков для каждого метода анализа данных. Полученные результаты забивались в таблицу (на примере расчета бутстреп-выборок для коэффициента корреляции).

Таблица 1 Параметры бутстрепа при расчете коэффициента корреляции на эталонном массиве между инновационным типом потенциальной работы и разнообразным типом текущей работы в ситуации наличия связи

Количество выборок	Значение коэффициента корреляции	Уровень значимости	Стандартная ошибка	Нижняя граница доверительного интервала	Верхняя граница доверительного интервала	Длина эталонного доверительного интервала
1000	0,372	0,000	0,037	0,313	0,422	0,109
10 000	0,372	0,000	0,029	0,315	0,426	0,111
100 000	0,372	0,000	0,029	0,315	0,426	0,111

Таблица 2 Сводная таблица результатов заполнения пропусков различными алгоритмами для расчета коэффициента корреляции в ситуации наличия связи

Структура пропусков	Количество созданных выборок	Алгоритм заполнения пропусков	Значение коэффициента корреляции	Искажение (bias)	Стандартная ошибка	Нижняя граница доверительного интервала	Верхняя граница доверительного интервала	Степень отклонения доверительных интервалов, %
10% пропусков по каждой переменной	1000	С помощью моды	0,259	0,001	0,030	0,197	0,325	195
		С помощью медианы	0,326	-0,002	0,029	0,265	0,389	74
		С помощью среднего арифметического	0,327	0,000	0,030	0,264	0,385	79
		Hot Deck заполнение	0,306	0,000	0,029	0,253	0,364	108
		С помощью модели регрессии	0,327	0,001	0,031	0,260	0,389	79
		EM-алгоритм	**0,360**	**0,002**	**0,031**	**0,300**	**0,419**	**15**
	1000 0	С помощью моды	0,259	0,000	0,030	0,198	0,316	205
		С помощью медианы	0,326	0,000	0,030	0,266	0,382	84
		С помощью среднего арифметического	0,327	0,000	0,030	0,270	0,385	77
		Hot Deck заполнение	0,306	0,001	0,030	0,245	0,364	119
		С помощью модели регрессии	0,327	0,000	0,031	0,264	0,386	82
		EM-алгоритм	**0,360**	**0,000**	**0,031**	**0,299**	**0,421**	**19**
	100 000	С помощью моды	0,259	0,000	0,030	0,198	0,317	204
		С помощью медианы	0,326	0,000	0,030	0,266	0,383	83
		С помощью среднего арифметического	0,327	0,000	0,030	0,268	0,385	80
		Hot Deck заполнение	0,306	0,000	0,030	0,245	0,365	118

		С помощью модели регрессии	0,327	0,000	0,031	0,264	0,387	81
		ЕМ-алгоритм	**0,360**	**0,000**	**0,031**	**0,297**	**0,420**	**22**
50% пропусков по каждой переменной	1000	С помощью моды	0,103	-0,001	0,027	0,046	0,155	490
		С помощью медианы	0,172	0,000	0,032	0,107	0,231	364
		С помощью среднего арифметического	0,170	0,000	0,031	0,104	0,232	366
		Hot Deck заполнение	0,155	0,000	0,032	0,093	0,216	391
		С помощью модели регрессии	0,208	0,000	0,032	0,145	0,264	299
		ЕМ-алгоритм	**0,377**	**0,002**	**0,031**	**0,315**	**0,436**	**15**
	10 000	С помощью моды	0,103	0,000	0,028	0,046	0,156	486
		С помощью медианы	0,172	0,000	0,032	0,108	0,232	361
		С помощью среднего арифметического	0,170	0,000	0,032	0,108	0,231	362
		Hot Deck заполнение	0,155	0,000	0,031	0,093	0,214	391
		С помощью модели регрессии	0,208	0,000	0,031	0,148	0,269	292
		ЕМ-алгоритм	**0,377**	**0,000**	**0,030**	**0,315**	**0,434**	**7**
	100 000	С помощью моды	0,103	0,000	0,028	0,046	0,156	485
		С помощью медианы	0,172	0,000	0,032	0,108	0,233	360
		С помощью среднего арифметического	0,170	0,000	0,032	0,108	0,231	362
		Hot Deck заполнение	0,155	0,000	0,031	0,094	0,216	388
		С помощью модели регрессии	0,208	0,000	0,031	0,146	0,269	293
		ЕМ-алгоритм	**0,377**	**0,000**	**0,030**	**0,315**	**0,436**	**9**
Исходная структура пропусков	1000	С помощью моды	0,240	-0,001	0,030	0,181	0,300	233
		С помощью медианы	0,313	0,000	0,031	0,250	0,374	102
		С помощью среднего арифметического	0,315	0,000	0,032	0,251	0,374	101
		Hot Deck заполнение	0,297	0,000	0,030	0,237	0,358	128
		С помощью модели регрессии	0,311	0,001	0,032	0,249	0,374	103
		ЕМ-алгоритм	**0,369**	**-0,002**	**0,033**	**0,308**	**0,431**	**13**
	10 000	С помощью моды	0,240	0,000	0,030	0,180	0,298	237
		С помощью медианы	0,313	-0,001	0,031	0,252	0,372	105
		С помощью среднего арифметического	0,315	0,000	0,031	0,253	0,374	103

Перейдем непосредственно к описанию результатов экспериментальной проверки. На схеме 3 изображено, какой алгоритм заполнения пропусков оказался наиболее эффективным для каждого исследуемого метода анализа данных.

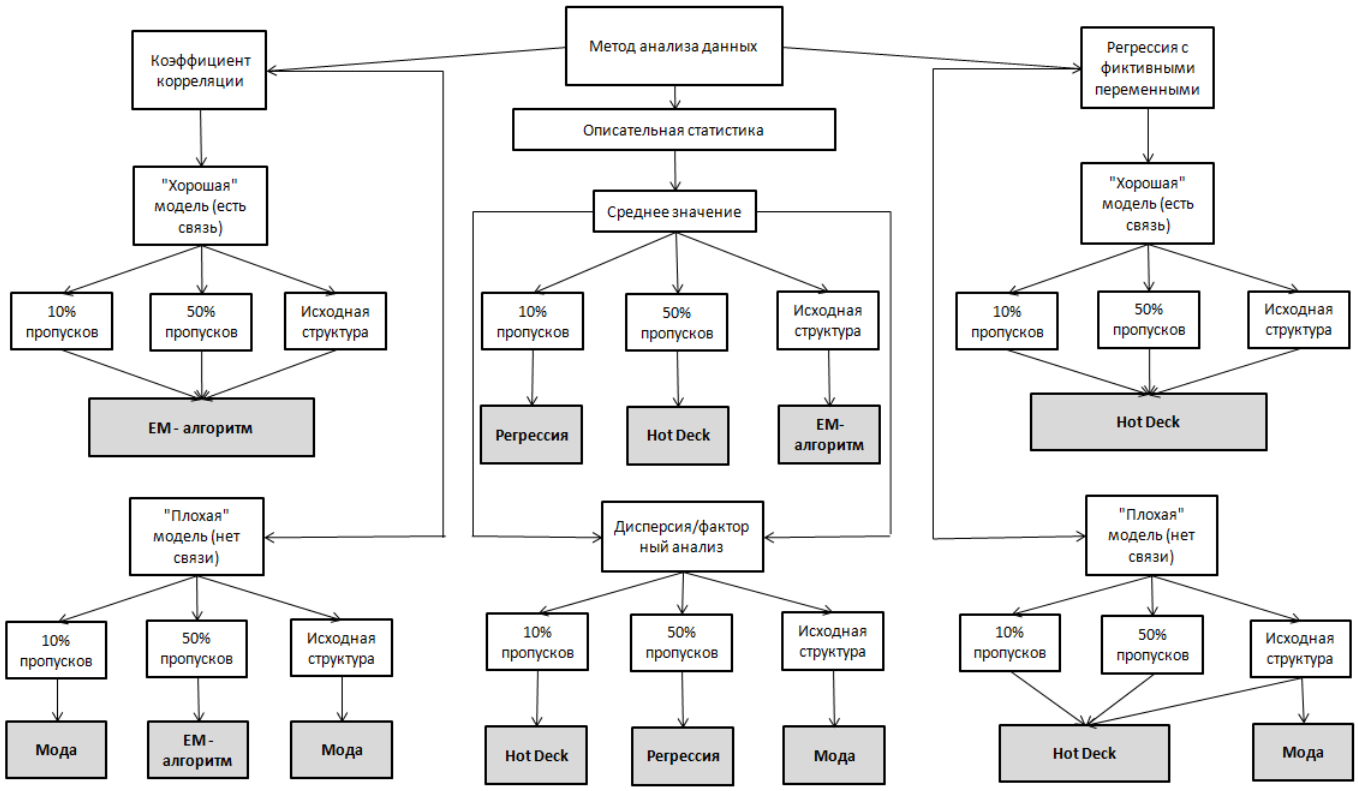

Рисунок 4 – Наиболее эффективные алгоритмы заполнения пропусков в данных в зависимости от используемого метода анализа данных (факторный анализ, регрессия с фиктивными переменными, коэффициент корреляции Пирсона, расчет среднего значения, расчет дисперсии)

На эталонном массиве с помощью коэффициента корреляции была исследована связь между 5 выявленными в ходе факторного анализа типами текущей работы и 3 типами работы потенциальной. В итоге было рассчитано 15 коэффициентов корреляции между каждыми выявленными факторами, однако для проведения бутстрепа было отобрано две противоположные ситуации: с наличием связи и с ее отсутствием, которые условно были названы хорошей моделью и плохой. Для любого количества пропусков для хорошей модели наиболее эффективным был выбран EM-алгоритм, для плохой модели EM-алгоритм оказался наиболее эффективным только для большого количества пропусков, а для 10% и исходной структуры пропусков наиболее эффективным стало заполнение с помощью моды.

Мы рекомендуем исследователям перед выбором конкретного алгоритма провести предварительные расчеты коэффициента корреляции на имеющихся данных, для того чтобы оценить характер связи между переменными и после этого, уже ориентируясь на схему выше, выбрать один из алгоритмов заполнения пропусков.

Применительно к описательной статистике выбор наиболее эффективного алгоритма заполнения пропусков для разного количества пропусков не столь однозначен, как для коэффициента корреляции.

Непосредственно для факторного анализа бутстреп не реализован, и в качестве гипотезы мы предположили, что для этого метода наиболее эффективным может быть тот алгоритм, после заполнения которым будет наименьшее искажение мер средней тенденции и разброса значения фактора.

По этим причинам бутстреп был проведен для расчета дисперсии факторной нагрузки. Как видно из схемы 3, для 10% пропусков наиболее эффективным оказался Hot Deck, для 50% – регрессия, для исходной структуры – мода. Поскольку выбор наиболее эффективного алгоритма заполнения пропусков для факторного анализа на основе дисперсии факторной нагрузки – это лишь предположение, принимать полученный результат в качестве непреложной истины было бы опрометчиво.

В случае с регрессией с фиктивными переменными по тому же принципу, что и в случае с расчетом коэффициента корреляции, для бутстрепа были отобраны две модели: хорошая – с наличием связи между зависимой переменной и предикторами и плохая – с ее отсутствием.

Для любого качества модели наиболее эффективным оказался Hot Deck алгоритм, для плохой же модели, помимо Hot Deck, на массиве с исходной структурой пропусков хорошие показатели дало также заполнение модой.

Результаты, полученные в ходе проведенного исследования, могут стать статистически верифицированными рекомендациями по выбору наиболее эффективного алгоритма заполнения пропусков для исследовательских целей, которые предполагают анализ данных с помощью регрессии с фиктивными переменными, корреляционного анализа, факторного анализа, описательной статистики. Мы надеемся, что эти рекомендации будут востребованы для представителей разных наук, так или иначе связанных с изучением социального мира.

Литература

1 Литтл Р., Рубин Д. Статистический анализ данных с пропусками. М. : Финансы и статистика, 1990.

2 Шитиков В. К., Розенберг Г. С. Рандомизация и бутстреп : стат. анализ в биологии и экологии с использованием R. Тольятти : Кассандра, 2013.

3 Abrahantes J. C., Sotto C. A comparison of various software tools for dealing with missing data via imputation // Journal of Statistical Computation and Simulation. 2011. Vol. 81, Nr 11. P. 1653–1675.

4 Acuna E., Rodriguez C. The treatment of missing values and its effect in the classifier accuracy // Classification, Clustering and Data Mining Applications. 2004. P. 639–648.

5 Davison A. C., Kuonen D. An introduction to the bootstrap with applications in R // Statistical Computing & Statistical Graphics Newsletter. 2002. Vol. 13, Nr 1. P. 6–11.

6 Efron B. Bootstrap methods : another look at the jackknife // The annals of statistics. 1979. Vol. 7, Nr 1. P. 1–26.

7 Elashoff R. M. Missing observations in multivariate statistics : I. Review of the literature // Journal of the American Statistical Association. 1966. Vol. 61, Nr 315. P. 595–604.

8 Horton N. J, Lipsitz S. R. Multiple imputation in practice: comparison of software packages for regression models with missing variables // The American Statistician. Vol. 55, Nr 3. P. 244–254.

9 Kromrey J.D., Hines C.V. Nonrandomly missing data in multiple regression : an empirical comparison of common missing-data treatments // Educational and Psychological Measurement. Vol. 54. P. 573–593.

10 Little R. J. Regression with missing X's : a review // Journal of the American Statistical Association. 1992. Vol. 87, Nr 420. P. 1227–1237.
11 Mitra R., Reiter J. P. Estimating propensity scores with missing covariate data using general location mixture models // Statistics in Medicine. 2011. Vol. 30. P. 627–641.
12 Moons K. J. M. Using the outcome for imputation of missing predictor values was preferred // Journal of Clinical Epidemiology. 2006. Vol. 59. P. 1092–1101
13 Peugh J. L., Enders C. K. Missing data in educational research : a review of reporting practices and suggestions for improvement // Review of Educational Research. Vol. 74, Nr 4. P. 525–556.
14 SOLAS for Missing Data Analysis // Statistical Solutions : [веб-сайт]. 2013. URL: http://www.solasmissingdata.com/software/features.