ТЕОРИЯ И МЕТОДОЛОГИЯ

DOI: 10.19181/socjour.2022.28.4.9313
EDN: YUZYJU

Д.В. ЛЕБЕДЕВ[1]
[1] Национальный исследовательский университет «Высшая школа экономики».
101000, Москва, ул. Мясницкая, д. 20, каб. 331.

ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ GPS-ПАРАДАННЫХ ДЛЯ КОНТРОЛЯ ПРОЦЕССА СБОРА ДАННЫХ: ОБЗОР СУЩЕСТВУЮЩИХ МЕТОДОВ И АНАЛИЗ КАЧЕСТВА ДАННЫХ[1]

Аннотация. Личное интервьюирование до сих пор широко распространено в социальных науках как метод сбора данных. Однако при подобных методах сбора информации велика опасность снижения качества данных из-за фабрикаций или фальсификаций со стороны интервьюеров. Поэтому появилось множество методов контроля процесса сбора данных, однако фокус их ограничен, и они не уделяют внимания поведенческим характеристикам интервьюеров. В этом контексте параданные и GPS-параданные являются важным новым инструментом контроля качества собранной информации, или методического аудита. Они позволяют не только потенциально выявить и предотвратить фальсификации или фабрикации со стороны интервьюеров, но и оценить корректность методических инструкций. В этой статье представлен обзор методов использования GPS-параданных, которые применяются на практике, а именно анализ отдельных локаций интервью и путей интервьюеров. При этом использование GPS-параданных для контроля процесса сбора информации не является безошибочным методом определения фабрикаций или фальсификаций интервьюеров, так как могут возникать технические неточности или их ненамеренные ошибки,

[1] Автор выражает огромную благодарность Айгуль Климовой за ценные советы в процессе проведения исследования и написания статьи, Батий Дине Вадимовне за полезные комментарии по улучшению стилистики и корректности отдельных формулировок и текста в целом, а также анонимным рецензентам «Социологического журнала» за ценные замечания и предложения по улучшению содержания и формы текста статьи.

Статья подготовлена при поддержке Российского фонда фундаментальных исследований. Грантовый проект «Аспиранты» № 20-311-90073.

но это точно полезный дополнительный метод контроля, помогающий выявить «подозрительные» интервью или интервьюеров, для которых требуется применение более ресурсозатратных методов контроля. Кроме того, в статье представлен анализ качества полученных GPS-параданных на примере 26-й волны RLMS–HSE. Он основан на анализе пропущенных данных и качества измерений. Результаты показывают, что качество GPS-параданных может зависеть как от региона проведения интервью, так и от характеристик самих интервьюеров.

Ключевые слова: GPS-параданные; параданные; качество GPS-параданных; мониторинг процесса сбора данных; CAPI; качество данных личных интервью; фальсификации и фабрикации интервьюеров.

Для цитирования: *Лебедев Д.В.* Возможности использования GPS-параданных для контроля процесса сбора данных: Обзор существующих методов и анализ качества данных // Социологический журнал. 2022. Том 28. № 4. С. 8–33. DOI: 10.19181/socjour.2022.28.4.9313 EDN: YUZYJU

Введение

Несмотря на широкое распространение онлайн-методов сбора данных, личное интервьюирование остается важной частью методического инструментария маркетинга и социальных наук. Однако еще в середине XX в. методологи указывали на возможность «обмана» со стороны интервьюеров [23; 49]. Впоследствии на этой проблеме все больше фокусировались в исследованиях методологии социальных наук. До сих пор проводится множество исследований, авторы которых пытаются понять причины фабрикации и фальсификации данных со стороны интервьюеров и найти способы их предотвращения [13; 36].

Вслед за В. Власовым [1, с. 11] и А. Ипатовой [2, с. 27–28] в этой работе фабрикации определяются как намеренные индивидуальные ошибки (отклонения от инструкций исследования) интервьюера, выражающиеся в заполнении анкеты без участия респондента. Фальсификации же подразумевают ситуации, когда фактическое интервью с респондентом имело место, но полученные данные или процесс их сбора были каким-то образом видоизменены интервьюером, чтобы извлечь из этого какую-то выгоду [8; 29]. Подобные ситуации при личном интервьюировании приводят к смещениям в получаемых данных и последующему снижению их качества. Это, в свою очередь, негативно влияет на результаты исследований, так как вносит сложно оцениваемые и зачастую не учитываемые смещения в ошибке выборки, покрытия, неответов и измерения [46], составляющих общей ошибки исследования [32].

Для снижения подобных смещений исследователи в академической, маркетинговой среде, а также сотрудники методологических отделов опросных компаний создали методы контроля работы ин-

тервьюеров. Все они направлены на получение информации о процессе сбора данных. Однако многие методы контроля основываются на самоотчетах интервьюеров и постопросной оценке их работы, что приводит к необходимости внедрения методов поведенческого контроля. Такие методы позволят оценить поведение интервьюеров, получить представление о том, как устроен процесс сбора данных, и даже узнать, как проводится интервью. Один из примеров такого рода данных — параданные.

К параданным относятся показатели, описывающие процесс сбора данных (характеристики интервьюеров, респондентов и ситуации интервью в целом). Они появляются в процессе сбора данных и направлены на оценку или повышение их качества [22; 41; 42]. Параданные включают объективные показатели времени, продолжительности, места интервью и т. д. Использование параданных для анализа личных интервью возможно в случаях применения компьютеризированных методов сбора данных (computer-assisted interviews — CAI), в том числе личных интервью с планшетом (computer-assisted personal interviews — CAPI).

Использование данных о местоположении проведения интервью, или GPS-параданных[2], — один из самых простых и интуитивно понятных способов контроля процесса сбора информации (или методического аудита [7]). Очевидно, если после интервью обнаруживается несовпадение заданного и фактического мест проведения интервью, это веский повод считать, что интервью было сфабриковано или сфальсифицировано (либо является «подозрительным», и необходима дополнительная проверка) [35]. Даже в случае отсутствия установленного места проведения интервью анализ маршрута интервьюера в процессе сбора данных может показать некоторые отклонения интервьюеров от инструкций по корректному сбору данных для обследования [60].

Это подводит к основной цели данного исследования: определить возможные способы использования GPS-параданных для контроля работы интервьюеров в процессе сбора данных и оценить, какие факторы связаны с качеством GPS-параданных, на примере данных Российского мониторинга экономического положения и здоровья населения НИУ ВШЭ (RLMS–HSE).

Статья структурирована следующим образом. Сначала описаны существующие методы контроля работы интервьюеров. Затем указаны возможности и ограничения GPS-параданных для контроля процесса сбора опросной информации. Далее представлен кейс анализа качества GPS-параданных на результатах 26-й волны RLMS–HSE. Наконец, кратко описаны результаты анализа факторов, связанных с качеством

[2] В тексте статьи синонимичным стоит считать термин «GPS-данные».

GPS-параданных, и приведены итоговые рекомендации по использованию GPS-параданных для контроля процесса сбора информации.

Контроль процесса сбора данных: текущие стратегии

Существующие методы контроля работы интервьюеров в процессе сбора данных можно разделить на следующие группы: наблюдение, повторный контакт, постопросный статистический анализ собранных данных и анализ параданных (включая методы метаанализа [11, с. 16]). Кратко опишем эти группы методов.

1. Наблюдение заключается в контроле за процессом коммуникации интервьюера и респондента с помощью анализа аудиозаписи интервью (полностью или частично) или записи изображения экрана планшета. В подобных методах контроля внимание акцентируют на успешности вопрос-ответной коммуникации, наличии коммуникативных сбоев, а также на когнитивных нарушениях в процессе предоставления ответа на вопрос, которые могут привести к смещениям и повышению ошибки измерения. Интервьюеры и респонденты должны быть предупреждены о возможной записи интервью, но не следует детально информировать о том, какая именно часть беседы будет записана [6; 9; 25; 31, с. 42; 58].

2. Повторный контакт включает связь с респондентом после проведения интервью и после передачи данных исследователям, чтобы узнать, проводилось ли интервью в целом, а также получить информацию о поведении интервьюера в процессе коммуникации. Общепринятая норма проверок проведенных интервью составляет 5–15% [45, с. 319]. Эта группа методов мониторинга полевых работ считается наиболее распространенной [3; 8, с. 7; 14, с. 1; 15; 38; 45, с. 316].

3. Методы, направленные на анализ полученных опросных данных, включают выявление фабрикаций и фальсификаций интервьюеров, основанное на изучении различных характеристик получаемых данных (как правило, с использованием статистических инструментов). Преимущество этих методов — возможность контролировать процесс сбора данных во время проведения полевого этапа, однако это доступно только в случае использования компьютерных методов сбора данных (в основном CAPI). Подобные методы контроля могут включать автоматический анализ получаемых ответов на вопросы о доходе каждого отдельного члена домохозяйства и домохозяйства в целом (сравнение суммы доходов каждого отдельного члена домохозяйства со значением указываемого дохода домохозяйства в целом) или постопросный анализ (например, применение закона Брэдфорда к ответам на открытые вопросы с числовыми значениями) [11; 14; 28; 44; 50–52; 57].

4. Методы, основанные на параданных, включают анализ процесса проведения интервью с учетом поведенческих характеристик

респондента и интервьюера (местоположение и передвижение, вокальные характеристики, навигация по анкете и т. п.). Они используют «побочные» данные, которые описывают процесс проведения интервью (параданные) [4; 12; 20; 22; 26; 35; 41–43; 47; 59; 60; 62].

Следует учитывать, что все стратегии контроля работы интервьюеров в процессе сбора данных специфичны для определенного типа обследования, дизайна выборки, процедуры сбора данных и т. д. [59]. Считается, что наиболее эффективным способом выявления фальсификаций и фабрикаций в работе интервьюеров является комбинация методов или их дополнение. Например, некоторые исследователи определяют выборку для повторного контакта или повторного интервью, используя методы мониторинга на основе полученных опросных данных [40]. В связи с этим распространена практика использования как можно большего набора методов контроля работы интервьюеров.

Контроль процесса сбора данных: GPS-параданные

Несколько десятилетий назад данные спутниковой системы навигации (Global positioning system — GPS) было трудно собрать и использовать для анализа. Для этого требовалась специализированная дорогостоящая техника [19]. Сейчас можно проводить точные измерения GPS-данных, используя только планшет или даже смартфон. Такие измерения уже доказали свою высокую точность в различных контекстах исследований [27; 63]. Благодаря повышению доступности использования GPS-данных и их эффективности становится возможным применение подобной информации при контроле процесса проведения полевого этапа опроса.

Данные географического позиционирования, или данные GPS, могут быть определены как цифровые записи, указывающие местоположение объекта (человека, животного, автомобиля и т. д.) в земном пространстве. Обычно они представлены совокупностью двух метрик — долготы и широты местоположения в определенный момент времени. Следует отметить, что данные GPS, как правило, изучаются либо как отдельные точки GPS (долгота и широта), либо как последовательность таких точек, из которых могут быть сформированы пути объекта в пространстве.

Важно понимать, что, несмотря на упрощение процесса сбора такой информации (подобные данные можно собирать без специального оборудования, используя встроенные в смартфоны и планшеты GPS-трекеры), все еще существуют трудности, связанные со сбором и анализом GPS-локаций. Одним из наиболее важных вопросов в этом отношении является качество GPS-данных.

Собранные GPS-данные могут включать значительную долю пропущенных замеров, которая может варьироваться от 12 до 30% от всех GPS-измерений и выше [18; 33; 64]. Пропуски замеров GPS-

данных могут быть вызваны различными причинами, включая ошибку субъекта, управляющего замером, технические проблемы устройства, окружающую среду (географическое положение хуже измеряется, когда место окружено высокими зданиями, что препятствует корректной передаче сигнала GPS (гипотеза «городских каньонов») [30; 39].

Более того, данные GPS могут быть подвержены ошибке измерения. Например, М. Китинг с коллегами обнаружили, что GPS-данные в ходе полевого исследования 18 тыс. домашних хозяйств были точны в пределах 8 метров в 95% и в пределах 25 метров в 99% случаев [35]. Ошибка измерения GPS-данных может быть вызвана плохим соединением (уровень заряда устройства, нахождение в помещении), окружающей обстановкой (здания, деревья), использованием вышек сотовой связи вместо спутников для определения местоположения [30; 39; 43], погодными условиями [43].

Для оценки качества измерения GPS-данных можно использовать следующие показатели:

– горизонтальное снижение точности (Horizontal Dilution of Precision — HDOP). Этот показатель используется в некоторых приложениях, позволяющих собирать данные (например, GPS Logger). Он отражает количество метров, на которое может отклоняться полученное местоположение от реального, и измеряется путем учета положения спутников, использованных для замера (чем дальше спутники друг от друга, тем выше качество замера). К. Олсон и Дж. Вагнер, авторы одной из немногих опубликованных работ об использовании GPS-параданных для контроля процесса сбора данных, получили только 21,6% измерений в пределах требуемого диапазона (5 метров) [60, с. 225];

– время, прошедшее между последовательными измерениями местоположения. Показатель также может использоваться для измерения качества GPS-данных. Конфигурации приложений, позволяющих собирать GPS-данные, могут различаться, но, как правило, местоположение измеряется каждые 60 секунд. Таким образом, временной интервал более 60 секунд между двумя последовательными GPS-замерами может указывать на некоторые проблемы с качеством их сбора. В исследовании К. Олсон и Дж. Вагнера 19% последовательных GPS-локаций имели разрыв более 120 секунд [60, с. 225];

– доля пропущенных GPS-измерений. С учетом автоматизма измерения местоположения в программах, направленных на сбор данных о путях (последовательность точек местоположений, которые складываются в маршрут), этот показатель может свидетельствовать о снижении качества GPS-измерений [18; 33; 64].

Важно отметить, что не все программные обеспечения, позволяющие замерять и собирать GPS-данные, также сохраняют показатели

качества измерений, поэтому стоит заранее проверять их наличие для возможности последующей корректировки и учета неточностей при анализе.

Несмотря на повышение интереса к параданным и различным стратегиям их применения (в том числе для контроля процесса сбора данных) [41], до сих пор представлено лишь несколько примеров использования GPS-параданных в контексте контроля процесса сбора данных [17; 53; 60].

В личных интервью с использованием планшета (CAPI) GPS-параданные могут быть собраны «реактивно» интервьюером: измерение местоположения в начале и в конце интервью, включение записи маршрута перед поездкой в исследуемый район и выключение после нее [60]. Или их собирают исследователи и руководители процесса сбора данных «нереактивно»: отслеживают пути интервьюеров с помощью установки специального программного обеспечения с непрерывной записью GPS-параданных [35; 59]. Соответственно, существуют также две основные доступные стратегии использования GPS-параданных для контроля процесса сбора информации: анализ отдельных точек GPS (местоположение интервьюеров при сборе данных, то есть проведении интервью) и анализ последовательности точек GPS (маршрут интервьюеров на полевом этапе).

Наиболее интуитивно понятный метод использования GPS-параданных для контроля процесса сбора данных — сравнение географического положения интервьюеров в начале и в конце интервью с последующим сравнением двух измеренных и записанных местоположений. Это показывает, было ли интервью начато в том же месте, где закончилось (такая проверка направлена на следование инструкциям заполнения всего опроса только в присутствии респондента) [42; 53; 61]. Из-за вероятности ошибки измерения GPS-параданных обычная процедура сравнения геопозиций включает построение виртуальных окружностей для каждой локации (диаметр зависит от качества измерений GPS, производимых устройством HDOP). Дальнейший анализ предполагает сравнение построенных окружностей на предмет наличия перекрытия между двумя окружностями локаций. Если такие области пересекаются, то можно сказать, что разница между местоположениями незначительная и присутствует только из-за неточности измерений. Если же такие виртуальные окружности вокруг двух локаций не пересекаются, это может указывать на необходимость дальнейшего разбирательства по поводу качества проведенного интервью, потому что в этом случае начало и конец интервью проводились в разных местах. Этот процесс иногда называют «построением геозоны» (geofencing) [48]. Сайкс представил близкий по логике применения GPS-параданных кейс — систему, которая используется в RTI International для проверки путем сравнения местоположения проведения интервью

с местоположением домохозяйства, включенного в выборку исследования (с помощью приложения Google Earth), а также геопозиции в начале и в конце интервью [55].

Другой метод использования GPS-параданных для обнаружения фальсификаций или фабрикаций со стороны интервьюеров был применен в переписи населения США 2010 г. для выявления скоплений местоположений интервью (curbstoning) (интервью проводились слишком плотно — более 6 ситуаций вопрос-ответной коммуникации на площади 12,2 х 12,2 метра в определенном для работы интервьюера регионе). В результате 13,82% интервьюеров не смогли пройти проверку качества данных. Дополнительный тип проверки качества данных включал сравнение расстояний между отобранными для опроса домохозяйствами с данными GPS, касающимися местоположения проведения самого интервью (strand length). В результате применения этого метода контроля процесса сбора данных 14,62% интервью были отмечены как «подозрительные» [16, с. 9]. В последующих переписях населения США такую стратегию мониторинга процесса сбора данных планируют использовать как после полевого этапа (для выявления сфабрикованных и фальсифицированных данных), так и в процессе сбора данных (для предотвращения такого поведения). Во время сбора данных местоположение интервьюера в начале интервью сравнивается с местоположением отобранного в выборку исследования домохозяйства; если расстояние окажется слишком большим, на планшете интервьюера появится предупреждение. Если 6 интервью подряд будут проведены в одной и той же локации (в процессе сбора данных) [24], интервьюера попросят подтвердить свое местоположение.

Л. Мохаджер и Б. Эдвардс описывают, как данные GPS используются в компании Westat [42, с. 272]. Физическое местоположение интервьюера в начале проведения интервью всегда записывается и используется для проверки тремя способами: местоположение интервьюера сравнивается с местоположением объекта выборки (респондент или домохозяйство) и с местом проживания самого интервьюера, а также анализируется время проведения замера локации интервью. Система EAGLE направлена на выявление подозрительного поведения интервьюеров для дальнейшего подробного разбирательства с использованием других методов мониторинга процесса сбора данных (повторный контакт, анализ полученных опросных данных и т. д.) [34].

Другая группа методов, связанная с использованием GPS-параданных для мониторинга процесса сбора данных, — анализ маршрутов интервьюеров в этот период для выявления любого рода фабрикаций и/или фальсификаций, которые могут заключаться даже в незначительных отклонениях от маршрутов (при использовании маршрутной выборки). Эти методы анализируют поведение интервьюеров таким образом, чтобы можно было выявить действия,

которые справедливо расценивать как потенциальные фальсификации или фабрикации данных (например, повторное прохождение выбранного домохозяйства без контакта, а не случайный выбор домохозяйств для обследования) [59, с. 339–347]. Несмотря на то что такой тип контроля процесса сбора данных имеет большой потенциал и может помочь глубже понять поведение интервьюеров, есть лишь несколько подробных примеров его использования [17; 47; 60].

GPS-параданные можно сопоставить с общедоступными картами, чтобы получить информацию о том, где происходил контакт интервьюеров с респондентами. Дж. Вагнер с коллегами исследовали связь между GPS-параданными и данными о контактах, чтобы выявить некоторые возможные ошибки интервьюеров. Для определения маршрутов интервьюеров они связали данные о местоположении с близлежащей улицей, а затем с полученными позициями на дороге, чтобы отобразить маршрут интервьюера. В итоге удалось получить важные сведения о поведении интервьюеров в процессе сбора данных. Повторное прохождение одного домохозяйства при избегании других и сами маршруты интервьюеров в процессе сбора данных отличались от предположений исследователей о поведении интервьюеров (нелинейные паттерны в сравнении с моделями путешествий внутри района). Эти представления о поведении интервьюеров на местах были выявлены благодаря анализу GPS-параданных на уровне интервьюеров. Проведенный анализ помог исследователям проанализировать поведение интервьюеров, отклоняющееся от методических инструкций по сбору данных, но повышающее эффективность сбора данных [59].

Краткое описание существующих методов использования GPS-параданных для контроля процесса сбора данных представлено в таблице 1.

Обычная стратегия работы с результатами проверки качества данных состоит в том, чтобы определить фабрикации и/или фальсификации после того, как процесс сбора данных завершен [34; 42; 48]. Однако методы curbstoning, geofencing и strand length могут быть использованы непосредственно в процессе сбора данных для предупреждения интервьюеров об их «ошибочном» поведении, что помогает изменить его и может привести к предотвращению снижения качества получаемых опросных данных [16; 24]. Тем не менее при применении методов анализа пути интервьюера проблематично использование стратегии предотвращения, поскольку для нее необходимы анализ слишком больших наборов данных, значительные временны*е* и технические ресурсы, связанные со специфическими для GPS-параданных методами анализа данных [42; 60].

Таблица 1

Методы использования GPS-параданных для контроля процесса сбора данных

Метод использования	Описание	Стратегия использования результатов контроля	Источники
	Анализ отдельных GPS-точек		
Strand length	Сравнение местоположений объекта выборки и интервью с учетом точности двух GPS-замеров для расчета достаточного расстояния, чтобы не считать его значимо отличным (учет ошибки измерения GPS-параданных). Анализ также может учитывать GPS-данные о месте проживания интервьюера.	Обнаружение и предотвращение (с предупреждением интервьюеров)	[16; 34; 42; 55]
Geofencing	Сравнение местоположений интервьюера в начале и в конце интервью. Наличие значимого расстояния между двумя точками (больше, чем потенциальная погрешность замера GPS-параданных) может указывать на наличие подозрительных интервью, потенциально сфабрикованных или сфальсифицированных интервьюерами.	Обнаружение и предотвращение	[17; 42; 43; 53; 61]
Curbstoning	Проверка на наличие слишком плотных групп мест проведения интервью (например, более 6 интервью на квадрате 12,2x12,2 метра в зоне выборки).	Обнаружение и предотвращение	[10; 16; 24]
	Анализ пути (последовательности GPS-точек)		
Соединение место-положений интервью	Проверка расстояния между местами проведения интервью, длины маршрута интервьюеров и случайности характера отбора домохозяйств для контакта.	Обнаружение	[17]
Анализ путей интервьюеров	Сбор и анализ маршрутов интервьюеров в процессе сбора данных для получения более глубокого понимания поведения интервьюеров.	Обнаружение	[47; 60]

Важно также отметить, что, несмотря на большой потенциал в использовании для контроля процесса сбора данных, GPS-параданные не являются панацеей для мониторинга полевых работ, они должны применяться вместе с другими методами и доступными данными. Использование на практике GPS-параданных для контроля процесса сбора данных заключается в выявлении подозрительного поведения интервьюеров, в связи с чем необходимо дальнейшее его изучение. Следует определить, является ли это фабрикацией и/или фальсификацией интервьюера либо чем-то другим (например, ошибкой измерения GPS, непреднамеренной ошибкой интервьюера, техническими трудностями при сборе данных и т. д.) [10; 16; 34]. Кроме того, способ использования GPS-параданных зависит от дизайна выборки проводимых исследований. Анализ путей можно использовать только в случае применения маршрутной выборки, тогда как анализ отдельных точек невозможен для метода сбора данных в местах скопления респондентов, попадающих под критерии выборки (в этом случае отсутствие разницы местоположений интервью заложено самой методикой полевого этапа).

Дж. Вагнер с коллегами представили единственный в своем роде и многообещающий способ анализа GPS-параданных, изучив маршруты интервьюеров в полевых условиях и связав их с данными о контактах [47; 60]. Стоит заметить, что проведенный ими анализ использовался не только как способ контроля интервьюеров или метод выявления подозрительного поведения, потенциально связанного с фабрикацией или фальсификацией, но и как возможность получить более полную картину процесса сбора данных. Такое использование GPS-параданных (анализ путей интервьюеров) вполне можно считать перспективным направлением в рамках методического аудита [7], которое позволяет лучше понять отклонения интервьюеров от методических инструкций в процессе проведения полевого этапа. Эти отклонения могут быть не только индикатором фальсификаций или фабрикаций, но и примером повышения эффективности сбора данных.

Исследования об использовании GPS-параданных для контроля процесса сбора данных немногочисленны. Необходимы практические примеры, как их можно применять и что нужно учитывать при анализе такого типа данных. Так, важно принимать во внимание качество GPS-параданных и факторы, провоцирующие его понижение и связанные с ошибкой измерения и ошибкой *неответов* (пропущенные данные из-за неправильного замера локаций).

Методология эмпирической части исследования

Для проверки наличия связи различных факторов с качеством GPS-параданных были использованы данные российского «Мониторинга экономического положения и здоровья населения

НИУ ВШЭ» (RLMS–HSE), панельного лонгитюдного обследования населения, проводимого ежегодно с 1992 г. (с некоторыми перерывами в 1990-х гг.) [37]. С 26-й волны руководство RLMS–HSE приняло решение постепенно переходить с метода сбора данных путем личного интервьюирования с бумажной анкетой на метод личного интервьюирования с использованием планшета (CAPI).

В 26-й волне 36 интервьюеров собирали часть интервью с помощью планшета, что позволило использовать и возможности GPS-контроля процесса сбора данных. Всего в период с ноября 2017 г. по февраль 2018 г. было проведено 448 интервью методом CAPI в 26-й волне RLMS–HSE. Распределение количества интервью по регионам представлено в таблице 2.

Таблица 2

Количество интервью, собранных с помощью планшета в 26-й волне RLMS–HSE, по регионам

Регионы	Количество интервью	Доля интервью, %
Соликамск	43	9,6
Казань	27	6,0
Курган	67	15,0
Вольск	33	7,4
Москва	100	22,3
Московская область	134	29,9
Бердск	44	9,8
Всего	448	100

Сбор данных осуществлялся с помощью планшетов Samsung Galaxy Tab A 16 SM-T355. Для записи данных использовалось открытое бесплатное программное обеспечение Survey Soulutions, которое также позволяло замерять местоположение планшета (GPS-параданные) в начале и в конце опроса посредством специального GPS-вопроса [56]. Запись GPS-координат планшета в начале и в конце интервью замерялась «реактивно» интервьюером (он должен был нажать на кнопку «замерить местоположение» при появлении на экране соответствующего GPS-вопроса). Благодаря программному обеспечению Survey Solutions также автоматически были получены значения точности GPS-измерений (HDOP) для каждого замера GPS-параданных.

В процессе анализа была использована информация о наличии пропущенных данных (либо в начале интервью, либо в конце), а также о качестве замеров (среднее значение показателя HDOP между двумя замерами для каждого отдельного интервью). Кроме того, перед на-

чалом полевого этапа интервьюеры заполняли бумажную анкету для предоставления релевантной информации относительно своих ожиданий об успешности использования планшетов для сбора данных, оценки собственных умений работы с планшетом, а также о возрасте, поле и других социально-демографических показателях. Это позволило при анализе также учитывать характеристики интервьюеров для оценки их связи с качеством измерений. Стоит отметить, что подобные данные использовались только при анализе пропущенных значений, замеряемых «реактивно», но не для анализа качества GPS-измерений. Такое решение было принято из-за отсутствия обоснованных предположений о возможности влияния интервьюеров и их характеристик на качество замеров (показатель HDOP), которое при успешном измерении зависит от местоположения, окружения и характеристик планшета [47].

Цель анализа данных — оценить связь различных факторов (характеристики интервьюеров и регион проведения интервью) с наличием пропущенных данных при измерении местоположения интервью в начале или в конце его проведения, а также с точностью измерений GPS-параданных.

В качестве факторов, для которых проверяется наличие связи пропуска данных с качеством замеров GPS-параданных (HDOP), использованы следующие переменные:

– возраст интервьюера — из доступных демографических переменных был выбран именно этот показатель на основе предыдущих исследований о качестве сбора данных интервьюерами с использованием планшетов [5]. При этом в анализ не был включен пол из-за отсутствия вариации (все 36 интервьюеров в 26-й волне RLMS–HSE, проводивших интервью с планшетом, были женского пола). Кроме того, нет оснований предполагать, что характеристика пола может быть связана с вероятностью наличия пропущенных GPS-измерений;

– наличие собственного планшета у интервьюера;

– самооценка уверенности в работе с планшетом — оценка интервьюером своих навыков как пользователя планшета по шкале от 0 до 7 (где 1 — «неуверенный пользователь, почти не умею пользоваться», 7 — «очень уверенный пользователь», 0 — «никогда не пользовался(-ась) планшетом»);

– индекс ожиданий относительно успешности применения планшета для сбора данных — сумма значений по четырем переменным (ожидания относительно длительности интервью, ожидания сложности/легкости проведения интервью с использованием планшетов, ожидания успешности проведения интервью, ожидания изменений в атмосфере проведения интервью; переменные оценивались по пятибалльной шкале, где более высокое значение отражает бóльшую

уверенность). Этот показатель показывает общий уровень ожиданий (см.: [5]);

– данные о регионе интервью.

Предполагается, что и пропущенные значения, и пониженное качество данных GPS будут характерны для крупных городов (Москва, Казань) в соответствии с гипотезой «городских каньонов» [30; 39; 43]. Также ожидается, что факт наличия пропущенных значений будет существенно связан с самооценкой уверенности в работе с планшетом и индексом ожиданий интервьюеров (в соответствии с результатами предыдущих исследований о влиянии ожиданий интервьюеров на качество собранных данных [6; 21; 54]), но не с возрастом, связь с которым может опосредоваться связью с уверенностью в использовании планшета [5].

Результаты эмпирической части исследования

В 26-й волне RLMS–HSE было 105 случаев пропуска в измерениях GPS-параданных (местоположения планшета в процессе проведения интервью) либо в начале, либо в конце интервью (22,3%). Для прогнозирования вероятности наличия пропущенных замеров GPS-данных на основе выбранных характеристик интервьюеров и региона была использована бинарная логистическая регрессия, построенная методом «Ввод», который предполагает одновременное включение в модель всех переменных.

Для включения показателя региона в модель переменная о регионе проведения интервью была переведена в фиктивный вид (отдельные дихотомические переменные для каждого региона). В качестве контрольной группы для набора фиктивных переменных, отражающих регион интервью, были выбраны интервью, проведенные в Москве (самый большой в анализе город/регион, в котором наблюдалась наибольшая доля пропущенных значений GPS-замеров в начале или в конце интервью).

По результатам построенных моделей подтверждается гипотеза о связи размера региона (наличие высоких зданий вокруг при замере GPS-параданных) и качества GPS-параданных. В сравнении с интервью, проведенными в Москве, при проведении интервью в Соликамске, Кургане, Вольске и Бердске значительно ниже вероятность наличия пропущенного GPS-замера в начале или в конце интервью (p-value $< 0{,}05$, exp(B) для всех перечисленных регионов меньше 1) (табл. 3).

Кроме того, частично подтверждается гипотеза о том, что с фактом наличия пропущенных значений GPS-измерений в начале или в конце интервью связаны уверенность интервьюера в работе с планшетом и индекс ожиданий от успешности использования планшета для сбора

данных. Возраст как независимая переменная перестает быть значимым при добавлении в модель взаимодействия возраста и уверенности интервьюеров в работе с планшетом (табл. 3, 4) При этом чем выше уверенность, тем ниже вероятность пропусков в данных GPS. Стоит отметить, что индекс ожиданий оказался незначимым предиктором (отсутствует значимая связь индекса ожиданий с вероятностью наличия пропущенного измерения GPS-местоположения в начале или в конце интервью).

Таблица 3

Регрессионные коэффициенты бинарной логистической регрессии для проверки связи характеристик интервьюеров и региона проведения интервью с фактом наличия пропусков в замерах GPS-параданных *(без взаимодействия таких показателей, как возраст и уверенность в работе с планшетом)*

Независимые переменные регрессионной модели	B	Значимость	Exp(B)	95% ДИ для Exp(b)	
				нижняя	верхняя
Константа	8,902	0,002	7344,991		
Возраст	–0,094	**0,001**	**0,910**	0,859	0,963
Наличие планшета	–0,302	0,216	0,739	0,458	1,193
Уверенность в работе с планшетом	–0,600	**0,000**	**0,549**	0,431	0,699
Индекс ожиданий	–0,102	0,238	0,903	0,762	1,070
Соликамск	–5,721	**0,000**	**0,003**	0,000	0,040
Казань	–20,294	0,998	0,000	0,000	0,000
Курган	–4,947	**0,000**	**0,007**	0,001	0,059
Вольск	–2,071	**0,002**	**0,126**	0,034	0,466
Московская область	–0,539	0,251	0,584	0,233	1,464
Бердск	–3,127	**0,000**	**0,044**	0,009	0,208
Оценка качества построенной модели	$X^2(10) = 110,8$; p = 0,000; –2 Log-правдоподобие = 200,5; R^2 Кокса и Снелла = 0,288; R^2 Найджелкерка = 0,468; Доля правильных предсказаний = 89%				

Следующая задача заключалась в проверке связи между регионом проведения интервью (GPS-замера в начале или в конце интервью) и точностью GPS-замеров. Для этого было использовано среднее значение показателя HDOP, отражающего точность GPS-измерения. Стоит отметить, что данный показатель выражается в метрах и демон-

стрирует, насколько сильно может отклоняться реальное местоположение от записанного. То есть повышение значения этого показателя свидетельствует о снижении качества GPS-измерения и увеличении ошибки измерения.

Таблица 4

Регрессионные коэффициенты бинарной логистической регрессии для проверки связи характеристик интервьюеров и региона проведения интервью с фактом наличия пропусков в замерах GPS-параданных
(с взаимодействием таких показателей, как возраст и уверенность в работе с планшетом)

Независимые переменные регрессионной модели	B	Значимость	Exp(B)	95% ДИ для Exp(b)	
				нижняя	верхняя
Константа	–0,295	0,056	244,740		
Возраст	0,110	0,105	0,952	0,913	1,365
Наличие планшета	–0,515	0,886	1,044	0,347	1,029
Уверенность в работе с планшетом	2,101	**0,000**	**0,642**	0,631	105,840
Индекс ожиданий	–0,221	0,488	0,940	0,640	1,004
Уверенность в работе с планшетом х возраст	–0,053	**0,004**	**0,986**	0,901	0,998
Соликамск	–6,176	**0,000**	**0,006**	0,000	0,037
Казань	–21,885	0,998	0,000	0,000	0,
Курган	–6,205	**0,000**	**0,008**	0,000	0,027
Вольск	–2,581	**0,004**	**0,142**	0,017	0,334
Московская область	–0,784	0,957	0,970	0,167	1,248
Бердск	–2,266	**0,000**	**0,065**	0,019	0,564
Статистики качества	$X^2(11) = 117$; p = 0,000; –2 Log-правдоподобие = 194,3; R^2 Кокса и Снелла = 0,301; R^2 Найджелкерка = 0,490; Процент правильных предсказаний = 91,4%				

Для решения этой задачи была построена линейная регрессионная модель (методом «Ввод», предполагающим одновременное включение в модель всех переменных), где в качестве зависимой переменной взято среднее значение HDOP для каждого опроса между замерами в начале и в конце интервью (общее среднее значение — 23,6 метра, стандартное отклонение — 11,3), а в качестве независимых переменных — набор фиктивных переменных, отражающих регион проведения интервью (Москва в качестве контрольной группы) (табл. 4). Полученная модель имеет низкое качество ($R^2 = 0,022$, ANOVA: F = 1,33, p = 0,242). При

интерпретации результатов построенной модели стоит учитывать ее низкое качество и вероятное наличие более важных характеристик, связанных с качеством GPS-измерений. Однако из-за отсутствия данных о технических характеристиках GPS-измерений и планшетов они не включены в модель. Тем не менее предполагается, что и полученный низкий уровень объясняемой дисперсии зависимой переменной позволяет оценить, какие нетехнические характеристики могут быть связаны со снижением качества измерения GPS-данных.

Содержательная гипотеза о наличии более качественных GPS-параданных в меньших регионах, чем Москва, отвергается (табл. 5). Несмотря на то что наблюдается значимая разница в среднем качестве GPS-параданных в Москве и Казани (в Казани качество данных выше, так как проведение интервью в этом городе связано со снижением отклонения GPS-измерения от реального местоположения (HDOP) на 6,4 метра, p-value < 0,05), в регрессионной модели отсутствуют другие значимые предикторы.

Таблица 5

Регрессионные коэффициенты линейной регрессии для проверки связи региона проведения интервью со средним значением качества измерений GPS-параданных (HDOP)

Независимые переменные регрессионной модели	Нестандартизованные коэффициенты	Стандартизованные коэффициенты	t	Sig.	95,0%-ный доверительный интервал для B	
	B	Бета			нижняя граница	верхняя граница
Константа	24,5 (1,4)		17,4	0,000	21,724	27,258
Соликамск	–0,5 (2,2)	–0,015	–0,2	0,813	–4,925	3,866
Казань	**–6,4 (2,5)**	–0,149	–2,5	**0,013**	–11,496	–1,337
Курган	–1,0 (1,9)	–0,036	–0,5	0,596	–4,930	2,837
Вольск	0,3 (2,5)	0,007	0,1	0,908	–4,721	5,310
Бердск	–1,0 (1,8)	–0,040	–0,6	0,571	–4,566	2,521
Московская область	0,6 (2,2)	0,017	0,2	0,786	–3,875	5,118

Таким образом, выдвинутые гипотезы подтверждаются частично. Действительно, наблюдается значимая связь между регионом проведения интервью и характеристиками интервьюера (уверенность в работе с планшетом) с фактом наличия пропущенных значений GPS-параданных в начале или в конце интервью. Чем выше уверенность в работе с планшетом, тем ниже вероятность пропуска GPS-замеров. Эта вероятность также ниже при проведении интервью в Соликамске, Кургане, Вольске и Бердске по сравнению с Москвой.

При этом снижение точности GPS-измерений не связано с тем, в каком регионе — малом, среднем или крупном — было проведено интервью. Единственное значимое отличие в регионах наблюдается при проведении интервью в Казани по сравнению с Москвой, что приводит к снижению ошибки измерения (HDOP).

Заключение

Использование методов личного интервью для сбора данных в лонгитюдных или срезовых обследованиях неразрывно связано с необходимостью включения интервьюеров в процесс получения данных. При этом методический аудит обследований в опросных компаниях и методологические исследования в связи с этим обязательно включают методы контроля процесса сбора данных, направленные на выявление фабрикаций и фальсификаций со стороны интервьюеров.

В статье описаны основные методы возможного использования GPS-параданных в контексте мониторинга процесса сбора данных, которые подразделяются на две основные группы: анализ отдельных точек (обычно не более двух) и анализ последовательных GPS-точек (существенно больше двух).

Анализ отдельных точек:

- Strand length;
- построение геозоны (Geofencing);
- Curbstoning.

Анализ последовательных GPS-точек:

- соединение местоположений интервью;
- анализ путей интервьюеров.

Важно отметить, что использование GPS-параданных, как и других методов контроля процесса сбора данных, не приводит к высокой надежности результатов. Это может быть связано как с тем, что отклонения в данных GPS могут появляться из-за непреднамеренных ошибок интервьюеров или технических проблем с планшетом/устройством для записи GPS, так и с тем, что GPS-параданные подвержены ошибке измерения и неответов из-за технических особенностей их измерения (снижение качества измерения в больших городах из-за отражения GPS-сигнала от стен высоких зданий, погодные условия, расположение спутников относительно места замера и др. [30; 39; 43]). Кроме того, на практике GPS-параданные зачастую используют не как единственный способ выявления фальсификаций или фабрикаций, а как метод определения подозрительных интервью или интервьюеров [16; 34]. При выявлении таким способом подозрительных интервью или интервьюеров их дополнительно проверяют с применением других методов контроля процесса сбора данных (например, методом повторного контакта) [15; 16; 34].

Задачей этой статьи была также оценка факторов, связанных с качеством GPS-параданных. На основе анализа пропущенных GPS-замеров в начале или в конце 448 интервью, собранных в рамках 26-й волны RLMS–HSE и проведенных методом личного интервьюирования с использованием планшетов, подтверждается гипотеза о значимом снижении вероятности наличия пропущенных данных при проведении интервью в регионах с меньшим количеством высоких зданий (в Соликамске, Кургане, Вольске и Бердске по сравнению с Москвой ниже вероятность появления пропущенного замера в начале или в конце интервью). Это согласуется с предыдущими результатами анализа подобных проблем в качестве GPS-параданных [30; 39; 43]. При этом по качеству измерений (показатель HDOP — ошибка измерения GPS-параданных) значимая связь наблюдается только в Казани в сравнении с Москвой (ошибка измерения снижается). В случае «реактивно» замеряемых GPS-параданных факт пропуска GPS-замеров может быть связан с характеристиками самих интервьюеров (повышение уверенности в работе с планшетом значимо связано с понижением вероятности наличия пропущенных замеров GPS-параданных в начале или в конце опроса) [47].

По результатам анализа, в целом качество GPS-параданных в 26-й волне RLMS–HSE, собранных в рамках 448 интервью, проведенных методом личного интервьюирования с использованием планшетов, можно считать удовлетворительным. Доля пропущенных GPS-замеров в начале или в конце интервью составила 22,3%, что не сильно отличается от результатов предыдущих кейсов использования GPS-параданных, описанных в литературе [18; 33; 64], как и средняя точность измерения, которая составила 23,6 метра (стандартное отклонение — 11,3) [35]. Это позволяет применять GPS-параданные для контроля процесса сбора данных. Необходимо учитывать имеющуюся связь с качеством GPS-параданных региона проведения интервью, а также характеристики интервьюеров (уверенность в работе с планшетом) в случае использования «реактивно» замеряемых GPS-параданных. Перед началом сбора данных следует провести подробный тренинг/серию тренингов для обучения интервьюеров работе с планшетом.

В статье не был проведен анализ самих GPS-параданных, так как это выходит за рамки поставленной задачи. При этом ограничением также стоит считать тот факт, что на практике интервьюеры могут использовать сторонние программы, в том числе направленные на фабрикацию самих GPS-параданных. Такие ситуации требуют отдельного изучения с разработкой стратегий для их выявления и предотвращения (например, ограничение возможности самостоятельной установки любых дополнительных программ интервьюерами, что применялось в случае RLMS–HSE). Ограничением при анализе эмпирической части

статьи является отсутствие доступа к дополнительным техническим показателям проводимых GPS-измерений на стороне использованного программного обеспечения для сбора данных о местоположении. Низкое качество построенной линейной регрессионной модели проверки взаимосвязи качества GPS-измерений и характеристик региона свидетельствует о том, что объясняемая дисперсия зависимой переменной находится на невысоком уровне и в модель не включены более важные характеристики (предположительно это технические характеристики связи планшета со спутниками или мобильными вышками). В следующих исследованиях оценки качества GPS-параданных стоит учесть необходимость добавления таких технических характеристик.

Таким образом, в статье описаны методы контроля процесса сбора данных, основанные на GPS-параданных. Использование таких методов в дополнение к традиционным методам мониторинга процесса сбора данных позволит более точно определять подозрительные интервью исходя из места их проведения при условии возможности измерения таких параданных и их релевантности с точки зрения дизайна выборочного процесса и методологических особенностей обследования в целом. Кроме того, GPS-параданные могут быть полезны не только в рамках поиска потенциальных фабрикаций и фальсификаций интервьюеров, но и для улучшения методических процедур процесса сбора данных (методического аудита) [7; 47; 60]. Стоит учитывать, что для применения некоторых методов использования GPS-параданных для контроля процесса сбора данных могут потребоваться значительные временны*е* и технические ресурсы [60]. Тем не менее использование GPS-параданных для контроля процесса сбора данных или методического аудита открывает возможности не только для снижения ресурсов, затрачиваемых на ручную перепроверку полученных данных и предотвращение фабрикаций или фальсификаций со стороны интервьюеров (снижение ошибки выборки, покрытия, неответов и измерения), но и для повышения эффективности методических процедур сбора данных в целом.

ЛИТЕРАТУРА НА РУССКОМ ЯЗЫКЕ[3]

1. *Власов В.* Вокруг плагиата // Медицинская газета. 2007. № 41. С. 11.
2. *Ипатова А.А.* Насколько разумна наша вера в результаты опросов, или Нарушение исследовательской этики в социологических исследованиях // Мониторинг общественного мнения: экономические и социальные перемены. 2014. № 3 (121). С. 26–39. DOI: 10.14515/monitoring.2014.3.02 EDN: SIGFPR
3. Контроль качества получаемой информации // ВЦИОМ. Дата обращения 13.07.2022. URL: https://old-ok.wciom.ru/kontrol_i_kachestvo_dannyh

[3] Полный список источников — в References.

4. *Лебедев Д.В.* Параданные: определение, типы, сбор и возможное применение // Мониторинг общественного мнения: экономические и социальные перемены. 2020. № 2 (156). С. 4–32. DOI: 10.14515/monitoring.2020.2.915 EDN: VTLLGY
5. *Лебедев Д.В., Богданов М.Б.* Переход с PAPI на CAPI: опыт интервьюеров и характеристики, влияющие на их ожидания // Мониторинг общественного мнения: экономические и социальные перемены. 2019. № 4 (152). С. 43–67. DOI: 10.14515/monitoring.2019.4.03 EDN: DPAZOU
6. *Мягков А.Ю., Журавлева И.В.* Эффект ожиданий интервьюера в персональном интервью // Социологический журнал. 2004. № 3–4. С. 6–26. EDN: PZQOLN
7. *Рогозин Д.М.* Зачем социологу изучать фабрикации // Телескоп: журнал социологических и маркетинговых исследований. 2017. № 5. С. 13–16. EDN: ZIDZHL
8. *Рогозин Д.М.* По(д)делки в «бумажном» поквартирном опросе // Мониторинг общественного мнения: экономические и социальные перемены. 2015. № 4 (128). С. 3–35. DOI: 10.14515/monitoring.2015.4.01 EDN: XBBCMF

СВЕДЕНИЯ ОБ АВТОРЕ

Лебедев Даниил Вадимович — стажер-исследователь, МЛИСИ НИУ ВШЭ, аспирант, школа по социологии НИУ ВШЭ; преподаватель, Кафедра МСиАСИ НИУ ВШЭ.
Электронная почта: zenon-daniil@yandex.ru

Дата поступления: 17.09.2022.

Sotsiologicheskiy Zhurnal = Sociological Journal. 2022. Vol. 28. No. 4. P. 8–33. DOI: 10.19181/socjour.2022.28.4.9313

Research Article
***Daniil V. Lebedev*[1]**
[1] HSE University.
20, Myasnitskaya str., 101000, Moscow, Russian Federation.

Using GPS-Paradata to Control the Data Collection Process: Review of Existing Methods and Analysis of GPS-Paradata Quality

Abstract. The use of face-to-face interviews is still a very common data collection method in social sciences. The danger associated with the use of such data collection methods is a reduction in the resulting survey data's quality due to interviewers' fabrications or falsifications, which in turn has led to the emergence of a large set of methods for controlling the data collection process, the focus of which is limited and bypasses the behavioral characteristics of interviewers. In this context, paradata and GPS-paradata are an important new tool for use in the process of quality control of collected data or as part of a methodological audit, allowing not only to potentially identify and prevent falsification

or fabrication by interviewers, but also to assess the correctness of methodological instructions. This article provides an overview of the available and practiced methods for using GPS-paradata in two main strategies (data point analysis and interviewer path analysis): geofencing, strand length, curbstoning, connecting interviews' locations and interviewer path analysis. The possibilities of using such control methods depend on the sample design and on the methodological features of the surveys in general. However, the use of GPS-paradata to control the data collection process is not in itself a surefire method for detecting interviewers' fabrications or falsifications, as it may be subject to technical inaccuracies or unintentional interviewer errors. It is a useful additional method aimed at identifying "suspicious" interviews which require the use of more resource-intensive methods of control (for example, repeated contact). In addition, the article presents an analysis of the quality of acquired GPS-paradata on the example of 26th wave of the RLMS–HSE, based on analyzing missing data and quality of measurements (HDOP). The results show that the quality of GPS-paradata can be related both to the region where the interview is conducted and to the characteristics of the interviewers.

Keywords: GPS-paradata; paradata; GPS-paradata quality; fieldwork monitoring; CAPI; face-to-face interview data quality; interviewers' falsifications and fabrications.

Acknowledgments. The author expresses deep gratitude to Aigul Klimova for her valuable advice during research and writing process.

Funding. The research was supported by the Russian Foundation for Fundamental Research. Grant project "Postgraduates" No. 20-311-90073.

For citation: Lebedev, D.V. Using GPS-paradata to control the data collection process: Review of existing methods and analysis of GPS-paradata quality. *Sotsiologicheskiy Zhurnal = Sociological Journal*. 2022. Vol. 28. No. 4. P. 8–33. DOI: 10.19181/socjour.2022.28.4.9313

REFERENCES

1. Vlasov V. Around plagiarism. *Meditsinskaya gazeta*. 2007. No. 41. P. 11. (In Russ.)
2. Ipatova A.A. Is our trust in survey results rational, or breaking the ethics in the social research. *Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsial'nye peremeny*. 2014. No. 3 (121). P. 26–39. DOI: 10.14515/monitoring.2014.3.02 (In Russ.)
3. Quality control of the information received. *VCIOM*. Accessed 13.07.2022. URL: https://old-ok.wciom.ru/kontrol_i_kachestvo_dannyh (In Russ.)
4. Lebedev D.V. Paradata: definition, types, collection, and possible uses. *Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsial'nye peremeny*. 2020. No. 2 (156). P. 4–32. DOI: 10.14515/monitoring.2020.2.915 (In Russ.)
5. Lebedev D.V., Bogdanov M.B. Transition from PAPI to CAPI: interviewers' experience and factors influencing their expectations. *Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsial'nye peremeny*. 2019. No. 4 (152). P. 43–67. DOI: 10.14515/monitoring.2019.4.03 (In Russ.)
6. Myagkov A.Yu., Zhuravleva I.V. The effect of interviewer expectations in a face-to-face interview. *Sotsiologicheskiy Zhurnal = Sociological Journal*. 2004. No. 3–4. P. 6–26. (In Russ.)
7. Rogozin D.M. Why should a sociologist study fabrications. *Teleskop: Zhurnal sotsiologicheskikh i marketingovykh issledovanii*. 2017. No. 5. P. 13–16. (In Russ.)
8. Rogozin D.M. Fabrication in paper-and-pencil door-to-door survey. *Monitoring obshchestvennogo mneniya: ekonomicheskie i sotsial'nye peremeny*. 2015. No. 4 (128). P. 3–35. DOI: 10.14515/monitoring.2015.4.01 (In Russ.)

9. Arceneaux T. Evaluating the computer audio-recorded interviewing (CARI) household wellness study (HWS) field test. *Proceedings of the American Statistical Association, Section on Survey Research Methods.* Chicago: ASA, 2007. P. 2811–2818.
10. Bhuiyan M.F., Lackie P. Mitigating survey fraud and human error: Lessons learned from a low budget village census in Bangladesh. *IASSIST Quarterly.* 2017. Vol. 40. No. 3. P. 20–26. DOI: 10.29173/iq398
11. Birnbaum B. *Algorithmic approaches to detecting interviewer fabrication in surveys.* Washington: University of Washington, 2012.
12. Birnbaum B., Borriello G., Flaxman A.D., DeRenzi B., Karlin A.R. Using behavioral data to identify interviewer fabrication in surveys. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* Paris: ACM, 2013. P. 2911–2920. DOI: 10.1145/2470654.2481404
13. Biruk C. Seeing like a research project: Producing "high-quality data" in AIDS research in Malawi. *Medical anthropology.* 2012. Vol. 31. No. 4. P. 347–366. DOI: 10.1080/01459740.2011.631960
14. Bredl S., Winker P., Kotschau K. A statistical approach to detect interviewer falsification of survey data. *Survey Methodology.* 2012. Vol. 38. No. 1. P. 1–10.
15. Bushery J.M., Reichert J.W., Albright K.A., Rossiter J.C. Using date and time stamps to detect interviewer falsification. *Proceedings of the American Statistical Association (Survey Research Methods Section).* Chicago: ASA, 1999. P. 316–320.
16. Cecchi R., Marquette R.J. 2010 Census Global Positioning System (GPS) Evaluation Report. *United States Census Bureau.* Accessed 13.07.2022. URL: https://www.census.gov/2010census/pdf/2010_Census_AC_Operational_Assessment.pdf
17. Choumert-Nkolo J., Cust H., Taylor C. Using paradata to collect better survey data: evidence from a household survey in Tanzania. *Review of Development Economics.* 2019. Vol. 23. No. 2. P. 598–618. DOI: 10.1111/rode.12583
18. Chung E.H., Shalaby A. A trip reconstruction tool for GPS-based personal travel surveys. *Transportation Planning and Technology.* 2005. Vol. 28. No. 5. P. 381–401. DOI: 10.1080/03081060500322599
19. Cornelius S.C., Sear D.A., Carver S.J., Heywood D.I. GPS, GIS and geomorphological field work. *Earth Surface Processes and Landforms.* 1994. Vol. 19. No. 9. P. 777–787. DOI: 10.1002/esp.3290190904
20. Couper M. Measuring survey quality in a CASIC environment. *Proceedings of the Survey Research Methods Section of the ASA at JSM1998.* Chicago: ASA, 1998. P. 41–49.
21. Couper M.P., Burt G. Interviewer attitudes toward computer-assisted personal interviewing (CAPI). *Social Science Computer Review.* 1994. Vol. 12. No. 1. P. 38–54. DOI: 10.1177/089443939401200103
22. Couper M.P., Kreuter F. Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society).* 2013. Vol. 176. No. 1. P. 271–286. DOI: 10.1111/j.1467-985X.2012.01041.x
23. Crespi L.P. Further Observations on the "Cheater" Problem. *Public Opinion Quarterly.* 1946. No. 4. Vol. 10. P. 646–649. DOI: 10.1086/265823.
24. Dajani N., Marquette R.Q. Reinterview Detection and Prevention at Census: New Initiatives. *Washington Statistical Society Curb-Stoning Seminar. Part III, Washington, DC.* Washington: WSS, 2015.
25. Durrant G.B., D'Arrigo J. Doorstep interactions and interviewer effects on the process leading to cooperation or refusal. *Sociological Methods & Research.* 2014. Vol. 43. No. 3. P. 490–518. DOI: 10.1177/0049124114521148
26. Durrant G.B., D'Arrigo J., Steele F. Analysing interviewer call record data by using a multilevel discrete time event history modelling approach. *Journal of the Royal*

Statistical Society: Series A (Statistics in Society). 2013. Vol. 176. No. 1. P. 251–269. DOI: 10.1111/j.1467-985X.2012.01073.x

27. Dwolatzky B., Trengove E., Struthers H., McIntyre J.A., Martinson N.A. Linking the global positioning system (GPS) to a personal digital assistant (PDA) to support tuberculosis control in South Africa: a pilot study. *International Journal of Health Geographics*. 2006. Vol. 5. No. 1. P. 1–6. DOI: 10.1186/1476-072X-5-34
28. Eyerman J., Murphy J., McCue C., Hottinger C., Kennet J. Interviewer falsification detection using data mining. *Proceedings: Symposium 2005, Methodological Challenges for Future Information Needs; Statistics Canada*. Toronto: Statistics Canada, 2005.
29. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS one*. 2009. Vol. 4. No. 5. P. e5738. DOI: 10.1371/journal.pone.0005738
30. Gong H., Chen C., Bialostozky E., Lawson C.T. A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems*. 2012. Vol. 36. No. 2. P. 131–139. DOI: 10.1016/j.compenvurbsys.2011.05.003
31. Groves B. Interviewer falsification in survey research: Current best methods for prevention, detection, and repair of its effects. *Survey Research*. 2004. Vol. 35. No. 1. P. 1–5.
32. Groves R.M., Lyberg L. Total survey error: Past, present, and future. *Public opinion quarterly*. 2010. Vol. 74. No. 5. P. 849–879. DOI: 10.1093/poq/nfq065
33. Haddaway S. *CAPI surveys on android devices in the developing world. Federal CASIC Workshops*. Washington, DC. Washington: CASIC, 2013.
34. Hasson M. Efficiency analysis through geospatial location evaluation (EAGLE). *Annual Federal Computer-assisted Survey Information Collection (FedCASIC) Workshop*. Suitland, MD. 2015.
35. Keating M., Loftis C., McMichael J., Ridenhour J. New dimensions of mobile data quality. *Federal CASIC Workshops. Washington, DC*. Washington: CASIC, 2014.
36. Kingori P., Gerrets R. Morals, morale and motivations in data fabrication: Medical research fieldworkers views and practices in two Sub-Saharan African contexts. *Social Science & Medicine*. 2016. Vol. 166. P. 150–159. DOI: 10.1016/j.socscimed.2016.08.019
37. Kozyreva P., Kosolapov M., Popkin B.M. Data Resource Profile: The Russia Longitudinal Monitoring Survey — Higher School of Economics (RLMS–HSE) Phase II: Monitoring the Economic and Health Situation in Russia, 1994–2013. *International journal of epidemiology*. 2016. Vol. 45. No. 2. P. 395–401. DOI: 10.1093/ije/dyv357
38. Krejsa E.A., Davis M.C., Hill J.M. Evaluation of the quality assurance falsification interview used in the census 2000 dress rehearsal. *Proceedings of the American Statistical Association (Survey Research Methods Section)*. Chicago: ASA, 1999. P. 635–640.
39. Lemmens M. *Geo-information: technologies, applications and the environment*. Vol. 5. N.Y.: Springer, 2011. DOI: 10.1007/978-94-007-1667-4
40. Li J., Michael Brick J., Tran B., Singer P. Using statistical models for sample design of a reinterview program. *Journal of Official Statistics*. 2011. Vol. 27. No. 3. P. 433.
41. McClain C.A., Couper M.P., Hupp A.L., Keusch F., Peterson G., Piskorowski A.D. & West B.T. A typology of web survey paradata for assessing total survey error. *Social Science Computer Review*. 2019. Vol. 37. No. 2. P. 196–213. DOI: 10.1177/0894439318759670
42. Mohadjer L., Edwards B. Paradata and dashboards in PIAAC. *Quality assurance in education*. 2018. Vol. 26. No. 2. P. 263–277. DOI: 10.1108/QAE-06-2017-0031
43. Montalvo J.D., Seligson M.A., Zechmeister E.J. Improving adherence to area probability sample designs: Using lapop's remote interview geo-locating of households in real-time (right) system. *Americas Barometer Methodological Note IMN004*. Nashville: IMN004, 2018.

44. Murphy J., Baxter R., Eyerman J., Cunningham D., Kennet J. A system for detecting interviewer falsification. *American Association for Public Opinion Research 59th Annual Conference*. Phoenix: AAPOR, 2004. P. 4968–4975.
45. Murphy J., Biemer P., Stringer C., Thissen R., Day O., Hsieh Y.P. Interviewer falsification: Current and best practices for prevention, detection, and mitigation. *Statistical Journal of the IAOS*. 2016. Vol. 32. No. 3. P. 313–326. DOI: 10.3233/SJI-161014
46. Olbrich L., Kosyakova Y., Sakshaug J.W. The reliability of adult self-reported height: The role of interviewers. *Economics & Human Biology*. 2022. Vol. 45. P. 101–118. DOI: 10.1016/j.ehb.2022.101118
47. Olson K., Wagner J. A feasibility test of using smartphones to collect GPS information in face-to-face surveys. *Survey Research Methods*. 2015. Vol. 9. No. 1. P. 1–13.
48. Robbins M., Johnson T.P., Pennell B.E., Stoop I.A.L., Dorer B. New frontiers in detecting data fabrication. *Advances in Comparative Survey Methods: Multicultural, Multinational and Multiregional Contexts (MC)*. Ed. by T.P. Johnson, B.-E. Pennell, I.A.L. Stoop, and B. Dorer. Hoboken, NJ: Wiley, 2018. DOI: 10.1002/9781118884997
49. Roth J.A. Hired hand research. *The American Sociologist*. 1966. Vol. 1. No. 4. P. 190–196.
50. Schäfer C., Schräpler J.P., Müller K.R., Wagner G.G. *Automatic identification of faked and fraudulent interviews in surveys by two different methods. DIW Discussion Papers*. Berlin: DIW, 2004. No. 441.
51. Schaefer C., Schräpler J.P., Müller K.R., Wagner G.G. Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch: Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts-und Sozialwissenschaften*. 2005. Vol. 125. No. 1. P. 183–193.
52. Schraepler J.P., Wagner G.G. Characteristics and impact of faked interviews in surveys — An analysis of genuine fakes in the raw data of SOEP. *Allgemeines Statistisches Archiv*. 2005. Vol. 89. No. 1. P. 7–20. DOI: 10.1007/s101820500188
53. Seeger J. A mobile, GPS-enabled listing application. *Federal CASIC Workshops*. Washington, DC: IEEE, 2011.
54. Shepherd J., Hill D., Bristor J., Montalvan P. Converting an ongoing health study to CAPI: Findings from the National Health and Nutrition Study. *Health survey research methods conference proceedings*. Baltimore: HSRM, 1996. P. 159–164.
55. Sikes N. Current trends in mobile technology for survey research. *Federal CASIC Workshops*. Washington, DC: CASIC, 2009.
56. Survey Solutions. *Survey Solutions*. Accessed 13.07.2022. URL: https://mysurvey.solutions/en/
57. Swanson D., Cho M., Eltinge J. Detecting possibly fraudulent or error-prone survey data using Benford's Law. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. Chicago: ASA, 2003. P. 937–941.
58. Thissen M.R. Computer audio-recorded interviewing as a tool for survey research. *Social Science Computer Review*. 2014. Vol. 32. No. 1. P. 90–104. DOI: 10.1177/0894439313500128
59. Thissen M.R., Myers S.K. Systems and processes for detecting interviewer falsification and assuring data collection quality. *Statistical Journal of the IAOS*. 2016. Vol. 32. No. 3. P. 339–347. DOI: 10.3233/SJI-150947
60. Wagner J., Olson K., Edgar M. Assessing Potential Errors in Level-of-E ort Paradata using GPS Data. *Survey research methods*. 2017. Vol. 11. No. 3. P. 219–233.
61. Wang K., Biemer P. The accuracy of interview paradata: Results from a field investigation. *Annual Meeting of the American Association for Public Opinion Research*. Chicago, IL (May). 2010.
62. Weisberg H.F. *The total survey error approach: A guide to the new science of survey research*. Chicago: University of Chicago Press, 2009.

63. Yabiku S.T., Glick J.E., Wentz E.A., Ghimire D., Zhao Q. Comparing paper and tablet modes of retrospective activity space data collection. *Survey research methods. NIH Public Access*. 2017. Vol. 11. No. 3. P. 329.
64. Zandbergen P.A. Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS*. 2009. Vol. 13. P. 5–25. DOI: 10.1111/j.1467-9671.2009.01152.x

Information about the author

Daniil V. Lebedev — research assistant ILSIR HSE University;
PhD student of Sociology, HSE University; faculty member,
Department of Sociological Research Methods, HSE University.
Email: zenon-daniil@yandex.ru

Received: 17.09.2022.