DOI: 10.14515/monitoring.2021.1.1760

М. Б. Богданов, И. Б. Смирнов

ВОЗМОЖНОСТИ И ОГРАНИЧЕНИЯ ЦИФРОВЫХ СЛЕДОВ И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В СОЦИОЛОГИИ

Правильная ссылка на статью:
Богданов М. Б., Смирнов И. Б. Возможности и ограничения цифровых следов и методов машинного обучения в социологии // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 1. С. 304—328. https://doi.org/10.14515/monitoring.2021.1.1760.
For citation:
Bogdanov M. B., Smirnov I. B. (2021) Opportunities and Limitations of Digital Footprints and Machine Learning Methods in Sociology. *Monitoring of Public Opinion: Economic and Social Changes.* No. 1. P. 304–328. https://doi.org/10.14515/monitoring.2021.1.1760. (In Russ.)

DOI: 10.14515/monitoring.2021.1.1760

ВОЗМОЖНОСТИ И ОГРАНИЧЕНИЯ ЦИФРОВЫХ СЛЕДОВ И МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В СОЦИОЛОГИИ

БОГДАНОВ Михаил Богданович — младший научный сотрудник Центра социологии культуры, Институт образования, аспирант департамента социологии, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: bogdanovmikle@mail.ru
https://orcid.org/0000-0001-6245-7178

СМИРНОВ Иван Борисович — ведущий научный сотрудник, заведующий лабораторией вычислительных социальных наук, Институт образования, Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
E-MAIL: ibsmirnov@hse.ru
https://orcid.org/0000-0002-8347-6703

Аннотация. В статье описываются возможности и ограничения использования в социологии новых источников данных и методов их сбора, обработки и анализа, а именно — цифровых следов и методов машинного обучения. Сначала обсуждаются недостатки классических источников данных — опросов, а затем, в контексте этих недостатков, на основе релевантных исследований анализируются возможности их преодоления с помощью цифровых следов. В качестве главных недостатков опросных данных, которым, в свою очередь, меньше подвержены цифровые следы, выделяются: реактивность, небольшой объем данных и редкая частотность. В контексте описания этих недостатков и способов их преодоления с помощью цифровых следов мы приводим типы исследова-

OPPORTUNITIES AND LIMITATIONS OF DIGITAL FOOTPRINTS AND MACHINE LEARNING METHODS IN SOCIOLOGY

Mikhail B. BOGDANOV[1] — Junior Research Fellow at the Centre for Cultural Sociology, Institute of Education; PhD student in Sociology
E-MAIL: bogdanovmikle@mail.ru
https://orcid.org/0000-0001-6245-7178

Ivan B. SMIRNOV[1] — Leading Research Fellow, Head of the Computational Social Science Lab, Institute of Education
E-MAIL: ibsmirnov@hse.ru
https://orcid.org/0000-0002-8347-6703

[1] National Research University Higher School of Economics, Moscow, Russia

Abstract. The article discusses the opportunities and limitations of using new data sources and methods of its collection, processing and analysis, namely, digital traces and machine learning in Sociology. At first, we examine the disadvantages of traditional data sources (surveys) and then, based on relevant and recent empirical studies, we discuss how these disadvantages can be overcome using digital traces. The main drawbacks of survey data are the reactivity, a small sample size, and rare frequency of surveys. Based on these drawbacks we identify types of research questions that can only be answered with digital traces. Finally, we also explore the disadvantages of digital traces: lack of representativeness, construct validity, external and internal interfering factors, and non-stationarity. Relying on recent

DOI: 10.14515/monitoring.2021.1.1760

тельских вопросов, на которые можно ответить только с помощью цифровых следов. После этого рассматриваем ограничения цифровых следов: нерепрезентативность, конструктную валидность, внешние и внутренние вмешивающиеся факторы, нестационарность. Затем, на основе актуальных методологических статей, мы описываем, как учитывать эти ограничения и по возможности корректировать их.

Ключевые слова: цифровые следы, большие данные, машинное обучение, предсказательное моделирование, вычислительные социальные науки, вычислительная социология, анализ данных, анализ текстов

Благодарность. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-311-90056.

methodological developments the paper explains how to take into consideration these limitations and how to adjust for them wherever possible..

Keywords: digital footprints, big data, machine learning, forecasting modeling, computational social sciences, computational sociology, data analysis, text analysis

Acknowledgments. The study was funded by the Russian Foundation for Basic Research (RFBR), project No. 20-311-90056.

Введение

Развитие цифровых технологий и широкое распространение интернета привели к стремительным изменениям в социальных науках, активному использованию новых источников данных, а также появлению новых методов их сбора, обработки и анализа. Возникшее таким образом направление исследований получило название «вычислительные социальные науки» (Computational Social Science) [Lazer et al., 2009]. Вычислительные социальные науки опираются на цифровые следы, то есть данные, получаемые из социальных сетей [Garcia et al., 2018; Garcia, Rimé, 2019; Jaidka et al., 2020; Sivak, Smirnov, 2019], поисковых систем [Bail, Brown, Wimmer, 2019; Stephens-Davidowitz, 2014], логи звонков [Blumenstock, Cadamuro, On, 2015] и других веб-сайтов [Abrahao et al., 2017; Lewis, 2013], показаний датчиков GPS [Stopczynski et al., 2014], оцифрованных текстов [Hills et al., 2019] и масштабных административных данных [Pierson et al., 2020]. Эти новые источники данных зачастую характеризуются большими объемами, высокой разрешающей способностью и слабой структурированностью, а для их анализа обычно используются алгоритмы машинного обучения и другие вычислительные методы. Исторически этим направлением занимались не социологи, большинство публикаций в этой сфере выходили в несоциологических журналах, и только в последние несколько лет это стало меняться [Edelmann et al., 2020; Golder, Macy, 2014; Lazer, Radford, 2017; Molina, Garip, 2019], поэтому для многих социологов

DOI: 10.14515/monitoring.2021.1.1760

в России остается непонятным, какие именно новые возможности открывают перед ними цифровые следы и вычислительные методы, а также какие трудности могут встретиться при работе с такого рода данными.

Задача этой статьи — ответить, опираясь на последние исследования в области вычислительных социальных наук, на следующие вопросы, которые современные социологи могут задавать сами себе: в каких случаях могут быть полезны цифровые следы? На какие новые исследовательские вопросы они позволяют ответить? Если я захочу использовать такие данные, то что нужно учитывать при планировании своего исследования?

В этом обзоре сначала кратко описываются ограничения таких классических источников данных для количественной социологии, как массовые опросы. Затем на примерах современных исследований описывается, как новые данные и подходы к их обработке и анализу могут помочь преодолеть эти ограничения. После этого мы описываем и обсуждаем типы исследовательских вопросов в количественной социологии, на которые нельзя было ответить с помощью опросов, но можно ответить с помощью новых данных. В конце приводятся и анализируются ограничения новых данных, а также даются некоторые общие рекомендации по сбору, обработке и анализу такого рода данных.

Ограничения опросных данных и возможности цифровых следов по их преодолению

Одним из главных ограничений классических массовых опросов является их *реактивность* — респонденты и испытуемые практически всегда знают, что участвуют в исследовании, а это, в свою очередь, может оказать эффект на результаты исследования [Lavrakas, 2008: 694]. Проблема в том, что оценить величину этого эффекта невозможно — мы не можем знать, как бы человек ответил на вопросы анкеты или как бы вел себя во время экспериментального воздействия, не дав ему или ей заполнить анкету или не поместив в экспериментальные условия. Цифровые же следы по своей природе *нереактивны* [Salganik, 2019: 23] или *малореактивны* [Девятко, 2018: 30], то есть не производятся специально в целях исследований, а представляют собой побочный продукт фиксации поведения людей в цифровых системах.

Проблема реактивности особенно актуальна в исследованиях сенситивных тем, так как реактивные методы, такие как массовые опросы, могут давать смещенные результаты из-за социально одобряемого поведения респондентов. В случае использования цифровых следов эта проблема становится менее выраженной. Например, данные о поисковых запросах в Google позволили проанализировать уровень расизма в разных штатах и административных единицах Америки. Затем эти данные использовались для предсказания результатов президентских выборов в США в 2008 и 2012 гг. Оказалось, что связь между уровнем расизма и результатами Барака Обамы на данных поисковых запросов была в 1,5—3 раза выше, чем по данным опросов [Stephens-Davidowitz, 2014].

С помощью этих агрегированных данных впоследствии была исследована взаимосвязь между расизмом и уровнем смертности среди темнокожего населения США [Chae et al., 2015], а также взаимосвязь между расизмом и неблагопри-

DOI: 10.14515/monitoring.2021.1.1760

ятными исходами родов среди темнокожих женщин — преждевременные роды и маленький вес ребенка при рождении [Chae et al., 2018].

На агрегированных по административным округам США данных поисковых запросов в Google была проанализирована причинно-следственная связь между антимусульманскими и протеррористическими настроениями в локальных сообществах [Bail, Merhout, Ding, 2018]. Исследование взаимосвязи этих феноменов с помощью опроса столкнулось бы с проблемами с обеих сторон: экстремисты не стали бы заявлять о своих радикальных взглядах в опросе, а люди с антимусульманскими установками могли бы преуменьшать их в своих ответах или не выражать совсем из-за социальной нежелательности таких взглядов [ibid.: 1].

В другом исследовании похожей тематики расовые предрассудки полицейских в США были исследованы на данных о почти 100 млн проверок машин дорожной полицией [Pierson et al., 2020].

Еще один пример связан с исследованием расовых предрассудков при выборе романтических партнеров на данных о поведении 126 тысяч пользователей онлайн-сервиса для знакомств «OkCupid» за 2,5 месяца [Lewis, 2013]. В этой работе анализировалась вероятность проявления и получения романтического интереса между двумя людьми в зависимости от их расовых и других социально-демографических характеристик. Подобное исследование было бы крайне сложно провести с помощью таких реактивных методов, как опрос или эксперимент: они позволили бы измерить намерения и интерес, а не реальное поведение в естественных условиях [ibid.: 18814].

Цифровые следы также могут служить валидацией результатов, полученных с помощью реактивных методов. Так, на данных сервиса краткосрочной сдачи и аренды жилья Airbnb было проведено исследование социальных предрассудков, в котором результаты онлайн-эксперимента валидировались с помощью анализа реальных поведенческих данных на сервисе [Abrahao et al., 2017].

Цифровые следы использовались и для изучения влияния социальной политики в области здравоохранения на уровень абортов [Reis, Brownstein, 2010], а также для изучения потребления наркотиков [Enghoff, Aldridge, 2019].

На данных 635 000 пользователей социальной сети «ВКонтакте» из Санкт-Петербурга было проанализировано гендерное неравенство в упоминании дочерей и сыновей родителями в социальных сетях [Sivak, Smirnov, 2019]. Оказалось, что родители, как матери, так и отцы, чаще упоминают в постах сыновей, чем дочерей. Более того, посты с упоминанием сыновей в среднем набирают на 50 % большей лайков [ibid.: 2040]. Авторы отмечают, что упоминание детей в постах представляет собой простую и напрямую измеримую поведенческую метрику гендерного неравенства, которую было бы сложно измерить с помощью опроса из-за социальной желательности таких вопросов [ibid.: 2040].

Другое существенное ограничение традиционных подходов — относительно *небольшой по сравнению с цифровыми следами объем* изучаемых выборок. Теоретически можно опросить всех людей, но на практике это сделать невозможно по множеству причин, среди которых ключевая — значительные затраты ресурсов, как финансовых, так и временны́х. Поэтому ученые ограничиваются выборочными опросами, размер выборки которых редко превышает несколько

DOI: 10.14515/monitoring.2021.1.1760

тысяч, иногда десятков тысяч, человек и не сравним с потенциальным размером данных цифровых следов.

Неоднородные зависимости

Сравнительно небольшой размер выборок в опросах ограничивает поиск неоднородных зависимостей и паттернов между изучаемыми концептами. Например, было показано, что использование социальных медиа не связано с удовлетворенностью жизнью у подростков [Orben, Dienlin, Przybylski, 2019]. Однако может оказаться, что связь существует, но она либо нелинейна, либо гетерогенна для разных групп подростков, либо нелинейна и гетерогенна одновременно. Возможность разбивать выборку на подгруппы по разным критериям и использовать модели, замеряющие нелинейные и гетерогенные по своей природе связи принципиально ограничена доступным размером выборки [Molina, Garip, 2019: 37]. Это ограничение может преодолеваться за счет использования как больших данных обособленно, так и вкупе с опросами.

Даже если какая-то информация не доступна в явном виде для большого количества людей, ее часто можно оценивать, используя модели машинного обучения на объединенных данных цифровых следов и анкет респондентов. Таким образом можно обогащать данные за счет предсказания значения переменных, измеренных с помощью опроса для тех людей, которые не участвовали в опросе, но чьи цифровые следы можно собрать. Такой подход некоторые исследователи называют усиленным вопрошанием (от англ. amplified asking) [Salganik, 2019: 122]. К примеру, многие исследования показали, что черты личности могут быть предсказаны по различным цифровым следам [Azucar, Marengo, Settann, 2018; Huang, 2019; Kosinski, Stillwell, Graepel, 2013; Settanni, Azucar, Marengo, 2018]. Также было показано, что образовательные результаты тоже могут быть предсказаны с помощью цифровых следов [Smirnov, 2020, 2018]. Другими словами, большие объемы данных позволяют анализировать данные более детализированно.

Кроме того, данные могут обогащаться не только на индивидуальном, но и на контекстуальном уровне, то есть на уровне групп, районов, регионов и т. д. Для социологии это представляет особый интерес, так как позволяет получить на агрегированном уровне информацию по тем переменным, которые не измеряются государственной статистикой. Так, на данных о дружеских связях в Facebook был создан индекс социальной связанности административно-территориальных регионов США, который измеряет относительное количество дружеских связей на Facebook между жителями разных административных округов США [Bailey et al., 2018].

С помощью агрегированных данных сервиса Facebook по созданию кампаний таргетированной рекламы было проанализировано гендерное неравенство в 217 странах [Garcia et al., 2018]. В похожем исследовании показано, что отношения количества женщин к количеству мужчин среди пользователей Facebook и Google в разных странах выступают сильными предикторами цифрового неравенства в стране [Kashyap et al., 2020].

На данных 1,53 млрд постов в Twitter было проанализировано географическое распределение субъективного благополучия в 1208 округах США [Jaidka et al., 2020]. При помощи методов обработки естественного языка и машинного обуче-

DOI: 10.14515/monitoring.2021.1.1760

ния исследователи оценили на данных Twitter показатели субъективного благополучия, агрегированные на уровне округов. Наилучшая оценка коррелировала с данными опроса 1,73 млн респондентов об удовлетворенности жизнью, счастье, беспокойстве и печали на уровне 0,51—0,64 [ibid.: 3]. Более того, эта модель показала свою устойчивость во времени: анализ подвыборок за 2012—2013 и 2015—2016 гг. продемонстрировал схожие результаты [ibidem]. Относительная стационарность во времени позволяет в будущем использовать эту модель для получения агрегированных данных о благополучии на уровне административных округов, а также о динамике этого феномена.

Другой хрестоматийный пример использования новых данных — это исследование географической гетерогенности межпоколенческой мобильности в США [Chetty et al., 2014]. В этом исследовании на индивидуальных данных налоговой статистики о 40 млн американцев была выявлена существенная гетерогенность в межпоколенческой мобильности по административным округам США. Если для Америки в целом вероятность попасть в верхний квинтиль по распределению доходов для ребенка 1980—1985 годов рождения из семьи из нижнего квинтиля равна 7,8%, то для разных округов эта вероятность может изменяться в несколько раз: от 4% до 13% [ibid.: 1596].

Также на данных административной статистики о смертности в штате Калифорния и данных 12 млн профилей на Facebook было выявлено, что социальная интеграция в сети, измеренная через социальные взаимодействия на Facebook (добавление в друзья, размещение постов и фотографий, отправка сообщений), связана со смертностью [Hobbs et al., 2016]. Более того, большой объем данных позволил проследить гетерогенность взаимосвязи для разных причин смертности и оказалось, что для рака такой связи нет, а для суицида и передозировки наркотиками есть [ibid.: 12983].

Предсказательное моделирование

Кроме того, наличие существенно больших по размеру данных позволяет ставить исследовательские вопросы другого характера. Речь идет прежде всего о *задачах и вопросах предсказания*.

Статистическое моделирование можно разделить на два типа: объяснительное (от англ. explanatory) и предсказательное (от англ. — predictive) [Shmueli, 2010: 290—291]. Если объяснительное моделирование — это применение статистических моделей для проверки гипотез о корреляциях и причинно-следственных связях между теоретическими конструктами [ibid.: 291], то предсказательное моделирование — это применение статистических моделей для предсказания новых (вневыборочных) наблюдений [ibid.: 292].

В социальных науках доминируют объяснительные модели, а вопросам предсказания уделяется существенно меньше внимания [Cranmer, Desmarais, 2017; Hofman, Sharma, Watts, 2017; Shmueli, 2010]. Вероятно, одна из причин заключается в том, что для использования более мощных с точки зрения предсказательной силы моделей машинного обучения (нейронные сети, бустинги и т. п.) необходимы существенно бóльшие выборки [Yarkoni, Westfall, 2017], чем те, которые обычно собираются с помощью опросов.

DOI: 10.14515/monitoring.2021.1.1760

Например, недавний конкурс по предсказательному моделированию на данных когортного лонгитюдного исследования «Fragile Families» показал, что на опросных данных, пусть даже очень хорошего качества, возможности предсказания существенно ограничены [Salganik et al., 2020]. Это исследование примечательно тем, что в нем сотни исследовательских команд со всего мира пытались предсказать с помощью имеющихся данных и самых разнообразных моделей машинного обучения различные образовательные и жизненные показатели 15-летних школьников. R-квадрат лучшей предсказательной модели среднего балла по школьным предметам составил всего 0,19, то есть лучшая модель смогла предсказать всего 19 % вариации зависимой переменной [Salganik и др., 2020: 8340].

Некоторые исследователи связывают такой невысокий результат с тем, что возможности классических, как по размеру, так и по качеству, опросных данных для предсказательного моделирования ограничены и следует апробировать этот подход на бóльших по размеру данных [Garip, 2020: 2]. Кроме того, вероятно, что те теоретические концепты, которые операционализируются в анкетные вопросы, недостаточно хорошо схватывают реальность [ibidem]. В этом контексте использование цифровых следов может дать заметно больше информации о тех нишах и тех аспектах реальности, которые не удается измерить с помощью анкетных вопросов.

Например, на данных 852 млн твитов 51,6 млн пользователей исследователям удалось предсказать количество ретвитов (репостов), которое наберет конкретный твит в зависимости от характеристик автора и текста этого твита [Martin et al., 2016]. Лучшая модель предсказывала почти половину дисперсии зависимой переменной — R-квадрат = 0,48.

Бóльшие выборки необходимы еще и потому, что для корректного измерения предсказательной силы моделей машинного обучения требуется разбиение выборки на подвыборки. Это связано с тем, что современные методы машинного обучения, такие как искусственные нейронные сети, используют модели, содержащие большое количество параметров, что может приводить к переобучению моделей [Yarkoni, Westfall, 2017: 1110—1111]. Чтобы избежать переобучения, необходимо разбивать исходную выборку как минимум на три части, тренировочный набор, валидационный набор и тестовый набор. Тренировочный набор используется для построения модели, предсказывающей зависимую переменную по информации о независимых. Валидационный набор используется для проверки, насколько точно эта модель предсказывает значения зависимой переменной на новых данных [Molina, Garip, 2019: 28—30; Yarkoni, Westfall, 2017: 1103—1104]. Исследователи обычно стремятся выбрать модель, которая максимизирует точность на валидационном наборе данных, но это в свою очередь может тоже привести к завышению оценки предсказательной точности модели. Поэтому для финальной оценки используется тестовый набор данных, который откладывается в самом начале исследования и используется только один раз для тестирования предсказательной силы финальной модели [Molina, Garip, 2019: 31—32; Yarkoni, Westfall, 2017: 1113].

Тем не менее в силу необходимости разделения выборки на обучающую и тестовую очевидно, что для задач и исследовательских вопросов предсказательного

DOI: 10.14515/monitoring.2021.1.1760

характера требуется больше данных, чем для классического статистического моделирования, в котором модели обучаются и оцениваются на одних и тех же данных.

Взаимодействия людей и социальные сети

С помощью выборочных опросов значительно сложнее исследовать *взаимодействия людей* и реконструировать масштабные *социальные сети*. Это связано с тем, что для точного воспроизводства структуры сети необходимы данные о каждом узле, в противном случае, если нет данных о ключевых узлах, например, соединяющих разные клики сети, вся структура реконструированной сети может быть некорректной [Kossinets, 2006].

Кроме того, люди плохо помнят, с кем они взаимодействовали, и могут по-разному воспринимать такие вопросы [Lazer, Radford, 2017: 23]. Тогда как социальные сети, построенные на цифровых следах (лайки, репосты, комментарии и т. п.), могут охватывать значительно бóльшие сети. Например, на данных о дружбе санкт-петербургских школьников из «ВКонтакте» была воссоздана сеть цифровой близости между школами Санкт-Петербурга и показано, что академические успехи школы сильно связаны с показателями ее цифровых школ-соседей, но совершенно не связаны с географическими [Smirnov, 2019: 5]. Очевидно, что невозможно было бы опросить всех или почти всех школьников города об их взаимодействиях друг с другом.

Цифровые репрезентации социальных сетей, построенных на данных популярных сетевых онлайн-платформ вроде Facebook, «ВКонтакте», Twitter и других, предоставляют гигантские возможности для тестирования классических социологических теорий и гипотез.

Например, теория Марка Грановеттера о силе слабых связей была апробирована на сетевых данных 6 млн американских пользователей Facebook [Gee, Jones, Burke, 2017]. В этой работе сравнивалось влияние сильных и слабых связей на трудоустройство. В качестве показателя связи выступала дружба на Facebook, а сила связей измерялась через количество совместных фотографий, постов друг у друга и общих друзей [ibid.: 493—496]. С помощью этих данных было показано, что, несмотря на важность слабых связей, наличие хотя бы одной сильной связи более ценно с точки зрения поиска работы и трудоустройства [ibid.: 485].

Впоследствии авторы использовали этот подход и сетевые данные Facebook о 17 млн дружеских связей пользователей из 55 стран и показали, что эта зависимость сохраняется во всех странах [Gee et al., 2017]. Более того, сила эффекта зависит от экономического неравенства в стране — чем оно выше, тем выше вероятность, что человек найдет следующую работу через сильную связь [ibid.: 370].

Также на данных о 1,5 млн обменов подарками на Facebook протестирована классическая теория Марселя Мосса о природе дарообмена [Mauss, 2000] и выявлено, что получение подарка увеличивает вероятность дарения в будущем на 56 % [Kizilcec et al., 2018].

Социальные сети влияют и на принятие решений на рынке недвижимости [Bailey et al., 2018]. На данных о дружбе в Facebook выявлено, что люди, чьи географически отдаленные друзья недавно испытали повышение цен на недвижимость в их районе, более склонны к переходу от съема жилья к покупке собственного, причем большего по размеру [Bailey et al., 2018: 2224].

DOI: 10.14515/monitoring.2021.1.1760

Это исследование также интересно тем, что в нем используются опросные данные для валидации и уточнения результатов, полученных на данных цифровых следов. После анализа данных из Facebook исследователи пришли к выводу, что рост цен на недвижимость друзей, живущих в других районах, может воздействовать на экономическое поведение в сфере недвижимости через улучшение ожиданий относительно покупки недвижимости в районе индивида как способа финансовых инвестиций. Для валидации этого вывода в Facebook был проведен вспомогательный опрос, в котором в том числе оценивалась привлекательность покупки недвижимости в районе проживания респондента как способ финансового инвестирования, а также собирались данные о дружеских сетях респондентов на Facebook. Вспомогательный анализ на данных опроса подтвердил вывод: увеличение цен на недвижимость в районах проживания друзей связано с более оптимистичными установками относительно покупки недвижимости [Bailey, Cao, Kuchler, Stroebel, 2018: 2266—2269].

На данных фотографий и дружеских сетей из Facebook также было проанализировано распространение культурных трендов. С помощью нейронных сетей из изображений была извлечена информация о предметах, местах и категориях, изображенных на них, затем композиция характеристик изображений сравнивалась с фотографиями из профилей друзей, было выявлено, что друзья склонны выкладывать фотографии со схожими культурными трендами. Сетевые данные также позволили отделить эффект социального влияния от эффекта гомофилии[1] и показать, что культурные тренды склонны распространяться по социальным сетям за счет воздействия одних людей на других.

В онлайн-эксперименте на 61 млн пользователей Facebook было выявлено, что сообщения с политической мобилизацией оказывают воздействие не только на самих получателей, но также и на их друзей и друзей друзей [Bond et al., 2012]. Очевидно, что исследование третичных эффектов, то есть воздействия сообщений на друзей друзей получателя, было бы невозможно осуществить с помощью опросных данных из-за проблем с неответами, неполной и неточной информацией о социальных связях и т. п.

Сетевые данные Facebook также использовали для изучения доверия. На основе опроса 6 тыс. пользователей и данных о группах на Facebook, в которых они состоят, и их характеристик было проанализировано общее доверие и доверие по отношению к участникам группы, из которой их рекрутировали [Ma et al., 2019]. Оказалось, что позиция индивида в дружеской сети участников группы связана с уровнем его/ее общего доверия [ibid.: 8].

Труднодоступные и маленькие группы

С помощью опросов сложно исследовать труднодоступные, закрытые и маленькие группы. Если размер группы небольшой и нет конкретных локаций ее обитания (как территориальных, так и виртуальных), по которым можно таргетировать опрос, то количество контактов, необходимых для достижения выборки нужного размера, может быть очень большим и, как следствие, стоимость проведения

[1] Гомофилия — свойство социальной сети, при котором узлы с похожими атрибутами склонны формировать между собой связи.

DOI: 10.14515/monitoring.2021.1.1760

такого опроса также может быть очень высокой. Кроме этого, члены закрытых групп могут быть не склонны участвовать в опросе. Например, вряд ли удастся исследовать с помощью опроса радикальных футбольных фанатов, состоящих в фанатских группировках и участвующих в организованных драках.

Цифровые же следы позволяют наблюдать за труднодоступными, закрытыми и маленькими группами через социальные медиа. Так, на данных 10 млн мигрантов в США и их сетях дружбы на Facebook был проанализирован уровень интеграции мигрантов из разных стран в американское общество [Herdağdelen et al., 2016]. В другом исследовании схожей тематики анализировалась культурная ассимиляция мигрантов из Мексики в США на данных сервиса Facebook для таргетированной рекламы о музыкальных предпочтениях той или иной аудитории [Stewart et al., 2019]. Размер и охват данных Facebook позволил проанализировать феномен культурной ассимиляции в разных социально-демографических группах [ibid.: 3259]. Кроме того, поскольку данные о музыкальных предпочтениях вносятся пользователями самостоятельно, это, в отличие от ограниченных анкетных вопросов, позволяет зафиксировать больший спектр музыкальных предпочтений и вкусов [ibidem].

Также цифровые следы можно использовать для таргетирования опросов на труднодоступные группы. Например, с помощью сервиса Facebook для создания таргетированной рекламы можно рекрутировать респондентов не только по их демографическим характеристикам, но и по поведению в сети, интересам и другим данным, собирающимся в интернете [Iannelli et al., 2018].

Редкая частотность

Еще одно существенное ограничение опросных данных — их относительно *редкая частотность*. Проведение опросов трудозатратно: нужно разработать анкету, сконструировать выборку и провести полевой этап сбора данных, длительность которого также зависит от метода сбора данных. Это, в свою очередь, ограничивает частоту проведения опросов и, следовательно, делает невозможным достаточно детализированное во времени изучение динамики некоторых феноменов. Опросные компании проводят еженедельные опросы (так называемые омнибусы), а иногда и ежедневные, но здесь возникают ограничения, связанные с объемами таких ежедневных выборок. При сравнительно небольших объемах выборочные показатели страдали бы от случайных флуктуаций, то есть от случайной ошибки выборки.

Кроме того, в долгосрочной перспективе на это может накладываться усталость людей от опросов, которая сейчас наблюдается в индустрии [Wojcik, Hughes, 2019]. А если говорить о лонгитюдном исследовании, то респонденты могут уставать от участия в опросе и отказываться от участия в исследовании. Если такое осыпание панели (от англ. panel attrition) происходит не случайным образом, то это может приводить к смещениям в выборке. Кроме того, сам опыт участия в лонгитюдном исследовании влияет на ответы респондентов (в англ. panel conditioning) [Warren, Halpern-Manners, 2012].

Преимущество цифровых следов заключается в том, что они всегда доступны и их можно собрать ретроспективно за весь интересующий промежуток времени. Таким образом, цифровые следы могут быть доступны в намного более частотном разрезе, чем опросы: в разрезе недель, дней и даже часов. Так, посты из Twitter

DOI: 10.14515/monitoring.2021.1.1760

позволили проанализировать динамику коллективных эмоций (то, что Дюркгейм называл «effervescence») до и после террористических атак в Париже в 2015 г. в разрезе месяцев, недель и даже дней [Garcia, Rimé, 2019]. Очевидно, что такое исследование было бы практически невозможно осуществить с помощью опроса по следующим причинам.

Во-первых, для замера эмоций до и после ключевого события было бы необходимо, чтобы уже на протяжении какого-то времени проводился опрос, замеряющий релевантные показатели. Ретроспективный опрос о динамике и изменениях коллективных эмоций видится слабым инструментом в силу давно известной неточности ретроспективных оценок респондентов, особенно оценок таких эфемерных концептов, как установки, отношение и восприятие [Smith, 1984]. Во-вторых, даже если бы такие замеры проводились, они все-равно не могли бы быть настолько детализированными во времени по описанным выше причинам. В-третьих, с помощью опроса было бы сложнее адекватно измерить взаимодействие людей в контексте обсуждения террористических атак и таким образом измерить коллективную составляющую эмоций.

В другой работе на данных 509 млн постов в Twitter 24 млн людей из 84 стран исследовалась динамика индивидуального настроения в течение дня и недели [Golder, Macy, 2011: 1879—1880]. Наличие таких данных позволило рассмотреть суточную и недельную динамику позитивных и негативных аффектов не просто для выборки в целом, но и в разрезе разных социокультурных сред [ibid.: 1879].

Такое исследование было бы невозможно провести с помощью опроса не только потому, что для этого потребовалось бы опрашивать существенное количество людей об их эмоциях каждый час на протяжении года, но это пришлось бы делать сразу в нескольких странах. Кроме того, неизвестно, как связано намерение участвовать в опросе и собственно настроение, если, например, люди в момент плохого настроения менее склонны участвовать в опросах, то это влечет за собой смещения в оценке изучаемого феномена. Однако стоит отметить, что и в случае постов в Twitter могут быть смещения — например, пользователи в зависимости от настроения могут быть более или менее склонны писать посты.

Цифровые следы использовались в качестве данных лонгитюдного формата и для ретроспективных страновых исследований. С помощью текстовых данных из оцифрованных книг, опубликованных за последние 200 лет, исследовалось субъективное благополучие на уровне страны и связанные с ним факторы [Hills et al., 2019]. На основе эмоционального оттенка слов, встречавшихся в этих текстах, был подсчитан национальный индекс валентности для США, Англии, Германии и Италии. Этот показатель валидировался с помощью данных Евробарометра — опроса, проводящегося с 1970 г. в этих странах. Корреляция между анкетным вопросом об удовлетворенности жизнью и индексом валентности составила 0,53 [ibid.: 1271—1272]. Это позволило проследить динамику субъективного благополучия в четырех странах с 1820 г. Также были построены модели, оценивающие влияние различных исторических, социальных и экономических факторов на уровень субъективного благополучия в стране [ibid.: 1274].

В похожем по дизайну исследовании на данных миллионов оцифрованных книг, опубликованных за последние 100 лет, изучалось, как изменились ассоциации

DOI: 10.14515/monitoring.2021.1.1760

и смыслы, вкладываемые в понятие «социальный класс», а также как эти изменения накладывались на социально-экономические трансформации [Kozlowski, Taddy, Evans, 2019]. В еще одном исследовании со схожим дизайном и методами исследовались динамика и изменение гендерных и этнических стереотипов в США за последние 100 лет, а результаты валидировались с помощью данных переписей населения и государственной статистики [Garg et al., 2018].

Еще один пример использования цифровых следов для исследования динамики и распространения социальных феноменов — работа Кристофер Бейла и коллег про глобальную диффузию вкусов, интересов и потребительских предпочтений на данных поисковых запросов в Google [Bail, Brown, Wimmer, 2019: 1496]. Это исследование примечательно еще и тем, что проверяет классическую теорию социальной имитации французского социального психолога Габриэля Тарда. В работе собраны данные о десяти самых популярных ежемесячных поисковых запросах в Google в 199 странах мира в период с 2004 по 2014 г. На этих данных построена сеть диффузии культурных вкусов и интересов между странами, а также модель, оценивающая взаимосвязь интенсивности диффузии культурных вкусов и интересов из одной страны в другую в конкретный месяц в зависимости от переменных, отражающих статус стран в различных областях (экономика, спорт, искусство, наука, политика, образование и т. д.), истории политического взаимодействия между странами, их географической, культурной и социальной близости и некоторых других факторов [ibid.: 1530—1539].

Для организации такого рода исследования с помощью опроса нужно было бы проводить ежемесячные опросы в 199 странах мира на протяжении десяти лет. Помимо огромных затрат просто на сбор данных, потребовались бы огромные усилия ученых и экспертов со всего мира по стандартизации и унификации анкет для 73 языков, разработке и постоянному обновлению списков культурных практик, предпочтений и интересов для самых разнообразных социокультурных контекстов, а также обработке и подготовке баз данных. Объем работы и финансовые затраты на организацию и проведение такого исследования можно сравнить с запуском нового «Всемирного исследования ценностей». Вышеупомянутое исследование К. Бейла и коллег было сделано усилиями трех ученых и девяти специально обученных ассистентов [ibid.: 1513].

Похожее исследование распространения культурных трендов, а именно моды, между городами всего мира было проведено на данных проекта GeoStyle, который собирает общедоступные фотографии, опубликованные в Instagram и Flickr [Al-Halah, Grauman, 2020].

Сотрудники Microsoft предложили подход к использованию цифровых следов как к несовершенным непрерывным панельным исследованиям [Diaz et al., 2016]. Поскольку сбор данных из социальных сетей не так трудозатратен, как проведение панельного опроса, по цифровым следам из социальных сетей можно наблюдать за некоторой выборкой пользователей на протяжении какого-то времени, таким образом, по сути образуя панельное исследование. Более того, поскольку, как мы уже говорили, цифровые следы можно собирать ретроспективно, то организация и проведение таких цифровых панельных исследований возможны постфактум.

DOI: 10.14515/monitoring.2021.1.1760

В качестве примера такого подхода можно привести работу Э. Кикимана и коллег, в которой на данных 658 млн постов в Twitter 63 тыс. студентов вузов за пять лет исследовалось влияние раннего употребления алкоголя в университете на различные жизненные и образовательные показатели [Kiciman, Counts, Gasser, 2018]. С помощью анализа текста авторам удалось вычленить среди всех англоязычных пользователей Twitter студентов, недавно поступивших в университеты, а также определить посты о потреблении алкоголя. Затем при помощи методов причинно-следственного анализа было выявлено, что при статистическом контроле многочисленных факторов, также измеренных на данных твитов, студенты, упоминавшие алкоголь во время первого семестра обучения, затем реже упоминали работу и успехи в учебе, но чаще упоминали рискованное поведение, проблемы с законом, их твиты выражали меньше позитивных эмоций [ibid.: 178].

Ограничения цифровых следов и способы их преодоления

Так как цифровые следы — сравнительно новый для современных социологов источник данных, особенно в России, освещение их недостатков и особенностей необходимо для лучшего понимания того, как работать с подобными данными. В этом разделе мы перечислим типичные проблемы, с которыми сталкиваются ученые при построении и реализации исследований с помощью цифровых следов и других новых источников данных.

Нерепрезентативность

Пользователи социальных медиа, таких как Facebook, Twitter, «ВКонтакте», Instagram, которые зачастую служат провайдерами цифровых следов для исследователей, не репрезентируют те популяции, к которым привыкли социологи, конструируя выборки массовых опросов: население страны, региона, города, определенные социальные группы и т. д. [Salganik, 2019: 29—33; Lewis, 2015: 1—2]. Зачастую пользователи таких платформ более молодые, образованные и обеспеченные по сравнению с населением страны, в которой они живут [Hargittai, 2020]. Пользователи Facebook в Америке значительно отличаются от всего населения США по данным переписи населения по возрасту, образованию и доходу [Ribeiro, Benevenuto, Zagheni, 2020]. В Европе были выявлены серьезные межпоколенческие различия в использовании сайтов социальных сетей среди пожилых людей, а также существенная межстрановая вариация в этих различиях [Sala, Gaia, Cerati, 2020].

В России также существуют значительные различия между пользователями социальных сетей [Богданов, Лебедев, 2017]. Так, использование «Одноклассников» выше в более возрастных группах и в малых городах и селах, а «ВКонтакте» популярнее среди более молодых поколений и в более крупных городах [там же: 138—139]. Доля пользователей Facebook больше в крупных городах, а также среди людей с высшим образованием [там же].

Но для разных платформ могут быть характерны разные смещения, варьируются смещения и от страны к стране [Hargittai, 2020: 11—19]. Кроме того, некоторые пользователи платформ могут быть более опытными и активными пользователями интернета и соответствующих платформ [ibid.: 16—19]. Неравномерное распре-

DOI: 10.14515/monitoring.2021.1.1760

деление активности пользователей на тех или иных платформах выражается, например, в том, что на 10 % активных пользователей Twitter из США приходится 80 % всех постов на этой платформе в этой стране [Wojcik, Hughes, 2019]. Это может приводить к тому, что анализ, основанный на таких данных, может в большей степени отражать закономерности поведения наиболее активных пользователей.

Что с этим можно сделать?

Методы корректировки смещений данных цифровых следов во многом основываются на методах, используемых для взвешивания невероятностных выборок [Beręsewicz et al., 2018: 17—89; Wang et al., 2019]. В качестве наиболее типичного подхода можно назвать постстратификацию, когда данные делятся по некоторым переменным на группы, а затем эти группы перевзвешиваются так, чтобы их размер (доля в выборке) совпадал с размером в генеральной совокупности [Kolenikov, 2016].

Например, для данных из Twitter была разработана модель, которая с помощью нейронных сетей предсказывает принадлежность аккаунта человеку или организации, а также пол и возраст на основе фотографии профиля, имени, никнейма и небольшого описания профиля, причем делает это для различных европейских языков [Wang et al., 2019]. На предсказанных данных, используется многоуровневое регрессионное моделирование для расчета постстратификационных весов для поло-возрастных групп по каждому региону внутри большинства европейских стран [ibidem]. Этот подход позволяет оценить вероятность присутствия человека определенного пола и возраста из определенного региона и страны в Twitter и на основе этих данных скорректировать смещенность выборки по полу и возрасту для конкретного региона или страны [ibid.: 2065—2066]. Предобученные нейронные сети для определения вышеупомянутых характеристик пользователя, а также код для использования этих моделей находятся в открытом доступе и могут быть свободно использованы[2] [ibid.: 2066].

Для корректировки смещений в данных из Facebook можно использовать подход, апробированный на американской популяции пользователей в [Ribeiro, Benevenuto, Zagheni, 2020].

Однако эти методы применимы только если склонность к присутствию или конкретной активности не связана с исследуемым феноменом. В противном случае методы корректировки смещенности данных, полученных с этой платформы, не исправят положение вещей. В этом смысле чрезвычайно важно различать две категории исследовательских вопросов: 1) исследование феноменов, происходящих на платформе социальных медиа, и производство выводов о пользователях платформы, 2) исследование с помощью данных о пользователях той или иной платформы феноменов, существующих вне этой платформы и, вероятно, охватывающих не только пользователей конкретной платформы [Olteanu и др., 2019: 3—4].

По-видимому, как в случае и с опросными данными, для цифровых следов пока не существует идеальных инструментов, позволяющих корректировать произвольные смещения в данных. Поэтому детальная и дотошная фиксация потенциальных

[2] Все материалы доступны по ссылке: https://github.com/euagendas/.

DOI: 10.14515/monitoring.2021.1.1760

смещений является витальным аспектом работы, в том числе и с данными цифровых следов [Ruths, Pfeffer, 2014: 1063].

Один из способов проверить надежность результатов исследования — апробировать исследование на данных других онлайн-платформ [ibidem].

Конструктная валидность

Цифровые следы и другие новые данные представляют собой побочный продукт функционирования различных онлайн-сервисов, мобильных приложений и интернет-платформ. Природа происхождения этих данных зачастую никак не связана с теми исследовательскими вопросами, на которые с помощью них могут пытаться ответить ученые [Salganik, 2019: 25—29].

Это влечет за собой проблему, связанную с конструктной валидностью, — теоретические концепты, на которых основывается исследование, могут не быть напрямую отражены в данных, процесс генерации которых никак не связан с целями исследования [Salganik, 2019: 25—29; Lazer, 2015; Rafaeli, Ashtar, Altman, 2019]. Если в случае анкетных вопросов операционализация, то есть переход от теоретического концепта к его измеримым показателям, происходит до исследования и контролируется исследователем, то в случае данных цифровых следов это происходит ad hoc, после того как данные уже собраны.

Очевидно, что проблема конструктной валидности в большей степени относится к таким латентным, ненаблюдаемым и неповеденческим концептам, как, например, дружба, ценности, социальные установки, предубеждения и т. д., и в меньшей степени к более однозначным переменным вроде возраста, пола, семейного статуса, дохода и т. п. [Lazer, Radford, 2017: 30; Olteanu et al., 2019: 4].

Так, было показано, что основанные на данных из Twitter метрики политической поддержки кандидатов и партий скорее говорят о временном повышении интереса общественности к политике и политическим фигурам, а не о намерении голосовать за тех или иных политических акторов [Jungherr et al., 2017].

Социальные сети дружбы, измеренные через поведение на онлайн-платформах, сильно разнятся в зависимости от используемых метрик интенсивности «дружбы» [Golder, Macy, 2014: 142], а также заметно отличаются от дружбы, измеренной через анкетные вопросы [Gilbert, Karahalios, 2009].

Что с этим можно сделать?

Один из способов проверки конструктной валидности — использование и сравнение нескольких метрик, потенциально измеряющих один и тот же конструкт. Например, в описанном в предыдущем разделе исследовании территориального распределения благополучия на данных Twitter сравнивалось сразу нескольких моделей предобработки и анализа текста [Jaidka et al., 2020: 2—3]. Причем сравнение проводилось с условно истинными значениями изучаемых концептов [разными аспектами благополучия], измеренными с помощью анкетных вопросов [ibid.: 3—4].

Быстро и без траты дополнительных ресурсов также можно проверить конструктную валидность исследования, сформулировав результаты не в терминах изучаемых концептов, а напрямую в тех метриках, которые использовались в анализе [Salganik, 2019: 25—30].

DOI: 10.14515/monitoring.2021.1.1760

Беспроигрышный способ добавить в анализ нужные переменные — собрать дополнительные данные [ibidem]. Например, можно попросить респондентов дать согласие на сбор их данных из социальных сетей и таким образом дополнить данные опроса цифровыми следами респондентов. Подробнее о возможностях, проблемах и способах соединения опросных данных и цифровых следов можно прочесть в [Baghal et al., 2019; Stier et al., 2019].

Внешние и внутренние вмешивающиеся факторы

Работая с цифровыми следами, нужно понимать, что эти данные в подавляющем большинстве случаев являются побочным продуктом функционирования коммерческих платформ, основная цель которых — это коммерческая деятельность [Salganik, 2019: 60—80; Olteanu et al., 2019: 10]. Поэтому процесс производства цифровых следов во многом определяется алгоритмами функционирования этих платформ и их изменениями. Если некоторые алгоритмы функционирования платформ и изменения в этих алгоритмах видны и доступны для исследователей и обычных пользователей, то другие нет и составляют коммерческую тайну компании.

Например, давно известный феномен социальных сетей людей к формированию закрытых триад — друг моего друга с большей вероятностью тоже будет моим другом — используется онлайн-платформами для рекомендации потенциальных друзей пользователям [Salganik, 2019: 60—80; Ruths, Pfeffer, 2014: 1063]. Однако мы не знаем, как именно онлайн-платформы используют это и на что еще опираются, рекомендуя друзей и подписки своим пользователям. В этом смысле сетевая структура, полученная на данных такой платформы, прямым образом зависит от алгоритмов, используемых на платформе, и вопрос об отделении эффекта алгоритмов от социальных эффектов при формировании сети остается открытым [Ruths, Pfeffer, 2014: 1063].

Еще один пример влияния алгоритмов онлайн-платформ — это персонализированные результаты поисковой выдачи в поисковиках [Olteanu et al., 2019: 11]. Такие алгоритмы являются собственностью компаний и засекречены, поэтому, например, исследования цифрового поведения на основе результатов выдачи по поисковым запросам могут быть смещены, и измерить это смещение будет чрезвычайно сложно, если вообще возможно.

Недавно Instagram при поиске фотографий по хештегу #море выдавал пользователям предупреждение о том, что «публикации со словами или тегами, которые вы ищете, зачастую поддерживают поведение, которое может привести к причинению вреда или даже смерти. Если вы столкнулись с какими-то трудностями, мы всегда готовы помочь вам»[3]. Такое предупреждение может быть результатом тестирования алгоритмов определения суицидальных намерений и их профилактики. Очевидно, что это нововведение может серьезным образом изменить поведение людей на онлайн-платформе и, как следствие, повлиять на процесс генерации данных.

Что с этим можно сделать? К сожалению, мы мало что можем сделать с теми факторами, о которых мало что знаем и которые зачастую составляют коммер-

[3] Инстаграм отобрал последнюю радость — возможность смотреть на фотографии моря 🌊. Приложение пишет, что такие снимки могут «привести к смерти» // Meduza. 2020. 5 мая. URL: https://meduza.io/shapito/2020/05/05/instagram-otobral-poslednyuyu-radost-vozmozhnost-smotret-na-fotografii-morya [дата обращения: 20.02.2021].

DOI: 10.14515/monitoring.2021.1.1760

ческую тайну. Однако что касается открытых и заметных для пользователей изменений в функционировании онлайн-платформ, то мы можем фиксировать эти изменения и учитывать при интерпретации результатов анализа. Кроме того, если это позволяют данные, то можно сравнивать результаты анализа при различных состояниях онлайн-платформ — например, до и после какого-то нововведения.

Нестационарность

Еще одно ограничение цифровых следов и других источников новых данных — нестационарность, то есть постоянно или периодически меняющийся процесс генерации данных и зависимостей внутри них. Это может быть связано как с изменениями внутренних алгоритмов функционирования онлайн-платформ, так и с другими факторами [Salganik, 2019: 70—80]. Например, композиция пользователей по их характеристикам на онлайн-платформах может меняться за счет роста или спада популярности той или иной платформы среди определенных пользователей. Также может меняться не только состав пользователей, но и паттерны поведения пользователей на платформе. Например, популяризация стикеров, эмоджи и GIF-изображений представленных на платформе, способна изменить то, как и какие тексты люди пишут на платформе.

Классический пример нестационарности цифровых следов — предсказание эпидемии гриппа в США на данных поисковых запросов в Google [Butler, 2013; Lazer et al., 2014]. Сначала точность предсказаний была высокая, то есть между поисковыми запросами и эпидемией наблюдалась связь, однако затем эти феномены перестали быть связанными и точность предсказаний серьезно снизилась [Lazer et al., 2014].

Что с этим можно сделать? Как и в случае с внутренними изменениями функционирования платформы, с нестационарностью процессов генерации данных на таких платформах мы мало что можем сделать. Фиксация изменений в характеристиках пользователей онлайн-платформ, из которых собираются данные, а также фиксация изменений в практиках пользования этими платформами могут быть разумными и прозрачными практиками для осмысления и сравнения результатов исследований с помощью цифровых следов.

Заключение

В этой работе, опираясь на актуальные исследования и публикации в передовых социологических и мультидисциплинарных журналах в области вычислительных социальных наук, мы описали возможности использования цифровых следов в контексте недостатков и ограничений массовых опросов. К новым источникам данных и методам их обработки и анализа пока существует некоторое недоверие, вызванное не столько их ограничениями (ибо они свойственны и другим данным и методам), сколько тем, что они новы и непривычны, а у многих социологов нет навыков работы с такими данными [Edelmann et al., 2020: 75]. При этом за последние несколько лет набралось уже достаточно исследований, которые мы частично описали в этой статье, позволяющих понять, для каких исследовательских задач эти данные и методы подходят, а для каких нет, а также как их нужно использовать. Правильное использование для решения релевантных исследовательских задач

DOI: 10.14515/monitoring.2021.1.1760

позволит преодолеть многие ограничения традиционных источников данных и откроет большие возможности для получения нового знания. Мы надеемся, что наш обзор вдохновит российских исследователей на то, чтобы активнее использовать новые источники количественных данных и методы их обработки и анализа.

Список литературы (References)

Богданов М. Б., Лебедев Д. В. Пользование сетью интернет в России в 2003—2015 гг. // Вестник Российского мониторинга экономического положения и здоровья населения НИУ ВШЭ (RLMS-HSE) / отв. ред.: П. М. Козырева. Вып. 7. М.: НИУ ВШЭ, 2017. С. 129—145.
Bogdanov M. B., Lebedev D. V. (2017) The use of the Internet in Russia from 2003 to 2015. *Russian Longitudinal Monitoring Survey — HSE.* No. 7. P. 129—145. (In Russ.)

Девятко И. Ф. Новые данные, новая статистика: от кризиса воспроизводимости к новым требованиям к анализу и представлению данных в социальных науках // Социологические исследования. 2018. № 12. С. 30—38.
Devyatko I. F. (2018) New Data, New Statistics: from Reproducibility Crisis toward New Requirements to Data Analysis and Presentation in Social Sciences. *Sociological Studies.* No. 12. P. 30—38. (In Russ.)

Abrahao B., Parigi P., Gupta A., Cook K. S. (2017) Reputation Offsets Trust Judgments Based on Social Biases Among Airbnb Users. *Proceedings of the National Academy of Sciences*. Vol. 114. No. 37. P. 9848—9853. https://doi.org/10.1073/pnas.1604234114.

Al-Halah Z., Grauman K. (2020) From Paris to Berlin: Discovering Fashion Style Influences Around the World. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. P. 10136—10145. http://arxiv.org/abs/2004.01316.

Azucar D., Marengo D., Settanni M. (2018) Predicting the Big 5 Personality Traits From Digital Footprints on Social Media: A Meta-Analysis. *Personality and Individual Differences*. Vol. 124. P. 150—159. https://doi.org/10.1016/j.paid.2017.12.018.

Baghal T. A., Sloan L., Jessop C., Williams M. L., Burnap P. (2019) Linking Twitter and Survey Data: The Impact of Survey Mode and Demographics on Consent Rates Across Three UK Studies. *Social Science Computer Review*. P. 517—532. https://doi.org/10.1177/0894439319828011.

Bail C. A., Brown T. W., Wimmer A. (2019) Prestige, Proximity, and Prejudice: How Google Search Terms Diffuse across the World. *American Journal of Sociology*. Vol. 124. No. 5. P. 1496—1548. https://doi.org/10.1086/702007.

Bail C. A., Merhout F., Ding P. (2018) Using Internet Search Data to Examine the Relationship Between Anti-Muslim and Pro-ISIS Sentiment in U. S. Counties. *Science Advances*. Vol. 4. No. 6. eaao5948. https://doi.org/10.1126/sciadv.aao5948.

Bailey M., Cao R., Kuchler T., Stroebel J. (2018) The Economic Effects of Social Networks: Evidence From the Housing Market. *Journal of Political Economy*. Vol. 126. No. 6. P. 2224—2276. https://doi.org/10.1086/700073.

DOI: 10.14515/monitoring.2021.1.1760

Bailey M., Cao R., Kuchler T., Stroebel J., Wong A. (2018) Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*. Vol. 32. No. 3. P. 259—280. https://doi.org/10.1257/jep.32.3.259.

Beręsewicz M., Lehtonen R., Reis F., Di Consiglio L., Karlberg M. (2018) An Overview of Methods for Treating Selectivity in Big Data Sources. Publications Office of the European Union. URL: https://ec.europa.eu/eurostat/documents/3888793/9053568/KS-TC-18-004-EN-N.pdf/52940f9e-8e60-4bd6-a1fb-78dc80561943 (accessed: 26.02.2021).

Blumenstock J., Cadamuro G., On R. (2015) Predicting Poverty and Wealth From Mobile Phone Metadata. *Science*. Vol. 350. No. 6264. P. 1073—1076. https://doi.org/10.1126/science.aac4420.

Bond R. M., Fariss C. J., Jones J. J., Kramer A. D. I., Marlow C., Settle J. E., Fowler J. H. (2012) A 61-Million-Person Experiment in Social Influence and Political Mobilization. *Nature*. Vol. 489. No.7415. P. 295—298. https://doi.org/10.1038/nature11421.

Butler D. (2013) When Google Got Flu Wrong. *Nature News*. Vol. 494. No. 7436. P. 155—156. https://doi.org/10.1038/494155a.

Chae D. H., Clouston S., Hatzenbuehler M. L., Kramer M. R., Cooper H. L. F., Wilson S. M., Stephens-Davidowitz S. I., Gold R. S., Link B. G. (2015) Association Between an Internet-Based Measure of Area Racism and Black Mortality. *PLOS ONE*. Vol. 10. No. 4. e0122963. https://doi.org/10.1371/journal.pone.0122963.

Chae D. H., Clouston S., Martz C. D., Hatzenbuehler M. L., Cooper H. L. F., Turpin R., Stephens-Davidowitz S., Kramer M. R. (2018) Area Racism and Birth Outcomes Among Blacks in the United States. *Social Science & Medicine*. Vol. 199. P. 49—55. https://doi.org/10.1016/j.socscimed.2017.04.019.

Chetty R., Hendren N., Kline P., Saez E. (2014) Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics*. Vol. 129. No. 4. P. 1553—1623. https://doi.org/10.1093/qje/qju022.

Cranmer S. J., Desmarais B. A. (2017) What Can We Learn From Predictive Modeling? *Political Analysis*. Vol. 25. No. 2. P. 145—166. https://doi.org/10.1017/pan.2017.3.

Diaz F., Gamon M., Hofman J. M., Kıcıman E., Rothschild D. (2016) Online and Social Media Data As an Imperfect Continuous Panel Survey. *PLOS ONE*. Vol. 11. No. 1. e0145406. https://doi.org/10.1371/journal.pone.0145406.

Edelmann A., Wolff T., Montagne D., Bail C. A. (2020) Computational Social Science and Sociology. *Annual Review of Sociology*. Vol. 46. P. 61—81. https://doi.org/10.1146/annurev-soc-121919-054621.

Enghoff O., Aldridge J. (2019) The Value of Unsolicited Online Data in Drug Policy Research. *International Journal of Drug Policy*. Vol. 73. P. 210—218. https://doi.org/10.1016/j.drugpo.2019.01.023.

DOI: 10.14515/monitoring.2021.1.1760

Garcia D., Mitike Kassa Y., Cuevas A., Cebrian M., Moro E., Rahwan I., Cuevas R. (2018) Analyzing Gender Inequality Through Large-Scale Facebook Advertising Data. *Proceedings of the National Academy of Sciences*. Vol. 115. No. 27. P. 6958—6963. https://doi.org/10.1073/pnas.1717781115.

Garcia D., Rimé B. (2019) Collective Emotions and Social Resilience in the Digital Traces After a Terrorist Attack. *Psychological Science*. Vol. 30. No. 4. P. 617—628. https://doi.org/10.1177/0956797619831964.

Garg N., Schiebinger L., Jurafsky D., Zou J. (2018) Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences*. Vol. 115. No. 16. P. E 3635—E 3644. https://doi.org/10.1073/pnas.1720347115.

Garip F. (2020) What Failure to Predict Life Outcomes Can Teach Us. *Proceedings of the National Academy of Sciences*. Vol. 117. No. 15. P. 8234—8235. https://doi.org/10.1073/pnas.2003390117.

Gee L. K., Jones J., Burke M. (2017) Social Networks and Labor Markets: How Strong Ties Relate to Job Finding on Facebook's Social Network. *Journal of Labor Economics*. Vol. 35. No. 2. P. 485—518. https://doi.org/10.1086/686225.

Gee L. K., Jones J. J., Fariss C. J., Burke M., Fowler J. H. (2017) The paradox of weak ties in 55 countries. *Journal of Economic Behavior & Organization*. Vol. 133. P. 362—372. https://doi.org/10.1016/j.jebo.2016.12.004.

Gilbert E., Karahalios K. (2009) Predicting Tie Strength With Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. P. 211—220. https://doi.org/10.1145/1518701.1518736.

Golder S. A., Macy M. W. (2011) Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science*. Vol. 333. No. 6051. P. 1878—1881. https://doi.org/10.1126/science.1202775.

Golder Scott A., Macy M. W. (2014) Digital Footprints: Opportunities and Challenges for Online Social Research. *Annual Review of Sociology*. Vol. 40. No. 1. P. 129—152. https://doi.org/10.1146/annurev-soc-071913-043145.

Hargittai E. (2020) Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*. Vol. 38. No. 1. P. 10—24. https://doi.org/10.1177/0894439318788322.

Herdağdelen A., State B., Adamic L., Mason W. (2016) The social ties of immigrant communities in the United States. *Proceedings of the 8th ACM Conference on Web Science — WebSci '16*. P. 78—84. https://doi.org/10.1145/2908131.2908163.

Hills T. T., Proto E., Sgroi D., Seresinhe C. I. (2019) Historical Analysis of National Subjective Wellbeing Using Millions of Digitized Books. *Nature Human Behaviour*. Vol. 3. No. 12. P. 1271—1275. https://doi.org/10.1038/s41562-019-0750-z.

Hobbs W. R., Burke M., Christakis N. A., Fowler J. H. (2016) Online Social Integration Is Associated With Reduced Mortality Risk. *Proceedings of the National Academy*

DOI: 10.14515/monitoring.2021.1.1760

of Sciences. Vol. 113. No. 46. P. 12980—12984. https://doi.org/10.1073/pnas.1605554113.

Hofman J. M., Sharma A., Watts D. J. (2017) Prediction and Explanation in Social Systems. *Science*. Vol. 355. No. 6324. P. 486—488. https://doi.org/10.1126/science.aal3856.

Huang C. (2019) Social Network Site Use and Big Five Personality Traits: A Meta-Analysis. *Computers in Human Behavior*. Vol. 97. P. 280—290. https://doi.org/10.1016/j.chb.2019.03.009.

Iannelli L., Giglietto F., Rossi L., Zurovac E. (2018) Facebook Digital Traces for Survey Research: Assessing the Efficiency and Effectiveness of a Facebook Ad — Based Procedure for Recruiting Online Survey Respondents in Niche and Difficult-to-Reach Populations. *Social Science Computer Review*. P. 462—476. https://doi.org/10.1177/0894439318816638.

Jaidka K., Giorgi S., Schwartz H. A., Kern M. L., Ungar L. H., Eichstaedt J. C. (2020) Estimating Geographic Subjective Well-Being From Twitter: A Comparison of Dictionary and Data-Driven Language Methods. *Proceedings of the National Academy of Sciences*. P. 10165—10171. https://doi.org/10.1073/pnas.1906364117.

Jungherr A., Schoen H., Posegga O., Jürgens P. (2017) Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support. *Social Science Computer Review*. Vol. 35. No. 3. P. 336—356. https://doi.org/10.1177/0894439316631043.

Kashyap R., Fatehkia M., Al Tamime R., Weber I. (2020) Monitoring Global Digital Gender Inequality Using the Online Populations of Facebook and Google. *Demographic Research*. Vol. 43. No. 27. P. 779—816. https://doi.org/10.4054/DemRes.2020.43.27.

Kiciman E., Counts S., Gasser M. (2018) Using Longitudinal Social Media Analysis to Understand the Effects of Early College Alcohol Use. *Proceedings of 12th International Conference on Web and Social Media (ICWSM-18)*. Vol. 12. No. 1. P. 171—180.

Kizilcec R. F., Bakshy E., Eckles D., Burke M. (2018) Social Influence and Reciprocity in Online Gift Giving. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems — CHI '18*. P. 1—11. https://doi.org/10.1145/3173574.3173700.

Kolenikov S. (2016) Post-Stratification or Non-Response Adjustment? *Survey Practice*. Vol. 9. No. 3. P. 1—12. https://doi.org/10.29115/SP-2016-0014.

Kosinski M., Stillwell D., Graepel T. (2013) Private Traits and Attributes Are Predictable From Digital Records of Human Behavior. *Proceedings of the National Academy of Sciences*. Vol. 110. No. 15. P. 5802—5805. https://doi.org/10.1073/pnas.1218772110.

Kossinets G. (2006) Effects of Missing Data in Social Networks. *Social Networks*. Vol. 28. No. 3. P. 247—268. https://doi.org/10.1016/j.socnet.2005.07.002.

DOI: 10.14515/monitoring.2021.1.1760

Kozlowski A. C., Taddy M., Evans J. A. (2019) The Geometry of Culture: Analyzing the Meanings of Class through Word Embeddings. *American Sociological Review*. Vol. 84. No. 5. P. 905—949. https://doi.org/10.1177/0003122419877135.

Lavrakas P. (2008) *Encyclopedia of Survey Research Methods*. Sage Publications, Inc. https://doi.org/10.4135/9781412963947.

Lazer D. (2015) Issues of Construct Validity and Reliability in Massive, Passive Data Collections. *The City Papers: An Essay Collection from the Decent City Initiative*. URL: http://citiespapers.ssrc.org/issues-of-construct-validity-and-reliability-in-massive-passive-data-collections/ (accessed 26.02.2021).

Lazer D., Kennedy R., King G., Vespignani A. (2014) The Parable of Google Flu: Traps in Big Data Analysis. *Science*. Vol. 343. No. 6176. P. 1203—1205.

Lazer D., Pentland A., Adamic L., Aral S., Barabási A.-L., Brewer D., Christakis N., Contractor N., Fowler J., Gutmann M. (2009) Computational Social Science. *Science*. Vol. 323. No. 5915. P. 721—723.

Lazer D., Radford J. (2017) Data ex Machina: Introduction to Big Data. *Annual Review of Sociology*. Vol. 43. No. 1. P. 19—39. https://doi.org/10.1146/annurev-soc-060116-053457.

Lewis K. (2013) The Limits of Racial Prejudice. *Proceedings of the National Academy of Sciences*. Vol. 110. No. 47. P. 18814—18819. https://doi.org/10.1073/pnas.1308501110.

Lewis K. (2015) Three Fallacies of Digital Footprints. *Big Data & Society*. Vol. 2. No. 2. P. 1—4. https://doi.org/10.1177/2053951715602496.

Ma X., Cheng J., Iyer S., Naaman M. (2019) When Do People Trust Their Social Groups? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems — CHI '19*. P. 1—12. https://doi.org/10.1145/3290605.3300297.

Martin T., Hofman J. M., Sharma A., Anderson A., Watts D. J. (2016) Exploring Limits to Prediction in Complex Social Systems. *Proceedings of the 25th International Conference on World Wide Web — WWW '16*. P. 683—694. https://doi.org/10.1145/2872427.2883001.

Mauss M. (2000) *The Gift: The Form and Reason for Exchange in Archaic Societies*. New York. NY: W. W. Norton.

Molina M., Garip F. (2019) Machine Learning for Sociology. *Annual Review of Sociology*. Vol. 45. P. 27—45.

Olteanu A., Castillo C., Diaz F., Kıcıman E. (2019) Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*. Vol. 2. P. 1—33. https://doi.org/10.3389/fdata.2019.00013.

Orben A., Dienlin T., Przybylski A. K. (2019) Social Media's Enduring Effect on Adolescent Life Satisfaction. *Proceedings of the National Academy of Sciences*. Vol. 116. No. 21. P. 10226—10228. https://doi.org/10.1073/pnas.1902058116.

DOI: 10.14515/monitoring.2021.1.1760

Pierson E., Simoiu C., Overgoor J., Corbett-Davies S., Jenson D., Shoemaker A., Ramachandran V., Barghouty P., Phillips C., Shroff R., Goel S. (2020) A Large-Scale Analysis of Racial Disparities in Police Stops Across the United States. *Nature Human Behaviour*. Vol. 4. No. 7. P. 1—10. https://doi.org/10.1038/s41562-020-0858-1.

Rafaeli A., Ashtar S., Altman D. (2019) Digital Traces: New Data, Resources, and Tools for Psychological-Science Research. *Current Directions in Psychological Science*. Vol. 28. No. 6. P. 560—566. https://doi.org/10.1177/0963721419861410.

Reis B. Y., Brownstein J. S. (2010) Measuring the Impact of Health Policies Using Internet Search Patterns: The Case of Abortion. *BMC Public Health*. Vol. 10. No. 1. P. 1—5. https://doi.org/10.1186/1471-2458-10-514.

Ribeiro F. N., Benevenuto F., Zagheni E. (2020) How Biased is the Population of Facebook Users? Comparing the Demographics of Facebook Users with Census Data to Generate Correction Factors. *12th ACM Conference on Web Science.* P. 325—334. http://arxiv.org/abs/2005.08065.

Ruths D., Pfeffer J. (2014) Social Media for Large Studies of Behavior. *Science*. Vol. 346. No. 6213. P. 1063—1064. https://doi.org/10.1126/science.346.6213.1063.

Sala E., Gaia A., Cerati G. (2020) The Gray Digital Divide in Social Networking Site Use in Europe: Results From a Quantitative Study. *Social Science Computer Review*. P. 1—18. https://doi.org/10.1177/0894439320909507.

Salganik M. (2019) *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press. URL: https://www.bitbybitbook.com/en/1st-ed/preface/ (accessed: 26.02.2021).

Salganik M. J., Lundberg I., Kindel A. T., Ahearn C. E., Al-Ghoneim K., Almaatouq A., Altschul D. M., Brand J. E., Carnegie N. B., Compton R. J., Datta D., Davidson T., Filippova A., Gilroy C., Goode B. J., Jahani E., Kashyap R., Kirchner A., McKay S., … McLanahan S. (2020) Measuring the Predictability of Life Outcomes With a Scientific Mass Collaboration. *Proceedings of the National Academy of Sciences*. Vol. 117. No. 15. P. 8398—8403. https://doi.org/10.1073/pnas.1915006117.

Settanni M., Azucar D., Marengo D. (2018) Predicting Individual Characteristics from Digital Traces on Social Media: A Meta-Analysis. *Cyberpsychology, Behavior, and Social Networking*. Vol. 21. No. 4. P. 217—228. https://doi.org/10.1089/cyber.2017.0384.

Shmueli G. (2010) To Explain or to Predict? *Statistical Science*. Vol. 25. No. 3. P. 289—310. https://doi.org/10.1214/10-STS330.

Sivak E., Smirnov I. (2019) Parents Mention Sons More Often Than Daughters on Social Media. *Proceedings of the National Academy of Sciences*. Vol. 116. No. 6. P. 2039—2041. https://doi.org/10.1073/pnas.1804996116.

Smirnov I. (2019) Schools Are Segregated by Educational Outcomes in the Digital Space. *PLOS ONE*. Vol. 14. No. 5. e0217142. https://doi.org/10.1371/journal.pone.0217142.

DOI: 10.14515/monitoring.2021.1.1760

Smirnov I. (2020) Estimating Educational Outcomes From Students' Short Texts on Social Media. *EPJ Data Science*. Vol. 9. No. 1. P. 1—11. https://doi.org/10.1140/epjds/s13688-020-00245-8.

Smirnov I. (2018) Predicting PISA Scores from Students' Digital Traces. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*. Vol. *3*. P. 360—365.

Smith T. W. (1984) Recalling Attitudes: An Analysis of Retrospective Questions on the 1982 GSS. *Public Opinion Quarterly*. Vol. 48. No. 3. P. 639—649. https://doi.org/10.1086/268865.

Stephens-Davidowitz S. (2014) The Cost of Racial Animus on a Black Candidate: Evidence Using Google Search Data. *Journal of Public Economics*. Vol. 118. P. 26—40. https://doi.org/10.1016/j.jpubeco.2014.04.010.

Stewart I., Flores R. D., Riffe T., Weber I., Zagheni E. (2019) Rock, Rap, or Reggaeton?: Assessing Mexican Immigrants' Cultural Assimilation Using Facebook Data. *The World Wide Web Conference*. P. 3258—3264. https://doi.org/10.1145/3308558.3313409.

Stier S., Breuer J., Siegers P., Thorson K. (2019) Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field. *Social Science Computer Review*. Vol. 38. No. 5. P. 503—516. https://doi.org/10.1177/0894439319843669.

Stopczynski A., Sekara V., Sapiezynski P., Cuttone A., Madsen M. M., Larsen J. E., Lehmann S. (2014) Measuring Large-Scale Social Networks with High Resolution. *PLOS ONE*. Vol. 9. No. 4. e95978. https://doi.org/10.1371/journal.pone.0095978.

Wang Z., Hale S., Adelani D. I., Grabowicz P., Hartman T., Flöck F., Jurgens D. (2019) Demographic Inference and Representative Population Estimates from Multilingual Social Media Data. *The World Wide Web Conference*. P. 2056—2067. https://doi.org/10.1145/3308558.3313684.

Warren J. R., Halpern-Manners A. (2012) Panel Conditioning in Longitudinal Social Science Surveys: *Sociological Methods & Research*. Vol. 41. No. 4. P. 491—534. https://doi.org/10.1177/0049124112460374.

Wojcik S., Hughes A. (2019) How Twitter Users Compare to the General Public. *Pew Research Center: Internet, Science & Tech*. URL: https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/ (accessed: 26.02.2021).

Yarkoni T., Westfall J. (2017) Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*. Vol. 12. No. 6. P. 1100—1122. https://doi.org/10.1177/1745691617693393.