БОЛЬШИЕ ДАННЫЕ И ОФИЦИАЛЬНАЯ СТАТИСТИКА: ОБЗОР МЕЖДУНАРОДНОЙ ПРАКТИКИ ВНЕДРЕНИЯ НОВЫХ ИСТОЧНИКОВ ДАННЫХ

Д.А. Плеханов

Революционный скачок в использовании компьютеров и прочих цифровых устройств и колоссальное увеличение информационных потоков привели к появлению новых источников информации об окружающей действительности. Эти источники, объединенные общим названием «большие данные», предоставляют исследователям уникальные возможности для проведения количественного анализа самых различных социально-экономических явлений.

В публикуемой статье представлен обзор исследовательских и пилотных проектов, которые были реализованы в последние годы международными организациями и национальными статистическими службами с целью изучения возможностей использования больших данных в официальной статистике. Анализ результатов проектов показывает, что внедрение больших данных в систему официальной статистики сталкивается с рядом серьезных ограничений, связанных с нерешенными вопросами в области методологии, обеспечения доступа к данным и сохранения конфиденциальности информации, технических требований к инфраструктуре.

По мнению автора, из-за существующих ограничений статистические показатели, которые рассчитывались бы исключительно на основе сбора и обработки больших данных, пока не получили распространения в официальной статистике. В настоящее время национальные статистические службы продолжают исследования возможностей совместного использования больших данных и традиционных источников информации.

Ключевые слова: официальная статистика, большие данные, конфиденциальность информации, статистическая инфраструктура.

JEL: C10, C80.

В современном мире все больше и больше информации генерируется и обрабатывается окружающими нас цифровыми устройствами. Активное использование компьютеров и различных цифровых устройств не только в деловой сфере, но и в повседневной жизни привели к экспоненциальному росту объема регистрируемых сведений, которые благодаря своим характеристикам стали называться большими данными. Появление новых источников информации в свою очередь существенно расширило возможности анализа тех процессов, которые происходят в самых различных сферах общества. Например, на потребительском рынке на протяжении последних лет отмечается тенденция более активного использования услуг интернет-магазинов и совершения покупок при помощи электронных платежей. Поскольку все эти процессы сопровождаются обменом большого объема информации, которая накапливается поставщиками предоставляемых услуг, у последних появляется уникальная возможность для мониторинга данных о потребительском поведении. Похожие процессы происходят и в других сферах человеческой деятельности (здравоохранении, транспорте и логистике, финансовой сфере и т. д.). В результате многие исследования, которые раньше могли проводиться только путем формирования определенной выборочной совокупности (всего населения, отдельных групп потребителей, предприятий и т. д.), в настоящее время могут осуществляться на основе данных непосредственного мониторинга. Формирование новых источников информации оказывает существенное влияние на все сферы, связанные с анализом данных, в том числе и на систему официальной статистики. С одной стороны, по мере распространения больших данных национальные статистические службы утрачивают свое монопольное положение как единственного поставщика информации о ситуации в обществе, а с другой стороны, они приобретают возможность доступа к данным, которые невозможно было получить раньше при помощи традиционных методов сбора информации.

Огромный объем информации - это не единственная характеристика больших данных. Кроме объема (*Volume*), исследователи в области больших данных, как правило, выделяют еще две дополнительные характеристики: скорость (*Velocity*)

Плеханов Дмитрий Александрович (plehanov@icss.ac.ru) - ведущий специалист, Институт комплексных стратегических исследований (ИКСИ) (г. Москва, Россия).

и многообразие (*Variety*) [1]. Скорость означает, что большие данные постоянно накапливаются, происходит постоянный процесс поступления новой информации о наблюдаемом объекте (новых транзакциях, действиях пользователя, изменениях в организме пациента, ситуации на дороге и т. д.). Под многообразием понимается разнообразие типов и источников поступающих данных - это различные количественные показатели, а также аудио- и видеоматериалы, текстовые сообщения и другая информация, которая может собираться об объекте наблюдения [2]. В некоторых случаях к характеристикам больших данных также причисляют изменчивость (*Variability*) и достоверность (*Veracity*). В то же время многие эксперты отмечают, что термин «большие данные» не имеет точного определения; содержание этого понятия может варьироваться в зависимости от конкретной области анализа данных [3].

Для целей публикуемой статьи мы будем рассматривать большие данные в узком понимании - как совокупность различных источников данных. Эксперты Европейской экономической комиссии ООН рассматривают большие данные в качестве дополнительной информации к той, которую предоставляет официальная статистика (описывающая определенную ситуацию и создающая картину страны, ее экономики, населения и т. д.), и предложили использовать следующее определение: это «данные большого объема, скорости и разнообразия, требующие затратоэффективных и инновационных видов обработки для совершенствования анализа и процесса принятия решений»[1]. В зависимости от источников большие данные могут быть условно разделены на несколько основных категорий: 1) данные, генерируемые человеком (например, сообщения в социальных сетях, поисковые запросы в сети Интернет, текстовые сообщения и т. д.); 2) данные, генерируемые в результате выполнения различных процессов (бизнес-процессы и работа веб-сайтов); 3) данные, генерируемые различными машинами (сенсорными устройствами, датчиками и компьютерными системами).

В публикуемой статье представлен обзор исследований и пилотных проектов, которые ведет международное статистическое сообщество с целью изучения возможностей использования новых источников данных в официальной статистике. Структура работы выглядит следующим образом. В первой части новые источники данных рассматриваются в качестве потенциальных конкурентов системы официальной статистики. Во второй части представлен краткий обзор деятельности международных и национальных статистических организаций в области исследования потенциала больших данных. Основное внимание уделяется достигнутым результатам и барьерам, которые стоят на пути внедрения новых источников данных. В третьей части сформулированы основные рекомендации по преодолению трудностей, связанных с внедрением больших данных в систему официальной статистики.

1. Большие данные как альтернативный источник информации о социально-экономических явлениях

Согласно декларации Статистической комиссии ООН, «официальная статистика является необходимым элементом информационной системы демократического общества, обеспечивая правительство, экономические круги и общественность данными об экономическом, демографическом, социальном и экологическом положении. С этой целью официальные статистические данные, имеющие практическую ценность, подготавливаются и распространяются на объективной основе государственными статистическими ведомствами для обеспечения уважения права граждан на общественную информацию»[2]. Таким образом, одной из основных целей деятельности статистических органов является предоставление обществу важной информации. Однако с появлением больших данных можно отметить, что все чаще информация, имеющая ценность для общества в целом, собирается и обрабатывается вне рамок системы официальной статистики (см. таблицу 1). Очевидно, такая ситуация - это определенный вызов для органов официальной статистики во всем мире, но, прежде всего, в развитых странах, в которых уровень развития современных

[1] Европейская экономическая комиссия ООН. Каково значение «больших данных» для официальной статистики? Записка секретариата. 61-я пленарная сессия Конференции европейских статистиков. Женева, 10-12 июня 2013 г. URL: https://statswiki.unece.org/download/attachments/77170614/Big%20Data%20Published%20version%20.RU.pdf?version=1&modificationDate=1370507714381&api=v2.

[2] Статистическая комиссия ООН. Основные принципы официальной статистики. URL: unstats.un.org/unsd/methods/statorg/fp-russian.pdf.

информационных технологий выше. Основная угроза для официальной статистики состоит в том, что правительства разных стран в будущем могут принимать решения о снижении расходов на содержание национальных статистических служб и направлять финансовые ресурсы на покупку необходимой информации у сторонних организаций.

Таблица 1

Потенциальные направления использования источников больших данных для расчета статистических показателей

Тип данных	Источники данных	Разделы официальной статистики
Данные, генерируемые человеком	Поисковые запросы в Интернете	Рынок труда Опережающие индикаторы экономической деятельности
	Сообщения в социальных сетях	Опережающие индикаторы экономической деятельности
	Онлайн-объявления о вакансиях и поиске работы	Рынок труда
Процессные данные	Интернет-магазины	Цены
	Веб-сайты недвижимости	Цены
	Веб-сайты компаний	Информационные и коммуникационные технологии
	Платежи по банковским картам	Опережающие индикаторы экономической деятельности
	Сканированные данные о продажах розничных сетей	Цены
Машинные данные	Дорожные датчики	Транспорт
	Счетчики электроэнергии	Потребление электроэнергии
	Данные сотовых операторов	Культура, отдых и туризм Демография
	Спутниковые снимки	Городское хозяйство Сельское хозяйство Окружающая среда
	Авиационные радары	Транспорт
	Система идентификации судов	Транспорт

Источник: составлено автором на основе Плана работ и дорожной карты Европейской статистической системы (ECC). URL: https://ec.europa.eu/eurostat/cros/content/ess-big-data-action-plan-and-roadmap-10_en.

В качестве примера того, как большие данные вторгаются в традиционные сферы официальной статистики, можно привести проект под названием PriceStats. В рамках этого проекта команда исследователей осуществляет в Интернете мониторинг цен на продукты более чем в 70 странах мира (в том числе в США, Японии, Австралии, Китае, России)[3]. На основе собранных данных рассчитываются показатели инфляции, которые затем предоставляются по подписке заинтересованным клиентам; кроме того, специалисты проекта занимаются проведением специализированных исследований по запросам. Индексы инфляции публикуются ежедневно (с временным лагом в три дня), в то время как большинство официальных показателей инфляции публикуется только раз в месяц. Методика расчета индексов основана на работах экономиста Альберто Ковалло, который в 2010 г. защитил в Гарвардском университете диссертацию, посвященную использованию данных о ценах в Интернете для расчета показателей инфляции (на примере Аргентины, Чили, Бразилии и Колумбии) [4].

Проект PriceStats оказался настолько успешным, что когда у представителей международного сообщества возникли сомнения в качестве официальной статистики инфляции в Аргентине, многие начали использовать показатели PriceStats. В частности, в феврале 2012 г. журнал Economist объявил о том, что прекращает публикацию официальных данных об инфляции в Аргентине, поскольку большинство альтернативных оценок инфляции в стране указывали на то, что официальные оценки занижены как минимум в два раза[4]. С тех пор вместо официальных данных журнал еженедельно публикует данные PriceStats. Пример Аргентины показывает, что органы официальной статистики больше не являются монополистами на рынке информации, и в случае возникновения сомнений в качестве публикуемых данных пользователи могут обратиться к альтернативным источникам информации.

Другим примером того, как современные технологии оказывают влияние на процесс сбора и распространения данных, является деятельность компании Genscape. Эта компания является глобальным поставщиком информации для участников сырьевого и финансового рынков. При этом в ходе своей работы компания активно прибегает к нетрадиционным способам сбора информации. В частности, Genscape проводит мониторинг работы нефтехранилищ при помощи малой авиации. Вертолет Genscape регулярно осуществляет фотосъем-

[3] Сайт проекта PriceStats - URL: https://www.pricestats.com/.

[4] Which of these is not like the others? // The Economist. February 24th. 2012. URL: https://www.economist.com/blogs/americasview/2012/02/measuring-inflation.

ку над одним из главных нефтехранилищ США в городе Кушинг (штат Оклахома) для определения того, сколько нефти находится в каждом из резервуаров[5]. Официальные данные об уровне запасов публикуются Энергетическим информационным агентством США (ЭИА), но клиенты Genscape имеют возможность получать эту информацию на несколько дней раньше. Данные Genscape оказались особенно востребованными осенью 2013 г. во время приостановки работы Федерального правительства США и публикации отчетов ЭИА. В этот момент компания фактически оказалась единственным поставщиком данных о запасах нефти. В связи с большим общественным интересом к этой информации Genscape даже опубликовала бесплатную версию своего пресс-релиза в тот самый день, когда должны были выйти официальные данные ЭИА[6].

2. Исследования возможностей использования больших данных в официальной статистике

Потенциал использования новых источников данных официально признан международным статистическим сообществом. Международные статистические организации обратили свое внимание на феномен больших данных на волне роста интереса к этому явлению в начале 2010-х годов (см. рис. 1). В 2013 г. руководители национальных статистических служб Европейского союза подписали так называемый Схевенингенский меморандум об изучении возможностей интеграции больших данных в систему официальной статистики[7]. Основные направления внедрения новых источников данных были конкретизированы в плане работ и дорожной карте, принятых Европейской статистической системой (ЕСС) в 2014 г.[8]. В этом же году Статистическая комиссия ООН создала рабочую группу для выработки стратегического видения, направления и координации глобальной программы изучения потенциала больших данных и возможностей его использования в официальной статистике[9]. С 2014 г. рабочая группа занимается, в частности, вопросами обеспечения доступа и налаживания партнерских отношений с представителями частного сектора; использования данных, передаваемых по сети мобильной связи; спутниковых изображений; сообщений в социальных сетях. Кроме того, Статистическая комиссия ООН при участии представителей статистического сообщества и экспертов из бизнеса проводит ежегодные международные конференции, посвященные вопросам использования больших данных в официальной статистике. В 2014 г. в рамках Европейской экономической комиссии ООН была создана так называемая «песочница больших данных» (Big Data Sandbox) - специальная платформа, предназначенная для взаимодействия национальных статистических служб и обмена опытом и информацией по вопросам использования новых источников данных [5].

Рис. 1. **Рост популярности запросов со словами «большие данные» и частоты упоминания термина на сайтах национальных статистических служб**

Примечание: индекс отражает долю запросов, содержащих указанные слова, в общем объеме поисковых запросов, обработанных **Google за месяц; максимальное значение** показателя за выбранный период времени приравнивается к 100 баллам.

Источник: расчеты автора на основе данных сервисов Google.

[5] Rothfeld M., Patterson S. Traders seek an edge with high-tech snooping // Wall Street Journal. December 18, 2013. URL: https://www.wsj.com/articles/traders-seek-an-edge-with-hightech-snooping-1387426263.

[6] Текст пресс-релиза на сайте компании Genscape от 16 октября 2013 г. URL: http://www.genscape.com/press-releases/genscape-publicly-release-its-highly-accurate-cushing-oil-stock-data-during.

[7] Scheveningen memorandum. Big data and official statistics. URL: http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13.

[8] The European Statistical System (ESS) big data action plan and roadmap 1.0. URL: https://ec.europa.eu/eurostat/cros/content/ess-big-data-action-plan-and-roadmap-10_en.

[9] Раздел на сайте Статистической комиссии ООН, посвященный большим данным: URL: https://unstats.un.org/bigdata/.

Какие преимущества могут предоставлять большие данные для официальной статистики? Во-первых, новые источники информации могут быть использованы для расширения возможностей статистических организаций при сборе данных. Большие данные позволяют взглянуть по-новому на многие традиционные сферы деятельности (например, мониторинг экономической активности населения на основе данных о транзакциях по банковским картам или анализ туристических потоков на основе данных сотовых операторов) или осуществлять мониторинг в совершенно новых областях, не охваченных ранее официальной статистикой (например, в сфере потребительского поведения населения в сети Интернет). Во-вторых, скорость публикации статистических показателей также имеет значение. Проведение статистических переписей, сплошных и выборочных наблюдений требует времени, в результате чего официальные показатели публикуются с временным лагом (от нескольких недель до одного года и более). В то же время источники больших данных могут быть доступны практически в режиме реального времени. В-третьих, большие данные могут быть использованы для повышения качества и точности существующих статистических показателей. В-четвертых, использование (даже частичное) источников больших данных вместо организации статистических наблюдений ведет к снижению нагрузки на респондентов, что является актуальной задачей для национальных статистических служб в настоящее время. Согласно результатам опроса, проведенного Статистической комиссией ООН в 2015 г., увеличение скорости публикации официальных статистических показателей и снижение нагрузки на респондентов являются основными преимуществами использования больших данных в официальной статистике с точки зрения представителей национальных статистических служб[10]. В-пятых, предварительные оценки показывают, что использование больших данных может быть выгодно национальным статистическим службам и с финансовой точки зрения (за счет снижения расходов на трудоемкие операции по сбору данных) [6].

Национальные статистические организации (в основном из стран - членов ОЭСР) начали изучение возможностей использования больших данных с проведения различных исследовательских и пилотных проектов в этой сфере. Примерами подобного рода проектов являются анализ данных дорожных датчиков и сообщений в социальных сетях (Центральное статистическое бюро Нидерландов), спутниковых снимков (статистическая служба Канады), статистики цен в онлайн-магазинах и сервисах (Национальная статистическая служба Великобритании). Результаты пилотных проектов, проведенных различными статистическими организациями, показывают, что большие данные представляют несомненный интерес для официальной статистики. В то же время примеры реальной интеграции новых источников данных в повседневную работу статистических служб по подготовке официальной статистической информации пока еще остаются довольно редкими. Наиболее популярным источником данных для национальных статистических служб остаются сканированные данные розничных сетей о продажах, которые рассматривались в качестве потенциального источника информации для официальной статистики еще в 1990-х годах (см. таблицу 2).

Таблица 2

Примеры реализации проектов национальных статистических служб по использованию больших данных в официальной статистике

Страна	Источник данных	Раздел официальной статистики	Год	Стадия проекта	Основные результаты и возможности использования
Норвегия	сканированные данные о продажах	цены	1997	интеграция	данные используются с 2005 г. при расчете инфляции; в структуре потребительской корзины доля товаров и услуг, цены по которым определяются на основе сканированных данных, составляет 22%

[10] UNSD. Report of the Big Data Survey 2015. URL: https://unstats.un.org/unsd/statcom/47th-session/documents/BG-2016-6-Report-of-the-2015-Big-Data-Survey-E.pdf.

Окончание таблицы 2

Страна	Источник данных	Раздел официальной статистики	Год	Стадия проекта	Основные результаты и возможности использования
Нидерланды	сканированные данные о продажах	цены	2002	интеграция	данные по отдельным товарным группам используются при расчете инфляции
	дорожные датчики	транспорт	2011	расчет	публикация с 2015 г. показателя загруженности дорог
	сообщения в социальных сетях	ожидания потребителей	2013	изучение потенциала	индексы, измеряющие настроение пользователей социальных сетей по их сообщениям, достаточно тесно коррелируют с индексом потребительской уверенности
	данные сотовых операторов	туризм	2013	изучение потенциала	данные могут быть использованы для оценок турпотоков за относительно короткие интервалы времени
Новая Зеландия	сканированные данные о продажах	цены	2006	интеграция	данные о ценах на потребительскую электронику используются с 2014 г. при расчете инфляции
Швеция	сканированные данные о продажах	цены	2008	интеграция	данные используются с 2012 г. при расчете инфляции
Швейцария	сканированные данные о продажах	цены	2008	интеграция	данные о ценах на продукты питания используются с 2008 г. при расчете инфляции
Португалия	сканированные данные о продажах	цены	2011	изучение потенциала	н. д.
Великобритания	сканированные данные о продажах	цены	2012	изучение потенциала	результаты расчета индексов цен на основе сканированных данных существенным образом зависят от выбора методологии построения индексов
	сайты интернет-магазинов	цены	2014	изучение потенциала	в 2016 г. проведены пробные расчеты индексов цен на продукты питания
Бельгия	сканированные данные о продажах	цены	2012	интеграция	данные используются с 2016 г. при расчете инфляции
	данные сотовых операторов	демография	2015	изучение потенциала	отмечена тесная корреляция с данными административного учета
Эстония	данные сотовых операторов	туризм / демография	2012	изучение потенциала	данные могут быть использованы в качестве дополнения к официальной статистике
Канада	спутниковые снимки	сельское хозяйство	2012	интеграция	с 2015 г. спутниковые данные используются при составлении прогноза урожая
Италия	данные сотовых операторов	демография	2013	изучение потенциала	отмечена тесная корреляция с результатами переписи и данными административного учета
	сайты интернет-магазинов	цены	2013	интеграция	при расчете индекса инфляции используются данные о ценах в Интернете на потребительскую электронику, услуги газоснабжения, финансовые услуги, транспорт
	сайты компаний	информационные технологии	2013	изучение потенциала	рассматривается возможность сбора данных о компаниях на основе информации, размещенной на их сайтах
	сканированные данные о продажах	цены	2014	изучение потенциала	рассматривается возможность использования данных при расчете инфляции
Австралия	сканированные данные о продажах	цены	2014	интеграция	с 2014 г. данные используются при расчете инфляции; в структуре потребительской корзины доля товаров и услуг, цены по которым определяются на основе сканированных данных, составляет 25%
Колумбия	спутниковые снимки	земельные ресурсы	2015	изучение потенциала	данные позволяют повысить качество статистики об использовании земельных ресурсов

Источник: составлено автором на основе обзора публикаций и выступлений представителей национальных статистических служб.

Лидером среди статистических служб в области анализа больших данных можно считать Центральное статистическое бюро Нидерландов (ЦСБ). В 2015 г. бюро опубликовало показатели транспортной статистики, которые были рассчитаны целиком на основе информации, полученной с 20 тыс. датчиков, расположенных на автомагистралях страны[11]. Оперативный характер данных позволяет национальной статистической службе публиковать актуальные показатели практически с минимальной задержкой по времени. В начале 2016 г., когда на севере страны на дорогах образовался сильный гололед, бюро провело сравнение интенсивности дорожного движения со средними показателями за 2012-2015 гг. и уже 8 января разместило пресс-релиз с результатами расчетов на официальном сайте[12]. В 2016 г. в рамках ЦСБ был создан Центр статистики больших данных с собственными мощностями и постоянным штатом[13]. Это было сделано для того, чтобы объединить все усилия по работе с большими данными в рамках одного структурного подразделения. Кроме того, ЦСБ принимает активное участие в международных проектах, связанных с большими данными. Так, бюро является участником рабочей группы Статистической комиссии ООН и группы высокого уровня при Европейской экономической комиссии ООН.

3. Ограничения использования больших данных в официальной статистике

Обзор пилотных и исследовательских проектов, проведенных за последние годы национальными статистическими службами разных стран, позволяет составить представление о том, какие основные препятствия стоят на пути более активного внедрения источников больших данных в официальную статистику. Проблемы, с которыми чаще всего сталкиваются исследователи при попытке использовать большие данные для расчета статистических показателей, могут быть разделены на три основные группы: 1) необходимость разработки новых методологических подходов; 2) обеспечение доступа к данным и требования сохранения их конфиденциальности; 3) наличие современной инфраструктуры информационных технологий, адекватной задачам исследования.

3.1 Методология анализа больших данных. Согласно результатам рассмотренных пилотных проектов, национальные статистические службы наиболее часто в качестве основных трудностей, возникающих при работе с большими данными, указывают на проблемы в области методологии. Коренное отличие больших данных от других сведений, используемых в официальной статистике, заключается в том, что при работе с первыми исследователь получает только доступ к информации, но никак не влияет на процесс ее сбора. Большие данные чаще всего собираются не для исследовательских целей, а для решения конкретных прикладных задач в сфере бизнеса и управления. В то же время сбору данных в официальной статистике, как правило, предшествует этап разработки методологии проведения исследования. На этом этапе определяют, каким образом будет проводиться сбор данных, какие объекты будут участвовать в проведении исследования и какие, собственно, данные будут собираться в ходе статистического наблюдения. Таким образом, в официальной статистике исследователь еще до получения данных имеет некоторое представление относительно их структуры. При работе с большими данными на первоначальном этапе основная задача исследователя заключается в непосредственном знакомстве с ними, изучении структуры и взаимосвязей, выявлении ошибок и пропусков данных [7]. Проверка информации также может осуществляться путем ее сопоставления с другими источниками. Так, Федеральная статистическая служба Швейцарии перед включением в расчет индекса потребительских цен сканированных данных о продажах новой розничной сети в обязательном порядке в течение шести месяцев проводит сравнение этих данных с результатами, полученными с использованием традиционных методов сбора информации о ценах в магазинах ритейлера[14].

[11] CBS. A13 busiest national motorway in the Netherlands. URL: https://www.cbs.nl/NR/rdonlyres/25CE3592-A756-42B7-BABF-C3E4C4E9375B/0/a13busiestnationalmotorwayinthenetherlands.pdf.

[12] Текст информационного сообщения на сайте Центрального статистического бюро Нидерландов. URL: https://www.cbs.nl/nl-nl/nieuws/2016/01/helft-minder-verkeer-in-noord-nederland-door-ijzel.

[13] URL: https://www.cbs.nl/en-gb/our-services/innovation/nieuwsberichten/big-data/cbs-launching-center-for-big-data-statistics.

[14] Müller R. Scanner data in the Swiss CPI: An alternative to price collection in the field. URL: https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2010/zip.10.e.pdf.

Еще одна важная методологическая проблема - это репрезентативность данных. Какими бы большими ни были данные, они охватывают только ограниченную часть объектов, которая может не совпадать с интересующей исследователей генеральной совокупностью (см. рис. 2). Поэтому при анализе больших данных необходимо ответить на вопросы о том, какие объекты описывают имеющиеся данные и насколько характеристики этих объектов соответствуют характеристикам генеральной совокупности. Однако в случае больших данных точно ответить на поставленные вопросы не всегда представляется возможным. Например, данные, полученные от сотовых операторов, в разные периоды времени могут охватывать разное количество абонентов в зависимости от того, как часто владельцы телефонов пользуются своими устройствами. Кроме того, один телефон может использоваться несколькими людьми; в то же время один человек может иметь несколько телефонов. Такого рода неопределенность затрудняет использование данных сотовых операторов для анализа туристских и миграционных потоков населения, особенно если речь идет о сопоставимости данных во времени.

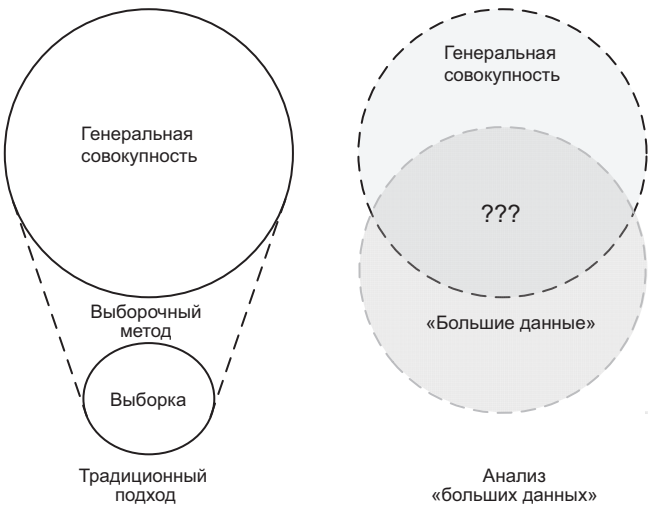

Рис. 2. **Проблема репрезентативности больших данных**

Поскольку отсутствует уверенность в репрезентативности больших данных, то при расчете на их основе показателей нельзя применять те же методы, которые традиционно используются в официальной статистике. Для проведения расчетов от исследователей требуется выдвижение обоснованных предположений относительно взаимосвязей между большими данными и характеристиками генеральной совокупности. Таким образом, выборочный метод в большей степени заменяется моделированием (вероятностное моделирование, эконометрические оценки и т. д.). Например, Центральное статистическое бюро Нидерландов при анализе данных, полученных с датчиков, регистрирующих движение транспортных средств на автомагистралях страны, выяснило, что качество данных непостоянно и в некоторых случаях информация может не фиксироваться в течение нескольких минут. В этой связи исследователи были вынуждены использовать байесовский рекурсивный фильтр для моделирования реальной дорожной ситуации [8].

Расчет показателей на основе больших данных требует от исследователей использования предположений и допущений в гораздо большей степени, чем при работе с традиционными источниками данных. Однако в целом такой подход не является чем-то новым для официальной статистики. Моделирование в той или иной степени применяется в настоящее время национальными статистическими службами при оценке малых выборок, восстановлении пропущенных значений, расчете показателей, очищенных от сезонных колебаний, и предварительной оценке макроэкономических показателей. Вместе с тем более активное использование модельных расчетов в официальной статистике потребует от национальных статистических служб более тщательной проверки используемых моделей и большей прозрачности методологии, поскольку в настоящее время пользователи, как правило, не имеют доступа к информации о том, какие предпосылки и допущения используются при расчете официальных показателей. В целом, анализ пилотных проектов показывает, что разнообразие источников больших данных и разница в используемых технологиях их сбора и обработки затрудняют разработку единых методологических подходов к использованию новых источников информации в официальной статистике [9].

3.2 Доступ к данным и вопросы конфиденциальности информации. Как правило, источники больших данных не включены в систему официальной статистики. Преимущественно эти данные носят конфиденциальный характер и находятся в распоряжении частных компаний. В этой связи

обеспечение доступа к данным и налаживание сотрудничества с поставщиками информации являются важной задачей для всех национальных статистических служб, приступающих к использованию больших данных в официальной статистике. В условиях отсутствия законодательного регулирования получение доступа к данным целиком и полностью зависит от инициативы национальных статистических служб. Кроме того, как отмечают представители национальной статистической службы Португалии, для успешной реализации совместных проектов требуется полная поддержка на высшем уровне от руководства всех заинтересованных организаций [10]. С точки зрения оптимизации временных и финансовых затрат, для национальных статистических служб предпочтительным является получение данных от одного единственного поставщика, а не от множества контрагентов. Например, национальная статистическая служба Италии получает сканированные данные о продажах розничных сетей от компании Nielsen, а также сотрудничает с национальной отраслевой ассоциацией[15].

Соблюдение конфиденциальности информации -это еще одно серьезное препятствие на пути использования больших данных в официальной статистике. В частности, исследования показывают, что данные мобильных телефонов потенциально могут найти применение во многих областях официальной статистики (оценка туристских потоков, миграции и численности населения). Однако передача этой информации для обработки вызывает вполне обоснованные опасения относительно сохранения конфиденциальности персональных данных абонентов. Опыт работы национальных статистических служб с большими данными показывает, что такие ограничения могут быть частично устранены за счет принятия дополнительных мер предосторожности. Во-первых, первичная информация может обрабатываться непосредственно поставщиками данных по запросу национальных статистических служб, которые в этом случае будут иметь доступ только к агрегированной информации, не содержащей персональные данные. Во-вторых, запросы на обработку данных могут передаваться через специально выбранного посредника[16]. В качестве примера можно привести опыт Эстонии, где национальная статистическая служба получает сведения, необходимые для оценки туристских потоков, от местной компании Positium LBS, имеющей доступ к данным сотовых операторов и специализирующейся на анализе данных мобильного позиционирования[17]. В-третьих, работа с данными может быть организована в форме частно-государственного партнерства (ЧГП) с привлечением всех заинтересованных сторон (стейкхолдеров) [11]. В-четвертых, процедура передачи данных может быть отрегулирована на законодательном уровне по аналогии с тем, как осуществляется регулирование передачи административных данных в официальной статистике[18].

3.3 Инфраструктура информационных технологий. Использование новых источников данных ставит вопрос о модернизации информационных технологий в системе официальной статистики. Обработка больших массивов данных требует наличия соответствующей вычислительной инфраструктуры. Представители национальных статистических служб, принимавших участие в проектах, связанных с использованием больших данных, отмечают, что существенную проблему может представлять не только объем данных, но также непрогнозируемый рост этого объема с течением времени до уровня, превышающего размер мощностей, выделенных для хранения данных. На стадии реализации пилотных проектов вопрос инфраструктуры является очень важным, так как низкая скорость вычислений и небольшой объем анализируемой информации могут стать препятствием на пути раскрытия потенциала больших данных.

[15] Brunetti A., Fatello S., Polidoro F. Preliminary results of scanner data analysis and their use to estimate Italian inflation. Workshop scanner data presentation. Rome,1-2 October 2015. URL: https://www.istat.it/it/files/2015/09/4.4-WS-Scanner-data-Rome-1-2-Oct_Brunetti_Fatello_Polidoro-Preliminary-results-of-scanner-data-analysis-and-their-use-to-estimate-Italian-inflation.pdf.

[16] Heerschap N., Ortega S., Priem A., Offermans M. Innovation of tourism statistics through the use of new big data sources. Presentation at the 12th Global Forum on Tourism Statistics. Prague, Czech Republic, May 15-16. 2014. URL: http://www.tsf2014prague.cz/assets/downloads/Paper%201.2_Nicolaes%20Heerschap_NL.pdf.

[17] Ahas R., Tiru M., Saluveer E., Demunter C. Mobile telephones and mobile positioning data as source for statistics: Estonian experiences. Presentation for NTTS, Brussels, 2011. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.3362&rep=rep1&type=pdf.

[18] UNECE. Outcomes of the UNECE Project on Using Big Data for Official Statistics. 2016. URL: http://www1.unece.org/stat/platform/download/attachments/77170975/Outcomes%20of%20the%20UNECE%20Project%20on%20Using%20Big%20Data%20for%20Official%20Statistics.docx?version=2&modificationDate=1456817253553&api=v2.

Внедрение новых информационных технологий связано с необходимостью осуществления дополнительных финансовых расходов, проведения организационных изменений и обучения персонала. При этом национальные статистические службы могут быть ограничены в возможностях использования аутсорсинга (в том числе облачных решений) при работе с большими данными в связи с конфиденциальным характером анализируемой информации. Универсальных решений в этом вопросе не существует, однако опыт реализации пилотных проектов показывает, что национальные статистические службы могут упростить свою задачу при работе с большими данными за счет кооперации с зарубежными коллегами и участия в международных проектах [12]. Использование гибридной инфраструктуры, сочетающей возможности традиционных реляционных баз данных с технологиями работы с большими данными типа Hadoop, также является одним из возможных вариантов внедрения новых технологий[20]. Кроме того, нагрузка на инфраструктуру может быть снижена за счет использования не полных наборов исходных данных, а агрегированной информации при условии организации сотрудничества с поставщиками данных или посредниками, готовыми взять на себя первичную обработку данных.

Заключение

В последние годы международные и национальные статистические организации приступили к активному исследованию проблемы использования новых источников информации в официальной статистике. Анализ проводился в рамках как исследовательских проектов, изучающих потенциальные возможности больших данных, так и пилотных проектов, предполагающих непосредственное внедрение новых источников информации в официальную статистику. Хотя опыт практического использования больших данных в официальной статистике относительно невелик, большинство представителей статистических организаций признают, что их применение может способствовать повышению качества статистических показателей.

Обзор проектов национальных статистических служб показывает, что дальнейшему внедрению больших данных в систему официальной статистики препятствуют ограничения, связанные с методологией, доступностью данных и соблюдением конфиденциальности информации, а также с наличием необходимой ИТ-инфраструктуры. По отдельности эти препятствия не кажутся непреодолимыми, однако взятые в совокупности, они, возможно, и являются основной причиной медленного внедрения больших данных в официальную статистику.

Принимая во внимание существующие ограничения, национальные статистические службы сконцентрировали свое внимание на интеграции новых данных с традиционными источниками официальной статистики (результатами опросов, административными данными и пр.) вместо попыток разрабатывать статистические показатели исключительно на основе больших данных. В качестве примера подобной интеграции можно привести изменение методологии прогнозирования урожая сельскохозяйственных культур статистической службой Канады. Традиционно прогноз строился на результатах опроса фермерских хозяйств о запасах полевых культур, размере посевных площадей и ожидаемой урожайности. С 2015 г. статистическая служба Канады при составлении прогноза, помимо опросных данных, стала использовать информацию о состоянии сельскохозяйственных земель, полученную со спутниковых снимков, а также климатические данные о температуре воздуха и количестве выпавших осадков. На международном уровне рост интереса статистического сообщества к вопросам совместного использования данных из различных источников информации выразился в запуске в 2016 г. специального проекта в рамках Группы высокого уровня по модернизации статистики Европейской экономической комиссии ООН[19]. Таким образом, вместо общих проектов, посвященных изучению потенциала больших данных, на первой план в официальной статистике выходят более конкретные проекты, направленные на интеграцию новых источников информации с традиционными.

[19] Virgillito A. Experiences in the use of Big Data for official statistics. Presentation at the international seminar «Think Big-Data innovation in Latin America. Chile, March 6, 2017. URL: http://www.cepal.org/sites/default/files/events/files/antonino_virgillito.pdf.

[20] URL: https://statswiki.unece.org/display/DI/Data+Integration+Home.

Литература

1. **Mayer-Schönberger V., Cukier K.** Big data: A revolution that will transform how we live, work, and think. Boston, MA: Houghton Mifflin Harcourt; 2013.

2. **McAfee A., Brynjolfsson E.** Big data: The management revolution // Harvard Business Review. October 2012. P. 59-68.

3. **Manyika J.** et al. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, McKinsey & Company. May 2011. URL: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

4. **Cavallo A.** Scraped data and prices in macroeconomics. Harvard University - Phd Dissertation, 2010.

5. **Vale S**. International collaboration to understand the relevance of big data for official statistics // Statistical Journal of the IAOS. 2015. Vol. 31. No. 2. P. 159-163. DOI: 10.3233/sji-150889.

6. **Norberg N., Sammar M., Tongur C.** A study on scanner data in the Swedish Consumer Price Index. Paper presented at the Ottawa Group Meeting on Prices. Wellington, 10-12 May 2011. URL: http://www.scb.se/Statistik/PR/PR0101/_dokument/KPI_namnden/A%20STUDY%20ON%20SCANNER%20DATA%20IN%20THE%20SWEDISH%20CPI.pdf.

7. **Daas P.J.H.** et al. Big data as a source for official statistics // Journal of Official Statistics. 2015. Vol. 31. Iss. 2. P. 249-262. DOI: 10.1515/jos-2015-0016.

8. **Braaksma B., Zeelenberg K**. «Re-Make/Re-Model»: Should big data change the modelling paradigm in official statistics? // Statistical Journal of the IAOS. 2015. Vol. 31. No. 2. P. 193-202. DOI: 10.3233/SJI-150892.

9. **Hackle P.** Big data: What can official statistics expect? // Statistical Journal of the IAOS. 2016. Vol. 32. No. 1. P. 43-52. DOI: 10.3233/SJI-160965.

10. **Dos Santos P.S., Lidonio F., Cardoso C.** Scanner data project: The experience of Statistics Portugal. Paper presented at the Workshop on Scanner Data. Stockholm, 7-8 June 2012. URL: http://www.scb.se/Statistik/PR/PR0101/_dokument/ScannerDataProject-the-experience-of-Statistics-Portugal.pdf.

11. **Florescu D.** et al. Will 'big data' transform official statistics? Paper presented at the European Conference on the Quality of Official Statistics. Vienna, Austria, 2-5 June 2014. URL: http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf.

12. **Struijs P., Braaksma B., Daas P.J.H.** Official statistics and big data // Big Data & Society. 2014. Vol. 1. Iss. 1. P. 1-6. DOI: 10.1177/2053951714538417.

BIG DATA AND OFFICIAL STATISTICS: A REVIEW OF INTERNATIONAL EXPERIENCE WITH INTEGRATION OF NEW DATA SOURCES

Dmitrii A. Plekhanov

Author affiliation: Institute for Complex Strategic Studies (ICSS) (Moscow, Russia). E-mail: plehanov@icss.ac.ru.

The growing use of digital devices and a massive increase in information flows in the modern world brought about new sources of information concerning our everyday life. These sources, collectively known as Big Data, provide unique opportunities for researchers to analyze quantitative data on various social and economic developments.

This paper provides a brief review of research and pilot projects, which have been carried out recently by national and international statistical organizations to analyze prospects of using Big Data sources in the official statistics. Results of surveyed projects indicate that use of Big Data in the official statistics is hindered by several serious impediments, such as unresolved questions concerning methodological approaches to new data collection and processing, data access and protection of data confidentiality, technical requirements to new IT infrastructure.

Therefore, the paper concludes that examples of statistical indicators calculated solely on the basis of Big Data sources are rare, and national statistical offices efforts are currently concentrated mainly on integration of Big Data sources with traditional sources of information.

Keywords: official statistics, big data, data confidentiality, statistical infrastructure.

JEL: C10, C80.

References

1. **Mayer-Schönberger V., Cukier K.** *Big data: A revolution that will transform how we live, work, and think*. Boston, MA: Houghton Mifflin Harcourt, 2013.

2. **McAfee A., Brynjolfsson E.** Big data: The management revolution. *Harvard Business Review*, October 2012, pp. 59-68.

3. **Manyika J.** et al. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, McKinsey & Company. May 2011. Available at: http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

4. **Cavallo A.** *Scraped data and prices in macroeconomics*. Harvard University - Phd Dissertation, 2010.

5. **Vale S**. International collaboration to understand the relevance of big data for official statistics. *Statistical Journal of the IAOS*, 2015, vol. 31, no. 2, pp. 159-163. DOI: 10.3233/sji-150889.

6. **Norberg N., Sammar M., Tongur C.** *A study on scanner data in the Swedish Consumer Price Index*. Paper presented at the Ottawa Group Meeting on Prices. Wellington, 10-12 May 2011. Available at: http://www.scb.se/Statistik/PR/PR0101/_dokument/KPI_namnden/A%20STUDY%20ON%20SCANNER%20DATA%20IN%20THE%20SWEDISH%20CPI.pdf.

7. **Daas P.J.H.** et al. Big data as a source for official statistics. *Journal of Official Statistics*, 2015, vol. 31, iss. 2, pp. 249-262. DOI: 10.1515/jos-2015-0016.

8. **Braaksma B., Zeelenberg K**. «Re-Make/Re-Model»: Should big data change the modelling paradigm in official statistics? *Statistical Journal of the IAOS*, 2015, vol. 31, no. 2, pp. 193-202. DOI: 10.3233/SJI-150892.

9. **Hackle P.** Big data: What can official statistics expect? *Statistical Journal of the IAOS*, 2016, vol. 32, no. 1, pp. 43-52. DOI: 10.3233/SJI-160965.

10. **Dos Santos P.S., Lidonio F., Cardoso C.** *Scanner data project: The experience of Statistics Portugal*. Paper presented at the Workshop on Scanner Data. Stockholm, 7-8 June 2012. Available at: http://www.scb.se/Statistik/PR/PR0101/_dokument/ScannerDataProject-the-experience-of-Statistics-Portugal.pdf.

11. **Florescu D.** et al. *Will 'big data' transform official statistics*? Paper presented at the European Conference on the Quality of Official Statistics. Vienna, Austria, 2-5 June 2014. Available at: http://www.q2014.at/fileadmin/user_upload/ESTAT-Q2014-BigDataOS-v1a.pdf.

12. **Struijs P., Braaksma B., Daas P.J.H.** Official statistics and big data. *Big Data & Society*, 2014, vol. 1, iss. 1, pp. 1-6. DOI: 10.1177/2053951714538417.