ПРОБЛЕМЫ И ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ БОЛЬШИХ ДАННЫХ В РОССИЙСКОЙ СТАТИСТИКЕ*

Н.С. Карпова,
А.Е. Суринов,
И.С. Ульянов

В статье раскрывается содержание понятия «большие данные», даются его характеристики, аргументируются возможности использования больших данных для подготовки официальной статистики и описываются связанные с этим проблемы и трудности. Представлен международный опыт работы по проектам больших данных в официальной статистике, а также перспективы их использования в российской статистике.

Авторами последовательно излагаются вопросы взаимосвязи больших данных и официальной статистики, которые полностью согласуются с основополагающими принципами официальной статистики, принятыми на 68-й сессии Генеральной ассамблеи ООН 23 января 2014 г. Анализируются результаты мониторинга, проведенного Статистическим отделом и ЕЭК ООН, в результате которого была получена информация о завершенных, действующих или потенциальных проектах (а также организационных условиях их осуществления) по большим данным в отдельных странах. Прокомментированы проблемы использования больших данных в официальной статистике, конкретизированы направления возможного использования больших данных как для замены существующей практики статистических наблюдений, так и в качестве дополнительного источника статистической информации и проверки полученных результатов.

Ключевые слова: большие данные, официальная статистика, статистическое наблюдение, административные данные, конфиденциальность информации, нагрузка на респондентов.

JEL: C80, C81, C82, M40.

В настоящее время основными источниками информации для официальной статистики являются данные, получаемые от респондентов в ходе проведения статистических обследований, и административные данные, предоставляемые органам статистики министерствами и ведомствами - держателями административных ресурсов. Третьим информационным источником в перспективе могут стать большие данные.

Большие данные и официальная статистика. *Большие данные* - это цифровая информация огромного объема, разнообразная по составу и имеющая неорганизованную структуру, поступающая в основном в режиме реального времени из многочисленных источников.

Характеристики больших данных определяют их ценность для официальной статистики. Так, большой объем данных может способствовать получению более детализированных статистических оценок различных явлений в обществе. В результате высокой скорости изменения информации увеличивается частота статистических оценок. Многообразие больших данных открывает возможности для организации статистики в новых областях (например, в области дистанционного зондирования и географических информационных систем) и получения информации о не-

Карпова Наталья Семеновна (Karpova@gks.ru) - начальник отдела организации научно-методологических работ управления организации статистического наблюдения и контроля Федеральной службы государственной статистики (г. Москва, Россия).

Суринов Александр Евгеньевич (fsgs_@gks.ru) - д-р экон. наук, профессор, руководитель Федеральной службы государственной статистики (г. Москва, Россия).

Ульянов Игорь Сергеевич (uliyanovIS@gks.ru) - д-р экон. наук, начальник управления организации статистического наблюдения и контроля Федеральной службы государственной статистики (г. Москва, Россия).

* Публикация подготовлена на основе текста выступления руководителя Росстата А.Е. Суринова на панельной сессии «Анализируй все. Революция больших данных» Петербургского международного экономического форума 2016 г.

наблюдаемых ранее официальной статистикой сферах. Разнородность источников происхождения больших данных может способствовать многоаспектности измерений, в результате чего повышается надежность статистических оценок, так как одно и то же явление рассматривается с разных точек зрения. Возможность получения сведений в масштабе реального времени решает проблему актуальности статистических оценок. Особо следует отметить, что использование больших данных существенно снизит нагрузку на респондентов, связанную с представлением статистической отчетности, и сократит расходы на официальную статистику. Таким образом, большие данные представляют собой источник новой информации, которую официальная статистика не может игнорировать.

Следует учитывать и тот факт, что в современном мире респонденты не очень охотно предоставляют сведения в рамках участия в статистических наблюдениях. А административные данные, например, могут искажаться в целях получения выгоды, занижения налоговой базы, нежелания выполнять требования законодательства, в частности Трудового кодекса, и т. д.

Использование больших данных для подготовки официальной статистики полностью согласуется с Основополагающими принципами официальной статистики, принятыми на 68-й сессии Генеральной ассамблеи ООН 23 января 2014 г. Так, в соответствии с пятым принципом данные для статистических целей могут собираться из всех видов источников с учетом их качества, своевременности, затрат и нагрузки, которая ложится на респондентов.

Большие данные и международное статистическое сообщество. Международное статистическое сообщество консолидирует усилия для решения общих вопросов, связанных с использованием больших данных для официальной статистики. Так, на 45-й сессии Статистической комиссии ООН в 2014 г. была учреждена Глобальная рабочая группа. Эта рабочая группа получила мандат на выработку стратегии и координацию глобальной программы использования больших данных для целей официальной статистики, в том числе в отношении показателей, включенных в Повестку дня в области устойчивого развития на период до 2030 года.

В рамках Глобальной рабочей группы было создано восемь специальных целевых групп по различным направлениям[1]:

- три группы по использованию данных соответственно мобильной телефонной связи, спутниковых изображений, сведений из социальных сетей;
- группа по обеспечению доступа к данным и налаживанию партнерских отношений с частным сектором и другими сообществами;
- группа по отслеживанию взаимосвязи между показателями устойчивого развития и видами использования больших данных;
- группа по учебной и профессиональной подготовке и укреплению потенциала;
- группа по классификации и системам обеспечения качества;
- группа по распространению информации о преимуществах и ценности больших данных (стратегия сбора средств для участия развивающихся стран в экспериментальных проектах).

За период, прошедший с момента создания Глобальной рабочей группы и до настоящего времени, национальными службами разных стран был выполнен ряд экспериментальных проектов по использованию данных мобильной телефонной связи и данных из социальных сетей, проведено апробирование статистической целевой программы для прогнозирования урожайности сельскохозяйственных культур на основе данных спутниковых изображений, методов дистанционного зондирования и геопространственных данных[2].

В 2014 г. Статистический отдел и ЕЭК ООН провели обследование проектов использования больших данных для подготовки официальной статистики, целью которого стал сбор информации о завершенных, действующих или потенциальных проектах, а также об организационных условиях их осуществления. Опросник был направлен в 78 национальных статистических организаций, 24 из которых представили информацию о 54 проектах. Респонденты отметили, что в рамках этих проектов один и тот же источник данных мог использоваться для разных отраслей статистики. Результаты этого обследования в

[1] Доклад Глобальной рабочей группы по вопросам использования больших данных для подготовки официальной статистики. ООН. Экономический и социальный совет. Статистическая комиссия. 47-я сессия. 8-11 марта 2016 г. С. 3.

[2] Там же.

части информации о потенциальных областях использования больших данных в официальной статистике представлены на рисунке[3].

Рисунок. **Возможности использования больших данных в различных областях официальной статистики** (в процентах)

Респонденты обследования отмечали, что в рамках этих проектов один и тот же источник данных может использоваться для разных отраслей статистики. Например, в Эстонии была разработана и применяется в различных сферах деятельности методология определения местонахождения с использованием сети мобильной связи, которая позволяет формировать статистику международных поездок и получать достоверную картину движения через границу физических лиц, выезжающих за рубеж или въезжающих в Эстонию. Также этими статистическими данными пользуется Центральный банк Эстонии для расчета объема импорта и экспорта услуг по организации поездок при составлении счета текущих операций в платежном балансе[4].

В 2015 г. был проведен опрос национальных статистических служб с целью выяснения их мнений относительно направлений использования больших данных для целей официальной статистики. В обследовании приняли участие в общей сложности 93 страны, в том числе и Россия. Были получены сведения о 115 связанных с большими данными проектах.

По результатам опроса было установлено, что национальные службы реализуют 42 проекта по использованию данных мобильной телефонной связи, 31 проект ориентирован на данные, извлеченные из веб-страниц, и 23 - на данные сканирования. В наибольшей степени большие данные используются: в статистике цен (на основе данных сканирования), затем в статистике туризма (данные мобильной телефонной связи), населения (данные мобильной телефонной связи), а также в статистике транспорта и труда (данные, извлеченные из веб-страниц)[5].

Группа высокого уровня по модернизации официальной статистики, работающая под эгидой Конференции европейских статистиков ЕЭК ООН, в течение последних трех лет организовала серию проектов международного сотрудничества по изучению потенциала больших данных для производства официальной статистики. Среди них проект «Sandbox» (песочница) по совместному использованию изолированной программной среды на основе веб-технологий для разработки и тестирования приложений, связанных с обработкой больших данных. Этот проект реализуется на базе Ирландского института высокопроизводительных вычислений (Irish Centre for High-End Computing - ICHEC). В его рамках выполнены следующие исследовательские работы:

- по использованию данных о просмотре страниц Википедии для получения информации о туристическом потенциале различных городов и курортов мира, мест культурного наследия ЮНЕСКО;

- по сбору информации о мобильности населения путем анализа данных социальной сети «Твиттер»;

- по использованию в статистических целях информации базы данных Всемирной торговой организации;

- по извлечению сведений о рабочих вакансиях с веб-сайтов организаций[6].

Проблемы использования больших данных в официальной статистике. Изучение феномена больших данных выявило ряд проблем в их использовании в официальной статистике.

[3] Доклад Глобальной рабочей группы по вопросам использования больших данных для подготовки официальной статистики. ООН. Экономический и социальный совет. Статистическая комиссия. 46-я сессия. 3-6 марта 2015 г.

[4] Доклад Генерального секретаря «Большие данные и модернизация статистических систем». ООН. Экономический и социальный совет. Статистическая комиссия. 45 сессия. 4-7 марта 2014 г.

[5] Доклад Глобальной рабочей группы по вопросам использования больших данных для подготовки официальной статистики. 47-я сессия. 8-11 марта 2016 г. С. 5-6.

[6] Outcomes of the UNECE Project on Using Big Data for Official Statistics. URL: http://www1.unece.org/stat/platform/display/bigdata/Big+Data+in+Official+Statistics.

Эти проблемы можно структурировать следующим образом:

- *юридические,* связанные с доступом к данным, находящимся в распоряжении организаций частного сектора, государственных учреждений или других организаций;

- связанные с *соблюдением конфиденциальности персональных данных*, то есть с созданием условий, при которых будут гарантированы неприкосновенность частной жизни и конфиденциальность персонифицированной информации, а также обеспечено доверие общественности к тому, каким образом будут использоваться большие данные в официальной статистике;

- *финансовые*: доступ к большим данным может быть весьма дорогим, поэтому необходима оценка затрат на получение и обработку данных в сопоставлении с получаемыми выгодами;

- *управленческие*: управление данными и обеспечение их защиты, возможное изменение организации работ статистической службы в связи с широкомасштабным использованием больших данных;

- *методологические:* при использовании данных методологическими проблемами являются их нестандартность, проблематичное качество, репрезентативность, изменчивость и размерность, а также пригодность существующих статистических методов для обработки таких данных;

- *технологические,* то есть вопросы, связанные с информационными технологиями, которые должны обеспечивать передачу и размещение огромных массивов данных, а также проведение масштабных вычислений при их обработке;

- *кадровые*: потребность в специалистах, обладающих навыками работы в области методологии, обработки и анализа больших данных, а также в области соответствующих информационных технологий.

Перспективы использования больших данных в российской статистике. Мы рассматриваем перспективы использования больших данных в российской статистике через призму сокращения статистической (отчетной) нагрузки на население и бизнес. Как показывает опыт, респонденты не испытывают особого желания участвовать в статистическом наблюдении. Об этом свидетельствуют отказы от участия в статистических обследованиях. Например, при проведении Всероссийской переписи населения 2010 г. более 1 млн жителей страны заявили переписчикам о своем отказе участвовать в переписи населения по личным (например, религиозным) соображениям; 2,6 млн человек переписчики не застали дома, поскольку они отсутствовали на протяжении всего периода переписи. Для обеспечения полноты учета населения в соответствии с процедурой, предусмотренной Федеральным законом «О Всероссийской переписи населения» (ст. 6, п. 3), из административных источников были получены и внесены в переписные листы данные о поле и возрасте 3,6 млн человек. Численность этой категории населения в 2002 г. составляла примерно 1,5 млн человек.

Если в 2012 г. в выборочном обследовании доходов населения и участия в социальных программах число отказавшихся участвовать респондентов составило 14,4% опрашиваемых домохозяйств, то в 2015 г. их было уже 15,1%.

В настоящее время перед Росстатом остро стоит проблема оптимизации баланса между информационными потребностями органов государственной власти и других потребителей и издержками бизнеса на ведение статистики. В этом вопросе незаменима роль административных и больших данных.

Уже сейчас информация по более 15% всего числа форм отчетности собирается Росстатом без привлечения хозяйствующих организаций и населения, то есть на основе административных данных, формируемых региональными и муниципальными органами власти.

Большие данные могут быть потенциально полезными как в качестве замены существующей практики статистических наблюдений, так и в качестве дополнительного источника статистической информации, используемого для различного рода досчетов, оценок, сопоставлений. А также и для поверки полученных результатов.

На сегодняшний день основным видом больших данных, используемых в области демографии, являются данные, получаемые на основе спутниковых изображений, причем используются не непрерывные потоки спутниковых изображений, а статические наборы данных. Для изучения вопросов, связанных с мобильностью людей, и их увязки с различными процессами, включая опасные природные явления, дорожно-транспортные происшествия и распространение инфекционных заболеваний, в мировой практике используются отчеты о звонках с мобильных телефонов. Разви-

ваются программные приложения, позволяющие увязать отчеты о звонках с демографическими показателями в целях оценки численности населения в отдельных небольших поселениях. В то же время их применение в глобальном масштабе до сих пор не стало реальностью отчасти из-за того, что ни одна компания не может предоставить отчетов о звонках при обеспечении полного охвата территории.

В рамках подготовки к Всероссийской переписи населения 2020 г. Росстат изучает возможность использования больших данных для расчета показателей *численности населения*, проживавшего на заданной территории в отчетном месяце; и *среднего количества человек*, находившихся на заданной территории у себя дома (на работе, на даче) в ночь на дату переписи. Мы считаем, что эта информация предоставит новые возможности контроля полноты охвата населения переписью и проверки качества заполнения переписных листов.

Что касается статистики торговли, то в настоящее время в России делаются первые шаги к использованию больших данных, формируемых в бизнесе. С августа 2014 г. на территории некоторых субъектов Российской Федерации (г. Москвы, Республики Татарстан, Московской и Калужской областей) Федеральной налоговой службой проводится эксперимент по применению контрольно-кассовой техники (ККТ) при наличных денежных расчетах и расчетах с применением платежных карт. Результаты эксперимента подтверждают техническую возможность организации передачи информации о расчетах в адрес налоговых органов через оператора фискальных данных, финансовую эффективность и удобство такой технологии.

Чтобы использовать эту информацию в статистических целях, статистическая служба должна разработать алгоритмы формирования статистических данных по группировкам национального классификатора видов продукции на основе информации, имеющейся у предприятий. Но проблема состоит в том, что предприятия для внутренних целей не обязаны использовать национальные (общероссийские) классификаторы, в частности классификатор продукции по видам экономической деятельности. Розничные торговые сети имеют собственные товарные справочники (классификации). Поэтому требуется разработка единой для всех ритейлеров классификации, построенной на использовании информации о штрих-кодах и согласованной с национальным классификатором.

Однако есть сегменты экономики, где влияние государства велико. В настоящее время в России единая номенклатура действует для алкогольной продукции, производство и оборот которой находятся под особым контролем государства. С 1 января 2016 г. (в соответствии с постановлением Правительства Российской Федерации от 29 декабря 2015 г. № 1459) вся информация об объемах розничной продажи алкогольной продукции от хозяйствующих субъектов поступает с использованием программно-аппаратных средств в единую государственную автоматизированную информационную систему (ЕГАИС). Обладателем информации, содержащейся в ней, является Российская Федерация, оператором - Федеральная служба по регулированию алкогольного рынка.

Таким образом, большие данные, содержащиеся в системе ЕГАИС, при организации соответствующего к ним доступа уже в ближайшее время могут стать источником для формирования официальной статистической информации об объемах розничной продажи алкогольной продукции на отечественном рынке.

Одним из наиболее перспективных направлений использования больших данных является статистика цен. Традиционная схема сбора информации о ценах предполагает их регистрацию - по определенному набору (корзине) товаров и услуг в местах их реализации - специалистами органов государственной статистики. При всей масштабности наблюдения возможности традиционных методов сбора ценовой информации ограничены.

Использование больших данных может способствовать повышению качества расчета индекса потребительских цен за счет учета в нем значительно большего объема данных, обеспечению возможности большей детализации информации о ценах и индексах цен, а также должно позволить снизить нагрузку на респондентов, сократить время на сбор информации и, в конечном итоге, привести к существенной экономии финансовых ресурсов.

В последние годы ряд стран, в частности Норвегия, Нидерланды, Италия, Дания, ведут работу по привлечению альтернативных источников для получения информации о потребительских ценах.

С предложением о разработке документа, который позволит определить единые понятия и принципы использования больших данных при расчете ИПЦ, к представителям ЕЭК, Евростата, МОТ обратились участники Совместного заседания ЕЭК/МОТ по индексам потребительских цен в г. Женеве в 2016 г., в котором приняли участие и представители Росстата.

Основными источниками информации являются данные сканирования, получаемые от ритейлеров, а также данные, собираемые в сети Интернет путем извлечения информации с вебстраниц. Использование сканированных данных в сфере розничной торговли позволяет практически в режиме реального времени получать информацию не только о ценах на продаваемые товары, но и данные о фактических объемах продаж.

Опыт стран, в которых применяются сканированные данные, показывает, что их использование дает возможность значительно расширить выборку товаров, получать единое описание их характеристик на основе штрих-кодов и обеспечивать сопоставимость данных. Кроме того, формируемая в рамках сканирования информация об объемах продаж может служить источником данных при формировании весов для расчета ИПЦ.

Вместе с тем при использовании сканированных данных выявились и проблемы, связанные, в частности, с обеспечением стабильного и своевременного получения информации, формирования справочников товаров с учетом штрих-кодов в условиях обработки больших массивов данных. Успешное использование сканированных данных возможно только в случае установления тесных и постоянных контактов с ритейлерами, предоставляющими свои базы данных о ценах на товары и проданном их количестве.

В настоящее время Росстат изучает возможности применения больших данных в статистике потребительских цен. Очевидно, что будет необходимо модернизировать методики расчета ИПЦ с учетом особенностей больших данных, включая частоту их обновления и источники. Предстоит решить вопросы сопряжения данных с действующими классификациями, оценки их качества и репрезентативности, потребуется более широкое внедрение в практику методов математического и статистического моделирования.

Также большие данные - это потенциальный элемент информационной статистической системы и один из возможных источников данных по таким статистическим направлениям, как цены, торговля и услуги, туризм, занятость, демография. Кроме того, использование больших данных согласуется с программой модернизации процесса подготовки статистической информации и укрепления потенциала российской статистики, в которой предусмотрены значительные изменения традиционных подходов к сбору, обработке и распространению официальной статистической информации, направленные на снижение респондентской нагрузки и экономии бюджетных средств, выделяемых на статистическую деятельность.

Для интеграции источников больших данных в информационную систему прежде всего необходимо решить проблему доступа к данным. Эта работа должна вестись по двум направлениям. Во-первых, это создание нормативно-правовой базы, регулирующее механизм доступа к большим данным и их защиты, включая соблюдение принципов конфиденциальности. В настоящее время законодательством Российской Федерации в области официального статистического учета не регламентировано использование больших данных при формировании официальной статистической информации.

Во-вторых, необходимо на взаимовыгодной основе налаживать партнерские отношения с государственными организациями и коммерческими компаниями, владеющими массивами больших данных.

По нашему мнению, официальная статистика нуждается в больших данных. Статистика и большие данные являются взаимодополняющими компонентами бурно развивающейся науки о данных (Data science). При этом статистика имеет немалый опыт развития и обладает сегодня несомненными преимуществами. Так, в процесс подготовки официальной статистики встроены системы обеспечения качества и методологии, принятые на международном уровне. Официальная статистика основана на принципах профессиональной независимости и доверия. Применение этих подходов, уже «отлаженных» официальной статистикой, полезно и в системах больших данных.

PROBLEMS AND POSSIBILITIES FOR USING BIG DATA IN THE RUSSIAN STATISTICS*

Natalia S. Karpova
Author affiliation: Federal State Statistics Service (Rosstat) (Moscow, Russia). E-mail: Karpova@gks.ru.

Alexander Ye. Surinov
Author affiliation: Federal State Statistics Service (Rosstat) (Moscow, Russia). E-mail: fsgs_@gks.ru.

Igor S. Uliyanov
Author affiliation: Federal State Statistics Service (Rosstat) (Moscow, Russia). E-mail: uliyanovIS@gks.ru.

The article outlines the definition of the concept of Big Data, presents its applicability for official statistics, and reviews problems and challenges associated with it. The paper introduces international experience in carrying out Big Data projects in statistics, as well as prospects of using this concept in the Russian statistics.

The authors give consecutive account of interdependence between Big Data and official statistics, which perfectly coincides with fundamental principles of official statistics adopted at the 68th General Assembly of the United Nations on January 23, 2014. There is an analysis of monitoring results conducted by the Statistics Division and Economic Commission for Europe which resulted in gathered information on completed, on-going and potential Big Data projects (as well as organizational conditions for their execution) in selected countries. The authors comment on challenges and problems which have to be overcome in order to use Big Data in official statistics; they specify implementation directions for the concept of Big Data not only to substitute the existing statistical observation practice, but also to use it as an additional source of statistical information and a way to check validity of the obtained results.

Keywords: Big Data, official statistics, statistical observation, administrative data, confidentiality of information, respondent burden.

JEL: C80, C81, C82, M40.

* This article was prepared on the basis of report made by A.Ye. Surinov, the Head of the Federal State Statistics Service (Rosstat), at the panel session «Analyzing everything: the continued big data revolution» of the 2016 St. Petersburg International Economic Forum.