

Большие данные в официальной статистике: взгляд на проблему

Александр Евгеньевич Суринов

Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ),
г. Москва, Россия

В статье рассматривается широко обсуждаемая в статистических кругах тема использования больших данных в официальной статистике. Несколько лет назад в подавляющем числе публикаций зарубежных и отечественных исследователей активно пропагандировалась идея замены новым источником информации — большими данными — показателей официальной статистики. В последние годы возобладали более трезвый взгляд на эту проблему, вопрос стоит об использовании массивов больших данных для получения официальных статистических оценок, встраивания в статистическое производство этого нетрадиционного источника. Большую роль в смене направленности дискуссии оказала активная позиция Статистической комиссии ООН. Именно Статистическая комиссия ООН, используя свой авторитет и возможности, смогла перевести обсуждение проблем применения больших данных в официальной статистике в конструктивную плоскость. В этот процесс были включены все региональные статистические площадки ООН, специализированные организации и программы, другие международные агентства. Более того, непосредственное участие в разработке, апробировании и реализации инновационных подходов в официальных статистических наблюдениях сразу стали принимать национальные статистики, работающие на правительство, в университетах, исследовательских центрах и бизнесе.

Автором проанализирован опыт внедрения больших данных в практическую деятельность национальных статистиков, описаны преимущества и недостатки больших данных с позиции возможного их использования в официальной статистике, направления этой деятельности, сформулированы предложения по участию российского статистического сообщества в этом процессе. Автор доказывает необходимость отказа от противопоставления больших данных в качестве информации об обществе и экономике данным официальной статистики, основанным на традиционных источниках. По его мнению, большие данные в настоящее время не являются полной альтернативой сведениям, собираемым на основе классических наблюдений, но могут быть встроены в национальные системы сбора информации об отдельных характеристиках социально-экономических феноменов. Расширение их использования в значительной степени зависит от уровня развития информационно-коммуникационных технологий и степени внедрения ИКТ в общественную жизнь.

Ключевые слова: официальная статистика, статистическое наблюдение, большие данные, традиционные источники данных, нетрадиционные источники данных, ИКТ.

JEL: C15, C55, C81, C82, C83.

doi: <https://doi.org/10.34023/2313-6383-2023-30-2-5-22>.

Для цитирования: Суринов А.Е. Большие данные в официальной статистике: взгляд на проблему. Вопросы статистики. 2023;30(2):5–22.

Big Data in Official Statistics: A View of the Problem

Alexandr Ye. Surinov

National Research University Higher School of Economics (HSE University), Moscow, Russia

This article addresses the use of Big Data in official statistics, a much-discussed topic in the statistical community. The idea of replacing official statistics' indicators with a new information source, Big Data, was actively promoted in an overwhelming number of publications by foreign and domestic researchers several years ago. A more realistic perspective of this issue has gained ground in recent years; the issue is how to incorporate this unconventional source into statistical production and use Big Data for official statistical evaluations. The UN Statistical Commission's proactive stance greatly contributed to the discussion's shift in focus. The UN Statistical Commission was able to make the conversation about issues with Big Data application in official statistics a constructive process, leveraging its authority and opportunities. This process included all regional statistical platforms of the UN, specialized organizations and programs and other international agencies. The creation, testing, and implementation of novel approaches in official statistical surveys also involved national statisticians working for the government in universities, research centers and business.

The article analyzes the practical application of Big Data in national statisticians' line of work, outlines the benefits and drawbacks of Big Data regarding its potential application in official statistics, outlines the directions of this activity, and formulates suggestions for the involvement of the Russian statistical community in this process. The author argues the need to avoid contrasting Big Data as information on society and economy versus official statistics based on traditional sources. In his opinion, Big Data is currently not a complete alternative

to data collected on the basis of classical natural phenomena but can be built into national systems for collecting information on specific characteristics of socio-economic phenomena. The expansion of their use largely depends on the level of development of information and communication technologies as well as ICT integration in public life.

Keywords: official statistics, statistical observation, Big Data, traditional data sources, unconventional data sources, ICT.

JEL: C15, C55, C81, C82, C83.

doi: <https://doi.org/10.34023/2313-6383-2023-30-2-5-22>.

For citation: Surinov A.Ye. Big Data in Official Statistics: A View of the Problem. *Voprosy Statistiki*. 2023;30(2):5–22. (In Russ.)

Введение

В последние 10 лет феномен больших данных подвергся многочисленным исследованиям. Первоначально большие данные признавались способными заменить всю социально-экономическую информацию, включая официальную статистику. Особенное восхищение вызывала быстрота получения оценок (режим реального времени) и высокая степень их детализации, позволяющая количественно измерить малые области изучения, например, рынки отдельных товаров и услуг, привязаться к географическим координатам или почтовому адресу при пространственном анализе. При этом не обращалось внимание на то, что сами массивы данных не постоянны, и поэтому зачастую нельзя понять, как развивается явление во времени. Они содержат информацию неструктурированную, из-за чего регистратор данных не всегда может идентифицировать их носителя в качестве объекта наблюдения. В большинстве случаев отсутствует метаданная. Сегодня проблемы больших данных в качестве источника сведений о жизни общества и об экономике хорошо изучены. И большинство исследователей, признавая эти проблемы, справедливо считают правильным все-таки прилагать усилия по включению больших данных в экономические измерения, анализ, прогнозирование и выработку политики.

Нам импонирует подход, рассматривающий большие данные в качестве «сырой» информации, которую полезно использовать в анализе современного общества. Эти огромные объемы быстроменяющихся сведений, представленных в различных форматах, требуют преобразования в данные, пригодные для использования при принятии решений. И если для выработки оперативных решений или краткосрочного прогноза часто можно обойтись фрагментарной информацией, на основе которой вырабатываются гипотезы, и разработать программу действий, то официальная

статистика требует дополнительных вложений в создание инструментов для такого преобразования. Это связано с тем, что именно официальная статистическая информация служит основой для прогноза и анализа, обеспечивая полный охват оцениваемого явления, сопоставимость оценок во времени и пространстве, их согласованность с различными источниками. Конечно, предварительные статистические оценки показателей могут быть построены на основе больших данных, и более того, учитывая оперативность измерений на их основе, они могут быть весьма полезны. Но в конце концов окончательные оценки должны базироваться на данных, надежность которых не вызывает сомнения у пользователей. В этом и состоит особенность включения в статистические наблюдения больших данных, замены ими традиционных методов сбора первичной информации.

Не стоит идеализировать первичные данные, собранные в ходе традиционных статистических наблюдений. Профессиональные статистики прилагают колоссальные усилия, чтобы снизить негативное влияние на их качество отказов респондентов от участия в обследованиях, смещения оценок, неполноты ответов, ошибок при обработке сведений, их хранении и пр. Но итоговые официальные статистические показатели в основе имеют первичные данные, прошедшие этап редактирования (чистки), включая исправление выявленных ложных данных, коррекцию на неполноту сообщенных сведений и т. д. Этот элемент системы обработки первичных данных является одним из главных в общей архитектуре современного статистического производства. Стоит также отметить и то, что официальные статистические наблюдения проводятся по прозрачной методологии, на регулярной основе, прилагаются огромные усилия по обеспечению сопоставимости результатов измерений и однородности единиц статистического наблюдения, применяются официальные классификации. Есть много и других плюсов.

Основываясь на материалах, обсуждаемых в статистическом мире, мы не видим заметных достижений официальной статистики в использовании больших данных, несмотря на значительные усилия, которые прилагаются в этом направлении. А вот в прогнозировании и анализе успехи в применении больших данных очевидны и по довольно широкому спектру тематики: экология, безопасность, финансы, рынок труда, маркетинг, метеорология, медицина, логистика, туризм, стихийные бедствия и пр.

В сентябре 2022 г. в г. Москве Евразийская экономическая комиссия организовала международный семинар «Большие данные и официальная статистика: вызовы и возможности на пространстве Евразийского экономического союза», где всесторонне обсуждались проблемы и опыт использования больших данных в официальной статистике (см. https://eec.eaeunion.org/comission/departament/dep_stat/events/stranitsy/meropriyatie.php?bitrix_include_areas=Y&clear_cache=Y). При подготовке данной статьи были использованы материалы, положенные в основу доклада автора на тему «Проблемные вопросы использования больших данных в социально-экономических измерениях», представленного автором на семинаре.

Большие данные и официальная статистика. Движение навстречу

Задача использования информации, содержащейся в массивах больших данных, для официальной статистики стала актуальной в связи с тем, что статистические наблюдения довольно дороги и национальные статистические службы практически всех стран живут в условиях ограниченного бюджета. А лица, принимающие решения о выделении ресурсов статистике, часто высказывают мнение о том, что большие данные являются надежным источником сведений, который может заменить собираемую традиционным способом информацию.

В своих рассуждениях мы исходим из того, что большие данные представляют собой огромные массивы многообразной неопределенно структурированной и неструктурированной информации, образующейся как побочный результат деятельности государственных структур, бизнеса и некоммерческих организаций. Эта ин-

формация охватывает многие стороны жизни общества, непредсказуемо возникает и исчезает, а ее создание производится многочисленными устройствами, фиксирующими объекты и события. Явление больших данных, как доступного источника информации для принятия решений, связано с появлением технологических возможностей собирать, хранить, передавать, преобразовывать и анализировать колоссальные массивы данных.

Часто вместо определения больших данных используются характеристики массивов информации. В настоящее время в качестве главных характеристик больших данных эксперты выделяют «пять V»: объем информации, как количество записей и их характеристик (volume), скорость прироста и обработки (velocity), многообразие источников, носителей и форматов данных (variety), достоверность (veracity) и ценность (value) [1–5].

Буквально несколько лет назад для идентификации больших данных ограничивались моделью из трех первых характеристик (модель трех V) [6].

В последних исследованиях на тему больших данных особо подчеркивается то, что они должны представлять ценность для достижения заданной цели. Если мы планируем использовать большие данные для формирования показателей официальной статистики, то ценностью обладает информация, которую можно преобразовать и получить соответствующие показатели, отвечающие требованиям конкретного статистического наблюдения и официальной статистики в целом. При этом очевидно, что разные данные требуют не одинаковых усилий для проведения таких манипуляций. Если посмотреть на массивы данных, которые можно отнести к классу больших данных, то львиная их доля требует значительной доработки, включающей систематизацию, проверку, чистку и, самое главное, определение полей статистической информации, которую они способны заменить без снижения качества результатов измерений. Такие данные в основном могут представлять ценность именно как первичная информация. И причины здесь не только те, которые перечислены выше, но и обеспечение охвата изучаемой совокупности, так как для измерений необходимо понимание того, какая часть исследуемой совокупности была включена в наблюдение. Так, например, в ЕС признается успешным проект по использованию данных о вакансиях, полученных с интернет-страниц работодателей [5].

Но все ли работодатели ищут работников таким образом? Можно ли быть уверенным, что с помощью этого инструмента зафиксированы все вакансии? Хотя, конечно, такие данные, да еще полученные в оперативном режиме, имеют несомненную ценность для разработчиков политики и мониторинга спроса на рабочую силу. Но в официальной статистике должны быть учтены или корректно оценены все вакансии.

Для официальной статистики, нацеленной на получение адекватного результата измерений, качество данных в многочисленных международных и национальных документах описывается такими их характеристиками, как согласованность с другими источниками или хотя бы непротиворечивость, полнота, приемлемая или хотя бы измеренная точность оценок. Из-за того, что большие данные не являются изначально продуктом некой деятельности, ориентированной на создание информации (то есть инфор-

мация является сопутствующим основной деятельности феноменом), требовать от создателей массивов больших данных их соответствия таким характеристикам вряд ли возможно. Но добиваться максимального достижения заданных параметров через системы обработки, чистки, исправления ложных сведений и заполнения неответов необходимо.

Большие данные не должны отождествляться с большими объемами информации, например, файлами микроданных, сформированными из сведений, собранных при переписях населения, промышленных и сельскохозяйственных переписях, или из регистров информации министерств и ведомств (так называемые административные данные — *administrative records*). В таблице приведено сопоставление основных характеристик больших данных с традиционными источниками информации, используемой в качестве основы для официальной статистики.

Таблица

Основные характеристики различных источников данных

Характеристики данных	Источники данных		
	Государственные (официальные) статистические наблюдения, включая ведомственные	Ведомственные регистры, реестры, административные записи	Большие данные (операторы мобильной связи, коммерческие банки, торговые сети, Интернет, службы наблюдения за перемещением транспортных средств, за погодными условиями и пр.)
Владелец данных	Государство (национальная статистическая система)	Государство (ведомства, отвечающие за функционирование административных регистров и реестров)	Частный сектор и отдельные государственные организации
Форма распространения данных	Готовые документы, регламентные таблицы, файлы обезличенных микроданных	Файлы обезличенных и персонифицированных микроданных	Файлы микроданных, позволяющие идентифицировать объект наблюдения
Тип данных	Структурированные	Структурированные частично или полностью	Неструктурированные
Возможность доступа к данным	Регламентируется	Регламентируется	Не регламентируется
Наличие утвержденных правил при создании и распространении данных	Нормативно-правовой акт	Нормативно-правовой или ведомственный акт	Решение организации-владельца/создателя
Скорость создания	Наличие временного лага после события	Наличие временного лага после события	В режиме реального времени или с коротким временным лагом
Обеспечение качества (достоверность, прозрачность методологии, своевременность)	Национальная статистическая система / Национальная статистическая служба	Ведомство, ответственное за функционирование регистра, реестра	Нет обязательства
Наличие утвержденной методологии	Обязательно	Желательно	Нет обязательства
Использование официальных классификаторов технико-экономической и социальной информации	Обязательно	Обязательно	Нет обязательства
Устойчивость во времени, смена методов сбора и обработки требует обеспечения сопряженности динамических рядов	Обязательна	Необязательна, зависит от изменения нормативной базы	Отсутствует

Большие массивы информации из административных регистров систем, как правило, пригодны для получения официальной статистики по причине открытости использованных форматов и структуры баз данных, а изменения этих параметров связаны с определенными легитимными публичными решениями. Национальные статистики имеют возможность подготовиться к внедрению новаций и обеспечить согласованность динамических рядов. Более того, в большинстве случаев государственные информационные ресурсы предполагают использование официальных классификаций, что, в свою очередь, дает возможность для унифицированной систематизации результатов измерений и их согласованности на концептуальном уровне. Важно отметить связь статистических концепций и определений (и даже терминологии) с действующим законодательством, так как благодаря этому достигается лучшее восприятие вопросов статистических анкет и форм отчетности со стороны респондентов и увязка результатов измерений, проведенных на основе статистических наблюдений, и административных записей. Без методологического единства со статистикой государственных финансов и платежным балансом составление, например, национальных счетов (самая сложная задача!) было бы более трудоемко, а результаты менее обоснованы. Другой пример: в странах, где статистическая отчетность предприятий гармонизирована с налоговой и таможенной статистикой, сбор сведений в статистических целях ограничен или отсутствует, так как все (или почти все) нужные данные содержатся в ведомственных информационных системах. Такого единства подходов к организации государственных информационных систем и файлов больших данных, созданных негосударственным сектором, нет.

Считается (и об этом свидетельствует большое число публикаций и нацеленность практических проектов), что использование больших данных в официальной статистике потенциально может обеспечить:

- отказ от проведения традиционных статистических наблюдений и, как следствие, снижение нагрузки на респондентов и стоимости «статистического производства»;
- возможность разработки новых статистических показателей и расширение тематики статистической программы;

- высокую степень детализации статистических оценок, расчет показателей в привязке к малым совокупностям объектов (территориям, отраслям, социальным стратам), не охваченных регулярными статистическими наблюдениями;

- сокращение времени получения итогов и создание информационной базы для оперативных оценок официальных показателей, расчетов опережающих индикаторов.

- повышение общего уровня качества официальной статистики через использование данных, полученных от регистрирующих систем, фиксирующих события независимо от желания респондента сотрудничать со статистикой.

Так, полезные результаты могут быть получены при использовании географических больших данных. Сегодня пространственная статистика стала одним из наиболее быстро развивающихся направлений. Но использование территориальных иерархических классификаторов территорий не всегда дает достаточную основу для оценки реального социально-экономического или демографического процесса. В последнее время аналитики все чаще формируют данные не по «юридически оформленным» территориям или поселениям, а по их группам, объединенным общностью характеристик изучаемого явления, спецификой пространственной концентрации. Примерами здесь являются исследования ОЭСР, которые предлагают систематизацию населенных пунктов на основе объективных характеристик, таких как численность жителей, плотность расселения, расстояние до населенного пункта-ядра агломерации и пр. В этой связи новые возможности анализа возникают при использовании, например, таких данных, как данные ночного освещения или застроенной земной поверхности, полученные со спутниковых изображений. Не менее важным направлением использования для определения «экономических» зон, образованных городским центром и окружающей его периферией, являются большие данные о времени в пути между ними [7].

При использовании больших данных в качестве источника исходной информации для официальной статистики национальная статистическая служба должна обеспечить качество данных, под которым понимается совокупность следующих характеристик распространяемой информации:

- целостность, как обязательность применения научно-обоснованной методологии;

- прозрачность оценок, как доступность метаинформации о методах сбора и обработки первичных данных;

- достоверность и точность, как степень адекватности оценок (доверительные интервалы, ошибки выборки, статистические расхождения);

- своевременность, как соответствие данных потребностям пользователей во времени;

- интерпретируемость, как возможность использования полученных результатов для характеристики явления;

- согласованность, как непротиворечивость результатов измерений, проведенных на данных из разных статистических наблюдений;

- полезность, как удовлетворение потребностей пользователей в информации;

- международная сопоставимость, как соответствие национальной методологии международным рекомендациям и ее практическая реализация.

Ответственность за качество результатов официальной статистики при переходе на новый источник информации (большие данные) с национальной статистической системы не снимается. Более того, признание массива таких данных в качестве источника сведений для официальных оценок означает, что национальная статистическая система принимает на себя обязательство обеспечить качество статистики в соответствии с критериями, описанными выше. Здесь должны быть решены следующие проблемы:

- взаимодействие с владельцами данных в части доступа к ним и к метаданным, включая описание методов сбора и обработки первичных сведений;

- уникальность массивов данных и спонтанность их возникновения/исчезновения;

- территориальная и динамическая несопоставимость данных;

- разработка механизма поиска данных, оценки возможности их использования (ценности) и методов преобразования;

- использование неофициальных классификаций и методов систематизации объектов и случаев наблюдений;

- неполнота охвата наблюдением изучаемых объектов и возможное наличие смещения в их характеристиках;

- сложность структур данных и многообразие форматов и архитектуры хранилищ, систем доступа.

Перечисленные выше характеристики имеют первостепенное значение для определения возможности использования в экономической статистике больших данных.

На наш взгляд, вопрос: могут ли большие данные полностью заменить официальную статистику, не стоит. А как использовать их в официальной статистике? Вот главный вопрос! Не замечать то, что для описания социально-экономической действительности появился новый источник оперативных и детализированных сведений нельзя. Аналитики будут его использовать, нравится это или нет статистикам.

Специалисты, работающие с информацией, понимают, что большие данные похожи на склад, где никогда не проводилась инвентаризация, или на свалку, где нужно найти то, что имеет ценность для решения конкретных вопросов измерений, анализа, прогнозирования. И далее необходимо определиться с инструментами преобразования необработанных сведений в тот продукт, который и станет основой для проведения расчетов официальных статистических показателей.

Большими данными, имеющими ценность для официальной статистики, являются только пригодные для использования в качестве первичных статистических данных. Их ценность и состоит в возможности эффективной замены традиционных статистических наблюдений, позволяющей обеспечить высокую оперативность, точность оценок и максимально возможную детализацию описания явлений.

Таким образом, использование больших данных в официальной статистике означает получение возможности доступа к информации из различных источников и реализация алгоритмов их преобразования в файлы микроданных, содержащие характеристики изучаемых объектов или событий. Преобразованные для дальнейшей обработки данные должны позволять идентифицировать изучаемые объекты таким образом, чтобы можно было получать информацию в привязке к периоду времени или дате, территории. Например, для экономической статистики важна отраслевая принадлежность предприятия, характеристики объемов выпуска по номенклатуре продукции, стоимость активов и обязательств, результаты финансово-хозяйственной деятельности и пр. Для социальной статистики важны демографические, образовательные, экономические признаки человека, условия жизни, состав семьи и домохозяйства.

Для таких характеристик в официальной статистике используются стандартные экономико-статистические классификаторы, гармонизированные с международными классификациями. Эти классификаторы составляют основу нормативно-справочной информации государственных информационных систем, в которые входят в качестве обязательного инструмента для систематизации результатов наблюдений.

А готовы ли владельцы больших данных использовать такие классификации? Если нет, то необходимо при преобразовании файлов больших данных в микроданные для включения в статистическое производство применять переходные ключи или убедить (обязать) их владельцев отказаться от собственных классификаций и перейти на официальные. Последнее довольно сомнительно. Для разработки переходных ключей должен быть известен алфавит и иные параметры кода, примененного владельцем больших данных. Захотят ли они предоставлять такую информацию? Есть у них обязанность делать это? Но без использования, например, территориального, отраслевого, продуктового классификаторов не будут доступны соответствующие оценки, системно описывающие структуру изучаемой совокупности. Сегодня во многих странах национальные статистические службы ведут активные переговоры с торговыми сетями для решения этого вопроса.

Эффективное применение больших данных в официальной статистике возможно только в том случае, если сведения, содержащиеся в них и признанные ценными для конкретных измерений, будут рассматриваться исключительно в качестве первичной информации. Эти данные должны быть подвергнуты самому тщательному изучению, по результату которого надо принять решение использовать этот источник данных, или нет. Если принимается решение использовать, то содержащаяся в них информация должна быть подвергнута процедуре редактирования, включающей определение полей ответов и ложной информации, нахождение эффективных алгоритмов поиска или конструирования донорской информации, ее импутации в массив данных, который и станет основой для дальнейших расчетов. Этот массив должен представлять собой не агрегированные данные по всей совокупности или по ее частям, а именно микроданные, которые «привязаны» к конкретной и идентифицируемой еди-

нице статистического наблюдения или к объекту, выбранному в качестве такового. Такое решение позволяет соединить сведения об объекте из всех источников: традиционных статистических наблюдений и административных регистров и полученные из массивов больших данных.

Усилия Статистической комиссии ООН по использованию больших данных в официальной статистике. История вопроса

Международное статистическое сообщество пристально отнеслось к феномену больших данных. Тем более, что основополагающие принципы официальной статистики, утвержденные в новой редакции Статистической комиссией ООН в 2014 г., подтвердили право статистиков выбирать различные источники сведений для статистических наблюдений.

Статистическая комиссия ООН, высший мировой форум официальной статистики, начала подробное рассмотрение темы больших данных на 45-й сессии в марте 2014 г. Данная тема стала обсуждаться на всех последующих сессиях, а также на сессиях региональных статистических форумов ООН, таких, как конференция европейских статистиков, Комитет по статистике Азии и Тихого океана и др. Для координации усилий национальных статистиков и международных статистических организаций Статистическая комиссия образовала в 2015 г. Глобальную рабочую группу по большим данным — Big Data UN Global Working Group (см. <https://unstats.un.org/bigdata/>). В зону ее ответственности вошли вопросы методологии, качества, технологий, доступа к данным, законодательства, конфиденциальности, управления применительно к использованию больших данных в официальной статистике. В мандате группы ее деятельность была увязана с организацией мониторинга хода достижения целей в области устойчивого развития [8].

В марте 2016 г. на заседании 47-й сессии Статистической Комиссии ООН Глобальная рабочая группа подготовила доклад «Report of the 2015 Big Data Survey», в котором приведены информационные и аналитические материалы о получении доступа к большим данным, о классификации больших данных и методах их обработки, в том числе для мониторинга достижения целей в области устойчивого развития, об итогах глобаль-

ного обследования по вопросам, касающимся больших данных в национальной статистической практике [9].

На всех форумах, проводимых под эгидой Статистической комиссии ООН, реализуется задача не противопоставлять большие данные официальной статистике, а направить усилия на разработку инструментов, дающих возможность их эффективного использования в целях официальной статистики (снижая затраты на организацию статистических наблюдений, повышая оперативность измерений, добиваясь максимальной детализации оценок). Например, в регионе Европейской экономической комиссии диалог по вопросам официальной статистики и больших данных организован на веб-странице <https://statswiki.unece.org/display/bigdata>.

В 2020 г. Глобальная рабочая группа создала Глобальную цифровую платформу, как инструмент международного сотрудничества для доступа к общемировым источникам больших данных, обмена опытом по разработке методов и алгоритмов их преобразования для нужд официальной статистики, распространения национального опыта и учебных материалов [10].

С 2021 г. эксплуатация, обслуживание и развитие Глобальной платформы возложено на неправительственное некоммерческое учреждение «Институт Глобальной платформы» при участии четырех региональных хабов (центров). Институт Глобальной платформы обеспечивает управление всеми сервисами, эксплуатируемыми платформой, включая их совместимость и защиту информационных ресурсов.

Региональные центры Глобальной платформы созданы [11]:

- в Ханчжоу при участии Национального бюро статистики Китая и правительства провинции Чжэцзян;
- в Кигали при участии Национального института статистики Руанды и Африканского центра статистики;
- в Дубае при участии Федерального управления по вопросам конкурентоспособности и статистики Объединенных Арабских Эмиратов;
- в Рио-де-Жанейро при участии Национального статистического управления Бразилии.

Глобальной группой декларировано, что региональные центры содействуют проектам по использованию больших данных для реализации мониторинга достижения ЦУР, организуют обмен

лучшими национальными практиками и реализуют программы обучения. Информация о мероприятиях, организованных центрами, приведена на странице веб-сайта Статистического отдела секретариата ООН в рубрике «The role of the Regional Hubs in mainstreaming Big Data and data science in official statistics» (<https://unstats.un.org/bigdata/>).

К настоящему времени Глобальной группой создано несколько целевых групп, которые разрабатывают методы, готовят справочники, проводят мероприятия по наращиванию потенциала по отдельным направлениям применения в официальной статистике больших данных [11]:

- по данным наблюдения за Землей;
- по данным мобильной телефонной связи;
- по данным сканирующих устройств;
- по данным отслеживания судов (AIS — Automatic Identification System);
- по методам защиты информации от несанкционированного доступа;
- по измерению доступа к всепогодным дорогам в сельской местности;
- по учебной подготовке, повышению квалификации и развитию потенциала статистических служб;
- по использованию больших данных для достижения целей в области устойчивого развития;
- по глобальному содействию получению данных из источников в частном секторе.

Именно целевые группы и реализуют большую часть мероприятий Комитета экспертов. Они занимаются методической работой и подготовкой рекомендаций, активно участвуют в проектах, реализуемых на Глобальной платформе, и взаимодействуют с региональными хабами.

На 52-й сессии Статистической комиссии ООН в марте 2021 г. Глобальную рабочую группу преобразовали в Комитет экспертов по использованию больших данных и инструментария науки о данных для целей официальной статистики. Анализ деятельности рабочих групп и Комитета экспертов показал, что она, по нашему мнению, несмотря на широкий охват тем и направлений статистики, не носит комплексного характера. В большей степени результаты направлены на распространение национального опыта, касающегося фрагментов статистических наблюдений, в виде сборников и докладов, на организацию учебных курсов и дискуссий, а также на проведение информационно-агитационных мероприятий (круглых столов, пресс-релизов,

видеороликов). Мы не видим сколько-нибудь заметного продвижения в направлении подготовки руководств или рекомендаций по использованию больших данных в статистике, замены ими традиционных статистических наблюдений, методики по оценке качества больших данных и их приемлемости для формирования официальной статистики. Не проявляют такого рода активности и специализированные организации ООН, ответственные за разные направления статистической деятельности.

Так, в 2021 г. целевая группа по данным наблюдения за Землей завершила работу над прошедшим экспертную оценку докладом, в котором приведена информация о проектах по использованию данных наблюдения за Землей для статистики сельского хозяйства. Целевая группа по данным мобильной телефонной связи подготовила пять справочников по использованию данных мобильной телефонной связи для составления статистики перемещения населения во время бедствий, динамического картирования населения, статистики информационного общества, статистики миграции и статистики туризма. Целевая группа по использованию данных сканирующих устройств составила электронный справочник по статистике цен и подготовила учебные материалы. Целевая группа по данным автоматической системы опознавания выпустила базовый электронный учебный курс по автоматической системе опознавания, организует несколько проектов по использованию системы в отдельных регионах мира. Целевая группа по использованию больших данных для достижения целей в области устойчивого развития создала хранилище данных, где содержится информация по шести показателям достижения ЦУР. Целевая группа по методам защиты информации от несанкционированного доступа продолжает работу над справочником, где описываются методы, стандарты и принципы защиты данных. Целевая группа по учебной подготовке, повышению квалификации и развитию потенциала разработала проект комплексной учебной программы по большим данным, внедрила платформу управления обучением [12].

Комитет экспертов на базе Глобальной платформы реализует следующие проекты:

- создание конвейера данных для обработки спутниковых данных в целях составления оценочной статистики урожайности сельскохозяйственных культур;

- создание системы информационных технологий для инструмента моделирования с использованием искусственного интеллекта для измерения параметров экосистемных услуг;

- проведение на еженедельной основе оценок числа заходов судов в порты (1200 портов) в целях измерения параметров деятельности, связанной с международной торговлей;

- обработка данных автоматической системы опознавания судов для еженедельных расчетов экономических показателей;

- многократное (ежедневное и еженедельное) генерирование статистических данных по выбросам от судоходной деятельности;

- расчеты статистических данных о заходе судов в порты на основе данных автоматической идентификационной системы;

- определение загруженности портов и фактических мест назначения судов;

- внедрение технологии STAT в целях размещения на Глобальной платформе национальной базы данных Камбоджи, Кыргызстана и Мальдивских островов;

- создание конвейера данных для расчета индексов потребительских цен на данных сканирующих устройств;

- создание глобального реестра групп, охватывающего 500 крупнейших многонациональных предприятий, и обслуживание содержащей этот реестр электронной базы данных;

- систематический анализ текстовой информации, содержащейся в отчетах о рациональном использовании ресурсов с использованием методов машинного обучения;

- оценка средней доли застроенной городской территории, относящейся к открытым для всех общественным местам, с указанием доступности в разбивке по полу, возрасту и признаку инвалидности;

- расчет средней доли застроенной территории городов, которая представляет собой открытые места общественного пользования, в разбивке по полу, возрасту и статусу инвалидности потребителей;

- доступ к информации частного сектора, связанной с глобальными цепочками создания стоимости;

- создание инфраструктуры для тестирования и внедрения сетевых технологий повышения конфиденциальности [11–13].

В проекте решения Статистической комиссии ООН (будет рассматриваться в марте 2023 г.) определены направления, где должны концентри-

роваться усилия международного статистического сообщества по эффективному использованию больших данных в официальной статистике. Выделены следующие [13]:

- создание глобальной сети ведущих специалистов национальных статистических служб и региональных центров и инструкторов, занимающихся вопросами больших данных, внедрение международной программы наставничества;
- расширение взаимодействия с сообществом специалистов по геопространственным данным;
- взаимодействие с частным сектором по вопросам, касающимся доступа к большим данным и использования инструментов с открытым исходным кодом;
- создание центра по использованию искусственного интеллекта для анализа данных об окружающей среде и устойчивом развитии (Artificial Intelligence for Environment & Sustainability — ARIES) с ориентацией на обеспечение совместимости различных источников информации для целей внедрения Системы природно-экономического учета.

Вопрос о больших данных в официальной статистике обсуждается и на 54-й сессии Статистической комиссии ООН (февраль—март 2023 г.). В докладе Комитета экспертов по использованию больших данных для целей официальной статистики содержится отчет о тех мероприятиях, которые были проведены в последние годы, описаны достижения целевых групп и их планы. Так, в частности, отмечено, что целевая группа по использованию данных систем мобильной связи для нужд официальной статистики, возглавляемая Международным союзом электросвязи, выпустила несколько методических руководств по применению информации систем мобильной связи для подготовки статистики перемещений населения и стихийных бедствий, оценок плотности населения, измерения информационного общества, статистики миграции и статистики туризма (<https://unstats.un.org/bigdata/task-teams/mobile-phone/>). В 2021 г. целевая группа по данным наблюдения за Землей выпустила материалы, содержащие рекомендации по обработке данных наблюдения за Землей в сельском хозяйстве (<https://unstats.un.org/unsd/statcom/53rd-session/documents/BG-3s-3uAgricultureAndBigData-E.pdf>). Целевая группа по использованию данных автоматической идентификационной системы публикует постоянно обновляемое руководство по ее использованию в статистике морско-

го транспорта. Целевая группа по данным сканирующих устройств работает над разработкой методических указаний по использованию данных о транзакциях, полученных с применением веб-скрейпинга, в статистике потребительских цен. Заявлено, что в 2023 г. целевая группа по методам сохранения конфиденциальности выпустит новое руководство, в котором будут представлены инновационные технологий, направленные на защиту конфиденциальности, представлены стандарты и описаны основные нормативно-правовые аспекты внедрения этих технологий [13].

Мировое статистическое сообщество продолжает совершенствовать структуру управления своей деятельностью в области использования больших данных в официальной статистике. Так, в 2023 г. Статистическая комиссия утвердила создание Организации сети лидеров в области науки о данных (The Data Science Leaders Network). Сеть должна объединить «лидеров в области науки о данных для проведения стратегических дискуссий по вопросам больших данных и науки о данных, обмена опытом и знаниями, укрепления процесса принятия решений и руководства в национальных статистических службах, а также для активизации научного сотрудничества и технических партнерств» [13]. Принятое решение объясняется необходимостью координации, выработки рекомендаций, определения приоритетов и направлений работы в области науки о данных для нужд официальной статистики. Статистическая комиссия в качестве таких лидеров видит руководителей национальных статистических служб, руководителей региональных и отраслевых центров больших данных. Сфера компетенции и обязанности Сети лидеров в области науки о данных будет в первую очередь сосредоточена на роли больших данных и науки о данных в модернизации глобальных и национальных статистических систем. Это направление деятельности будет охватывать весь спектр новых источников данных, методов, технических средств и технологий, которые статистические службы могут использовать для повышения качества и экономической эффективности своих производственных процессов. Из текста доклада Комитета экспертов, выступившего в 2022 г. с этой инициативой, пока не ясны конкретные направления работы Сети лидеров [12]. Важно, что целевые группы, работающие в предметных областях применения больших данных в официальной статистике, свою деятельность продолжают.

Практика использования больших данных в официальной статистике. Достижения и проблемы

Содержательный анализ состояния использования для официальной статистики больших данных по источникам их образования приведен в докладе Группы друзей председателя Статистической комиссии ООН по Основополагающим принципам официальной статистики, представленном статистическим отделом секретариата ООН в 2019 г. на 50-й ее сессии. Этот доклад «*Supplementing the United Nations Fundamental Principles of Official Statistics: Implementation Guidelines*» содержал предложения по дополнению Основополагающих принципов официальной статистики в отношении нетрадиционных источников данных для официальной статистики [14]. Предложения были сформулированы на основе анализа имеющихся сведений о национальном опыте включения в производство официальной статистики больших данных. Данный документ получил развитие в докладе «*Fundamental Principles of Official Statistics: Mapping and Guidance for the United Nations Fundamental Principles of Official Statistics against non-conventional and non-traditional data sources*», представленном на 51-й сессии Статистической комиссии ООН в марте 2020 г. [15].

В качестве нетрадиционных и необычных (non-traditional and non-conventional) источников больших данных, которые могут быть полезны для формирования показателей официальной статистики, в этих документах рассматривались следующие:

- данные мобильных телефонов Системы операторов мобильной связи;
- данные с интеллектуальных счетчиков потребления электроэнергии;
- данные спутниковых снимков;
- данные из социальных сетей;
- данные, извлеченные из Интернета;
- данные датчиков дорожного движения и данные датчиков слежения за пассажирами;
- данные сканеров и кассовых аппаратов (пример для супермаркетов);
- данные видеонаблюдения (видеозаписи без опасности / наблюдения).

Здесь же были приведены многочисленные примеры привлечения больших данных в качестве источника сведений для официальных статисти-

ческих оценок [14, 15]. В части использования данных мобильного телефона был описан опыт Новой Зеландии для понимания («to understand short-term population movement») краткосрочных перемещений населения после стихийных бедствий. Другие страны (Соединенное Королевство и Франция) оценивают движение населения между центральным городом и пригородами. Но каких-либо фундаментальных примеров замены данных о населении, полученных при переписи населения или из регистра не приведено.

Мы согласны с мнением, высказанным О.С. Чудиновских о том, что большие данные для измерения миграции могут использоваться только при оценках «...краткосрочной мобильности населения и сдвигов в его размещении в определенные моменты или периоды времени. В больших данных нет возможности применить критерии учета мигрантов и миграции, которые используются в официальной статистике, в первую очередь концепции обычного места жительства. Сделан вывод о том, что большие данные пока не могут быть альтернативой традиционным источникам информации для разработки надежной и понятной статистики миграции» [16]. Эта позиция основана на детальном анализе информации, которая создается разнообразными источниками больших данных, фиксирующих территориальные перемещения людей.

В качестве примера использования данных интеллектуальных счетчиков в официальной статистике приведен опыт Эстонии, где на их основе рассчитывается потребление электроэнергии населением в привязке к данным из регистров о характеристиках домашних хозяйств, числе жильцов, размере жилья и количестве комнат. Обследование дало информацию о пустующих помещениях. Другой пример — Соединенное Королевство, где дистанционно проверялись адреса жилых помещений и также выявлялось свободное жилье.

В докладе отмечается, что применение данных спутниковых снимков в официальной статистике довольно широко распространено. Эта информация используется при планировании переписей в привязке к территории, а также при составлении отчетности об использовании земли. В частности, Австралийское бюро статистики составило по этим данным в сочетании с другой информацией экспериментальный земельный счет, основываясь на методологии Системы при-

родно-экономического учета, включающий оценки стоимости земли. Статистическое управление Канады разработало земельные счета для отслеживания изменений почвенного покрова. В ЕС развивается сельскохозяйственная статистика с использованием спутниковых снимков Европейского космического агентства в направлении оценки воздействия экстремальных природных явлений (наводнение, засуха, заморозки и пр.) на валовые сборы сельскохозяйственных культур.

Использование данных социальных сетей в официальной статистике осуществляется Статистическим управлением Нидерландов для расчета альтернативного показателя потребительского доверия (*alternative consumer confidence indicator*). Институт географии и статистики Мексики использует Twitter для оценивания настроений по различным аспектам человеческой жизни (туризм, психическое здоровье, мобильность населения, доверие потребителей, безопасность).

Использование данных, собранных в Интернете, в официальной статистике показано на примере Статистического управления Польши, которое в этой работе ориентировано на применение больших данных для измерения индекса потребительских цен. В качестве источника необходимой информации служат данные из информационных систем торговых сетей (*scanned data*), данные, полученные от частных компаний, данные о розничных ценах из Интернета, в том числе с использованием *web scraping techniques*. Еще одним примером здесь являются попытки сбора данных о рынке труда в режиме реального времени с порталов работодателей, размещающих информацию о вакантных рабочих местах. Здесь сразу необходимо провести раздел между официальными статистическими оценками и использованием больших данных для принятия решений на рынке труда. Так, например, пособие, опубликованное Европейским Фондом образования в 2019 г., где описаны национальные практики по применению больших данных для анализа, не содержит рекомендации по замене большими данными официальной статистической информации о занятости, безработице, вакансиях, стоимости рабочей силы, об условиях труда и т. п. Не подвергается сомнению необходимость в сборе данных о рынке труда на базе выборочных обследований рабочей силы и отчетности работодателей. Авторы пособия признают, что большие данные — это основа оперативных оценок и быстрых решений, а офи-

циальная статистика ориентирована на стратегические решения, базирующиеся на системном анализе и прогнозе, охватывающих все сегменты рынка труда, которые не были замечены источниками больших данных [5].

В докладе также приведен опыт Соединенного Королевства, Австралии, Новой Зеландии и Португалии, где статистики используют Интернет для сбора данных о потребительских ценах для измерения инфляции, разрабатывают методики, проводят экспериментальные расчеты. Наиболее далеко продвинулось Австралийское бюро статистики, которое, в частности, внедрило данные, полученные из Интернета, в расчет индекса потребительских цен, заменив ими данные, собранные на местах [17].

С одной стороны, идея получения данных о ценах на товары и услуги из Интернета представляется понятной, и ее практическая реализация должна принести пользу, так как экономит ресурсы на сбор информации, обеспечивает единство качества товаров-представителей. Но может ли статистик быть уверенным, что выбранный таким образом для регистрации цен товар занимает основную или заметную долю рынка? Для развитых стран можно исходить из гипотезы, что динамика цен на один и тот же товар в сетевых магазинах (именно здесь достигнут наибольший успех) соответствует динамике цен на него и в других сегментах рынка. Но для многих стран это будет весьма большим допущением из-за присутствия на рынке разнообразных форм организации торгового бизнеса, наличия большого числа мелких торговцев, неформальной экономики. Еще более проблематично принять такую гипотезу для данных об уровнях цен. А ведь информация об уровнях цен есть основа расчета стоимости различных типов потребительских корзин, включая минимальную, признаваемую во многих странах порогом бедности, и также является основой расчета территориальных индексов стоимости жизни.

В докладе приведен опыт Статистического управления Польши, которое использует в официальной статистике данные о дорожном движении (*road traffic data*) — информации, считываемой с датчиков Автоматической системы идентификации судов (AIS) и Электронной системы взимания платы за проезд (аналог отечественных транспондеров для платных дорог). Это, сказано в докладе, позволяет получить новую информацию (*new information*) об объеме передвижения,

работе транспорта, объеме выброса транспортными средствами загрязняющих веществ с целью обеспечения оперативного (в режиме реального времени) мониторинга транспортной политики. Но очевидно, что статистику пассажиро- и грузооборота такие данные не заменят.

Использование данных сканера (scanner data) в официальной статистике показано в докладе на примере опыта статистиков Польши, которые разрабатывают методы использования этих сведений для расчета индекса потребительских цен, опыта статистиков Новой Зеландии, включившей в этот расчет информацию о ценах на бытовую электронику (consumer electronics), считанную со сканеров, для оценки инфляции. В Нидерландах полностью отказались от традиционной регистрации цен во всех супермаркетах и перешли на данные сканеров. Во Франции ведутся аналогичные работы. В Австралии статистики довольно далеко продвинулись по использованию сведений, полученных от сканеров в розничной торговле для измерения инфляции, в частности, в наибольшей степени при расчетах инфляции данные сканеров используются по продуктам питания, табаку и бытовым товарам длительного пользования. В Австралии данные сканеров используются и в других областях: для расчета показателей розничной торговли, при составлении национальных счетов и проведении обследований здоровья.

В качестве примера использования данных камер видеонаблюдения (CCTV data / Security videos) приведен опыт полиции Бермудских островов, которая на основе полученных видеозаписей, имеет возможность более полной регистрации статистики преступности. Но это не отменяет необходимости проведения статистических обследований процесса превращения лица в жертву преступного посягательства на основе опроса жертв преступлений.

Никаких примеров успешности проектов в этой области не приведено в докладе «Региональные и национальные усилия по использованию больших данных в интересах официальной статистики в Азиатско-Тихоокеанском регионе», рассмотренном на седьмой сессии комитета по статистике ЭСКАТО ООН в 2020 г. [18]. Здесь дана информация о встречах экспертов и учебных семинарах по проблемам больших данных. Кроме этого, упомянуто, в частности, то, что правительство Бангладеш и правительство Непала обозначили оценки масштабов нищеты в каче-

стве одного из своих приоритетов и планируют использование в целях мониторинга спутниковых данных о ночном освещении. А правительство Монголии заинтересовано в получении оценок урожайности на спутниковых данных наблюдения за поверхностью Земли.

Пример применения больших данных. Статистика туризма

Одним из перспективных направлений использования больших данных в целях официальной статистики, о котором много пишут, является статистика туризма. Здесь реализуются многочисленные проекты в разных странах. Считается, что именно по этому направлению большие данные могут принести заметный эффект и сократить затраты на традиционный сбор данных. Причиной является то, что процессы, связанные с производством и потреблением туристского продукта, довольно латентны, с одной стороны. С другой стороны, имеется значительное разнообразие источников больших данных, содержащих сведения о туристской мобильности населения и его активности на рынках туристских товаров и услуг.

Статистики пытаются привлечь к расчетам показателей туризма такие источники больших данных, как транспортные компании и продавцы пассажирских билетов, агрегаторы средств размещения, банки, операторы камер и датчиков слежения за дорожной сетью, операторы мобильной связи. Кроме того, с позиций получения доступа к данным о путешественниках рассматривается возможность использования в этих целях веб-сайтов и социальных сетей.

Важной особенностью больших данных для применения именно в статистике туризма является возможность наряду с оперативностью и получать показатели в привязке к небольшим географическим областям. Это объясняется тем, что в официальной статистике часто показатели дают оценки явления в соответствии с административным устройством страны, а географические ареалы, привлекающие туристов, не связаны с формальными границами провинций и муниципалитетов.

Значительный опыт использования больших данных в статистике туризма связан с операторами мобильной связи и агрегаторами транспортных и гостиничных услуг [19]. Данные мо-

бильных операторов предоставляют возможности получения оценок различных статистических показателей; в частности, они позволяют оценить численность населения в привязке к месту проживания на момент времени, а также его географическую мобильность. Это связано с тем, что данные мобильных операторов предоставляют возможность оценивания передвижения человека (точнее телефона) в привязке к временным и пространственным параметрам, а это, в свою очередь, позволяет фиксировать продолжительность его пребывания в определенных местах. Последнее и является основанием для оценки частоты и продолжительности перемещений. При оценках исходят из гипотезы о том, что нерегулярные, редкие и относительно дальние перемещения за пределы обычной зоны проживания признаются туристскими. Но для использования сведений мобильных операторов для выявления и, соответственно, для включения в оценки туристского потока однодневных туристских поездок нет оснований, так как такие визиты могут быть не связаны с туризмом.

При измерении туристского потока на основании данных операторов мобильной телефонной сети следует помнить и о том, что первичная (исходная) информация привязана к характеристикам зарегистрированного в договоре физического лица, а это устройство может использоваться другим лицом и двумя и более лицами в целях экономии расходов, может и не использоваться вообще по тем же причинам. Поэтому признать какие-либо характеристики населения, использующего мобильные телефоны вне зоны обычного проживания, данными, относящимися ко всему изучаемому населению, будет очевидной ошибкой. Есть и иные проблемы, ограничивающие использование такой информации, например, отказ от использования мобильных телефонов в международном роуминге, использование альтернативных систем связи [20]. Таким образом, видна условность результатов замеров численности населения и размеров (не говоря о характеристиках) туристских потоков.

Для такой информации должен быть предложен иной показатель, характеризующий туристский поток. Такой информацией нельзя полностью заменить «медленные» данные официальной статистики, которые обеспечивают полноту охвата изучаемого явления и точное описание объекта и единиц статистического наблюдения. Хотя

конечно, для анализа и выработки мер политики, особенно в краткосрочной перспективе, таких данных может быть вполне достаточно, так как они позволяют оценить приблизительно количество поездок, продолжительность пребывания, страну происхождения, регион постоянного проживания, маршрут.

Использование в целях официальной статистики сбора данных из веб-страниц подразумевает наличие перечня веб-сайтов для сбора данных, из которых их планируется извлекать. Процедура извлечения данных имитирует запрос пользователя сайта и преобразует полученные неструктурированные данные в структурированный формат. В настоящее время имеется определенный опыт получения информации о размещении туристов в гостиницах на основе такой технологии [21]. Информация, которая может быть использована в официальной статистике туризма, охватывает данные об общем количестве мест размещения, их заполнении и стоимости размещения. Необходимо сразу оговориться, что такие данные относятся только к тем фирмам, которые размещают соответствующую информацию в сети Интернет. И здесь большие данные явились бы хорошим, с точки зрения оперативности, заменителем официальной статистики, но не охватывают всего объема предложения и удовлетворенного спроса на рынке услуг по размещению путешественников.

Данные об операциях клиентов банков и других кредитных организаций, проведенные с использованием пластиковых карт, также являются одним из перспективных направлений применения больших данных в официальной статистике вообще, в частности, для измерения туристского потока, его направленности, объема вовлеченных денежных средств и пр. Важно, что такая информация позволяет стратифицировать туристов по демографическим характеристикам, стране (региону) постоянного проживания, а также по характеру и месту совершения операций. Но не стоит забывать про одно ограничение — банки не фиксируют того путешественника, который не пользовался их услугами. Поэтому нельзя сказать, что эти данные могут оценить весь туристский поток и совершаемые туристами денежные операции в полном объеме.

Для измерения туристского потока с детализацией путешественников по различным характеристикам может быть привлечена информация стра-

ховых компаний. Здесь, правда, надо отметить, что обязательное страхование путешественников реализуется не во всех странах и распространяется, как правило, на международный туризм. При условии выполнения требования по страхованию международного туриста можно оценить полноту охвата числа выезжающих лиц и, соответственно, быть уверенным в адекватности описания их характеристик.

Довольно много усилий тратится на использование в качестве источника данных о туристских потоках информации, аккумулируемой датчиками камер слежения за дорожной сетью. В аналитических целях эта информация бесспорно обладает большой ценностью, но в официальной статистике для получения показателей по туризму успехи, судя по материалам Статистической комиссии ООН, весьма скромные. Главная проблема состоит в том, что благодаря этим устройствам фиксируется не число лиц, передвигающихся по дорогам, а число транспортных средств. Переход от этих данных к оценкам путешественников пока остается нерешенной задачей. Но нагрузка на дорожную сеть во времени и в пространстве — важная информация. Она представляет данные о величине транспортного потока, в составе которого выделяется транспорт, зарегистрированный в разных регионах страны и разных странах. Определяются и типы транспортных средств — пассажирские, грузовые автомобили, мотоциклы или автобусы, выполняющие междугородные или международные перевозки. Но кто перемещается и в каких целях при этом нельзя выяснить.

Ряд с статистических служб пытается использовать большие данные для получения показателей статистики туризма, но примеров их применения как главного источника первичной информации, лежащей в основе официальных статистических наблюдений, мы не обнаружили. Дело ограничивается фрагментарным их использованием для решения частных задач не на регулярной основе. Гораздо больше успехов в применении в этих целях административных данных, например, пограничной службы, таможенных органов, учреждений санитарного контроля и т. п.

Действительно, про использование больших данных для измерения и анализа туристских потоков пишут довольно много. И это можно понять, поскольку рынок этих услуг довольно динамичен, а фактор оперативности оценок здесь играет

важную, если не первостепенную роль, так как туроператоры хорошо знают из прошлого опыта туристский спрос и предложение по стоимости и товарной номенклатуре. А вот даже общие сведения о направлениях потоков и их масштабах (хотя бы в оценках «больше/меньше») бывают весьма полезны.

Поэтому большие данные являются полезным дополнением к традиционным источникам информации. Но есть и трудности в их использовании, которые связаны с конкуренцией на этом рынке, подразумевающей нежелание раскрывать коммерческие базы данных, возможные смещения в оценках из-за исключения из наблюдения отдельных сегментов рынка, необходимость инвестирования в инструменты поиска нужных данных и их обработки под нужды аналитики. Также имеет смысл отметить, что вряд ли можно серьезно обсуждать замену специфических туристских обследований большими данными, или способность последних дать полную информацию для составления таблиц вспомогательного счета туризма. Здесь стоит посмотреть на опыт стран ЕС [22].

Заключение

Вопросы экономии ресурсов для официальной статистики всегда стояли на повестке дня, а с расширением цифровизации государственного управления и революции данных, проявляющейся в феномене больших данных и возросших возможностей ИКТ, стали еще более острыми. В этой связи официальная статистика должна прилагать максимальные усилия для широкого использования административных и больших данных в интересах расширения мониторинга демографических, социальных, экономических и экологических процессов.

Приведенная выше информация свидетельствует о высокой активности международного статистического сообщества по включению больших данных как нового источника информации об обществе и экономике в официальную статистику. Сегодня мы не видим больших достижений в этом вопросе, усилия в основном направлены на модернизацию фрагментов отдельных статистических наблюдений. Но тем не менее движение вперед есть. И это движение идет широким фронтом, если судить по тематике деятельности целевых групп и проектов, реализуемых на базе

Глобальной платформы. Например, последние новации коснулись разработки подходов к оценке глобальных цепочек добавленной стоимости, одного из самых сложных направлений в макроэкономической статистике и в статистике производства.

И здесь следует с сожалением отметить низкую международную активность российской статистики и статистических служб стран СНГ, за редким исключением. Хотя, если посмотреть на участников проектов и целевых групп, то обращает на себя внимание представленность статистиков из стран, где национальные статистические системы не отличаются высоким потенциалом. Необходимо активизировать усилия статистиков, бизнеса и академических кругов России и партнеров из СНГ в данном процессе, включая участие в Сети лидеров в области науки о данных. Имеет смысл проработать вопросы создания в регионе СНГ регионального хаба по наращиванию потенциала применения больших данных в официальной статистике, который стал бы площадкой для обсуждения международных инициатив, лучших зарубежных практик и опыта стран региона по данной тематике.

Литература

1. Сухобоков А.А., Лахвич Д.С. Влияние инструментария Big Data на развитие научных дисциплин, связанных с моделированием // Наука и Образование: научное издание МГТУ им. Н.Э. Баумана. Сетевое издание. 2015. № 3. С. 207–240. URL: <http://engineering-science.ru/doc/761354.html> (дата обращения 28.08.2022).
2. Thelin R. What is Big Data? Characteristics, Types, and Technologies. Dec. 15, 2020. URL: <https://www.educative.io/blog/what-is-big-data> (дата обращения 28.11.2022).
3. Furht B., Villanustre F. Introduction to Big Data // Big Data Technologies and Applications. Springer, 2016. P. 3–11. doi: https://doi.org/10.1007/978-3-319-44550-2_1 (дата обращения 28.12.2022).
4. Rai A. What is Big Data – Characteristics, Types, Benefits & Examples. August 16, 2022. URL: <https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/> (дата обращения 23.11.2022).
5. Большие данные в контексте рынка труда. Ознакомительное пособие / сост.: М. Мещанцаника и Ф. Меркорио. Европейский фонд образования, 2019. URL: https://www.etf.europa.eu/sites/default/files/2020-11/guide_big_data_lmi_ru_0.pdf (дата обращения 21.07.2022).
6. Sicular S. Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s // Forbes. March 27, 2013. URL: <https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/?sh=5e22c3b042f6> (дата обращения 12.11.2022).
7. Redefining 'Urban'. A New Way to Measure Metropolitan Areas. Paris: OECD Publ., 2012. doi: <https://doi.org/10.1787/9789264174108-en> (дата обращения 10.07.2022).
8. Terms of Reference and Mandate of the Global Working Group on Big Data for Official Statistics // Report of the Global Working Group on Big Data for Official Statistics: note / by the Secretary-General. Statistical Commission Forty-Sixth Session, 3–6 March 2015. E/CN.3/2015/4. URL: <https://unstats.un.org/bigdata/documents/TOR%20-%20GWG%20-%202015.pdf> (дата обращения 04.09.2022).
9. UN Statistical Commission. Report of the 2015 Big Data Survey. Statistical Commission Forty-Seventh Session, 8–11 March 2016. URL: <https://unstats.un.org/unsd/statcom/47th-session/documents/BG-2016-6-Report-of-the-2015-Big-Data-Survey-E.pdf> (дата обращения 08.09.2022).
10. UN Statistical Commission. Report of the Global Working Group on Big Data for Official Statistics: note / by the Secretary-General. Statistical Commission Fifty-First Session, 3–6 March 2020. E/CN.3/2020/24. URL: <https://documents-dds-ny.un.org/doc/UNDOC/GEN/N19/412/39/PDF/N1941239.pdf?OpenElement> (дата обращения 04.09.2022).
11. Статистическая комиссия ООН. Доклад Глобальной рабочей группы по использованию больших данных для целей официальной статистики: записка Генерального секретаря. Пятьдесят вторая сессия, 1–3 и 5 марта 2021 г. E/CN.3/2021/14. URL: <https://daccess-ods.un.org/tmp/8787885.90431213.html> (дата обращения 02.09.2022).
12. Статистическая комиссия ООН. Доклад Комитета экспертов по использованию больших данных и обработке и анализу данных для целей официальной статистики: записка Генерального секретаря. Пятьдесят третья сессия, 1–4 марта 2022 г. E/CN.3/2022/25. URL: <https://unstats.un.org/unsd/statcom/53rd-session/documents/2022-25-BigData-R.pdf> (дата обращения 04.09.2022).
13. Статистическая комиссия ООН. Доклад Комитета экспертов по использованию больших данных и обработке и анализу данных для целей официальной статистики: записка Генерального секретаря. Пятьдесят четвертая сессия, 28 февраля – 3 марта 2023 г. E/CN.3/2023/17. URL: https://unstats.un.org/UNSDWebsite/statcom/session_54/documents/2023-17-BigData-R.pdf (дата обращения 04.09.2022).
14. UN Statistical Commission. Supplementing the United Nations Fundamental Principles of Official Statistics: Implementation Guidelines: note / by the Secretary-General. Statistical Commission Fiftieth Session, 5–8 March 2019.
15. UN Statistical Commission. Supplementing the United Nations Fundamental Principles of Official Statistics

tics: Mapping and Guidance for the United Nations Fundamental Principles of Official Statistics Against Non-Conventional and Non-Traditional Data Sources. Statistical Commission, Fifty-First Session, 3–6 March 2020. URL: https://unstats.un.org/unsd/statcom/51st-session/documents/BG-Item3q-Supplementing_non-traditional-E.pdf.

16. **Чудиновских О.С.** Большие данные и статистика миграции // Вопросы статистики. 2018. Т. 25. № 2. URL: <https://voprstat.elpub.ru/jour/article/view/629/545> (дата обращения 04.07.2022).

17. Australian Bureau of Statistics (ABS). Information Paper: Making Greater Use of Transactions Data to Compile the Consumer Price Index. 29.11.2016. URL: <https://www.abs.gov.au/ausstats/abs@.nsf/mf/6401.0.60.003>.

18. ESCAP. Regional and National Efforts to Embrace Big Data for Official Statistics in Asia and the Pacific Economic and Social Commission for Asia and the Pacific. Note by the Secretariat. Committee on Statistics Seventh Session Bangkok, 26–28 August 2020. ESCAP/CST/2020/4. URL: https://www.unescap.org/sites/default/files/ESCAP.CST_2020.4_Regional_and_national_efforts_to_embrace_big_data_for_official_statistics.pdf (дата обращения 04.07.2022).

19. **Kim J.K., Wang Z.** Sampling Techniques for Big Data Analysis // International Statistical Review. 2019. Vol. 87. No. S1. P. S177–S191. doi: <https://doi.org/10.1111/insr.12290> (дата обращения 04.01.2023).

20. **Grassini L., Dugheri G.** Mobile Phone Data and Tourism Statistics: A Broken Promise? // National Accounting Review. 2021. Vol. 3(1). P. 50–68. doi: <https://doi.org/10.3934/NAR.2021002> (дата обращения 04.09.2022).

21. **Saluveer E.** et al. Methodological Framework for Producing National Tourism Statistics from Mobile Positioning Data // Annals of Tourism Research. 2020;81. doi: <https://doi.org/10.1016/j.annals.2020.102895> (дата обращения 14.10.2022).

22. Tourism Satellite Accounts in Europe. 2019 Edition. Luxembourg: Publications Office of the European Union, 2019. URL: <https://ec.europa.eu/eurostat/documents/7870049/10293066/KS-FT-19-007-EN-N.pdf/f9cdc4cc-882b-5e29-03b1-f2cee82ec59d> (дата обращения 17.07.2022).

Информация об авторе

Суринов Александр Евгеньевич — д-р экон. наук, профессор, руководитель Департамента статистики и анализа данных, директор Центра экономических измерений и статистики, заведующий Научно-учебной лабораторией измерения благосостояния факультета экономических наук, Национальный исследовательский университет «Высшая школа экономики». 101000, г. Москва, ул. Мясницкая, д. 20. E-mail: surinov@hse.ru. ORCID: <https://orcid.org/0000-0002-0294-2881>.

Финансирование

Статья подготовлена по материалам исследования, осуществленного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2022 г.

References

1. **Suhobokov A.A., Lakhvich D.S.** The Big Data Tools Impact on Development of Simulation-Concerned Academic Disciplines. *Science and Education of Bauman MSTU*. 2015;(3):207–240. (In Russ.) Available from: <http://engineering-science.ru/doc/761354.html> (accessed 28.08.2022).

2. **Thelin R.** *What is Big Data? Characteristics, Types, and Technologies*. Available from: https://doi.org/10.1007/978-3-319-44550-2_1 (accessed 28.11.2022).

3. **Furht B., Villanustre F.** Introduction to Big Data. In: *Big Data Technologies and Applications*. Springer; 2016. Pp. 3–11. Available from: https://doi.org/10.1007/978-3-319-44550-2_1 (accessed 28.12.2022).

4. **Rai A.** *What is Big Data – Characteristics, Types, Benefits & Examples*. Available from: <https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/> (accessed 23.11.2022).

5. **Mezzanzanica M., Mercorio F.** *Big Data for Labour Market Intelligence: An Introductory Guide*. European Training Foundation; 2019. Available from: <https://www.etf.europa.eu/sites/default/files/2019-06/Big%20data%20for%20LMI.pdf> (accessed 21.07.2022).

6. **Sicular S.** Gartner's Big Data Definition Consists of Three Parts, Not to Be Confused with Three "V"s. *Forbes*. 27.03.2013. Available from: <https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/?sh=6c704f642f68> (accessed 12.11.2022).

7. *Redefining "Urban". A New Way to Measure Metropolitan Areas*. Paris: OECD Publishing; 2012. Available from: <https://doi.org/10.1787/9789264174108-en> (accessed 10.07.2022).

8. Terms of Reference and Mandate of the Global Working Group on Big Data for Official Statistics. In: *Report of the Global Working Group on Big Data for Official Statistics. Note by the Secretary-General. Statistical Commission Forty-Sixth Session, 3–6 March 2015. E/CN.3/2015/4*. Available from: <https://unstats.un.org/bigdata/documents/TOR%20-%20GWG%20-%202015.pdf> (accessed 04.09.2022).

9. UN Statistical Commission. *Report of the 2015 Big Data Survey. Note by the Secretary-General. E/CN.3/2020/24. Statistical Commission Forty-Seventh Session, 8–11 March 2020*. Available from: <https://unstats.un.org/unsd/statcom/47th-session-2020/documents/2020-03-11-report-of-the-2015-big-data-survey-note-by-the-secretary-general.pdf> (accessed 04.09.2022).

10. **Saluveer E.** et al. Methodological Framework for Producing National Tourism Statistics from Mobile Positioning Data // *Annals of Tourism Research*. 2020;81. doi: <https://doi.org/10.1016/j.annals.2020.102895> (дата обращения 14.10.2022).

2016. Available from: <https://unstats.un.org/unsd/statcom/47th-session/documents/BG-2016-6-Report-of-the-2015-Big-Data-Survey-E.pdf> (accessed 08.09.2022).

10. UN Statistical Commission. *Report of the Global Working Group on Big Data for Official Statistics. Note by the Secretary-General. Statistical Commission Fifty-First Session, 3–6 March 2020. E/CN.3/2020/24*. Available from: <https://documents-dds-ny.un.org/doc/UNDOC/GEN/N19/412/39/PDF/N1941239.pdf?OpenElement> (accessed 04.09.2022).

11. UN Statistical Commission. *Report of the Global Working Group on Big Data for Official Statistics. Note by the Secretary-General. United Nations Statistical Commission Fifty-Second Session, 1–3 and 5 March 2021. E/CN.3/2021/14*. Available from: <https://unstats.un.org/bigdata/documents/reports/GWG%20report%20-%202021-14-BigData-E.pdf> (accessed 02.09.2022).

12. United Nations Statistical Commission. *Report of the Committee of Experts on Big Data and Data Science for Official Statistics. Note by the Secretary-General. Statistical Commission Fifty-Third Session, 1–4 March 2022. E/CN.3/2022/25*. Available from: <https://unstats.un.org/bigdata/documents/reports/UNCEBD%20report%20-%202022-25-BigData-E.pdf> (accessed 04.09.2022).

13. UN Statistical Commission. *Report of the Committee of Experts on Big Data and Data Science for Official Statistics. Note by the Secretary-General. Statistical Commission Fifty-Fourth Session 28 February – 3 March 2023. E/CN.3/2023/17*. (In Russ.) Available from: https://unstats.un.org/UNSDWebsite/statcom/session_54/documents/2023-17-BigData-R.pdf (accessed 04.09.2022).

14. UN Statistical Commission. *Supplementing the United Nations Fundamental Principles of Official Statistics: Implementation Guidelines. Note by the Secretary-General. Statistical Commission, Fiftieth Session, 5–8 March 2019*.

15. UN Statistical Commission. *Supplementing the United Nations Fundamental Principles of Official Statistics: Mapping and Guidance for the United Nations Fundamental Principles of Official Statistics Against Non-Conventional and Non-Traditional Data Sources. Statistical Commission,*

Fifty-First Session, 3–6 March 2020. Available from: https://unstats.un.org/unsd/statcom/51st-session/documents/BG-Item3q-Supplementing_non-traditional-E.pdf.

16. **Chudinovskikh O.S.** Big Data and Statistics on Migration. *Voprosy Statistiki*. 2018;25(2):48–56. (In Russ.) Available from: <https://voprstat.elpub.ru/jour/article/view/629/545> (accessed 04.07.2022).

17. Australian Bureau of Statistics (ABS). *Information Paper: Making Greater Use of Transactions Data to compile the Consumer Price Index*. Australia; 2016. Available from: <https://www.abs.gov.au/ausstats/abs@.nsf/mf/6401.0.60.003>.

18. Economic and Social Commission for Asia and the Pacific. *Regional and National Efforts to Embrace Big Data for Official Statistics in Asia and the Pacific Economic and Social Commission for Asia and the Pacific. Note by the Secretariat. Committee on Statistics Seventh Session Bangkok, 26–28 August 2020. ESCAP/CST/2020/4*. Available from: https://www.unescap.org/sites/default/files/ESCAP.CST_.2020.4_Regional_and_national_efforts_to_embrace_big_data_for_official_statistics.pdf.

19. **Kim J.K., Wang Z.** Sampling Techniques for Big Data Analysis. *International Statistical Review*. 2019;87(S1):S177–S191. Available from: <https://doi.org/10.1111/insr.12290> (accessed 04.01.2023).

20. **Grassini L., Dugheri G.** Mobile Phone Data and Tourism Statistics: A Broken Promise? *National Accounting Review*. 2021;3(1):50–68. Available from: <https://doi.org/10.3934/NAR.2021002> (accessed 04.09.2022).

21. **Saluveer E.** et al. Methodological Framework for Producing National Tourism Statistics from Mobile Positioning Data. *Annals of Tourism Research*. 2020;81. Available from: <https://doi.org/10.1016/j.annals.2020.102895> (accessed 14.10.2022).

22. *Tourism Satellite Accounts in Europe. 2019 Edition*. Luxembourg: Publications Office of the European Union; 2019. Available from: <https://ec.europa.eu/eurostat/documents/7870049/10293066/KS-FT-19-007-EN-N.pdf/f9cdc4cc-882b-5e29-03b1-f2cee82ec59d> (accessed 17.07.2022).

About the author

Alexandr Ye. Surinov – Dr. Sci. (Econ.), Professor, Department Head, Department of Statistics and Data Analysis, Director, Economic Statistics Centre of Excellence, Laboratory Head, Laboratory for Wealth Measurement, Faculty of Economic Sciences, National Research University Higher School of Economics (HSE University). 20, Myasnitskaya Str., Moscow, 101000, Russia. E-mail: surinov@hse.ru. ORCID: <https://orcid.org/0000-0002-0294-2881>.

Funding

This work is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University) in 2022.