

Большие данные и официальная статистика

Шьям Упадхьяя

Организация Объединенных Наций по промышленному развитию (ЮНИДО), г. Вена

Большие данные – одна из составляющих четвертой промышленной революции. Глубокое внедрение цифровой технологии в экономику способствовало тому, что информация стала неотъемлемым элементом производственного процесса. Большие данные создаются в процессе работы машины, взаимодействия человека с машиной и взаимодействия между людьми.

В статье последовательно рассматриваются вопросы, вытекающие из ее названия. Прежде всего, раскрывается содержание понятия «большие данные», отмечается, что это не только цифры в традиционном понимании, но и текстовая часть, аудио- и видеозаписи в социальных сетях, фотографии, спутниковые изображения, электронные письма, программы, приложения и многое другое. Автор проводит различие между неструктурированными и структурированными данными, отмечая, что последние – это в основном количественные данные, которые представлены в базе данных с заранее определенной моделью для их хранения, обработки и распространения.

Анализируя такой новый и, безусловно, революционный источник информации, каковым являются большие данные, автор оценивает их с точки зрения соответствия основным критериям и базовым принципам качества данных, таким как достоверность, возможность обеспечения сопоставимости, точность и надежность, правильное использование методологии.

Отмечается, что тема больших данных вызывает чрезвычайный интерес у статистиков, которые рассматривают их как дополнительный источник сведений в условиях бурного развития информационных технологий. При этом некоторые пользователи переоценивают их потенциал и часто трактуют большие данные как предстоящую замену официальной статистики. Однако, по мнению автора, такое заключение является преждевременным; использовать большие данные необходимо с определенной осторожностью. Автор статьи отмечает два важных момента. Во-первых, для значительного числа пользователей интерес представляет только часть больших данных, а именно структурированные данные, в результате чего объем первых значительно сокращается. Во-вторых, как наукой, так и практикой доказано, что для получения достоверных результатов достаточно наблюдать небольшое число единиц, отобранных на основе случайной выборки (выборочной совокупности).

В статье также дается критическая оценка больших данных с точки зрения других национальных базовых принципов, принятых ООН для обеспечения качества статистических данных. Особо выделяются проблемы, связанные с мониторингом достижения Целей устойчивого развития (ЦУР). Автор отмечает, что за данные, предоставляемые официальной статистикой, несут ответственность национальные статистические управления (НСУ). В отсутствие какой-либо институциональной ответственности надежность больших данных может быть поставлена под сомнение.

В заключение подчеркивается, что пригодность больших данных определяется обоснованностью предположений, которые устанавливаются в ходе трансформации неструктурированного массива информации для проведения некоторого количественного измерения. В противном случае, по мнению автора, возможно проникновение в информационное поле потока нестатистической количественной информации в большом объеме, которая может дезинформировать общество и привести органы государственного управления и бизнеса к принятию неверных решений.

Ключевые слова: большие данные, качество данных, Цели устойчивого развития (ЦУР), официальная статистика.

JEL: C18, C80.

doi: <https://doi.org/10.34023/2313-6383-2019-26-12-5-14>.

Для цитирования: Упадхьяя Ш. Большие данные и официальная статистика. Вопросы статистики. 2019;26(12):5-14.

Big Data and Official Statistics

Shyam Upadhyaya

UN Industrial Development Organization (UNIDO), Vienna

Big data is a component of the Fourth Industrial Revolution. The deep penetration of digital technology has turned data into an essential component of the production process. Data are automatically generated by machines during the course of operation and during interactions with humans.

This paper describes the concept and composition of big data. Most of the big data are unstructured and include text, audio-video files, images, emails, log files, etc. Statisticians are more interested in structured data presented in a pre-defined database model. Big data offer new sources and opportunities that cannot be discounted. However, the use of big data requires proper assessment in terms of quality dimensions such as accuracy, comparability and methodological soundness. Against the backdrop of arguments regarding big data, some users view big data as a replacement of official statistics. Such a conclusion is premature for at least two reasons: first, only a small part of big data can be used for decision-making. Second, theory and practice prove that a small sample based on scientific methods can yield much more reliable and accurate estimates than the results obtained from the processing of large amounts of unstructured data.

The paper assesses the possibility of using big data for Sustainable Development Goals (SDG) monitoring, which is a nationally owned process, and NSOs are accountable for the SDG data they report. If the data are derived from a big data source, irrespective of the level of technical sophistication used in data transformation, the reliability of such data might be questioned by the national institutions.

The paper concludes that the reliability of data obtained from big data sources hinges on the quality of tools and methods applied to data transformation. Statisticians can play an important role in alerting society, decision-making bodies of the government and businesses about the reliability of information derived from the different sources.

Keywords: big data, data quality, Sustainable Development Goals (SDG), official statistics.

JEL: C18, C80.

doi: <https://doi.org/10.34023/2313-6383-2019-26-12-5-14>.

For citation: Upadhyaya S. Big Data and Official Statistics. *Voprosy Statistiki*. 2019;26(12):5-14. (In Russ.)

Огромный массив больших данных, поступающих в реальном времени, создал у пользователей повышенные ожидания относительно того, что они способны заменить дорогостоящую официальную статистику, публикуемую со значительным отставанием. Пока спор о больших данных продолжается, некоторые руководители международных организаций уже сейчас видят возможность сотрудничества с крупными транснациональными компаниями, работающими в сферах цифровой экономики и информационных технологий, для получения недостающих данных, необходимых для мониторинга прогресса в достижении Целей устойчивого развития (ЦУР). Схожая тенденция наблюдается и в практике национальных статистических служб.

Действительно, получение информации традиционными методами статистических обследований, проводимых органами официальной статистики, является достаточно кропотливой и трудоемкой работой. Формирование данных - это сложный и длительный процесс, включающий их передачу, редактирование и агрегирование в местных и центральных органах государственной статистики.

Несмотря на желание некоторых специалистов упростить этот сложный процесс, мы должны серьезно задаться вопросом: являются ли большие данные реальной альтернативой официальной статистике? В этой статье сделана попытка дать критическую оценку больших данных с точки зрения их качества. С этой целью рассматриваются такие важнейшие критерии качества данных, как точность измерения, согласованность, сопоставимость и отчетность.

1. Миф об объеме больших данных

Большие данные имеют так называемые 3-V (*volume, variety, velocity*) свойства: объем, многообразие и скорость. Трудно спорить с тем, что благодаря новейшим технологиям большие данные действительно отличаются своим многообразием и создаются с небывалой ранее скоростью, а процесс их возникновения можно наблюдать в реальном времени. Специалисты говорят, что 90% всех ныне существующих данных было создано за последние два-три года. По некоторым оценкам, ежедневно в мире производится около 2,5 эксабайта данных (1 эксабайт равен 1 миллиарду гигабайт). Понятие «данные» в свою очередь тоже изменилось, поскольку большие данные включают в себя не только цифры в традиционном понимании, но и текстовую часть, аудио- и видеозаписи в социальных сетях, фотографии, спутниковые изображения, электронные письма, программы и приложения и многое другое. Поэтому различают структурированные и неструктурированные данные.

Структурированные данные - это в основном количественные данные, которые представлены в базе данных с заранее определенной моделью для их хранения, обработки и распространения. Именно поэтому эти данные представляют большой интерес для статистиков.

Подавляющая часть больших данных не имеет ни количественного выражения, ни какой-либо структуры. Они включают в себя текстовую часть, аудио- и видеозаписи, фотографии и прочие компоненты, перечисленные выше.

Именно за счет них образуется такой колоссальный массив данных. Статистические данные, упорядоченные по определенным форматам, не являются столь объемными. Как представлено на рис. 1, объем неструктурированных данных, по оценке автора, может вырасти в четыре раза за период с 2010 по 2020 г., в то время как рост структурированных данных останется не столь значительным.

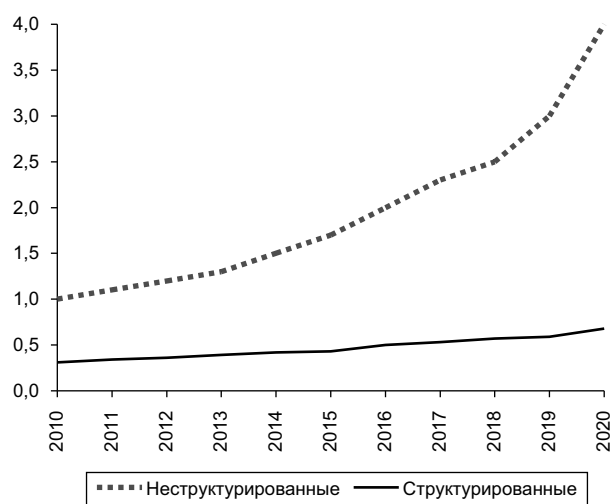


Рис. 1. Расчетный объем больших данных по видам (зетабайт)

Источник: расчеты автора на основе данных из разных источников, в том числе Oracle Advanced Analytics. URL: <https://www.oracle.com/database/technologies/>.

Для сравнения: объем базы данных ЮНИДО, которая содержит сведения о добывающей и обрабатывающей промышленности, а также электроэнергии, газо- и водоснабжении почти 200 стран и территорий за более чем 50 лет, составляет всего 1 гигабайт (файлы в сжатом виде). В то же время один фильм с двухчасовой трансляцией на *You Tube*-канале или *Netflix* занимает 2,5 гигабайта. Общий объем данных, содержащихся в видеохостинге *You Tube*, оценивается в 1 эксабайт (эквивалентно 1 млрд гигабайт).

Поэтому чрезвычайно быстрое увеличение объема больших данных не означает наличия большого объема статистических данных. Наоборот, чем больше растет объем и многообразие больших данных, тем сложнее, более трудоемким и дорогостоящим становится их использование, особенно учитывая тот факт, что национальные статистические службы не имеют необходимых навыков и программного обеспечения для управления колоссальным объемом данных,

наличие которых не обещает быть пригодным в анализе.

2. Достоверность важнее объема

Для того чтобы подчеркнуть, что правильно проведенная случайная выборка небольшого количества объектов дает гораздо более достоверные результаты в отличие от наблюдения большого числа статистических единиц, часто приводят пример опроса избирателей в США, проведенного в 1936 г. Тогда журнал *Literary Digest* провел самый крупный опрос в истории США с участием 2,4 млн респондентов, отдававших свое предпочтение одному из двух кандидатов: Альфреду Лэндону или Франклину Рузвельту, занимавшему в то время пост президента. Результаты опроса однозначно указали на внушительную победу Лэндона. Однако на выборах одержал убедительную победу Рузвельт, обыграв своего оппонента в 46 из 48 штатов. В это же время Американским институтом общественного мнения был проведен другой опрос, охват которого составил менее 2% респондентов от числа опрошенных *Literary Digest*. Результаты этого опроса верно указали на победу Рузвельта. Разница в проведении двух опросов заключалась в том, что *Literary Digest* посылал опросные листы своим читателям, большинство из которых отдавали свои предпочтения Лэндону. Опрос же Американского института общественного мнения был основан на случайной выборке, соответственно, выборочная совокупность объективно отражала структуру генеральной совокупности.

Достоверность результатов не гарантируется лишь включением большого количества единиц в статистическое обследование. Чтобы понять эту логику, достаточно вспомнить теорему П.Л. Чебышева [1]. При достаточно большом количестве объектов n случайной выборки можно с вероятностью, близкой к единице, утверждать, что отклонение выборочной средней от генеральной средней будет сколь угодно малым:

$$P\left[|\bar{x} - \bar{X}| < \frac{tS}{\sqrt{n}}\right] > 1 - \frac{1}{t^2}, \quad (1)$$

где $\bar{x}$, $\bar{X}$ — значения средней для выборочной и генеральной совокупности соответственно; n — размер выборки (число выбранных единиц).

Необходимый размер выборки, вытекающий вследствие неравенства Чебышева:

$$n = \frac{t^2 s^2}{a^2}, \quad (2)$$

где s^2 - дисперсия (обычно берется выборочная дисперсия, так как генеральная дисперсия неизвестна); a - ошибка выборки; значение $t = 1,96$ для вероятности 0,95.

В этом соотношении число единиц в генеральной совокупности полностью отсутствует, и оно не играет главной роли в определении размера выборки. Также генеральная дисперсия, как правило, неизвестна, и она заменяется выборочной дисперсией. Определение размера выборки еще более упрощается, если исследователь имеет целью получить оценку доли альтернативного признака.

Для оценки доли p размер выборки при заданной ошибке с уровнем значимости a определяется как:

$$n = \frac{t^2 p(1-p)}{a^2}. \quad (3)$$

Как известно, дисперсия достигает своего максимального значения 2500, когда доля обоих альтернативных признаков одинакова, то есть равна 50%. Тогда необходимая численность выборки для предельной ошибки в 5% с вероятностью 0,95 составит всего 400 единиц (при использовании метода случайной выборки с повторным отбором) [2].

Неудивительно, что опрос, в котором приняли участие менее 1000 избирателей [3], мог дать достаточно точные результаты президентских выборов в США в 2012 г. (при наличии всего двух кандидатов метод альтернативного признака работает достаточно эффективно).

Поэтому как бы странно это ни звучало для неспециалистов, необходимое число респондентов в выборке для получения оценки, скажем, доли населения, не имеющего доступа к Интернету, или доли жителей, проживающих ниже прожиточного минимума, для таких абсолютно разных городов по численности населения, как Москва и Якутск, будет одинаковым. Однако необходимым условием является то, чтобы выборочная совокупность правильно отражала структуру исследуемых элементов в генеральной совокупности. К сожалению, в случае с большими данными это не так.

В качестве примера можно привести данные социальных сетей, которые считаются одним

из главных источников информации для проведения социальных обследований. Установлено, что число подписчиков социальных сетей давно достигло нескольких миллиардов. На своих страницах они размещают разную информацию о своем месте жительства, уровне образования, выкладывают фотографии разных стран, в которых они побывали. Из этой информации, как предполагают, можно извлечь большой объем данных для социальной статистики, например статистики образования, туризма и т. д. Для проверки мы обработали данные социальных сетей, наиболее популярных в России¹, и сопоставили их с официальными данными Росстата. Отклонения оценок соотношения мужчин и женщин по данным пользователей социальных сетей и других интернет-сайтов от оценок Росстата представлены на рис. 2.



Рис. 2. Вариация оценок соотношения мужчин и женщин по источникам данных [данные Росстата (все население) = 100]

Данные об интернет-пользователях получены только для возрастной группы 12-65 лет. Соотношение мужчин и женщин в этой группе оказалось выше (92,4 мужчин на 100 женщин), чем соотношение мужчин и женщин в общей численности населения России. Однако по данным социальных сетей, это же соотношение колеблется от 83,0 (ВКонтакте) до 91,1 (пользователи разных интернет-сайтов).

Как видно, данные, полученные из открытых интернет-источников, в силу разных причин неверно отражают структуру генеральной совокупности, хотя они основаны на довольно большом количестве наблюдений. Например, число активных пользователей ВКонтакте составляет 70 млн человек. Были рассмотрены также данные, взятые с других сайтов, таких как Mail.Ru, Google.ru и

¹ На основе расчета Ярослава Сергеевского по данным Media Scope. URL: <https://webindex.mediascope.net/>.

Youtube.com, и результаты тоже оказались несопоставимыми. Это означает, что если проводится какое-нибудь обследование на основе этих данных, то статистическая оценка будет смещенной, иными словами, такое обследование приведет к систематическому завышению или занижению оценок исследуемых переменных.

В то же время НИУ ВШЭ уже несколько лет успешно проводит обследование экономического положения и здоровья населения России², которое охватывает 6000 случайно отобранных домашних хозяйств (размер выборки увеличивался в последующих обследованиях), при этом производились достаточно надежные оценки по ряду социально-экономических показателей. Росстат тоже проводит несколько регулярных обследований, основанных на случайных выборках. Количество выборочных единиц в этих обследованиях, естественно, намного больше, потому что оценка производится не только на национальном, но и на региональном уровнях. Соответственно, генеральная совокупность состоит из ряда подсовкупностей, из которых формируются выборки для каждого региона.

3. Обеспечение сопоставимости данных

Пример, приведенный на рис. 2, указывает на то, что большие данные не удовлетворяют критерию сопоставимости, который является одним из главных базовых принципов качества данных, принятых Статистической комиссией ООН [4]. Во-первых, они несопоставимы между собой, поскольку данные, полученные из одних источников, отличаются от других. Во-вторых, большие данные несопоставимы с данными официальной статистики. При этом расхождения

вполне могут быть неизмеримы, то есть разность величин может происходить за счет применения разных методологий, классификаций и т. д. На какие данные должен полагаться пользователь?

В системе официальной статистики один государственный орган проверяет и сверяет все данные в рамках единой системы методики и классификации. Когда данные поступают из неофициальных источников, каковыми являются большие данные, они сами по себе не проходят такую проверку. Требуются огромные усилия для приведения больших данных в соответствие с требованиями принятой национальной методики и международных рекомендаций, что, в свою очередь, уменьшает преимущества, которые могут иметь большие данные в отношении своевременности и экономической эффективности.

Часто указывают на возможность использования больших данных для проведения оценок некоторых величин, отсутствующих в результатах статистического обследования. Примером могут послужить данные *Google Trends*, которые предоставляют информацию о том, как часто выполнялся поиск определенного термина по отношению к общему количеству поисковых запросов. *Google Trends* позволяет отслеживать изменение спроса на определенные промышленные товары, отражающее предпочтения потребителей. Однако платформа *Google Trends* может быть использована только для наблюдения краткосрочных изменений. Ее пригодность для анализа долгосрочных процессов остается весьма спорной [5].

На рис. 3 представлено сопоставление статистических данных, содержащихся в INDSTAT2 (базе ЮНИДО), с данными, полученными из *Google Trends*.

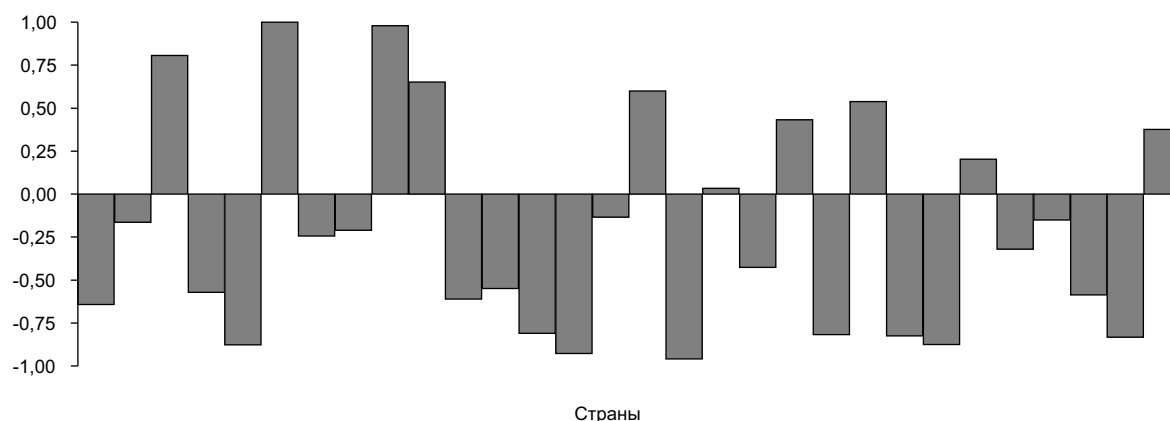


Рис. 3. Корреляция между данными *Google Trends* и базой данных INDSTAT2

² Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ. URL: <https://www.cpc.unc.edu/projects/rhms-hse/project/sampling>.

Предварительный анализ согласованности данных из этих двух источников показал отсутствие какой-либо связи между ними.

4. Обеспечение точности и надежности

Производство статистических данных не является самоцелью – они должны приносить пользу обществу. На их основе органы государственного управления разрабатывают необходимую политику. Потребность в тех или иных данных определяется сложившимися приоритетами социально-экономического развития. В отличие от традиционных методов наблюдения большие данные собираются без какой-либо определенной цели. Поэтому возникает сомнение: данные, может быть, и существуют, но пригодны ли они [6]?

На самом деле, большие данные редко дают точное измерение интересующих исследователя явлений. Их пригодность определяется обоснованностью предположений, которые устанавливаются в ходе трансформации неструктурированного массива информации для проведения некоторого количественного измерения. В настоящее время, когда государства – члены ООН сообща работают над выполнением ЦУР, потребность в данных для мониторинга целей и задач устойчивого развития резко возросла. Однако национальные статистические службы даже развитых стран, не говоря уже о большом числе развивающихся стран, не могут регулярно предоставлять требуемые данные по многим показателям. В сложившейся ситуации некоторые эксперты предлагают пользоваться информацией из источников больших данных, которая имеет весьма отдаленную связь с реальным положением дел. Например, в брошюре, выпущенной *UN Global Pulse*³, предложено пользоваться данными о перемещении пользователей мобильных телефонов, для того чтобы предсказать распространение инфекционных заболеваний. Предполагается, что после вспышки заболевания люди перемещаются из места его распространения, что можно отследить по данным мобильных телефонов. Этими данными можно пользоваться, только если верить, что каждый пользователь мобильного телефона заражен.

Другой пример связан с почтовыми данными, которые якобы пригодны для оценки роста ВВП.

Предполагается, что объем и частота почтовых отправок тесно коррелируют с доходом на душу населения, поэтому по данным о почтовых отправлениях можно судить об экономическом росте той или иной страны. Если даже верить в такое предположение, почтовые данные в лучшем случае могут лишь приблизительно указать на характер динамики, но никак не могут определить темпы роста ВВП. Необходимо учесть еще тот факт, что требуются значительное время и программные навыки для перевода данных из исходных компьютерных материалов в некоторую структуру. Даже после этих преобразований большие данные не могут достичь той степени точности и надежности, чтобы на их основе можно было бы принимать важные решения.

5. Ответственность за разработку и использование правильной методологии

Одним из национальных базовых принципов обеспечения качества данных является ответственность за применение правильной методологии. Официальная статистика производит данные на основе утвержденной методологии, которая во многом соответствует международным стандартам и рекомендациям. Гармонизация методологии национальной статистики обеспечивает сопоставимость статистических данных во времени и пространстве (по регионам страны, по странам). Официальная статистика использует механизм проверки согласованности данных, полученных из разных источников. Например, насколько соответствует показатель количества предприятий в бизнес-регистрации результатам годовых обследований; число работников в отраслях экономики совпадает с данными о трудовых ресурсах, полученными в ходе переписи населения, и т. д. Сводные данные регулярно сверяются с основными макроэкономическими показателями, в том числе с ВВП. Очевидно, есть необходимость создать такой же механизм для оценки и использования больших данных в официальной статистике. Главной целью этой работы является обеспечение контроля качества данных, полученных из неофициальных источников.

В начале этого процесса национальные статистические управления (НСУ) могли бы выработать методику использования данных из неофициальных источников, в том числе из источников

³ UN Global Pulse: Big Data and SDGs. URL: https://www.un.org/en/pdfs/Bigdata_SDGs_single_spread_2017.pdf.

больших данных, в составлении показателей, принятых в официальной статистике. Ученые часто указывают на возможность обращения к большим данным для получения некоторых экстраполяторов, которые впоследствии можно было бы применять для оценки добавленной стоимости на отраслевом уровне или других показателей на макроэкономическом уровне. Для этого необходимо предварительно анализировать исходные данные с точки зрения определения, охвата, сезонности и весовых коэффициентов. В настоящее время такие методологии имеют только несколько НСУ промышленно развитых стран. Большинству других НСУ лишь предстоит запустить процесс, без которого использование больших данных для принятия важных решений социально-экономического развития или мониторинга ЦУР невозможно.

Мониторинг ЦУР является процессом, согласованным с государствами - членами ООН, который не может осуществляться вне официальной статистики. НСУ несут ответственность за данные, которые они предоставляют. Если данные непосредственно предоставлены частными компаниями или исследовательскими центрами из источников больших данных, НСУ может отказаться от ответственности за использование таких данных независимо от уровня программного обеспечения. На сессиях Статистической комиссии ООН бывали случаи, когда некоторые НСУ отказывались от принятия Доклада о человеческом развитии по причине того, что сведения, опубликованные в докладе ООН, отсутствовали в национальной базе данных.

6. Большие данные и глобальный мониторинг ЦУР

Завершается уже четвертый год с момента принятия ООН Повестки дня в области устойчивого развития на период до 2030 г. Статистика играет незаменимую роль в мониторинге прогресса стран в достижении ЦУР. Разработана схема поступления национальных данных в глобальную базу. Требуются данные по 230 показателям, которые были разработаны экспертной группой, состоящей из представителей национальных статистических ведомств и международных организаций, и были одобрены Статистической комис-

сией ООН. Однако, как отмечено выше, данные по многим странам отсутствуют. Глобальная база данных ЦУР пока содержит информацию по 176 показателям из 230, но это сильно варьируется в зависимости от страны, так как некоторые развивающиеся страны Азии и Африки не могут предоставить данные более чем по 30% показателей. Международные организации, активно выполняющие задачи в области ЦУР, часто сталкиваются с нехваткой данных, необходимых для разработки разного рода проектов и программ развития. Предвидя эту проблему, Генеральная ассамблея ООН изначально отмечала: «Мы признаем, что по нескольким задачам исходные данные по-прежнему отсутствуют, и призываем увеличить помощь, оказываемую в деле активизации сбора данных и укрепления потенциала в государствах-членах... Мы обязуемся устранить этот пробел в сборе данных с тем, чтобы наладить более надежное информационное обеспечение»⁴.

Однако международное сообщество не смогло мобилизовать необходимые финансовые средства на развитие статистики для выполнения задач ЦУР. Доля ассигнования на статистические нужды в общем объеме официальной помощи в целях развития составляет всего 0,3%. По оценке Paris-21 [7], ежегодный финансовый дефицит, исходя из нужд, определенных Кейптаунским глобальным планом действий в отношении данных в области устойчивого развития [8], составляет 700 млн долларов США. Оценка проведена с учетом того, что требуются дополнительные ресурсы на развитие национальных статистических систем во многих развивающихся странах. Представители национальных статистических управлений едины в том, что для мониторинга ЦУР должны использоваться официальные статистические данные. Следовательно, необходимо провести работу по улучшению качества статистических данных, полученных из официальных источников. Однако нехватка финансовых средств является большим препятствием в этой работе.

В то же время некоторые влиятельные коммерческие организации и исследовательские центры активно предлагают свои статистические и аналитические услуги. Среди них можно выделить *Sustainable Development Solutions Network (SDSN)* - частный исследовательский институт, возглавляемый профессором Джефффри Саксом

⁴ Резолюция 70/1, принятая Генеральной ассамблеей 25 сентября 2015 г., «Преобразование нашего мира: Повестка дня в области устойчивого развития на период до 2030 года». URL: <https://undocs.org/ru/A/RES/70/1>.

из Колумбийского университета. *SDSN* публикует ежегодный доклад устойчивого развития (*Sustainable Development Report*), основанный на своем собственном выборе показателей ЦУР и вторичных данных [9]. Другим подобным институтом является *International Institute for Sustainable Development (IISD)* под руководством Лина Вагнера. Этот институт представляет собой центр онлайн-информации по ЦУР. Еще один институт по ЦУР - *Big Data for Development (BD4D)* - определяет себя как партнера по проведению соответствующих исследований больших данных, связанных с целями в области устойчивого развития.

На содержание работы этих институтов, целью которых не является оказание помощи национальным статистическим управлениям, уходят достаточно большие средства. Однако нет конкретных примеров или перечня определенного числа показателей, источником которых могли бы стать большие данные. Например, ЮНИДО

является одним из ответственных международных агентств по шести показателям ЦУР-9, связанным с индустриализацией. Однако ни один из этих показателей не может быть составлен из источников больших данных. Поэтому большие данные во многих случаях остаются пока лишь предметом исследовательского интереса, а не конкретным практическим решением.

Наличие больших данных и возможность их использования зависят от уровня доступа к Интернету. Несмотря на колоссальный рост больших данных, особенно связанный с широким распространением Интернета вещей (англ. *Internet of Things, IoT*), страны сильно различаются по своим возможностям доступа к Интернету. По данным Международного союза электросвязи (*ITU*)⁵, число интернет-пользователей в наименее развитых странах на каждые 100 человек населения составляет всего 21 по сравнению с 85 в промышленно развитых странах (см. рис. 4).

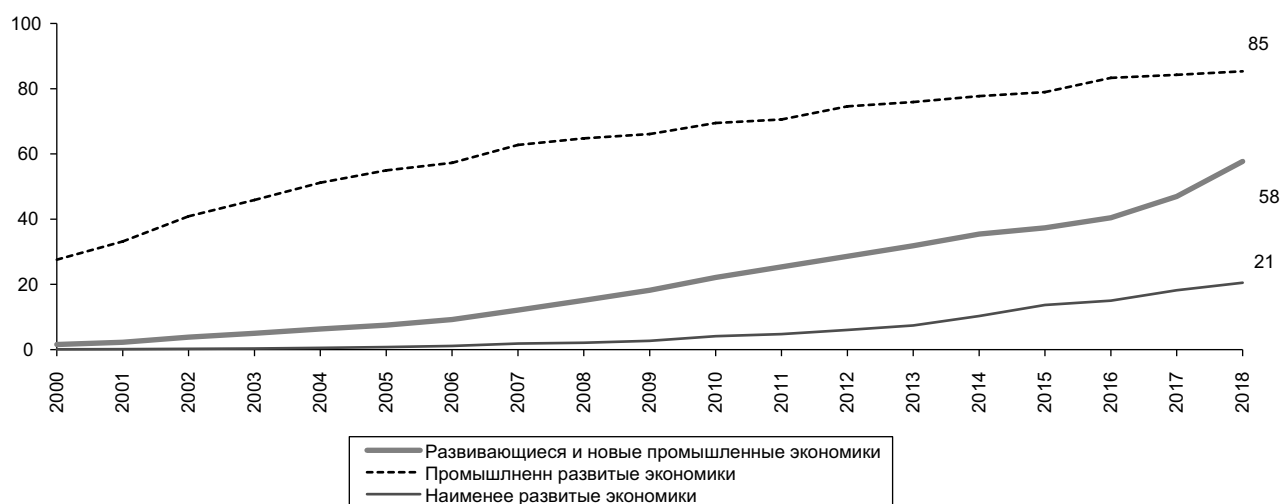


Рис. 4. Число интернет-пользователей на 100 человек населения, по группам стран

Таким образом, возможности использования больших данных в развитых странах гораздо выше, но эти страны особо не нуждаются в них в силу своей достаточно сильной официальной статистики. Однако и в тех развивающихся странах, где официальная статистика слабая, возможности использования больших данных ограничены. Так, например, трансформация больших данных требует специалистов по работе с данными и специалистов высокого класса в области цифровых технологий, а в развивающихся странах наблюдается их недостаток.

Цели устойчивого развития - это глобальная программа, охватывающая все страны в мире независимо от их географического расположения или стадии развития. В их основе - принцип инклюзивности, при котором также учтены интересы малых и островных государств.

* * *

Статистики не отрицают те большие перемены, которые произошли в информационном обществе благодаря крупному технологическому прогрессу.

⁵ URL: <https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx>.

Революция данных - это реально произошедшее событие, вследствие которого они стали частью современного производства, социального общения и индивидуальной деятельности. Эта статья не имеет целью в какой-либо мере уменьшить значение тех широких возможностей, которые открывают большие данные. Но в то же время не следует переоценивать их значение и преждевременно рассматривать как возможную замену официальной статистики. В этой связи также необходимо предостеречь о возможном проникновении в информационное поле потока нестатистической количественной информации в большом объеме, которая может дезинформировать общество и поставить органы государственного управления и бизнеса на путь неверных решений. Статистики могут сыграть очень важную роль в том, чтобы правильно воспользоваться предоставленной технологией.

Статистики также вправе выбирать источники данных, основываясь на разных приоритетах, таких как точность измерения, своевременность или затраты на получение данных. Если, например, большие данные окажутся своевременными, но дорогостоящими и не достаточно точными, статистический орган может от них отказаться. Поэтому дискуссия должна идти не столько о наличии или объеме больших данных, сколько о методах и моделях их обработки и трансформации в сопоставимую и достоверную статистическую информацию. Наиболее эффективным исполнителем этого процесса могли бы стать НСУ, которые отвечали бы за качество данных независимо от того, из каких источников они получены первоначально.

Что касается мониторинга ЦУР, существует механизм, принятый Статистической комиссией ООН, который определяет способ передачи данных из официальной системы национальной статистики на глобальную платформу. Попытка

некоторых влиятельных коммерческих организаций миновать этот процесс несовместима с общим пониманием задач статистики в достижении ЦУР.

Литература

1. Венецкий И.Г., Венецкая В.И. Основные математико-статистические понятия и формулы в экономическом анализе. М.: Статистика, 1974. 279 с.
2. Cochran W. Sampling Techniques. New York: John Wiley & Sons, 1999. 428 p.
3. Kaplan R.M., Chambers D.A., Glasgow R.E. Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias // CTS Journal. 2014. Vol. 7. Iss. 4. P. 342-346. doi: <https://doi.org/10.1111/cts.12178>.
4. Национальные базовые принципы обеспечения качества. Доклад Генерального секретаря ООН. 2012. URL: <https://unstats.un.org/unsd/statcom/doc12/2012-13-NQAF-R.pdf>.
5. Upadhyaya S., Kynclova P. Big Data - Its Relevance and Impact on Industrial Statistics. Vienna: UNIDO, 2017.
6. Cox D.R., Kartsonaki C., Keogh R.H. Big Data: Some Statistical Issues // Statistics & Probability Letters. 2018. Vol. 136. P. 111-115. doi: <https://doi.org/10.1016/j.spl.2018.02.015>.
7. Calleja R., Rogerson A. Financing Challenges for Developing Statistical Systems: A review of Financing Options. PARIS21 Discussion Paper, No. 14. Paris, 2019. URL: <https://paris21.org/sites/default/files/2019-01/Financing%20challenges%20for%20developing%20statistical%20systems%20%28DP14%29.pdf>.
8. UN Statistical Commission. Cape Town Global Action Plan for Sustainable Development Data. Adopted by the UN Statistical Commission at its 48th Session, March 2017. URL: https://unstats.un.org/sdgs/hlg/Cape_Town_Global_Action_Plan_for_Sustainable_Development_Data.pdf.
9. Sachs J. et al. Sustainable Development Report 2019. New York: Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN), 2019. URL: https://s3.amazonaws.com/sustainabledevelopmentreport/2019/2019_sustainable_development_report.pdf.

Информация об авторе

Упадхья Шьям - Ph.D. (экономическая статистика), главный статистик Организации Объединенных Наций по промышленному развитию, ЮНИДО. Vienna International Centre, Wagramer Strasse 5, P.O. Box 300, A-1400 Vienna, Austria. E-mail: s.upadhyaya@unido.org. ORCID: <https://orcid.org/0000-0002-1463-240X>.

Благодарность

Автор выражает искреннюю благодарность д-ру экон. наук, профессору Игорю Ульянову за комментарии и помощь в редактировании статьи, а также главному специалисту - эксперту Росстата Артему Душечкину за содействие при подготовке данной статьи.

References

1. **Venetskii I.G., Venetskaya V.I.** *Basic Mathematical and Statistical Concepts and Formulas in Economic Analysis*. Moscow: Statistics Publ.; 1974. 279 p. (In Russ.)
2. **Cochran W.** *Sampling Techniques*. New York: John Wiley & Sons; 1999. 428 p.
3. **Kaplan R.M., Chambers D.A., Glasgow R.E.** Big Data and Large Sample Size: A Cautionary Note on the Potential for Bias. *CTS Journal*. 2014;7(4):342-346. Available from: doi: <https://doi.org/10.1111/cts.12178>.
4. *National Quality Assurance Frameworks. Report of the Secretary-General. 2012*. Available from: <https://unstats.un.org/unsd/statcom/doc12/2012-13-NQAF-R.pdf>. (In Russ.)
5. **Upadhyaya S., Kynclova P.** *Big Data - Its Relevance and Impact on Industrial Statistics*. Vienna: UNIDO; 2017.
6. **Cox D.R., Kartsonaki C., Keogh R.H.** Big Data: Some Statistical Issues. *Statistics & Probability Letters*. 2018;(136):111-115. Available from: doi: <https://doi.org/10.1016/j.spl.2018.02.015>.
7. **Calleja R., Rogerson A.** *Financing Challenges for Developing Statistical Systems: A review of Financing Options*. PARIS21 Discussion Paper, No. 14. Paris; 2019. Available from: <https://paris21.org/sites/default/files/2019-01/Financing%20challenges%20for%20developing%20statistical%20systems%20%28DP14%29.pdf>.
8. UN Statistical Commission. *Cape Town Global Action Plan for Sustainable Development Data*. Adopted by the UN Statistical Commission at its 48th Session, March 2017. Available from: https://unstats.un.org/sdgs/hlg/Cape_Town_Global_Action_Plan_for_Sustainable_Development_Data.pdf.
9. **Sachs J.** et al. *Sustainable Development Report 2019*. New York: Bertelsmann Stiftung and Sustainable Development Solutions Network (SDSN), 2019. Available from: https://s3.amazonaws.com/sustainabledevelopment.report/2019/2019_sustainable_development_report.pdf.

About the author

Shyam Upadhyaya - Ph.D. in Economic Statistics, Chief Statistician of UNIDO. Vienna International Centre, Wagramer Strasse 5, P.O. Box 300, A-1400 Vienna, Austria. E-mail: s.upadhyaya@unido.org. ORCID: <https://orcid.org/0000-0002-1463-240X>.

Acknowledgments

The author is grateful to Doctor of Economic Sciences, Professor Igor Ulyanov for the comments and support in editing the article, as well as to the Chief Specialist-Expert of Rosstat Artem Dushechkin for his assistance in the preparation of this article.