Организация Объединенных Наций ECE/CES/GE.41/2016/10

Экономический и Социальный Совет

Distr.: General
19 July 2016
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Группа экспертов по переписям населения и жилищного фонда

Восемнадцатое совещание
Женева, 28−30 сентября 2016 года
Пункт 7 предварительной повестки дня
Возможные виды использования новых источников данных (например, «больших данных») в целях переписей

Существует ли возможность использования «больших данных» в ходе переписи населения 2020 года?

Записка Центрального статистического управления Польши (ЦСУ)

Резюме

Возможности новых источников и технологий «больших данных» имеют весьма важное значение для официальной статистики, однако способны ли «большие данные» иметь значительные последствия для следующего цикла переписей? Большинство государств хотели бы провести перепись 2020 года с меньшими издержками по сравнению с предыдущей переписью 2010 года, сохранив при этом высокое качество результатов. Большой объем данных не является чем-то новым для многих статистических управлений, поскольку в ходе предыдущей переписи 2010 года многие страны использовали административные данные. В то же время «большие данные» являются весьма сложной и сравнительно новой темой.

Источники «больших данных» имеют значительный потенциал в контексте переписей в качестве важнейших статистических обследований. Однако раунд переписей 2020 года настолько близок, что существует высокая вероятность того, что источники «больших данных» не будут надлежащим образом изучены в требуемые сроки, изложены и практически использованы для статистических исследований в меньших масштабах. Существует много методологических и правовых проблем, проблем качества и трудностей в области ИТ в

GE.16-12448 (R) 030816 040816

1612448

плане использования «больших данных» в официальной статистике, при этом продолжение работы в данной области является одной из приоритетных задач. Таким образом, дальнейшая работа по использованию «больших данных» по-прежнему необходима, и в случае получения положительных и обнадеживающих результатов может быть рекомендовано использование этих источников в ходе следующих (т.е. после 2020 года) раундов переписей.

I. Введение

1. Явление «больших данных» получает все большее распространение по всему миру. Оно представляет собой глобальную тенденцию к постоянным изысканиям нового, неизменную потребность в развитии и совершенствовании мира вокруг нас. Это явление является наиболее заметным в предпринимательской деятельности частного сектора, главной целью которого является получение прибыли. Оно затрагивает также и государственный сектор. Органы государственного управления, хотя и преследуют другие цели, такие, как оптимизация государственных расходов, повышение эффективности функционирования государственных учреждений или информирование граждан, также осознают потенциал «больших данных». Кроме того, огромные возможности «больших данных» видят и органы официальной статистики. Путем сочетания новых источников данных и новых экспертных знаний статистическое сообщество в состоянии проанализировать это, пока что еще не изученное, явление и прийти к выводам, о которых ученые еще 10 лет назад и не мечтали.

II. Использование «больших данных» в официальной статистике

2. Обычные источники данных могут дать статистикам разнообразные возможности. В настоящее время официальная статистика, а также всевозможные показатели и метрические характеристики, основываются на данных, полученных из государственных регистров, и на информации, поступившей от предпринимателей, респондентов или собранной на основе наблюдений счетчиков и экспертов. Регистры, наблюдения и отчеты, поступившие от компаний, предлагают достоверные данные, в то время как опросы респондентов позволяют получить сведения о мнениях и настроениях, а также субъективные данные.

3. Вместе с тем в мире постоянно происходят изменения, при этом возникают новые явления, которые также требуют описания с использованием статистических процессов. В силу этого, невозможно ограничится лишь известными источниками данных и следует постоянно изыскивать новые пути и решения. Это, кроме того, позволит по-новому взглянуть на нынешние методы работы.

4. В качестве примера можно привести разработку обследования по вопросам ИКТ в Польше. В настоящее время для изучения таких явлений используют, среди прочего, вопросники. Респонденты отвечают на вопросы о том, как они пользуются компьютерами и Интернетом. Использование «больших данных» в этой конкретной области означает, например, создание комплекса роботов, сканирующих веб-сайты и изучающих их содержание на основе ранее зафиксированных моделей. Полученные результаты дают, среди прочего, ответы на вопросы относительно использования ИКТ на предприятиях, типов веб-сайтов, об онлайн-магазинах, видах имеющихся возможностей регистрации данных и т.д.

5. «Большие данные» также широко применяются в отношении рынка труда. Посредством надлежащего анализа веб-сайтов, социальных сетей и новостных сайтов, например путем использования некоторых ключевых слов или алгоритмов поиска данных о занятости с разбивкой по роду занятий, регионам (провинциям) секторам ответственности, статистические органы получают возможность дополнить данные о рынке труда, особенно в контексте спроса на рабочую силу, текучести кадров и забастовок.

6. С другой стороны, данные операторов мобильной телефонной связи помогают в изучении явления перемещения населения в дневное и ночное время, что является прекрасной основой для более подробного анализа причин этих миграционных потоков.

7. Дорожное движение, в том числе данные о количественных потоках и типе транспортных средств, можно контролировать с помощью камер и датчиков, установленных на разных дорогах. В результате, например, обследований, проводимых на пограничных контрольно-пропускных пунктах, можно свести к минимуму или по крайней мере сократить число счетчиков, занимающихся подсчетом транспортных средств.

8. Использование «больших данных» не только дополнило бы данные для статистики, но в отдаленном будущем позволило бы частично заменить имеющиеся в настоящее время обычные обследования. В результате этого граждане заполняли бы меньше вопросников и сократилось бы количество статистических счетчиков, занимающихся сбором таких вопросников, при сохранении уровня качества существующих данных.

III. Отдельные международные проекты

9. Работа по использованию «больших данных» для целей официальной статистики на международном уровне продолжается на протяжении последних нескольких лет. В ней участвуют учреждения как глобального, так и европейского уровня, а также представители деловых и научных кругов. Среди нескольких инициатив следует упомянуть инициативу ЕЭК ООН. ЕЭК ООН разработала проект на основе совместного подхода, в котором приняли участие более 70 экспертов из национальных и международных статистических организаций всего мира, с тем чтобы выявить и устранить основные проблемы, связанные с использованием источников «больших данных» для целей официальной статистики. Этот проект осуществлялся в период с января по декабрь 2014 года и преследовал три основные цели:

- выявить, изучить и подготовить методические указания для статистических организаций по основным стратегическим и методологическим вопросам, связанным с использованием «больших данных» при подготовке официальной статистики;
- продемонстрировать целесообразность эффективной подготовки как новой статистической продукции, так и «традиционной» официальной статистики с использованием источников «больших данных» и возможность копирования этих подходов в разных национальных условиях;
- содействовать обмену знаниями, техническим опытом, инструментами и методами между организациями в целях подготовки статистики с использованием источников «больших данных».

10. Этот проект включал следующие целевые группы, отвечающие за соответствующие вопросы:

- Целевая группа по партнерствам;
- Целевая группа по частному сектору;
- Целевая группа по вопросам качества;

- Целевая группа по экспериментальным вопросам. Работа этой группы была продолжена и в 2015 году.

11. В результате осуществления данного проекта был подготовлен ряд документов, касающихся различных аспектов, которыми занимались целевые группы, а также создан перечень «больших данных».

12. Еще одной важной инициативой является проект «больших данных» сети ЕСС, охваченный Планом действий и дорожной картой в области «больших данных», цель которых состоит в том, чтобы интегрировать ПДБД в Концепцию Европейской статистической системы (ЕСС) до 2020 года.

13. Этот проект имеет две цели. Первая цель заключается в проведении экспериментальных проектов по изучению потенциала отдельных источников «больших данных» для подготовки или содействия подготовке официальной статистики. Вторая цель ориентирована на горизонтальные вопросы, такие, как методология, качество, инфраструктура ИТ и метаданные, которые закладывают условия для будущего использования этих источников данных в рамках Европейской статистической системы.

14. Общая цель проекта состоит в том, чтобы подготовить ЕСС для интеграции источников «больших данных» в производство официальной статистики. Критерии отбора ФПА предусматривают, что проект должен быть ориентирован в первую очередь на проведение экспериментальных проектов по изучению потенциала отдельные источников «больших данных» для подготовки или содействия подготовке официальной статистики. Задача этих проектов состоит в том, чтобы получить практический опыт в области использования «больших данных» в целях официальной статистики.

15. Конкретные цели проекта «больших данных» сети ЕСС включают:

- проведение экспериментальных проектов для подготовки статистических данных из источников «больших данных» на уровне ЕСС;
- анализ портфеля конечных данных, полученных из источников «больших данных»;
- разработка/обзор методологических рамок и основ качества для источников «больших данных» в официальной статистике;
- выявление, определение и внедрение инфраструктуры ИТ для обработки «больших данных»;
- обеспечение доступа к источникам «больших данных», выявление и подготовка условий неправового и правового характера в отношении доступа к «большим данным» в рамках ЕСС и их использования;
- обмен информацией с заинтересованными сторонами в рамках статистической системы и с исследовательским сообществом.

IV. Переписи

16. Тенденция к внедрению «больших данных» для целей официальной статистики становится все более очевидной, при этом потенциал источников «больших данных» огромен. В данном случае возникает естественный вопрос: если доказано, что «большие данные» могут быть успешно включены в официальную статистику на примере обследований, то можно ли использовать их уже в раунде переписей 2020 года?

17. Для ответа на этот вопрос сначала необходимо рассмотреть конкретные аспекты переписей. Перепись дает наиболее подробную информацию о численности населения, его территориальном нахождении, демографической, социальной и профессиональной структуре, а также о социально-экономических характеристиках домохозяйств и семей и их ресурсах, о жилищных условиях на всех уровнях территориального деления страны – на национальном, региональном и местном уровнях. Особое внимание уделяется получению сведений относительно изменений в демографических и социальных процессах, особенно в связи с ростом миграции. Результаты переписей используются непосредственно для нужд официальной статистики в качестве основы для создания основ выборки для последующих обследований, проводимых с использованием выборки домохозяйств. Наиболее важным моментом является получение информации по вопросам, которые были охвачены переписью в предыдущем раунде. Такая преемственность необходима для проведения сравнительного анализа явлений во времени и для описания изменений, которые произошли в демографических, социальных и экономических процессах с точки зрения народонаселения, состояния жилья и зданий; домохозяйств и семей применительно к жилищным условиям. Этот подход позволяет сохранить временные ряды и обеспечить сопоставимость с результатами предыдущих переписей.

18. Надежность результатов переписи зависит от выполнения следующих условий:

- универсальность – переписи должны проводиться на всей территории страны без исключения какого-либо региона или района;
- собранные данные должны относиться к конкретному моменту времени (определенный день и определенное время);
- непосредственный характер – информацию следует получать непосредственно от респондентов или из проверенных источников, таких, как высококачественные административные регистры;
- проведение только для статистических целей – это означает, что перепись имеет конфиденциальный характер, и полученные данные не могут быть использованы для выполнения других задач.

19. В силу этого представляется, что перепись является гораздо более важным мероприятием по сравнению с другими видами обследований. Периодичность ее проведения, которая зачастую составляет десять лет, означает, что после сбора данных они будут использоваться в качестве источника сведений в некоторых областях на протяжении десятилетия. По этой причине результаты переписи зачастую имеют решающее значение для разработки экономических и демографических стратегий. Таким образом, перепись должна проводиться на справедливой основе, аккуратным и профессиональным образом. Для выполнения этих условий необходимо иметь соответствующую методологию, инструментарий и знания, а сотрудники статистических органов должны располагать необходимой компетенцией.

V. Трудности

20. В связи с тем, что «большие данные» являются относительно новым явлением, многие страны еще не разработали четкие руководящие принципы для сбора, обработки и распространения таких данных. В силу этих ограничений многие аспекты данного явления не были проверены на практике. Даже в том

случае, если будут осуществлены различные практические экспериментальные проекты с использованием источников «больших данных», в силу нехватки времени до сих пор отсутствует надежный, устойчивый метод обработки и анализа данных. Качество продуктов, разработанных на основе «больших данных», разумеется, повышается, однако в случае такого уникального обследования, каким является перепись, никто не может позволить себе ошибки или неопределенность в отношении качества.

21. Еще одной проблемой, возникающей как на международном, так и на национальном уровнях, является отсутствие экспертов – так называемых специалистов по обработке и анализу данных. Стремление охватить различные данные, получаемые компаниями или людьми, например на веб-сайтах социальных сетей, данные государственных учреждений, данные, генерируемые всевозможными устройствами, машинами, стало стимулировать рост спроса на специалистов, обладающих исключительными навыками. В последнее время в результате объединения знаний, компетенции и навыков в таких областях, как информационные технологии, процессы, оптимизация предпринимательской деятельности, математика и статистика, была создана современная наука о данных, и специалисты – от «традиционных» аналитиков данных/предпринимательской деятельности до лиц, способных обрабатывать большие объемы данных (инженеры по данным/инженеры по интеллектуальному анализу данных/архитекторы данных), которые способны распространять свои знания, навыки и компетентность на другие области науки, предпринимательской деятельности, а также новые технологии, источники и каналы передачи информации, превратились в специалистов по обработке и анализу данных/инженеров по «большим данным».

22. Специалист по обработке и анализу данных является примером современного специалиста. В отличие от специалиста в обычном смысле, который является экспертом в одной области, этот новый специалист, будучи экспертом в своей основной деятельности, например в компьютерных науках, расширяет свои познания в принципиально иных областях, например в психологии. Сочетание этих, как представляется совершенно не связанных между собой навыков, является весьма плодотворным. Оно позволяет анализировать неструктурированные, динамические данные столь же динамичным и нестандартным образом. Именно по этой причине чрезвычайно трудно найти людей, которые удовлетворяют таким критериям. Существуют, конечно, ученые, обладающие соответствующими навыками, однако для того, чтобы извлечь выгоды из «больших данных», требуется нечто большее. Таким образом, один из ключевых вопросов заключается в том, чтобы организовать подготовку в области информационных технологий, охватив аналитические, методологические аспекты и т.д. В таком случае, каким образом при нехватке специалистов с надлежащей квалификацией проводить такое серьезное обследование, которым является перепись? Это было бы весьма рискованным шагом и, в конечном счете, могло бы привести к ошибкам.

23. Одним из главных препятствий является отсутствие достаточной правовой основы для сбора, анализа и хранения «больших данных». Масштабы этой проблемы весьма различаются в зависимости от страны. Каждое государство имеет свои собственные нормы, которые – для одних в большей степени, чем для других – затрудняют получение нестандартных данных, не предусмотренных существующей правовой системой. Кроме того, на уровне секторов действуют принципы конфиденциальности, которые зачастую не позволяют владельцам обмениваться своими данными для статистических целей. И хотя проведение переписи с использованием «больших данных» было бы, несомненно,

интересным проектом, необходимо рассмотреть вопрос о том, можно ли осуществить его в 2020 году или же данный проект целесообразно отложить до следующего раунда переписей. Помимо описанных выше проблем, связанных с качеством, нехваткой специалистов и отсутствием методологии, следует подумать о том, будут ли вообще в 2020 году иметься в наличии данные, без которых невозможно произвести подсчеты.

VI. Заключение

24. В завершение можно сказать, что использование «больших данных», при их многочисленных преимуществах, связано, тем не менее, и с некоторыми ограничениями, которые в большинстве своем носят критический характер, а иногда и сдерживают дальнейшую работу.

25. К сожалению, внедрение технологий «больших данных» в государственном секторе, включая статистику, является непростой задачей. В этом контексте государственный сектор сталкивается с гораздо более трудной задачей, чем сектор частный. В рамках предпринимательской деятельности, в которой основное внимание уделяется прибыли, вопрос о защите данных клиентов связан в основном с имиджем компании. В эпоху Интернета, когда неблагоприятные мнения распространяются со скоростью лесного пожара, ни одна компания не может позволить себе потерять репутацию. Это, в сочетании с различного рода правовыми гарантиями, обеспечивает защиту данных, не препятствуя развитию компании. Таким образом, при наличии надлежащих мер по защите данных клиентов, технологии «больших данных» могут быть использованы в различных формах.

26. Иная ситуация наблюдается в государственном секторе. Административные органы ориентированы не на прибыль, а на обслуживание граждан. В этом контексте государственные учреждения должны быть безупречными, поскольку они олицетворяют собой достоинство государства. Это означает, что необходимость в обеспечении безопасности данных в их случае еще выше, чем в случае предприятий. Гарантом безопасности является закон, и такой закон может носить весьма ограничительный характер.

27. Таким образом, официальная статистика не может строиться на необдуманных и поспешных решениях. Она должна обеспечивать качество и достоверность всех собираемых данных, в частности – данных переписей. Однако это не означает прекращения поиска новых путей. «Большие данные» – это хорошая руководящая идея, просто она слишком новая. На основе всех тех инициатив, о которых говорилось выше, вскоре будет получен ответ на вопрос о том, насколько значителен потенциал «больших данных» для статистики. После более тщательного и точного анализа постепенно выяснится, какие – на первом этапе ограниченные по масштабам и несложные – обследования могут быть дополнены «большими данными». Со временем перечень обследований, для которых будут использоваться «большие данные», будет расширен. Существует вероятность того, что через несколько лет мы станем свидетелями того, что вместо существующих методов сбора данных возникнут новые возможности. С учетом вышесказанного использование «больших данных» для раунда переписей 2020 года не представляется возможным, однако такое положение после 2024 года изменится.