Организация Объединенных Наций ECE/CES/2017/24

Экономический и Социальный Совет

Distr.: General
4 April 2017
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Шестьдесят пятая пленарная сессия
Женева, 19–21 июня 2017 года
Пункт 4 предварительной повестки дня
Следующее поколение статистиков и ученых по данным

Система производства статистических данных 4.0

Записка Статистического управления Швеции

Резюме

В этом документе рассматриваются требования в отношении квалификации и компетенции статистиков в условиях изменения производства статистических данных, а также роль официальных статистиков в этом контексте. Для обсуждения и определения этого следует рассмотреть системы производства статистических данных национальных статистических управлений. В этом документе также рассматриваются пользователи и их потребности в данных, а также вопрос о том, какие ответные меры могут принять статистические управления.

Настоящий документ представляется на рассмотрение семинара Конференции европейских статистиков на тему «Следующее поколение статистиков и ученых по данным».

GE.17-05383 (R) 260417 270417

1705383

I. История вопроса

1. Обследования на основе теории рандомизации были и остаются сегодня одним из основных источников для подготовки статистических данных в национальных статистических управлениях (НСУ). Одна из главных проблем в деле применения этой теории заключается в отсутствии ответов. Хотя теория предполагает полный объем ответов, в ходе практической работы НСУ некоторые единицы выборки не дают ответов. Проблема отсутствия ответов пока не нашла удовлетворительного решения (например, Brick, 2013) и сегодня является одной из основных угроз для действительности статистических данных, полученных в ходе выборочных обследований в силу высоких коэффициентов отсутствия ответов. Отчасти в силу этой проблемы и высокой стоимости выборочных обследований НСУ рассматривают «большие данные» в качестве альтернативного источника данных для подготовки статистических данных.

2. Существует множество сложных проблем в области использования «больших данных» для подготовки официальных статистических данных, однако «большие данные» также имеют много интересных особенностей (например, Daas et al., 2015; Japec et al., 2015). Сбор данных рассматривается как более дешевый и более быстрый метод по сравнению с традиционными выборочными обследованиями, и в некоторых случаях он может использоваться для подготовки статистических данных практически непрерывно. Считается, что при использовании «больших данных» основное внимание в ходе обследования переносится со сбора данных на анализ данных, при этом также проблема редактирования данных становится более сложной задачей (например, Tang, 2014).

3. Очевидно, что перевод процессов производства данных в НСУ с выборочного обследования на использование альтернативных источников данных предполагает необходимость в ином профиле компетенции. Что касается статистических знаний и навыков, акцент смещается со сбора данных на анализ данных, и статистики должны обладать навыками и профессиональными знаниями для проведения этих анализов.

4. Термин «большие данные» еще не был четко определен. В настоящее время некоторые НСУ в значительной степени полагаются на использование регистров и административных данных. Иногда «большие данные» рассматриваются в качестве административных данных, а иногда нет. Поэтому вместо «больших данных» будет использоваться термин «альтернативные источники данных», который определяется как другие источники данных, помимо вероятностных выборочных обследований и регистров, где реестры определяются как в Wallgren and Wallgren (2014).

5. Для того чтобы иметь возможность обсудить требования в отношении навыков и компетентности и роли статистиков, необходимо определить систему производства данных в НСУ. Существует также необходимость рассмотрения вопросов, касающихся пользователей и проблем, связанных с принятием ими решений, а также того, какую продукцию можно ожидать от НСУ.

II. Системы производства статистических данных 1.0–3.0

6. Прочные позиции теории рандомизации в НСУ могут объясняться свойствами получаемых статистических данных. Неспециалисту может показаться, что статистические данные являются однородными и их можно толковать как таковые. Это, разумеется, не соответствует действительности, и следует надлежащим образом учитывать, каким образом были получены статистические данные. На практике, к сожалению, и в академических кругах редко ставится вопрос «Где находится точка статистического умозаключения?». Ответ на этот вопрос может быть сформулирован как интервальные оценки, т.е. «Особенность статистического умозаключения…», если цитировать Chatterjee (2003, p. 30).

7. Статистическое умозаключение предоставляет интервальные оценки, которые не только обеспечивают оценку изучаемого параметра, но и содержат информацию о неопределенности оценок. Интервальные оценки приводятся в виде доверительных интервалов в рамках частотного подхода к толкованию вероятности. В настоящее время в рамках теории рандомизации доверительные интервалы, полученные в ходе выборочных обследований, имеют объективный характер, в том смысле, что они могут толковаться без ссылки на какие-либо предположения. Это особенно привлекательно в контексте роли НСУ в обществе. В практической работе необходимо применять определения относительно единиц обследования, народонаселения, рамок, переменных и т.д.

8. Статистические управления, использующие выборочные обследования в качестве основного источника данных, характеризуются в настоящем документе как имеющие систему производства статистических данных 2.0 (СП 2.0). Статистическое управление Швеции в 1950-х годах начало использовать вероятностные выборки и в последующие десятилетия разработало процесс производства. Система, использовавшаяся до 1950-х годов, называется СП 1.0.

9. В конце XX века административные данные государственных учреждений начали приобретать интерес для системы производства статистических данных на основе регистров. Административные данные имеют ряд преимуществ по сравнению с выборочными обследованиями для производства статистических данных (см. Wallgren and Wallgren, 2014), и в настоящее время большинство НСУ работают над созданием систем производства статистических данных на основе регистров. Такая система хорошо развита в Статистическом управлении Швеции и сегодня, и она имеет равное, если не более важное значение для производства статистических данных.

10. Несмотря на то, что Статистическое управление Швеции разработало как хорошо функционирующую систему производства статистических данных на основе регистров, так и хорошо функционирующую систему на основе выборочных обследований, они не используются с полной отдачей, поскольку статистические данные производятся главным образом при помощи одного из этих способов, а не путем их сочетания.

11. Wallgren and Wallgren (2014, Sec. 15.1.3) рассматривают системный подход в отличие от поляризации между обследованиями на основе регистров и выборочных обследований. Система на основе регистров может использоваться для подготовки статистических данных из регистров, но она также имеет важное значение для разработки выборочных обследований. Они (ibid) разрабатывают концепцию «выборочных обследований на основе регистров». Полное использование имеющихся регистров может улучшить качество оценок на основе выборочных обследований в плане актуальности и точности. Таким образом, статистическая СП 3.0 здесь определяется как система производства статистических данных на основе регистров с использованием данных как из регистров, так и из выборочных обследований. При подготовке проектов обследований используются одновременно и регистры и выборочные обследования.

III. Будущий контекст операций

12. В своем докладе Комиссия о будущем Швеции (2013, p. 244) заявляет: «Мы живем в неспокойном мире, в котором границы становятся менее важными, в котором страны становятся все более взаимозависимыми, в котором темпы перемен являются, пожалуй, более быстрыми, чем когда-либо ранее, в котором вследствие постоянно расширяющегося потока информации становится все труднее определить, что является наиболее актуальным и чему, таким образом, должно уделяться приоритетное внимание, и в котором наблюдается небывало жесткая глобальная конкуренция». Далее в контексте демократических задач она заявляет (ibid, page 258): «Другая проблема касается растущей сложности процессов принятия решений. Это объясняется в основном тем, что многие

проблемы и трудности, с которыми сталкивается Швеция, проходят через границы, будь то на местном, региональном или национальном уровне между Швецией и другими частями мира. В то же время необходимо скорее расширять трансграничное сотрудничество для решения проблем в будущем. Существует опасность того, что процесс принятия решений станет еще более сложным».

13. Статистические данные производятся для принятия решений, и общая идея этих заявлений подразумевает необходимость быстрого принятия решений в более сложных условиях, зачастую с международными аспектами. Подготовка статистических данных должна быть пригодной для принятия решений в будущем. Безусловно, одним из аспектов является периодичность и своевременность, и можно также ожидать роста спроса на сопоставимость и согласованность. Можно также ожидать, что потребуются новые и специальные статистические данные, для того чтобы пролить свет на новые возникающие проблемы, имеющие временный характер.

14. В течение последних двух десятилетий появились «большие данные», которые проявили себя как потенциальный новый источник для подготовки статистических данных. В будущем можно ожидать появления новых дополнительных источников данных и методов сбора данных. Опыт Статистического управления Швеции свидетельствует о росте объема источников административных данных в различных правительственных ведомствах. Можно ожидать, что эта тенденция продолжится и позволит выработать новые типы обследований. Это, разумеется, также верно и в отношении частного сектора, в котором имеются многочисленные источники, потенциально полезные для подготовки статистических данных. Одним из важных вопросов здесь является доступ к данным, который может быть сопряжен с юридическими проблемами, и владельцы могут быть не заинтересованы в предоставлении данных.

15. Для сбора данных в ходе выборочных обследований, как правило, используются вопросники, составленные в той или иной форме, либо интервью по телефону или личные беседы. Развитие технологий будет предлагать альтернативные средства сбора данных. Мобильные телефоны сегодня могут использоваться в ходе обследований привычных маршрутов. Операции по кредитным картам можно использовать для обследования бюджетов домашних хозяйств. К числу других примеров относятся подключение транспортных средств к Интернету и проект «Амазон прайм эйр». В будущем появятся новые технические разработки, и некоторые из них будут полезными для сбора данных. Таким образом, будет не только больше имеющихся источников данных, но будет также больше возможностей для сбора персональных данных.

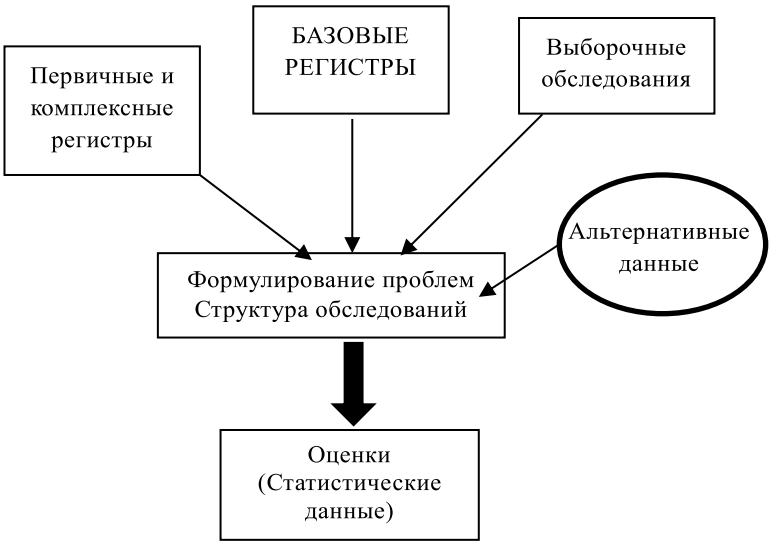
IV. Система производства статистических данных 4.0

16. С учетом вышеизложенного требование к будущим системам производства статистических данных можно охарактеризовать одним словом – гибкость. Гибкость в деле удовлетворения новых потребностей, гибкость при производстве статистических данных с быстрым распространением результатов, гибкость в сборе, извлечении и обработке данных из новых источников, гибкость в реорганизации обследования, когда появляются новые источники данных или когда исчезают старые, и гибкость в сообщении результатов. В нашем распоряжении имеются, по крайней мере, три источника данных: регистры, выборочные обследования и альтернативные источники данных (например, «большие данные»).

17. Схематическое представление системы на основе этих источников данных приводится на рис. 1. В центре этой системы стоят базовые регистры, например регистры населения и коммерческие регистры. Они имеют исключительно важное значение для значимости и ценность всей статистики. Они не только отслеживают единицы в составе населения, но и объединяют производство в плане категорий и определений, что сказывается на актуальности, точно-

сти, согласованности и сопоставимости. Слева располагаются вторичные регистры информации о единицах учета в базовых регистрах, например регистры занятости и доходов, связанные с регистром народонаселения, а также регистры налога на добавленную стоимость и внешней торговли, связанные с коммерческим регистром. Эти регистры представляют собой ресурсную базу системы и могут использоваться для получения статистических данных или как поддержка для других обследований на основе регистров или выборочных обследований. Некоторые статистические данные вырабатываются непосредственно на основе базовых регистров.

Рис. 1
Система производства статистических данных 4.0 в действии



18. Еще одним источником являются выборочные обследования, показанные справа. Есть две причины, по которым они показаны в двух квадратах. Первая заключается в самой методологии, а другая – выборочные обследования, которые были проведены ранее или которые будут проводиться. Круг справа означает альтернативные источники данных, например «больших данных».

19. Основной принцип сбора данных может показаться более простым в случае альтернативных источников данных по сравнению с традиционными методами сбора данных для выборочных обследований. Однако на практике существуют три основные проблемы, требующие рассмотрения. Первая из них связана с методами сбора, хранения и редактирования данных, вторая проблема связана с правовыми и этическими вопросами, а третья – с готовностью владельцев данных обмениваться данными. Когда альтернативные данные уже собраны, они должны быть обработаны и отредактированы, после чего проводится соответствующий анализ. Это может включать в себя гораздо более сложное статистическое моделирование, чем в настоящее время проводится в НСУ. Эти части также включают комбинирование и интеграцию данных с другими наборами данных.

20. Сейчас, когда все эти ресурсы имеются в наличии, можно использовать их в различных сочетаниях для производства статистических данных в той или иной области. Они могут быть основаны только на регистрах, исключительно на основе выборочных обследований, или же они могут использовать все ресурсы.

21. К сожалению, теорию и методологию выборочных обследований зачастую рассматривают как теорию и методологию для изучения в НСУ групп субъектов, вызывающих интерес, например групп граждан, домашних хозяйств, предприятий и т.д. Теория и методология выборочных обследований представляют собой общие инструменты и могут использоваться в специальных целях в

рамках альтернативных систем обследования с использованием как регистров, так и альтернативных источников данных.

22. Например, для статистики туризма часто используются открытые обследования, в которых непрямое обследование иностранных посетителей проводится на выборке пунктов перехвата, например в аэропортах, портах и дорожных пограничных контрольно-пропускных пунктах. Такие системы следуют стандартному формату определения единиц, представляющих интересов, группы единиц, структуры выборки и прямого сбора данных с помощью, например, собеседований. Если цель исследования заключается в том, чтобы представить данные для вторичных счетов, было бы достаточно провести оценку общего числа посещений и средних расходов на поездку. Это позволяет использовать различные системы, в которых составление выборки обследования играет иную роль. Так, например, для оценки общего числа посещений может использоваться сочетание данных мобильных устройств о местоположении и выборочных обследований, проводимых для проверки (калибровки). Аналогичным образом для оценки средних расходов может использоваться сочетание данных с кредитных карт и данных выборочного обследования. При таком методе выборочные обследования играют иную роль, объединяя элементы необходимой информации, которых не имеется в основном источнике альтернативных данных.

23. Интересное место в этой системе производства занимает «Формулирование проблем/структура обследований». Как свидетельствует пример выше, благодаря пересмотру формулировки проблемы могут появиться альтернативные методы, позволяющие более эффективно использовать имеющиеся ресурсы. Предпочтительно, но не всегда возможно, сделать еще один шаг назад и пересмотреть проблемы, лежащие в основе решения, лежащего в основе обследования. Часть «Формулирование проблем/структура обследований» будет иметь исключительно важное значение для успешного использования альтернативных источников данных.

24. Предварительным условием для системы производства, способной решать будущие потребности и вызовы, является более тесное сотрудничество между департаментами НСУ; специалисты по конкретным вопросам, методологии, специалисты по ИТ и другие должны сотрудничать друг с другом в поиске надлежащих решений в отношении структуры обследования. Сочетание знаний и навыков, имеющихся в различных департаментах, имеет важнейшее значение для определения и тестирования различных вариантов структуры обследования. Понимание потребностей пользователей и контекста определяет решение в отношении того, какие статистические данные необходимо получить. Понимание имеющихся ресурсов и методов их использования определяет, каким образом статистические данные могут быть получены. Необходимо признать, что этот процесс носит интерактивный характер.

V. Роль статистиков и ученых, работающих с данными, в будущем

25. С учетом темы семинара многие новые функции и компетенции будущих НСУ не охватываются в настоящем докладе. Их обсуждение здесь не проводится, но они признаются в описании СП 4.0.

26. Наука о данных и ученые, работающие с данными – это новые концепции, и можно найти несколько определений этих терминов. Они часто ассоциируются с научными исследованиями и с прикладными программами для ведения деловых операций. В этом докладе они также увязываются с процессом получения новых знаний (Boyd and Crawford, 2012). Основная роль НСУ заключается не в том, чтобы получить новые знания наравне с теориями и моделями, объясняющими реальные мировые явления. НСУ дает описание состояния дел в стране, например уровень безработицы, инфляции, ВВП и т.д. Таким образом, НСУ не нужно адаптировать новый подход к получению «знаний» или «пони-

мания», но они могут использовать научные инструменты и знания ученых, работающих с данными, для выявления скрытой структуры в крупных и смешанных наборах данных. Она в свою очередь используется для разработки обследования, дающего официальные статистические данные при соблюдении традиционных статистических критериев качества. Это показывает, что официальная статистика является надлежащей сферой применения науки о данных, поскольку действительность результатов, полученных благодаря основанному на модели выборочному обследованию, не зависит от «истинности» модели.

27. Как показано в описании СП 4.0, ключевым элементом является формулирование проблемы и разработка структуры обследований. В этой связи необходимо, чтобы специалисты разного профиля работали совместно в рамках «целевая группа», в которой ученые, работающие с данными, и статистики играют центральную роль. Основной особенностью их компетенции являются обширные знания в своих областях в отношении того, что может быть достигнуто с использованием различных методов и технологий. Необходима стратегия для обеспечения доступа к соответствующей компетенции.

28. В будущем статистические данные будут также производиться при помощи выборочных обследований при поддержке административных данных и данных из альтернативных источников. В других проектах выборка обследования будет использоваться для поддержки обследований, основанных на административных данных и данных из альтернативных источников.

29. Кроме того, в будущих системах производства потребуются статистические знания для обработки и использования административных данных, и, вероятно, они приобретут еще большее значение. Правительственные учреждения постоянно производят новые административные источники данных. В других государственных органах также имеются данные, которые не используются в настоящее время, но, возможно, будут представлять интерес в будущем. Кроме того, данные, собранные в организациях частного сектора, требуют навыков в работе с регистровыми данными.

30. Последний аспект, касающийся будущей роли статистиков и ученых, работающих с данными, заключается в коммуникации и поощрении пользователей. Традиционный способ публикации точечных оценок с указанием пределов ошибок является достаточным, если пользователь имеет достаточную подготовку в теории статистики. Обычно это не так, и пользователи не знают, что делать с погрешностью, и игнорируют ее. Отказ от толкования статистики в пределах традиционной статистической точности открывает возможности для альтернативных производителей на рынке, что позволяет получать менее дорогостоящие статистические данные, но такие данные гораздо менее качественны и потенциально могут причинить вред. В этой связи будет важно найти способы распространения статистических данных таким образом, чтобы пользователь мог правильно интерпретировать данные и в то же время получить знания о свойствах точности статических данных. Самая первая и простая мера состоит в том, чтобы заменить сочетание точечных оценок и пределов ошибок на доверительный интервал, назвав его «интервальная оценка».

VI. Справочная информация

Boyd, D. and Crawford K. (2012). *Critical Questions for Big Data*, Information, Communication & Society, 15:5, 662-679.

Brick, J.M. (2013). *Unit Nonresponse and Weighting Adjustments: A Critical Review*, Journal of Official Statistics, 29:3, 329-353.

Chatterjee, S.K. (2003). *Statistical Thought - A Perspective and History*, Oxford University Press, New York.

Commission on the Future of Sweden (2013). Future Challenges for Sweden, Prime Minister’s Office, Sweden, Ds 2013:19 (English translation).

Daas, P.J.H., Puts, M.J., Buelens, B. and P.A.M. van den Hurk (2015). *Big Data as a Source for Official Statistics*, Journal of Official Statistics, 31:2, 249-262.

Japec, L., Kreuter, F., Berg, M., Biemer, P., et al. (2015). *Big Data in Survey Research: AAPOR Task Force Report*, Public Opinion Quarterly, 79:4, 839-880.

Tang, N. (2014). Big data cleaning. In L. Chen, Y. Jia, T. Sellis, and G. Liu (eds) *Web technologies and applications, Lecture notes in computer science*, 8709, 13–24. Cham: Springer International Publishing.

Wallgren and Wallgren (2014). *Register-based Statistics – Statistical Methods for Administrative Data, 2nd Ed*, Wiley, New York.