Организация Объединенных Наций ECE/CES/2017/25

Экономический и Социальный Совет

Distr.: General
11 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

Статистики или ученые по данным? Будущее официальной статистики в эпоху новых технологий и современных источников данных[1, 2]

Записка Центрального статистического управления Израиля

Резюме

Значительные успехи в технологии и доступность больших данных выдвигают новые требования в плане более детализированной, более точной и более своевременной официальной статистики. Возникающие в этом отношении технологические и методологические проблемы, которые повлияют на характер производства официальной статистики в предстоящие годы, будут иметь серьезные последствия для работы национальных статистических управлений. В настоящем документе будут рассмотрены такие вопросы, как сбор и использование больших данных в целях подготовки официальной статистики, конфиденциальность и защита больших данных, повышение доступности данных, при одновременном сохранении частного характера и конфиденциальности, возможность использования веб-панелей, влияние способа сбора данных, использование данных социальных сетей и интеграция административных данных с данными обследований. Возникает вопрос, в какой степени университеты готовят студентов к работе в современных статистических управлениях. В настоящем документе рассматриваются некоторые из этих вопросов на основе национального и международного опыта.

Настоящий документ представляется для обсуждения участникам семинара Конференции европейских статистиков по теме «Следующее поколение статистиков и ученых по данным».

[1] На основе статьи «Методологические вопросы и проблемы, возникающие при составлении официальной статистики», подготовленной национальным статистиком Израиля профессором Дэнни Пфефферманном для двадцать четвертой ежегодной лекции им. Морриса Хансена и опубликованной в «Журнале статистики и методологии обследований» в декабре 2015 года. Помощник национального статистика Израиля г-н Йоэль Финкель отредактировал эту сокращенную версию.

GE.17-05876 (R) 050517 080517

1705876

I. Введение

1. Термин «официальная статистика» имеет широкое применение, однако его формальное определение отсутствует. В настоящем документе под официальной статистикой понимается любой набор публикаций национальных статистических управлений (НСУ), которые основаны на обследованиях, переписях, административных данных или их сочетании. Вместе с тем данное определение носит весьма ограниченный характер, поскольку в настоящее время ведется обширная исследовательская работа по использованию «больших данных» для производства официальной статистики. Большие данные, как правило, не являются результатом обследования и, как правило, гораздо больше, динамичнее и могут отображаться в самых разных форматах, по сравнению с данными традиционно относимыми к категории административных. Использование больших данных для производства официальной статистики является, вероятно, самой интригующей задачей, стоящей перед НСУ. Эта проблема будет обсуждена в последующих разделах настоящего документа.

2. Об официальной статистике люди наслышаны больше, чем о любом другом виде статистики. Каждый месяц до них доводят новые показатели безработицы, доходов и уровня бедности, индексов цен, достижений в области образования, статистики здравоохранения и окружающей среды и многие другие связанные с ними показатели. Для большинства людей официальная статистика и является настоящей статистикой. Более того, официальная статистика – это то, что политики используют (или должны использовать) для планирования и принятия решений, которые влияют на жизнь нашего общества. Когда центральный банк решает изменить процентную ставку, такое решение основано на официальной статистике. То же самое можно сказать и о решениях, касающихся государственного финансирования, строительства новых школ, социальных и медицинских программ, и даже о политических решениях. В этом случае очевидно, насколько важно иметь своевременную, достоверную официальную статистику по каждому аспекту нашей жизни. Тем не менее мир постоянно меняется, развиваются новые передовые технологии, в то время как бюджеты на соответствующие нужды постоянно сокращаются.

3. Цель настоящего документа состоит в том, чтобы обсудить то, что можно рассматривать как некоторые из основных методологических проблем, с которыми сталкиваются производители официальной статистики, а в некоторых случаях – предложить способы их решения. В настоящем сокращенном варианте статьи для Конференции европейских статистиков 2017 года рассматриваются следующие проблемы:

a) сбор и использование больших данных в целях подготовки официальной статистики;

b) интеграция компьютерной науки для подготовки официальной статистики на основе больших данных;

c) доступность данных, их частный характер и конфиденциальность;

d) интеграция статистики и геопространственной информации.

4. В связи с признанием огромного значения официальной статистики неизбежно возникает вопрос о том, готовят ли университеты студентов к работе в НСУ. Как будет отмечено в заключительной части настоящего документа, как правило это не так. Фактически, положение в этой области за последнее десятилетие по-видимому ухудшилось. В настоящее время лишь несколько университетов предлагают базовые курсы официальной статистики, например по выборкам или по другим вопросам. Это особенно тревожно, если учитывать, что НСУ входят в число крупнейших работодателей для экономистов и статистиков.

II. Сбор и использование больших данных для производства официальной статистики

5. Под «большими данными» обычно понимаются огромные массивы высокоскоростных масштабных данных, которые отличаются сложностью, переменчивостью с точки зрения структуры, источников и формата, но которым в то же самое время присуща некоторая неопределенность, влияющая на их достоверность (определение «пять V» (value, velocity, variety, veracity, volume), существует и определение «семь V»). Типичными примерами являются данные, собранные о человеческом геноме и мозге, данные, связанные с социальными сетями и торговлей по Интернету, показания спутников, сведения датчиков климата, данные об использовании мобильных телефонов и т.д. Серьезные проблемы, стоящие перед учеными в области использования и анализа таких данных, обсуждаются в многочисленных других публикациях[3]. И хотя в этих превосходных докладах о производстве официальной статистики почти не говорится, вполне очевидно, что НСУ не могут игнорировать потенциальные преимущества больших данных для статистики. В этом направлении уже осуществляются различные инициативы. Например, Статистическая комиссия в 2014 году учредила глобальную рабочую группу, уполномоченную «обеспечивать стратегическое видение, направленность и координацию глобальной программы использования больших данных для подготовки официальной статистики и поощрять практическое использование источников больших данных для официальной статистики» (ООН, 2014 год).

6. Ниже приведены некоторые важные аспекты возможного использования больших данных для производства официальной статистики с учетом проблем подсчетов и вопросов конфиденциальности, затрагиваемых в последующих разделах:

a) Тип данных. Важно различать данные, получаемые от датчиков, камер, сотовых телефонов, в виде изображений со спутников, которые, как правило, структурированы и точны и относятся к определенной категории или области, и данные, полученные из социальных сетей, от субъектов электронной торговли, веб-рекламы и т.д., которые очень разнообразны и неструктурированы, представляются нерегулярными и не имеют отношения к какой-либо категории. Как утверждается в документе Национального исследовательского совета (2013 год), структура (или состояние отсутствия таковой) может меняться весьма быстро, и НСУ должны быть готовы к такой возможности. В общем данные из разных источников могут кодироваться в разных форматах, поступая в разное время с разной степенью надежности и, возможно, будучи определены по-разному. Еще более тревожным является то, что некоторые виды больших данных могут неожиданно исчезнуть, что требует быстрых изменений в производстве статистических данных, основанных на этом источнике. Например, та или иная компания сотовой связи может внезапно разориться.

b) Опубликование. Общей чертой наборов данных в двух примерах, описанных выше, и многих других возможных наборов больших данных является то, что эти данные как правило имеются на каждый момент времени. В настоящее время официальные статистические публикации публикуются ежегодно, ежемесячно или могут относиться к определенной дате. В этой связи возникают три интересных вопроса:

[3] Хорошее резюме обсуждений содержит недавний доклад «Статистика и наука» (2013 год). В этом докладе кратко изложены лекции и дискуссии, имевшие место в ходе специального семинара по будущему статистических наук, который был проведен в Лондоне, Соединенное Королевство, в 2013 году с участием 100 приглашенных лиц, в рамках празднования «года статистики». Другой, еще более обширный (и носящий более технический характер) доклад – это документ Национального исследовательского совета (2013 год), подготовленный Национальной академией наук Соединенных Штатов Америки.

i) Какую статистику следует составлять и публиковать? Должны ли официальные публикации из источников больших данных, которые постоянно измеряются, в основном представляться в форме (онлайновых) графиков и изображений? НСУ уже используют хорошо разработанные инструменты визуализации данных, но их исходные данные обычно намного проще.

ii) Если предположить, что агрегированные (средние) оценки будут по-прежнему использоваться для планирования и принятия решений, каким образом будет осуществляться преобразование динамических (непрерывно измеряемых) входных данных, например в ежемесячные агрегированные показатели? Следует ли статистикам выбирать непрерывно измеряемые данные путем выборки или с использованием других, более сложных, методов?

iii) Представляется очевидным, что случайная выборка будет продолжать играть важную роль в эпоху больших данных, однако выборка из больших динамических данных будет отличаться от выборки из конечных категорий данных. Это потребует разработки новых алгоритмов выборки, не только уменьшающих объем хранилищ данных, но также создающих управляемые наборы данных, которые можно «прогнать» через алгоритмы для получения оценок и для которых могут быть адаптированы модели, например для решения проблемы выборки на основе социальных сетей. Должны ли НСУ использовать такие данные для подготовки официальной статистики – это уже отдельный вопрос[4]. Кроме того, использование выборки помогает обеспечить конфиденциальность (раздел 4.1 ниже).

c) Алгоритмическая оценка. Применительно к традиционной выборке, используемой при обследовании, различают сконструированный алгоритм оценивания, зависящий от модели алгоритм оценивания и использующий модель алгоритм оценивания. В последнем случае такой алгоритм выбирается на основе модели, но его свойства изучаются в рамках распределения случайной выборки. При использовании больших данных появляется новый класс алгоритмов оценивания, которые можно назвать алгоритмическими оценками. Эти алгоритмы оценивания получают в результате применения вычислительного алгоритма к исходным данным. Например, в Израиле постоянно запрашивают сведения о степени религиозности еврейского населения и просят представить демографическую и социально-экономическую информацию для различных секторов, определяемых такой характеристикой. В одной неопубликованной работе[5] говорится о том, что 12 различных административных файлов были объединены с регистром населения Израиля по состоянию на начало 2006 года. Эти файлы содержали около 6 млн. записей, при этом был применен сложный иерархический алгоритм, с помощью которого был присвоен балл религиозности в диапазоне [1,3] каждому лицу в регистре. Данный объединенный регистр охватывал приблизительно 95% лиц в возрасте от 0 до 64 лет.

d) Мера погрешности. НСУ стремятся определять меру погрешности (неопределенности) опубликованных статистических данных в виде стандартных ошибок или доверительных интервалов. Большие данные предположительно не содержат ошибок выборки (если только выборка не производится). Является ли по-прежнему мера погрешности проблемой в случае больших данных? Следует ли сконцентрировать внимание на измерении систематической ошибки и качественных показателях (ошибки измерения), а не на различиях? Как оценить систематическую ошибку? Должны ли статистики в определенный момент

[4] См. обсуждение в Национальном исследовательском совете (2013 год).

[5] Портной (2007 год), Центральное статистическое управление Израиля (ЦСУИ).

времени делать это, сравнивая алгоритмы оценивания с оценками, полученными в ходе традиционного обследования?[6]

e) Систематическая ошибка. Возможность возникновения значительной систематической ошибки является одной из основных проблем при использовании больших данных для производства официальной статистики. Систематическая ошибка с точки зрения охвата или выбора имеет место тогда, когда существующие данные не охватывают или должным образом не представляют всю изучаемую категорию. Например, цены продажи домов, рекламируемые в Интернете, явно не отражают всех продажных цен за указанный месяц (систематическая ошибка с точки зрения охвата). Если данные собираются таким образом, что предпочтение отдается более крупным субъектам (скажем, более крупным предприятиям), то систематическая ошибка возникает с точки зрения выбора. Мнения, выражаемые в социальных сетях, часто сильно отличаются от мнений общественности. Малозатратным способом борьбы с систематической ошибкой с точки зрения охвата, когда известно, что он существует, является пересмотр изучаемой категории. Например, категорию «дома на продажу» можно ограничить категорией «рекламируемые в Интернете дома», но будет ли это представлять какой-либо интерес? В других ситуациях сведения о наличии систематической ошибки с точки зрения покрытия или отбора могут отсутствовать, и, как упоминалось выше, потенциальным способом обнаружения и оценки такой ошибки будет сравнение оценок, полученных из источников больших данных, с не имеющими систематической ошибки оценками, полученными в ходе традиционных обследований (при условии, что они продолжают существовать).

f) Увязка данных. НСУ не только производят и публикуют сводные оценки на национальном уровне, но весьма часто производят оценки при значительно более высоких разрешениях, определяемых по возрасту, полу, этнической принадлежности, области проживания, виду отрасли и т.д. Вместе с тем имеющиеся большие данные могут не содержать всю эту информацию, для чего потребуется широкомасштабная привязка, если отсутствующая информация может быть получена из каких-либо других источников. Это, однако, свидетельствует о еще одном возможном недостатке больших данных в виде отсутствия идентификаторов, которые позволяли бы связывать разные файлы. Например, данные о покупках в супермаркетах не содержат никакой информации о покупателях, в отличие от данных, собранных в обследованиях расходов семьи. Единственными идентификаторами, которые могут связывать покупки с покупателями, являются номера кредитных карт, но будут ли компании-эмитенты кредитных карт предоставлять такие данные НСУ?

7. На основании вышеприведенного перечня можно сделать вывод о том, что использование больших данных для производства официальной статистики может потребовать новых методов, например для увязки данных, при их получении из разных источников. Кроме того, необходимо разработать новые методы редактирования и анализа, позволяющие обрабатывать доступную (возможно динамичную) информацию с достаточной скоростью и точностью, усовершенствованные методы визуализации и новые методы оценки и определения ошибок. Таковы лишь некоторые этапы на пути вперед.

8. В следующем разделе рассматриваются вопросы компьютерной инженерии и разработки программного обеспечения, без которых использование больших данных НСУ было бы невозможным. Раздел 4 посвящен обсуждению конфиденциальности данных и контролю за раскрытием информации.

9. В настоящем документе говорится о тех огромных проблемах, которые стоят перед учеными-компьютерщиками и статистиками в плане использования

[6] Более подробное обсуждение возможных ошибок измерений, связанных с большими данными, см. в документах Национального исследовательского совета (2013 год) и Американской ассоциации исследований общественного мнения (ААИОМ) (2015 год).

больших данных. С другой стороны, было бы безответственно игнорировать потенциальные преимущества больших данных для производства официальной статистики с точки зрения своевременности (в некоторых случаях возможно и в реальном времени) и гораздо большей универсальности, охвата и точности (хотя при этом возможна систематическая ошибка). Большие данные будут продолжать существовать при постоянно растущем объеме. Использование больших данных не требует разработки основы в виде выборки, вопросников, опросов и всех других необходимых компонентов выборочных обследований. В конечном итоге их использование может привести к значительному сокращению расходов. С учетом того факта, что доля ответивших на вопросники в рамках традиционных обследований постоянно снижается, использование больших данных в качестве альтернативного или дополнительного источника информации для официальной статистики представляется неизбежным.

III. Интеграция компьютерной науки для производства официальной статистики на основе больших данных

10. Огромный объем и разнообразие больших данных требует новых мощных аппаратных и программных технологий для хранения данных, а также для обработки и анализа данных, которые в настоящее время, как правило, не доступны НСУ. При решении вопроса о хранении данных на наших ноутбуках мы обычно оперируем гигабайтами (примерно 109 байт). Однако большие данные обычно измеряются в терабайтах (1 012 байт) или петабайтах (1 015 байт), при этом упоминаются также экзабайты (1 018 байт) и йоттабайты (1 024 байт). С тем чтобы просто дать представление о том, что все это означает, генеральный директор «Гугл» Эрик Шмидт как-то заявил, что каждые два дня мы создаем столько информации, сколько не было накоплено с момента начала цивилизации вплоть до 2003 года[7]. Это составляет около пяти экзабайт данных. Согласно сообщениям, сеть магазинов «Уолмарт» в США каждый час обрабатывает миллион операций, питая базу данных в 2,5 петабайта, что почти в 170 раз превышает объем данных, хранящихся в Библиотеке Конгресса США. Поэтому неудивительно, что «стандартные» аппаратные и программные инструменты баз данных не могут хранить, обрабатывать и анализировать такие большие данные. Более того, наблюдается феноменальный рост количества датчиков и машин, генерирующих данные. Примерами таковых, некоторые из которых уже упоминались, являются датчики погоды/загрязнения, дорожные и мобильные датчики и спутниковые системы.

11. На первый взгляд привлекательным решением для таких высоконагруженных вычислительных систем является использование облачной обработки данных для получения доступа к очень большим наборам данных и управления ими и поддержка мощных элементов инфраструктуры (хранилищ и вычислительных мощностей). На ее основе создаются виртуальные машины с огромным объемом памяти и вычислительной способностью. Такая архитектура состоит из совокупности виртуальных машин, которые позволяют обрабатывать данные путем разбивки на множество параллельных процессов. Пользователи (компании) могут использовать облачную инфраструктуру для своих сервисов больших данных, не загружая собственную инфраструктуру. Фактически, облачные пользователи не управляют облачной инфраструктурой и платформой, на которой выполняется приложение. Она является источником всего необходимого программного обеспечения, а также может хранить и обрабатывать звуковую информацию, что является привлекательным вариантом, когда речь идет о производстве официальной статистики.

12. Значение облачной обработки данных будет, вероятно, возрастать с точки зрения как обработки больших данных, как хранения и обеспечения доступа, а также анализа, хотя в настоящее время это все еще сопряжено с трудностями

[7] http://techcrunch.com/2010/08/04/schmidt-data/.

при передаче наборов больших данных. В этой связи представляется, что такая обработка данных может оказаться привлекательным способом использования больших данных НСУ, особенно с учетом возможного повышения производительности в тех случаях, когда несколько пользователей могут одновременно работать над одними и теми же данными. Вместе с тем это, в свою очередь, выдвигает на первый план серьезную проблему защиты данных. Теоретически, защита данных может быть улучшена путем централизации данных, однако при наличии нескольких пользователей и при распространении данных на более широкую область или на большое количество устройств, риск раскрытия данных определенно увеличивается. Вместо этого следует изучить возможность создания частной облачной среды (центра обработки данных), включающей все локальные вычислительные устройства для хранения данных и произведения распределенных вычислений при централизованном управлении. Это может стать одной из основных задач, которые НСУ предстоит решать в ближайшем будущем.

13. При изложении вышеупомянутых моментов были затронуты только самые общие проблемы в плане того, что необходимо для больших данных с точки зрения вычислений, и того, что может предложить современная вычислительная техника. Однако они указывают на то, насколько огромными будут требования к вычислительным ресурсам для НСУ, если планируется использовать большие данные в рамках обычной производственной системы, с точки зрения хранения, аппаратных средств, программного обеспечения, навыков вычислений, анализа данных и контроля за раскрытием информации.

14. Большинство видов традиционной информационной инфраструктуры, имеющихся в распоряжении НСУ, не рассчитано на такие требования, что предполагает приобретение новых надлежащих инфраструктурных средств. В этой связи важно упомянуть, что известные программные пакеты, которые регулярно используются НСУ, такие как SAS, SPSS и R, уже содержат некоторые программные процедуры, рассчитанные на использование больших данных, однако в том случае, если речь будет идти о производстве официальной статистики на основе больших данных, сотрудники всех уровней должны будут освоить новые навыки работы на компьютере. Несмотря на это, если будут использоваться услуги по облачной обработки данных, все, о чем говорилось выше, будет трактоваться по-иному. В этом случае основное внимание будет уделяться конфиденциальности и защите данных при надзоре и регулировании со стороны правительства. Кроме того, возникшие затраты могут оказаться и меньше. В любом случае потенциальное использование больших данных для производства официальной статистики в ближайшие годы несомненно потребует от НСУ значительных усилий. В ЦСУИ, как и во многих других странах, над изучением возможностей использования наборов больших данных, которые могут поступить в наше распоряжение, работает специальная целевая группа.

IV. Доступность, частный характер и конфиденциальность данных

A. Предисловие

15. НСУ находятся под постоянным давлением со стороны исследователей, лиц, принимающих решения, журналистов и широкой общественности, которые настаивают на предоставлении данных с высоким разрешением и, если возможно, индивидуальных данных. Это, конечно, вступает в противоречие с необходимостью защиты частной жизни и обеспечения конфиденциальности. Без поддержания доверия к таким органам невозможно провести ни одно исследование, и на соответствующую обязанность ссылаются в каждом письменном вопроснике и при любом опросе.

16. Существует два разных аспекта этой проблемы: защита данных от злоумышленников, также известная как «кибербезопасность», и гарантии того, что данные, переданные лицам за пределами НСУ, не могут быть использованы для раскрытия конфиденциальных данных частного характера.

a) Защита данных от злоумышленников связана с компьютерными технологиями, что является огромной проблемой, которая становится все более острой и которая, естественно, не ограничивается данными, хранящимися в НСУ. Мы обязаны регулярно приобретать новое компьютерное оборудование и связанное с ним оборудование для повышения степени защиты данных от злоумышленников. Как всегда в случае проблем такого рода через три–четыре года статистики могут снова услышать, что наши данные более не являются защищенными и что для обеспечения их защиты необходимы новые дорогие вычислительные устройства.

b) Второй аспект, известный как контроль за раскрытием статистических данных, привлекает внимание статистиков и компьютерщиков на протяжении многих десятилетий. В приведенном ниже изложении кратко описаны некоторые новые методы и связанные с ними показатели качества, применяемые в настоящее время, с упором на новые проблемы, обусловленные использованием больших данных.

B. Риски раскрытия информации

17. Традиционно НСУ публикуют результаты своей деятельности либо в виде микроданных, полученных в основном в ходе социальных обследований, либо в форме табличных данных с указанием частоты подсчетов или данных о размерах, которые обычно собираются в рамках обследований предпринимательской деятельности, таких как совокупные поступления. Проводились многочисленные исследования с целью количественной оценки риска раскрытия применительно к каждому из этих обычных результатов в рамках определенного метода контроля за раскрытием статистических данных, а также оценки воздействия примененного метода на полезность данных, например в плане того, содержат ли по-прежнему опубликованные данные информацию, необходимую для проведения исследований и принятия решений. Вполне очевидно, что чем лучше данные защищены от раскрытия, тем меньше их полезность, и наоборот.

18. Новый вид риска раскрытия информации – это логическое раскрытие, под которым понимается выявление новых признаков с высокой степенью вероятности. Например, регрессионная модель с очень высокой прогнозирующей способностью способна генерировать логическое раскрытие даже для лиц, не охваченных соответствующим набором данных. Другим примером логического раскрытия является раскрытие путем дифференцирования, когда на основе одного источника данных получают сразу несколько наборов. Например, таблицы переписи могут быть дифференцированы/обработаны для выявления данных по отдельным лицам. Такой вид раскрытия лучше всего контролировать, ограничивая его фиксированным набором переменных и категорий, тем самым не допуская дифференцирования не включенных в выборку групп лиц.

19. Близкой логическому раскрытию является концепция дифференцированной конфиденциальности, которую ученые-компьютерщики активно изучают для целей защиты результатов[8]. Дифференцированная конфиденциальность направлена на предотвращение получения информации на основе логического раскрытия путем обеспечения того, чтобы злоумышленник не мог узнать признаки конкретной целевой единицы в базе данных с высокой степенью вероятности в том случае, когда было изменено только одно значение в базе данных, а он располагает полной информацией о всех других единицах в базе данных (сценарий «наихудшего случая»). Такое довольно жесткое определение обеспе-

[8] Более подробно см. Динур и Ниссим (2003 год) и Дворк и др. (2006 год).

чивает контроль за раскрытием информации в результате дифференцирования или использования моделей с высокой степенью предсказуемости, что становится более проблематичным в связи с увеличением количества запросов к системам онлайновых запросов для распространения статистических данных по сравнению с использованием, как это было в прошлом, печатных материалов. Решение, предложенное учеными-компьютерщиками для обеспечения дифференцированной конфиденциальности, заключается в добавлении шума/помех к результатам запросов при определенных параметризациях, хотя это, конечно, снижает ценность данных для логических построений. В связи с этим постоянно изучаются другие средства защиты конфиденциальности, некоторые из которых рассматриваются в следующих подразделах, после чего следует краткое обсуждение.

C. Защита данных путем использования анклавов данных

20. За последние два десятилетия многие НСУ во всем мире создали у себя исследовательские (безопасные) помещения, также известные как анклавы данных. Анклав данных – это защищенная среда, в которой исследователи могут получить доступ к конфиденциальным данным. Защищенные серверы не подсоединяются к принтерам или Интернету, и доступ к ним имеют только прошедшие авторизацию пользователи. Никакие данные не могут быть удалены из такого анклава, и исследователи проходят соответствующую подготовку в целях ознакомления с правилами безопасности. Исследователям предоставляется статистическое программное обеспечение, такое как SAS, STATA или R, и весь поток информации находится под контролем. Все результаты, полученные из анклава данных, вручную проверяются на предмет риска раскрытия информации, например путем оценки числа малых ячеек, остаточных графических значений, которые могут указывать на отклоняющиеся показатели, или ядерной оценки плотности при малой ширине полосы.

21. Очевидными недостатками анклавов данных являются необходимость поездки исследователей в НСУ и дополнительное бремя для сотрудников НСУ в плане подготовки необходимых файлов данных и управления анклавом. В последнее время некоторые НСУ расширили концепцию анклавов данных, включив в нее дистанционный доступ через виртуальные анклавы данных. Такие виртуальные анклавы данных позволяют пользователям регистрироваться на защищенных серверах и получать доступ к данным со своего персонального компьютера, причем вся деятельность регистрируется и контролируется на уровне нажатия клавиш. Подобная защищенная лаборатория данных должна быть одобрена соответствующими учреждениями, а полученные результаты дистанционно изучаются сотрудниками по обеспечению конфиденциальности перед тем, как они будут отправлены исследователям в виде защищенного файла. Очевидно, что использование виртуальных анклавов данных требует более высокой степени доверия со стороны НСУ, при этом возможности для контроля оказываются более ограниченными, чем в случае «домашних» анклавов данных.

D. Статистический контроль за раскрытием данных для веб-приложений

22. Ориентируясь на спрос со стороны разработчиков политики и исследователей на рассчитанные на конкретных пользователей специализированные таблицы статистических данных и, в частности, на данные переписей, некоторые НСУ создали серверы для генерирования адаптивных таблиц, которые позволяют пользователям формировать и создавать свои собственные таблицы. Пользователи получают доступ к таким серверам через Интернет и строят необходимые таблицы, используя предлагаемые переменные и категории.

23. В целом существует два основных подхода в плане применения метода контроля за раскрытием статистических данных к таблице результатов: предшествующий построению таблицы подход и подход на основе уже построенной таблицы. В рамках первого подхода метод контроля за раскрытием статистической информации применяется к исходным данным, а полученные на их основе таблицы считаются безопасными для распространения. Второй подход предусматривает сначала получение исходных данных, а затем к таблице применяют метод контроля за раскрытием статистических данных. Применение такого последующего подхода в значительной степени строится на используемом в компьютерной науке определении дифференцированной конфиденциальности, речь о которой идет в разделе 4.2. Можно также использовать сразу оба подхода, хотя это способно привести к чрезмерной защите и, следовательно, к снижению полезности данных.

24. Для генерирования адаптивных таблиц сервер должен количественно оценить риск раскрытия в исходной таблице, применить метод контроля за раскрытием статистических данных и затем повторно оценить такой риск раскрытия. Очевидно, что риск раскрытия информации будет зависеть от того, получены ли исходные данные в результате переписи и нулевые значения являются реальными или же данные представляют собой результаты обследования, а нулевые значения носят случайный характер. После того, как защита таблицы будет обеспечена, сервер должен также рассчитать влияние применения метода контроля за раскрытием статистических данных на полезность этих данных путем сравнения «возмущенной» таблицы с исходной таблицей. На таком сервере, создающем таблицы, для оценки риска раскрытия информации и полезности данных могут быть использованы основанные на теории информации средства.

25. Устройство серверов для дистанционной генерации таблиц обычно предусматривает включение множества специальных предварительных правил контроля за раскрытием статистических данных, которые могут быть легко запрограммированы в пределах системы, с тем чтобы исключить таблицы, которые не должны передаваться. Такие правила контроля за раскрытием статистической информации могут включать в себя ограничение числа измерений в таблице, установление минимальных пороговых значений категорий, таких как средние размеры ячеек или количество мелких ячеек, обеспечивающих последовательные и вложенные категории переменных, с целью избежать раскрытия информации путем дифференцирования и т.д.[9].

26. Предшествующие созданию таблицы методы контроля за раскрытием статистических данных могут включать взаимную замену записей, при которой производится обмен признаками между двумя записями, имеющими схожие характеристики по набору контрольных переменных. Методы, используемые после создания таблицы, могут включать в себя «возмущение» ячейки, такое как случайное округление, или использование пострандомизационного метода, который предусматривает «возмущение» показателя количества ячеек на основе матрицы вероятностей переходов. Метод контроля за раскрытием статистических данных должен обеспечивать сохранение достаточных статистических сведений, таких как маргинальные совокупные значения, и поддержание согласованности между одними и теми же ячейками, сгенерированными в разных таблицах, с тем чтобы избежать необходимости восстановления данных после применения этого метода.

E. Серверы для дистанционного анализа

27. Сервер для дистанционного анализа – это онлайновая система, которая принимает запрос от исследователя, исполняет его в безопасной среде с исполь-

[9] Касательно правил и методов контроля за раскрытием статистических данных применительно к серверам для дистанционной генерации таблиц см. Шломо, Антал и Эллиотт (2015 год).

зованием соответствующих данных и возвращает конфиденциальный результат, при этом отсутствует необходимость вмешательства человека для ручной проверки результатов на риски, связанные с раскрытием информации. Как и в случае с генераторами адаптивных таблиц, запросы передаются через дистанционный интерфейс, и исследователи не имеют непосредственного доступа к данным. Такие запросы могут включать в себя анализ поисковых данных, определение показателей ассоциации, регрессионный анализ и статистическое тестирование. Их можно исполнять на основе исходных или конфиденциальных данных, а результаты могут быть ограничены и проверены в зависимости от требуемого уровня защиты[10].

F. Синтезированные данные

28. В последние годы НСУ предпринимают шаги по созданию синтезированных микроданных на основе моделей, которые сохраняют важные статистические свойства исходных данных. Синтезированные данные хранятся как общедоступные файлы. Производство синтезированных данных становится все более популярным, поскольку, как обсуждалось в разделе 4.3, доступ к реальным данным с дистанционных серверов может быть запрещен. В последнее время наблюдается тенденция к тому, чтобы дать исследователям возможность самостоятельно готовить свои работы и писать соответствующий программный код на основе синтезированных данных, а затем адаптировать программное обеспечение под реальные данные в такой защищенной среде, какой является анклав данных.

29. Для получения синтезированных данных в учреждении модель адаптируют под исходные данные, а затем синтезированные данные выбираются из соответствующего апостериорного распределения, по аналогии с методом множественного восстановления. Для получения значимых статистических оценок могут быть использованы несколько выборок синтезированных данных[11]. Синтезированные данные могут быть частично дополнены реальными данными, с тем чтобы опубликовать смесь реальных и синтезированных данных[12]. Однако следует отметить, что частично синтезированным наборам данных могут попрежнему угрожать риски с точки зрения раскрытия информации, которые необходимо проверить до их распространения. Если модели, используемые для статистического анализа, представляют собой подмодели модели, используемой для генерации данных, обязательно ли анализ синтезированных данных будет давать действительные логические результаты, при условии, конечно, что исходная модель является «правильной»?

30. Помимо этого, для табличных данных существуют методы, которые позволяют разрабатывать таблицы синтезированных величин, связанных со статистикой предпринимательской деятельности. В ходе контролируемой корректировки таблицы производится исключение определенных ячеек, заменяемых вмененными значениями, которые сохраняют определенные статистические свойства[13].

31. Должно ли использование синтезированных данных стать обычным способом защиты микроданных? Суть статистики заключается в том, что аналити-

[10] О'Киф и Гуд (2008 год) описывают регрессионное моделирование с использованием сервера для дистанционного анализа. О'Киф и Шломо (2012 год) сравнивают результаты на основе исходных данных и двух подходов к контролю за раскрытием статистических данных: результаты на основе конфиденциальных микроданных и конфиденциальные результаты, полученные из исходных данных через сервер для дистанционного анализа.

[11] Более подробную информацию и обсуждения см. Рейтер (2005 год) и Абоуд и Вильгубер (2008 год).

[12] См. Литтл и Лю (2003 год).

[13] См. Дандекар и Кокс (2002 год).

ки должны использовать реальные данные, а не данные, генерируемые моделью, хотя можно утверждать, что «возмущенные» данные также не являются «реальными данными». Крупный недостаток синтезированных данных заключается в том, что они полностью зависят от того, как модель адаптируется к исходным данным, а это является субъективной процедурой, при этом модель может не охватить все взаимосвязи между переменными, особенно в подкатегориях. Кроме того, трудно воспроизвести возможные аномалии в данных. Так, исходные данные могут содержать наблюдения, значения которых выходят за пределы отдельных наборов единиц (например, предприятий). Если синтезированные данные должны напоминать исходные данные, то и они в отношении аналогичных наборов единиц должны содержать наблюдения отклоняющихся значений с такими же масштабом и характеристиками. Сохраняется ли конфиденциальность данных даже в том случае, если наблюдения отклоняющихся значений несколько изменены? И наконец, что делать с большими данными? Собираемся ли мы генерировать множество наборов больших синтезированных данных, тем самым усугубляя проблему хранения и обработки больших данных в силу нескольких факторов?

G. Обсуждение

32. Существует очевидное противоречие между спросом со стороны исследователей на более подробную информацию и обязанностью НСУ защитить конфиденциальность респондентов. Такое противоречие порождает огромную потребность в методах контроля за раскрытием статистических данных, которые позволяют решить двойную задачу: с очень высокой степенью вероятности гарантировать конфиденциальность и сохранять полезность данных, передаваемых исследователям. Это повлекло тесное сотрудничество ученых-компьютерщиков, которые разрабатывают формальные определения рисков, связанных с раскрытием, в частности с раскрытием результатов логических выводов, и статистиков, разрабатывающих методы контроля за раскрытием статистической информации, которые гарантируют конфиденциальность. Использование таких методов предполагает, что исследователям придется иметь дело с «возмущенными» данными при проведении статистического анализа, что в свою очередь требует повышения уровня знаний о статистических выводах при наличии погрешности измерений. Существует очевидная необходимость дальнейшего изучения основанных на возмущениях методов контроля за раскрытием статистических данных, обеспечивающих полезность данных, что позволяет производить последовательную и объективную оценку статистических моделей.

33. А какова ситуация с большими данными? Проблема распространения данных становится все острее, что требует налаживания еще более тесного сотрудничества с компьютерными аналитиками. В первую очередь, нам придется иметь дело с гораздо большими объемами сложных данных высокой размерности, с многочисленными дополнительными переменными и категориями, по сравнению с традиционными обследованиями. Одним из способов является использование исключительно выборок, полученных на основе больших данных, причем не только в качестве средства уменьшения размеров данных, но и средства сохранения их конфиденциальности.

34. В отношении распространения больших данных НСУ существуют проблемы с точки зрения как государственной политики, так и этики, которые могут потребовать специального законодательства. При проведении обследования или переписи имеется четкое обязательство сохранять конфиденциальность данных. Однако такое обязательство отсутствует применительно к большим данным, собранным с помощью датчиков, через компании мобильной связи или социальные сети. Должны ли будут компании, собирающие большие данные, передавать их НСУ? Будет ли общественность согласна с передачей личных данных исследователям, а затем с возможным распространением, даже при условии применения методов контроля за раскрытием статистической инфор-

мации? Кроме того, в связи с возможностью доступа к нескольким наборам данных, а также их увязки повышается вероятность нарушения конфиденциальности. Это особенно верно в отношении наборов данных, отслеживающих деятельность человека, например при идентификации одного и того же лица в нескольких источниках социальных медиа.

V. Интеграция статистических данных и геопространственной информации

35. Использование географических информационных систем (ГИС) позволяет придать собранным данным пространственное измерение и, следовательно, получить о них иное представление[14]. В качестве общеизвестного примера можно привести карты бедности, выпускаемые Всемирным банком и другими организациями, каждый географический регион которых окрашен различными цветами, представляющими различные уровни бедности. Другим примером является карта дорожно-транспортных происшествий. В этом случае каждый участок дороги окрашивается в соответствии с числом таких происшествий. Важным преимуществом этого вида указывающих на «горячие точки» карт является не только то, что на них с первого взгляда видно, какие географические местоположения (регионы, участки дороги и т.д.) требуют дополнительного внимания, но также и то, что они показывают пространственное сходство между соседними местоположениями, если таковое существует. Иными словами, они преобразуют дискретные оценки в пространственно-временной процесс.

36. Использование ГИС имеет много других важных преимуществ:

- оно улучшает структуру выборочных обследований путем определения границ пластов, отбора ячеек и т.д. Оно также предоставляет всю необходимую информацию для районированной выборки;
- оно обеспечивает эффективное распределение квот выборки для проводящих опросы лиц и построение оптимальных навигационных маршрутов для прибытия в выбранные места;
- использование ГИС позволяет отслеживать такие явления, как изменение во времени социально-экономических условий отдельных лиц или семей, проживающих в данной местности, и увязывать их с географическими характеристиками, такими как расстояние от большого города, возможности для поездок на работу и с работы и перемещение в другие районы;
- ГИС весьма существенно повышает разрешение данных, что позволяет изучать и создавать новые группы (кластеры), которые не были известны ранее.

37. Быстрое развитие технологий открывает путь для сбора новых (больших) данных в очень высоком разрешении, при этом использование ГИС осуществлять очень точную привязку таких данных к географическим местам.

38. До этого момента в настоящем документе отмечалось, что в будущем статистик должен будет не только изучить классическую статистическую теорию, но и пройти подготовку в области информатики и кибербезопасности. К числу трех дополнительных важных вопросов для НСУ, с последствиями для будущих статистиков, относятся: возможность использования веб-панелей для производства официальной статистики; преодоление различий в используемых форматах при проведении обследований в смешанном режиме; сочетание административных данных с оценкой малых районов. Весьма подробно эти вопросы обсуждаются в исходном документе, и в настоящем укороченном его варианте сложные статистические методы, представленные в статье, не излагаются.

[14] Эта тема была более подробно освещена Майклом Гудчайлдом в ходе лекции им. Морриса Хансена в 2006 году (Гудчайлд, 2007 год).

VI. Готовят ли университеты студентов к работе в национальных статистических управлениях?

A. Предисловие

39. Во введении был поднят вопрос о том, готовят ли университеты студентов к работе в НСУ. Как правило, это не происходит, однако о потребности в более обширной и более качественной подготовке фактически говорилось в рамках некоторых других форумов, при этом в данной области наблюдаются определенные положительные моменты. Отправной точкой для обсуждения этого вопроса является то, что никто не ставит под сомнение важность НСУ и других подобных организаций и что НСУ относятся к числу крупнейших работодателей для статистиков и экономистов.

40. В настоящем разделе будут обсуждены следующие три основные темы в работе НСУ и рассмотрен вопрос о том, в каком объеме они преподаются в университетах.

1. Выборочное обследование

41. Вполне очевидно, что в этой статье не нужно подробно объяснять значение выборочного обследования для работы НСУ. Невозможно разработать опрос, отредактировать (очистить) исправленные данные на отклоняющиеся значения и неполученные сведения и подготовить надлежащие расчеты и оценки погрешности без хорошего знания теории выборочного обследования. Ниже приводится учебный план одногодичного курса по проектированию и анализу выборочных обследований, предлагаемого Гарвардским университетом (три академических часа в неделю):

Методы проектирования и анализа выборочных обследований; инструментарий построения функций выборки и их использование в оптимальных стратегиях проектирования; методы определения весов выборки и оценки дисперсии, включая методы повторного отбора; краткий обзор нестатистических аспектов методологии обследования, таких как проведение обследования и вопросники.

2. Сезонная корректировка (СК) и оценка тенденций

42. Для изучения тенденций и обнаружения изменений (поворотные точки) в социально-экономической деятельности используются временные ряды социально-экономических данных. Вместе с тем такой процесс обучения оказывается невозможным в том случае, если наблюдаемый временной ряд охватывает не только интересующий компонент трендов-циклов, но и сезонные изменения, воздействие количества рабочих дней, «плавающие» праздники и нерегулярные влияния. В том случае, если планируется изучить соответствующую тенденцию, эти дополнительные компоненты должны быть оценены и удалены из наблюдаемого ряда. Первую оценку тренда получают путем вычитания сезонного воздействия, что известно как сезонная корректировка. В литературе для сезонной корректировки и оценки тенденции было предложено несколько модельно-зависимых и непараметрических процедур, которые используются на регулярной основе. Многие НСУ практикуют публикацию скорректированных на сезонные колебания или отражающих тенденции рядов параллельно с первоначальной (наблюдаемой) изучаемой серией. Разница между этими двумя отражающими компоненты сериями заключается в том, что ряд, учитывающий сезонные колебания, содержит также отклоняющиеся значения, которые сглаживают для оценки тренда. Отражающий тренд ряд является более сглаженным, но на конечном участке он может скрывать важные поворотные точки. И наоборот, учитывающий сезонные колебания ряд может демонстрировать ложные поворотные точки.

3. Национальные счета

43. Значительная часть деятельности НСУ посвящена производству высококачественной экономической статистики. Примерами таковой являются национальные счета, платежный баланс, статистика государственных финансов, статистика цен, статистика международной торговли, сателлитные счета (производство и потребление энергии, социальное обеспечение, образование и т.д.), а также эколого-экономические счета. Система национальных счетов (СНС) является одним из наиболее важных систем рядов и ведется в соответствии со строгими международными стандартами. Она дает пользователям и принимающим решения лицам всестороннее представление об экономической деятельности и ее эволюции. В Израиле, как и во многих других странах, аналитики, принимающие решения лица, средства массовой информации и широкая общественность с нетерпением ожидают публикацию новых оценок НС каждый раз, когда они выходят. В этой связи можно было бы предположить, что такой важный базовый материал макроэкономической статистики будет преподаваться на каждом факультете экономики всех университетов. Но так ли это на самом деле?

B. Ответ на поставленный вопрос

44. Приобретают ли студенты в университетах базовые знания, необходимые для работы в этих трех важных областях? С тем чтобы ответить на этот вопрос мы просмотрели учебную программу подготовки бакалавров и аспирантов в области статистики и экономики 25 лучших университетов мира в соответствии с Шанхайским академическим рейтингом университетов мира http://www.shanghairanking.com/ (см. приложение). Вот что мы выяснили:

45. Только в 11 из 25 лучших университетов в той или иной форме предлагается вводный курс по выборочным обследованиям.

46. И хотя почти все университеты предлагают один или несколько курсов по теме «Анализ и прогнозирование временных рядов», отсутствуют курсы, в рамках которых достаточное количество времени уделялось бы изучению сезонной корректировки или оценке тренда. Фактически в описании только трех посвященных временным рядам курсов упоминается вопрос сезонности.

47. Еще реже встречаются специализированные курсы по национальным счетам. В лучшем случае национальные счета иногда упоминаются в рамках макроэкономических курсов, в основном в связи с валовом внутренним продуктом (ВВП). Как представляется, только Международный валютный фонд и его партнеры организуют комплексные курсы по национальным счетам в дополнение к нескольким магистерским программам по официальной статистике (см. ниже).

48. В связи с такими выводами возникает вполне естественный вопрос: какова на самом деле роль государственных и частных университетов? Предназначены ли они вообще для того, чтобы готовить своих студентов к трудовой деятельности, или же они должны сосредоточиться исключительно на научных исследованиях и подготовке новых поколений исследователей? Эта дискуссия продолжается уже много веков и, вполне очевидно, выходит за рамки данного документа. Тем не менее необходимо рассмотреть следующие моменты:

a) упомянутые выше три темы и многие другие темы, лежащие в основе работы НСУ, требуют обширных знаний в теории статистики и экономики. Выборочные обследования, сезонная корректировка и оценка тенденции предполагают знакомство с углубленными темами теоретической статистики вместе с очень важными приложениями, в силу чего преподавание курсов по этим темам не противоречит той точке зрения, что университеты должны сосредоточиться на исследованиях. Некоторые ведущие ученые мира, занимающиеся вопросами математической статистики, участвуют в исследованиях по оценке ма-

лых районов, что является еще одним очень важным видом продукции НСУ. Приводимый далее список включает в себя несколько примеров тем, лежащих в основе работы НСУ, которая основана на классической теории статистики: методы выборочных обследований, адаптация моделей к данным комплексных обследований, процедуры увязки данных, контроль за раскрытием статистических данных, выделение сигнала на основе моделей АРИМА, оценка среднеквадратической ошибки (СКО) сезонно скорректированных оценок, использование методов повторной выборки для уменьшения систематической ошибки и оценка дисперсии;

b) возможно, что причиной малого числа курсов по темам, связанным с работой НСУ, является отсутствие исследователей-экспертов для их преподавания. По этой причине во время учебы студенты не знакомятся с выборкой обследования и другими важными проблемами, лежащими в основе работы НСУ, и, следовательно, они не рассматривают эти проблемы в своих работах или позже в ходе научных исследований;

c) примерно за последние десять лет статистика значительно изменилась, и на смену тому, что известно как «классическая статистика», приходят новые сложные, требующие значительных компьютерных ресурсов методы анализа больших данных и подобных явлений. Изменились также и выборочные обследования и анализ временных рядов, при этом вполне очевидно, что курсы по таким темам должны быть перестроены с учетом современных разработок в этих и других областях, что, как мы надеемся, повысит их привлекательность;

d) в конечном итоге немало университетов в ряде стран в своей преподавательской и исследовательской деятельности делают упор на выборочные обследования. Кроме того, существует несколько университетских программ подготовки магистров в области официальной статистики. Двумя такими известными программами являются Совместная программа в области методологии обследования Соединенных Штатов и компании «Вестат» и магистерская программа в области официальной статистики Университета Саутгемптона в Соединенном Королевстве. Первая представляет собой результат сотрудничества Университета штата Мэриленд и Университета штата Мичиган, а вторая спонсируется Бюро национальной статистики (БНС) Соединенного Королевства. Национальный институт статистики и экономических исследований (НИСЭИ) во Франции и Бразильский институт географии и статистики (БИГС) имеют школы (колледжи) статистики, которые предлагают специальные программы для получения степеней бакалавра и магистра в области официальной статистики;

e) наконец, Европейский союз недавно принял решение об учреждении европейской магистерской программы в области официальной статистики (EMOC), при этом уже направлены заявки на заключение договоров с европейскими НСУ и университетами. Европейская магистерская программа в области официальной статистики (EMOC) представляет собой сеть магистерских программ послевузовского образования в области официальной статистики на европейском уровне. EMOC является совместным проектом университетов и производителей данных в Европе. После двукратного приглашения заявить о своем интересе участниками этой сети стали более 20 программ в 14 странах;

f) EMOC была создана в целях укрепления сотрудничества в рамках научного сообщества и производителей официальной статистики. Она помогает готовить специалистов, способных работать с европейскими официальными данными на разных уровнях в рамках быстро меняющейся системы производства данных в условиях XXI века. Степень магистра EMOC базируется на результатах обучения, в ходе которого выпускники вузов знакомятся с системой официальной статистики, моделями производства данных, статистическими методами и способами распространения информации;

g) предлагающие степень магистра EMOC университеты активно сотрудничают с НСУ, с тем чтобы уменьшить разрыв между теорией и практикой.

Таким образом, академические учреждения все больше признают значение производства официальной статистики, и эта тенденция может распространиться и на другие университеты.

VII. Список литературы[15]

[Только на английском языке]

AAPOR (2010). *Report on online survey panels*. http://poq.oxfordjournals.org/content/early/2010/10/19/poq.nfq048.full.html?

AAPOR (2015). *Report on big data.* American Association for Public Opinion Research. http://www.aapor.org/AAPORKentico/AAPOR_Main/media/Task-Force-Reports/BigDataTaskForceReport_FINAL_2_12_15.pdf

Abowd, J.M., and Vilhuber, L. (2008). *How protective are synthetic data?* In: PSD'2008 Privacy in Statistical Databases, (Eds. J.Domingo-Ferrer and Y. Saygin). Springer LNCS 5262, 239-246.

Cavallo, A., and Rigobon, R. (2010). *The Billion Prices Project@MIT*. (http://bpp.mit.edu).

Cavallo, A. (2012). *Online vs official price indexes: measuring Argentina's inflation*. Journal of Monetary Economics, 1-14.

Chaudhuri, S., Handcock, M.S. and Rendall, M.S. (2010). *A conditional empirical likelihood approach to combine sampling design and population level information.* Technical report No. 3/2010, National University of Singapore, Singapore, 117546.

Couper, M.P. (2000). *Web surveys, a review of issues and approaches*. Public Opinion Quarterly 64, 464-494.

Couper, M.P. (2008). *Designing Effective Web Surveys.* Cambridge University Press.

Daas P.J.H., Puts M.J. Buelens B. and van den Hurk, P.A.M. (2013*). Big Data and official statistics. Proceedings of the NTTS*, Euro Stat, Brussels. http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_76.pdf

Dandekar, R.A., and Cox L.H. (2002). *Synthetic tabular data: an alternative to complementary cell suppression.* Manuscript, Energy Information Administration, U. S. Department of Energy.

De Leeuw, E. (2005*). To mix or not to mix? Data collection modes in Surveys*. Journal of Official Statistics, 21, 1-23.

Dillman, D.A., and Christian, L. (2005). *Survey mode as a source of instability in response across surveys.* Field Methods, 17, 30-52.

Dinur, I., and Nissim, K. (2003). *Revealing Information While Preserving Privacy*. PODS 2003, 202-210.

Dwork, C., McSherry, F. Nissim, K. and Smith, A. (2006). *Calibrating Noise to Sensitivity in Private Data Analysis. In Theory of Cryptography TCC* (eds. S. Halevi and R. Rabin). Heidelberg: Springer, LNCS 3876, 265-284.

Fay, R.E., and Herriot, R. (1979). *Estimates of income for small places: an application of James–Stein procedures to census data.* Journal of the American Statistical Association, 74, 269–77.

Feder, M., and Pfeffermann, D. (2015). *Statistical inference under non-ignorable sampling and nonresponse- an empirical likelihood approach.* Southampton Statistical Sciences Research Institute, http://eprints.soton.ac.uk/id/eprint/378245

Goodchild, M.F. (2007). *The Morris Hansen Lecture 2006: Statistical perspectives on social science.* Journal of Official Statistics, 23, 1–15.

Hartley, H.O., and Rao, J.N.K. (1968). *A new estimation theory for sample surveys*. Biometrika, 55, 547-557.

[15] Перечисленные ссылки относятся к статьям, опубликованным в «Журнале статистики и методологии обследований», декабрь 2015 года.

Lee, J., and Berger, J.O. (2001). *Semiparametric Bayesian Analysis of Selection Models*. Journal of the American Statistical Association, 96, 1397-1409.

Lee, S. (2006). *Propensity score adjustment as a weighting scheme for volunteer panel web surveys.* Journal of Official Statistics, 22, 329-349.

Lee, S., and Valliant, R. (2009). *Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment*. Sociological Methods & Research, 37. 319-343.

Little, R.J.A., and Liu, F. (2003). *Selective multiple imputation of keys for statistical disclosure control in microdata*. The University of Michigan Department of Biostatistics Working Paper Series. Working Paper 6.

National Research Council (2013). *Frontiers in Massive Data Analysis*. Washington D.C.: The National Academies Press. (http://www.nap.edu)

New Zealand (2012). *Using cellphone data to measure population movements*. (http://www.stats.govt.nz/tools_and_services/earthquake-info-portal/using-cellphone-data-report.aspx).

Nirel, R. and Glickman, H. (2009*). Sample surveys and censuses.* In: Handbook of Statistics 29A. Sample Surveys: Design, Methods and Application Eds., D. Pfeffermann and C.R. Rao. Amsterdam: North Holland, 539-565.

O'Keefe, C.M. and Good, N. (2008*). A remote analysis server – What Does Regression Output Look Like?* In PSD'2008 Privacy in Statistical Databases, (Eds. J.Domingo-Ferrer and Y. Saygin), Springer LNCS 5262, 270-283.

O'Keefe, C.M. and Shlomo, N. (2012). *Comparison of Remote Analysis with Statistical Disclosure Control for Protecting the Confidentiality of Business Data*. Transactions on Data Privacy, 5, 403-432.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd edition). New York: Cambridge University Press.

Pfeffermann, D. (2013*). New important developments in small area estimation*. Statistical Science, 28, 40-68.

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998). *Parametric distributions of complex survey data under informative probability sampling*. Statistica Sinica, 8, 1087-1114.

Pfeffermann, D. and Landsman, V. (2011*). Are private schools better than public schools?* Appraisal for Ireland by methods for observational studies. The Annals of Applied Statistics, 5, 1726–1751.

Pfeffermann, D., Moura, F. A. S. and Nascimento-Silva, P.L. (2006). *Multilevel modeling under informative sampling.* Biometrika, 93, 943-959.

Pfeffermann, D. and Sikov A. (2011). *Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information.* Journal of Official Statistics, 27, 181-209.

Pfeffermann, D. and Sverchkov, M. (1999). *Parametric and semi-parametric estimation of regression models fitted to survey data.* Sankhya, 61, 166-186.

Pfeffermann, D. and Sverchkov, M. (2003). *Fitting generalized linear models under informative probability sampling.* In: Analysis of Survey Data, eds. R. L. Chambers and C. J. Skinner, New York: Wiley, pp. 175-195.

Pfeffermann., D. and Sverchkov, M. (2007). *Small area estimation under informative probability sampling of areas and within the selected areas*. Journal of the American Statistical Association, 102, 1427-1439.

Rao, J.N.K. (2003). *Small Area Estimation*. Wiley, Hoboken, NJ.MR1953089

Reiter, J.P. (2005). *Releasing multiply imputed, synthetic public-use microdata: an illustration and empirical study.* Journal of the Royal Statistical Society, A, 168, 185-205.

Rivers, D. (2007). *Sampling for web surveys*. Joint Statistical Meeting, Proceedings of the Section on Survey Research Methods, Salt Lake City, UT, USA.

Rosenbaum, P.R. and Rubin, D.B. (1983). *The central role of the propensity score in observational studies for treatment effects.* Biometrika, 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984). *Reducing bias in observational studies using subclassification on the Propensity score.* Journal of the American Statistical Association, 79, 516-524.

Rotnitzky, A. and Robins, J. (1997). *Analysis of Semi-Parametric Regression Models With Non-Ignorable Non-Response*. Statistics in Medicine, 16, 81-102.

Shlomo, N., Antal, L. and Elliot, M. (2015). *Measuring disclosure risk and data utility for flexible table generators*. Journal of Official Statistics, 31, 305–324.

Smith T.M.F. (1994). *Sample surveys 1975-1990; an age of reconciliation*? International Statistical Review, 62, 5-19.

Statistics and Science (2013). *A report of the London workshop on the future of the statistical sciences.* http://www.worldofstatistics.org/wos/pdfs/Statistics&Science-TheLondonWorkshopReport.pdf

Sverchkov, M., and Pfeffermann, D. (2004). *Prediction of finite population totals based on the sample distribution.* Survey Methodology, 30, 79-92.

UN (2014). *Report of the global working group on big data for official statistics.* United Nations, E/CN.3/2015/4. http://unstats.un.org/unsd/statcom/doc15/2015-4-BigData-E.pdf

Waksberg, J. and Goldfield, E. D. (1996). Morris Howard Hansen, 1920-1990. *A biographical memoir.* National Academy of Sciences, Washington D.C., U.S.A.

Vaccari, C. (2014). *Big Data in Official statistics. School of advanced studies*, University of Camerino, Italy. https://www.academia.edu/7571682/PhD).

Vannieuwenhuyze, J.T.A; Loosveldt, G and Molenberghs, G. (2014). *Evaluating mode effects in mixed-mode survey data using covariate adjustment models*. Journal of Official Statistics, 30, 1-21.