Большие Данные, Официальная Статистика и Некоторые Инициативы Австралийского Бюро Статистики

Сиу-Минг Там и Фредерик Кларк (Siu-Ming Tam and Frederic Clarke)

Отдел методологии и управления данными
Австралийское бюро статистики

ABS House, 45, Benjamin Way, Belconnen, ACT 2615, Австралия

E-mail: Siu-Ming.Tam@abs.gov.au

Перевод: Статкомитет СНГ

Аннотация

Специалисты, работающие в системе официальной статистики, на протяжении многих десятилетий имели дело с разнообразными источниками данных. Однако новые источники информации, связанные с использованием больших данных, дают возможность статистикам более эффективно оказывать услуги по производству статистической информации. В статье приведены некоторые соображения, которые должны учитывать статистики при принятии решения относительно использования какого-то источника больших данных для регулярного производства официальной статистики. Основными критериями являются релевантность, польза для бизнеса и обоснованность использования источника для производства официальной статистики, получения оценок изучаемой совокупности или аналитических выводов. В статье также описан Флагманский проект по использованию больших данных Австралийского бюро статистики (АБС), осуществление которого даст возможность АБС приобрести практический опыт в оценке экономических, статистических, технических, вычислительных и других вопросов, рассмотренных в данной статье. Кроме того, участие АБС в национальных и международных мероприятиях, связанных с использованием больших данных, также поможет поделиться опытом и знаниями, а сотрудничество с представителями научного сообщества поможет АБС развить способности для решения своих производственных задач, используя большие данные в качестве одного из подходов.

Ключевые слова: стандарт качества данных**,** конфиденциальность, общественное благо, обоснованность статистических выводов.

1. Вступление

Недавние обсуждения в открытой печати о возможностях, предоставляемых *большими данными*, охватили сферу официальной статистики. Среди недавних значительных событий можно отметить обсуждение Статистической комиссией ООН статьи, озаглавленной «Большие данные и модернизация официальной статистики» (2014), и принятие Схевенингенского меморандума по вопросу «большие данные и официальная статистика» главами европейских статистических служб (Евростат, 2013). Поскольку в официальной статистике для производства официальной статистической информации долгое время использовались административные данные и бизнес данные, которые являются одними из многих источников больших данных, то специалисты в целом осторожно – что можно понять - относятся к распространению этих принципов на другие виды больших данных.

Что такое большие данные? В данной статье мы определяем большие данные с точки зрения официальной системы статистики как собирательный термин для все более широкого диапазона источников данных, которые становятся доступными в «сети, где найдется все». Сами по себе большие данные не являются хорошо определенным материальным объектом, и потенциальное использование больших данных для решения проблем, действительно, зависит от того, каковы эти проблемы, каков может быть вклад больших данных в решение проблемы, и являются ли какие-то внутренние искажения или ошибки измерения в этих источниках достаточно существенными, чтобы данные были непригодны для использования.

Почти всегда широкое обсуждение больших данных сосредоточено на обсуждении информационных и коммуникационных технологий (ИКТ), при этом в основном внимание уделяется вычислительной инфраструктуре, системам и методам, необходимым для эффективной реализации *«трех V»*: объёма (*Volume)*, скорости (*Velocity*) и многообразия (*Variety)* возникающих источников больших данных. На языке официальной статистики это означает повышение технологического потенциала Национальных статистических служб (НСС) для извлечения, хранения, обработки и анализа больших данных в целях производства статистической информации. В ходе этих дискуссий возникает ряд существенных вопросов для официальной статистики, которые обсуждаются ниже в порядке возрастания их важности.

Во-первых: являются ли *технологии больших данных* достаточно зрелыми для того, чтобы национальные статистические службы могли осуществлять инвестиции в них? В широко используемом хайп-цикле[1] компании Гартнер (Rivera and van der Meulen, 2013), где оценивается зрелость новых технологий, большие данные в 2013 году помещены на «пик завышенных ожиданий». Считается маловероятным, что эта технология достигнет «плато продуктивности» в течение ближайших пяти лет.

Во-вторых: в чем состоит возможная выгода от использования *больших данных* в официальной статистике в дополнение к административным данным и некоторым типам бизнес данных? Тогда как исследовательский анализ новых источников больших данных, безусловно, может быть полезен в принципе, утверждение о том, что

[1] http://habrahabr.ru/post/198506/
В 1995 году исследовательская компания Gartner предложила hype cycle — кривую зрелости технологии, графически представляющую стадии, через которые проходит технологическое новшество в ходе своего становления.

производитель статистических данных будет регулярно приобретать такие массивы данных без явной производственной необходимости равносильно решению по поиску проблем. НСС, которые испытывают все большее бюджетное давление, не готовы инвестировать в разработку больших данных, если отсутствует сильный стимул для инвестиций.

Наконец, как можно использовать большие данные для производства надежных статистических продуктов? Кроуфорд (Crawford, 2013) считал, что «скрытые искажения, как на стадии сбора, так и на стадии анализа, представляют собой значительные риски и столь же важны для уравнения с большими данными, как и сами цифры…» Предположение о том, что большие массивы данных каким-то образом находятся ближе к «истине», не принимается статистиками, поскольку объективная «истина» сильно зависит от того, насколько репрезентативен конкретный источник больших данных для изучаемой совокупности, а также от характера статистических выводов на основе этих данных. Другие вопросы, касающиеся использования больших данных, рассмотрены в работе Даас и Путс (Daas ad Puts (2014)).

Несмотря на эти вопросы и проблемы, мы считаем, что большие данные, семантическая статистика (Clarke and Hamilton, 2013), и трансформация бизнеса на основе статистической оценки (HLG BAS, 2012; Pink et al, 2009; и Tam and Gross, 2013) – это три наиболее многообещающие инициативы для радикальной трансформации будущей бизнес-модели и информационного влияния НСС. Для представителей официальной статистики задача в области больших данных состоит в том, чтобы обнаружить и использовать эти нетрадиционные массивы данных, которые могут расширить или заменить существующие источники, для эффективного производства официальной статистической информации для конкретных целей. Некоторые международные и национальные статистические организации уже начали исследовать потенциал больших данных (UN Statistical Commission, 2014; Eurostat, 2013).

Цели этой работы состояли в следующем:

- описать некоторые концепции больших данных и выразить озабоченность относительно ценности для бизнеса, методологической обоснованности и технологических возможностей использования больших данных для производства официальной статистической информации;
- представить план статистических мероприятий АБС по оценке случаев, в которых можно использовать определенные виды больших данных для замены существующего источника данных, создания новых статистических продуктов, или повышения эффективности деятельности Австралийского бюро статистики (АБС)

2. Определение, использование и источники больших данных

Что такое большие данные? В отчете о конфиденциальности больших данных (Podesta et. al., 2014), говорится: « … существует много определений больших данных, которые отличаются в зависимости от того, являетесь ли вы специалистом по компьютерным наукам, финансовым аналитиком или венчурным предпринимателем…».

Википедия дает следующее определение: «… широкий термин для обозначения любых массивов данных, таких больших и сложных, что их трудно обрабатывать с

использованием имеющихся инструментов управления базами данных или традиционных приложений для обработки данных» (Wikipedia, 2014).

В качестве определяющих характеристик для больших данных отмечают следующие (Daas and Puts, 2014):

- Объем (Volume) – количество записей данных, их атрибуты и связи.
- Скорость (Velocity) – как быстро прирастают и изменяются данные и как быстро они должны быть получены, обработаны и осмыслены;
- Многообразие (Variety) – разнообразие источников, форматов, носителей и содержания данных.

Кто является пользователями больших данных? Маниика и др. (Manyika et. al., 2011) считают, что:

> «…существует пять направлений, где использование больших данных может принести пользу. Во-первых, значительная польза может состоять в том, что информация станет более прозрачной и готовой к использованию с большей частотой. Во-вторых, поскольку организации создают и хранят больше данных об операциях в цифровой форме, они могут собирать более точную и подробную информацию о деятельности - от запасов товарно-материальных средств до пропуска дней по болезни - и таким образом демонстрировать разнообразие и повысить производительность. В-третьих, использование больших данных позволит проводить более узкую сегментацию клиентов и, следовательно, создавать продукты, лучше соответствующие конкретным потребностям в товарах или услугах. В-четвертых, опытные аналитики могут существенно улучшить процесс принятия решений. И наконец, большие данные могут быть использованы для совершенствования разработки следующих поколений продуктов и услуг».

Другое потенциальное преимущество использования больших данных состоит в обеспечении более регулярной и своевременной информации по интересующим вопросам, например, для раннего выявления эпидемий (например, *Google Flu Trends*), выявление экономических подъемов и спадов, роста безработицы или роста покупок жилья и проч., благодаря более низким удельным затратам по приобретению источников больших данных, по сравнению с традиционными методами прямого сбора данных, используемым НСС. Прекрасный пример приведен в работе Вариан и Чой (Varian and Choi (2011)), которые изобрели термин "nowcasting" для описания процесса предсказания настоящего на основе информации из *Google Trends*. В блоге издания Washington Post, Муи (Mui (2014)) говорит о том, что распространенность статистики, предлагаемой большими данными - имеющейся в наличии в качестве побочного продукта от других операций по сбору информации – и то, как данные могут быть «добыты» по конкретным интересующим вопросам, являются преимуществами больших данных по сравнению с традиционными источниками. С другой стороны, Харфорд (Harford, 2014) считает, что хотя выявлять корреляцию на основе больших данных дешево и просто, но корреляция – на что всегда стараются указывать статистики – это не то же самое, что причинная связь, и «… теоретически не обоснованный анализ простых корреляций неизбежно является очень хрупким».

С точки зрения официальной статистики большие данные можно определить как источники статистической информации, включающие традиционные источники и новые источники данных, которые становятся доступными в сети. Объемы и скорость больших данных огромны и поэтому пока выходят за рамки текущих возможностей управления данными и обработки данных, мы считаем, что НСС не обязательно должны использовать полные наборы этих данных для производства официальной статистики, поскольку выборочные методы могут применяться для производства статистической информации, удовлетворяющей поставленным целям.

Хотя не все разнообразные виды больших данных подходят для производства официальной статистики, у них есть потенциал для увеличения эффективности затрат НСС, разработки новых статистических продуктов и услуг и увеличения частоты производства официальных статистических данных при небольших дополнительных затратах для НСС. Большие данные могут предоставить возможность НСС лучше выполнять свою миссию по производству официальных статистических данных для принятия обоснованных решений. Однако мы утверждаем, что решение о том, какой источник больших данных использовать, включая решения по поводу объемов, скорости и разнообразия, должны оцениваться с учетом критерия затрат и выгод, который будет описан ниже в данной работе.

В целом, источники возникновения больших данных можно классифицировать следующим образом (UN Statistical Commission, 2013):

- источники данных, связанные с осуществлением программы, будь то государственной или иной, например электронные медицинские карты, ведомости приема клиентов больничными учреждениями, учетные страховые документы, учетные банковские документы и продовольственные банки. В официальной статистике данные из государственных источников традиционно считались административными данными;
- коммерческие или операционные источники данных, связанные с совершением операций между двумя сторонами, например, операции по кредитным карточкам и онлайновые операции (в том числе совершаемые с помощью мобильных устройств);
- источники данных, связанные с работой сенсорных сетей, например данные от изображений, полученных со спутников, данные с автодорожных датчиков и метеорологические данные от измерительных устройств;
- источники данных, связанные с работой регистрирующих устройств, например регистрация данных из сети мобильной телефонной связи и из Глобальной системы определения координат (GPS);
- источники данных, связанные с поведением пользователей, например, данные поиска в Интернете (по тому или иному продукту, услуге или по любому другому виду информации) и данные о просмотрах веб-страниц;
- источники данных, связанные с выражением пользователями своих мнений, например данные из комментариев в социальных сетях

Переписи и обследования, а также первые два источника, то есть административные данные и, в ограниченной степени, данные бизнеса (например,

отсканированные данные из супермаркетов, данные о продаже автомобилей и пр.) в настоящее время являются главными источниками для производства официальной статистики. Большие данные открывают возможности для НСС для получения новых источников данных.

Некоторые виды больших данных являются идентифицируемыми, например данные спутникового зондирования с географическими координатами точек, а многие другие – неидентифицируемыми, например, цены на товары и услуги в интернете, данные сканирования или коммерческих сделок. Как идентифицируемые, так и неидентифицируемые данные имеют свои области использования в официальной статистике, например, данные спутникового зондирования могут быть объединены с данными, предоставляемыми фермерами в рамках сельскохозяйственных обследований на уровне единиц обследования, тогда как данные о ценах на товары при онлайн покупках могут быть использованы для получения относительных цен для использования при расчете индекса потребительских цен. Задача для системы официальной статистики состоит в том, чтобы найти эффективные и обоснованные пути для использования источников больших данных, где их применение для целей регулярного формирования официальной статистики является оправданным.

3. Большие данные и официальная статистика

Многие национальные статистические службы в мире, включая АБС (Австралийское Бюро Статистики, 2013), обладают значительными знаниями и опытом в области сбора и обработки больших массивов данных, например:

- в соответствии с законом имеют право добиваться предоставления данных поставщиками информации для целей производства официальной статистики;
- являются уполномоченными интеграторами чувствительных данных в соответствии с законами о статистике;
- обладают уникальной возможностью оценивать качество и «репрезентативность» источников больших данных;
- в состоянии производить статистические данные высокого качества, так что пользователи могут быть уверены, что используемая ими информация соответствует целям, и
- являются независимыми, обеспечивают высокие требования к безопасности данных и непредвзяты. Большинство НСС публикуют описания концепций, источников, методов и результаты всех сборов данных и предоставляют равноправный доступ для всех пользователей официальной статистики.

Наряду с высоким уровнем доверия пользователей к органам официальной статистики (см., например, ABS, 2010b), эти качества обеспечивают хорошие позиции НСС для экспериментирования и изучения потенциального использования больших данных.

4. Возможности и задачи для официальной статистики

Чтобы продолжать повышать ценность предложения статистических данных, многие НСС стремятся снизить стоимость производства статистики, улучшить своевременность и частоту предложений данных, а также создавать новые или более богатые статистические данные, которые соответствуют возникающим потребностям

пользователей. В рамках своих программ трансформации бизнеса некоторые НСС (например, АБС, ЦБС Нидерландов, ИСТАТ и др.) предпринимают инициативы для изучения возможностей, предоставляемых большими данными.

Мы считаем, что некоторые области применения больших данных могут быть определены, если провести параллели с хорошо развитыми направлениями использования административных данных в официальной статистике при условии, что источники отвечают критериям выгодности и являются статистически значимыми, как отмечено в данной статье. Эти направления использования включают:

- создание основы выборки или регистра – определение единиц обследования и/или предоставление вспомогательной информации, например, о переменных для стратификации;
- полное замещение данных – замена сбора данных через обследование;
- частичное замещение данных для подгруппы совокупности – уменьшение объема выборки;
- частичное замещение данных для некоторых элементов данных – сокращение продолжительности обследования или обогащение массива данных без необходимости статистической увязки;
- импутация пропущенных элементов данных – замещение на ту же или подобную единицу;
- редактирование – помощь в обнаружении и трактовке аномалий в данных обследования;
- увязка с другими данными – создание более богатых наборов данных или временной перспективы;
- сопоставление данных – обеспечение значимости и согласованности данных обследования;
- создание новых аналитических подходов – улучшение измерения и описания экономических, социальных и экологических явлений.

Хотя основное внимание уделяется использованию больших данных, в основном, для формирования более разнообразных и своевременных статистических продуктов, большие данные, сформированные в самой организации, могут также быть использованы для повышения эффективности статистической деятельности в НСС (Groves and Heeringa, 2006), а именно:

- улучшения опыта поставщиков и потребителей данных;
- повышения эффективности экономической деятельности; и
- мониторинга безопасности интернета и сетей и сетевого поведения пользователей.

5. Выгода для бизнеса

Решение об использовании конкретного источника больших данных для формирования статистических данных должно приниматься строго в соответствии с производственной необходимостью и на основании будущих выгод, оцененных для каждого случая: как это может улучшить результаты статистической деятельности с точки зрения объективных критериев «затраты-выгоды», то есть при сравнении затрат при использовании нового источника данных с такими выгодами, как снижение нагрузки на поставщика данных, устойчивость нового источника, а также точность, релевантность, согласованность, интерпретируемость и своевременность

статистических результатов в соответствии с требованиями стандарта качества данных (ABS, 2010a; Brackstone,1999; OECD, 2011).

В качестве примера можно оценить полное замещение данных, полученных в ходе обследований, на данные спутникового зондирования для производства показателей сельскохозяйственной статистики, таких как растительный покров и урожайность:

- Затраты – Каковы вероятные затраты на приобретение, очистку и приведение данных спутникового зондирования в форму, пригодную для последующей обработки органами официальной статистики, отмечая что потребности в вычислительных мощностях для приобретения, передачи, обработки, интегрирования и анализа больших массивов изображений в настоящее время неизвестны, но вероятно будут снижаться со временем? Каковы затраты на разработку статистической методологии для трансформации этих данных в показатели урожайности и на развитие статистической системы для обработки и распространения данных спутникового зондирования? Каковы эквивалентные затраты на прямой сбор данных, и как они сравнимы друг с другом?

- Сокращение нагрузки на поставщика данных – Насколько сократится нагрузка на респондентов, если прямой сбор данных будет заменен использованием данных спутникового зондирования? Насколько важно добиться этого сокращения нагрузки, учитывая существующий опыт поставщиков данных и государственную политику по сокращению бюрократического регулирования? Какова в настоящее время степень кооперации фермеров и насколько вероятно, что она изменится к лучшему или к худшему в будущем?

- Устойчивость получения статистических результатов. Доступен ли источник данных органам статистики для регулярного формирования официальных статистических данных? Насколько вероятно, что этот источник в будущем исчезнет?

- Точность, релевантность, согласованность, интерпретируемость и своевременность. – Каков новый источник данных по сравнению с текущим источником при оценке по критериям, указанным в стандарте качества данных? (Australian Bureau of Statistics, 2010a; Brackstone,1999; OECD, 2011). Хотя спутниковое зондирование обеспечивает точное измерение «отражения» - измерение света, отраженного от объектов - имеются пропущенные данные с миссий спутников Landsat 7 (см. ниже) и облачного покрова. Являются ли эти проблемы более серьезными или менее серьезными, чем пропущенные данные при прямом сборе данных? Кроме того, для трансформации данных об отражении в показатели производства сельскохозяйственных культур статистикам необходимо применять научное или статистическое моделирование, что далеко не всегда практикуется в НСС и может вызвать вопросы к интерпретируемости данных. Поскольку данные спутникового зондирования доступны раз в две недели, у них, безусловно, имеется явное преимущество перед данными, собираемыми напрямую один или два раза в год, с точки зрения частоты формирования статистических показателей об урожайности культур.

6. Значимость статистических выводов

Массивы, полученные с использованием больших данных, не обязательно являются случайными выборками из целевой совокупности. Основанные на дизайне выборки статистические выводы, принятые большинством НСС для оценки параметров конечной совокупности, таких как средние, итоги и квантили, получают на основе случайных выборок, то есть механизм отбора не зависит от значений единиц, не попавших в выборку (Sarndal и др., 1977; Kish, 1965; Rubin, 2006); или статистических моделей для корректировки или трактовки ошибок отбора при неслучайной выборке (Puza and O'Neill, 2006).

В качестве примера можно привести социальные сети (такие как Twitter), которые являются богатым источником данных для измерения общественного мнения. Однако имеется мало проверяемой информации относительно пользователей этих услуг, и сложно определить, являются ли профили пользователей репрезентативными для населения в целом. Следует ожидать, что некоторые подгруппы населения будут недостаточно представлены в любой выборке данных из социальных сетей в связи с разной степенью освоения новых технологий. Таким образом, оценки мнения населения, формируемые на основе таких источников, без корректировки могут содержать искажения (Smith, 1983).

Вообще говоря, поскольку НСС являются хранителями и распорядителем большого количества разнообразных статистических фондов, то только они имеют возможность оценить репрезентативность совокупности, представленной большими данными. В некоторых случаях большие данные могут нуждаться в дополнении данными обследований для обеспечения охвата сегментов совокупности, не представленных в выборке. В других случаях может быть полезно, публиковать статистические данные, которые описывают подгруппы. Здесь проблема может состоять в том, что статистический анализ больших, сложных гетерогенных наборов данных неизбежно выявляет значительно больше ложных и зависящих от модели корреляций, чем можно ожидать при использовании традиционных источников данных. Это может усилить любое искажение, связанное с моделированием, подталкивая к выбору неправильных переменных, алгоритмов и параметров.

Приведем в качестве примера приложение *Google Flu Trends*, где используется количество поисковых запросов в качестве меры распространения гриппа в общей популяции и где были приведены ошибочные оценки, что пик заболеваемости гриппом достиг уровня 11% от населения США в эпидемическом сезоне 2012 года. Это было почти в два раза выше официальной оценки в 6%, опубликованной органами здравоохранения. Служба *Google Trends* объяснила такую завышенную оценку тем, что «… повышенное внимание средств массовой информации к тяжести заболеваемости гриппом в данном сезоне привело к тому, что пользователи направляли запросы по ключевым словам в течение более долгого времени, и мы посчитали, что это связано с уровнями заболеваемости» (Google Trends, 2013). Это указывает на важность оценки того, при каких условиях и для каких приложений использование больших данных требует или не требует корректировки для получения статистических оценок такого же уровня качества как официальная статистика, регулярно публикуемая НСС.

Купер (Couper (2013)) указывает на некоторые существенные стороны, которые следует учитывать аналитику, использующему большие данные для своих выводов;

среди них: ошибки охвата, ошибки выборки (репрезентативности), ошибки измерения и ошибки ответов.

7. Конфиденциальность и доверие общественности

Ситуация с конфиденциальностью данных кардинально изменилась в связи с появлением больших данных. Существует очевидное разногласие между систематическим использованием больших данных, где это оправдано, для поддержки принятия решений органами государственного управления и признаваемой необходимостью установить и поддерживать общественное доверие к использованию персональных данных государственными учреждениями.

Национальные статистические службы в своей деятельности руководствуются законодательством о статистике, там же закреплены их полномочия по сбору данных. Эти законы устанавливают основные правила по получению, объединению, защите, распространению, представлению, анализу и сохранению таких данных. Законодательство и соответствующие стратегии призваны содействовать укреплению доверия и конфиденциальности, а использование источников больших данных будет дальнейшей проверкой наших решений в плане приверженности этим положениям.

Важным нерешенным вопросом является угроза раскрытия информации в результате накопления данных. Каждый человек предоставляет уникальную мозаику публично видимых характеристик и частной информации. В богатом информационном мире некоторые элементы данных, которые не представляют риска для конфиденциальности сами по себе, могут раскрыть частную информацию при их объединении – ситуация, известная в разведывательном сообществе как «мозаичный» эффект. Использование больших данных в большой степени усиливает мозаичный эффект, поскольку большие богатые массивы данных обычно содержат много видимых характеристик, так что сами по себе или в комбинации могут привести к спонтанному узнаванию отдельных лиц и соответствующему раскрытию их частной информации. Это будет важным вопросом при распространении массивов микроданных из источников больших данных.

8. Владение данным и доступ к данным

Владение данными и доступ к данным являются ключевыми вопросами для НСС, причем эта область обычно недостаточно регулируется на законодательном уровне. Задача состоит в том, чтобы извлечь общественное благо из данных, собранных частным образом, при этом защищая коммерческие интересы хранителей данных. Во многих случаях владельцы первичных и производных неофициальных массивов данных придают им коммерческую ценность, поскольку либо производство таких данных является частью их бизнеса, либо обладание ими составляет существенный элемент их конкурентного преимущества. Возникает вопрос, как НСС может приобретать чувствительные данные, имеющие коммерческую ценность, для своего статистического производства, особенно если эти статистические продукты конкурируют непосредственно с информационными продуктами, создаваемыми собственниками данных, или же они ставят под угрозу их позиции на рынке. Этот вопрос еще более усложняется в связи с тем, что может быть несколько сторон с

некоторой формой коммерческих прав в отношении массива данных, таких как собственность, владение или лицензионное соглашение.

Большая часть контента в сети не структурирована и неуправляема: метаданные, описывающие его использование и источники (происхождение, получение, история, хранение и контекст) являются либо неполными, либо несоответствующими. И в самом деле, долгосрочная надежность источников больших данных может быть проблемой для продолжающегося статистического производства. Авторитетная статистика для выработки политики и оценки услуг обычно требуется на продолжительные периоды времени, часто на много лет. Однако большие массивы данных из динамичных сетей изменчивы: источники данных могут изменить свою природу и исчезнуть со временем. Такая быстротечность потоков и источников данных подрывает надежность статистического производства и публикацию значимых временных рядов.

9. Эффективность расчетов

Использование больших данных окажет существенное влияние на требования к ресурсам ИКТ для приобретения, хранения, обработки, интеграции и анализа данных. Существующие вычислительные модели для наиболее общих статистических задач в типичной НСС очень плохо масштабируются с учетом количества, разнообразия и изменчивости элементов данных, а также характеристик и связей источников больших данных.

В частности, традиционные подходы использования реляционных баз данных не являются достаточно гибкими для обработки динамичных и по-разному структурированных массивов больших данных для эффективного проведения вычислений, а выполнение сложных статистических алгоритмов для того масштаба задач, который характерен для использовании больших данных, вероятно, потребуют объемов памяти и ресурсов процессоров, превышающих те, которые обеспечиваются существующими платформами. Например, вероятностная увязка данных в модели Феллеги-Сантера (Fellegi & Sunter, 1969) обычно трактуется как задача условного максимального правдоподобия с использованием алгоритмов симплекс-метода. Сложность этой задачи составляет, по крайней мере, $O(N^3)$, и ее невозможно решить на базе существующих вычислительных мощностей, когда размер массива данных N определяется масштабом больших данных.

Один из возможных подходов состоит в том, чтобы аналитикой занимался владелец данных на условиях аутсорсинга. Статистическое управление Новой Зеландии собирается сделать так с данными сканирования, поскольку собственник данных обладает необходимой вычислительной инфраструктурой, а проводить анализ в месте хранения данные дешевле и легче. Дополнительное и важное преимущество такого подхода состоит в том, что собственнику данных не нужно делиться первичными данными, которые могут быть очень чувствительными. Необходимы совместные усилия методологов и технических специалистов для разработки методов сокращения объема и сложности данных при сохранении их обоснованности, а также для улучшения разрешимости и эффективности алгоритмов. Это будет включать явное пере формулирование существующих задач в формы, более подходящие для применения распределенных расчетов, более широкое использование методов аппроксимации и применение эвристических моделей прогнозирования в соответствующих обстоятельствах.

10. Технологическая инфраструктура

Технология использования больших данных возникла на базе огромных объемов обработки данных в интернете и за последнее десятилетие получила применение во все большем количестве областей деловой жизни. Поддерживаемые отраслью разработки технологий с открытым исходным кодом быстро достигли такого уровня зрелости, что обработка «корпоративного класса» - в сочетании с традиционными технологиями обработки данных – предоставляет более интегрированный набор технологических возможностей. Независимые и «точечные» решения по использованию больших данных сокращаются, поскольку они интегрируются в более широкую архитектуру решений. Большинство поставщиков технологий теперь включают решения, связанные с большими данными, в свой продуктовый портфель. Инфраструктура и инструментарий для использования больших данных постоянно развиваются, и в будущем будут продолжать существовать собственные и независимые решения.

Обработка больших данных также требует новых подходов к представлению данных (семантические данные и графовые базы данных), выводов (аналитические методы на основе искусственного интеллекта в сочетании с надежным статистическим анализом), визуализации (для сложных сетевых взаимоотношений), аналитических языков (такие как R и SAS) и использования масштабируемых стандартных аппаратных средств. Некоторые из этих технологий имеют значение при применении в "традиционных" методах обработки и анализа.

11. Инициативы АБС в области использования больших данных

Австралийское национальное централизованное статистическое агентство – Австралийское бюро статистики (АБС) производит официальные статистические данные по широкому кругу социальных, демографических, экономических и экологических тем для поддержки информированного принятия решений, проведения исследований и обсуждений с органами государственного управления и общественностью. Основными законодательными актами, определяющими функции и ответственность АБС, являются *Закон об Австралийском бюро статистике 1975 года* и *Закон о Переписях и статистике 1905 года*.

В настоящее время разрабатывается несколько инициатив для построения будущего потенциала использования источников больших данных и для обеспечения позиций АБС как внутри страны, так и на международной арене в качестве ведущего агентства в использовании новых методов анализа данных.

11.1 Стратегия использования больших данных

Для обеспечения позиций АБС в деле развития возможностей, предоставляемых большими данными, в АБС был подготовлен документ о стратегии в области использования больших данных (Australian Bureau of Statistics, 2014), который был одобрен руководством АБС. Целью Стратегии является создание интегрированного многогранного потенциала для систематического использования возможной ценности больших данных для производства официальной статистики.

Этот потенциал включает:

• квалифицированных работников, которые в состоянии интерпретировать информационные потребности и доносить идеи, почерпнутые из богатых данных;

• передовые методы, инструменты и инфраструктура для представления, хранения, обработки, интегрирования и анализа больших сложных массивов данных;

• разнообразные виды государственных, частных и открытых источников данных, имеющихся для использования в статистических целях;

• безопасный и адекватный открытый доступ к массивам микроданных и статистических решений, полученных на основе ряда источников данных; и

• тесное мульти дисциплинарное сотрудничество между органами государственного управления, отраслями, академическим сообществом и статистиками.

11.2 Флагманский проект по использованию больших данных

Флагманский проект АБС по использованию больших данных – это инициатива методологов АБС, и она направлена на координацию усилий по проведению исследований и разработок, которые позволят построить крепкую методологическую базу для широкого использования больших данных для производства и анализа статистических данных. Желаемые результаты осуществления проекта таковы:

• способствовать лучшему пониманию концепций, возможностей, практических вопросов и проблем больших данных в АБС;

• поощрять методологическую строгость в использовании различных источников больших данных для статистического производства;

• разработать продуктивные подходы к изучению, объединению, визуализации и анализу больших, сложных и изменчивых массивов данных;

• and развивать тесные связи с сетями экспертов в области больших данных в органах государственного управления, промышленности, научных кругах и международном статистическом сообществе;

• улучшить национальные и международные позиции АБС в области использования больших данных для получения статистических результатов.

В рамках проекта запланированы следующие виды работ:

• анализ ситуации и возможностей – изучение источников больших данных на предмет потенциального использования в статистическом производстве, выявление проблем и «болевых точек», для анализа которых можно было бы использовать нетрадиционные источники данных и аналитические методы;

• использование удаленного зондирования для производства сельскохозяйственной статистики – исследование возможности применения данных спутниковой съемки для производства показателей сельскохозяйственной статистики, таких как использование земли, тип культур и урожайность культур;

• использование данных локализации мобильных устройств для анализа мобильности населения – исследование возможностей использования сервисов по локализации мобильных устройств или глобального позиционирования для измерения мобильности населения;

• моделирование для прогнозирования безработицы – исследование применения машинного обучения для построения прогнозных моделей для безработицы для малых областей на основе увязки данных обследований и административных;

• визуализация для разведочного анализа – исследование новых технологий визуализации данных для разведочного анализа сложных многомерных массивов данных;

• анализ множества соединений в связанных данных – исследовать связанные открытые методы анализа данных для многосвязных объектов данных на различных уровнях детализации;

• прогнозные модели для оценки неответов в обследовании – исследование применения машинного обучения для построения прогнозных моделей для оценки неответов с использованием параданных из прошлых обследований; и

• автоматический контент-анализ сложных административных данных - исследование методов для автоматического извлечения и разрешения концепций, единиц и фактов из мульти структурного контента массивов административных данных.

11.3 Участие в инициативах Австралийской государственной службы (APS) по анализу данных

Центр передового опыта по анализу данных Австралийской государственной службы был создан в конце 2013 года в ответ на рекомендации Стратегии в области информационных и коммуникационных технологий на период 2012-2015 гг. Ее цель состоит в развитии потенциала для совместной работы органов государственного управления в области использования передовых методов анализа данных посредством:

• передачи технических и экономических знаний, инструментария и методов, повышения квалификации и применения производственных стандартов, таких как протоколы конфиденциальности и управление информацией;

• изучения и выявления возможностей для повышения ценности для бизнеса в результате использования аналитических материалов по следующим направлениям: развитие практики управления информацией и знаниями, развитие методов анализа, инфраструктуры и программного обеспечения, аттестация и профессиональное развитие специалистов в области анализа данных занятых в секторе государственного управления; и

• выявления важных вопросов и проблем и консультирование Комитета ИТ-директоров относительно аналитического потенциала, барьеров на пути эффективного использования больших данных, пилотных проектов по использованию больших данных и других мероприятий, указанных в Стратегии использования больших данных APS (Department of Finance and Regulation, 2013).

В настоящее время в Центре заканчивается разработка руководства по использованию больших данных/анализу на основе больших данных, в котором представлена общегосударственная стратегия по использованию и внедрению больших данных в государственных учреждениях Австралии.

11.4 Сотрудничество с научным сообществом

АБС создает сеть для сотрудничества с ведущими австралийскими учеными-исследователями в области анализа данных для достижения научных целей

Флагманского проекта по использованию больших данных. В частности, к проекту будут привлечены исследователи группы обработки изображений и удаленного зондирования, которые работают в кампусе Университета Нового Южного Уэльса в Канберре и сотрудники Института передовых методов анализа данных Технологического университета в Сиднее для работы в таких областях, как спутниковое зондирование и прогнозное моделирование.

АБС также сотрудничает с Центром передового опыта для изучения математических и статистических границ больших данных, больших моделей и новых идей, который возглавляет выдающийся специалист в области математической статистики, профессор Питер Халл из Университета Мельбурна. Этот Центр, где работает мульти дисциплинарная команда, включающая статистиков, математиков, специалистов по вычислительной математике и вычислительной технике, был основан Австралийским научно-исследовательским советом с финансированием в объеме 20 млн. австралийских долларов на 7 лет. В качестве партнера в данной сфере деятельности АБС удавалось повлиять на включение в исследовательскую программу таких тем, как слияние и интеграция данных, которые представляют значительный интерес для АБС.

12. Заключение

Специалисты, работающие в системе официальной статистики, на протяжении многих десятилетий имели дело с разнообразными источниками данных. Хотя новые источники информации, связанные с использованием больших данных, дают возможность статистикам более эффективно оказывать услуги по производству статистической информации, при принятии решения о том, стоит ли использовать конкретный источник больших данных, следует принимать во внимание некоторые соображения, а именно производственную необходимость, выгоду и обоснованность использования данного источника для целей производства официальной статистики для получения оценок совокупности или аналитических выводов. Стандарт качества данных является полезным инструментом для оценки качества источников больших данных и для оценки соответствия цели исследования использованию больших данных.

До недавнего времени успехи АБС в области больших данных сводились в основном лишь к обзору и мониторингу разработок в отрасли и участии в мероприятиях по развитию стратегий и концепций за пределами организации. Осуществление Флагманского проекта по использованию больших данных дает возможность приобрести практический опыт в оценке экономических, статистических, технических, вычислительных и других вопросов, рассмотренных в данной статье. Участие АБС в национальных и международных мероприятиях, связанных с использованием больших данных, также поможет поделиться опытом и знаниями, а сотрудничество с научными кругами поможет АБС развить способности для решения своих производственных задач, используя большие данные в качестве одного из подходов. Наконец, эти и другие инициативы по данной теме описаны в Стратегии АБС по использованию больших данных (Australian Bureau of Statistics, 2014).

Выражение признательности

Авторы выражают признательность следующим специалистам: Брайан Студман (Brian Studman), Петер Радизих (Peter Radisich), главному редактору и двум рецензентам за их полезные замечания по более ранней версии данной статьи. Авторы также благодарны Ноэль Кресси (Noel Cressie) за полезные обсуждения. Взгляды, выраженные в данной статье, принадлежат авторам и не обязательно отражают мнение АБС.

Литература

Australian Bureau of Statistics. (2010a). The ABS Data Quality Framework. Available at: https://www.nss.gov.au/dataquality/aboutqualityframework.jsp. Accessed November 2014

Australian Bureau of Statistics. (2010b). Measuring trust in official statistics: the Australian experience. The OECD, Stat.News., 50, 9-11.

Australian Bureau of Statistics. (2012). Big data and official statistics. ABS Ann. Report, 13, 27-31.

Australian Bureau of Statistics. (2014). Big data strategy. Unpublished report.

Brackstone, G. (1999). Managing data quality in a statistical agency. Surv. Methodol., 25, 139-149.

Clarke, F. & Hamilton, A. (2013). From metadata to meaning: Semantic statistics in the ABS. Unpublished ABS manuscript.

Couper, M. (2013). Is the sky falling? New technology, changing media, and the future of surveys. Surv. Res. Meth., 7, 145-156.

Crawford, K. (2013). The hidden biases in Big Data. Harvard Business Review Blog. Available at: http://blogs.hbr. org/2013/04/the-hidden-biases-in-big-data/. Accessed November 2014.

Daas, P. & Puts, M. (2014). Big data as a source of statistical information. The Surv. Stat., 69, 22-31.

Department of Finance and Regulation. (2013). Big Data Strategy - Issues paper. Available at: http://www.finance.gov.au/files/2013/03/Big-Data-Strategy-Issues-Paper1.pdf. Accessed November 2014.

Eurostat. (2013). Scheveningen Memorandum on Big Data and Official Statistics. Available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf. Accessed November 2014.

Fellegi, I.P. & Sunter, A.B. (1969). A theory of record linkage. J. Am. Stat. Assoc., 64, 1183-1210. Groves, R. & Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs.

Harford, T. (2014). Big data: Are we making a big mistake? Financial Times. Available at: http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz36aRk9emv. Accessed November 2014.

Kish, L. (1965). Survey Sampling New York: Wiley.Landsat.

Manyika, L., Chui, M., Brown, B., Bughin, J., Dobbs, R. & Roxburgh, C. (2011). Big Data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute, McKinsey & Company.

Mui, Y. (2014). The weird Google searches about unemployment and what they say about the econ-omy. The Washington Post. Available at: http://www.washingtonpost.com/blogs/wonkblog/wp/2014/05/30/the-weird-google-searches-of-the-unemployed-and-what-they-say-about-the-economy/. Accessed November 2014.

OECD. (2011). Quality dimensions, core values for OCED statistics and procedures for planning and evaluating statistical activities. Available at: http://www.oecd.org/std/21687665.pdf. Accessed November 2014.

Pink, B., Borowik J. & Lee G. (2009). The case for an international statistical innovation program – Transforming national and international statistics systems. Stat. J. Int. Assoc. Offcial Stat., 26, 125-133.

Podesta, J., Pritzker, P., Monitz, E., Holdren, J. & Zients, J. (2014). Big Data: Seizing opportunities, preserving values. Washington: Executive Office of the President. Available at: http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf. Accessed November 2014.

Puza, B. & O'Neill, T. (2006). Selection bias in binary data from volunteer surveys. Math. Sci., 31, 85-94.

Rivera, J. & van der Meulen, R. (2013). Gartner Hype Curve. Available at: http://www.gartner.com/newsroom/id/2575515. Accessed November 2014.

Rubin, D.B. (1976). Inference and missing data. Biometrika, 63, 581-592.

Sarndal, C.E., Swensson, B. & Wretman, J. (1977). Model-Assisted Survey Sampling. New York: Springer-Verlag.

Smith, T.M.F. (1983). On the validity of inference from non-random samples. J Roy. Stat. Soc. A, 146, 394-403.

Tam, S.M. & Gross B. (2013). Discussion. J Official Stat., 29, 209-211.

UN Statistical Commission. (2013). Big data and modernization of statistical systems. Report of the Secretary-General. Available at:http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf.

UN Statistical Commission. (2014). Big data and modernisation ofstatistical systems. Available at: http://unstats.un. org/unsd/statcom/doc14/2014-11-BigData-E.pdf. Accessed November 2014.

Varian, H. & Choi, H. (2011). Predicting the present with Google Trends. Google Research. Available at: http://people. ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf . Accessed November 2014.

Wikipedia. (2014). Big data. Available at: .http://en.wikipedia.org/wiki/Big_data.

[Получено в феврале 2014, принято в апреле 2015]