Организация Объединенных Наций ECE/CES/2013/29

Экономический и Социальный Совет

Distr.: General
16 April 2013
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Шестьдесят первая пленарная сессия
Женева, 10−12 июня 2013 года
Пункт 4 b) предварительной повестки дня
Каким образом национальные статистические управления должны реагировать − переход от избежания рисков к управлению рисками

Инновационный доступ к микроданным − динамическое обеспечение конфиденциальности

Записка Австралийского бюро статистики*

Резюме

В настоящем документе описываются инновационные подходы к анализу микроданных. Он содержит обзор ключевых факторов, которые обусловили переход Австралийского бюро статистики на эти новые методы. В нем описывается комплект серверов дистанционного анализа и "динамический" подход. Он также содержит краткий обзор международных подходов к предоставлению доступа к микроданным, и в заключение в нем предлагаются направления будущих исследований.

Австралийское бюро статистики признает необходимость "переосмысления" доступа к микроданным с целью сделать возможным удовлетворение спроса пользователей на доступ к более подробным данным регистрационных единиц по более широкой совокупности наборов данных. Необходимы более гибкие подходы, чем простое обеспечение конфиденциальности данных до предоставления доступа к ним. Исходя из этого, Австралийское бюро статистики разработало среду дистанционного выполнения для доступа к микроданным, включая сервис построителя и анализа таблиц данных обследований. В сопоставлении с традиционными процедурами обеспечения конфиденциальности этот новый метод оказывает меньшее влияние на качество данных благодаря внесению необходимых модификаций только в конечный продукт, а не в первичные данные. Все конечные продукты подвергаются "динамическому" процессу обеспечения конфиденциальности с целью защиты конфиденциальности данных при одновременном сохранении уровня детализации и качества.

* Данный документ был представлен с опозданием из-за задержки с представлением материалов из других источников.

I. Введение

1. Статистические учреждения собирают большие объемы микроданных в рамках переписей, обследований и из административных источников. Такие микроданные могут использоваться в целях разработки и оценки политики на благо общества или для оказания ему полезных услуг. Спрос на расширение доступа к таким микроданным продолжал расти и после сессии Конференции европейских статистиков (КЕС) 2003 года, посвященной теме конфиденциальности микроданных и доступа к ним.

2. Задача Австралийского бюро статистики (АБС) заключается в "оказании помощи и содействия принятию информированных решений, проведению информированных исследований и обсуждений в рамках органов власти и обществе в целом, обеспечивая функционирование высококачественной, объективной и ответственной национальной статистической службы". АБС, как и многим другим национальным статистическим организациям (НСО), потребовалось "переосмыслить" доступ к микроданным, столкнувшись с задачей поиска компромисса между правовыми обязательствами по обеспечению снижения до минимума вероятности раскрытия информации о конкретном лице или организации и необходимостью публикации более подробных микроданных для информированного принятия решений, проведения информированных исследований и обсуждений на благо общества.

3. Управление риском раскрытия, как правило, называют статистическим контролем за раскрытием (СКР). Даже после удаления персональной идентификационной информации, такой как имя и адрес, сохраняется риск раскрытия на основе микроданных (см., например, Willenborg and de Waal, 2001). Вследствие непрерывного роста объема данных, доступных в административных и связанных между собой источниках, ставшего возможным благодаря технологическому прогрессу, риск раскрытия на основе микроданных, несомненно, постоянно увеличивается.

4. В настоящее время разработан ряд программных продуктов для обеспечения конфиденциальности данных. Большинство из них предназначены для использования НСО в целях обеспечения конфиденциальности данных до предоставления к ним доступа. Они предназначены для работы либо с табличными данными (например, Tau-Argus[1] и sdcTable[2]) или с микроданными (например, Special Uniques Detection Algorithm[3] и sdcMicro[4]). Многие НСО также разрабо-

[1] Tau-Argus представляет собой бесплатное программное обеспечение для защиты статистических таблиц, которое можно загрузить по адресу: http://neon.vb.cbs.nl/casc/tau.htm.

[2] sdcTable представляет собой бесплатный пакет статистического ограничения данных для защиты табличных данных, который можно загрузить по следующему адресу: http://cran.r-project.org.

[3] Special Uniques Detection Algorithm представляет собой систему обнаружения и ранжирования особых уникальных характеристик. Она необходима для обеспечения конфиденциальности наборов данных за счет обнаружения всех особых уникальных записей для их последующих маскировки или удаления.

[4] SDCMicro представляет собой бесплатное программное обеспечение для генерирования защищенных микроданных для исследователей и широкого использования, которое можно загрузить по адресу: http://cran.r-project.org/web/packages/sdcMicro/index.html.

тали свои собственные индивидуальные процессы и программные продукты с учетом требований своего законодательства.

5. АБС постепенно разработало набор методов, процессов и приложений для подготовки широкого ассортимента продуктов, предоставляющих доступ к его статистическим данным, включая предоставление статистических табличных данных на веб-сайте АБС, предоставление подготовленных с учетом требований заказчика таблиц через сервис справочной информации и через сервис анализа предварительно обезличенных файлов микроданных, называемых "Файлы обезличенных единичных записей" (ФОЕЗ). ФОЕЗ предоставляются либо в форме базовых ФОЕЗ аналитикам для использования в их собственной исследовательской среде, либо в виде расширенных ФОЕЗ, которые предоставляются по удаленным запросам в Лаборатории удаленного доступа к данным (ЛУДД) или в одной из онсайт лабораторий данных АБС (ЛДАБС). Авторизованные пользователи обычно направляют запросы в SAS, STATA или SPSS[5]. В случае ЛУДД результаты запросов автоматически проверяются и одобренные продукты предоставляются пользователям через их настольные ПК. Кроме того, на постоянной основе формируются выборки запросов для ручной проверки. В случае ЛДАБС все продукты, подлежащие выдаче из лабораторий, одобряются вручную. Уровень предварительного обезличивания, применяемого к каждому файлу ФОЕЗ, зависит от метода распространения, начиная с глубоко обезличенных базовых ФОЕЗ и кончая более подробными, менее обезличенными расширенными ФОЕЗ. Подготовка ФОЕЗ в любой форме требует значительных затрат труда и времени.

6. В ответ на многочисленные требования перемен, включая возросшую сложность наборов данных, в случае которых, как представляется, традиционные подходы не работают, АБС в последнее время посвящает значительные ресурсы разработке комплекта исследовательских приложений, позволяющих аналитикам направлять через Интернет запросы, которые исполняются с использованием первичных микроданных, подвергающихся динамическому обезличиванию в реальном времени, после чего обезличенные продукты возвращаются аналитику. На международном уровне такие приложения обычно называют серверами дистанционного анализа (СДА). Они позволяют пользователям специфицировать конкретные продукты, которые они хотят извлечь из набора данных. Задача статистических управлений заключается в предоставлении СДА для разработки различных продуктов.

7. В разделе II настоящего документа приводится обзор ключевых факторов, обусловивших переход АБС на эти новые подходы к анализу микроданных. В разделе III описывается комплект серверов удаленного анализа АБС и динамический подход к СКР. Раздел IV содержит краткий обзор различных международных подходов к предоставлению доступа к микроданным, а раздел V – перечень направлений будущих исследований и возможностей международного сотрудничества.

[5] Эти статистические программные решения общего назначения используются в процессах статистического производства и анализа: SAS (Statistical Analysis System), STATA and SPSS (Statistical Package for the Social Sciences).

II. Движущие факторы перемен

8. Многие НСО, включая АБС, занимаются пересмотром своих подходов к предоставлению микроданных исследователям. К ключевым движущим факторам перемен относятся:

a) растущий спрос на более эффективный, гибкий и оперативный доступ к подробным микроданным;

b) накопление опыта работы пользователями в области предоставления удобных, опирающихся на меню интерфейсов, которые не требуют от пользователей навыков статистического программирования;

c) снижение издержек, связанных с существующими ручными и ресурсоемкими подходами к распространению микроданных, такими как процесс создания ФОЕЗ;

d) повышение оперативности доступа к микроданным (ФОЕЗ в настоящее время предоставляется с почти шестимесячным лагом по сравнению со статическими таблицами);

e) смягчение возросшего риска раскрытия в результате наращивания вычислительных ресурсов (как аппаратного, так и программного обеспечения), увеличения объема распространяемых продуктов и расширения доступности больших наборов данных;

f) упрощение анализа новых источников данных (включая операционные, административные и интегрированные данные), в случае которых традиционные подходы к СКР являются недостаточными для смягчения возросшего риска идентификации;

g) все более широкое признание того, что не все основные статистические активы находятся у НСО, что обусловливает необходимость разработки методов и инфраструктуры, которые могли бы использоваться другими организациями;

h) эволюционирующая модель типичного аналитика данных в направлении организаций, стремящихся к возросшей доступности продуктов благодаря более эффективному процессу межмашинного запроса информации, таких как использование веб-сервисов SDMX (обмен статистическими данными и метаданными).

9. Эти движущие факторы перемен заставили многие НСО, включая АБС, приступить к разработке отвечающих современным требованиям серверов дистанционного анализа.

III. Серверы дистанционного анализа Австралийского бюро статистики − динамическое обеспечение конфиденциальности

10. Комплект серверов дистанционного анализа АБС позволяет АБС предоставлять полный набор подробных характеристик наборов данных, одновременно сводя к минимуму потерю полезности благодаря применению СКР, учитывающего особенности каждого конкретного продукта. Для пользователя это означает, что он может специфицировать конкретные продукты, которые он хочет извлечь из набора данных. Это является фундаментальным сдвигом в процес-

се − с традиционной парадигмы, в рамках которой АБС принимало решения по всем продуктам, которые оно будет предоставлять, к парадигме, в случае которой пользователи могут специфицировать, что им требуется и когда.

11. При распространении табличных и аналитических продуктов существует реальный риск раскрытия, который необходимо смягчать. По этому вопросу был опубликован ряд документов, включая предлагаемые подходы к управлению риском раскрытия. В отношении аналитических продуктов см. работы Gomatam et. al (2005), Bleninger et. al (2011) и Sparks et. al (2008), а в отношении табличных продуктов − Shlomo (2007). Цель данной литературы заключается в описании методов защиты от злонамеренных действий в отношении данных, которые предусматривают использование аналитиком материалов с сервера анализа, включая диагностику графиков и моделей, для восстановления признаков одной или нескольких записей, которые, в случае удачи, могли бы использоваться для возможной идентификации. Задача АБС заключается в обеспечении динамического СКР в отношении различных возможных продуктов.

12. Весьма упрощенная модель сервера дистанционного анализа является следующей:

a) учреждение предоставляет файл микроданных исследователям. Безопасность его хранения обеспечивается учреждением. Чувствительные микроданные обычно невидимы аналитику;

b) аналитик направляет запрос через Интернет на сервер анализа учреждения, который обрабатывает его с использованием чувствительных микроданных;

c) статистический продукт (например, коэффициент регрессии или таблица) запроса модифицируется с использованием специального метода обеспечения конфиденциальности, соответствующего данному анализу, в целях СКР;

d) сервер анализа направляет модифицированный продукт через Интернет аналитику. Некоторые продукты могут ограничиваться исходя из того, что они могут позволить аналитику восстановить признаки случайной записи.

13. Следующее поколение серверов дистанционного анализа АБС включает в себя три приложения − Census TableBuilder, ABS TableBuilder и ABS DataAnalyser. Каждое из них описывается в нижеследующих пунктах. Ни в один момент времени аналитик не видит первичные микроданные.

14. Для целей переписи населения и жилищного фонда 2006 года АБС совместно с компанией Space-Time Research Pty Ltd разработало приложение **Census TableBuilder**. Данное приложение представляет собой веб-продукт, который доступен аналитикам за рамками АБС. Census TableBuilder опирается на метод создания помех (Fraser and Wooton (2005)) для автоматической защиты таблиц данных переписи. Данный метод был разработан с целью снижения риска раскрытия при обработке запросов на схожие таблицы, повторных запросов на идентичные таблицы и повторных запросов в отношении одной и той же ячейки таблицы в рамках различных таблиц. Этот метод призван обеспечить более широкий доступ к данным в отношении различных подсовокупностей, а также разработку веб-систем, позволяющих пользователям определять свои собственные параметры таблиц. Все табличные продукты переписи 2006 года защищаются с использованием одного и того же метода, включая таблицы, создаваемые сотрудниками АБС для публикаций. По этой причине существуют внутренние

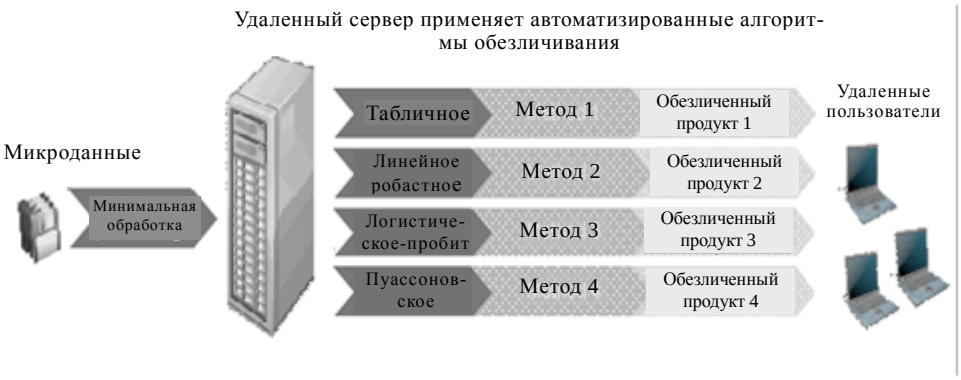
системы применения данного метода, а также размещенное в Интернете приложение Census TableBuilder для аналитиков.

15. Исходя из этой основы, АБС продолжает разработку набора индивидуализированных алгоритмов и процедур защиты конфиденциальности, которые работают в динамическом режиме и снижают до минимума утрату полезности, одновременно обеспечивая, чтобы производимые продукты не позволяли идентификации респондентов. Каждый метод обеспечения конфиденциальности учитывает особенности конкретного продукта, снижая уровень требуемого обезличивания. Для сравнения существующие подходы ФОЕЗ требуют более громоздких процедур обеспечения конфиденциальности для защиты всего набора потенциальных продуктов.

16. Приложение **ABS TableBuilder,** которое является преемником Census TableBuilder, также было разработано совместно с компанией Space Time Research Pty Ltd. Оно содержит динамические процедуры обеспечения конфиденциальности в отношении взвешенных данных наблюдений, которые выходят за итоги совокупности на ключевую сводную статистику из данных порядка величины (такую как установленные классы величины, итоги, средние, медианы и квантили). В дополнение к помехам в него также встроен ряд процедур защиты/ограничения. Они включают в себя ограничения на совместное табулирование определенных комбинаций единиц данных, ограничения на использование рассеянных таблиц, в которых существует большое число малых ячеек, и предъявление требования минимального размера совокупности для расчета медиан и квантилей. Приложение ABS TableBuilder также облегчает межмашинное направление запросов в отношении микроданных через вебсервис SDMX.

17. Программное обеспечение **ABS DataAnalyser** было разработано с целью облегчения исследовательского анализа данных и регрессивного моделирования. Приложение ABS DataAnalyser представляет собой безопасную систему меню для проведения статистического анализа через удаленный интерфейс пользователя. Система позволяет пользователям дистанционно оценивать параметры статистических моделей, адаптированных к данным АБС, одновременно обеспечивая защиту конфиденциальности провайдеров. Все статистические продукты, которые могут просматриваться пользователями в автоматическом режиме, являются обезличенными с использованием различных методов контроля раскрытия, включая искажение уравнения оценки. Данное искажение само по себе не является достаточным, и поэтому в систему был включен набор ограничений и процедур защиты от конкретных злонамеренных действий, включая обезличенные графические изображения для оказания помощи пользователям в диагностике пригодности модели. Резюме метода постановки помех и дополнительных мер защиты можно найти в работе Chipperfield, Gare and Yu (2011). Исходная версия ABS DataAnalyser позволяет пользователям осуществлять трансформацию и обработку данных, составление таблиц, исследовательский анализ данных и линейное робастное, логистическое, пробитрегрессионное, пуассоновское или мультиноминальное моделирование. Данная система в настоящее время тестируется пользователями, и публикация полностью рабочей версии запланирована на третий квартал 2013 года. Диаграмма 1 служит иллюстрацией приложения ABS DataAnalyser.

Диаграмма 1
Приложение DataAnalyser с динамическим процессом обеспечения конфиденциальности Австралийского бюро статистики



18. Преимущества подхода АБС к серверам дистанционного анализа включают в себя следующие:

a) анализ осуществляется с использованием реальных микроданных, сохраняющих сложные взаимосвязи в данных;

b) статистический продукт подвергается модификации до уровня, который конкретно отвечает типу осуществляемого анализа, а также до уровня, который сводит до минимума потерю информации;

c) после завершения настройки сервер способен обрабатывать множественные виды анализа в реальном времени;

d) все представленные программы можно регистрировать и проверять, и в случае выявления попытки идентификации аналитику может быть закрыт доступ на сервер;

e) интерфейс меню "укажи и выбери" означает, что от пользователей не требуется значительной подготовки и обучения новому языку программирования.

19. Недостатки подхода АБС включают в себя следующие:

a) аналитик вынужден ограничиваться использованием только тех методов анализа, преобразования и обработки данных, которые поддерживаются сервером;

b) анализ через удаленные серверы может потребовать больше времени, чем анализ микроданных, имеющихся на персональном компьютере аналитика;

c) для разработки каждой новой аналитической функции требуются значительные затраты времени и денег.

IV. Международные подходы к предоставлению доступа к микроданным

20. Национальные статистические организации используют весьма различные подходы к предоставлению доступа к микроданным, которые в значительной степени обусловлены различиями в требованиях законодательства. Ряд

НСО публикуют файлы общего доступа. Эти файлы предназначены для общего использования и характеризуются глубоким обезличиванием.

21. НСО также широко используют исследовательские центры данных, схожие с ЛД АБС, для анализа подробных микроданных. Недостатком этих подходов является то, что продукты, полученные из этих центров, должны, как правило, проверяться вручную. Это является весьма ресурсоемкой процедурой, ведущей к ограничениям на число исследователей, которым может предоставляться доступ. АБС стремится найти решение, обеспечивающее максимально широкий, по возможности, доступ.

22. Некоторые исследовательские центры являются центрами "он-сайт", в то время как другие используют передовые технологии для создания виртуальных исследовательских центров, доступ на которые возможен с неинтеллектуальных терминалов, расположенных в других организациях. Уровень детализации доступных данных также является различным в зависимости от требований законодательства, регулирующего деятельность НСО. В некоторых случаях авторизованным или "надежным" исследователям предоставляются те же права доступа к подробным микроданным, что и сотрудникам НСО. Законодательство, регулирующее деятельность АБС, запрещает делать это в Австралии.

23. Для сравнения веб-системы АБС были разработаны для предоставления удаленного доступа внешним пользователям. Данный подход не требует глубокого обезличивания микроданных до их предоставления в целях анализа и опирается на методы обеспечения конфиденциальности, которые применяются в режиме реального времени. Признавая преимущества серверов дистанционного анализа, предусматривающих процесс динамического обеспечения конфиденциальности, ряд НСО приступили к осуществлению программ исследований и разработок в этой области. В качестве двух весьма передовых разработок следует отметить проект Höninger (2011) Государственного статистического института Берлина-Бранденбурга и проект Lucero et. al (2011), реализуемый Бюро переписей США.

V. Будущие направления и возможности международного сотрудничества

24. Серверы дистанционного анализа АБС обладают преимуществом предоставления высококачественных продуктов, созданных на основе файлов микроданных, а также удобного для пользователей доступа, обеспечивая при этом конфиденциальность данных. Однако многое предстоит еще сделать.

25. Текущие исследования сосредоточены на оценке существующих подходов к связанным наборам данных, которые характеризуются повышенным риском раскрытия, а также изучении их эффективности. Будущая исследовательская работа будет посвящена разработке динамических методов обеспечения конфиденциальности при распространении коммерческих и лонгитюнидальных наборов данных через серверы дистанционного анализа и созданию показателей потери полезности для исследователей (Marley and Leaver 2011).

26. АБС также проявляет значительный интерес к изучению потенциала синтетических микроданных, хотя бы для того, чтобы позволить аналитикам протестировать и пересмотреть свои модели до их применения к реальным данным, которые безопасно размещены в приложении ABS DataAnalyser.

27. АБС также проявляет интерес к сотрудничеству с международным статистическим сообществом с целью проведения исследований по изучению проблем, связанных с такими наборами данных. АБС также хотела бы представить статистические методы и алгоритмы, лежащие в основе этих методов, международному статистическому сообществу, если последнее пожелает включить данные методы в свои приложения.

VI. Справочные материалы

Bleninger, P., Drechsler, J. and Ronning, G. 2011, "Remote Data Access and the Risk of Disclosure from Linear Regression: An Empirical Study", *Privacy in Statistical Databases*, Springer. See http://www.idescat.cat/sort/sortspecial2011/DataPrivacy.1.bleninger-etal.pdf).

Chipperfield, J., Gare, M. & Yu, F. 2011, "Providing access to microdata for statistical purposes − experiences of the Australian Bureau of Statistics with Remote Analysis Servers", paper presented to the Statistics Canada 2011 Methodology Symposium, Ottawa, Canada, 1−4 November.

Fraser, B. & Wooton, J. 2005, "A proposed method for confidentialising tabular output to protect against differencing", paper presented to the Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva, Switzerland, 9−11 November.

Gomatam, S., Karr, A. F., Reiter, J. P., & Sanil, A. P. 2005, "Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk – Utility Framework for Remote Access Analysis Servers", *Statistical Science*, 20, pp. 163−177.

Höninger, J. 2011, "An Innovative Approach to Remote Data Access", 58th International Statistical Institute World Statistics Congress, Dublin, Ireland, 21−26 Aug 2011.

Lucero, J., Zayatz, L., Singh, L., You, J., DePersio, M. and Freiman, M. 2011, "The Current Stage of the Microdata Analysis System at the U.S. Census Bureau", 58th International Statistical Institute World Statistics Congress, Dublin, Ireland, 21−26 Aug 2011.

Marley, J. & Leaver, V. 2011, "A method for confidentialising user-defined tables: statistical properties and a risk-utility analysis", paper presented to the International Statistical Institute session, Dublin, Republic of Ireland, 22−26 August.

Shlomo, N. 2007, "Statistical disclosure control methods for census frequency tables", International Statistical Review, 75, (2), 199−217.

Sparks, R., Carter, C. Donnelly, J., O'Keefe, C.M., Duncan, J., Keighley, T. and McAullay, D. (2008), "Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving AnalyticsTM", *Computer Methods and Programs in Biomedicine* 91, pp. 208−222.

Willenborg. L. and de Waal, T. 2001, "Elements of Disclosure Control", Lecture Notes in Statistics, Vol 155, ISBN 978-0-387-95121-8, Springer.