КОМИССИЯ ПО СТАТИСТИКЕ и ЭКОНОМИЧЕСКАЯ КОМИССИЯ ДЛЯ ЕВРОПЫ

КОНФЕРЕНЦИЯ СТАТИСТИКОВ ЕВРОПЫ

КОМИССИЯ ЕВРОПЕЙСКИХ СООБЩЕСТВ

ЕВРОСТАТ

Совместный рабочий семинар ЭКЕ и ЕВРОСТАТа по конфиденциальности статистической информации
(Скопье, бывшая югославская республика Македония, 14-16 марта 2001 г.)

Рабочий доклад №2

Тема I: Применение методологии по контролю за соблюдением конфиденциальности статистической информации и программное обеспечение в коммерческой и социально-демографической статистике.

НОВЫЕ ИНСТРУМЕНТЫ ПОДАВЛЕНИЯ ЯЧЕЕК В ПРОГРАММЕ TAU-ARGUS: ОДНА ЧАСТЬ ПРОГРАММЫ РАБОТ ПО ПРОЕКТУ CASC

Запрошенная работа

Представлена Федеральным статистическим бюро Германии[1]

Аннотация: В ходе финансируемого ЕС проекта CASC предполагается расширить компьютерную программу τ-ARGUS таким образом, чтобы превратить ее в универсальный, стандартный инструмент для защиты табулярных данных. Необходимые для этого модификации повлияют на средства по обеспечению защиты от (остаточного) риска нарушения конфиденциальности, структуру данных и интерфейс пользователя. Требуется осуществить меры по защите сложных иерархических таблиц и их связок, особенно в контексте систем баз данных, предназначенных для общественного доступа. Необходимо модифицировать методологию вторичного подавления ячеек, лежащую в основе данного программного пакета. Программа будет работать в частности с гиперкубическим алгоритмом GHQUAR и, наконец, будут добавлены инструменты по пертурбации таблиц.

Ключевые термины: проект CASC, τ-ARGUS, защита табулярных данных, иерархические таблицы, защита связок таблиц, вторичное подавление ячеек.

I. ВВЕДЕНИЕ

1. Одним из проектов, финансируемых Европейским Союзом в рамках Европейской программы исследований в области официальной статистики (EPROS) и входящим в 5-ю базовую программу, является проект CASC (Вычислительные аспекты конфиденциальности в статистике). Общее руководство проектом осуществляет Анко Хандепул из Статистического бюро Нидерландов.

2. Работа над проектом будет осуществляться с января 2001 г. по декабрь 2003 г. Данный проект является продолжением проекта SCD из 4-ой базовой программы в том смысле, что в нем получат дальнейшее развитие достижения этого проекта и будут использоваться его результаты и конечные продукты. Одной из основных задач является дальнейшее совершенствование программного обеспечения ARGUS, например, программы •- ARGUS для создания безопасных файлов микроданных и программы τ-ARGUS для защиты табулярных данных. Цель данной работы заключается в описании предполагаемой работы над этим проектом с точки зрения методологии τ-ARGUS. Автор является членом управляющего комитета CASC и в этом качестве

[1] Подготовила Сара Гиссинг (e-mail: sarah.giessing@statistik-bund.de).

будет контролировать исследование и разработку дополнительной методологии по защите табулярных данных в программе τ-ARGUS.

3. Что касается защиты табулярных данных, работа по проекту CASC будет осуществляться статистическими институтами Нидерландов, Германии и земли Северный Рейн-Вестфалия, Германия (CBS, StBA, LDS, NRW), а также специальными группами Университета Ла Лагуна (ULL), Технического университета Илменау (TUI), Политехнического университета Каталонии (UPC). CBS будет осуществлять работу с программой τ-ARGUS, ULL обеспечит оптимальные поисковые алгоритмы, StBA предложит стратегию методологии, LDS/NRW предоставят гиперкубический алгоритм GHQUAR, UPC представит альтернативный алгоритм оптимизации, а основной задачей TUI будет разработка алгоритмов по конкретным методам пертурбации таблиц. StBa осуществляет координирующую роль в исследованиях, а CBS – в разработке программного обеспечения и исследованиях по линейному программированию.

II. ЦЕЛИ ПРОЕКТА CASC В ОТНОШЕНИИ МЕТОДОЛОГИИ ЗАЩИТЫ ТАБУЛЯРНЫХ ДАННЫХ

4. Что касается защиты табулярных данных, то целью проекта является разработка программного обеспечения, подходящего для использования в качестве стандартного инструмента по контролю конфиденциальности агрегированных показателей. Это значит, что такое программное обеспечение должно быть способно оперировать таблицами любого размера и структурной сложности, должны быть разработаны гибкие, несложные для пользователя и удобные способы решения специфических проблем в конкретных ситуациях. Программа должна быть легкодоступной и полезной и, самое главное, хорошо уравновешена с точки зрения качества и количества. То есть, программа должна предлагать, в зависимости от конкретных ситуаций (например, размера и сложности конкретной задачи), самые оптимальные модели подавления (с точки зрения потери информации в результате такого сокрытия), которые можно эффективно применять (с точки зрения требований к вычислительным ресурсам).

5. Конкретные действия по достижению главной задачи данной рабочей программы сводятся к следующему:

(i) Уточнение и помощь по интегрированию необходимых качеств и средств существующих программных систем по защите табулярных данных в программу τ-ARGUS.

(ii) Интегрирование самой последней версии программы GHQUAR, что обеспечит широкие возможности применения программы к (связанным) таблицам любого размера и структурной сложности.

(iii) Существенное совершенствование уже имеющихся в программе алгоритмов по подавлению ячеек на основе линейного программирования, разработка альтернативных методов на основе методологии сетевых потоков, а также предоставление вспомогательной эвристической методологии.

(iv) Предоставление информации по эффективности различных алгоритмов для вторичного подавления ячеек, которые должны быть включены в конечный программный пакет. Такая информация поможет модифицировать программу и также будет полезна при принятии внутренних решений в ходе работы над проектом.

(v) Обеспечение методов и инструментов по максимальному увеличению информационного содержания таблиц с подавленными единицами, т.е. после завершения процедуры подавления.

(vi) Приобретение опыта работы с новыми средствами по контролированию выбора вторичных подавлений и передача этого опыта потенциальным пользователям. В частности, решить проблему вторичного подавления ячеек на «европейском уровне», например, каким

образом можно упростить и обеспечить подходы по координации моделей подавления по всей Европе, как это предлагалось, к примеру, «Евростатом» применительно к данным структурных коммерческих обследований,- см. [(Doc. Eurostat/D2/SBS-T/NOV99/03)].

III. НОВЫЕ СРЕДСТВА ПО ОПРЕДЕЛЕНИЮ РИСКА НАРУШЕНИЯ КОНФИДЕНЦИАЛЬНОСТИ

6. Если только вообще не имеет место распространение каких-либо данных, всегда остается некоторая доля риска нарушения конфиденциальности даже защищенных данных. Прежде чем выполнять процедуру по контролю за риском нарушения конфиденциальности, пользователь должен ввести в систему допустимый для него уровень риска нарушения конфиденциальности.

III.1 Первичная оценка риска нарушения конфиденциальности

7. Первый шаг в процессе предотвращения нарушения конфиденциальности в отношении таблиц заключается в оценке риска нарушения конфиденциальности, связанного с выпуском каждой ячейки внутри данной таблицы. Такая проверка обычно осуществляется с помощью применения к данным определенных правил по определению их чувствительности. Ячейка, считающаяся чувствительной в соответствии с примененным правилом чувствительности, не публикуется, т.е. «подавляется». Вместе с так называемыми «первичными подавлениями» также должны быть подавлены и остальные ячейки (так называемые «вторичные» или «дополнительные» подавления) с тем, чтобы пользователи опубликованных табличных данных не имели возможности точно вычислить первичные подавления или достаточно точно расчитать значения благодаря линейным отношениям между опубликованными и подавленными ячейками таблицы.

8. В настоящее время программа τ-ARGUS предлагает использовать правило концентрации (например, правило (n,k)-доминанты) в сочетании с правилом минимального количества респондентов, где параметры n и k правила доминанты и минимальное количество респондентов определяются пользователем. Эти правила являются лишь специфическими случаями (хотя и наиболее известными и распространенными) “мер по верхней линейной чувствительности” более общего класса (см. напр. [1], [11]). Новая версия τ-ARGUS позволит объединить несколько мер по верхней линейной чувствительности. В частности, будет предложено применять правило p% или правило (p,q).

III.2 Интервал защиты

9. С помощью линейных отношений между опубликованными и подавленными ячейками пользователи опубликованных данных могут расчитать верхнюю и нижнюю границы истинной величины любого подавленного показателя. Интервал между этими пороговыми значениями называется «интервалом подавления». Для правильного выбора дополнительных подавлений распространитель должен определить безопасные границы «интервала защиты». Процедура подавления обеспечит невозможность определения модели подавления, если только соответствующий интервал подавления не включает интервал защиты для каждой чувствительной ячейки.

10. В настоящей версии τ-ARGUS пользователя просят указать границы интервала защиты. Однако для процедуры подавления очень важно, чтобы эти границы были определены надлежащим образом. Иначе может возникнуть риск нарушения конфиденциальности или так называемое “переподавление”.

11. Когда для обеспечения первичной конфиденциальности используется такое правило концентраций как (n,k)-доминанта или правило p%, то, согласно этому правилу, имеется недопустимый риск нарушения конфиденциальности, так как верхняя граница интервала подавления не достаточно превышает истинное значение чувствительной ячейки. Если разница между верхней границей и истинным значением ниже определенного минимального уровня, то

эту верхнюю границу можно использовать для расчета отдельных показателей чувствительной ячейки, которые слишком близки по используемому критерию чувствительности. Формулы для (верхних) границ интервала защиты, отвечающего этому критерию, можно легко получить на основе результатов [1] и они представлены в приложениях к самым распространенным правилам определения чувствительности. Новая версия τ-ARGUS, как предполагается, должна предложить заданную по умолчанию возможность расчета интервала защиты согласно этим формулам.

12. К сожалению, в некоторых случаях даже при надлежащем определении интервала защиты, как это описывалось выше, тем не менее существует некоторый остаточный риск нарушения конфиденциальности. Если дополнительные подавления также являются чувствительными или если ячейки, являющиеся частями одной и той же общей ячейки, имеют общих респондентов, как поясняется в III.3 ниже, то модель подавления должна также обеспечивать невозможность раскрытия (общего) показателя данных респондентов на основе верхней границы истинного значения (подавленной) совокупной ячейки. Этого нельзя достичь только за счет правильного определения интервала защиты для чувствительной ячейки и, к примеру, за счет того, что дополнительное подавление удовлетворяет условию некоего минимального количества. Эту проблему необходимо решать и найти решения как это предлагается, например, в [7] и [10].

III.3 Проблема общих респондентов

13. Статистики часто строят одну и ту же таблицу для различных переменных показателей в ответах респондентов. В наборе таких переменных может присутствовать дополнительное отношение – например, одна из переменных может фактически представлять собой сумму всех остальных. Примером таких взаимоотношений является такое отношение, как «общая сумма инвестиций = инвестиции в строительство + инвестиции в землю + инвестиции в техническое оборудование + другие инвестиции».

14. Распространенный метод защиты таблиц, используемый в таких случаях, заключается в защите таблицы в отношении только одного из переменных показателей ответов респондентов (например, для «общей суммы инвестиций») с последующем подавлением в этом примере тех показателей в таблицах «строительство», «земля», «техническое оборудование» и «другие инвестиции», которые соответствуют подавленным показателям в таблице «общая сумма инвестиций». Хотя такой подход, конечно, полезен с точки зрения облегчения защиты данных, его лучше не использовать, если информация, представленная какими-либо другими переменными, считается и чувствительной и идентифицирующей. «Идентифицирующая» – значит, что «взломщики» предположительно могут не только идентифицировать респондентов, предоставивших необычайно крупные показатели в «общий» переменный показатель (например, общая сумма инвестиций), но и догадаться о крупнейших респондентах в других переменных. Такое может иметь место и в нашем примере с инвестициями. Если, например, таблица представляет собой агрегированные данные низкого уровня, а один из респондентов имеет дорогое новое здание, в то время как остальные респонденты не имеют значительных расходов на строительство, то в этом случае многие из этих респондентов могут знать на кого приходится наибольшая доля из общего опубликованного показателя по инвестициям в строительство и могут определить эту отдельную долю. Поэтому эту ячейку необходимо подавить, даже если «общие инвестиции» могут оказаться безопасными. В данном случае все таблицы по инвестициям необходимо защитить вместе как одну таблицу, а отношения между различными категориями инвестиций представить в одном измерении этой таблицы.

15. Таблицы, построенные подобным образом, имеют типичное свойство: ячейки, представляющие собой части одной (общей) ячейки, характеризуются общими респондентами. Если общая доля респондента в общей ячейке считается конфиденциальной информацией, такой, скажем, как «инвестиции в строительство и землю», то, как говорилось в III.2 выше, модель подавления должна обеспечивать невозможность определения общего вклада каких-либо респондентов на основе верхней границы истинного значения (подавленной) общей ячейки.

IV. НОВЫЕ СТРУКТУРЫ ДАННЫХ

16. Настоящая версия τ-ARGUS не может оперировать таблицами с иерархической подструктурой и таблицами с расчлененной структурой переменных показателей респондентов, которые описывались в разделе III.3 выше. Отсутствует возможность защиты связанных таблиц и, если в результате децентрализованной организационной структуры статистической системы типа европейской или германской, потенциальный пользователь программного обеспечения не в состоянии предоставить массив микроданных, лежащих в основе таблицы, которую он хочет защитить, то он не сможет пользоваться этой системой. Все эти возможности будут иметься в новых версиях, которые,несомненно, потребуют модификацию и новые концепции структур данных.

17. Необходимо разработать новые структуры данных для ввода табулярных показателей, модифицировать структуру ввода микроданных, а также модифицировать или создать заново файлы для метаданных, списки кодов и прочую структурную информацию по таблицам.

IV.1 Перспективы для пользователей: что ожидать относительно новых интерфейсных возможностей τ-ARGUS

18. В данном разделе будут кратко рассмотрены альтернативные варианты входных данных пользователя для определения иерархических таблиц и управления процедурой защиты связанных таблиц.

IV.1.1 Определение иерархических таблиц

19. В случае иерархических таблиц пользователь должен ввести информацию по структуре подлежащей защите таблицы. Естественно, можно вообразить целый ряд возможных способов представления такой информации. С точки зрения ввода табулярных данных, т.е. пользователей, предоставляющих табулярные данные, одна из возможностей заключается в извлечении такой информации из файлов данных. В этом случае входной файл табулярных данных для иерархической таблицы должен указывать ячейки с помощью иерархических кодов для любых иерархически структурированных переменных. Затем программа создаст список кодов по каждому измерению в том виде, в котором они представлены в файле данных. Это, конечно, не пригодно для пользователей микроданными. В этом случае самый простой вариант заключается в запросе пользователю предоставить список иерархических кодов для каждого параметра таблицы. Конечно, такой список должен соответствовать кодам данного параметра в файле микроданных. Программа проверяет выполнение соответствующих условий.

20. Многие пользователи были бы довольны входной информацией такого рода. Иерархические списки кодов будут иметься для самых подробных переменных показателей – таких, например, как КДЕС (NACE) или NUTS (Номенклатура территориальных единиц для статистических целей = НТЕС) для переменных «отрасль» или «область». Если для переменных такого списка кодов нет (обычно это такие переменные, как «размерность класса», «год образования», и т.д.), то создать такой список не очень сложно, так как эти переменные, как правило, имеют простые иерархические структуры (если они есть вообще), состоящие обычно из не более 20 категорий. Таким образом, будучи разработанной в соответствии с этими принципами, программа будет тем не менее полезна, хотя и без достаточной гибкости и несмотря на достаточно трудоемкие организационные действия, необходимые для подготовки данных, особенно при решении нестандартных задач или нестрогом соответствии входных данных заложенным требованиям.

21. Учитывая тот факт, что мы стремимся разработать стандартный инструмент, который можно было бы использовать в самых разнообразных обстоятельствах, имеющих место по крайней мере в рамках Европейской статистической системы, нам необходима по возможности более гибкая и несложная для пользователя программа. Такая программа предоставляла бы пользователю необходимые средства и руководство по разработке иерархической структуры

переменных показателей и ее модификации (например, если таблица не должна указывать данные по всем кодам NACE, а только по части из них), а также одновременно недопускала бы ввод неправильных спецификаций, которые могут привести к нарушению конфиденциальности или нарушению процедуры вторичного подавления ячеек. Инструментарий такого рода был бы полезен прежде всего тем пользователям, которые стремятся оптимизировать конструкцию таблиц «играя с данными», вновь и вновь перекраивают таблицы, убирают промежуточные суммарные показатели или вводят новые и так далее, пока, наконец, не получат желаемую таблицу и приемлемую модель подавления. Возможно, это превышает требования «новичков» с маленьким опытом по автоматизированной защите табулярных данных, однако по мере приобретения опыта, особенно если решается задача защиты не одной, а множественных таблиц, или даже с обеспечением систем общественного доступа к статистическим базам данных, пользователи, думается, смогут по достоинству оценить такую гибкость.

22. Наша задача состоит в том, чтобы удовлетворить уже имеющиеся спецификационные стандарты для таблиц, обеспечить несложность программы для пользователя и в то же время не тратить имеющиеся научно-технические ресурсы на «изобретение велосипеда», поскольку уже существуют компьютерные программы, способные оперировать кодами переменных показателей.

IV.1.2 Определение процедур защиты связанных таблиц и базы данных

23. По сравнению с усилиями, необходимыми для создания удобного инструмента для разработки спецификаций иерархических таблиц, как это описывалось выше, усилия, связанные с разработкой программных средств для ввода пользователем спецификаций по процедуре защиты связанных таблиц, можно считать довольно незначительными. (Однако, в этом замечании игнорируется тот факт, что удобные инструменты для спецификации иерархических таблиц типа описанных выше, наряду со значительными усилиями по разработке связанного с этим программного обеспечения, становятся действительно полезными главным образом в контексте защиты связанных таблиц). Для активизации функции защиты группы таблиц пользователю надо будет создать эти таблицы и затем ввести их определения в систему также, как и в случае защиты одной единственной таблицы. Затем он вводит список таблиц для защиты и дело сделано.

24. Аналогичные замечания относительно усилий, необходимых для их создания, можно сделать и в контексте средств для защиты базы данных. Планируется создать опции по спецификации комплекта (защищенных или незащищенных) таблиц, объединенных в один файл. В таком общем файле будет содержаться только одна запись для каждой ячейки в данном комплекте таблиц, в частности для ячеек, присутствующих в более, чем одной таблице. Будут иметься общие метафайлы с хронологической информацией, например, указывающие, какие таблицы уже содержатся в файле, какие из них уже защищены, с указанием для защищенных таблиц данных по использованным параметрам и журналом регистрации, созданном во время их защиты. И, наконец, будут иметься некоторые простые опции для переноса данных из общего файла в готовые таблицы.

25. Хотя это звучит достаточно просто, следует еще раз отметить, особенно в контексте защиты базы данных, что потребуются и другие средства по модернизации проведения такой процедуры, которые в противном случае могли бы и не понадобиться. Мы определим их как:

IV.1.3 Средства предпочтения

26. «Средства предпочтения», как предполагается, предоставят системе возможность предпочесть неподавление определенных ячеек или даже «заставить» определенные ячейки остаться неподавленными, а также наоборот – обеспечить первоочередное подавление определенных ячеек. Будут разработаны опции по определению таких «предпочтений» в автоматическом режиме, наподобие своего рода выключателя для ячеек определенного класса - таких, например, как ячейки промежуточных суммарных показателей, некоторые ячейки из перекрещивающихся частей таблиц, ячейки, использовавшихся в качестве вторичного подавления в предыдущие периоды в таблицах, воспроизводящих результаты периодических обследований, и

т.д. С другой стороны, возможна разработка опций по определению «предпочтений» в отношении определяемых пользователем групп ячеек, где в крайнем случае такая группа может содержать всего одну ячейку. Такие ячейки могут быть специфицированы как (под)таблицы в качестве перекрещивающихся комбинаций подгрупп общих кодовых групп, используемых для определения изначальных таблиц. Конечно, в этом случае должны также иметься средства отбора подгрупп из списка кодов.

27. Конечно, мы должны будем предоставить детальное руководство и параметры по умолчанию для «новичков» и менее опытных пользователей. Если в ходе работы над проектом CASC выяснится, что масштаб проекта (главным образом: имеющиеся возможности по разработке программного обеспечения) недостаточен для создания всех удобных для пользователя средств, которые могут быть полезны, то в этом случае мы по крайней мере попытаемся спроектировать достаточно открытую систему с тем, чтобы опытный пользователь мог вводить свои собственные параметры (извне) или же в ручном режиме выполнять процедуру подавления, отвечающую его специфическим требованиям.

IV.2 Перспективы для разработчиков: методологические стратегии расширения τ-ARGUS

28. В данном разделе рассматриваются стратегии по использованию алгоритма подавления для неструктурированных и иерархических таблиц, по осуществлению защиты связанных таблиц, а также по эффективному расширению защиты связанных таблиц в контексте систем запросов для баз данных.

IV.2.1 Стратегии защиты иерархических таблиц

Однотабличный подход

29. С чисто методологической точки зрения фактически наилучшая стратегия защиты иерархических таблиц заключается в том, чтобы относиться к ним не как к иерархическим и заранее преобразовать такие таблицы в одну единственную (!) неиерархическую таблицу. Это всегда можно и довольно нетрудно сделать. Фактически, это единственный метод надлежащей защиты такой таблицы, предотвращающий описанный ниже риск нарушения конфиденциальности. Задача при этом состоит в обеспечении достаточного ускорения алгоритма подавления для одиночных, неструктурированных таблиц, чтобы он был достаточно мощным для использования в реальных по размеру таблицах. В этом контексте следует отметить, что в не иерархическом виде таблица становится намного, намного больше, нежели в иерархическом. Принимая во внимание тот факт, что вводимый алгоритм подавления на основе методологии линейного программирования представляет собой главным образом набор уравнений, который (для простоты) можно рассматривать как описывающий линейные отношения между (потенциально) подавленными и неподавленными ячейками таблицы, на определенном этапе процесса подавления ячеек конечно придется отбросить из набора уравнений, относящихся к неиерархическим структурам, все незначимые уравнения, - например, идентичные и появляющиеся более одного раза (или даже вообще не создавать таковых).

30. Известно, что проблема вторичного подавления ячеек является сложной с вычислительной точки зрения. Количество вычислений, необходимых для решения задачи вторичного подавления ячеек, известной как задача дискретного линейного программирования (ДЛП), возрастает экспоненциально с увеличением размера таблицы. Поэтому уже заранее ясно, что такая возможность не распространяется на защиту очень больших таблиц, содержащих несколько сотен тысяч ячеек и более, с использованием методологии линейного программирования при таком однотабличном подходе.

Стратегия обратного прослеживания

31. Общепринятый подход заключается в разбивке таблицы на подтаблицы и защите таких подтаблиц по-отдельности. При этом, конечно, необходимо учитывать, что такие подтаблицы

одной и той же таблицы имеют общие ячейки. Иначе может получиться так, что одна и та же ячейка подавляется в одной подтаблице, поскольку используется в качестве вторичного подавления, а в другой таблице остается неподавленной. Пользователь, сравнивающий обе подтаблицы, в этом случае может раскрыть конфиденциальные ячейки в первой таблице. Обычный метод решения этой проблемы состоит в регистрации всех вторичных подавлений, относящихся также к одной из остальных подтаблиц, подавлении их также и в этой таблице, после чего повторяется процедура подавления ячейки для данной таблицы Такой подход иногда называют *«обратным прослеживанием»,* и мы будем придерживаться этого определения и в дальнейшем. Несмотря на то, что в процессе обратного прослеживания подавление ячеек обычно повторяется для каждой подтаблицы несколько раз, необходимое для защиты количество вычислений намного меньше, чем при одноразовой защите всей таблицы. Необходимо, однако, подчеркнуть, что процедура обратного прослеживания не дает стопроцентной гарантии. К примеру, даже если каждая подавленная ячейка таблицы надлежащим образом защищена во всех соответствующих подтаблицах, вполне может случиться так, что показатели ячейки будут вскрыты именно при анализе всех линейных отношений между ячейками всей таблицы. Эту проблему можно проиллюстрировать на примере простой двухмерной таблицы без подструктуры. Мы можем рассматривать такую таблицу как ряд взаимосвязанных одномерных таблиц, например, набор всех строк и столбцов таблицы. Затем мы можем защитить каждую строку и каждый столбец по-отдельности и убедиться в том, что в конечном итоге каждая строка и каждый столбец защищены надлежащим образом, то есть каждая строка и каждый столбец имеют по крайней мере два или ноль подавлений. Однако, это не является достаточным критерием надежной модели защиты для двухмерных таблиц (см. противоположный пример в [3]). Эта же проблема характерна и для *n*-мерных таблиц (см. [8]).

32. Скорость процесса обратного прослеживания можно увеличить, если ячейкам, появляющимся в более чем одной подтаблице, присваивается низкая вероятность их выбора для вторичного подавления, которое может осуществляться, например, с помощью схем динамического взвешивания (см. IV.2.2 «Применение стратегии предпочтения» ниже).

33. Естественный способ разбивки иерархической таблицы на подтаблицы состоит с разбивке таблицы на ряд подтаблиц, не имеющих структуры, например, построенных таким образом: для каждой объясняющей переменной мы выбираем одну конкретную категорию не нижнего уровня. Затем мы создаем «подпеременную». Такая подпеременная состоит только из выбранной на первом этапе категории, а также из тех категорий более низкого иерархического уровня, которые принадлежат к данной категории. Таблица, заданная через такой ряд объясняющих (под)переменных, в этом случае не имеет структуры и представляет собой подтаблицу изначальной таблицы – например, все ячейки в подтаблице принадлежат также и изначальной таблице. Повторяя эту процедуру для любой комбинации категорий не нижнего уровня по каждому измерению, мы разделим первоначальную таблицу на ряд подтаблиц без подструктуры.

34. Вместо создания не имеющих какой-либо подструктуры подтаблиц, мы можем также построить таблицы с менее сложной подструктурой. В этом случае мы не будем строить вышеописанные «подпеременные» для каждой объясняющей переменной, а только для некоторых из них. Таблица, заданная с помощью первоначальных (иерархических) переменных в одних измерениях и подпеременных в других, будет тогда иметь менее сложную структуру по сравнению с первоначальной таблицей.

IV.2.2 Стратегии по защите связанных таблиц и баз данных

Защита связанных таблиц

36. Иногда некоторые таблицы из группы множественных таблиц, публикуемых на основе одного источника (например, данных обследования), перекрещиваются. Допустим, таблица Т1, к примеру, отражает «движение трудовых ресурсов по классам количества работников на предприятиях», таблица Т1.1 отражает « движение трудовых ресурсов по классам количества работников на предприятиях и по классификации КДЕС», а таблица Т1.2 показывает «движение

трудовых ресурсов по классам количества работников на предприятиях и по юридическим формам предприятий». В этом случае T1 является подтаблицей как T1.1, так и T1.2, если все категории «класса количества работников» для таблиц T1.1 и T1.2 идентичны. Ячейка перекрещивающейся таблицы T1 будет чувствительной ячейкой в T.1.1, если и только если данная ячейка является чувствительной также в T1.2.

37. При осуществлении подавления ячейки для T1.1 и T1.2 по-отдельности вполне может оказаться, что в T1.1 будут неподавленные ячейки T1, в то время как в T1.2 они будут являться дополнительными подавлениями, и наоборот. Любой пользователь, получивший доступ к обеим таблицам T1.1 и T1.2, сможет раскрыть эти показатели и, следовательно, перерасчитать значения чувствительных ячеек.

38. Конечно, существуют возможности предотвратить такое развитие событий. Можно защитить «полную» таблицу T1.3: «движение трудовых ресурсов по классам количества работников на предприятиях, КДЕС и юридическим формам предприятий», а в таблицах T1.1 и T1.2 подавить все ячейки, подавленные в T1.3. Теперь предположим, что в результате слишком подробной классификации по количеству работников, эту таблицу невозможно защитить за одну операцию ввиду ограниченных вычислительных ресурсов компьютера. Как и при описанной выше процедуре обратного прослеживания, компьютерная программа прежде всего применит вторичное подавление ячеек, например, к таблице T1.1, затем к таблице T1.2, при этом регистрируя все вторичные подавления в перекрещивающейся таблице T1. Вторичные подавления в T1, проведенные в результате защиты T1.1, будут рассматриваться в качестве первичных подавлений при защите таблицы T1.2, и наоборот. Эта процедура будет повторяться вновь и вновь, пока не будет достигнут такой этап цикличности, когда в T1 не останется ячеек для очередного вторичного подавления. По окончании процесса защиты связанных таблиц, каждая из ячеек перекрещивающейся таблицы T1 будет или подавлена как в T1.1, так и в T1.2, или останется неподавленной как в T1.1, так и в T1.2. Более того, никакие из этих подавленных ячеек нельзя будет раскрыть на основании дополнительных связей между подавленными и неподавленными ячейками в T1.1 или T1.2.

Защита связанных таблиц в контексте систем запроса баз данных.

39. В идеальном варианте процедура защиты связанных таблиц должна применяться к полному комплекту таблиц, публикуемых на основании одного источника данных. Однако похоже, что это становится все менее реальным на практике. В настоящее время процесс предоставления данных все больше диктуется требованиями пользователей и все труднее поддается заблаговременному планированию – даже до такой степени, что внедряются системы запроса баз данных общественного пользования. Пока это еще не сказывается серьезно на подавлении ячеек. Такую ситуацию можно в некоторой степени поправить, если данные сводятся вместе для регистрации подавлений в тех таблицах, которые уже были опубликованы, а другие еще только создаются. Будущие версии τ-ARGUS будут уметь создавать такой “пул” данных. Можно попытаться использовать пул данных, созданный как описано ниже, в качестве основы для систем запросов баз данных общественного или научного назначения. Здесь стоит подчеркнуть, однако, что создание такой системы для запроса баз данных вовсе не входит в программу данного проекта. Мы не заявляем также, что пользователи или разработчики систем запросов баз данных будут полностью удовлетворены уровнем детализированной информации или пропорцией неподавленных ячеек нижнего уровня в таком пуле данных.

40. Пул данных, соответствующий определенной базе микроданных, будет содержать одну и только одну запись по каждой ячейке любой уже защищенной таблицы. В записи будет указан статус подавления ячейки. Если необходимо защитить новую таблицу, программа проанализирует пул данных для всех ячеек такой таблицы. Теперь предположим, что в пуле данных действительно содержатся какие-либо данные по этой ячейке. Если эта ячейка уже использовалась как вторичное подавление в одной из ранее обработанных таблиц, тогда, как и при процессе обратного прослеживания, в новой таблице она будет рассматриваться в качестве первичного подавления. И наоборот, если согласно данным пула ячейка имеет статус неподавленной, то программа будет

стремиться не избирать данную ячейку как вторичное подавление в новой таблице насколько это возможно. Поскольку это будет возможно не всегда, порой пользователь будет вынужден отказаться от новой таблицы, по крайней мере – от ее части, или допустить несоответствие между моделями подавления в уже опубликованной и новой таблице и, следовательно, смириться с риском нарушения конфиденциальности.

Использование стратегии предпочтения

41. Стратегии такого рода будут осуществляться посредством «замораживания» ранее опубликованных ячеек, например, придавая им статус неподпадающих под подавление. Более слабый вариант – по сравнению с полным «замораживанием» этих ячеек – состоит в придании им низкой вероятности их выбора для вторичного подавления. Средства такого рода могут быть обеспечены посредством манипулирования (например, в сторону увеличения) «ценой» или «весом», которые регулярно присваиваются при подавлении таких ячеек.

42. Однако, иногда может быть желательно предпочесть некоторые ячейки как подавления. Если, к примеру, ячейки необходимо включить в таблицу, так как вторичное подавление требует полноты таблицы, содержащей, например, весь комплект линейно-связанных ячеек, хотя эти ячейки и не предназначаются действительно для публикации. Или если ячейка, являющаяся частью перекрещивающейся таблицы из ряда связанных таблиц, подходит или с большой долей вероятности будет отобрана для вторичного подавления в одной из остальных таблиц, которые будут обрабатываться позднее. Это можно осуществить уменьшением регулярных «цен», присвоенных данной ячейке.

43. В качестве еще одного примера использования метода предпочтения представьте себе ситуацию, когда какая-либо таблица публикуется периодически – например, ежемесячно, ежеквартально или ежегодно. Некоторые чувствительные ячейки могут оставаться таковыми при каждой публикации. В качестве простого примера представим себе таблицу без подструктуры с некоторыми чувствительными ячейками, которые являются таковыми «всегда». Предположим также, что существует более одной вероятной модели подавления и что «цены» этих моделей отличаются друг от друга очень незначительно. Если ничего не предпринимать, то очень вероятно, что время от времени модели подавления будут меняться, что было бы нежелательным и могло бы привести к возникновению риска нарушения конфиденциальности, если различие в значениях вторичных подавлений в разные периоды незначительно. В нашем примере эту проблему можно преодолеть с помощью предпочтительного подавления подавлений предыдущих периодов.

Координация моделей подавления как особая задача

44. Особые проблемы появляются, если данные публикуются на разных уровнях региональных классификаций (например, на национальном и сверхнациональном (ЕС) уровне, или на областном и национальном уровне), когда вторичные подавления определяются фактически разными организациями (например, национальным статистическим управлением и Евростатом, или же областными и национальными статистическими органами). Эта проблема, возникающая в результате децентрализованной организации официальной статистики в Европе, может быть решена с помощью «средств предпочтения» в программном обеспечении. Реальные возможности различных подходов по улучшению ситуации будут изучаться с особым акцентом на практическую пригодность предлагаемых методов. Такие методы будут опробированы на нескольких реальных базах данных, имеющихся на национальном и областном уровне. Наиболее многообещающие методы должны поддерживаться данным пакетом программного обеспечения – например, в программу можно включить специальные опции.

V. РАСШИРЕНИЕ ЯДРА ПРОГРАММЫ τ-ARGUS

45. Для расширения программы τ-ARGUS необходимо будет значительно улучшить алгоритм подавления ячеек в существующей версии или же включить новые алгоритмы. В качестве способа

оптимизации информационного содержания защищенных данных должны иметься алгоритмы, поддерживающие предоставление интервалов для подавленных показателей, а также предоставление пертурбированных значений вместо подавленных первоначальных.

V.1 Качество или эффективность: алгоритмы подавления в будущих версиях

46. В настоящей версии отбор вторичных подавлений в программе τ-ARGUS осуществляется с помощью методологии (дискретного) линейного программирования, описанной в [2]. Этот метод, предоставленный Университетом Ла Лагуна, должен быть усовершенствован группой проф. Х.Х. Салазара, чтобы он мог быть применим к иерархическим таблицам при разумных затратах времени на вычисления. Ввиду огромных вычислительных нагрузок, которые указывались выше в IV.2.1, даже с помощью новой версии может стать невозможным его применение к реальным большим многомерным таблицам. Поэтому, одним из возможных решений в будущих версиях может быть стратегия обратного прослеживания, разбивающая таблицу на подтаблицы и применяющая процедуры линейного программирования к подтаблицам по-отдельности в ходе повторяющихся циклов. Однако и это может оказаться очень медленным процессом применительно к большим рядам крупных многомерных таблиц, подробным иерархическим таблицам, которые разрабатываются, например, в результате крупных экономических обследований.

47. Чтобы обеспечить практическую решаемость и таких больших задач программный пакет может включать метод гиперкуба GHQUAR Р.Д. Репсилбера из Института обработки данных и статистики земли Северный Рейн-Вестфалия, Германия. Этот метод описан в [8] и [9], а также более кратко изложен в [4], [5] и [11]. Используя тот факт, что подавленную ячейку в какой-либо *n*-мерной таблице нельзя раскрыть с достаточной точностью, если подавленная ячейка находится в модели подавленных, не нулевых ячеек, образующих угловые точки гиперкуба, этот алгоритм анализирует только модели подавления гиперкуба. В настоящей версии[2] алгоритм GHQUAR разделяет *n*-мерные таблицы с иерархической структурой на соответствующий ряд *n*-мерных подтаблиц без подструктуры (ср. IV.2.1). Эти подтаблицы последовательно защищаются в ходе интерактивного процесса, начиная с высшего уровня. Вместо определенной проверки применимости (потенциальной) модели подавления с помощью решения ряда линейных задач, расчитывается просто нижняя граница количества защиты. Это, конечно же, влияет на эффективность метода. Иногда отвергается «более лучшая» модель подавления, поскольку эта граница находится вне порогового значения, хотя действительное количество защиты и достаточно. Это, наряду с тем фактом, что метод гиперкуба иногда пропускает «наилучшее решение» (поскольку «критерий гиперкуба» является достаточным, но не необходимым критерием надежной модели подавления, и поэтому время от времени «лучшая» модель подавления не представляет собой ряд гиперкубов), приводит к тенденции некоторого переподавления, характерной для этой модели. Экспериментальные результаты (ср. [6], Табл.3, стр.125) выявляют на 30% больше подавлений по сравнению с методом линейного программирования ARGUS.

48. В зависимости от результатов исследований, которые будут проводиться группой проф. Дж. Кастро из Каталонского политехнического университета, будущие версии ARGUS, возможно, будут предлагать алгоритм сетевого потока в качестве альтернативного метода отбора вторичных подавлений.

V.2 Тестирование

49. Для успешного завершения проекта безусловно очень важно, чтобы исследователи хорошо понимали проблемы, которые необходимо решить с помощью разрабатываемых ими алгоритмов, - в противном случае они смогут успешно решить упрощенные задачи и предложить решения, не срабатывающие в сложных реальных ситуациях. Самый важный способ правильно ознакомить их

[2] Сейчас разрабатывается новая версия этого метода, которая сможет оперировать и таблицами с иерархической подструктурой в рамках единой задачи без их расчленения на подтаблицы. ARGUS будет включать и эту новую версию, если она к тому времени будет готова.

с имеющимися проблемами – это предоставить им для изучения примеры из реальной жизни для опробирования новых методов. Такие примеры должны быть как можно более реалистичными. Наша идея заключается в преобразовании реальных, незащищенных табулярных данных в табулярные данные с соответствующими характеристиками оригинальных данных в отношении отбора вторичных подавлений – такими, как структура таблицы, количество и местонахождение первичных подавлений в таблице, объем необходимой защиты при первичном подавлении, количество и местонахождение нулевых ячеек. Таким образом, с одной стороны эти данные должны казаться незащищенными, а сдругой стороны необходимо недопустить возможность раскрыть данные первоначальной таблицы в опубликованном виде (с подавлениями), даже при их непосредственном сравнении. Тот факт, что тестируемые данные совсем необязательно должны быть пригодны для статистического анализа, конечно, облегчает задачу создания таких данных. Мы можем буквально превратить яблоки в апельсины, если нам этого захочется.

50. Любые алгоритмы, разработанные в ходе проекта для отбора вторичных подавлений, будут опробованы на наборе таблиц из такой экспериментальной библиотеки. Результаты по определенным ключевым вопросам (потеря информации в смысле количества и/или общей величины подавлений, и т.д., требуемое время на вычисления) будут регистрироваться. Эта информация может пригодиться для принятия внутренних решений в ходе работы над проектом. В последствии это поможет в распространении пакета, предоставляя потенциальным пользователям информацию по функциональным возможностям программы и конкретным алгоритмам до ее приобретения.

V.3 Методы пертурбации таблицы

51. Как указывалось выше в разделе III.2, используя линейные отношения между опубликованными и подавленными ячейками, пользователи опубликованной таблицы могут в принципе определить верхнюю и нижнюю границы истинного значения любого подавленного показателя. Процедура подавления в τ-ARGUS поэтому должна обспечивать отбраковывание любой модели подавления, если только раскрытие всех этих границ может привести к возникновению риска раскрытия данных индивидуального респондента. Учитывая это распространитель данных, защищающий табулярную информацию с помощью τ-ARGUS, вполне может публиковать также и эти границы вместе с защищенными данными. Поэтому планируется включить средства по расчету и публикации этих границ. Пакет программного обеспечения будет разрабатываться в этом направлении с добавлением средств по вычислению пертурбированных значений, замещающих истинные показатели ячейки. Пертурбированные значения будут находиться между нижней и верхней границами (которые должны быть одобрены) и вписываться в промежуточные и итоговые суммарные показатели защищенной таблицы, таким образом подразумевая сохранение первоначальных дополнительных отношений между ячейками таблицы. Алгоритмы, призванные решать эти вопросы, должны быть разработаны группой проф. К. Луна из Университета Илменау, который, совместно с StBA, также отвечает за выполнение задач, описанных в разделе V.2.

VI. ЗАКЛЮЧИТЕЛЬНЫЕ ЗАМЕЧАНИЯ

52. В данной работе предложена и проиллюстрирована методология, которую необходимо включить в τ-ARGUS в ходе проекта CASC с тем, чтобы решить задачи, указанные в разделе II. Автор, однако, хотел бы подчеркнуть, что на этапе написания этой работы все данные относительно концепции и разработки программы τ-ARGUS являются сугубо предварительными. В ходе работы над проектом CASC будут предлагаться средства, описанные в данной работе, однако эти предложения необходимо будет согласовать с другими участниками проекта. Кроме того, возможности проекта конечно ограничены, в частности с точки зрения ресурсов по разработке программного обеспечения. Это может привести к тому, что будет использована лишь часть предлагаемой здесь методологии, даже если она будет одобрена всеми участниками проекта. В этом случае мы постараемся предложить не столь строгие подходы.

Список литературы

[1] Кокс, Л. (1981), “Меры линейной чувствительности в контроле за соблюдением конфиденциальности в статистике”, Журнал планирования и выводов, 5, 153-164, 1981 г.

[2] Фишетти, М и Салазар, Х.Х. (1988), “Моделирование и решение задачи по подавлению ячеек для линейно-ограниченных табулярных данных”, в Дж. Доминго-Феррер и др., *Защита статистических данных ’98, документы конференции, 25-27 марта 1998 г., Лиссабон, Португалия.*

[3] Гуртс, Дж. (1992), “Эвристика для подавления ячеек в таблицах”, рабочий доклад, Центральное статистическое бюро, Нидерланды.

[4] Гиссинг, С. (1998), “Поиск эффективных автоматизированных систем вторичного подавления ячеек: сравнение программного обеспечения”, Журнал исследований в официальной статистике 2/1998 г.

[5] Гиссинг, С. (1999), “Анализ пакетов для автоматизированного вторичного подавления ячеек”, Федеральное статистическое бюро Германии, документы Рабочего семинара Евростата/ООН-ЭКЕ по конфиденциальности статистических данных, 1999 г.

[6] Гиссинг, С. (1999), “Сравнение систем программного обеспечения для автоматического осуществления вторичного подавления”, *Форум по федеральной статистике, т.31/1999г.: Методы обеспечения конфиденциальности статистических данных,* (на немецком языке).

[7] Джуитт, Р. (1993), “Анализ нарушений конфиденциальности при экономической переписи 1992 г.”, неопубликованная работа. Отделение методов экономической статистики и программирования, Бюро переписи, Вашингтон, О.К.

[8] Репсилбер, Р.Д. (1994), “Сохранение конфиденциальности агрегированных данных”, доклад, представленный на Втором международном семинаре по конфиденциальности в статистике, Люксембург, 1994 г.

[9] Репсилбер, Д. (1999), “Метод прямоугольного параллелепипеда”, *Форум по федеральной статистике, т.31/1999г.: Методы обеспечения конфиденциальности статистических данных,* (на немецком языке).

[10] Робертсон, Д. (2000), “Модернизация программного обеспечения (CONFID) статистического управления Канады для подавления ячеек”, *документы Конференции Компстат 2000, 21-25 августа, Утрехт, Нидерланды.*

[11] Уилленборг Л., де Ваал, Т. (2000) “Элементы контроля за соблюдением конфиденциальности в статистике”, Спрингер, записи лекций по статистике.

Приложение

Возможные верхние границы интервала защиты

Правило чувствительности	Возможная верхняя граница интервала защиты
правило (1,k)	$\frac{100}{k} x_1$
правило (2,k)	$\frac{100}{k}(x_1 + x_2)$
правило (n,k)	$\frac{100}{k}(x_1 + x_2 + \ldots + x_n)$
правило p%	$\frac{p}{100} x_1 + (x_1 = x_2)$
правило (p,q)	$\frac{p}{q} x_1 + (x_1 = x_2)$