

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

Рабочий доклад №3

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

**КОНТРОЛЬ ЗА СОБЛЮДЕНИЕМ КОНФИДЕНЦИАЛЬНОСТИ В СТАТИСТИКЕ НА  
ПРАКТИКЕ:  
НЕКОТОРЫЕ ПРИМЕРЫ ИЗ ОФИЦИАЛЬНОЙ СТАТИСТИКИ  
СТАТИСТИЧЕСКОГО БЮРО НИДЕРЛАНДОВ**

**Запрошенная работа**

Представлена Статистическим бюро Нидерландов<sup>1</sup>

***Аннотация:** В данной работе рассматривается использование двух взаимосвязанных пакетов программного обеспечения для выпуска безопасных с точки зрения риска нарушения конфиденциальности данных. Пакет  $\tau$ -ARGUS используется для табулярных данных, а его близнец  $\mu$ -ARGUS – для микроданных. Основные методы, используемые для защиты чувствительной информации, - это глобальная перекодировка и локальное подавление. Истинные исследователи, которым необходима дополнительная информация, имеют возможность посещать Статистическое бюро Нидерландов и работать на месте в зоне безопасности на территории Статистического бюро Нидерландов. Приведены некоторые примеры из официальной статистики, полученные благодаря использованию методов контроля за соблюдением конфиденциальности статистических данных.*

***Ключевые термины:** микроданные,  $\mu$ -ARGUS, программное обеспечение, контроль за соблюдением конфиденциальности статистических данных, таблицы,  $\tau$ -ARGUS.*

**I. ВВЕДЕНИЕ**

1. Задачей статистических управлений является разработка и публикация статистической информации об обществе. Собранные данные в конечном итоге выпускаются в приемлемой форме для политических деятелей, исследователей и общественности в статистических целях. Выпуск подобной информации может иметь нежелательные последствия, заключающиеся в раскрытии информации об отдельных лицах вместо информации о достаточно больших группах отдельных лиц. Следовательно, возникает вопрос: каким образом можно так модифицировать имеющуюся информацию, чтобы выпускаемые статистические данные были полезны и не нарушали права отдельных лиц на конфиденциальность своих сведений. Для решения проблемы публикации и выпуска как можно более подробной информации без раскрытия сведений об индивидуальных лицах используется теория контроля за соблюдением конфиденциальности статистических данных (Уилленборг и Де Ваал, 1996).

---

<sup>1</sup> Подготовил Эрик Шульте Нордхолт. Мнения, выраженные в данной работе, принадлежат автору и не обязательно отражают точку зрения Статистического бюро Нидерландов.

2. В данной работе рассматриваются имеющиеся инструменты защиты данных и возможность исследователей работать на месте в Статистическом бюро Нидерландов. Таблицы, разрабатываемые Статистическим бюро Нидерландов на базе микроданных обследований, должны быть защищены от риска нарушения их конфиденциального характера. Поэтому к полученным таблицам можно применить пакет программного обеспечения  $\tau$ -ARGUS (Хандепул и др., 1998). Более подробная информация о программе  $\tau$ -ARGUS и возможностях ее применения приводится в разделе II. В разделе III объясняются возможности выпуска микроданных для исследовательских целей и файлов микроданных общественного доступа с помощью пакета программного обеспечения  $\mu$ -ARGUS (Хандепул и др., 1998b). Возможность работы истинных исследователей с более крупными файлами микроданных на месте рассматривается в разделе IV. В разделе V приведены некоторые примеры из официальной статистики, полученные благодаря использованию методов контроля за соблюдением конфиденциальности статистических данных. И в заключение в VI разделе обсуждаются имеющиеся версии и возможности модернизации пакетов ARGUS.

## II. ВЫПУСК ТАБЛИЦ С ПОМОЩЬЮ ПРОГРАММЫ $\tau$ -ARGUS

3. Многие таблицы разрабатываются на основе обследований. Поскольку эти таблицы необходимо защитить от риска раскрытия конфиденциальных данных, для этого можно использовать пакет программного обеспечения  $\tau$ -ARGUS (Хандепул и др., 1998a). Две распространенные стратегии защиты от риска нарушения конфиденциальности заключаются в перестройке таблиц и подавлении индивидуальных показателей. Показатели ячеек в таких таблицах необходимо скрывать, так как публикация этих показателей (их приближенных величин) может привести к их раскрытию. Такие сокрытия называются первичными подавлениями. Для определения подлежащих подавлению ячеек используется правило доминанты. Согласно этому правилу ячейка считается небезопасной для опубликования если  $n$  основных респондентов по этой ячейке составляют не менее  $p$  процентов от суммарного показателя ячейки. Концепция, лежащая в основе этого правила, заключается в том, что по таким небезопасным ячейкам крупнейшие респонденты могут с достаточно большой точностью определить доли своих конкурентов. В программе  $\tau$ -ARGUS в качестве параметра по умолчанию для  $n$  используется число 3, а для  $p$  в качестве параметра по умолчанию используется показатель 70%, однако эти значения можно легко изменять, если пользователь предпочтет какие-либо другие значения. С помощью такого выбранного правила доминанты программа  $\tau$ -ARGUS показывает пользователю какие ячейки являются небезопасными. В публикациях небезопасные показатели ячеек обычно замещаются крестиками (X).

4. Поскольку выпускаются не только показатели ячеек, но и суммарные показатели строк, необходимо подавить и некоторые другие ячейки, чтобы первоначальные подавленные значения нельзя было рассчитать по суммарному показателю. Даже если точно рассчитать значения подавленных ячеек не представляется возможным, часто это можно сделать с достаточным приближением. На практике значение любой ячейки не может быть отрицательным и, следовательно, не может превышать суммарный показатель строки или столбца. Если интервал такого приближения достаточно маленький, необходимо подавить и некоторые другие ячейки для того, чтобы сделать эти интервалы достаточно крупными. Такой интервал называется диапазоном безопасности и в программе  $\tau$ -ARGUS по умолчанию задана нижняя граница диапазона безопасности 70% и верхняя граница 130% от показателя ячейки, однако пользователь по своему усмотрению может менять эти параметры. Пользователь таблицей не может определить, является ли данное подавление первичным или вторичным: обычно все подавленные ячейки обозначаются крестиками (X). Отсутствие информации о причинах подавления ячеек помогает предотвратить раскрытие конфиденциальных данных.

5. Вторичные подавления предпочтительно осуществлять наиболее оптимальным образом, однако определение оптимальности также представляет собой интересную проблему. Часто наиболее оптимальным считается сведение количества вторичных подавлений до минимума. Другие возможности заключаются в сведении до минимума общей суммы подавленных значений или общего числа отдельных слагаемых в значении подавленных ячеек. Сведение до минимума

общего показателя подавленных значений, конечно, имеет смысл только если значения всех ячеек не являются отрицательными. В программе  $\tau$ -ARGUS опция по минимизации суммарного показателя подавленных значений выполняется по умолчанию. В версии 2.0 программы  $\tau$ -ARGUS также можно минимизировать общее число слагаемых в подавленных ячейках. Если предпочтение отдается такому критерию, то для выполнения вторичных подавлений в версии 2.0 программы  $\tau$ -ARGUS используется так называемая переменная цены, равная 1 для каждой записи. Однако сама опция по минимизированию количества вторичных подавлений еще не введена. В будущих версиях программы  $\tau$ -ARGUS предполагается ввести больше опций с тем, чтобы можно было сравнивать различные результирующие группы вторичных подавлений.

6. Если процесс вторичного подавления осуществляется непосредственно для большей части имеющихся подробных таблиц, это обычно приводит к большому количеству локальных подавлений. Поэтому лучше попытаться объединить категории охватывающих (объяснительных) переменных. Таблица, перестроенная посредством сжатия уровней, будет иметь меньшее количество строк или столбцов. При объединении двух безопасных ячеек получается одна безопасная ячейка. При объединении двух ячеек, одна из которых не является безопасной, невозможно предсказать, будет ли результирующая ячейка безопасной или нет, но это можно легко проверить в последствии в программе  $\tau$ -ARGUS. Однако, остающиеся ячейки с большой численностью предприятий имеют тенденцию эффективнее защищать индивидуальные данные и это значит, что за счет сжатия уровней процентное количество неблагонадежных ячеек имеет тенденцию к уменьшению. Следовательно, практическая стратегия защиты таблицы начинается с объединения строк или столбцов. В программе  $\tau$ -ARGUS это можно легко осуществить. Небольшие изменения в объяснительных переменных легче всего осуществить ручным редактированием в блоке перекодировки программы  $\tau$ -ARGUS, а крупные модификации эффективнее осуществляются в отдельном файле перекодировки вне программы, а затем без каких-либо проблем импортируются в  $\tau$ -ARGUS. По завершении процесса реконструирования в программе  $\tau$ -ARGUS можно выполнить локальные подавления при заданных значениях  $n$  и  $p$ , а также нижней и верхней границ диапазона безопасности.

7. Поскольку многие таблицы разрабатываются на основе обследований, а используемое для защиты данных программное обеспечение базируется на отдельных таблицах, остается риск, что хотя каждая ячейка и будет безопасной, сочетание данных в этих таблицах может раскрыть индивидуальные сведения. Это может произойти, если таблицы имеют общие охватывающие и респондентные переменные. Настоящая версия  $\tau$ -ARGUS не поддерживает связанные таблицы. Хотя и имеется возможность защиты таких таблиц, в настоящей версии это не предусматривается. Однако задача состоит в таком расширении программы  $\tau$ -ARGUS, чтобы она могла оперировать важным подклассом связанных таблиц и, в частности, иерархическими таблицами. Иерархическая таблица – это обычная таблица с суммарными показателями строк, а также с дополнительными промежуточными суммарными показателями. В случае иерархических таблиц приходится решать намного более сложные задачи по оптимизации, нежели для одиночных таблиц. Для оптимального решения таких задач существуют некоторые методы приближения. Новые версии программы  $\tau$ -ARGUS с расширениями предполагается разработать в ходе проекта CASC (Вычислительные аспекты конфиденциальности в статистике). Проект CASC финансируется в соответствии с пятой базовой программой Европейского Союза.

### **III. ВЫПУСК МИКРОДАНЫХ ДЛЯ ИССЛЕДОВАНИЙ И ФАЙЛОВ МИКРОДАНЫХ ОБЩЕГО ДОСТУПА С ПОМОЩЬЮ ПРОГРАММЫ $\mu$ -ARGUS**

8. Многие пользователи результатов обследований удовлетворены надежными таблицами, выпускаемыми Статистическим бюро Нидерландов, однако некоторым пользователям необходима дополнительная информация. Для исследовательских целей выпускаются микроданные по многим обследованиям. Пакет программного обеспечения  $\mu$ -ARGUS (Хандепул и др., 1998b) помогает разработать такие микроданные для исследователей. В отношении микроданных для исследовательских целей Статистическое бюро Нидерландов руководствуется следующими правилами:

- 1) Прямые идентификаторы не выпускаются.
- 2) Косвенные идентификаторы подразделяются на экстремально идентифицирующие переменные, очень идентифицирующие переменные и идентифицирующие переменные. Только непосредственно региональные переменные считаются экстремально идентифицирующими. Каждая комбинация значений экстремально идентифицирующей переменной, очень идентифицирующей переменной и идентифицирующей переменной должна встречаться не менее 100 раз среди данного населения.
- 3) Максимальный уровень подробностей по роду занятий, фирме и уровню образования определяется на основании самых подробных прямых региональных показателей. Это правило не заменяет правило №2, а скорее является его продолжением.
- 4) Регион, который можно определить в микроданных, должен иметь не менее 10 000 жителей.
- 5) Если микроданные относятся к перечню лиц, непосредственные региональные данные не выпускаются. Это правило недопускает раскрытия индивидуальных сведений за счет списочного характера микроданных.

9. В отношении большей части коммерческой статистики Статистического бюро Нидерландов предприятия-респонденты по закону об официальной статистике обязаны предоставлять данные Статистическому бюро Нидерландов. Этот закон был принят еще в 1936 г. и переработан в 1996 г. не изменяя обязанности предприятий предоставлять сведения. При опубликовании результатов таких коммерческих обследований недопускается раскрытие индивидуальных данных. Закон запрещает предоставление микроданных таких обследований для исследовательских целей. Таким образом, статистическое бюро Нидерландов может предоставлять два вида информации по таким обследованиям: таблицы и файлы микроданных для общественного пользования. Файлы микроданных для общественного пользования содержат намного меньше подробных показателей, чем микроданные для исследовательских целей. Программа  $\mu$ -ARGUS (Хандепул и др., 1998b) также помогает разработать файлы микроданных для общественного пользования. В отношении файлов микроданных для общественного пользования Статистическое бюро Нидерландов придерживается следующих правил:

- 1) Возраст выпускаемых микроданных должен быть не менее одного года.
- 2) Прямые идентификаторы не выпускаются. Также не предоставляются непосредственные региональные переменные, данные по национальности, стране происхождения и этнической принадлежности.
- 3) Может предоставляться только один вид косвенных региональных переменных (например, размерность класса места проживания). Комбинации значений косвенных региональных переменных должны быть в достаточной степени разбросаны, т.е. каждая область, которую можно выделить, должна иметь не менее 200 000 жителей в искомом населении и, кроме того, должна состоять из муниципалитетов по крайней мере шести из двенадцати провинций Нидерландов. Количество жителей муниципалитета в области, которую можно выделить, должно составлять менее 50% от общей численности населения в этой области.
- 4) Количество идентифицирующих переменных в микроданных не должно превышать 15.
- 5) Чувствительные переменные не выпускаются.
- 6) Должно быть невозможно вывести дополнительную идентифицирующую информацию по объему выборки.
- 7) Не менее 200 000 человек из всего населения должны соответствовать каждому значению идентифицирующей переменной.

- 8) Не менее 1000 человек из всего населения должны соответствовать каждому значению перекрещивающихся двух идентифицирующих переменных.
- 9) Для каждого домашнего хозяйства, из которого в опросе приняли участие более одного человека, общее число домашних хозяйств, соответствующих какому-либо сочетанию значений переменных по домашним хозяйствам, в микроданных должно быть не менее пяти.
- 10) Записи микроданных должны предоставляться в произвольном порядке.

10. В соответствии с этим сводом правил файлы для общественного доступа защищены намного жестче, чем микроданные для исследований. Надо отметить, что в случае микроданных для исследований необходимо проверить определенные трехвариантные комбинации значений идентифицирующих переменных, а в файлах для общественного пользования достаточно проверить двухвариантные комбинации. Однако, в файлах данных для общественного пользования нельзя предоставлять прямые региональные переменные. Если в массиве микроданных для исследовательских целей отсутствуют прямые региональные переменные, то в этом случае, согласно правилам контроля за обеспечением конфиденциальности статистических данных, необходимо проверить только некоторые двухвариантные комбинации значений идентифицирующих переменных. В случае же соответствующих файлов для общественного пользования необходимо проверять все двухвариантные комбинации значений идентифицирующих переменных.

11. С помощью компьютерной программы  $\mu$ -ARGUS можно идентифицировать и защитить небезопасные комбинации в необходимом файле микроданных. Таким образом, с помощью  $\mu$ -ARGUS можно проверить соблюдение правила №2 для микроданных для исследовательских целей и правил №7 и №8 для файлов микроданных для общественного доступа. Глобальная перекодировка и локальное подавление – это два способа защиты данных, которые используются для предоставления безопасных файлов микроданных. В случае глобальной перекодировки несколько категорий идентифицирующей переменной сжимаются в одну. Этот метод применяется ко всему массиву данных, а не только к его ненадежной части, и этим обеспечивается однородная категоризация каждой идентифицирующей переменной.

12. Если какая-либо идентифицирующая переменная желательна во многих категориях это значит, что остальные идентифицирующие переменные могут иметь меньше категорий. В идеальном варианте все идентифицирующие переменные имели бы так мало категорий, что в микроданных больше бы не осталось никаких «опасных» комбинаций и локальное подавление было бы не нужно. При использовании локального подавления подавляется одна или более значений в небезопасных комбинациях, т.е. они замещаются недостающей величиной. Такие недостающие значения можно приписать, но обычно это не практикуется, поскольку неудачные приписки дают пользователям недостоверную информацию, а удачные могут привести к раскрытию конфиденциальных сведений респондентов. Таким образом, локальное подавление ограничивает возможность анализа данных, поскольку больше нет прямоугольных файлов данных для анализа. Однако на практике при разработке защищенных микроданных (микроданных для исследовательских целей или же файлов микроданных для общественного доступа) очень трудно ограничить уровень подробностей в идентифицирующих переменных и часто возникает необходимость в локальных подавлениях для удовлетворения условий по защите данных. Следовательно, после перекодировки идентифицирующих переменных в интерактивном режиме в программе  $\mu$ -ARGUS остальные небезопасные комбинации необходимо защитить посредством подавления некоторых значений. Компьютерная программа  $\mu$ -ARGUS автоматически и оптимально определяет необходимые локальные подавления, т.е. количество значений, которые необходимо подавить, сводится к минимуму. Таким образом можно быстро подготовить микроданные для исследователей, а также файлы микроданных для общественного доступа.

13. Небольшие изменения в идентифицирующих переменных легче всего провести в режиме ручного редактирования в блоке перекодировки программы  $\mu$ -ARGUS, а крупные модификации

эффективнее осуществляются в отдельном файле перекодировки вне программы, а затем без каких-либо проблем импортируются в  $\mu$ -ARGUS. После такой глобальной перекодировки остающиеся ненадежные комбинации подавляются программой  $\mu$ -ARGUS и на выходе обеспечиваются защищенные микроданные. Из этого массива данных уже нельзя получить никакие другие защищенные данные, так как меры по защите данных могут быть обойдены при объединении информации. Поэтому, прежде чем выпускать защищенные микроданные, необходимо тщательно предусмотреть, какие переменные могут быть включены в файлы и каким образом можно перекодировать идентифицирующие переменные, включенные в эти файлы. Такой файл можно получить только один раз.

14. В области микроданных в ходе проекта CASC (Вычислительные аспекты конфиденциальности в статистике) будет рассмотрен ряд новых методов. Проект CASC финансируется по пятой базовой программе Европейского Союза. Такие новые методологии как пост-рандомизация (PRAM), микроагрегирование и добавление шума будут включены в новые версии программы  $\mu$ -ARGUS, которые должны выйти в ближайшем будущем. Это позволит поэкспериментировать этими новыми методами. Для оценки качества применяемых методов также будут разработаны модели по риску нарушения конфиденциальности и потери информации.

#### **IV. РАБОТА НА МЕСТЕ В БЕЗОПАСНОМ УЧАСТКЕ СТАТИСТИЧЕСКОГО БЮРО НИДЕРЛАНДОВ**

15. Некоторым исследователям необходимо больше информации, чем ее содержится в выпускаемых микроданных для исследовательских целей или файлах микроданных для общественного доступа. Поскольку выпуск более подробной информации запрещен, индивидуальные исследователи могут проводить свои исследования с более подробными микроданными в помещении Статистического бюро Нидерландов. Исследователи имеют возможность работать на месте в специальном безопасном участке Статистического бюро Нидерландов. Исследователи могут по своему усмотрению выбрать один из двух офисов Статистического бюро Нидерландов: в Воорбурге на западе страны или в Хеерлене на юге. Выносить какую-либо информацию можно только по разрешению ответственного сотрудника по статистике. Исследователи могут пользоваться как статистическим, так и своим собственным программным обеспечением. Как и все служащие Статистического бюро Нидерландов исследователи, работающие на месте, должны принести присягу в том, что они не будут раскрывать индивидуальные данные респондентов (Куиман, Нобель и Уилленборг, 1999).

16. Исследователи, работающие на месте с экономическими микроданными, должны соблюдать правила Центра Статистического бюро Нидерландов по исследованию экономических микроданных (CEREM). Наиболее важные из них следующие:

- исследователи должны иметь отношение к признанному исследовательскому институту (например, университету);
- должен иметься план исследований, соответствующий действующим научным стандартам;
- исследователь и его руководитель должны подписать гарантии по конфиденциальности;
- исследователь получает доступ только к тем данным, которые необходимы для его исследований;
- данные не должны содержать информации о наименованиях и адресах предприятий;
- данные за последние два года не предоставляются;
- запрещено выносить за пределы Статистического бюро Нидерландов данные или незащищенные промежуточные результаты;

- все планируемые публикации должны проверяться в отношении риска нарушения конфиденциальности;
- все публикации должны быть доступны общественности;
- в публичном регистре регистрируются: имя/имена исследователя/исследователей, исследовательский проект, публикация/публикации и предоставленные базы данных.

17. Помещение предоставляется не бесплатно. Как правило, исследователь должен оплатить затраты за предоставление необходимых данных. Кроме того, взимается пошлина за использование служебного помещения.

## **V. ПРИМЕРЫ ИЗ ОФИЦИАЛЬНОЙ СТАТИСТИКИ**

18. В сентябре 2000 г. Статистическое бюро Нидерландов перешло на новую организационную схему. Большая часть данных теперь разрабатывается отделениями коммерческой статистики и социально-пространственной статистики.

19. В отделении коммерческой статистики самые важные обследования относятся к производственной статистике. На основании этих обследований разрабатывается множество таблиц. Разработка согласованной стратегии защиты таких таблиц нелегкая задача. Для решения этой проблемы используются определенные специально разработанные модули. Идея заключается в интегрировании некоторых таких модулей в программу  $\tau$ -ARGUS, чтобы ими многие могли пользоваться. Некоторые исследователи хотят осуществлять специальные проекты исследований и работать с микроданными производственной статистики на месте в безопасном участке Статистического бюро Нидерландов. В ряде проектов эти микроданные сопоставляются с другими обследованиями. В этих случаях сопоставление производит Статистическое бюро, после чего исследователи могут анализировать полученные массивы данных без прямых идентификаторов.

20. В отделении социально-пространственной статистики проводится множество более мелких обследований. Самым крупным из них является ежегодное обследование занятости и оплаты труда (ASEE). В работе Шульца Нордхольта (2000 г.) описывается процесс сбора данных по оплате труда для ASEE. В массивах данных ASEE содержится огромное количество записей и большая информация по оплате труда.

21. Проблема заключается в том, как обрабатывать связанные таблицы ASEE в настоящей версии  $\tau$ -ARGUS. Поскольку от риска раскрытия конфиденциальных данных необходимо защитить все таблицы, настоящая версия  $\tau$ -ARGUS применяется к трем основным таблицам. Это намного меньше, чем количество публикуемых таблиц, однако многие специальные таблицы можно построить из защищенных базовых таблиц, которые также автоматически станут безопасными. Поэтому остается проблема одновременной защиты различных базовых таблиц. Поскольку решение проблемы оптимального подавления одновременно для двух и более таблиц в настоящей версии программы не предусмотрено, необходимо было найти практичную стратегию защиты.

22. На практике проблема защиты данных для наших связанных таблиц ASEE несколько усложняется двумя факторами. Во-первых, публикуются не только значения ячеек и сводные показатели, но и многочисленные промежуточные суммы. Следовательно, процесс необходимо осуществлять на уровне базовой подтаблицы. Во-вторых, если есть возможность выбора места размещения крестика вторичного подавления, то в соответствии с принятой практикой, крестик помещается в ячейке, которая была подавлена также и в прошлом году. В противном случае базовая подтаблица по каждому году может быть безопасной, однако сочетание таких таблиц за ряд последовательных лет может привести к раскрытию индивидуальных данных. Многие показатели ячеек лишь незначительно меняются из года в год и многие респонденты по этим ячейкам также остаются одними и теми же, на основании чего можно неплохо рассчитать значения

подавленных ячеек, если одна из них оказывается неподдавленной в таблице за предшествующий или последующий год.

23. Сейчас в массиве данных ASEE содержится информация примерно о 50% всех трудовых ресурсов Нидерландов. Задача состоит в увеличении количества записей по оплате труда всех работников в Нидерландах в течение нескольких следующих лет. Для достижения этой цели очень помогло бы использование информации регистра застрахованных лиц, в котором содержится множество сведений и в котором очень хорошо представлен частный сектор. Недостаток этого регистра в том, что количество переменных меньше, чем в ASEE, но эту проблему можно решить с помощью методов приписки (см. к пр. Шульте Нордхолт, 1998). Конечно, более крупное обследование ставит новые задачи в области контроля за соблюдением конфиденциальности статистических данных.

24. Еще одно недавнее новшество – это разработка матриц (агрегированных микроданных), которые публикуются в Statline. Statline – это продукт Статистического бюро Нидерландов для удобного просмотра данных и предоставления пользователям возможности строить свои собственные таблицы. Пользователи Statline могут создать по матрице любую таблицу по своему желанию и поэтому надо быть очень осторожными относительно того, какую информацию можно вводить в такие матрицы. Следовательно, количество (категорий) идентифицирующих переменных на каждую матрицу должно быть ограниченным, а для дополнительной защиты индивидуальных чувствительных данных используется округление. Такие матрицы сейчас разрабатываются для статистики по социальному обеспечению, образованию и занятости.

## VI. ОБСУЖДЕНИЕ

25. Пакеты программного обеспечения  $\tau$ -ARGUS и  $\mu$ -ARGUS появились в ходе выполнения проекта по соблюдению конфиденциальности в статистике (SDC), осуществлявшегося в соответствии с четвертой базовой программой Европейского Союза. Судя по всему, эти компьютерные программы очень помогли с точки зрения практического контроля за соблюдением конфиденциальности. Благодаря программам ARGUS можно решить многие задачи по защите статистических данных. Некоторые из них упоминались в этой работе.

26. Работы (Хандепул и др., 1998а и b) очень помогают пользователям пакетов программ ARGUS. Однако, всегда хочется чего-то еще. В случае  $\tau$ -ARGUS очень помогла бы более автоматизированная обработка связанных таблиц и, в частности, иерархических. Необходимость в этом уже назрела. Фактически, предварительная разработка параметров по обработке связанных таблиц в настоящей версии  $\tau$ -ARGUS уже готова. Отдел статистических методов Статистического бюро Нидерландов сейчас разрабатывает специализированную компьютерную программу (на основе DLL –оптимизации  $\tau$ -ARGUS) для оперирования иерархическими таблицами. Необходимо также провести больше исследований по вопросам защиты от раскрытия данных одного и того же обследования, проводящегося в течение ряда последовательных лет. И, наконец, было бы хорошо иметь больше выбора по способам осуществления вторичных подавлений. С точки зрения  $\mu$ -ARGUS важно более четко разграничить в программе защиту микроданных для исследовательских целей от защиты файлов микроданных для общественного пользования. Поскольку  $\mu$ -ARGUS можно использовать со множеством различных критериев защиты, важно помочь пользователям понять способы осуществления различных стратегий на практике. Недавно проводились исследования метода пертурбации добавлением в микроданные дополнительного стохастического шума. Было бы хорошо, если бы в программе  $\mu$ -ARGUS имелась возможность пертурбации данных как метод защиты.

27. Можно прийти к заключению, что в области контроля за соблюдением конфиденциальности статистической информации еще предстоит провести много исследований. Надеемся, что в скором времени будут выпущены новые версии пакетов ARGUS (включающие результаты текущих исследований). Выпуск этих новых версий является частью проекта CASC (Вычислительные аспекты конфиденциальности в статистике). Проект CASC финансируется в соответствии с пятой базовой программой Европейского Союза. В целях дальнейшей разработки

результатов статистических проектов по четвертой базовой программе Европейского Союза, по пятой базовой программе финансируется также проект AMRADS (Сопутствующие меры в исследованиях и разработках по статистике). Будут проводиться многочисленные курсы и конференции, и в том числе по проблеме контроля за соблюдением конфиденциальности статистических данных.

### Список литературы

Хандепул, А.Дж., Уилленборг, Л.С.Р.Дж., Ван Гемерден, Л., Уэсселз, А., Фишетти, М., Салазар, Х.Х. и Капрара, А. (1998а), *τ-ARGUS, Руководство для пользователя*, версия 2.0.

Хандепул, А.Дж., Уилленборг, Л.С.Р.Дж., Уэсселз, А., Ван Гемерден, Л., Тюрин, С. и Харкенс, С. (1998b), *μ-ARGUS, Руководство для пользователя*, версия 3.0.

Куиман, П., Нобель, Дж.Р. и Уилленборг, Л.С.Р.Дж. (1999), “Защита статистических данных в Статистическом бюро Нидерландов” в *Официальной статистике Нидерландов*, том 14, весна 1999 г., стр. 21-25.

Шулте Нордхолт, Е. (1998), “Приписывание: методы, моделирующие эксперименты и практические примеры” в *Международном статистическом ревю*, том 66, №2, стр. 157-180.

Шулте Нордхолт, Е. (2000), “Контроль в Статистическом бюро Нидерландов за соблюдением конфиденциальности данных по занятости и оплате труда” в *Конфиденциальности статистической информации, документы совместного рабочего семинара Евростата/ООН-ЭКЕ по конфиденциальности статистических данных, состоявшегося в Тессалониках в марте 1999 г.*, Европейские Сообщества, 1999 г., стр. 3-13.

Уилленборг, Л.С.Р.Дж. и Де Ваал, А.Г. (1996), “Контроль за соблюдением конфиденциальности в статистике на практике”, *записи лекций по статистике III*, Спрингер-Верлаг, Нью-Йорк.