

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

Рабочий доклад №15

Тема II: Влияние новых технологических разработок в программном обеспечении, средствах связи и вычислительных процессах на SDC (Контроль за соблюдением конфиденциальности статистической информации)

## **ОПЫТ ОГРАНИЧЕНИЯ РИСКА НАРУШЕНИЯ КОНФИДЕНЦИАЛЬНОСТИ НА ОСНОВЕ МОДЕЛИРОВАНИЯ**

### **Представленная работа**

Представлена организацией «Истат», Италия<sup>1</sup>

**Аннотация:** При анализе отсутствующих ответов респондентов национальные статистические учреждения обычно прибегают к методам приписок на основе статистических моделей. Исследования в этой области ведутся очень интенсивно, поскольку на этом базируется выпуск как можно более точных экономических данных. Задача состоит в том, чтобы на основе опыта, приобретенного в области методологии приписывания данных, восполнить пробел между этой областью исследований и ограничением риска нарушения конфиденциальности статистических данных. В данной работе анализируется приобретенный нами опыт по способам ограничения риска нарушения конфиденциальности на основе моделирования. В целом данные способы позволяют подставлять рассчитанные значения через статистическую модель наблюдаемых значений определенной переменной. Обсуждаются, в частности, обнаруженные проблемы и возможности, присущие двум различным моделям: модели древовидной регрессии (Бриман и др., 1984) для категориальных переменных (Романо и Сери, 2000) и иерархической модели для непрерывных переменных (Франкони и Стандер, 2000).

**Ключевые термины:** Коммерческие микроданные, конфиденциальность, иерархические модели, древовидная регрессия.

### **I. ВВЕДЕНИЕ**

1. В настоящее время в Италии исследователи имеют возможность анализировать экономические микроданные официальной статистики исключительно в помещении «Истата». Это, конечно, не оптимальная ситуация; поэтому идет поиск альтернативных решений на основе как ограничения доступа, так и использования новых методик ограничения риска нарушения конфиденциальности. В данной работе мы рассмотрим только последний вариант.

2. Под коммерческими микроданными мы понимаем базы данных как по крупным, так и малым предприятиям. Ввиду различной структуры таких предприятий массив данных имеет двойной характер: для небольших предприятий это выборка, а для больших предприятий – перепись. По этой причине, а также ввиду высокой узнаваемости крупных предприятий, большинство встречающихся проблем по ограничению риска нарушения конфиденциальности при

---

<sup>1</sup> Подготовили Луиза Франкони, Алессандра Капобианчи, Сильвия Полеттини и Джованни Сери.

распространении коммерческих микроданных относится именно к крупным предприятиям (см. Кокс (1995)).

3. Для предотвращения раскрытия конфиденциальных данных предлагается несколько пертурбативных методов. Из широкого спектра методов матричного маскирования (Кокс, 1994) можно упомянуть методы добавления постороннего шума (Тендик, 1991), перестановки данных (Далениус и Райсс, 1982), микроагрегирования (Дефейз и Нанопоулос, 1992; Доминго Феррер и Матео Санц, 1998). Тем не менее, ограничение риска нарушения конфиденциальности в случае коммерческих микроданных является трудной задачей. Если говорить о методе добавленного шума, то Уинклер (1999) сообщает о неудачной попытке создать безопасные и полезные данные, а использование метода перестановки может значительно исказить коммерческие микроданные.

4. Недавно проводившиеся в «Истате» эксперименты по одноосному микроагрегированию позволили создать массивы микроагрегированных данных из системы счетов предприятий. Однако с точки зрения конечного пользователя методы такого рода не являются вполне удовлетворительными. В основе этой неудовлетворенности лежит тот факт, что применение таких методов может изменить экономические характеристики некоторых единиц до такой степени, что они не будут правильно отражать исходные данные предприятия. По этой причине выпуск микроагрегированных данных рассматривается главным образом только как начальный эксперимент. В ходе других исследований «Истата» в сотрудничестве с Плимутским университетом анализировалась возможность разработки способов ограничения риска нарушения конфиденциальности коммерческих микроданных на основе специальных статистических моделей. Под ограничением риска нарушения конфиденциальности мы имеем ввиду замещение истинного значения определенной переменной значением, полученным в процессе расчета статистической модели.

5. В порядке общепринятой практики национальные статистические учреждения широко применяют к незаполненным опросным листам или полностью отсутствующим данным предприятий методы приписывания. Исследования в этой области ведутся очень интенсивно, поскольку на этом базируется выпуск как можно более точных экономических данных. На наш взгляд большая часть методов приписывания в отношении недостающих данных, если не все эти методы, основываются на статистических моделях (Калтон и Каспржик, 1986). Задача состоит в том, чтобы на основе опыта, приобретенного в области методологии приписывания данных, восполнить пробел между этой областью исследований и ограничением риска нарушения конфиденциальности статистических данных. Естественно, такой шаг подразумевает решение ряда вычислительных и методологических задач. В последнее время национальные статистические учреждения сталкиваются с проблемой множественных приписок (Рубин, 1987) и ищут способы включения этой методологии в процесс разработки официальных статистических данных. Кенникель (1999) сообщает об опыте применения множественных приписок для ограничения риска нарушения конфиденциальности. И хотя результаты не являются вполне удовлетворительными, дальнейшая разработка таких концепций представляется многообещающей областью исследования, Файнберг и др. (1998).

6. В данной работе мы кратко опишем практический опыт, приобретенный «Истатом» в области ограничения риска нарушения конфиденциальности на основе моделей. В частности, мы остановимся на работе Романо и Сери (2000), в которой предлагается модель древовидной регрессии (Бриман и др., 1984) для ограничения риска нарушения конфиденциальности данных Сообщества по обследованию инноваций. Мы также рассмотрим работу Франкони и Стандера (2000), которые предлагают иерархическую модель в рамках бейсиановских моделей с эффектами произвольных участков. Более простой подход, рассматривающий простые регрессии подлежащих защите переменных, представлен в еще одной работе этих же авторов: Франкони и Стандер (2001).

7. Общая характеристика всех таких моделей заключается в простоте подхода. Задача носит двойственный характер: во-первых, изучить возможности, предлагаемые такими методами в отношении ограничения риска нарушения конфиденциальности, а во-вторых обеспечить легкость и простоту их применения в программном обеспечении  $\mu$ -Argus (Уилленборг и Хандепул, 1999) в

качестве составляющей проекта CASC (Вычислительные аспекты конфиденциальности в статистике), финансируемого Европейским Союзом.

8. В Разделе II описываются различные подходы к ограничению риска нарушения конфиденциальности на основе моделей, исходящие из вида коммерческих переменных, используемых при обследовании. В Разделе III обсуждается модель древовидной регрессии, а в Разделе IV мы представляем иерархическую модель. В Разделе V приводятся выводы и предложения по дальнейшей работе.

## **II. РАЗЛИЧНЫЕ ПЕРСПЕКТИВЫ ДЛЯ РАЗЛИЧНЫХ ОБСЛЕДОВАНИЙ**

9. Применение какого-либо метода по ограничению риска нарушения конфиденциальности должно быть тщательно откорректировано в зависимости от типа переменных, присутствующих в коммерческом обследовании. Мы прежде всего разграничиваем различные виды переменных, поскольку риск нарушения конфиденциальности и последующая стратегия его ограничения в значительной степени определяются рассматриваемыми переменными. После этого сравнение обследования инноваций Сообщества (CIS) и ежегодного обследования системы счетов предприятий выявит различия между возможными подходами.

10. Во-первых, существуют переменные, которые на первый взгляд невозможно пертурбировать, поскольку они в этом случае полностью изменяют структуру изучаемого явления. Такими переменными являются классификация NACE (КДЕС) и географические районы предприятий. Учитывая важность таких переменных, единственный способ ограничить с их помощью риск нарушения конфиденциальности заключается в сокращении содержащейся в них информации посредством глобальной перекодировки. Так, например, вместо публикации полной пятизнаковой классификации NACE публикуется только ее двухзнаковый уровень. Это, конечно, зависит от количества предприятий, принадлежащих к этому уровню и, следовательно, от структуры экономики. Что же касается географических районов, обычно пользователи запрашивают самую подробную региональную информацию, однако уровень подробных данных, которые можно сообщить, вновь определяется количеством предприятий, входящих в искомый уровень агрегирования. С точки зрения ограничения риска нарушения конфиденциальности необходима очевидная сбалансированность между классификацией NACE и региональными данными.

11. В целом, следовательно, в отношении применения различных методов ограничения риска нарушения конфиденциальности классификация NACE и географические районы рассматриваются главным образом в качестве многоуровневых переменных. Тем не менее, хотя в отношении географических районов большинство методов ограничения риска нарушения конфиденциальности уже заранее подразумевают модель агрегирования независимо от структурных различий между экономическими районами, методы на основе моделирования могут легко предложить возможные способы агрегирования посредством эффекта фиксированных районов (Франкони и Стандер, 2001) или более сложного эффекта произвольных районов (Франкони и Стандер, 2000).

12. В широком смысле защита коммерческих данных методами приписочного типа может осуществляться различными способами. Можно сохранить нетронутыми структурные переменные, относящиеся к предприятиям, т.е. такие общедоступные переменные, как количество работников, и одновременно смоделировать все прочие переменные, имеющиеся в обследовании. Такой подход предлагается в частности Рубином (1993) с помощью множественного приписывания. Идея заключается в сохранении истинной структуры, т.е. истинной структурной информации о предприятиях, но публикации только смоделированных данных вместо конфиденциальной информации. Таким образом, хотя идентификация предприятия становится несложной при использовании метода сопоставления, результаты таких сопоставлений не нанесут ущерба респондентам. Этот подход является наиболее оптимальным если в обследовании содержатся главным образом количественные переменные, поскольку с чисто интуитивной точки зрения моделирование (а также пертурбирование) меньше затрагивают информационное содержание количественных переменных. Готовятся эксперименты по применению такого

подхода к ежегодным обследованиям системы счетов предприятий, в ходе которых собираются данные годовых балансов.

13. С другой стороны, если большинство присутствующих в коммерческом обследовании переменных являются категориальными, можно воспользоваться более бережным подходом. Фактически для категориальных переменных характерен меньший риск нарушения конфиденциальности, нежели для количественных переменных. В качестве альтернативы поэтому можно применить процесс пертурбирования на основе моделей ко всем переменным, которые прямо или косвенно могут способствовать идентификации предприятия. Все такие переменные являются общедоступными и количественные чувствительные переменные могут дать подсказку по размеру предприятия. Знание таких переменных как оборот, экспорт и затраты, а также общедоступных переменных, может привести к раскрытию конфиденциальных сведений. Обследование инноваций в странах Сообщества, т.е. обследование технологических нововведений на производственных и обслуживающих предприятиях Европы, - является типичным примером обследований такого второго класса. В частности, перед каждым предприятием в этой выборке ставятся вопросы по самым важным экономическим переменным показателям, а также ряд вопросов в отношении нововведений. Многие вопросы по инновациям, которые задаются предприятиям, предусматривают ответ скорее в форме частного мнения, нежели точных числовых показателей. Например, по вопросу о целях инноваций возможны следующие варианты ответов: 0 – не имеет значения, и 1, 2 или 3 в зависимости от степени важности конкретных целей согласно приведенному списку. Как следствие, большая часть характеризующих заинтересованность ответов вряд ли приведет к идентификации. Дополнительная ценность такого подхода заключается в том, что категориальные переменные, относящиеся к инновациям, т.е. переменные, которые могут заинтересовать пользователя, остаются неизменными.

### **III. ЗАЩИТА НА ОСНОВЕ МОДЕЛИРОВАНИЯ: ДРЕВОВИДНАЯ РЕГРЕССИЯ**

14. Использование процедуры, специально предназначенной для обработки и анализа качественных ответов, которые допускают гибкость в группировании и синтезировании категориальных данных, мотивируется введением методики древовидных классификаций (Бримен и др., 1984):

- i) в качестве процедуры группирования;
- ii) в качестве метода классифицирования, рассматриваемого как инструмент приписывания для категориальных данных.

Мы проверили эффективность этих концепций на примере национальных данных из обследования инноваций в Сообществе (CIS).

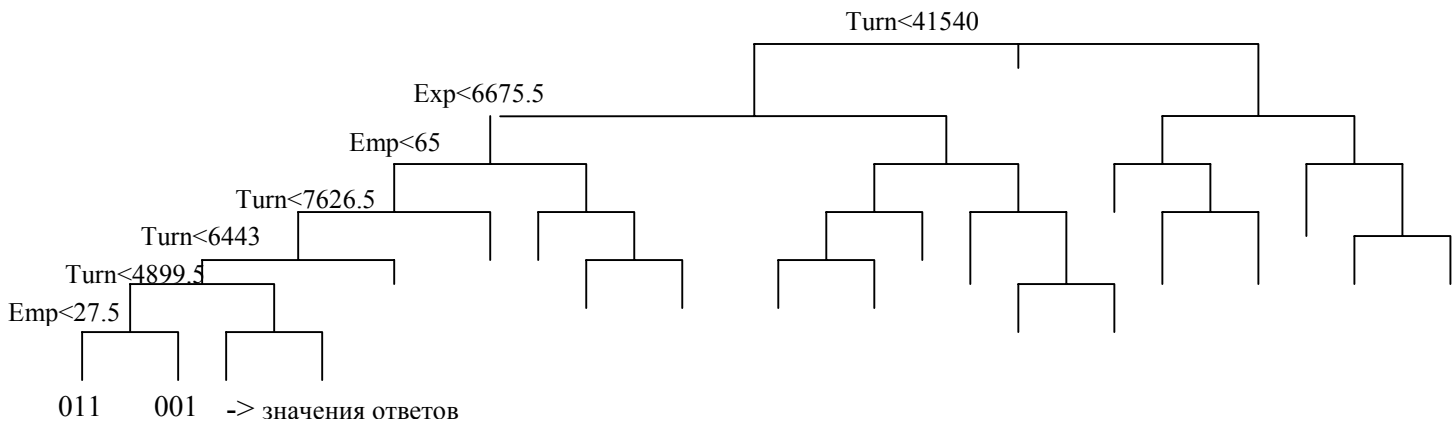
15. В целях ясности мы кратко изложим методы древовидной классификации и регрессии (CART), ряд непараметрических методов, полезных для изучения структуры данных и решения проблем классифицирования при обработке как категориальных, так и количественных переменных.

16. Имея население из  $N$  индивидуальных лиц, для которых обследованы  $M$  объясняющих переменных плюс 1 ответ, метод CART предназначен для разбивки населения на классы, являющиеся однородными с точки зрения переменной ответов респондентов. Классы строятся таким образом, чтобы определенный уровень разброса был минимальным внутри групп и максимальным между группами. Разделение осуществляется иерархически последовательными бинарными расщеплениями, каждое из которых детерминируется критическим значением одной объясняющей переменной. На первом этапе все возможные бинарные расщепления сканируются и выбирается то из них, которое максимизирует однородность ответов в двух образованных подгруппах. Эта же процедура применяется последовательно для каждой подгруппы. Алгоритм продолжает процесс расщепления до получения одной группы или достижения критерия для остановки.

17. Вышеописанный алгоритм расщепления можно представить как бинарное дерево, изображенное на рис. 1. Конечные узлы разветвления дерева обозначают классы искомого расщепления. Каждому узлу разветвления присваивается репрезентативное значение ответов респондентов (например, мода группы, медиана или среднее значение, в зависимости от характера

переменной в ответе), т.е. классифицирующая категория. «Неправильное классифицирование» происходит всякий раз, когда категория индивидуальной единицы отличается от категории, присвоенной группе, к которой эта единица принадлежит. Узел разветвления считается «чистым», если в нем не содержится неправильно отклассифицированных единиц. Более крупные расщепления можно получить за счет «прищипывания» классифицирующего дерева.. «Прищипанное» дерево можно получить удалением узла расщепления (но не корневого узла) и всех его отростков.

**Рисунок 1:** Схема классифицирующего дерева: прищипанное дерево с 21 узлом расщепления



18. Основной задачей данного обследования было изучение отношения предприятий к инновациям. Эти отношения грубо разбивались на категории тремя бинарными вопросами по внедрению новой продукции, новых производственных процессов, и совершенствованию имеющейся продукции и/или процессов соответственно. По крайней мере один положительный ответ характеризует предприятие как стремящееся к инновациям. В модели CART, примененной к данным CIS, ответы респондентов представляют собой булеву комбинацию трех вышеуказанных переменных по инновациям. В качестве объясняющих переменных мы выбрали основные числовые переменные: оборот (Turn), количество работающих (Emp) и объем экспорта (Exp).

19. Задача модели двойственная. Во-первых, при необходимости разработки микроагрегирования метод CART можно использовать для процедуры группирования; для каждой переменной, входящей в модель, в каждой полученной группе можно рассчитать синтезированные показатели (средние объясняющие переменные и прогнозируемое значение ответа респондента). Этот подход отличается от стандартных процедур микроагрегирования для группирования, предназначенного для специального учета категориальных переменных в ответах. Во-вторых, методы CART могут служить способом приписывания. Ответам вновь можно приписать прогнозируемое значение, а критические значения, генерирующие расщепления по объясняющим переменным, можно использовать для публикации интервалов.

20. Самый существенный недостаток обоих подходов – это ошибки при классифицировании. Более того, в первом примере агрегированные значения могут быть недостаточно безопасными для публикации; во втором случае предложенные в модели интервалы могут быть неприменимы непосредственным образом. В работе Романо и Сери (2000) сообщается о дальнейших исследованиях по сравнению качества выпущенных с помощью древовидной регрессии и микроагрегированных данных. Использование метода к данным CIS привело к слишком большому количеству ошибочных классифицирований (свыше 25%); это побудило нас отказаться от использования метода CART для микроагрегирования; однако защита на основе моделей

остается актуальной задачей, требующей дальнейших исследований. Нижеприведенный пример иллюстрирует еще один опыт, приобретенный в этом направлении.

#### IV. ЗАЩИТА НА ОСНОВЕ МОДЕЛИРОВАНИЯ: ИЕРАРХИЧЕСКИЕ МОДЕЛИ

21. Первоначальная концепция - (Франкони и Стандер, 2000) заключалась в улучшении использования интервалов, предлагаемых моделью древовидной регрессии (использование интервалов для ограничения риска нарушения конфиденциальности не является чем-то новым, см., например, (Гопал, Гоуз и Гарфинкель, 1999)). Новая особенность предлагаемого метода заключается в возможности публикации интервала, базирующегося на прогнозируемом распределении для данной статистической модели; см. пример бейсианской модели с использованием прогнозируемого распределения (Дункан и Ламберт, 1986). Предлагаемая модель является обычной авторегрессией с вариацией логарифма переменной (оборот), ковариацией логарифма переменной (экспорт) и логарифма (работников), независимо от того, относятся ли инновации на предприятии к продукции или процессам и относится предприятие к группе и соответствующему уровню классификации NACE или нет. В модели также используется географический район, к которому принадлежит предприятие. Эта географическая переменная вводится в модель посредством как структурированных, так и неструктурированных произвольных эффектов на том предположении, что смежные районы должны иметь близкие значения. Это достигается модификацией условной авторегрессионной процедуры, обсуждавшейся, например, в (Бесаж и др, 1991) и (Моллье, 1996). Дополнительным результатом этого метода является лучшее понимание географической структуры, лежащей в основе данных, благодаря пространственным моделям. Такое понимание географических эффектов подсказывает использование более широкой категоризации при распространении общедоступных описательных географических переменных, которые вряд ли ведут к уменьшению потери информации.

22. Выводя следствия из этой модели Франкони и Стандер (2000) используют распределение Гиббса. Распределение Гиббса является примером марковского процесса алгоритма Монте-Карло; далее см., к примеру, (Гилкс и др., 1996). Простая причина такого выбора кроется в простоте его осуществления. Из распределения Гиббса мы получили последовательность векторов  $\mathbf{G} = 1000$  с параметрами  $\theta^{(i)}$  для нашей модели. Для устранения влияния начальных условий на процесс мы отбрасываем первые  $\mathbf{B} = 500$  значений. Логический вывод основывается на данной последовательности. Распространяемые данные базируются на прогнозируемой плотности  $p(\mathbf{y}^{\text{новы́е}} | \text{данные})$ , где  $\mathbf{y}^{\text{новы́е}}$  – предсказанное значение вектора лог(оборот). Реализацию данной прогнозируемой плотности можно легко получить формированием вектора из  $p(\mathbf{y}^{\text{новы́е}} | \theta^{(i)})$  для каждой  $t = \mathbf{B}+1, \dots, \mathbf{G}$ . Таким образом, для каждого наблюдаемого параметра мы получаем вектор реализаций  $(y_{ij}^{(\mathbf{B}+1)}, \dots, y_{ij}^{(\mathbf{G})})$  из соответствующей прогнозируемой плотности.

Из этого вектора можно получить прогнозируемый интервал  $(1 - \gamma)\%$  посредством его

сортировки и вычисления значения  $\mathbf{th}$ -ных элементов:  $\mathbf{floor} \left\{ \frac{\gamma}{2} (\mathbf{G} - \mathbf{B}) \right\}^{\mathbf{th}}$  и  $\mathbf{ceiling} \left\{ \left( 1 - \frac{\gamma}{2} \right) (\mathbf{G} - \mathbf{B}) \right\}^{\mathbf{th}}$

где  $\mathbf{floor}(x)$  и  $\mathbf{ceiling}(x)$  - ближайшее целое меньшее (или большее) чем  $x$ .

23. Чтобы защитить истинный показатель оборота и, следовательно, уменьшить вероятность идентификации предприятия, мы предлагаем вместо этого публиковать соответствующие интервалы. Конечно, на основе прогнозируемого интервала, в пределах которого может находиться показатель оборота предприятия, можно рассчитать его истинное значение, например, по средней точке интервала. Возможно, лучше публиковать точечный ряд прогнозируемой плотности вместо интервала. Примерами такого точечного ряда могут служить прогнозируемые средние и усредненные показатели.

24. Данная модель применялась к выборке микроданных CIS Италии, соответствующих двум различным секторам NACE: сектору 18 (производство одежды) и сектору 28 (производство металлопродукции). Анализ защиты с помощью этого метода выявил более благоприятные результаты, нежели результаты, полученные с помощью одноосного микроагрегирования, где рассматривается только переменная оборота. Тем не менее, эксперимент по сопоставлению

потребовал бы также наличия общедоступных переменных по количеству работников. Возможным способом распространения этой переменной в данном контексте было бы использование интервала. Результаты были обнадеживающими, но не до конца удовлетворительными. Результаты улучшаются при использовании одной модели для каждой подлежащей защите переменной, как это предлагается в работе Франкони и Стандера (2001).

## **V. ВЫВОДЫ И ДАЛЬНЕЙШИЕ ИССЛЕДОВАНИЯ**

25. В данной работе мы обсуждаем ограничение риска нарушения конфиденциальности на основе моделей и аргументируем различные стратегии защиты. В целом, основные значения количественных переменных создают ряд проблем в отношении ограничения риска. Как необычайно большие, так и необычайно маленькие значения легко опознаются специалистами в соответствующей отрасли. Ограничение риска нарушения конфиденциальности на основе моделей решает эту проблему лишь частично. Практический подход мог бы предполагать использование модельного метода ограничения риска с последующим применением дополнительных способов защиты данных для тех предприятий, уровень безопасности данных которых не является вполне удовлетворительным. Однако было бы целесообразно разработать общую модель, которая могла бы автоматически решать все проблемы, характерные для коммерческих микроданных. Другие важные вопросы, имеющие жизненно важное значение для ограничения риска нарушения конфиденциальности, относятся к оценке уровня безопасности распространяемого файла данных и количественный расчет возможных искажений и потерь информации в защищенных данных. Оба эти вопроса подлежат дальнейшим исследованиям в рамках проекта CASC. Первый из них предполагает изучение более сложных способов привязки записей, а второй – изучение матрицы для оценки различных способов пертурбирования.

26. Работы, проводившиеся в «Истате», выявили возможности, проблемы и ограничения защиты данных на модельной основе. Они также предопределили другие и более радикальные способы защиты файлов коммерческих микроданных. В действительности распространение единственного файла, защищенного с помощью модельного метода ограничения риска нарушения конфиденциальности или какого-либо другого пертурбативного метода, неизбежно приведет к недооценке исходных переменных в массиве данных. Однако, чтобы сохранить такую информацию национальные статистические учреждения должны быть готовы к применению сложных методов моделирования, а пользователи должны быть готовы смириться с публикацией ряда смоделированных массивов данных по одному и тому же обследованию. Планируется проведение дальнейших экспериментов по определению возможностей создания файлов псевдомикроданных с помощью множественного приписывания. Это необходимо для того, чтобы удостовериться в возможных преимуществах моделирующего подхода и последствиях его применения для конечного пользователя.

### **Благодарность**

Данная работа частично финансировалась проектом IST-2000-25069 CASC (Вычислительные аспекты конфиденциальности в статистике) Европейского Союза.

Выраженные в работе мнения принадлежат авторам и не обязательно отражают точку зрения Национального статистического института Италии.

### **Список литературы**

Бесажа, Ж., Йорк, Дж. и Моллье, А. (1991), Бейссианово восстановление изображения с двумя примерами из пространственной статистики (с обсуждениями), *Анналы Института статистической математики*, 43, 1-59.

Бриман Л., Фридман Дж.Х., Олсен Р.А. и Стоун С.Дж. (1984), *Древовидное классифицирование и регрессия*, «Уодзворт Интернэшнл Групп».

Кокс, Л.Х. (1994) Методы матричного маскирования для ограничения риска нарушения конфиденциальности микроданных. *Методология исследований*, 20, 165-169.

Кокс, Л.Х. (1995) Защита конфиденциальности в коммерческих обследованиях. В *Методах коммерческих обследований* (ред. Б. Г. Кокс, Д.А. Биндер, Б.Н. Чиннаппа, А. Кристиансон, М.Дж. Коллидж и П.С. Котт), стр. 443-473. Нью-Йорк: «Уайли».

Далениус, Т. и Райсс, С.П. (1982) Перестановка данных: способ контроля за соблюдением конфиденциальности. *Журнал статистического планирования и выводов*, 6, 73-85.

Дефейз, Д., Нанопулос, Ф., (1992), Списки предприятий и конфиденциальность: метод малых агрегатов, *Документы Симпозиума-92 по статистике Канады, Планирование и анализ продольных обследований*, 195-204.

Доминго-Феррер, Дж., Матео-Санц, Дж.М., (1998), Практическое, ориентированное на данные микроагрегирование для контроля за соблюдением конфиденциальности в статистике, *Report de Recerca*, DEI-RR-98-005, Факультет технической информации, Университет Ровира-и-Виргили, Испания.

Дункан, Г.Т. и Ламберт, Д. (1986). Распространение ограничивающих риск нарушения конфиденциальности данных (с обсуждениями). *Журнал Американской статистической ассоциации*, 81, 10-28.

Франкони, Л. и Стандер, Дж. (2000). Ограничение риска нарушения конфиденциальности коммерческих микроданных на основе моделей. *Материалы II Международной конференции по институциональным исследованиям*, 17-21 июня 2000 г., Буффало, Нью-Йорк. В процессе публикации.

Франкони, Л. и Стандер, Дж. (2001). Микроагрегирование и ограничение риска нарушения конфиденциальности на основе моделирования применительно к коммерческим микроданным. Работа передана для публикации.

Файнберг, С.Е., Маков, У.Е. и Стил, Р.Дж. (1998). Ограничение риска разглашения с помощью пертурбационных и аналогичных методов для категориальных данных. *Журнал официальной статистики*, 14, 485-502.

Гилкс, У.Р., Ричардсон, С. и Спигельхолтер, Д.Дж. (1996). Внедрение монтекарловского алгоритма марковской цепи. *Монтекарловский алгоритм марковской цепи на практике* (ред. У.Р. Гилкс, С. Ричардсон и Д.Дж. Спигельхолтер), стр.1-19. Лондон: «Чэпмен энд Холл».

Гопал, Р., Гоуз, П. и Гарфинкель, Р. (1999). Конфиденциальность через камуфлирование: SVC подход к управлению запрошенными базами данных. *Материалы Конференции по защите статистических данных*, 25-27 марта 1998 г., Лиссабон, стр. 19-28.

Калтон, У., и Каспржик, Д. (1986). Обработка недостающих данных обследований. *Методология обследований*, 12, 1-16.

Кенникелл, А.Б. (1999). Множественное приписывание и защита от нарушения конфиденциальности, *Материалы Конференции по защите статистических данных*, 25-27 марта 1998 г., Лиссабон, стр. 381-400.

Моллье, А. (1996). Бейсиановское картографирование болезней. *Марковский процесс алгоритма Монте-Карло на практике* (ред. У.Р. Гилкс, С. Ричардсон и Д.Дж. Спигельхолтер), стр.359-379. Лондон: «Чэпмен энд Холл».

Романо, Д. и Сери, Дж. (2000). Применение методов древовидной регрессии для защиты исходных данных предприятий. *40-е научное заседание Итальянской статистической ассоциации*, Флоренция, 26-28 апреля 2000 г. В процессе публикации.

Рубин, Д.Б. (1987). *Множественное приписывание отсутствующих данных в обследованиях*. Нью-Йорк: «Уайли».

Рубин, Д.Б. (1993). Обсуждение ограничения риска нарушения конфиденциальности в статистике. *Журнал официальной статистики*, 9, 461-468.

Тендик, П. (1991). Оптимальное добавление шумов для сохранения конфиденциальности многовариативных данных. *Журнал статистического планирования и выводов*, 27, 342-353.

Уилленборг, Л. и Хандепул, А. (1999). ARGUS: программное обеспечение по проекту SDC. *Конфиденциальность статистических данных: Материалы Совместного рабочего семинара Евростат/ООН-ЭКЕ по конфиденциальности статистических данных*, 8-10 марта 1999 г., Тессалоники, стр.87-98.

Уинклер, У., И. (1999), Методы переидентифицирования для оценки конфиденциальности аналитически значимых микроданных. *Материалы Конференции по защите статистических данных*, 25-27 марта 1998 г., Лиссабон, стр. 319-335.