

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

Рабочий доклад №5

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

## **ЭМПИРИЧЕСКОЕ СРАВНЕНИЕ МЕТОДОВ SDC ДЛЯ НЕПРЕРЫВНЫХ МИКРОДАНЫХ С ТОЧКИ ЗРЕНИЯ ПОТЕРИ ИНФОРМАЦИИ И РИСКА НАРУШЕНИЯ КОНФИДЕНЦИАЛЬНОСТИ**

### **Представленная работа**

Представлена Университетом Ровира-и-Виргили, Каталония, Испания<sup>1</sup>

**Аннотация:** Мы представляем в этой работе первый опыт эмпирического сравнения методов SDC в отношении непрерывных микроданных. На основе экспериментов по переидентификации мы делаем попытку оптимизировать сбалансированность между потерей информации и риском нарушения конфиденциальности. Рассматриваемые методы SDC включают ввод дополнительного шума, искажение за счет распределения вероятностей, микроагрегирование, перевыборку, перестановку показателей местами, а также новый подход, основанный на допускаящем потери сжатии данных. Определяются меры по измерению суммарных потерь информации (не предназначенной для специфических пользователей данными) и используются два метода эмпирической переидентификации: евклидова привязка записи и вероятностная привязка записи.

**Ключевые термины:** Контроль за соблюдением конфиденциальности статистических данных, непрерывные микроданные, привязка записи, эксперименты по переидентификации, измерение потерь информации.

### **I. ВВЕДЕНИЕ**

1. В данной работе описываются некоторые предварительные результаты проекта OTTILIE-R (Оптимизация сбалансированности между потерями информации и риском нарушения конфиденциальности для непрерывных микроданных), финансируемого Бюро переписи США и осуществляемого Университетом Ровира-и-Виргили.

2. Целью экспериментальной работы, проводящейся по проекту OTTILIE-R, является демонстрация методологии по оптимизации сбалансированности между потерями информации и риском нарушения конфиденциальности. При этом используется следующий подход:

- *Анализ литературы.* Изучена литература по SDC для микроданных с целью идентификации методов, необходимых для защиты непрерывных микроданных. Кроме того, введен SDC для непрерывных микроданных на основе допускаящего потери сжатия.
- *Данные для эксперимента.* Данные для эксперимента получены из общедоступных файлов микроданных.

---

<sup>1</sup> Подготовили Джозеп Доминго-Феррер и Джозеп М. Матео-Санц, Факультет вычислительной техники и математики (e-mail {jdomingo, jmateo}@etse.urv.es).

- *Оценка риска нарушения конфиденциальности.* Для определения риска нарушения конфиденциальности, связанного с конкретным методом SDC, использовались два алгоритма по привязке записи. Кроме того, был определен интервал для возможного нарушения конфиденциальности.
- *Определение метрики.* Потери информации в действительности определяются назначением используемых замаскированных данных. Поскольку виды использования данных находятся вне рамок проекта OTTILIE-R, мы определили ряд характерных робастных метрик потери информации, с помощью которых можно попытаться обнаружить структурные различия между файлами исходных и скрытых данных.
- *Эмпирическая работа.* Проведенные эксперименты были нацелены на получение группы характеристик (*метод, пармы, риск, потеря*), где *пармы* – вводимые параметры для *метода*, *риск* – процент переидентифицированных записей в тестируемом массиве данных, а *потеря* – потери информации.

3. В разделе II рассматриваются соответствующие методы SDC для защиты непрерывных микроданных. В разделе III приводится перечень мер по измерению потерь информации, которые учитывались при проведении экспериментальных работ. В разделе IV описаны подходы по привязке записей для оценки риска нарушения конфиденциальности. В разделе V приводятся фактические результаты сравнения. Раздел VI – выводы.

## II. СООТВЕТСТВУЮЩИЕ МЕТОДЫ SDC ДЛЯ НЕПРЕРЫВНЫХ МИКРОДАНЫХ

4. Методы выборки вполне пригодны для категориальных микроданных, однако их адекватность в отношении непрерывных микроданных в общем сценарии нарушения конфиденциальности менее очевидна. Причина этого в том, что такие методы оставляют непрерывную переменную непertурбированной для всех индивидуальных респондентов в выборке. Таким образом, если переменная  $V_i$  присутствует во внешнем административном общедоступном файле, очень высока вероятность ее идентификации по уникальности, так как для непрерывной переменной (даже в числовом представлении) крайне маловероятно, что  $V_i(o_1) = V_i(o_2)$  если  $o_1 \neq o_2$ . Поэтому мы будем рассматривать только пертурбативные методы.

5. Рассматриваемые пертурбативные методы представляют собой подряд значимых для непрерывных микроданных методов:

- *Добавляемый шум* (сокращенно: Noisep). Гауссов шум добавляется к оригинальным данным, чтобы получить скрытые данные [Kim86]. Если стандартное отклонение оригинальной переменной составляет  $s$ , то шум генерируется с помощью  $N(0,ps)$ . Значения  $p$ , рассматриваемые в нижеописанном эксперименте, составляют 0.01, 0.02, 0.04, 0.06, 0.08 и до 0.2 с инкрементом 0.02.
- *Искажение данных посредством распределения вероятностей* (сокращенно: Distr, [Liew85]). Для каждой переменной из оригинальных переменных найдено наиболее оптимальное распределение; затем подходящее распределение используется для генерирования массива замаскированных данных. Параметры отсутствуют.
- *Перевыборка.* Берем  $t$  независимых выборок  $X_1, \dots, X_t$  значений оригинальной переменной  $V_i$ . Сортируем все выборки с помощью одного и того же критерия ранжирования. Строим замаскированную переменную  $V'_i$  беря в качестве первого значения среднее арифметическое первых значений выборок, в качестве второго – среднее арифметическое вторых значений и так далее. Перевыборка проверялась для  $t=1$  (Resamp1) и  $t=3$  (Resamp3).
- *Микроагрегирование.* Записи распределяются на небольшие агрегаты или группы размером не менее  $k$  [Defa93, Domi02]. Вместо публикации переменного показателя конкретного индивидуального респондента публикуется среднее значение переменных в группе, к которой данный респондент принадлежит. Рассмотренные варианты микроагрегирования включают: индивидуальное ранжирование (MicIRk); микроагрегирование предполагаемых данных с помощью z-счетной проекции (MicZk) и проекции основных компонентов (MicPCPk); микроагрегирование предполагаемых многовариантных данных при одновременном рассмотрении двух переменных (Mic2mulk), при одновременном рассмотрении трех переменных (Mic3mulk), четырех переменных (Mic4mulk), или всех переменных одновременно (Micmulk). Рассматривались значения  $k$  от 3 до 10.

- *Допускающее потери сжатие (JPEGq)*. Этот метод является новым и предложен данными авторами для непрерывных микроданных. Суть состоит в рассмотрении файла числовых микроданных в качестве изображения (где строки – это записи, а столбцы – переменные). Затем к изображению применяется допускающее потери сжатие или, в частности, алгоритм JPEG [JPEG], и сжатое изображение интерпретируется как файл замаскированных данных. В зависимости от используемого алгоритма допускающего потери сжатия необходимо ввести соответствие между рядами переменных и цветовым спектром. Качество  $q$  JPEG было выбрано в качестве параметра со значениями от 5% до 100% с инкрементом 5%.
- *Перестановка рангов (Rankp)*. Хотя этот метод первоначально был описан только для порядковых переменных, его можно использовать для любых числовых переменных [Moog96]. Первые значения  $V_i$  ранжируются в возрастающем порядке; затем каждое отранжированное значение  $V_i$  меняется местами с другим произвольно выбранным отранжированным значением в ограниченном диапазоне (например, ранги двух переставленных значений не могут отличаться более, чем на  $p\%$  от общего количества записей). При проведении экспериментов рассматривались следующие значения  $p$ : 1, 2, 3, 4, 5, 6, 7 и 10.

### III. МЕРЫ ПОТЕРЬ ИНФОРМАЦИИ

6. Для оценки потерь информации, вызванных применением метода SDC к массиву непрерывных микроданных, мы хотели изучить насколько отличается массив замаскированных данных от массива оригинальных данных. Мы исходим из того, что потери информации незначительны, если структура массива замаскированных данных мало отличается от структуры массива оригинальных данных. Фактически, мотивация сохранения структуры массива данных заключается в обеспечении аналитической пригодности и полезности массива замаскированных данных. В действительности мы можем опробовать несколько дополнительных способов для оценки сохранности структуры оригинального массива данных:

- Сравнением данных в оригинальном и замаскированном массивах данных. Чем больше метод SDC походит на функцию идентификации, тем меньше различий (но выше риск нарушения конфиденциальности!).
- Сравнением некоторых статистических показателей, рассчитанных на основе оригинального и замаскированного массивов данных.

7. Допустим, что  $X$  и  $X'$  являются оригинальным и замаскированным массивами данных. Допустим, что  $V$  и  $V'$  – это ковариантные матрицы  $X$  и  $X'$  соответственно; аналогично допустим, что  $R$  и  $R'$  – это корреляционные матрицы. В таблице 1 приводятся предлагаемые метрики. В этой таблице:  $p$  – количество переменных,  $n$  – количество записей, а компоненты матриц представлены соответствующими строчными буквами (например:  $x_y$  – компонент матрицы  $X$ ). В отношении мер  $X-X'$  также имеет смысл рассчитывать их по средним значениям переменных, а не по всем данным (см. строку  $\bar{X} - \bar{X}'$  в таблице 1). Аналогично, в отношении метрик  $V - V'$  также имеет смысл сравнивать только вариации переменных, т.е. сравнить диагональ ковариантных матриц, а не все матрицы полностью (см. строку  $S - S'$  в таблице 1).

Таблица 1. Измерение потерь информации

	Средняя квадратная погрешность	Средняя абсолютная погрешность	Среднее отклонение
$X-X'$	$\frac{\sum_{j=1}^p \sum_{i=1}^n (x_{ij} - x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^p \sum_{i=1}^n  x_{ij} - x'_{ij} }{np}$	$\frac{\sum_{j=1}^p \sum_{i=1}^n \frac{ x_{ij} - x'_{ij} }{ x_{ij} }}{np}$
$\bar{X} - \bar{X}'$	$\frac{\sum_{j=1}^p (\bar{x}_j - \bar{x}'_j)^2}{p}$	$\frac{\sum_{j=1}^p  \bar{x}_j - \bar{x}'_j }{p}$	$\frac{\sum_{j=1}^p \frac{ \bar{x}_j - \bar{x}'_j }{ \bar{x}_j }}{p}$

V-V'	$\frac{\sum_{j=1}^p \sum_{l \leq i \leq j} (v_{ij} - v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{l \leq i \leq j}  v_{ij} - v'_{ij} }{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{l \leq i \leq j} \frac{ v_{ij} - v'_{ij} }{ v_{ij} }}{\frac{p(p+1)}{2}}$
S-S'	$\frac{\sum_{j=1}^p (v_{jj} - v'_{jj})^2}{p}$	$\frac{\sum_{j=1}^p  v_{jj} - v'_{jj} }{p}$	$\frac{\sum_{j=1}^p \frac{ v_{jj} - v'_{jj} }{v_{jj}}}{p}$
R-R'	$\frac{\sum_{j=1}^p \sum_{l \leq i \leq j} (r_{ij} - r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{l \leq i \leq j}  r_{ij} - r'_{ij} }{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^p \sum_{l \leq i \leq j} \frac{ r_{ij} - r'_{ij} }{ r_{ij} }}{\frac{p(p-1)}{2}}$

#### IV. МЕРЫ РИСКА НАРУШЕНИЯ КОНФИДЕНЦИАЛЬНОСТИ

8. Оценка качества метода SDC не может ограничиваться потерями информации; риск нарушения конфиденциальности – это еще одна величина, которую необходимо измерить. Самым лучшим оказывается метод, который оптимизирует сбалансированность между обеими этими показателями в соответствии с требованиями пользователя.

9. Литература по риску нарушения конфиденциальности в основном рассматривает методы выборки и публикацию части оригинального массива данных. Риск нарушения конфиденциальности измеряется как вероятность того, что уникальность в выборке является уникальностью для населения [Skin94]. Если размер выборки равноценен численности всего населения, то такая вероятность может быть опасно высокой; в этом случае взломщик, который обнаруживает уникальный показатель в опубликованной выборке, может быть почти уверен в наличии только одного индивидуального респондента во всем населении с таким показателем. Это может привести к идентификации такого респондента.

10. Вышеуказанное свойство уникальности уже не является значимым для пертурбативных методов, поскольку в этом случае публикуется весь массив микроданных, но с некоторыми искажениями. Литературы по риску нарушения конфиденциальности, которую можно было бы использовать для широкого класса пертурбативных методов, немного; определение риска нарушения конфиденциальности обычно зависит от конкретных методов (измерения, описанные в [Adam89] по-прежнему актуальны). Эмпирические методы типа метода привязки записей предполагают более согласованный подход к оценке риска нарушения конфиденциальности для пертурбативных методов. Ниже мы кратко опишем два подхода к привязке записей и один способ определения меры (интервала) риска нарушения конфиденциальности.

##### IV.1 Дистанционная привязка записей

11. Данный подход по привязке записей описан в [Pagl98] для частного случая маскирования микроагрегирования при использовании евклидовой метрики. Однако его можно сделать более общим для любого пертурбативного метода при условии, что расстояние между оригинальным и замаскированным значением поддается определению. Как и в любом другом контексте привязки записи предполагается, что взломщик имеет внешний массив данных, содержащий в качестве ключевых переменных те же самые переменные, которые присутствуют в опубликованном массиве замаскированных данных. Предполагается, что взломщик попытается связать массив замаскированных данных с внешним массивом данных.

12. Привязка осуществляется посредством расчета расстояний между записями в оригинальном и замаскированном массиве данных. Используются стандартизированные расстояния, чтобы избежать масштабирования. Для каждой записи в массиве замаскированных

данных рассчитывается расстояние до каждой записи в оригинальном массиве данных. Затем анализируются «ближайшая» и «вторая ближайшая» записи в оригинальном массиве данных. Запись из массива замаскированных данных помечается как «привязанная», если ближайшая запись в оригинальном массиве данных имеет тот же номер, что и соответствующая оригинальная запись. Запись из массива замаскированных данных помечается как «привязанная ко 2-ой ближайшей», если вторая ближайшая запись в оригинальном массиве данных имеет тот же номер. Во всех остальных случаях запись в массиве замаскированных данных помечается как «непривязанная». Процент «привязанных» и «привязанных ко второй ближайшей» является мерой риска нарушения конфиденциальности.

## IV.2 Вероятностная привязка записи

13. В [Jago89] метод вероятностной привязки записи описан и проиллюстрирован на примере переписи 1985 г. в Тампе, Флорида. Алгоритм привязки использует модель присвоения линейной суммы к «паре» записей из двух файлов, которые необходимо связать (в нашем случае – это оригинальный и замаскированный файл). Процент правильно «спаренных» записей является мерой риска нарушения конфиденциальности.

14. Хотя этот подход и не такой простой по сравнению с описанным в предыдущем разделе евклидовым методом, он является привлекательным, поскольку от пользователя требуется ввести всего две вероятности: одну для верхней границы вероятности неправильного соответствия и одну для верхней границы вероятности неправильного несоответствия. Вышеописанный евклидов метод требует перемасштабирования переменных, а также предположений о весе переменных для расчета расстояний: например, в предложении [Pagl98] все переменные имеют одинаковый вес.

15. В экспериментальных работах использовалась вероятностная привязка записи используемая Бюро переписи США, предоставленная У. Уинклером [USBC, Wink98] (с некоторыми дополнениями).

## IV.3 Интервал нарушения конфиденциальности

16. Для какой-либо записи из массива замаскированных данных возьмем интервал ранга, отцентрированный по значениям данной записи, следующим образом: каждая переменная ранжируется независимо и интервал ранга определяется относительно значения переменной, которое она принимает в каждой записи; ранги значений в пределах этого интервала для переменной записи  $r$  должны отличаться менее, чем на  $p\%$  от общего количества записей, а ранг в центре интервала должен соответствовать значению переменной в записи  $r$ . Тогда мерой будет относительная доля исходных значений, которые входят в интервал, отцентрированный вокруг их соответствующего исходного значения. 100% доля означает, что взломщик абсолютно уверен в том, что истинные значения находятся внутри интервала, охватывающего замаскированное значение (интервал нарушения конфиденциальности). В ходе эксперимента рассматривались значения  $p$  от 1% до 10%.

## V. РЕЗУЛЬТАТЫ СРАВНЕНИЯ

17. Массив микроданных строился с использованием системы извлечения данных (DES) Американского бюро переписи (<http://www.census.gov/DES>). Было отобрано 13 непрерывных переменных и 1080 записей с тем, чтобы не было большого количества повторяющихся значений для любой из переменных (в принципе, в непрерывных переменных не предполагаются повторяющиеся значения, однако повторения имели место в массиве данных). В таблице 2 приводится ранжирование методов, описанных в Разделе 2 (параметрические значения, описанные в этом разделе, применялись для каждого метода). Колонка потери информации (I.L.) рассчитывается по среднему значению средних переменных  $X-X'$ ,  $\bar{X} - \bar{X}'$ ,  $V-V'$ ,  $S-S'$  и средней абсолютной погрешности  $R-R'$ ; результирующая средняя величина умножалась на 100. Колонка риска раскрытия конфиденциальности по дистанционной привязке (D.L.D.) содержит процентный показатель привязанных записей с помощью дистанционного метода привязки. Аналогично, колонка риска раскрытия конфиденциальности по вероятностной привязке (P.L.D.) содержит

процентный показатель привязанных записей с помощью вероятностного метода привязки. В колонке интервала нарушения конфиденциальности (I.D.) указывается средний процент исходных значений, попадающих под интервалы, охватывающие свои соответствующие скрытые значения (средние показатели рассчитывались по всем параметрическим значениям, т.е. от 1% до 100% с инкрементом 1%). И, наконец, колонка оценки (Score) используется для ранжирования таблицы 2 и ее значения рассчитывались следующим образом:

$$\text{Оценка} = 0.5(I.L.) + 0.125(D.L.D.) + 0.125(P.L.D.) + 0.25(I.D.).$$

18. Идея такого расчета заключается в придании потерям информации одинакового веса (0.5), также как и риску нарушения конфиденциальности. Вес риска нарушения конфиденциальности (0.5) равномерно распределяется по I.D. (0.25) и привязке записи. Вес привязки записи 0.25 равномерно распределяется между обеими подходами к привязке записей. Корреляция между D.L.D. и P.L.D. фактически составляет 0.962 – то есть оба подхода очень похожи. Корреляции (I.L.,D.L.D.), (I.L., P.L.D.), (I.L., I.D.) составляют –0.605, –0.551 и –0.807; таким образом, чем ниже потери информации, тем выше риск нарушения конфиденциальности, как и можно было предполагать. В колонках ранжирования I.L.Rank, D.L.D.Rank, P.L.D.Rank, I.D.Rank указан ранг каждого метода по отношению к I.L., D.L.D., P.L.D., I.D.; чем ниже ранг, тем лучше срабатывает метод (т.е., тем меньше потери информации и риск нарушения конфиденциальности).

**Таблица 2. Сравнительные результаты**

Метод	I.L.	D.L.D.	P.L.D.	I.D.	Score	I.L. Rank	D.L.D. Rank	P.L.D. Rank	I.D. Rank
Rank10	13.4	3.9	0.4	53.2	20.5	39	14	7	35
Rank7	9.2	7.5	1.1	68.7	22.9	30	29	31	51
Rank6	7.9	9.0	2.8	73.8	23.9	26	31	44	59
Mic3mul7	11.1	19.3	4.7	72.3	26.6	36	56	53	57
Rank5	6.8	16.8	13.6	78.9	26.9	22	46	58	65
Mic3mul9	13.5	19.2	3.4	69.9	27.0	40	55	48	53
Mic3mul0	14.8	18.0	3.4	68.6	27.3	43	52	47	50
Mic4mul4	12.1	19.8	6.7	71.8	27.3	37	57	56	56
Mic4mul5	14.5	17.4	5.4	69.1	27.4	42	49	54	52
Mic3mul8	13.5	20.8	4.2	70.7	27.5	41	59	51	54
Mic4mul8	18.9	17.8	3.3	62.8	27.8	47	50	46	44
Mic3mul6	10.2	20.4	13.9	74.0	27.9	33	58	59	60
Mic4mul7	19.4	17.1	2.1	64.4	28.2	48	48	41	46
Mic4mul6	17.9	17.8	4.0	66.4	28.3	45	51	50	48
Mic4mul9	21.4	15.9	2.0	61.7	28.3	50	45	40	43
Mic4mul0	23.0	16.9	2.4	60.6	29.0	51	47	43	40
Mic3mul5	9.7	23.8	18.3	76.6	29.3	31	64	61	62
Mic3mul4	7.5	23.5	22.8	79.1	29.3	24	63	63	67
Mic4mul3	10.7	22.9	16.7	76.9	29.5	35	62	60	63
Rank4	5.9	22.8	22.8	84.1	29.7	20	61	64	74
Micmul3	27.7	14.3	1.9	57.2	30.2	53	42	38	37

Micmul4	31.7	13.7	1.4	52.4	30.9	55	41	36	34
Mic3mul3	6.3	29.7	29.1	83.0	31.2	21	67	68	73
Micmul5	35.1	11.7	1.1	48.4	31.3	58	34	32	31
Micmul7	37.7	13.2	1.2	43.5	31.5	60	40	33	26
Micmul6	38.8	13.0	1.2	45.8	32.6	61	38	34	28
Micmul8	41.5	13.1	1.0	42.7	33.2	63	39	30	24
Rank3	5.1	31.7	36.9	89.5	33.5	18	68	71	81
Mic2mul0	10.7	49.4	27.3	77.4	34.3	34	74	66	64
Micmul10	44.7	14.7	0.5	40.4	34.3	64	43	16	21
Noise0.16	32.6	15.6	4.7	64.4	34.9	56	44	52	45
Micmul9	46.0	12.8	0.8	41.0	34.9	67	37	28	22
Mic2mul9	9.9	51.0	33.0	78.9	35.2	32	75	69	66
Mic2mul8	8.6	54.3	33.7	79.8	35.2	27	76	70	68
Mic2mul7	7.5	54.7	37.4	81.4	35.6	25	77	72	71
Noise0.12	25.2	22.2	22.4	71.6	36.1	52	60	62	55
Noise0.1	21.1	27.7	29.0	75.2	36.5	49	66	67	61
Mic2mul6	7.0	56.4	42.0	82.9	36.5	23	78	74	72
JPEG80	34.0	19.1	6.9	66.3	36.8	57	53	57	47
Noise0.14	35.1	19.2	6.2	67.6	37.7	59	54	55	49
Noise0.18	41.1	12.0	3.5	61.0	37.7	62	35	49	41
Noise0.08	17.4	36.1	39.8	79.8	38.2	44	70	73	69
Rank2	2.9	47.3	57.5	94.6	38.2	11	73	78	84
JPEG70	44.9	9.7	2.3	57.3	38.3	64	32	42	38
Noise0.2	46.0	10.0	1.0	57.6	38.8	66	33	29	39
Mic2mul5	5.9	59.0	56.8	85.4	38.8	19	80	77	76
JPEG85	29.5	23.8	24.5	72.8	39.0	54	65	65	58
Mic2mul4	4.9	61.5	60.7	87.3	39.5	17	82	79	77
JPEG90	18.2	35.4	47.0	80.9	39.6	46	69	75	70
Noise0.06	13.0	45.5	56.2	84.2	40.3	38	72	76	75
Mic2mul3	3.3	67.0	64.8	90.5	40.7	15	83	80	82
Noise0.04	8.9	58.5	65.63	89.0	42.2	28	79	82	78
JPEG75	50.4	12.7	2.9	61.3	42.5	68	36	45	42
JPEG95	9.1	60.1	66.6	89.2	42.7	29	81	84	80
Resamp1	3.1	67.9	67.6	96.8	42.7	14	84	85	85
Rank1	2.3	69.2	66.3	99.5	43.0	9	85	83	94
JPEG65	57.8	7.0	1.9	53.9	43.5	69	28	39	36
Noise0.02	4.2	77.3	71.3	94.4	44.3	16	87	86	83
Resamp3	3.1	75.4	71.9	98.4	44.6	13	86	87	87
MicPCP3	69.6	3.2	0.8	38.4	44.9	72	7	27	19
JPEG55	63.7	5.6	1.3	49.7	45.1	71	27	35	32
Noise0.01	2.6	85.2	74.1	97.0	45.5	10	88	91	86
JPEG100	3.1	87.1	73.0	99.1	46.3	12	89	88	89
MicIR10	1.2	97.4	74.1	99.1	46.8	8	90	90	88
MicIR8	1.0	97.8	74.1	99.3	46.8	6	96	89	91
MicIR9	1.1	98.0	74.4	99.2	46.9	7	97	92	90
MicIR6	0.9	97.7	75.3	99.5	46.9	5	94	93	93
MicIR5	0.7	97.6	76.0	99.6	46.9	3	92	94	95
MicIR3	0.5	97.4	79.0	99.8	74.2	1	91	95	97
MicIR4	0.6	97.6	79.8	99.7	47.4	2	93	96	96
MicIR7	0.8	97.8	88.1	99.4	48.5	4	95	97	92
MicPCP4	78.8	3.4	0.6	36.0	48.9	75	9	21	17
JPEG50	73.2	4.3	0.7	48.0	49.2	74	21	24	30
JPEG60	71.2	7.7	1.5	51.7	49.7	73	30	37	33
MicPCP5	82.5	3.9	0.7	34.1	50.4	76	15	26	15
MicPCP7	89.3	4.0	0.6	32.6	53.4	79	17	22	12

MicPCP9	90.8	4.5	0.3	31.4	53.8	82	24	3	9
MicPCP6	90.3	3.4	0.5	33.4	54.0	81	8	17	14
MicZ3	90.2	3.2	0.6	35.7	54.5	80	6	20	16
JPEG35	88.8	3.7	0.4	43.2	55.7	78	10	13	25
JPEG45	87.5	4.2	0.7	46.8	56.1	77	20	25	29
MicZ4	94.9	3.7	0.5	33.0	56.3	84	11	19	13
MicPCP8	96.9	4.0	0.3	32.0	57.0	85	16	6	10
MicPCP10	97.8	4.1	0.5	31.2	57.3	86	19	14	7
JEPG40	91.0	3.7	0.7	45.0	57.3	83	12	23	27
MicZ7	102.9	4.3	0.4	30.5	59.7	87	22	10	6
MicZ6	103.9	3.9	0.4	30.4	60.1	88	13	11	5
MicZ5	104.1	4.0	0.4	31.3	60.4	89	18	12	8
MicZ8	107.9	4.6	0.5	29.6	62.0	90	25	18	4
MicZ10	109.8	4.8	0.4	28.2	62.6	91	26	8	1
MicZ9	110.9	4.4	0.4	28.4	63.1	93	23	9	2
Distr	58.6	43.1	64.9	89.0	65.0	70	71	81	79
JPEG30	110.5	3.0	0.5	41.8	66.1	92	5	15	23
JPEG25	155.2	2.1	0.3	38.8	87.6	94	4	4	20
JPEG20	164.9	1.0	0.3	36.1	91.7	95	3	5	18
JPEG15	202.7	1.1	0.1	32.1	109.5	96	2	1	11
JPEG10	269.4	0.9	0.2	28.4	141.9	97	1	2	3

## VI. ВЫВОДЫ

19. Существует широкий ряд методов по ограничению риска нарушения конфиденциальности микроданных. В данной работе был определен и описан ряд предложений для непрерывных микроданных. Также описаны способы измерения потери информации. Результаты экспериментов, приведенные в таблице 2, говорят сами за себя. Стоит только особо отметить, что перестановка рангов параметра в области 10% представляет собой очень хорошую возможность; затем следует многовариантное микроагрегирование одновременно трех или четырех переменных; для микроагрегирования размер группы не имеет существенного значения. Искажение данных за счет распределения вероятностей, похоже, очень хорошо срабатывает. Для большинства методов эффективность их применения зависит от выбранных параметров, даже если некоторые методы являются больше параметро-зависимыми, чем другие.

### Благодарность

Данная работа частично финансировалась Бюро переписи США по контракту №OBLIG-2000-29158-0-0. Спасибо Франческу Себэ за его помощь в автоматизации программного обеспечения по вероятностной привязке записей и проведение экспериментов.

### Список литературы

[Adam89] Эдам, Н.Р., Уортманн, Дж.С., (1989), Методы контроля безопасности в отношении статистических баз данных: сравнительный анализ, *Компьютерные исследования АСМ*, вып. 21(4):515-556.

[Defa93] Дефейз, Д., Нанопоулос, Ф., (1993), Списки предприятий и конфиденциальность: метод малых агрегатов, в *документах 92-го Симпозиума по разработке и анализу продольных обследований*, Оттава: Статистическое бюро Канады, 195-204.

[Domí02] Доминго-Феррер, Дж., Матео-Санц, Дж.М., (2002), Практическое, ориентированное на данные микроагрегирование для контроля за соблюдением конфиденциальности в статистике, *Операции IEEE по знанию и обработке данных*, (планируется, март 2002 г.).

[Jaro89] Джаро, М.А., (1989), Развитие методологии привязки записей применительно к соотношению с переписью 1985 г. в Тампе, Флорида, *Журнал Американской статистической ассоциации*, вып.84: 414-420.

[JPEG] Объединенная экспертная группа по фотографированию, стандарт IS 10918-1 (ITU-T T.81) <http://www.jpeg.org>.

[Kim86] Ким, Дж.Дж., (1986), Метод ограничения нарушения конфиденциальности в микроданных на основе случайного шума и трансформирования, в *документах Отдела ASA по методике изучения обследований*, стр. 303-308.

[Liew85] Лью, С.К., Чой, Ю.Дж., Лью, С.Дж., (1985), Искажение данных распределением вероятностей, *Операции АСМ на системах баз данных*, вып. 10: 395-411.

[Moog96] Мор, Р., (1996), Методы контролируемой перестановки данных для маскирования массивов микроданных общего доступа, Бюро переписи США (неопубликованная работа).

[Pagl98] Паглиуча, Д., Сери, Дж.Ж (1998), Некоторые результаты метода индивидуального ранжирования для системы ежегодного обследования счетов предприятий, Проект Esprit SDC, выдается MI-3/D2.

[Skin94] Скиннер, С., Марш, С., Опеншоу, С., Уаймер, С., (1994), Контроль за соблюдением конфиденциальности микроданных по переписи, *Журнал официальной статистики*, вып. 10: 31-51.

[USBC] Бюро переписи С.Ш.А., (2000), Программное обеспечение по привязкам записей: документация для пользователя. Можно получить в Бюро переписи США.

[Wink98] Уинклер, У., (1998), Методы переидентифицирования для оценки конфиденциальности аналитически значимых микроданных, в *Защите статистических данных*, Люксембург: Бюро официальных публикаций Европейских Сообществ, 1999 г. Журнальный вариант в *Исследованиях по официальной статистике*, выпуск 1(2): 50-69, 1998 г.