Distr.
GENERAL

ECE/CES/SEM.54/4 (Summary)
3 April 2006

RUSSIAN
Original: ENGLISH AND
RUSSIAN ONLY

**СТАТИСТИЧЕСКАЯ КОМИССИЯ и ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ ОРГАНИЗАЦИИ ОБЪЕДИНЕННЫХ НАЦИЙ**
**КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ**

**ЕВРОПЕЙСКАЯ КОМИССИЯ**
**СТАТИСТИЧЕСКОЕ УПРАВЛЕНИЕ ЕВРОПЕЙСКИХ СООБЩЕСТВ (ЕВРОСТАТ)**

**ОРГАНИЗАЦИЯ ЭКОНОМИЧЕСКОГО СОТРУДНИЧЕСТВА И РАЗВИТИЯ (ОЭСР)**
**СТАТИСТИЧЕСКИЙ ДИРЕКТОРАТ**

Совместный семинар ЕЭК ООН/Евростата/ОЭСР по вопросам управления статистическими информационными системами (УСИС)
София, Болгария, 21-23 июня 2006 года

Тема i): Изменения в статистических процессах

# ПРИМЕНЕНИЕ ЭВОЛЮЦИОННЫХ АЛГОРИТМОВ ДЛЯ КОНТРОЛЯ ЗА ДОСТУПОМ К ИНФОРМАЦИИ

Специальный документ, подготовленный Андреа Тониоло Стаггемейром и Виком Дуоба, Управление национальной статистики Соединенного Королевства Великобритании и Северной Ирландии

## Резюме

1. Перед статистическими организациями (МСИ и НСИ) стоит трудная задача защиты своих источников информации от несанкционированного доступа при одновременном наращивании объема общедоступных и информативных данных. Одна из широко используемых для этих целей методологий называется Технология исключения. Она заключается в попытке отыскать идеально оптимизированный вариант размещения вторичных исключенных ячеек, позволяющий свести к минимуму общую потерю информации при сохранении аддитивности таблиц и выполнении требования относительно уровней защиты. Другая используемая в этой области методология называется Управляемым округлением. При использовании этой методологии производится округление номинальных указанных в таблице величин в сторону увеличения или уменьшения до ближайшего значения интеграла соответствующей

GE.06-22072 (R) 120606 130606

базисной величины, и при этом также сохраняется требуемый уровень защиты каждой ячейки и аддитивность всей таблицы. Оба эти метода присутствуют в программе Tau-Argus, разработка которой финансировалась в рамках проекта EU CASC Европейского союза. Статистическое управление Нидерландов (Анко Хундерпол) и университет города Ла-Лагуна на Тенерифе (профессор Саласар) внесли важный вклад в распространение методов математического программирования в контексте Европейского союза. Кроме того, дополнительную помощь по линии программного обеспечения оказала Германия.

2. Хотя в текущей практике применения упомянутых двух методологий имеются определенные достоинства, остаются нерешенными практические проблемы, связанные с большими размерами некоторых таблиц, с которыми приходится иметь дело Управлению национальной статистики и внешним источникам, представляющим данные этому Управлению. В связи с применением алгоритмов математического программирования, ориентированных на вычисление идеальных решений, требуются вычислительные ресурсы настолько большого объема, что часто обеспечение конфиденциальности крупных таблиц выходит за рамки временных нормативов или создает избыточную нагрузку на ресурсы иным образом. Это обусловливает потребность в проведении исследований, которые бы позволили преодолеть числовые и вычислительные трудности, связанные с применением методов, предусматриваемых программой Tau-Argus.

3. Гибридные подходы, в которых используются менее формализованные математические определения проблем, обычно называются эвристическими методами. Эти методы весьма эффективны для быстрого нахождения близких к оптимальным решений крупных задач, но, с другой стороны, их применение также может столкнуться с проблемой ограниченных возможностей конкретной вычислительной среды, а качество полученных с применением этих методов результатов не может быть повышено путем увеличения машинного времени. Это - классическая проблема большинства алгоритмов оптимизации.

4. Для нахождения адекватных решений по обеспечению конфиденциальности крупных таблиц с применением эвристических методов к соответствующим эвристическим процессам добавляется так называемая "мета"-фаза. Эта дополнительная фаза позволяет решать крупные задачи в рамках разумных лимитов машинного времени. Часто этот подход к ее решению сложных задач увязывается с применением математического программирования, а именно с применением формул линейного программирования, которые обеспечивают оценку реалистичности каждого решения.

5. Метаэвристические подходы хорошо известны в области исследования (ИО) и их с успехом применяют для решения различных задач оптимизации и комбинаторных задач. Однако это является относительно новым подходом в области статистической защиты от

идентификации (СЗИ). В работе Cox. et al. (2006) был впервые предложен комбинированный подход с использованием точных методов, а именно линейного программирования и эвристического и метаэвристического методов. Например, для защиты от идентификации применялись процедуры "Табу на поиск информации" и "Команда на рассеивание поиска" с использованием методологии, отличной от применяемой в "оптимальных" методах программы Tau Argus, которая называется Контролируемой корректировкой таблиц (в Бюро переписей США).

6. Это исследование содействовало развитию идеи использования в Управлении национальной статистики другого метафорического подхода, называемого Эволюционным алгоритмом (ЭА). Процедуры ЭА имитируют процесс естественной эволюции, представляя решение проблемы в виде генетического материала, который в результате процессов селекции, воспроизводства и мутации дает совокупность решений, приближающихся к наилучшим результатам. Процесс обучения выражается в том, что последующие популяции наследуют положительные черты каждого из родителей, а впоследствии каждому индивидуальному решению может быть добавлена определенная степень мема. В контексте защиты от идентификации речь идет о возникновении определенной популяции подавленных характеристик конкретной таблицы, так что в конце процесса обработки данных мы получим наилучшие индивидуальные решения, позволяющие представить решение проблемы в алгоритмической форме.

7. В Управлении национальной статистики изучаются также и другие метафорические подходы, такие, как использование Принципов муравейника или Адаптивные процедуры жадной рандомизации с целью нахождения наилучшего способа подавления данных в таблице.

8. В настоящем документе подчеркивается важное значение поддержания тесных связей между МСИ, НСИ и университетами с целью применения творческого подхода к решению проблем, возникающих в связи с крупными таблицами с многочисленными иерархическими связями и с различной плотностью нулевых и чувствительных ячеек. Работа в этой области разворачивается в тесном сотрудничестве с двумя университетами Соединенного Королевства, а именно с Университетом Западной Англии (Бристоль, Соединенное Королевство) и Университетом Кардиффа. Д-р Алистер Р. Кларк и д-р Джеймс Смит (оба из университета Западной Англии в Бристоле) возглавляют работу по направлению эволюционных алгоритмов, а доктор Джонатан Томсон (Кардифф) - работу по методам оптимизации по принципу муравейника и с применением алгоритмов GRASP. В рамках настоящего документа мы сосредоточимся лишь на одном метаэвристическом подходе, а именно - методе эволюционных алгоритмов. За дополнительной информацией о методах оптимизации по принципу муравейника или методологии GRASP можно обратиться к авторам настоящего документа.

9. УНС оказывает поддержку и участвует в новых разработках в области применения эволюционных алгоритмов для статистической защиты от идентификации, используя для этого последние достижения научных исследований и разработок, близкие к практическим решениям. Исходя из соображений удобства, мы уделим основное внимание описанию нашего подхода к решению проблемы подавления ячеек, а также описанию того, каким образом можно применять методологию эвристических подходов. Мы также полагаем, что по завершении начального этапа исследования его можно будет распространить на проблему контролируемого округления.

10. Недавно был достигнут баланс между оптимальными решениями, полученными путем применения имеющихся методологий оптимального (т.е. оптимального в рамках конкретной математической модели) подавления в программе Tau-argus (Fischetti and Salazar, 2001), и решениями, полученными в результате применения новых технологических решений, что свидетельствует об отсутствии снижения безопасности подавленной информации (как она определена в "оптимальном" методе). Для проверки практической применимости решения, а также в качестве критерия оценки того, насколько хорошим или плохим является предложенное решение с точки зрения общей потери информации, используется стандартная Модель атаки нарастающей интенсивности (Fischetti and Salazar, 2001). Первые результаты показывают, что при решении небольших задач требуемые вычислительные мощности относительно невелики и позволяют весьма эффективно находить оптимальные решения, в то время как в случае больших таблиц может понадобиться, например, 30 секунд, чтобы найти решение для таблицы, в которой насчитывается 10 000 ячеек и 700 первичных позволяющих идентификацию ячеек. В целях проверки первоначальных обнадеживающих выводов будет проведено дальнейшее тестирование этой методологии по широкому набору источников данных.

11. Если результаты последующего тестирования окажутся сопоставимыми с первоначальными результатами, то это будет означать, что УНС нашло более эффективный подход к обеспечению конфиденциальности более крупных массивов информации (при этом мы подчеркиваем, что необходимо провести подтверждающие эксперименты и тесты под стрессовой нагрузкой). При новом подходе, реализуемом в указанных исследованиях, требование полной оптимальности снижается в пользу достижения защиты информации путем применения некой модели подавления информации, которая, вероятно, будет близка к оптимальной. Это может быть достигнуто путем применения эвристических и/или метаэвристических методов. УНС намерено сообщать о любых успешных разработках в применении эвристических методов для защиты крупных таблиц за счет использования методики Tau Argus. В документе обсуждаются внесенные нами усовершенствования в методику необходимых практических проверок и процесс управления эволюционным алгоритмом с использованием операторов репродукции и мутации. В заключительной части документа

также описываются первые результаты и направления дальнейшей работы в рамках этого исследования.

12. В документе особо подчеркивается важное значение поддержания тесных связей с национальными статистическими управлениями, с университетами, а также с другими НСИ для стимулирования творческого и новаторского подхода к решению трудных проблем с помощью современных методов прикладной математики (ИО) и мощных вычислительных платформ.

**Справочные материалы**

Cox, L.H., Glover, F., Kelly, J.P. & Patil, R.J. (2006), "Confidentiality Protection By Controlled Tabular Adjustment: An Analytical and Empirical Investigation of Exact, Heuristic and Metaheuristic Methods," Decision Sciences Institute, в печати.

M. Fischetti, J.J. Salazar (2001). "Solving the Cell Suppression Problem on Tabular Data with Linear Constraints", Management Science, vol. 47, no. 7, pp. 1008-1026.

---