

Distr.
GENERAL

CES/SEM.47/15 (Summary)
30 January 2002

RUSSIAN
Original: ENGLISH

**СТАТИСТИЧЕСКАЯ КОМИССИЯ и
ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ**

**КОМИССИЯ ЕВРОПЕЙСКИХ
СООБЩЕСТВ**

КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ

ЕВРОСТАТ

**Совместный семинар ЕЭК ООН/Евростата
по интегрированным статистическим информационным
системам и связанным с ними вопросам (ИСИС-2002)**
(17–19 апреля 2002 года, Женева, Швейцария)

Тема II: Надежные средства связи и конфиденциальность данных

ВОПРОСЫ КОНФИДЕНЦИАЛЬНОСТИ, СВЯЗАННЫЕ С СИСТЕМАМИ ЗАПРОСОВ К СТАТИСТИЧЕСКИМ БАЗАМ ДАННЫХ

Специальный документ

Представлен Национальным центром статистики здравоохранения
Соединенных Штатов Америки¹

Резюме

1. Национальные статистические управления (НСУ) отвечают за сбор, проверку и обработку статистических данных с целью предоставления директивным органам и общественности надежной информации. В соответствии с законодательством или нормативной или этической практикой НСУ должно делать это таким образом, чтобы обеспечить сохранение *конфиденциальности* данных, касающихся индивидуальных

¹ Автор: Лоуренс Х. Кокс (LCOX@CDC.GOV).

субъектов, таких, как физические лица, предприятия или медицинские учреждения. Примечательным является то, что требования конфиденциальности не распространяется на статистические данные, касающиеся правительственных единиц.

2. До 60-х годов прошлого столетия НСУ разрабатывали статистическую информацию главным образом в форме расчетных и оценочных *таблиц*, определяемых комбинацией признаков одной, двух или небольшого числа переменных. НСУ решало, какие таблицы следует публиковать в первую очередь в печатной форме, а затем также и в электронной. Защита конфиденциальности, называемая в настоящее время "*ограничением доступа к статистическим данным*", обеспечивалась путем исключения или комбинирования избранных таблиц или всего набора таблиц либо, что встречалось менее часто, путем значительного изменения таблиц путем округления или включения случайных помех. По сути, НСУ сначала определяло, какие таблицы достойны опубликования, а затем публиковало соответственно меньше информации с учетом соображений конфиденциальности и качества данных.

3. В 60-е годы, начав публикации выборки непрерывного трудового стажа Управления социального обеспечения США, а затем общедоступных выборок из итогов десятилетних переписей 1960 года США, НСУ приступили к публикации *файлов статистических микроданных*, содержащих записи, касающиеся индивидуальных субъектов (главным образом физических лиц). Пользователи данных в настоящее время могут создавать любые мыслимые обобщения на основе данных индивидуальных записей и, что является в равной степени важным, адаптировать статистические, демографические или эконометрические модели к микроданным. В области ограничения доступа к статистическим данным акцент был перенесен на изменение или исключение избранных файлов микроданных. Продольные данные создавали проблемы с конфиденциальностью, которые в значительной степени остаются нерешенными и до сих пор. В настоящее время начаты исследования в направлении адаптации данных к сложным статистическим моделям и публикации вместо разработанных на основе моделей данных *синтетических микроданных* и/или самих моделей. Ограничение доступа к таблицам и микроданным, несомненно, является сложной проблемой как с теоретической, так и с вычислительной точек зрения.

4. НСУ в настоящее время занимаются изучением возможности предоставления пользователям данных прямого доступа к статистическим базам данных либо на основе открытого, либо ограниченного доступа через *систему запросов к статистической базе данных*. Это ведет к усилению беспокойности вопросами и проблемами конфиденциальности и способно стимулировать исследования в области ограничения доступа к статистическим данным в предстоящее десятилетие. В настоящем документе мы на основе примеров анализируем некоторые проблемы конфиденциальности и

полезности данных, возникшие в связи с системами запросов к базам статистических данных. Основное внимание уделяется двум парадигмам запросов: формирование таблиц из базы данных, организованной в виде крупной многомерной таблицы сопряженности признаков, и расчет простых статистических моделей, разработанных на основе базы данных, а именно моделей регрессии методом наименьших квадратов и наилучшего линейного несмещенного прогноза (*кригинга*) в отношении пространственных данных.
