РОБОЧИЙ ДОКУМЕНТ
ІЗ СТАТИСТИЧНОЇ ПОЛІТИКИ 22 (Версія друга, 2005)

Звіт по методології обмеження статистичного розкриття інформації

Федеральний комітет із статистичної методології

Початково підготовлений підкомітетом з методології
обмеження розкриття інформації від 1994

Перевірений Комітетом з конфіденційності і доступу до даних 2005 року

Політика в області наукових досліджень і статистики
Управління інформації і нормативно-правового регулювання
Служба управління і бюджету

Грудень 2005

Федеральний комітет із статистичної методології
(Грудень 2005)

Учасники

Браян А. Херіс-Кожетін, Голова, Служби управління і бюджету

Венді Л. Елві, Секретар, Бюро перепису населення США

Лінда Карлсон, Національний науковий фонд

Стівен Б. Коен, Агентство з досліджень та оцінки якості медичного обслуговування

Стів Х. Коен, Бюро трудової статистики

Лоуренс Х. Кокс, Національний центр медичної статистики

Роберт Е. Фей, Бюро перепису населення США

Рональд Фексо, Національний науковий фонд

Деніс Фікслер, Бюро економічного аналізу

Джеральд Гейтс, Бюро перепису населення США

Беррі Граубард, Національний інститут раку

Вільям Івіг, Національна служба із сільськогосподарської статистики

Артур Кеннікел, Рада керуючих Федеральної резервної системи

Ненсі Дж. Кіркенделл, Управління з інформації в області енергетики

Сюзан Щечтер, Служба управління і бюджету

Рольф Р. Шмітт, Федеральне управління автомобільних доріг

Мерлін Сісторм, Національний центр зі статистики освіти

Монро Г. Сіркен, Національний центр медичної статистики

Ненсі Л. Спруілл, Міністерство оборони

Клайд Такер, Бюро трудової статистики

Алан Р. Тупек, Бюро перепису населення США

Г. Девід Вільямсон, Центри з контролю і профілактики захворювань США

Консультант-спеціаліст

Роберт Гровс, Мічиганський Університет та Об’єднана Програма з Методології Опитування.

Передмова

Федеральний комітет із статистичної методології (FCSM) було організовано Службою управління і бюджету (OMB) в 1975 для дослідження питань якості даних, що впливають на Федеральну статистику. Члени комітету, що обираються OMB на основі їх особистої компетенції і зацікавленості в статистичних методах, служать швидше як приватні особи ніж в якості представників агентства. Комітет здійснює свою роботу через підкомітети, що організовуються для вивчення певних питань. Робочі документи із статистичної політики підготовляються членами підкомітету, і переглядаються та утверджуються членами FCSM.

Комітет з конфіденційності і доступу до даних (CDAC) – це підкомітет FCSM, що представляє спеціальні інтереси, і який було сформовано в 1995 році в результаті наданих рекомендацій, що містяться в оригінальному Робочому документі із статистичної політики 22. Комітет складається в основному із статистиків, що працюють у федеральних агентствах, і які причетні до вирішення питань, що мають відношення до захисту конфіденційності даних, і надавання вибіркового та обмеженого доступу до конфіденційних даних. CDAC надає доступ до унікального форуму для обговорення цих питань та обміну інформацією, а також дослідження ідей поміж федеральними агентствами. До веб-сайту CDAC можна отримати доступ за адресою http://www.fcsm.gov/committees/cdac. Перегляд Робочого документу із статистичної політики 22 від 2005 року – це друга версія праці 1994 року, виконаної Підкомітетом з Обмеження розкриття і методології. Підкомітет з Методології обмеження розкриття було сформовано в 1992р. для опису та оцінки існуючих методів обмеження розкриття для файлів у формі таблиць і мікроданими, і для оновлення інформації в попередній роботі, що представлена в Робочому документі із статистичної політики 2, «Звіт із методик статистичного розкриття та уникнення розкриття», опублікованому в 1978р. Див. Титульний і вступний матеріал у версії Робочого документу із статистичної політики 22 від 1994 року для обговорення Підкомітетом із Методології обмеження розкриття.

Звіт із Методології обмеження статистичного розкриття, Робочий документ із статистичної політики 22, обговорює як таблиці, так і мікродані та описує поточну практику основних Федеральних статистичних агентств. Оригінальний звіт містить: навчальний посібник, керівні вказівки і рекомендації для здійснення рекомендованих норм; рекомендації для подальших досліджень; анотовану бібліографію. У 2004 Комітет із конфіденційності і доступу до даних (CDAC), переглянув Робочий документ із статистичної політики 22, який повинен включати в себе дослідження і нові методології, що було розроблено протягом останніх десяти років, і відображати поточну практику агенції. Анотовану бібліографію було частково оновлено. Члени CDAC, які працювали над переглядом:

Джейкоб Бурмазіан, Управління з Інформації в Області Енергетики
Ненсі Кіркендал, Управління з Інформації в Області Енергетики
Стів Коен, Бюро Трудової Статистики
Філіп Стіл, Бюро Перепису Населення
Алван О. Зарате, Національний Центр Медичної Статистики
Арнольд Рехнек, Бюро Перепису Населення
Пол Массел, Бюро Перепису Населення

Висловлення подяки

Ми дякуємо представникам агентства CDAC за їх внески у цей робочий документ та оновлення до опису практики агентства в Розділі 3.

Зміст

РОЗДІЛ І − Вступ

А. Предмет і мета цього звіту

Від федеральних агентств та їх підрядників, що розголошують статистичні таблиці чи файли з мікроданими, часто вимагається законом чи усталеною практикою захищати конфіденційність індивідуальної інформації. Ця вимога щодо конфіденційності застосовується до розкриття даних громадськості; вона також може застосовуватись до розголошень іншим агентствам чи навіть іншим підрозділам цього ж агентства. Необхідний захист досягається застосуванням процедур обмеження статистичного розкриття, метою яких є забезпечення того, щоб ризик розкриття конфіденційної інформації про осіб, що піддаються ідентифікації, суб'єктів підприємницької діяльності чи їх підрозділів був дуже малим.

Протягом 2004 року Комітет з конфіденційності і доступу до даних (CDAC), комітет із спеціальними інтересами щодо конфіденційності даних питань доступу для Федерального комітету із статистичної методології (FCSM), переглянув Робочий документ із статистичної політики 22 для включення нових вдосконалень в методологіях обмеження статистичного розкриття, та для оновлення процедур і практик агентства з конфіденційності даних. Опис CDAC та їх діяльності міститься в Додатку Г. Робочий документ із статистичної політики 22 було написано в 1994 Підкомітетом із методології обмеження розкриття. Метою підкомітету 1994 року було переглянути та дати оцінку методам обмеження статистичного розкриття, що використовуються федеральними статистичними агентствами і для розробки рекомендацій для їх вдосконалення. Опис цього підкомітету міститься в **Роз'яснювальному** і вступному матеріалі оригінального Робочого документа із статистичної політики 22 від 1994 року.

Закони, прийняті Конгресом після первинного випуску Робочого документа із статистичної політики 22 у 1994, було додано до потреб федеральних агентств для захисту конфіденційності даних, які вони збирають. Акт захисту і права переказу медичної страховки (HIPPA), вперше приведений в дію у 1996 році, мав серйозний вплив на встановлення вимог щодо захисту даних з охорони здоров'я. Акт захисту конфіденційної інформації і статистичної ефективності (CIPSEA) від 2002 року створив новий механізм для агентств захисту конфіденційності даних, і в той же час обмежив діяльність із обміну інформацією лише до статистичних цілей. Протягом цього часу, зацікавленість у федеральних статистичних даних у межах спільнот, що користуються даними і досліджують їх, продовжувала зростати. Потреба для більшого доступу до даних призвела до розробки нових методів з уникнення розкриття для того, щоб більше даних могло розголошуватись громадськості, в той час як агентства підтримують захист інформації опитуваного. Цей перегляд Робочого документа із статистичної політики 22 доповнює обговорення цих питань включаючи поточне дослідження і нові розробки в цій галузі.

Цілі перегляду цього звіту були наступними:

- описати та оцінити існуючі методи обмеження розкриття для таблиць і файлів з мікроданими;
- надати рекомендації та керівні вказівки для обрання і використання ефективних методик обмеження розкриття;
- сприяти розробці, обміну і використанню програмного забезпечення для прикладних програм із методів обмеження розкриття;

-
- і підтримувати дослідження щодо розвитку покращених методів обмеження статистичного розкриття для табличних файлів, а також файлів із мікроданими для громадського користування.

Кожне агентство чи підрозділ у його межах, що розголошує статистичні дані, повинно бути спроможне обирати і застосовувати відповідні процедури обмеження розкриття стосовно всіх даних, які воно розголошує. Кожне агентство повинно мати одного чи більше працівників із чітким розумінням методів і теорії, що лежать в їх основі. Цей звіт орієнтований в основному на працівників федеральних агентств та їх підрядників, які залучені у зібранні та розповсюдженні статистичних даних, особливо на тих, що безпосередньо несуть відповідальність за обрання і використання процедур обмеження розкриття. Цей звіт також корисний для працівників із схожими обов'язками в інших організаціях, що оприлюднюють статистичні дані, і для користувачів даних для того, щоб вони могли краще розуміти і використовувати пристрої обробки і передачі даних, захищених від розкриття.

Б. Деякі визначення

Для того, щоб уточнити обсяг застосування цього звіту, ми тут визначаємо та обговорюємо деякі основні терміни, які будуть використовуватись у всьому звіті.

Б.1. Конфіденційність і розкриття

Визначення **конфіденційності** було подано Президентською комісією з федеральної статистики (1971:222):

> [Конфіденційний повинно означати, що розповсюдження] даних у спосіб, що зробить можливим публічну ідентифікацію респондента чи у будь-який інший спосіб буде шкідливим для нього, забороняється, і що ці дані мають імунітет від судового провадження. Дункан і співавтори, 1993, *«Особисте життя і державна політика»,* с. 24.

Конфіденційність відрізняється від приватності, тому що вона застосовується як до комерційної діяльності, так і до фізичних осіб. Приватність − це право особи, тоді як конфіденційність часто застосовується з приводу даних по організаціях та фірмах. Другий елемент визначення, імунітет від обов'язкового розкриття через судовий процес, представляє собою правове питання і знаходиться поза межами застосування цього договору.

Друге визначення також надається для підтримки користувачів в розумінні цієї концепції.

«Конфіденційність має відношення до трактування інформації, яку розкрила фізична особа у відносинах довіри і з очікуванням того, що вона не буде оприлюднена іншим у спосіб, що є несумісним із розумінням оригінального розкриття без дозволу». Посібник IRB, Частина III.Г,

Міністерство охорони здоров'я і соціальних служб, Управління захисту від досліджень на людях.

Потреба управління в захисті конфіденційності даних, які вона збирає, основана на різноманітних вимогах законодавства. Статистичне розкриття має місце, коли розголошені статистичні дані (незалежно від того, чи це табличні чи індивідуальні записи) оприлюднюють конфіденційну інформацію про окремого респондента. Цей документ розглядає мінімізацію ризику **розкриття** (публічної ідентифікації) характерних ознак окремих одиниць звітності та інформації про них.

Стаття 512 Розділу V Акту статистичної ефективності і захисту конфіденційної інформації від 2002 (CIPSEA) вимагає, щоб всі федеральні агентства захищали дані або інформацію, що здобувається агентством згідно із порукою щодо конфіденційності винятково для статистичних цілей та у такій формі, що виключає розкриття, що сприяє ідентифікації. Стаття 502 CIPSEA визначає **«форму, що підлягає ідентифікації»** як будь-яке представлення інформації, що дозволяє належним чином визначити за допомогою прямих чи непрямих засобів особу респондента, якого стосується ця інформація.

Правило приватності щодо Акту захисту і права переказу медичної страховки (HIPAA) було введено в дію 14 квітня 2003 року. Це правило зобов'язує «осіб, найбільш забезпечених грошовим покриттям», таких як постачальників послуг «Медікер» (федеральні програмі медичного страхування для населення старшого віку), для захисту конфіденційності інформації щодо охорони здоров'я, якою вони володіють. Правило приватності підпорядковує постачальників інформації щодо охорони здоров'я певним вимогам для захисту конфіденційності даних, що розголошуються. Незважаючи на основні компоненти, що використовуються для захисту конфіденційності, федеральні статистичні агентства, так як і деякі приватні організації зі збору інформації, пов'язані з інформацією щодо охорони здоров'я, повинні врівноважити два завдання: надавати корисну статистичну інформацію користувачам даними, і забезпечити, щоб відповіді осіб були захищені.

Закон про права сім'ї на освіту і недоторканність приватного життя (FERPA) (Кодекс Сполучених Штатів 20, § 1232g; 34 CFR Частина 99) був прийнятий для захисту приватності записів про освіту студента. Закон застосовується до всіх шкіл, що отримують фінансування згідно із діючою програмою Міністерства освіти США. FERPA надає батькам і правочинним студентам (тобто студентам, віком старше 18 або які відвідують школу, нижчу рівня середньої) певні права по відношенню до їх записів щодо освіти. В загальному, школи повинні мати письмовий дозвіл від батьків чи правочинного студента для того, щоб розголошувати будь-яку інформацію із записів про освіту студента. Однак, FERPA дозволяє школам розкривати ці записи без згоди певним призначеним для цього сторонам, або ж при наявності особливих умов. Школи можуть також розкривати без згоди «довідкову» інформацію, таку як ім'я студента, адресу, номер телефону, дату і місце народження, особисті відзнаки та нагороди, а також дні відвідування. Однак, школи повинні повідомляти батькам і правочинним студентам про довідкову інформацію і надавати батькам і правочинним студентам прийнятну кількість часу для подання запиту, щоб школа не розкривала довідкову інформацію про них.

Розголошення статистичних даних неминуче оприлюднює деяку інформацію про суб'єкта індивідуальних даних. Розкриття має місце тоді, коли розголошується конфіденційна інформація. Інколи розкриття може відбуватись на основі самих розголошених даних; в інших випадках розкриття може бути результатом від комбінування розголошених даних із публічно доступною

інформацією; а також іноді розкриття можливе лише шляхом комбінування розголошених даних із детальними джерелами зовнішніх даних, що можуть або не можуть бути доступними для громадськості. Доступ і/або приєднання громадськості до баз електронних даних створює деякий ступінь ризику, що розкриття конфіденційної інформації може мати місце навіть якщо особисті ідентифікатори усуваються із файлу. Як мінімум, кожне статистичне агентство повинно гарантувати, що ризик розкриття через розголошені дані при комбінації з іншою відповідною публічно доступними даними дуже низький.

Було запропоновано декілька різних визначень розкриття і різних типів ризику розкриття. Дункан і співавтори (1993: 23-24) надає визначення, що вирізняє три типи розкриття:

> **Розкриття** стосується неналежного умовного нарахування інформації суб'єкту даних, незалежно від того чи це фізична особа чи організація. Розкриття має місце коли суб'єкт даних ідентифікується із розголошеного файлу **(розкриття особи),** приватна інформація про суб'єкт даних оприлюднюється через розголошений файл **(розкриття реквізитів),** або якщо розголошені дані зробили можливим визначення значення деяких характеристик фізичної особи, точніше, ніж було б можливо за інших обставин **(логічно виведене розкриття).**

Зверніть увагу, що кожен тип розкриття може траплятись у зв'язку з розголошенням або таблиць, або мікроданих. Визначення і висновки, що виникли внаслідок цих типів розкриття вивчаються більш детально в наступному розділі.

Б.2. Таблиці, мікродані, та системи запитів в режимі реального часу.

Вибір методів обмеження статистичного розкриття залежить від характеру пристроїв обробки і передачі даних, конфіденційність яких повинна захищатись. Більшість статистичних даних розголошуються у формі таблиць, файлів з мікроданими, або через системи запитів в режимі реального часу. Таблиці в подальшому можна розділити на дві категорії: таблиці частотних (рахункових) даних і таблиці порядкових даних. Для обох категорій дані можна представляти в формі чисел, пропорцій або процентних відношеннях.

Файл з мікроданими складається з індивідуальних записів, кожен з яких містить значення змінних для окремої особи, комерційного підприємства чи іншої організаційної одиниці. Деякі файли з мікроданими включають в себе прямі ідентифікатори, такі як ім'я, адреса чи номер соціального забезпечення. Усунення будь-якого з цих ідентифікаторів є очевидним першим кроком в підготовці до розголошення файлу, для чого повинна захищатись конфіденційність індивідуальної інформації.

Історично, методи обмеження розкриття для таблиць застосовувались безпосередньо до них. Методи включають: переробку таблиць, приховування, контрольоване чи довільне округлення. Більш сучасні методи зосередились на захисті мікроданих, що лежать в основі таблиць з використанням деяких технік захисту мікроданих. В цей спосіб всі таблиці, що розробляються із захищених мікроданих також захищаються. Це можна здійснювати незалежно від того, чи є намір розкривати мікродані чи

ні. Це особливо корисний спосіб захисту таблиць, що розробляються на основі систем запитів в режимі реального часу.

Б.3. Закриті дані та обмежений доступ

Конфіденційність індивідуальної інформації може захищатись обмеженням кількості інформації, що надається, або корегування даних у розголошених таблицях і файлах з мікроданими **(закриті дані)** або накладанням умов щодо доступу до приладів обробки і передачі даних **(обмежений доступ),** або певною комбінацією цих двох методів. Число федеральних агентств, які запровадили програми з обмеженого доступу, зросло протягом останніх десяти років, і цей звіт надає деякі довідкові матеріали. Проте, основною метою цього звіту є обговорення методів обмеження розкриття, що забезпечують захист конфіденційності обмежуючи доступ до даних. Той факт, що цей звіт в основному має справу із процедурами обмеження розкриття, що обмежують доступ або корегують зміст даних, не повинен тлумачитись таким чином, що означатиме ніби процедури обмеженого доступу є менш важливими. Читачі, які зацікавлені в останньому можуть знайти детальну інформацію у книзі Дункана і співавторів, 1993, «*Особисте життя і державна політика»*,. с. 157 і «Процедури обмеженого доступу», виданої Комітетом з конфіденційності і доступу до даних (квітень 2002) за посиланням http://www.fcsm.gov/committees/cdac/cdacra9.doc.

Якщо описувати коротко, існує чотири основних методи, які використовують агентства для забезпечення обмеженого доступу до конфіденційних даних: Центри даних дослідження (RDC), Дистанційний доступ, Співдружності наукових співробітників і Програми для співробітників, що мають ступінь доктора наук, і Ліцензійні договори. **RDC** дозволяють використовувати конфіденційні файли у фізично захищеному середовищі із спеціалізованим обладнанням. Користувачі погоджуються на положення та умови, що регулюють доступ і використання конфіденційних даних. Продукти дослідження перевіряються агентством для гарантії, щоб жодну конфіденційну інформацію не було розголошено. **Дистанційний доступ** через безпечні електронні лінії до спеціалізованих комп'ютерів є другим методом. Користувачі можуть застосовувати статистичні методи до конфіденційних даних. Статистична продукція розглядається агентством для гарантії, що жодних конфіденційних даних не було розкрито. **Співдружності і програми для співробітників, що мають ступінь доктора наук** є третім методом, і дослідники підписують договори, що дозволяє їм бути трактованими як співробітники агентства, і підпорядковуватись такими ж обмеженням як і співробітники. Подібно до доступу RDC, дослідникам може надаватись обмежений доступ, а продукти дослідження переглядаються агентством для гарантії, що жодних конфіденційних даних не буде розголошено. Крім того, **ліцензійні договори** дозволяють досліднику використовувати конфіденційну інформацію за межами обчислювального центру, але згідно із строгими обмежувальними умовами, викладеними у договорі, що має обов'язкову юридичну силу. Домовленості, що накладають обмеження щодо того, хто має доступ, в якому місцезнаходженні, і для яких цілей дозволяється доступ, зазвичай вимагають письмових угод між агентством і користувачами. Ці договори зазвичай підпорядковують користувачів дії штрафів, забороні доступу в майбутньому і/або іншим покаранням за неналежне розкриття індивідуальної інформації та інші порушення узгоджених умов щодо використання. Користувачі можуть підлягати зовнішнім аудиторським перевіркам, що здійснюються агентством для гарантії, що умов договору дотримуються. Від користувачів, що здійснили порушення, може вимагатись оплатити штраф або підлягати правовим мірам покарання.

Більшість пристроїв обробки і передачі даних **суспільного користування** випускаються

статистичними агентствами будь-кому, зазвичай без обмежень з користування чи інших умов, крім виплати комісійних винагород для купівлі публікацій або файлів даних в електронній формі. Як NCHS так і NCES вимагають від користувачів файлів даних суспільного користування виразити, що вони не будуть використовувати дані, які є їм доступні, щоб намагатись ідентифікувати окремого респондента. Агентства вимагають, щоб ризики розкриття для пристроїв обробки і передачі даних суспільного користування були дуже низькими. При дотриманні цієї вимоги застосування методів обмеження розкриття, описаних у цьому документі, можуть значним чином обмежити вміст даних, до тієї міри, коли дані вже не мають вартості для певних цілей. Національний центр статистики освіти надає пристрої обробки і передачі даних суспільного користування, що включають в себе доступ до конфіденційних даних. Хоча ці дані призначені для «Суспільного користування», користувачі повинні підписати договори, які надають гарантію, що вони будуть підтримувати конфіденційність даних. Користувачі можуть проходити аудиторську перевірку для того, щоб впевнитись, що вони дотримуються відповідних процедур.

B. Організація звіту

Розділ II, «Методи обмеження статистичного розкриття: Підручник», подає просте визначення і зразки методів обмеження розкриття, що можуть використовуватись для обмеження ризику розкриття при розголошенні таблиць і мікроданих.

Розділ III, «Поточна практика федеральних статистичних агентств», описує методи обмеження розкриття, що використовуються 14 (чотирнадцятьма) головними федеральними статистичними агентствами і програмами. Поміж факторами, що пояснюють варіації в практиці агентств, є також відмінності в типах даних і респондентів, різні правові вимоги і політика для захисту конфіденційності, інший технічний персонал, а також окремі історичні підходи до питань конфіденційності.

Розділ IV, «Методи для табличних даних» надає систематичний і детальний опис та оцінку методів обмеження статистичного розкриття для таблиць частотних і порядкових даних. Розділ V, «Методи для файлів з мікроданими суспільного користування», описує різноманітні методи обмеження статистичного розкриття, що використовуються для захисту конфіденційності при публічному розголошенні файлів з мікроданими. Ці розділи будуть представляти найбільший інтерес для читачів, що несуть пряму відповідальність за використання методів обмеження, розкриття, або проводять дослідження для оцінювання і покращення існуючих методів чи розробки нових.

Частково завдяки стимулу, що надається звітами попереднього підкомітету (включаючи Робочі документи статистичної політики 2 і 22), покращені методи обмеження розкриття було розроблено і використано деякими агентствами протягом останніх 25 років. На основі перегляду цих методів надаються керівні вказівки в Розділі VI в якості практичних рекомендацій для всіх агентств. Розробка і виробництво файлів з мікроданими суспільного використання продовжує розширюватись, і підвищує потребу в перегляді можливості приєднання даних до зовнішніх файлів, та ролі ідентифікаторів для файлів.

Ці додатки також включаються. Додаток А містить технічні замітки щодо практик, які статистичні агентства вважають корисними для розширення правил першочергового приховування на інші звичайні ситуації. Додаток Б це перелік веб-сайтів та державних довідкових документів щодо статистичного розкриття. Додаток В це бібліографія. Додаток Г містить опис CDAC і його досягнення.

Г. Основні предмети цього звіту

П’ять основних предметів лежать в основі керівних вказівок в Розділі VI:

Існують відмінності між вимогами обмеження розкриття, що застосовуються до федеральних агентств. Федеральні агентства, що мають спеціальне законодавство, яке охоплює їх діяльність зі збору даних, зобов’язані підтримувати конфіденційність всіх відповідей, отриманих в ході опитування. Інші агентства, що не мають спеціального законодавства, яке охоплює їх діяльність збору даних, можуть визначати які дані потребують захисту. Незважаючи на це, агентствам, що потребують захищати дані, необхідно рухатись якнайбільше у напрямку використання малої кількості стандартизованих методів обмеження розкриття, ефективність яких було продемонстровано.

Методи обмеження статистичного розкриття було розроблено і запроваджено окремими агентствами протягом останніх 40 років. Інформацією і дослідженнями в цій галузі необхідно обмінюватись серед всіх федеральних установ. Документацією і відповідним програмним забезпеченням, що використовується статистичним агентством, необхідно обмінюватись між федеральними агентствами.

Результати досліджень, обмежені для розкриття, повинні підлягати аудиторській перевірці для визначення того, чи вони відповідають цілям процедури із захисту даних, яка застосовувалась. Наприклад, програмне забезпечення для лінійного програмування може використовуватись для здійснення аудиторських перевірок розкриття для деяких типів табличних даних. Наприклад, програмне забезпечення для лінійного програмування може використовуватись для здійснення аудиторських перевірок розкриття для деяких типів табличних даних. Разом з цим, корисність даних із результатів досліджень, обмежених для розкриття, повинні оцінюватись як частина аналізу застосовуваної процедури.

Декілька агентств сформували наглядові ради, статистичні або аналітичної комісії, і призначили чиновників із конфіденційності агентства для гарантії того, щоб відповідна політика і практика обмеження розкриття була в наявності, і належним чином використовувалась. Кожне агентство повинно централізувати свій контроль і перегляд застосування методів обмеження розкриття через розвиток стандартизованого переліку питань або галузей дослідження. «Контрольний список із потенціалу розкриття запропонованих видачі даних», опублікований CDAC і розміщений за посиланням

http://www.fcsm.gov/committees/cdac/resources.html це корисний довідник для агентств, щоб формувати структуру їх перегляду.

РОЗДІЛ II – Методи обмеження статистичного розкриття: підручник

Цей розділ надає вступний курс до методик обмеження розкриття, що широко використовуються для обмеження можливості розкриття ідентифікуючої інформації про респондентів у таблицях і файлах з мікроданими. Методики проілюстровані з прикладами. Таблиці або файли з мікроданими, створені з використанням цих методів, зазвичай, надаються в розпорядження громадськості без жодних подальших обмежень. Стаття Б представляє деякі з основних визначень, що використовуються в цих статтях і подальших розділах. Вона включає в себе обговорення відмінностей між таблицями частотних даних і таблицями порядкових даних, визначення розмірності таблиці, та ієрархічні змінні, а також стислий виклад різних типів розкриття. Стаття В обговорює методи обмеження розкриття, що застосовуються до таблиць підрахунків і частот. Стаття Г звертається до таблиць порядкових даних, Стаття Е підсумовує розділ. Читачі, які вже ознайомлені з методологією обмеження статистичного розкриття, можуть переходити до Розділу III, що описує усталені практики агентства, Розділу IV, що представляє обговорення (з більш математичним нахилом) методологій обмеження розкриття, що використовувалось для захисту таблиць, або Розділу V, що надає більш детальне обговорення методологій обмеження розкриття, що застосовуються до мікроданих.

А. Передумови

Однією з функцій федерального статистичного агентства є збирати дані, що піддаються

індивідуальній ідентифікації, обробляти їх і надавати статистичні висновки і/або файли з мікроданими публічного користування громадськості. Деякі з даних вважаються респондентами приватними.

З іншого боку не всі дані, що систематизуються і публікуються урядом, підлягають техніці обмеження розкриття. Деякі дані щодо бізнесу, що збираються для цілей врегулювання, вважаються публічними. Крім того, деякі дані не вважаються чутливими і не збираються згідно із запевненням конфіденційності. Технології обмеження статистичного розкриття, описані в цьому документі, застосовуються незалежно від того, коли б не вимагалась конфіденційність, а дані або підсумкові показники стають публічно доступними. Всі методи обмеження розкриття призводять до певної втрати інформації, та іноді публічно доступні дані не можуть бути належними для певних статистичних досліджень. Однак, головним наміром є надати якнайбільше інформації якщо це можливо, не розкриваючи дані, що підлягають ідентифікації в індивідуальному порядку. (Див. Розділ I для короткого обговорення використання обмеженого доступу в порівнянні із закритими даними).

Найбільш звичний метод надання даних громадськості це через статистичні таблиці. Із розвитком потужних комп'ютерів, великою місткістю пам'яті, та високою швидкістю обробки даних, агентства почали надавати в користування систему запитів в режимі реального часу із доступом до статистичної бази даних. Користувачі даних створюють свої власні табличні дані за допомогою пристосованих питань. У більшості з цих систем лише дані, до яких вже застосовувалось обмеження розкриття, доступні для користувачів. Якщо незахищені мікродані використовуються в якості основи для системи обробки запитів, то правила обмеження розкриття повинні застосовуватись автоматично до таблиць, стосовно яких подається запит. Проблема із останнім підходом полягає в тому, що користувачам може надаватись можливість розпізнати конфіденційні дані, якщо вони використовують послідовність запитів, в яких обмеження розкриття застосовується незалежно.

Файли з мікроданими представляють інший спосіб, за допомогою якого агентства намагаються надати результати досліджень, орієнтовані на користувача. Ці результати досліджень стали незамінними для наукової спільноти, коли виріс рівень розголошення файлів з мікроданими для публічного користування. У файлі з мікроданими кожен запис містить набір змінних, що пов'язані з єдиним респондентом і мають відношення до значень, повідомлених респондентом. Однак імена, адреси та інші **прямі ідентифікатори** усуваються з файлу, а дані можуть бути певним чином змінені для того, щоб гарантувати, що окремі елементи даних не були пов'язані лише з певним респондентом.

Б. Визначення

Кожен запис у статистичній таблиці представляє загальне значення кількості по відношенню до всіх одиниць аналізу, що належать до унікального статистичного мінімального елементу даних. Наприклад таблиця, що представляє підрахунки осіб по 5-річних вікових категоріях і загальний річний дохід поетапно по $10,000, складається із статистичних мінімальних елементів даних, таких як ця (вік 35-39 років, $40,000 до $49,999 щорічного доходу). Число у комірці таблиці – це підрахунок або частота кількості людей в населенні із характеристиками цієї комірки. Таблиця, що відображає вартість будівельних робіт, що проводяться протягом певного періоду у штаті Меріленд окружними групами і 4-циферними групами Статистичної класифікації господарської діяльності в Північній Америці (NAICS) складаються з таких комірок, як {NAICS 4231, Округ Прінс-Джорджес}. У цьому

випадку число в комірці представлятиме середню вартість (або загальну вартість) будівельних робіт для компаній серед населення із характеристиками цієї комірки.

Б.1. Таблиці порядкових даних в порівняні з таблицями частотних даних

Вибір техніки обмеження статистичного розкриття для даних, представлених у таблицях **(табличні дані),** залежить від того, чи дані представляють частотності або послідовності. Таблиці **даних з підрахунку частотності** представляють одиниці аналізу в комірці. Відповідно, дані можуть бути представлені у відсотках розділенням підрахунків на загальне число, представлене у таблиці (або підсумкове число у ряді чи колонці) і множенням на 100. Таблиці **порядкових даних** представляють сукупність «кількості відсотків», що застосовується до одиниць аналізу в комірці. Рівноцінно, дані можна представляти в якості середнього значення розділяючи загальне значення на число одиниць у комірці.

Для того, щоб офіційно розрізняти **дані з підрахунку частотності** і **порядкові дані,** «кількість відсотків» повинна вимірюватись дещо інше, ніж належність до комірки. Таким чином, таблиці кількості організацій у межах виробничого сектору, що створюються групою із Стандартної промислової класифікації та округом у межах штату, представлені таблицями підрахунку частотності, тоді як таблиці, що представляють загальну вартість поставок для тих самих комірок представлені таблицями порядкових даних.

Б.2. Розмірність таблиці

Якщо значення, представлені у комірках статистичної таблиці є сукупностями двох змінних, то таблиця представлена **двохвимірною** таблицею. Обидва приклади описових комірок, представлені вище, (вік 35-39 років, $40,000-$49,999 щорічного доходу) і (NAICS 4231, Округ Прінс-Джорджес) є з двохвимірних таблиць. Типово, категорії однієї змінної надаються в колонках, а категорії іншої змінної надаються рядами.

Якщо значення, представлені в комірках статистичної таблиці є середніми значеннями над трьома змінними, то таблиця представлена **тривимірною** таблицею. Якщо дані у першому прикладі, представленому вище, було надано округом у штаті Меріленд, результатом може бути інформативна комірка така як (вік 35-39 років, $40,000-$49,999 щорічного доходу, округ Монтґомері). Для другого прикладу, якщо дані було також представлено за роком, результатом могла би бути інформаційна комірка, така як (NAICS 42, округ Прінс Джорджес, 2002). Вважається, що перші двовимірні дані представляються в рядах і колонках, а третя змінна «шарами» або «сторінками», і ці шари представлені окремою таблицею для кожної категорії третьої змінної.

Б.3. Ієрархічна структура змінних

Більшість таблиць це перехресні комбінаційні таблиці двох чи більше класифікаційних змінних, таких як географічні. Класифікаційні змінні можуть мати ієрархічну структуру. Структура

ієрархічного кодування створює проміжні підсумки із структурою кодування змінної. Наприклад, класифікаційні змінні згідно із Статистичною класифікацією господарської діяльності в Північній Америці (NAICS) – це змінні з ієрархічною структурою. Коди чотиризначних підгалузей економіки можуть стягуватись до тризначних кодів для головних галузей промисловості і двозначних для промислових груп. Внутрішня комірка таблиці може мати відношення до спеціального коду NAICS на 4 цифри, із проміжними підсумками, представленими кодами NAICS на 3 цифри, і порядковим підсумком, представленим відповідним кодом на 2 цифри. Ідентифікація будь-якої ієрархічної структури у межах класифікаційних змінних на файлі необхідна для застосування технік обмеження розкриття і для оцінювання захисту.

Географією, зазвичай, називають змінну з ієрархічною структурою. Проте, це не завжди може бути технічно правильним в залежності від структури класифікації. Якщо б географія розбивалась на штати, області, а також на державному рівні, то географія була б ієрархічною змінною, тому що кожен штат класифікується у межах конкретних областей. Проте, якщо географічна класифікація надає населений пункт, столичну зону, округ, штат і область, то класифікація не обов'язково повинна бути ієрархічною, тому що округи, населені пункти, і муніципальні райони не можуть бути складовими частинами один одного.

Б.4. Що таке розкриття?

Хоча визначення розкриття, надане в Розділі, є широким, цей звіт документує методологію, що використовується для обмеження розкриття, і займається лише розкриттям конфіденційної інформації через публічне розголошення результатів дослідження даних. В Розділі І три типи розкриття, подані в книзі Дункана та співавторів (1993) були стисло представлені. Це розкриття особи, розкриття реквізитів і логічно виведене розкриття.

Розкриття особи має місце, коли третя сторона може ідентифікувати суб'єкта або респондента на основі розголошених даних. Розголошення того факту, що фізична особа є респондентом або суб'єктом збору інформації може або ж не може порушити вимоги щодо конфіденційності. Для таблиць даних, що розголошують особу в загальному не представляють собою розкриття, за винятком якщо така ідентифікація призводить до оприлюднення конфіденційної інформації (розкриття реквізитів) про тих, кого ідентифіковано. Для мікроданих ідентифікація в загальному розглядається як розкриття, тому що записи мікроданих зазвичай настільки детальні, що ідентифікація автоматично викриє додаткову характерну інформацію, що не використовувалась при ідентифікації запису. Таким чином, методи обмеження розкриття, що застосовуються до файлів з мікроданими, обмежують або змінюють інформацію, що може використовуватись для ідентифікації конкретних респондентів або суб'єктів даних.

Розкриття реквізитів трапляється, коли конфіденційна інформація про суб'єкт даних розголошується і її можна приписати суб'єкту. Розкриття реквізитів трапляється, коли конфіденційна інформація про особу чи ділові операції фірми розголошуються, або якщо можна здійснити їх чітку оцінку. Таким чином, розкриття реквізитів містить ідентифікацію суб'єкта та оприлюднення конфіденційної інформації, що має відношення до суб'єкта.

Розкриття реквізитів є головною проблемою для більшості статистичних агентств, при вирішенні

того, чи розголошувати табличні дані. Методи обмеження розкриття, що застосовуються до таблиць, гарантують, що дані про респондента публікуються лише як частина цілого із достатнім числом інших респондентів для зміни реквізитів окремого респондента.

Третій тип розкриття, **логічно виведене розкриття,** трапляється коли індивідуальну інформацію можна логічно вивести з високою достовірністю із статистичних якостей розголошених даних. Наприклад, дані можуть показувати високе співвідношення між доходом і ціною купівлі будинку. Так як ціна купівлі будинку є типово публічною інформацією, третя сторона може використати цю інформацію щоб зробити висновок про дохід суб'єкта даних. Існує дві основних причини чому деякі статистичні агентства не переймаються логічно виведеним розкриттям в табличних або мікроданих. Першою причиною є те, що головною метою статистичних даних є надати користувачам можливість зробити висновок і зрозуміти співвідношення між змінними. Якщо статистичні агентства прирівняли розкриття до логічного виведення, дуже мало даних буде розголошено. Другою причиною є те, що логічні виведення призначені для передбачення сукупної поведінки, а не окремих характерних рис, і таким чином, часто представляють негативне передбачення значень особистих даних. Логічно виведене розкриття все ще залишається проблемою там, де наявні випадки винятково точних статистичних асоціацій, а також, де можуть використовуватись регресійні моделі для генерування передбачень. Логічно виведене розкриття є важливим фактором для перегляду результатів аналітичних досліджень, отриманих або з дослідницького інформаційного центру або через дослідницький проект з програмою агентства щодо даних з обмеженим доступом. Ризик розкриття може існувати в регресійних моделях, що містять лише повністю інтерактивні набори фіктивних змінних в якості незалежних змінних. В таких випадках агентства потребують надалі вивчати потенційні ризики розкриття через використання певних регресійних моделей.

B. Системи обробки запитів в режимі реального часу

Розповсюдження даних через наявність систем обробки запитів в режимі реального часу вимагає спеціального застосування методів обмеження розкриття. Системи обробки запитів в режимі реального часу можуть мати можливості багатоцільового застосування. Найпростішою формою застосування є та, коли система здійснює доступ до файлів із стислим викладом, що містять сумарні дані, які вже було перевірено на конфіденційність і методи обмеження розкриття, що застосовуються. Іншою здатністю є розповсюдження таблиць даних від запитів в режимі реального часу щодо файлів з мікроданими, які вже підлягали захисту. Прикладні програми, що здійснюють доступ до незахищених мікроданих, можуть представляти ризик розкриття особи, коли опитування обмежується до маленької географічної зони чи категорії. Це представляє значну проблему для послідовності незалежних опитувань про маленькі географічні зони або категорії. Спеціалізовані таблиці даних, створені в результаті опитувань і закладені у незахищених файлах з мікроданими, повинні перейти через серію фільтрів, де застосовуються правила обмеження розкриття.

Чотири агентства розробили системи опитування в режимі реального часу з різноманітними можливостями для користувачів, та генерування спеціальних таблиць даних. Центри з контролю і профілактики захворювань розробили «CDC Wonder» ((Широкомасштабні інтерактивні дані для епідеміологічних досліджень (WONDER)), що знаходиться за посиланням http://www.cdc.gov/nchs/index.htm. Система CDC wonder дозволяє користувачам подавати опитування в пакети даних публічного користування про мораль (смерть), захворюваність на рак, ВІЛ і СНІД, поведінкові фактори ризику, діабет, народжуваність (народження), а також дані перепису населення в

системному блоці CDC і запитувані дані без затримки підлягають підсумовуванню. Дані попередньо перевіряються на конфіденційність за допомогою методів обмеження розкриття, що застосовуються до того, як додаються до бази даних. Користувачі системи «CDC wonder» підлягають дії обмежень агентства щодо використання даних, що забороняють з'єднання даних з іншими пакетами даних чи інформації для цілей ідентифікації особи. Бюро трудової статистики також має систему інтерактивного опитування, доступну за посиланням http://www.bls.gov/data/sa.htm, яке надає користувачам доступ до зведених даних першого рівня (застосовується обмеження розкриття) для створення спеціалізованих таблиць.

Служба економічних досліджень спільно з Національною службою сільськогосподарської статистики розробили систему, доступну за посиланням http://www.ers.usda.gov/Data/ARMS/ користувачам для створення спеціалізованих таблиць з даними здійснюючи доступ до даних з програми Опитування з управління сільськогосподарськими ресурсами (ARMS). У системі ARMS, обмеження розкриття вже було застосовано до мікроданих. Бюро перепису населення розробило «Американський шукач фактів», доступний за посиланням **http://www.census.gov,** що надає користувачам доступ як до зведених табличних даних, так і до файлів з мікроданими. Передова система опитування американського шукача фактів, має правила конфіденційності і методи розкриття, вбудовані в систему для того, щоб подані користувачами опитування пройшли перевірку на предмет розкриття перед тим, як користувач зможе переглянути результати. В Національному центрі освітньої статистики (NCES) всі дані вибіркового опитування щодо середньої спеціальної освіти доступні через використання засобів аналізу даних, що формують таблиці до трьох вимірів і надають кореляційні матриці. Крім того, дані щодо початкової і середньої освіти із Національної оцінки прогресу освіти (NAEP) також доступні в інтерактивних засобах збору даних. Більш детальний опис інтерактивних систем опитування міститься в Розділі 4 Статті B.

Г. Таблиці підрахунків або частотностей

Дані, що збираються з більшості опитувань про людей, публікуються в таблицях, які показують підрахунки (число людей за категоріями) або частотностей (частка або відсоток людей за категоріями). Частина таблиці, опублікованої із вибіркового опитування домашніх господарств, що збирає інформацію щодо споживання електроенергії, показана в Таблиці 1 нижче в якості прикладу.

Г.1. Вибірка як метод обмеження статистичного розкриття

Один метод із захисту конфіденційності даних полягає у проведенні вибіркового опитування замість перепису населення. Техніки обмеження розкриття не застосовуються в Таблиці 1, хоч, опитуваним надавалось зобов'язання конфіденційності, тому що це широкомасштабне **вибіркове** опитування. Оціночні показники обчислюються множенням даних респондента на вибіркову вагу і потім об'єднанням всіх зважених відповідей. Коли дані використовуються для того, щоб здійснювати оцінку стосовно населення, з якого робиться вибірка, вони в основному врегульовуються вибірковою

вагою, що бере до уваги особливості процедури вибірки. Зважені суми замінюють собою робочих частот в публікованих таблицях. Використання вибіркових ваг робить дані окремих респондентів менш піддатливими для ідентифікації із опублікованих підсумків, коли самі значення ваг не розкриваються. Зокрема, якщо зважування відповідей опитування є комплексним, то опублікована оцінка може приховувати той факт, що існує лише одна чи двоє осіб, що роблять внески в комірку. Через те, що зважені числа представляють всі домашні господарства в Сполучених Штатах, підрахунки в Таблиці 1 подаються в одиницях мільйонів домашніх господарств. Вони були виведені із вибіркового опитування менш ніж 7000 домашніх господарств. Це ілюструє захист, що надається окремим респондентом через здійснення вибірки та попереднє оцінювання.

Таблиця 1: Зразок без розкриття

Число домашніх господарств по загальній опалюваній площі і сімейному доходу (Мільйонів домашніх господарств США)

1997 Сімейний дохід

Загальна опалювана площа кв.фут	Підсумок	Менше ніж $10000	$10000 до $24999	$25000 до $49999	$50000 або більше	Нижче прожиткового мінімуму	Допущений до федеральної допомоги
Менше ніж 600	7.9	2.9	3.1	1.6	0.3	2.7	4.9
600 до 999	21.5	4.3	8.6	6.0	2.6	4.6	10.2
1000 до 1599	30.4	2.8	9.7	10.8	7.0	3.7	9.9
1600 до 1999	15.3	.6	3.2	5.4	6.1	0.9	2.8
2000 до 2399	7.9	.2	1.2	2.5	4.0	0.3	1.1
2400 до 2999	5.3	Q	0.3	1.4	3.4	0.2	0.5
3000 або більше	4.1	Q	0.3	.9	2.8	Q	0.4

ЗАМІТКА: Q – Дані, що утримуються, тому що відносна стандартна помилка перевищує 50% або вибірку було здійснено менше ніж з 10 домашніх господарств.

ДЖЕРЕЛО: «Характеристики житлового забезпечення за 1997», Опитування щодо побутового споживання електроенергії, Управління з інформації в області енергетики, DOE/EIA-0632(97), сторінка 58.

Коли стає точно відомо, що особа є респондентом в дослідженні, завдання визначення особи і його/її характерних рис є значно простішим, ніж коли є висока ймовірність того, що особа не представлена в таблиці або мікроданих взагалі. Якщо все-таки повні розрахункові дані виявлять, що лише респондент використовував інформацію про те, що фізична особа була респондентом, то його або її ідентичність було б підтверджено а їх характерні риси розголошено. Зібрання даних, що базується на вибірці осіб, є захисним, тому що присутність записів даних осіб не є певним а респондент, який виявляється унікальним, може бути не тією особою за яку він/вона вважається.

Крім того, багато агентств вимагає, щоб попередні оцінки досягали нормативної точності перед тим, як вони публікуються. У Таблиці 1 комірки з «Q» нічого відображають, тому що відносна стандартна помилка більша, ніж 50 відсотків. Вимоги щодо точності вибіркового опитування, такого як це, призводять до того, що більше комірок утримуються від публікації, ніж це вимагає обмеження розкриття. У Таблиці 1 значення у комірках, позначені Q, можуть виводитись вилучаючи інші комірки у ряді із граничного підсумку. Мета Q це не обов'язково приховувати значення комірки від громадськості, але швидше вказати, що будь-яка виведена таким чином кількість не відповідає вимогам агентства щодо точності.

Здійснення вибірки може знизити ризики розкриття від опублікованих даних в залежності від частоти вибірки, число і детальність змінних, що приводяться у формі таблиці, і чи взагалі існує публічний перелік повного населення, з якого здійснюється вибірка. Вибірка повинна бути вільна від будь-яких посторонніх значень, таких як фізичні особи чи організації із незвичайними характеристиками. Використання методології здійснення вибірки не гарантує, що опубліковані дані вільні від ризиків розкриття, а будь-які опубліковані таблиці із вибірки повинні все-ще перевірятись.

Г.2. Визначення комірок, що містять конфіденційну інформацію

В дискусії нижче ми виділяємо два класи правил обмеження розкриття для таблиць підрахунків або частотностей. Перший клас складається із спеціальних правил, розроблених для певних таблиць для захисту проти потенційної шкоди агентству, або респонденту через розкриття конфіденційної інформації. Такі правила відрізняються від агентства до агентства і від таблиці до таблиці. Ці спеціальні правила в загальному призначені для забезпечення захисту даних, що вважаються агентством особливо чутливими до розкриття. Другий клас є більш загальними, де число в комірці, як вважається, представляє неприпустимий ризик розкриття, такий як: комірка визначається як конфіденційна, якщо число респондентів менше ніж вказана гранична величина (правило граничної величини).

Г.2.а Спеціальні правила

Спеціальні правила накладають обмеження на рівень деталізації, що може надаватись у таблиці. Наприклад, правила Адміністрації соціального забезпечення (SSA) забороняють таблиці даних, в яких значення комірки всередині ряду чи колонки таблиці рівне граничному підсумку, або який би дозволив користувачам визначити вік фізичної особи у межах п'ятирічного інтервалу, прибутки у межах інтервалу в $1000 або привілеї у межах інтервалу в $50. Таблиці 2 і 3 ілюструють ці правила.

Вони також ілюструють метод структурної перебудови таблиць і комбінування категорій для обмеження розкриття в таблицях.

Таблиця 2 представляє двовимірну таблицю, що показує число бенефіціарів по країнах і розміру доходів. Цю таблицю не можна розголошувати громадськості, тому що дані, показані для країн Б і Г, порушують правила розкриття Соціального страхування. Для країни Г в наявності є лише одна комірка з позитивним значенням, а бенефіціар в цій країні, як відомо, отримує доходи від $40 до $59 на місяць. Це порушує два правила. Перше – значення комірки, що надає детальну інформацію, дорівнює підсумку ряду; і друге – це розкриває той факт, що всі бенефіціари в країні отримують від $40 до $59 на місяць допомогою. Цей інтервал менший, ніж необхідний в $50. Для країни Б існує 2 комірки з позитивними значеннями, але діапазон можливих грошової допомоги коливається від $40 до $79 на місяць, інтервал менший, ніж необхідні $50.

Таблиця 2: Зразок – З розкриттям

Число бенефіціарів по сумі щомісячної грошової допомоги і країнах

Сума щомісячної грошової допомоги

Країна	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Підсумок
А	2	4	18	20	7	1	52
Б	--	-	7	9	-	-	16
В	--	6	30	15	4	-	55
Г	-	-	2	--	-	-	2

ДЖЕРЕЛО: Робочий документ із статистичної політики 2 FCSM (Федеральний комітет із статистичної методології).

Для захисту конфіденційності, Таблицю 2 можна було б реструктуризувати, а ряди або колонки комбінувати (іноді іменовані як «категорії згортання» або «стягування»). Комбінування ряду для округу Б із рядом для округу Г все ще буде розголошувати, що діапазон грошової допомоги від $40 до $79. Комбінування А з Б і В з Г не забезпечують необхідного захисту, як ілюстровано в Таблиці 3.

Таблиця 3: Приклад – Без розкриття

Число бенефіціарів по сумі щомісячної грошової допомоги та округах

Сума щомісячної грошової допомоги

Округ	$0-19	$20-39	$40-59	$60-79	$80-99	$100+	Підсумок
А і Б	2	4	25	29	7	1	68

В і Г	--	6	32	15	4	-	57

ДЖЕРЕЛО: Робочий документ із статистичної політики 2 FCSM (Федеральний комітет із статистичної методології)..

Г.2.б. Правило граничного значення

Із дотриманням правила граничного значення, комірка в таблиці частотностей визначається як **чутлива**, якщо число респондентів менше, ніж певне визначене число. Деякі агентства вимагають принаймні 5 респондентів на комірку, тоді як інші вимагають 3. За певних обставин це число може бути значно більшим. Вибір мінімального числа в загальному робиться з огляду на: (а) конфіденційність інформації, яку агентства обдумує опублікувати, (б) рівень захисту, який агентство визначає як необхідний, беручи до уваги ступінь точності, необхідний для досягнення розкриття.

Г.3. Захист чутливих комірок після зведення в таблиці

В таблицях частотних даних, якщо її комірки було ідентифіковано як чутливі, агентство повинне приймати міри для захисту конфіденційних даних. В загальному є два підходи для здійснення цього. Один полягає у внесенні змін до самої таблиці. Це здійснюється як частина, або після зведення в таблиці. Ці методи включають: реструктуризацію таблиць і комбінування категорій (як продемонстровано вище), приховування комірки, довільне округлення, контрольоване округлення, або контрольоване коригування у формі таблиці. Другий підхід, що недавно розвинувся, це застосування методів з мікроданими до файлу з даними перед зведенням у таблиці. Ці методи особливо ефективні для використання з інтерактивними системами опитування або ж коли з одного файлу з даними будуть створюватись багато таблиць. Цей підхід проілюстровано в розділі Г.4 цього ж розділу.

Таблиця 4 це умовний приклад таблиці з розкриттями. Набір умовних даних складається х інформації, що стосується дітей-правопорушників. Комірки в Таблиці 4 з менш ніж 5 респондентами визначаються як чутливі, і виділяються зірочкою. Ця таблиця використовується для ілюстрації приховування комірки, довільного округлення, контрольованого округлення, і контрольованого врегулювання в табличній формі в статтях нижче.

Таблиця 4: Приклад – Із розкриттям

Число дітей-правопорушників по округах і рівню освіти голови домашнього господарства

Рівень освіти голови домашнього господарства

Округ	Низький	Середній	Високий	Дуже високий	Підсумок
Альфа	15	**1***	**3***	**1***	20
Бета	20	10	10	15	55
Гамма	**3***	10	10	**2***	25
Дельта	12	14	7	**2***	35

Підсумок	50	35	30	20	135

ДЖЕРЕЛО: Числа взяли із Кокс, Макдональд, і Нельсон (1986). Назви, заголовки ряду і колонки є умовними.

Г.3.а. Приховування

Один з найбільш відомих методів захисту конфіденційних комірок здійснюється за допомогою **приховування.** У рядку чи колонці із відкинутою конфіденційною коміркою принаймні одна додаткова комірка повинна бути відкинута, або значення в чутливій комірці може точно вираховуватись через віднімання від граничного підсумку. З цієї причини певні нечутливі комірки також повинні бути відкинуті. Вони іменуються як **додаткові** приховування. Попри те, що є можливим вибирати комірки для додаткового приховування вручну мало не у всіх випадках, важко гарантувати, щоб результат надавав належний захист.

Таблиця 5 показує приклад системи відкинутих комірок для Таблиці 4, що має принаймні дві відкинуті комірки в кожному рядку і колонці. Ця таблиця, як виявляється, пропонує захист для чутливих комірок, проте, детальніший розгляд показує, що розкриття конфіденційних даних все ще має місце.

Таблиця 5: Зразок – З розкриттям, не захищене розкриттям

Число дітей-правопорушників по країнах і рівень освіти голови домашнього господарства

Рівень освіти голови домашнього господарства

Округ	Низький	Середній	високий	Дуже високий	Підсумок
Альфа	15	D1	D2	D3	20
Бета	20	D4	D5	15	55
Гамма	D6	10	10	D7	25
Дельта	D8	14	7	D9	35
Підсумок	50	35	30	20	135

ПРИМІТКА: Г вказує дані, утримані для обмеження розкриття.

ДЖЕРЕЛО: Числа взяті із Кокс, Макдональд, і Нельсон (1986). Назви, заголовки ряду і колонки є умовними.

Прийміть до уваги наступну лінійну комбінацію введених даних в рядках і колонках: Рядок 1 (округ Альфа) + Рядок 2 (округ Бета) − Колонка 2 (середній рівень освіти) − Колонка 3 (високий рівень освіти), можуть писатись як

(15 + D1 + D2 + D3) + (20 + D4 + D5 + 15) − (D1 + D4 + 10 + 14) − (D2 + D5 + 10 + 7) = 20 +

55 − 35 − 30.

Це зводиться до D3 = 1.

Цей приклад показує, що вибір комірок для додаткового приховування це складний процес. Математичні методи лінійного програмування використовуються для автоматичного обирання комірок, та для додаткового приховування, а також для здійснення **аудиту** запропонованої моделі приховування (наприклад, Таблиця 5) для того, щоб побачити чи вона забезпечує необхідний захист. Розділ IV надає більше деталей з математичних завдань щодо обрання додаткових комірок і здійснення аудиту моделей приховування.

Таблиця 6 показує нашу таблицю із системою відкинутих комірок, що не надає належного захисту для чутливих комірок. Однак, Таблиця 6 демонструє одну з проблем із приховуванням. Із загальної кількості в 16 внутрішніх комірок, лише 7 комірок публікуються, тоді як 9 приховуються.

Таблиця 6: Приклад – Без розкриття, захищається приховуванням

Число дітей-правопорушників по округах і рівень освіти голови домашнього господарства

Рівень освіти голови домашнього господарства

Округ	Низький	Середній	Високий	Дуже високий	Підсумок
Альфа	15	D	D	D	20
Бета	20	10	10	15	55
Гамма	D	D	10	D	25
Дельта	D	14	D	D	35
Підсумок	50	35	30	20	135

ПРИМІТКА: D вказує на дані, що утримуються для обмеження розкриття.

ДЖЕРЕЛО: Числа беруться з Кокс, Макдональд, і Нельсон (1986). Назви, заголовки рядків і колонок є умовними.

Г.3.б. Довільне округлення

Для того, щоб зменшити кількість втрат даних, що трапляються через приховування комірок в чутливих таблицях, методи збурень альтернативних даних, такі як довільне і контрольоване округлення, доступні для захисту чутливих комірок у таблицях, які показують частотні дані. При **довільному округленні** значення комірки округлюються, але замість використання стандартних правил з округлення, приймається довільне рішення стосовно того, чи вони будуть округлені в більшу або в меншу сторону. (Більш теоретичне обговорення цього методу міститься в книзі «Елементи контролю статистичного розкриття», написаній Леоном Вілленборгом і Тон де Ваалом, 2011).

В цьому прикладі прийнято вважати, що кожна комірка буде округлена до величини, кратної 5. Кожна кількість комірок, X, можна писати у формі

$$X = 5q + r,$$

де q це невід'ємне ціле число, а r це залишок (який може прийняти одне із 5 значень: 0, 1, 2, 3, 4). Цей підрахунок буде округлений в більшу сторону до 5*(q+1) з імовірністю r/5; і буде округлений в меншу сторону до 5*q з імовірністю (1-r/5). Можливий результат продемонстрований у Таблиці 7.

Таблиця 7: Приклад – Без розкриття, захищений довільним округленням

Число дітей-правопорушників по округах і рівень освіти голови домашнього господарства

Рівень освіти голови домашнього господарства

Округ	Низький	Середній	Високий	Дуже високий	Підсумок
Альфа	15	0	0	0	20
Бета	20	10	10	15	55
Гамма	5	10	10	0	25
Дельта	15	15	10	0	35
Підсумок	50	35	30	20	135

ДЖЕРЕЛО: Числа беруться з Cox, McDonald, and Nelson (1986). Назви, заголовки рядів і колонок є умовними.

Оскільки округлення здійснюється окремо для кожної комірки в таблиці, рядки і стовпчики не обов'язково додаються до опублікованих підсумків для рядків і стовпчиків. У Таблиці 7 підсумок для першого рядка це 20, але сума значень для внутрішніх комірок у першому рядку це 15. Таблиця, підготовлена з використанням довільного округлення могла б призвести до того, що громадськість втратить довіру до чисел: щонайменше вона виглядає так, ніби агентство не може додати нові значення.

Г.3.в. Контрольоване округлення

Для того, щоб розв'язати проблему адитивності, було розроблено процедуру, що називається **контрольованим округленням**. Це форма довільного округлення, але вона обмежена до володіння сумою опублікованих введених даних в кожному рядку і стовпчику, що відповідає відповідним опублікованим граничним підсумкам (див. Кокс та Ернст, 1982). Методи лінійного програмування використовуються для ідентифікації контрольованого округлення для таблиці. Контрольоване округлення використовується Адміністрацією соціального забезпечення у статистичних таблицях, що показують підрахунки частот. Таблиця 8 ілюструє контрольоване округлення, де сума значень

комірки у кожному рядку і стовпчику обмежені для того, щоб відповідати сумі опублікованих підсумків.

Г.3.г. Контрольоване корегування у формі таблиць

Контрольоване коригування у формі таблиць це відносно новий підхід, подібний до контрольованого округлення, але він найбільш цінний, коли застосовується до таблиць порядкових даних. Цей метод спочатку було іменовано як «узагальнені табличні дані». Він був описаний як контрольоване коригування у формі таблиць у наступній роботі (Кокс і Дандекар, 2004). Для порядкових даних, правило лінійної чутливості використовується для визначення того, які комірки є чутливими до розкриття. Із контрольованим табличним коригуванням кожне оригінальне чутливе значення таблиці замінюється безпечним значенням, яке знаходиться на «достатній відстані» від правильного значення; значення нечутливої комірки, мінімально відкориговані для гарантії того, щоб опубліковані граничні підсумки були адитивними. Під «достатньою відстанню» від правдивого значення мають на увазі значення, яке необхідно додати до підсумку комірки, що б зробило комірку нечутливою відповідно до правила лінійної чутливості при його застосуванні. Для частотних даних, більшість правил лінійної чутливості прирівнюються до правила граничного значення для 3 респондентів, а «достатня відстань» від правдивого значення включатиме в себе зміну значення або на 1, або на 2. Тобто значення чутливих комірок будуть змінені або на 0, або на 3. Це ідентично до округлення до бази 3.

Таблиця 8: Приклад – Без розкриття, захищена контрольованим округленням

Число дітей-правопорушників по округах і рівень освіти голови домашнього господарства

Рівень освіти голови домашнього господарства

Округ	Низький	Середній	Високий	Дуже високий	Підсумок
Альфа	15	0	5	0	20
Бета	20	10	10	15	55
Гамма	5	10	10	0	25
Дельта	10	15	5	5	35
Підсумок	50	35	30	20	135

ДЖЕРЕЛО: Числа беруться з Cox, McDonald, and Nelson (1986). Назви, заголовки рядків і колонок є умовними.

Таблиця 9 ілюструє спрощений спосіб для реалізації контрольованого коригування у формі таблиці, як це описано у книзі Дандекар (2004). Внутрішні чутливі комірки спочатку перелічені в порядку зменшення від найбільш до найменш конфіденційних (2, 2, 1, 1). Коригування застосовуються послідовно, починаючи з першої комірки. Перша комірка змінюється навмання до 0 або 3 (або віднімаючи 2, або додаючи). Подальші коригування будуть реалізовуватись з альтернативними

знаками. Отже, якщо перша комірка змінюється додаванням 1, то друга комірка змінюється відніманням 2, третя змінюється додаванням 2, а остання змінюється відніманням 1. Як тільки внутрішні чутливі комірки було змінено, не потрібно жодних додаткових змін у внутрішніх нечутливих комірках (як це типово робиться з контрольованим округленням). Підсумки в таблиці граничних значень перераховуються для того, щоб відзвітувати за зміни, внесені до внутрішніх чутливих комірок. Ці зміни потрібні для того, щоб додавались таблиці. У Таблиці 9 граничні підсумкові значення коригуються для того, щоб мінімізувати відсоток, на який змінюються комірки. У цьому прикладі не потрібно вносити жодних змін до загального підсумку.

Таблиця 9: Приклад -- Без розкриття – Захищена контрольованим коригуванням в табличній формі

Число дітей-правопорушників по округах і рівень освіти голови домашнього господарства

Рівень освіти голови домашнього господарства

Округ	Низький	Середній	Високий	Дуже високий	Підсумок
Альфа	15	**1* − 1 = 0**	3	**1* + 2 = 3**	**20 + 1 = 21**
Бета	20	10	10	15	55
Гамма	3	10	10	**2* − 2 = 0**	**25 − 2 = 23**
Дельта	12	14	7	**2* + 1 = 3**	**35 + 1 = 36**
Підсумок	50	**35 − 1 = 34**	30	**20 + 1 = 21**	135

Контрольовані коригування в табличній формі для окремих значень комірок показані **Напівжирним** шрифтом.

Г.4. Захист чутливих комірок перед зведенням у таблиці.

Табличні дані можуть бути захищеними застосуванням методів захисту розкриття до файлів з мікроданими, що лежать в їх основі, для гарантії, що будь-які таблиці які формуються із файлів з мікроданими є повністю захищеними. Цей підхід є особливо ефективним, якщо в наявності є багато табличних зведень, що створюються із тих самих даних.

Бюро перепису населення було лідером в застосування методів мікроданих для захисту файлів, основаних на Переписі, що проводиться з десятирічними інтервалами. Обмін даними ілюстровано в секції II.F.2.c, а також описано в Domingo-Ferrer (2002). Перепис, що проводиться з десятирічними інтервалами, збирає вихідні дані зі всіх домашніх господарств в США. Він збирає більш вичерпні дані через звітну форму із зразка домашніх господарств США. Обидва блоки даних підлягають процедурі обміну даними. Ця техніка використовувалась для короткочасних даних в переписі

населення 1990 року, але її було переглянуто і розширено до довготривалих даних у 2000 році. Процедура тепер приймає цільовий підхід щодо обміну даними, що збільшує ефективність процедури із деякими витратами по відношенню до зміщеності дисперсії. Всі Табличні зведення, що повторюються кожні десять років, походять від файлів, отриманих в результаті обміну, що гарантує узгодженість таблиць та уникає проблем, пов'язаних із захистом взаємозв'язаних таблиць.

В 1990 було використано іншу процедуру в редагуванні конфіденційності для вибіркових даних, що називалась «незаповнений і приписаний», див. секцію II.F.2.г. В цій техніці обрані записи мають певні значення, які видалені або трактуються як відсутні. Оскільки, зазвичай, в наявності є попередні процедури для вирахування втрачених даних, «незаповнений і приписаний» має певні переваги в економії. Однак, процедура зменшує ефективний розмір вибірки, а компенсації при обчисленні дисперсії іноді важко досягти. У певному значенні, «незаповнений і приписаний» це попередник технологій узагальнених даних, що на даний час досліджуються в Бюро перепису населення та в інших місцях (Раунтан і співавтори, 2003). Перевага обміну даними полягає в тому, що вона максимізує інформацію, яка може бути надана в таблицях. Крім того, всі таблиці захищені значним чином.

Д. Таблиці порядкових даних

Таблиці, що показують порядкові дані, мають унікальний набір проблем розкриття. Порядкові дані в загальному представляють собою невід'ємні кількості, що доповідаються в оглядах або переписах комерційних підприємств, ферм або закладів. Розподіл цих опублікованих значень ймовірно буде асиметричним, і кілька суб'єктів будуть мати дуже великі значення. Обмеження розкриття в цьому випадку зосереджується на тому, щоб переконатись, що опубліковані дані не могли бути використані для оцінки занадто малого діапазону значень, які повідомляються найбільш значним респондентом. Захищаючи опубліковані значення, ми в результаті маємо можливість захищати всі значення.

Правила лінійної чутливості використовуються для ідентифікації комірок, що є «чутливими» і потребують захисту. Нещодавне дослідження зосередилось на використанні захисту до файлу з мікроданими перед зведенням в таблиці. Це забезпечує перевагу, особливо якщо зведення в таблиці будуть надаватись через систему опитування. Історично приховування комірки використовувалось для захисту чутливих комірок в таблицях. Приховування комірки здійснюється як частина формування таблиці.

Д.1. Визначення чутливих комірок – Правила лінійної чутливості

Для порядкових даних менш ймовірним є те, що саме здійснення вибірки забезпечить захист від розкриття, тому що більшість планів вибірки для економічних опитувань містять групу організацій більшого об'єму, що із впевненістю обираються. Таким чином, елементи, що є найбільш видимими через їх розмір, не отримують будь-якого захисту від здійснення вибірки. Для таблиць із порядковими даними, правила, що називаються **правилами початкового приховування** або **міри лінійної чутливості,** були розроблені для визначення того, чи дана комірка таблиці могла б розголосити інформацію про окремого респондента. Комірки, що не проходять тест на лінійну чутливість, визначаються як **чутливі** комірки, та утримуються від публікування.

Правила початкового приховування, що найчастіше використовуються для ідентифікації чутливих комірок державними органами представлені **(n) правилом граничного значення, «правилом (n, k)»,** і «**p-відсоткок»** або «правилами **pq»**. Див. кокс, (1981). Всі вони основані на бажанні ускладнити одному респонденту оцінювання значень, що занадто детально повідомляються іншим респондентом. Найбільше опубліковане значення найімовірніше буде точно оцінене. Правила початкового приховування можуть застосовуватись до частотних даних. Але, оскільки всі респонденти вносять однакові значення до підрахунку частотності, правила граничних значень порушуються, а комірка є чутливою, якщо вона має занадто мало респондентів. «Правила p% і pq» порушують правило граничного значення 3, коли застосовуються до рахункових даних. Правила початкового приховування обговорюються більш детально в Секції VI.Б.1.

Д.2 Захист чутливих комірок після зведення в таблиці

Таблиці для публікації поширюються із файлів з мікроданими. Під час накопичення, правило лінійної чутливості використовується для ідентифікації будь-яких чутливих комірок. Як тільки чутливі комірки було ідентифіковано, є 3 опції: реструктурувати таблицю і стягувати комірки, поки не залишиться жодних чутливих комірок, використати приховування комірок, або застосувати контрольоване табличне коригування. Із приховуванням комірки, як тільки чутливі комірки було ідентифіковано, вони утримуються від публікації. Вони називаються **початковими приховуваннями.** Інші комірки, що називаються **додатковими приховуваннями,** обираються і приховуються для того, щоб чутливі комірки не можна було вивести доданням або відніманням від опублікованих граничних підсумків. Проблеми, пов'язані із прихованням комірок для таблиць із рахунковими даними, було проілюстровано в Секції В.3.а цього розділу. Ті ж проблеми існують для таблиць порядкових даних.

Контрольоване табличне коригування було проілюстровано для частотних даних в Секції В.3.г. цього розділу. Для порядкових даних, «достатня відстань» це сума, яку необхідно буде додати до підсумку комірки для того, щоб правило лінійної чутливості класифікувало комірку як нечутливу.

Адміністративний спосіб уникнення приховування комірки використовується деякою кількістю агентств. Вони отримують письмовий дозвіл, або **«інформовану згоду»** публікувати конфіденційні комірки, отримані від респондентів, що роблять внесок у комірку. Письмовий дозвіл називається «відмовою» від обіцянки захищати конфіденційні комірки і спеціального дозволу або згоди, щоб агентство публічно розголошувала конфіденційну інформацію. В цьому випадку агентство подає запит респондентам добровільно дати свою згоду після інформування про потребу розголошення інформації, а також запропонованого статистичного або нестатистичного використання інформації. Цей метод найбільш придатний для малих опитувань або блоків таблиць, що включають в себе лише кілька маленьких комірок, де необхідно лише кілька відмов. Звичайно, респонденти повинні бути проінформовані про запропоноване використання даних перед наданням своєї згоди.

Д.3. Захист конфіденційних комірок перед зведенням у таблиці

Існує декілька результатів досліджень з мікроданими для зборів статистичних даних по підприємствах через викривлений характер населення. Проте застосування методів мікроданих для захисту файлів зборів статистичних даних по підприємствах перед зведенням у таблиці спростило захист табличних даних і надало нові результати досліджень з даними.

Бюро перепису населення першим застосувала методи мікроданих для захисту файлів даних рівня

підприємства перед зведенням у таблиці. Технічний прийом накопичення шуму, секція II.E.2.б, був початковим використаним методом, в поєднанні з іншими методами. Зокрема, накопичення шуму використовувалось для захисту щоквартальних показників трудових ресурсів, випущених проектом Longitudinal Employer Household Dynamics (Довготривалі динамічні характеристики домашнього господарства роботодавця). Порядкові дані для підприємств мають тенденцію до викривлення і домінування великими компаніями. Це також може призвести до ситуації, де застосування правил лінійної конфіденційності позначає багато комірок для захисту від розкриття. Накопичення шуму додає шум до кожних даних опитуваного підприємства на невелике процентне відношення. Сума відхилення опублікованих значень залежить від послідовності опублікованих даних, і значення правила лінійної конфіденційності для комірок, що містять дані цього респондента. Якщо комірка містить лише одне підприємство, або якщо окреме підприємство домінує в комірці, то опубліковане значення у комірці не буде навіть наближено рівним значенню домінуючого підприємства, тому що до цього значення були додані шуму. Правдиве опубліковане значення домінуючого підприємства захищається накопиченням шуму. Важливо зазначити, що всі підприємства мають свої значення, помножені на відповідний фактор шуму, або відкориговану масу, перед тим як дані зводяться у таблицю. Помножувачі шуму можуть довільно призначатись для контролювання ефектів шуму на різних типах комірок в межах таблиці.

Накопичення шуму також використовувалось Службою економічних досліджень Міністерства сільського господарства США для захисту опублікованих значень у їх щорічному Дослідженні з управління сільськогосподарськими ресурсами (ARMS), яке доступне через інтерактивну систему опитування. Значення коригуються, що чергується між доданням і відніманням шуму після послідовності спостережень в наборі даних, так щоб підсумкові значення комірки були приблизно такими ж як і після застосування накопичення шуму.

Метод має декілька переваг над приховуванням комірки, які полягають в тому, що вони надають деяку інформацію у більшій кількості комірок таблиці, і він виключає потребу координувати моделі приховування комірки. Ця методологія забезпечує узгодженість у таблицях, що формуються із мікроданих, але важливо те, щоб початкові мікродані були достатньо збурені для того, щоб сформовані таблиці були безпечними для розголошення. Одне обмеження цієї методології полягає в тому, що граничні значення можуть показувати великі зміни в результаті коригування вагових функцій, що лежать в їх основі. Відносини між фактичними значеннями не коригованої комірки і значення коригованої комірки з використанням накопичення шуму повинні переглядатись перед розголошенням даних.

E. Мікродані

Інформація, що збирається про підприємства, в основному представлена послідовними даними. Ці дані ймовірно дуже перекручені, і можливо, будуть опитувані з високим ризиком, яких можна легко ідентифікувати через іншу публічно доступну інформацію. В результаті, особливої обережності треба дотримуватись при обдумуванні розголошення файлів з мікроданими, що містять дані про підприємство. Приклади публічного розголошення файлів з мікроданими із зборів статистичних відомостей по підприємствах включають в себе дані з Огляду споживання електроенергії адміністративною будівлею, що надається Управлінням з інформації в області енергетики, і файли з Сільськогосподарського перепису від 1997, що надаються Бюро перепису населення. Захист від розкриття забезпечується використанням технологій, описаних нижче на додачу до усунення змінних, що служать як прямі ідентифікатори респондентів в огляді.

Довгий час визнавалось, що важко захищати блок мікроданих від розкриття через можливість підбору до зовнішніх джерел даних (Bethlehem, Keller and Panekoek, 1990). Крім того, не існує жодних прийнятих мір ризику розкриття для файлу з мікроданими, отже, немає жодного «стандарту»,

який можна застосувати для гарантії того, щоб захист був належним. «Контрольний перелік з потенціалу розкриття запропонованих розголошень даних» було розроблено Комітетом з конфіденційності і доступу до даних для підтримки агентств в перегляді потенціалу розкриття запропонованих файлів з мікроданими публічного користування, і він є доступний для завантаження за посиланням http://www.fcsm.gov/committees/cdac/. Бюро трудової статистики, Бюро статистики транспорту, Національний центр медичної статистики, Бюро перепису населення та Адміністрація соціального забезпечення використовують контрольний перелік CDAC або інший вдосконалений формат контрольного переліку для перегляду випусків запропонованих даних для будь-якого потенціалу розкриття. Національний науковий фонд також використовує контрольний перелік CDAC в якості інструкцій для своїх підрядників, яких вони повинні дотримуватись при перегляді запропонованого файлу для публічного розголошення. Описані нижче методи для захисту файлів з мікроданими використовуються всіма агентствами, що надають файли з даними публічного користування. Для зменшення потенціалу для розкриття, більшість файлів з мікроданими публічного користування:

1. Включають в себе дані лише із вибірки населення,
2. Не включають очевидні ідентифікатори,
3. Обмежують географічні деталі,
4. Обмежують число і детальна розбивка категорій у межах змінних у файлі.

Додаткові методи використовуються для викривлення змінних високого ризику включають в себе:

1. Відсікання граничних кодів для певних змінних (Верхнє або нижнє кодування),
2. Перекодування в інтервали або округлення,
3. Додавання або множення на довільні числа (шум),
4. Обмін або обмін категоріями (також називається переходом),
5. Обирання записів довільно, викреслювання обраних змінних та умовного нарахування на їх рахунок (також називаються викресленими і приписаними),
6. Об'єднання маленьких груп респондентів і заміна опублікованого значення однієї фізичної особи середнім (також називається спотворенням).

Ці явища будуть ілюструватись з умовним прикладом, який ми використали в попередній секції.

E.1. Здійснення вибірки, усунення ідентифікаторів та обмеження географічних деталей

Перше: включайте лише дані із вибірки населення. Для цього прикладу ми використали 10 відсоткову вибірку населення з дітей-правопорушників. Друге: усуньте ідентифікатори, що прямо встановлюють респондентів, такі як ім'я, адреса, та ідентифікаційні номери. У цьому випадку ідентифікатор це ім'я дитини. Третє: візьміть до уваги географічні деталі. Ми вирішили, що ми не можемо показувати дані про окремий округ для округу, що має менше ніж 30 дітей-правопорушників в населенні. Таким чином, дані з Таблиці 4 показують, що ми не можемо надавати географічні дані для округів Альфа і Гамма. В результаті, округи Альфа і Гамма комбінуються і показуються як АльфГам в Таблиці 9. Ці маніпуляції призводять до умовних файлів з мікроданими, показаних в Таблиці 10.

В цьому прикладі ми обговорювали лише 5 змінних для кожної дитини. Можна уявити, що ці 5 змінних було обрано із більш завершеного блоку даних, включаючи імена батьків, імена і кількість рідних братів та сестер, вік дитини, вік рідних братів і сестер, адреса, школа і так далі. Оскільки

більше змінних включаються у файл з мікроданими для кожної дитини, унікальні комбінації змінних роблять більш ймовірним те, що певна дитина може бути ідентифікована обізнаною особою. Обмеження числа змінних до 5 робить таку ідентифікацію менш ймовірною.

E.2. Змінні високого ризику

Існує ймовірність того, що інформація, доступна для інших серед населення, може використовуватись із даними про дохід, показаними в Таблиці 10, лише для ідентифікації сім'ї дитини-правопорушника. Наприклад, роботодавець голови домашнього господарства в загальному знає свою точну заробітну плату. Змінні, такі як дохід, раса та вік представляють змінні **високого ризику** і вимагають додаткового захисту.

Таблиця 10: Умовні мікродані – Відібрані, ідентифікатори усунено

Географічні деталі усунено – Діти-правопорушники

Число	Округ	Освіта НН	Дохід НН	Раса
1	АльфГам	Високий	61	Б
2	АльфГам	Низький	48	Б
3	АльфГам	Середній	30	Ч
4	АльфГам	Середній	52	Б
5	АльфГам	Дуже високий	117	Б
6	Бета	Дуже високий	138	Ч
7	Бета	Дуже високий	103	Б
8	Бета	Низький	45	Б
9	Бета	Середній	62	Б
10	Бета	Високий	85	Б
11	Дельта	Низький	33	Ч
12	Дельта	Середній	59	Ч
13	Дельта	Середній	59	Б
14	Дельта	Високий	72	Ч

ПРИМІТКА: НН означає голову домашнього господарства. Дохід доповідається у тисячах доларів. Округ АльфГам означає або Альфа, або Гамма.

E.2.a. Верхнє кодування, нижнє кодування, перекодування в інтервали

В цьому прикладі великі значення доходу підлягають **верхньому кодуванню** показуючи лише, що дохід більший ніж 100,000 доларів на рік. Маленькі значення доходу підлягають **нижньому кодуванню** показуючи лише, що дохід менший ніж 40,000 доларів на рік. Крім того, значення доходу **перекодовуються** представленням доходу в інтервалах в 10,000 доларів. Результат цих маніпуляцій приносить файл з умовними даними публічного користування в Таблиці 11. Верхнє кодування, нижнє кодування і записування в інтервали знаходяться у часто використовуваних методів для захисту змінних високого ризику у файлах з мікроданими.

Таблиця 11: Умовні мікродані – Відібрані, ідентифікатори усунено

Географічні деталі обмежені, верхня межа доходу, нижні і записані – Діти-правопорушники

Обмежені географічні дані – Діти-правопорушники

Число	**Округ**	**Освіта НН**	**Дохід НН**	**Раса**
1	АльфГам	Високий	60-69	Б
2	АльфГам	Низький	40-49	Б
3	АльфГам	Середній	<40	Ч
4	АльфГам	Середній	50-59	Б
5	АльфГам	Дуже високий	>100	Б
6	Бета	Дуже високий	>100	Ч
7	Бета	Дуже високий	>100	Б
8	Бета	Низький	40-49	Б
9	Бета	Середній	60-69	Б
10	Бета	Високий	80-89	Б
11	Дельта	Низький	<40	Ч
12	Дельта	Середній	50-59	Ч
13	Дельта	Середній	50-59	Б
14	Дельта	Високий	70-79	Ч

ПРИМІТКА: НН означає голову домашнього господарства. Дохід доповідається в тисячах доларів. Округ АльфГам означає або Альфа або Гамма.

Е.2.б. Додавання довільного шуму

Альтернативний метод спотворення змінних високого ризику, таких як змінних доходу, полягає в додаванні або множенні на довільні числа. Наприклад, у вказаному вище прикладі, припустимо що ми додамо випадкову величину, розподілену за нормальним законом із середнім 0 і стандартним

відхиленням 5 до доходу. Разом із здійсненням вибірки, усуненням ідентифікаторів та обмеженням географічних деталей, це може призвести до створення файлу з мікроданими, такого як Таблиця 12. Для сформування цієї таблиці, було вибрано 14 довільних чисел із визначеного нормального розподілу, і їх було додано до даних про доходи в Таблиці 10.

Таблиця 12: Умовні мікродані – Вибірка здійснена, ідентифікатори усунено

Географічні деталі обмежено, довільний шум додано до доходу – Діти-правопорушники

Число	**Округ**	**Освіта НН**	**Дохід НН**	**Раса**
1	АльфГам	Високий	61	Б
2	АльфГам	Низький	42	Б
3	АльфГам	Середній	32	Ч
4	АльфГам	Середній	52	Б
5	АльфГам	Дуже високий	123	Б
6	Бета	Дуже високий	138	Ч
7	Бета	Дуже високий	94	Б
8	Бета	Низький	46	Б
9	Бета	Середній	61	Б
10	Бета	Високий	82	Б
11	Дельта	Низький	31	Ч
12	Дельта	Середній	52	Ч
13	Дельта	Середній	55	Б
14	Дельта	Високий	61	Ч

ПРИМІТКА: НН означає голову домашнього господарства. Дохід звітується в тисячах доларів. Округ АльфГам означає або Альфа або Гамма.

Е.2.в. Обмін даними та обмін рангом

Обмін включає в себе обрання зразку записів, пошук відповідності у базі даних на сукупності заздалегідь визначених змінних і обмін всіма іншими змінними. Обмін проілюстрований в секції Д.2.д. В цьому прикладі записи було ідентифіковано із різних округів, що підходили за расою, статтю і доходом, а змінними щодо ім’я дитини та рівень освіти в домашньому господарстві обмінювались. Для цілей надання додаткового захисту змінній доходу у файлі з мікроданими, ми можемо вирішити замість цього знайти відповідність в іншому округу по рівню освіти в домашньому господарстві і расі для обміну змінними доходу.

Обмін пропонує можливість обрати певну статистику, яка буде зберігатись протягом операції з обміну. Це супроводжується зобов'язуючим договором між парами по змінних, що обмінювались, і які залучені в цій статистиці. Національний інститут статистичних наук (NISS) має пакет програм, який здійснює та аналізує обмін даними в змінних категорійних даних, що є доступний на їх веб-сайті за посиланням http://www.niss.org/software/dstk.html. Техніка NISS використовує довільний обмін; це надає можливість визначити вплив на статистику, сформовану із набору даних, що пройшли обмін. Для наборів даних із точним показником ризику рівня запису можна застосовувати варіацію, що називається цільовим обміном. Ці записи із високим ризиком автоматично обираються для сполучення в процесі обміну. При цільовому обміні меншу кількість записів залучено, а рівень захисту в загальному вищий. Однак цільова процедура є необ'єктивною, а спроможність представляти загальне твердження щодо якості даних є дуже обмежена.

Обмін рангом надає можливість використання безперервних змінних для визначення пар записів щодо обміну. Замість наполягання на тому, щоб змінні підходили (точно погоджувались), вони визначаються як такі, що тісно основані на їх близькості один до одного по списку, посортованому за безперервною змінною. Записи, близькі по рангу із відсортованої змінної, позначаються як пари для обміну. Нерідко при обміні рангом змінна, що використовується у цьому типі, є тією, яку будуть обмінювати.

Перетасовка даних є іншим методом для модифікації мікроданих, які застосовувались до числових даних. Процедура включає два кроки: спочатку значення конфіденційних змінних модифікуються з використанням техніки загального збурення, а потім застосовується процедура перетасування даних з використанням збурених значень конфіденційних змінних на файлі. Збурені значення сортуються від найнижчого до найвищого значення в повторно перетасованому файлі. Потім збурене значення замінюється оригінальним значенням конфіденційної змінної, основаної на пріоритетності оригінальних значень із конфіденційної змінної. Перед тим, як дані збурюються, умовний розподіл між конфіденційним і не конфіденційними змінними є похідним. Цей метод зберігає співвідношення порядку пріоритетності між конфіденційними і не конфіденційними ознаками, і запобігає втрату корисності даних, які можуть відбутись через застосування методології обміну даними або обміну пріоритетністю. Перетасування даних обговорюється більш детально в Розділі V.

Обмін даними використовувався для захисту конфіденційності 2000 табличних зведень Перепису. Процедура виконувалась на базових мікроданих, а всі табличні зведення із 100% (коротка форма) та із зразка (довга форма) даних було створено із файлів, що підлягали обміну. Це вплинуло на пари домашніх господарств (або партнерські домашні господарства), де одне чи обидва із цих домашніх господарств мали ризик розкриття. Сукупність домашніх господарств, що підлягали перепису, і які вважались такими, що мають ризик розкриття, було обрано із файлів з даними внутрішнього перепису. Ці домашні господарства були унікальними в їх географічній зоні (блок для 100% даних і блокова група для вибіркових даних), основаній на певних характеристиках. Обмін даних із цих домашніх господарств здійснювався на дані із партнерських домашніх господарств, які мали ідентичні характерники по певному набору ключових змінних, але які були із різних географічних місць розташування. Те, які домашні господарства підлягали обміну, не є публічною інформацією. Процедуру обміну було здійснено незалежно для 100% даних і вибіркових даних. Для підтримки якості даних, в наявності був максимальний відсоток записів, які було поміняно на кожен штат за 100% даних та інший максимальний відсоток за вибіркові дані.

Для ілюстрації набору процедур обміну даних, які застосовувались до 100 відсоткового файлу з даними ми використовуємо умовні записи для 20 фізичних осіб в окрузі Альфа, які зробили внесок від Таблиці 4 до 8 включно. Таблиця 13 показує 5 змінних для цих фізичних осіб. Нагадаємо, що

попередні таблиці показували підрахунки фізичних осіб по округах і рівню освіти голови домашнього господарства. Мета обміну даними полягає в надаванні захисту від розкриття для таблиць частотних даних. Однак для того, щоб цього досягнути, здійснюються коригування до файлу з мікроданими перед тим як створюються таблиці. Приймаються наступні кроки для застосування процедур з обміну даними:

1. Візьміть випадкову вибірку записів із файлу з мікроданими (такого як 10% зразок). Припустимо, що записи 4 і 17 було обрано як частину із нашого зразку на 10%.

Таблиця 13: Умовні мікродані

Всі діти-правопорушники в округу Альфа

Число	**Дитина**	**Округ**	**Рівень освіти НН**	**Дохід НН**	**Раса**	**Стать**
1	Джон	Альфа	Дуже високий	201	Ч	Ч
2	Джейкоб	Альфа	Високий	103	Б	Ч
3	Сью	Альфа	Високий	75	Ч	Ж
4	Піт	Альфа	Високий	61	Б	Ч
5	Рамеш	Альфа	Середній	72	Б	Ч
6	Данте	Альфа	Низький	103	Б	Ч
7	Ларі	Альфа	Низький	91	Ч	Ч
8	Мерлін	Альфа	Низький	84	Б	Ж
9	Стів	Альфа	Низький	75	Б	Ч
10	Пол	Альфа	Низький	62	Ч	Ч
11	Рене	Альфа	Низький	58	Б	Ж
12	Вірджінія	Альфа	Низький	56	Ч	Ж
13	Мері	Альфа	Низький	54	Ч	Ж
14	Лаура	Альфа	Низький	52	Б	Ж
15	Том	Альфа	Низький	55	Ч	Ч
16	Ел	Альфа	Низький	48	Б	Ч
17	Майк	Альфа	Низький	48	Б	Ч
18	Філ	Альфа	Низький	41	Ч	Ч
19	Браян	Альфа	Низький	44	Ч	Ч
20	Ненсі	Альфа	Низький	37	Б	Ж

ПРИМІТКИ: НН позначає голову домашнього господарства. Дохід показано в тисячах доларів.

2. Оскільки ми потребуємо таблиць по округах і рівню освіти, ми знаходимо відповідник в іншому окрузі по інших змінних раси, статі і доходу. (В результаті підбору за расою, статтю і доходом, підсумки округу для цих змінних не будуть мінятися через обмін.) Відповідник для запису 4 (Піт) знаходиться в окрузі Бета. Відповідник знаходиться з Альфонсо, голова домашнього господарства в якому має дуже високий рівень освіти. Запис 17 (Майк) пристосовується до Джорджа в окрузі Дельта, голова домашнього господарства якого має середній рівень освіти. Крім того, частина 10% відбірки, що обирається довільно, із інших округів відповідають записам в окрузі Альфа. Один запис із округу Дельта (Джун із високим рівнем освіти) співставляється із Вірджинією, номер запису 12. Один запис із округу Гамма (Хезер із низьким рівнем освіти) співставляєтьтся із Ненсі, із номером запису 20.

3. Після того, як здійснюються всі співставлення, обмін впливає на підібрані записи. Відкоригований файл з мікроданими, після то як цими характеристиками обмінюються, появляється в Таблиці 14.

4. Використовуйте файл з даними, що пройшли обмін, прямо для формування таблиць. Див. Таблицю 15.

Застосування набору процедур із обміну даними має велику перевагу у тому, що багатовимірні таблиці можна легко формувати і захист від розкриття, що застосовується, завжди буде сумісним.

Таблиця 14: Умовні мікродані

Діти-правопорушники після обміну – Показано лише округ Альфа

Число	**Дитина**	**Округ**	**Рівень освіти НН**	**Дохід НН**	**Раса**	**Стать**
1	Джон	Альфа	Дуже високий	201	Ч	Ч
2	Джейкоб	Альфа	Високий	103	Б	Ч
3	Сью	Альфа	Високий	75	Ч	Ж
4*	**Альфонсо**	**Альфа**	**Дуже високий**	**61**	**Б**	Ч
5	Рамеш	Альфа	Середній	72	Б	Ч
6	Данте	Альфа	Низький	103	Б	Ч
7	Ларі	Альфа	Низький	91	Ч	Ч
8	Мерлін	Альфа	Низький	84	Б	Ж
9	Стів	Альфа	Низький	75	Б	Ч
10	Пол	Альфа	Низький	62	Ч	Ч
11	Рене	Альфа	Низький	58	Б	Ж
12*	**Джун**	**Альфа**	**Високий**	**56**	**Ч**	Ж

13	Мері	Альфа	Низький	54	Ч	Ж
14	Лаура	Альфа	Низький	52	Б	Ж
15	Том	Альфа	Низький	55	Ч	Ч
16	Ел	Альфа	Низький	48	Б	Ч
17*	**Джордж**	**Альфа**	**Середній**	**48**	**Б**	Ч
18	Філ	Альфа	Низький	41	Ч	Ч
19	Браян	Альфа	Низький	44	Ч	Ч
20*	**Хізер**	**Альфа**	**Низький**	**37**	**Б**	Ж

Дані: ім’я і рівень освіти, що пройшли обмін в умовному файлі з мікроданими з іншої країни.

ПРИМІТКИ: НН позначає голову домашнього господарства. Дохід показаний в тисячах доларів.

Таблиця 15: Таблиця захищена обміном даних

Число дітей-правопорушників по округах і рівень освіти голови домашнього господарства

Округ	Низький	Середній	Високий	Дуже високий	Підсумок
Альфа	13	2	3	2	20
Бета	18	12	8	17	55
Гамма	5	9	11	0	25
Дельта	14	12	8	1	35
Підсумок	50	35	30	20	135

ДЖЕРЕЛО: Умовні мікродані.

Е.2.г. Метод незаповнених полів і умовного нарахування для записів, що довільно обираються.

Метод незаповнених полів та умовного нарахування включає видалення значень для вибраних змінних окремих респондентів із файлу з мікроданими та їх заміну значеннями для таких же змінних від інших респондентів або через моделювання. Ця техніка ілюстрована з використанням даних, показаних в Таблиці 16.

Таблиця 16: Умовні мікродані – Вибірку проведено, ідентифікатори усунено

Географічні деталі обмежено, використовуючи метод незаповнених полів та умовного нарахування – Діти-правопорушники

Число	**Округ**	**Рівень освіти НН**	**Дохід НН**	**Раса**
1	АльфГамм	Високий	61	Б
2	АльфГамм	Низький	***63***	Б

3	АльфГамм	Середній	30	Ч
4	АльфГамм	Середній	52	Б
5	АльфГамм	Дуже високий	117	Б
6	Бета	Дуже високий	***52***	Ч
7	Бета	Дуже високий	103	Б
8	Бета	Низький	45	Б
9	Бета	Середній	62	Б
10	Бета	Високий	85	Б
11	Дельта	Низький	33	Ч
12	Дельта	Середній	59	Ч
13	Дельта	Середній	***49***	Б
14	Дельта	Високий	72	Ч

ПРИМІТКА: НН означає голову домашнього господарства. Дохід звітується в тисячах доларів. Округ АльфГамм означає або Альфа або Гамма.

Спершу, один запис обирається довільно із кожного округу, придатного для видання, АльфГамм, Бета і Дельта. В обраному записі значення доходу замінюється умовно нарахованим значенням. Якщо довільно обраними записами є 2 в округу АльфГамм, 6 в округу Бета і 13 в округу Дельта, значення доходу, що фіксуються в цих записах, можуть замінюватись 63, 52 і 49 відповідно. Ці числа є також умовними, але ви можете припустити, що умовно нараховані значення було вирахувано, як середнє по всіх домашніх господарствах в окрузі із такою ж расою і рівнем освіти. Метод незаповнених полів використовувався як частина редагування конфіденційності для таблиць частотних даних з файлів з вибірковими даними Перепису 1990 (містить інформацію із великої звітної форми Перепису, що проводиться з десятилітніми інтервалами).

Е.2.д. Викривлення

Викривлення замінює опубліковане значення на середню величину. Існує багато можливих способів введення в дію викривлення. Групи записів для отримання середніх значень можуть формуватись співставленням з іншими змінними або сортуванням шуканої змінної. Число записів в групі (з даних якої буде отримано середні значення) можуть бути встановленими або довільними. Середнє значення, що має відношення до певної групи, може бути закріпленим за всіма членами групи, або за «середнім» членом (як у випадку змінного середнього значення). Воно може здійснюватись на більш ніж одній змінній, з різними групуваннями для кожної змінної.

В нашому прикладі ми ілюструємо цю техніку викривляючи дані про доходи. У повному файлі з мікроданими ми могли б співставити важливі змінні, такі як округ, раса і дві групи рівня освіти (дуже високий, високий) і (середній, низький). Потім викривлення може включати в себе виведення

середніх значень домашніх господарств в кожній групі рівня освіти, наприклад по двоє на раз. В округу Альфа (див. Таблиця 9) це означатиме, що дохід домашнього господарства для групи, яка складається з Джона і Сью, буде замінений середнім значенням їх доходів (139), а дохід домашнього господарства для групи, що складається з Джима і Піта, буде замінено їх середнім значенням (82), і так далі. Після викривлення, файл з даними може підлягати вибірці, усуненню ідентифікаторів, та обмеженню географічних деталей для подальшого зниження ризику ідентифікації.

E.2.e. Цільове приховування

Хоча **приховування** це один з найчастіше використовуваних способів захисту конфіденційних комірок в таблицях, воно може також використовуватись щодо записів у файлах з мікроданими. Коли запис містить граничні значення, або унікальні значення, які не можна належним чином захистити, може бути необхідним видалити окремий запис в його повноті, або ж приховати конфіденційні значення для певних зареєстрованих змінних.

Є. Зведення

Цей розділ описує стандартні методи обмеження розкриття, що використовуються федеральними статистичними агентствами для захисту як таблиць, так і мікроданих. Він значною мірою покладається на прості приклади для ілюстрації понять. Проблема при оцінюванні різних методів полягає в тому, що записи, які підлягають обміну, викривленню та умовному нарахуванню, а також методології зведення не визначаються (або позначаються) на файлі у будь-який спосіб. Це означає, що відкориговані записи не лише захищені, але й високий ступінь невизначеності вводиться в такий спосіб, що які б методи не використовувались для ізоляції будь-якого окремого запису, користувач не буде спроможний визначити із впевненістю, що ізольований запис містить фактичні і не замінені, приписані або викривлені значення. Математичні обґрунтування щодо застосування методології обмеження розкриття в таблицях а також мікродані доповідаються більш детально в Розділах IV і V відповідно. Методики агентства щодо обмеження розкриття описано в Розділі III.

РОЗДІЛ III – Поточна практика федерального статистичного агентства

Цей розділ надає огляд політики, практики і процедур для статистичного обмеження розкриття для 14 Федеральних агентств. Агентства уповноважені, або від них вимагається захищати дані, що ідентифікуються в індивідуальному порядку, відповідно до безлічі статутів, правил і політики. Методи обмеження статистичного розкриття застосовуються агентствами для обмеження ризику розкриття індивідуальної інформації, коли статистика розповсюджується в форматах таблиць або мікроданих.

Цей перегляд практик агентства оснований на трьох джерелах. Перше джерело це Джабін (1993б), документ оснований частково на інформації, що надається статистичними агентствами у відповідь на запит, поданий в 1990 Радою з конфіденційності і доступу до даних, Комітетом з національної статистики. Інше джерело практик агентства було з 1991, коли кожне статистичне агентство просили надати опис поточних практик, стандартів, і планів дослідження з розкриття для табличних і мікроданих. 12 статистичних агентств дали відповідь на цей запит.

Третє джерело було з 2004, коли кожному агентству було подано запит від Комітету з конфіденційності і доступу до даних, підкомітету Федерального комітету із статистичної методології, щодо перегляду і доповнення їх відповідей стосовно поточних практик і стандартів з розкриття, і надати коментарі щодо будь-яких положеннях для доступу дослідника. Таким чином, матеріал в цьому розділі поточний станом на дату публікування.

Перша секція цього розділу підсумовує практики обмеження розкриття для 14 Федеральних статистичних агентств, як показано в Статистичних програмах уряду Сполучених Штатів: Фінансовий рік 2004 (Адміністративно-бюджетне управління). Зведення агентства супроводжуються оглядом поточного статусу політики, практики, і процедур обмеження статистичного розкриття, основаних на наявній інформації. Конкретні методології і стан програмного забезпечення, що використовуються, обговорюються в тій мірі, в якій їх було включено у відгуки окремих агентств.

A. Зведення агентства

A.1. Міністерство сільського господарства

A.1.a. Служба економічних досліджень (ERS)

Практики обмеження розкриття ERS задокументовані у декларації «Політика ERS щодо розповсюдження статистичної інформації» від 28 вересня 1989. Ця декларація передбачає, що: оціночні показники не будуть опубліковані із вибіркових опитувань, за винятком якщо: (1) достатня кількість ненульових звітів, не буде отримано для пунктів у даному класі або комірці з даними для надання статистично чинних результатів, які вільні від ймовірності розкриття інформації про окремих респондентів. У всіх випадках приблизно три дослідження повинні бути в наявності, хоч більш обмежені правила можуть застосовуватись до конфіденційних даних, (2) другою умовою не буде застосування правила концентрації (n, k) або правила домінування для гарантії того, що нерозгорнуті дані для будь-якого одного респондента не опирається до визначеної граничної величини. Для кожного опублікованого значення комірки, респондент повинен представити менш ніж 60 відсотків від підсумку, що публікується, крім випадку коли від респондента отримується письмовий дозвіл. В цьому прикладі (n, k) = (1, 0.6). Обидві умови застосовуються до порядкових даних, в той час як перша умова також застосовується до підрахунків.

Згідно із ERS, доступ до неопублікованих, конфіденційних даних контролюється начальником відповідної галузі. Уповноважені користувачі повинні підписати форми з сертифікації конфіденційності. Обмеження вимагають, щоб дані були підсумовані таким чином, щоб окремі звіти не могли бути розголошені.

ERS не розголошує файли з мікроданими публічного користування. ERS забезпечує доступ до мікроданих через їх програмне забезпечення «віддаленого центру обробки даних» для уповноважених користувачів. ERS буде ділитися даними для статистичних цілей із урядовими агентствами, університетами, та іншими організаціями згідно із угодами про співпрацю, як це описано нижче для Національної служби сільськогосподарської статистики (NASS). Запити організацій згідно із угодами про співпрацю з ERS для зведення у таблиці даних, які початково збиралися NASS, підлягають перегляду NASS.

A.1.б. Національна служба сільськогосподарської статистики (NASS)

NASS підтримує серію Меморандумів з політики і норм (PSM), які документують політику і норми, встановлені для всіх програм Агентства. PSM 12 регулює правила атрибутивного і логічно виведеного розкриття разом із положеннями для ведення спеціальних справ. PSM 7 документує політику NASS щодо розголошення неопублікованих зведених даних та оцінок, а також доступу до файлів з мікроданими. PSM 6 охоплює використання основи для вибірки із списку, включаючи

розкриття особи. PSM 4 представляє правове зобов'язання NASS захищати конфіденційну інформацію і визначає процедури для сертифікації конфіденційності працівників і спеціальних агентів.

Оціночні показники з сільського господарства містять звіти щодо урожаю, поголів'я, звіти про стан навколишнього середовища, та економічні звіти, які NASS регулярно підготовляють через Раду з сільськогосподарської статистики. Програма Оціночних показників сільського господарства визначає першочергові приховування використовуючи правило граничного значення трьох, і правило домінування (n, k). Значення n і k встановлюються в адміністративному порядку і, крім кількох винятків, є сумісними у всіх публікаціях. Статистики NASS несуть відповідальність за ідентифікацію першочергових приховувань і їх доповнень, а також за гарантування того, щоб всі моделі приховування є сумісними протягом певного часу. Приховування можуть представлятись окремо або як сукупності. PSM 12 дозволяє використання інформованої згоди (відмов) для програми Оціночних показників сільського господарства, якщо воно визнається таким, що представляє інтерес для промисловості. Всі сторони, що потрапляють під ризик, повинні дати згоду на те, щоб оціночні значення були опубліковані і мали право скасувати свою згоду. Угоди оновлюються кожні п'ять років.

Для Сільськогосподарського перепису Пуерто-Рико, подальших програм з перепису, включаючи Опитування щодо зрошення ферм і ранчо, а також Перепису водного господарства, NASS використовує правило p-відсотка для ідентифікації комірок з конфіденційними даними, що підлягають ризику розкриття. Правило граничного значення також застосовується до всіх порядкових даних для гарантування того, щоб мінімальне число ферм було представлено в кожній опублікованій комірці. Всі порядкові дані, що мають відношення до комірок із менш ніж трьома фермами, також приховуються. Додаткове приховування обирається, використовуючи методологію потоку в мережі. Дані підрахунку частотності не вважаються конфіденційними і не підлягають приховуванню. Крім того, NASS не дозволяє використання інформованої згоди від респондентів для Сільськогосподарського перепису і його подальших програм.

Оскільки політика NASS полягає в тому, щоб не розголошувати файли з мікроданими, NASS управляє Лабораторією даних у межах Вашингтонського центрального апарату. Окремі дослідники можуть подати пропозицію щодо дослідження і просити видати дозвіл на ведення спеціалізованих моделей і зведень у таблиці по відношенню до певних файлів з мікроданими у межах лабораторії. Запити адресуються та схвалюються чи не схвалюються Начальником управління в кожному окремому випадку. Робочий персонал NASS здійснює нагляд за лабораторією і всі матеріали, що залишають лабораторію, підлягають перегляду на предмет розкриття. Фізичні особи, що використовують лабораторію даних, підписують форми з конфіденційності, так як агенти NASS зобов'язані до виконання законодавчих актів, що обмежують незаконне використання і розкриття даних. NASS врегулює всі питання, що стосуються лабораторії даних у будь-якому з його 46 регіональних відділень, коли це потрібно. Користувачі даними можуть також подати запит щодо спеціальних зведень в таблиці через Лабораторію даних. Ці зведення в таблиці здійснюються робочим персоналом NASS і виключають потребу для доступу до файлів з мікроданими. Результати кожного зведення в таблиці вважаються суспільним надбанням і є доступними для будь-якого користувача даними.

NASS і Служба економічних досліджень спільно надають інтерактивні веб-додатки із вбудованим переглядом розкриття і фільтруванням, що дозволяє окремим дослідникам вести табличні зведення і спеціальний аналіз стосовно мікроданих із Опитування щодо управління сільськогосподарськими

ресурсами. Процедури доступу відображають ці Лабораторії даних. Окремі дослідники можуть подати пропозицію щодо дослідження і подати запит на Ідентифікаційний номер засвідченого доступу. Конфіденційність даних захищена застосуванням підходу, основаному на шумі, до базових даних перед тим як формуються табличні дані. Параметри, що використовуються для створення шуму, залишаються конфіденційними. «Правило р-відсотку» також застосовується до зведених показників для тестування комірки таблиці на домінування над окремою організацією.

NASS проводить певну кількість опитувань, що підлягають компенсації, для урядових та академічних організацій, і розробила спеціальні процедури конфіденційності для цих опитувань. В таких ситуаціях NASS чітко ідентифікує спонсорську організацію і ціль опитування для опитуваних перед збиранням їх добровільних відгуків. У цих випадках NASS може надати файл з мікроданими, позбавлений від ідентифікаторів, спонсорській організації для своїх аналізів. Файл з мікроданими повинен знаходитись на фізично захищеному сайті згідно із заходами безпеки, схваленими NASS. Всі фізичні особи, які матимуть доступ до файлу, повинні підписати фори з конфіденційності, так як агенти NASS зв'язані законодавчими актами, що обмежують незаконне використання і розкриття даних.

У лютому 1993, Юридичний департамент (OGC) Міністерства сільського господарства США переглянув закони і нормативні акти, що мають відношення до розкриття конфіденційних даних NASS. Таким чином, інтерпретація OGC законодавчих актів дозволяє ділитись даними з іншими агентствами, університетами, і приватними юридичними особами, якщо тільки вона посилює завдання USDA та укладає договір співпраці, договір про відшкодування витрат, контракт, меморандум про взаємопорозуміння. Такі юридичні чи фізичні особи, що отримують дані, також зобов'язані законодавчими актами, які обмежують незаконне використання і розкриття даних. Поточна політика NASS полягає в тому, що розподіл інформації для статистичних цілей буде мати місце на індивідуальній основі, якщо необхідно, для звернення до схваленого визначеного або для публічних потреб, або згідно із вищеописаними спеціалізованими ситуаціями.

За умови, що майбутні випадки використання даних відомі на момент збору даних, вони пояснюються респонденту, а також подається запит на дозвіл ділитися даними між різними користувачами. Запит на цей дозвіл подається в письмовій формі з бланком дозволу на публікацію, підписаного кожним респондентом.

A.2. Міністерство торгівлі

A.2.a. Бюро економічного аналізу (BEA)

Діяльність BEA з обмеження розкриття в основному відноситься до даних, які воно збирає по зарубіжних прямих інвестуваннях і торгівлі послугами. Ці дані збираються із комерційних підприємств США — які належать як США, так і іноземним власникам — в обов'язкових оглядах, що здійснюються на підставі Акту з опитування зарубіжного інвестування і торгівлі послугами (P.L. 94-472, в новій редакції). Право на огляди торгівлі фінансовими послугами також надається на підставі Вичерпного акту з торгівлі і конкуренції від 1988. Так як це вимагається Актом щодо опитування, зібрані дані залишаються конфіденційними і публікуються у такий спосіб, що запобігає ідентифікації окремих відповідей. Діяльність з обмеження розкриття також здійснюється для певних даних по регіональній економічній діяльності, які отримуються із Бюро трудової статистики. BLS здійснює діяльність із обмеження розкриття для своїх власних цілей і надає копію результатів для

BEA.

З огляду на дані щодо прямомго інвестування і торгівлі послугами, загальне правило для першочергового приховування містить розгляд даних для основного доповідача, другого оповідача і всіх інших доповідачів в даній комірці. Якщо дані для всіх доповідачів, крім двох основних, додаються не більше ніж до певного відсотка даних основного доповідача, то комірка представляє першочергове приховування. Це застосування правила р-відсотка.

Це правило захищає основного доповідача від другого доповідача, захищає другого доповідача від основного доповідача, і автоматично приховує інформацію у будь-якій комірці з одним або двома доповідачами. В поодиноких випадках респонденти можуть, за запитом BEA, надавати право відмови від конфіденційності.

При застосуванні загального правила, абсолютні значення використовуються якщо елемент даних може бути негативним (наприклад, чистий дохід). Якщо доповідач має більше ніж один запис даних у тій самій комірці, ці записи консолідуються а приховування здійснюється на рівні доповідача.

На додачу до застосування загального правила, можуть застосовуватись декілька спеціальних правил, що охоплюють округлені оціночні показники, зведені показники по округу і промисловості, і приховування «основного елементу» (розглядаючи набір відповідних елементів, як групу і приховуючи всі елементи, якщо основний елемент приховується).

Додаткове приховування здійснюється частково через комп'ютер, а частково через людське втручання. Комп'ютерні програми, що використовуються, містять рутинні операції, які вивчають різні комбінації комірок для гарантії, що приховування не можуть бути відкриті через вирахування лінійних комбінацій рядків і колонок.

Деякі таблиці публікуються за числами компаній, таких як число закордонних філій компаній США в різних країнах або промисловостях. Ці підрахунки числа не вважаються конфіденційними і не аналізуються на предмет розкриття, або не приховуються.

Згідно із Акту опитування зарубіжного інвестування і торгівлі послугами, обмежений обмін даними з іншими Федеральними агентствами, і з консультантами та підрядниками BEA дозволений, але лише для статистичних цілей і лише для того, щоб виконувати конкретні функції відповідно до Акту. До них включені також «Спеціальні співробітники, що принесли присягу», яким надано локальний доступ до мікроданих рівня компанії для цілей дослідження, і які присягаються підтримувати конфіденційність даних на таких самих підставах як і звичайні співробітники BEA. На певні типи обміну даними з іншими Федеральними агентствами також надається право Актом щодо прямих іноземних інвестицій і покращень міжнародних фінансових даних від 1990, та Актом щодо захисту конфіденційної інформації і статистичної ефективності від 2002. Цей обмін даними призначений лише для статистичних цілей, і від будь-якого робочого персоналу цих агентств, який повинен розглянути неприховані дані BEA у зв'язку з цією діяльністю, вимагається отримувати статус Спеціального співробітника BEA, що приніс присягу.

В іншій програмній зоні Регіональне відділення економічного вимірювання BEA опубліковує оціночні показники особистого доходу обмеженого району по основному джерелу, основаному на даних по заробітній платі робітників і службовців на рівні округу, які воно отримує від Федеральної/державної програми ES-202 Бюро трудової статистики (BLS). Від BEA вимагається дотримуватись правил обмеження статистичного розкриття, що відповідають вимогам BLS. Для запобігання або прямому, або непрямому розкриттю конфіденційної інформації, BEA використовує файл BLS з державного та окружного нерозголошення для захисту конфіденційної інформації в даних ES-202, що поставлялись для BEA. Файл з нерозголошення визначає конфіденційні комірки,

які необхідно захистити для уникнення розголошення конфіденційної інформації.

BEA використовує таку кількість комірок нерозголошення BLS як це можливо, але не може використовувати деякі з них з різних причин. Найбільш важливими причинами є те, що промислова і географічна структура, опублікована BEA не точно відповідають промисловим і географічним деталям, наданими BLS, і що BEA не використовує дані ES-202 для сільськогосподарського сектору. Для цих випадків, BEA повинен обирати додаткові комірки для запобігання розкриття конфіденційної інформації. Для того, щоб визначити які оціночні показники повинні бути прихованими, файл з сукупними заробітними платами робітників і службовців, а також файл нерозголошення заробітних плат робітників і службовців використовуються для підготовки багатовимірної матриці. Ця матриця випробовується, а оціночні показники, що повинні приховуватись, обираються. Додаткові приховування, якщо необхідно, генеруються комп'ютером і перевіряються для гарантії, що вони є належними.

A.2.б. Бюро перепису населення (BOC)

Бюро перепису населення проводить свої статистичні програми згідно із державним законодавством, таким як Закон про охорону прав особистості, Закон про свободу інформації (FOIA), і Закон про захист конфіденційної інформації і статистичну ефективність (CIPSEA) від 2002; і спеціальне законодавство для агентства, таке як Розділ 13, Кодекс Сполучених Штатів, від 1954.

Розділ 13 Кодексу Сполучених Штатів визначає основу для норм Бюро перепису населення для конфіденційності. Дані, що ідентифікують фізичних осіб, комерційну діяльність, а також інші організації, не повинні розділятись з будь-ким, за винятком якщо ця особа прийняла клятву підтримувати конфіденційність Перепису, і має ділову потребу в цих даних. Бюро перепису населення захищає конфіденційні дані через використання технологічних мір безпеки, захисту статистичних даних, і через обмежений доступ. Методи, що використовуються, включають програмне забезпечення для шифрування, спеціальні виділені лінії, так як і методи паролю і мережевий екран.

Бюро перепису населення має правомочність за законом проводити опитування для інших агентств відповідно до Розділу 13 або Розділу 15 Кодексу Сполучених Штатів. Агентство, що надає фінансування, із договором, що відшкодовується згідно із Розділом 13, може використовувати зразки та основи вибірки, розроблені для різноманітних опитувань і переписів Розділу 13. Це б зберегло спонсору додаткові видатки, які б могли накопичитись, якби він мусив розробити свою власну основу вибірки. Проте дані, розголошені агентству про те, що спонсори опитування, яке може бути відшкодоване, згідно із Розділом 13 підлягають положенням з конфіденційності будь-якого файлу з мікроданими публічного користування, або таблицями Бюро перепису населення; наприклад, Бюро перепису населення не буде розголошувати або ідентифіковані мікродані, або дані невеликої площі. Ситуація згідно із Розділом 15 є набагато іншою. При здійсненні опитувань згідно із Розділом 15, Бюро перепису населення може розголосити спонсорам інформацію, що піддається ідентифікації, так як і дані по малій площі. Проте інші джерела, крім опитувань і переписів, що охоплюються Розділом 13, повинні використовуватись для складання вибірок. Коли фінансуюча організація формує основу, дані збираються згідно із Розділом 15, і застосовуються правила конфіденційності фінансуючої організації.

Наглядова рада з розкриття (DRB) переглядає характеристики і пропозиції, що має відношення до

кожного розголошення даних згідно з Розділом 13, призначеного для публічного користування. DRB гарантує дотримання основних принципів «Контрольного списку DRB Бюро перепису населення» і будь-яких інших критеріїв, попередньо встановлених DRB. Вона сполучає політику обмеження розкриття з керівниками програми, посадовими особами Бюро перепису населення, користувачами даними, потенційними спонсорами і широкою публікою. DRB ініціює і координує дослідження з потенціалу розкриття в мікроданих, табличних даних та інших результатів статистичного обчислення; і з ефективності методів уникнення розкриття, які застосовуються до таких результатів. Члени Групи з дослідження уникнення розкриття у Відділі статистичних досліджень проводять дослідження із відповідних методів захисту даних для матеріалів, що публікуються.

Деякі механізми існують для надаття доступу до більш детальної інформації на обмежених основах. Сюди входять Дослідницькі центри збору даних для ухвалених дослідників із Спеціальним статусом особи, що прийняла присягу, так як і віддалений інтерактивний доступ в Державних центрах збору даних та Інформаційних центрах перепису через Розширену систему опитування для таблиць, що задаються користувачем, із Перепису від 2000. Вказана система дозволяє користувачам подавати запит на певні типи таблиць і потім автоматично переглядає таблиці для уникнення розкриття конфіденційної інформації. Користувачі отримують лише таблиці, які пройшли перегляд на предмет розкриття.

Деякі мікродані доступні для ухваленння науковцями у Дослідницьких центрах збору даних Бюро перепису населення (RDC). Метою Центру економічних досліджень (CES) і RDC є збільшити вигідність та якість результатів досліджень Бюро перепису населення. Використання мікроданих може звертатись до важливих питань з політики без потреби додаткових зборів інформації. Крім того, це найкращий засіб, за допомогою якого Бюро перепису населення може перевіряти якість даних, які воно збирає, редагує і зводить у таблиці. Ці безпечні науково-дослідні центри розташовані на різноманітних об’єктах по всій країні. Доступ строго обмежений для дослідників і робочого персоналу, уповноваженого Бюро перепису населення. Весь аналіз повинен здійснюватись у межах безпечного науково-дослідного центру RDC. Забезпечення безпеки на RDC має декілька аспектів: нагляд за проектом, фізично захищений комплекс, безпека персоналу, захищене обчислювальне середовище, співробітник Перепису населення в дослідному центрі, і застосування правил уникнення розкриття по відношенню до аналітичних результатів, представлених громадськості.

Для кожного економічного перепису, що здійснюється кожних п’ять років, і для пов’язаних з ним опитувань Бюро перепису населення використовує правило p% для ідентифікації конфіденційних комірок в таблицях, але не публікує значення p. Конфіденційні комірки приховуються, а додаткові приховування ідентифікуються використовуючи техніку потоку в мережі (який можна розглядати як особливий випадок лінійного програмування), що по відношенню до обчислення є дуже швидким, або лінійне програмування, яке є повільнішим. Потік в мережі є ідеальним для 2-вимірних таблиць. Його також було застосовано до 3D таблиць, хоча для таких таблиць лінійне програмування є більш прийнятним методом з теоретичної точки зору; тобто повний захист конфіденційних комірок повністю гарантований, що позбавляє від необхідності проводити програму аудиту розкриття для перевірки рівня досягнутого захисту.

Для Економічного перепису 2002, потік в мережі використовувався для всіх 2-вимірних таблиць і більших 3-вимірних таблиць. Програми приховування, основані на лінійному програмуванні, використовувались для менших 3-вимірних таблиць. Певні опитування мають 4-вимірні або 5-вимірні дані, а програми, основані на лінійному програмуванні, можуть використовуватись для цих таблиць, якщо час виконання завдання не є надмірним. Аудиторські програми використовуються за

необхідності.

Для демографічних даних, отриманих не шляхом перепису, Бюро перепису населення чином основному використовує комбінацію географічних, граничних, даних щодо населення та збільшення. Мікродані не можуть показувати географію нижче рівня населення в 100,000. Для більш детальних мікроданих, це граничне значення підвищується до 250,000 або вище. Деякі опитування зводять в таблиці лише на державній, регіональній або Переписній ділянці. Для результатів досліджень з даними, що виходять за межі основних публікацій, граничне значення може застосовуватись на рівні комірки або до населення. Багатовимірні табличні дані щодо конкретних публікацій повинні мати мінімальне число незважених випадків, зазвичай 50. Мінімальне граничне значення комірки, що найчастіше використовується це 3 незважені фізичні особи із 3 різних домашніх господарств. Збільшення використовується для уникнення використання граничних значень. Для маленького населення, або рідкісних характеристик шуму можуть додаватись до ідентифікаційних змінних, даними можуть обмінюватись, або умовне нарахування може застосовуватись до характеристики. Дані перепису, яким бракує компоненту захисту, що надається вибіркою, застосовує цільовий обмін на додачу до комбінації проекту таблиці і граничних значень, описаних вище.

Більшість поточних практик обмеження статистичного розкриття і досліджень підсумовуються у трьох наукових публікаціях Заяц (2002), Заяц, Массель, і Стіл (1999) Гавала, Заяц, і Роуланд (2004). Інші довідкові документи знаходяться в цих трьох публікаціях.

A.3. Міністерство освіти: Національний центр статистики освіти (NCES)

Національний центр статистики освіти (NCES) здійснює вагому законодавчу діяльність, що вимагає від установи захищати конфіденційність своїх зібраних даних. Спочатку, відповідно до Змін щодо покращення початкової і середньої школи Хоукінса-Стаффорда від 1988, і потім згідно з Національним актом щодо статистики освіти від 1994, від NCES вимагалось підтримувати конфіденційність всіх даних, що піддаються індивідуальній ідентифікації, про фізичні особи (наприклад, дані про ректора, вчителя чи студента). Хоч закон чітко не захищав інституційних даних, захист даних про фізичні особи у межах інституцій часто призводив також до захисту даних про освітні заклади. Закон про реформу в педагогіці від 2002 чітко вимагає від NCES захищати конфіденційність даних, що піддаються індивідуальній ідентифікації, про студентів, їх сім'ї і школи. По відношенню до цих законів, NCES має статистичну норму щодо дотримання конфіденційності (Статистичний стандарт NCES 4-2 http://nces.ed.gov/statprog/2002/std4_2.asp). Ця норма підсумовує відповідні закони, встановлює зобов'язання працівника і підрядника при обробленні конфіденційних даних, описує альтернативні методи, які можуть використовуватись для захисту даних NCES від розкриття, і включає в себе повідомлення із згодою, яке слід помістити у файли з даними публічного користування NCES. Крім того, Наглядова рада з розкриття (DRB) переглядає плани з аналізу і розкриття, та запропоновані розголошення даних публічного користування для захисту конфіденційності окремих опублікованих значень.

Більшість зборів даних NCES включають деякі інституційні дані, але додатково містять дані з будь-якої комбінації голів, вчителів, бібліотекарів, студентів навчального закладу, або їхніх батьків. Необхідно захищати саме дані фізичних осіб. Ці бази даних можна зробити публічно доступними або через файл публічного користування, або через систему аналізу даних (DAS) після застосування аналізу розкриття, ухваленого DRB, і вирішення будь-яких наявних ризиків розкриття. Цей процес описано нижче.

Файл публічного користування – це файл або серія зв'язаних файлів, які: 1) містять відповіді

фізичних осіб про себе, і 2) пройшли через аналіз на предмет розкриття, схвалений DRB. Вся безпосередня інформація, що піддається індивідуальній ідентифікації (наприклад, назва школи, ім'я фізичної особи, адреси) видаляється із файлу публічного користування. Безперервні змінні підлягають верхньому і нижньому кодуванню для захисту від ідентифікації значень, що різко відхиляються. Після здійснення цього, єдиний спосіб в який зловмисник, що хоче отримати дані, може ідентифікувати окремого респондента, це спочатку ідентифікувавши відібраний навчальний заклад про наявність фізичної особи.

Щоб запобігти ідентифікацію відібраного закладу, збираються всі відомі і публічно доступні переліки освітніх закладів, які містять назви та адреси. Кожен перелік пристосовується до вибраного файлу-зразка використовуючи всі загальні змінні між двома файлами. Якщо заклад можна ідентифікувати з точністю до 2 інших закладів використовуючи відповідну міру відстані, тоді є ризик розкриття, і його потрібно вирішити перед розголошенням даних.

Якщо отримують багато ризиків розкриття, то загальна змінна(ні) можуть приховуватись у файлі публічного користування, або змінна(ні) може збільшитись. Якщо існує лише декілька ідентифікованих ризиків розкриття, то належною дією буде вибірково збурити набір загальних змінних, поки всі ризики розкриття не будуть розв'язані. Цей аналіз повторюється послідовно для кожного файлу з переліком, поки він не повториться для кожного файлу з переліком без ідентифікації будь-яких ризиків розкриття.

Аналіз відповідності, описаний вище, призначений для запобігання, щоб випадковий «шукач» даних не зміг визначити учасників опитування. Умовно, якщо заклад не може бути ідентифікований, то фізичні особи у межах цього закладу також не можуть бути ідентифікованими. Проте зловмисники, що намагаються заволодіти даними, і з детальними знаннями заклад, зможуть ідентифікувати цю установу, у такий спосіб збільшуючи ймовірність ідентифікації фізичної особи. Для зменшення ймовірності правильного виконання цієї процедури, необхідна додаткова дія розкриття.

У всіх випадках, коли дані про керівника закладу, вчителя, студента, або батьків збираються в одну групу, вимагається взяти підвибірки з респондентів. Дані від респондентів, що обираються в цій підвибірці, переглядаються з використанням додаткового редагування розкриття. Редагування це або: 1) метод незаповнених полів та умовне нарахування, або обмін даними вибірки зібраних конфіденційних елементів; або 2) обмін даними основної ідентифікаційної змінної респондента чи закладу. Кількість редагування встановлюється на рівні, достатньо високому для захисту конфіденційності респондента, в той самий час не поступаючись аналітичною корисністю файлу з даними.

Важливий аспект цього редагування полягає в тому, що всі респонденти мають вибір. Зазвичай респондентам із вищим ступенем ризику надається більша можливість вибору. Якщо ж хтось все-таки подумає, що вони ідентифікували респондента, вони не можуть бути впевненими, що ці дані справді відповідають цьому респонденту.

Інший спосіб, в який NCES розповсюджує дані, це через Систему уникнення розкриття (DAS). DAS це програма генерування таблиць, яка може генерувати пропорції, середні величини або коефіцієнти кореляції з відповідними стандартними помилками, які були обчислені беручи до уваги процедури комплексних вибірок, що використовуються в опитуваннях NCES. DAS приєднаний до файлу з даними, але всі елементи даних маскуються для того, щоб сам файл був непридатним для читання для будь-чого або будь-кого крім програми для генерування таблиць. Дані також захищаються через процес здійснення вибірки для огляду (тобто, будь-яка обрана одиниця ймовірно буде мати багато інших подібних одиниць в генеральній сукупності). Проте, існує мало контролю над типом і числом генерованих таблиць, подальший захист від розкриття застосовується через збурення даних (наприклад, обмін даними) і збільшення даних.

Для того, щоб DAS було розголошено, файл з ключовими даними повинен містити серію редагувань конфіденційності DRB: або метод незаповнених полів чи умовно нарахування, або обмін даними щодо здійснення вибірки зібраних конфіденційних елементів; або обмін даними основної ідентифікаційної змінної респондента або закладу.

Всі таблиці NCES використовують або метод збурень (тобто підхід редагування конфіденційності), або процес стягування комірок, поки всі комірки не будуть містити значення, що мають відношення до хоча б трьох респондентів. Підхід редагування конфіденційності застосовується до файлів з мікроданими обмеженого користування. Таблицю можна підготувати без застосування жодного додаткового методу обмеження розкриття.

A.4. Міністерство енергетики: Управління з інформації в області енергетики (EIA)

EIA встановило статистичні норми (http://www.eia.doe.gov/smg/Standard.pdf), включаючи норми для захисту даних, доступності і нерозголошення. Стандарт 2002-22, «Нерозголошення даних, що ідентифікують компанію, в сукупних комірках», містить процедури і політику для гарантування того, щоб значення комірок з конфіденційними даними бути прихованими (тобто, утримані від публічного розголошення) для захисту даних конфіденційного опитування. EIA також вимагає додаткового навчання з питань конфіденційності для тих, хто має доступ до даних, захищених згідно з CIPSEA.

Початковий метод EIA для гарантії захисту конфіденційності це застосування правила комбінації. Незважаючи на обрані параметри, правило гарантує, що комірки з даними ненульового значення повинні базуватись на трьох чи більше респондентів. Правило комбінації – це правило pq в поєднанні з деякими іншим лінійним правилом лінійного приховування. Значення параметру чутливості pq представляє максимально допустиме здобуття інформації, коли одна компанія використовує підсумкове значення опублікованої комірки і своє власне значення для створення кращих оціночних показників значень його конкурента. Значення параметр pq, які обираються для конкретних опитувань, не публікуються і вважаються конфіденційними. Додаткове приховування застосовується до інших комірок для гарантії, щоб конфіденційне значення не було реконструйовано з опублікованих даних. Для інформації, що збирається відповідно до завірення про конфіденційність, EIA публічно не розголошує імен або інших ідентифікаторів учасників опитування, приєднаних до їх поданих даних.

Для багатьох опитувань EIA, що використовують правил pq, додаткові приховування обираються вручну. Одна система опитування, що публікує складні таблиці цін та об'ємів для сирої нафти і продукції з очищеної нафти використовує програмне забезпечення для обрання додаткових приховувань. Вона гарантує, що в наявності є принаймні дві приховані комірки в кожному напрямку, комірки нульового значення виключаються як варіанти для приховання, і що обрані комірки представляють менший інтерес для користувачів даними.

Норма 2002-22 також містить окремі додаткові матеріали із керівними вказівками для розуміння і застосування правила pq. Керівні вказівки включаються для ситуацій, коли всі значення є негативними; деякі дані є приписані; опубліковані значення є істинними значеннями (різниця між позитивними числами); і опубліковані значення є середньозваженими величинами (такі як середньозважені ціни). Багато такої ж інформації надається в Додатку А в цьому звіті.

В обраних програмних зонах EIA не використовує методи обмеження розкриття по відношенню до

статистичних даних. Для певних даних з енергопостачання, число компаній, що надають інформацію є відносно малим і/або розподіл компаній з енергопостачання є дуже викривлений з відносно малою кількістю великих компаній. Статистичні дані для географічних підзон Сполучених Штатів (наприклад, Штати, Нафтове управління по оборонних районах, Нафтопереробні райони) типово включають деякі чутливі значення, які б не були опубліковані, якщо б застосовувались методи обмеження розкриття. Якщо застосовувались методи обмеження розкриття, що використовують початкове і додаткове приховування, результатом буде значна кількість втрат інформації. Ця втрата інформації для користувачів даними серйозно підірве вартість даних для публічної і приватної домовленості та аналізу енергопостачання.

В цих програмних зонах EIA використовує повідомлення Федерального реєстру для оголошення запропонованої політики щодо відмови від використання методів обмеження розкриття і вимагає відкритих коментарів. Після розгляду відкритих коментарів EIA вирішує, чи робити свою політику офіційною. Якщо політика не має наміру використовувати такі методи, EIA пояснює політику в той час, коли збір інформації проходить процес ухвалення Адміністративно-бюджетним управлінням і коли матеріали опитування надаються потенційним опитуваним в той час, коли подається запит на інформацію. Пояснення стверджує, що процедури обмеження розкриття не застосовуються до статистичних даних, опублікованих з інформації цього опитування. Пояснення переходить до твердження, що в наявності може бути деяка сформована статистика, основана на даних від менш ніж трьох респондентів, або в якій переважають дані від одного або двох значних респондентів. В цих випадках для обізнаної особи може бути можливим оцінити інформацію, що доповідається конкретним респондентом.

EIA не має норми для таблиці адрес частотних даних. Однак, існує лише дві первинні публікації частотних даних в таблицях EIA. Ці публікації представлені публікацією Характеристик домашнього господарства Опитування щодо споживання електроенергії в комунальному секторі (RECS) і публікації Характеристик будівлі Опитування щодо споживання електроенергії адміністративними будівлями (CBECS). В обох публікаціях, комірки приховуються з причин точності, а не з причин ймовірності розкриття. Для першої публікації, значення комірок приховуються якщо в наявності є менше ніж 10 респондентів або ж Відносні середньоквадратичні помилки (RSE) складають 50 відсотків або більше. Щодо другої публікації, значення комірок приховуються якщо в наявності є менше ніж 20 респондентів або ж RSE' складають 50 відсотків або більше. Не використовується жодного додаткового приховування.

EIA не має норми для методик обмеження статистичного розкриття для файлів з мікроданими. Єдині файли з мікроданими для конфіденційних даних, розголошені EIA є для RECS і CBECS. У цих файлах, різноманітні стандартні процедури обмеження статистичного розкриття використовуються для захисту конфіденційності даних для окремих домашніх господарств і будівель. Ці процедури включають в себе: усунення ідентифікаторів, обмеження географічних деталей, упущення або стягування елементів даних, верхнє кодування, нижнє кодування, інтервальне кодування, округлення, заміна чисел середньозваженого значення (викривлення), і запровадження шуму через метод коригування даних, який довільно коригує дані рівня респондента у межах контрольованого рівня максимального відсотка наближено до фактичного опублікованого оціночного показника. Після застосування методу коригування із внесеним елементом випадковості до даних, середні значення для широких груп населення, основаних на скоригованих даних, є такими ж як середні значення, генеровані із даних без коригування.

A.5. Міністерство охорони здоров'я і соціальних служб

A.5.a. Агентство досліджень і оцінки якості медичного обслуговування (AHRQ)

Процедури обмеження розкриття, що використовуються AHRQ подібні до процедур NCHS. Огляд ради з медичних видатків (MEPS), проведений AHRQ використовує Національне анкетування з питань здоров'я в якості своєї основи вибірки. Таким чином, процедури обмеження розкриття, що використовуються AHRQ для файлів з даними публічного користування MEPS, відповідають процедурам, що застосовуються NCHS для MEPS. Від всіх розголошень файлів з даними публічного користування вимагається, щоб вони були переглянуті та ухвалені Наглядовою радою з розкриття NCHS перед тим, як вони будуть розголошені. AHRQ також переглядає і виправляє розголошення файлів публічного користування із NHIS.

AHRQ заснував локальний центр збору даних у межах Центру фінансування, доступу і тенденцій цін (CFACT) для полегшення дослідникам доступу до обраних даних MEPS внутрішнього користування.

Центр збору даних CFACT це фізичний простір на AHRQ, розташований в Роквіллі, штат Меріленд де дослідникам із ухваленими проектами надається доступ до файлів з даними, які недоступні для широкого розповсюдження. Ці дані класифіковані як «закриті» і містять інформацію, яка не розголошується громадськості. Ці набори даних можуть містити географічні змінні на нижчому рівні, ніж ті, які випущені для публічного користування, більш детальну інформацію про умови, може складатися не із відредагованих сегментів бази даних, які ще підготовлені для публічного розголошення. Ці набори обмежених даних не містять інформації, що прямо ідентифікує респондента (ім'я, номер соціального забезпечення, поштова адреса).

Дослідникам надано доступ лише до інформації, необхідної для завершення проекту. Жоден дослідник не може видаляти будь-які матеріали із Центру збору даних, поки ці матеріали не були переглянуті спеціальним персоналом CFACT для уникнення розкриття. Лише результати зведення (таблиці, рівняння) можуть усуватись із Центру збору даних. Жодні файли з даними не дозволяється видаляти із Центру збору даних.

Всі матеріали, що мають видалятись із центру збору даних, підлягають перегляду на предмет розкриття. Персонал CFACT несе відповідальність за гарантію конфіденційності даних, що використовуються в центрі збору даних. У випадку користувачів, що знаходяться на об'єкті, персонал CFACT переглядає результати досліджень, або таблиці перед тим, як матеріал залишить Центр збору даних. У випадку дослідників, що використовують Центр збору даних дистанційно, персонал CFACT проведе перегляд матеріалів на предмет розкриття перед відправкою результатів обчислень досліднику. Розробка формальних критеріїв для перегляду табличних матеріалів це постійний процес.

Для користувачів, Керівник Центру збору даних CFACT це контактна особа для вирішення конфліктів щодо перегляду на предмет конфіденційності. Кожна спроба буде здійснюватись для того,

щоб працювати із дослідником для розробки специфікацій для таблиць, які «пройдуть» перегляд на предмет конфіденційності. Проекти із питаннями конфіденційності, що все ще актуальні, будуть обговорюватись із провідними спеціалістами CFACT перед винесенням остаточного рішення.

Будь-які вихідні дані, що можуть потенційно ідентифікувати (або прямо, або шляхом логічного виведення) респондентів або малі географічні зони, не можуть бути видалені із центру збору даних. Таблиці з географічними зонами, як одне із табличних зведень (крім тих, що ідентифікуються на файлах публічного користування), не можна видаляти, так як і таблиці, що містять комірки з менш ніж 100 спостереженнями. Користувачам Центру збору даних ніколи не дається доступ до файлів із прямими ідентифікаторами, такими як ім'я чи адреса. Користувачам може надаватись доступ до файлів із фіктивними кодами для позицій. Однак, оскільки центри збору даних не мають потреби розпізнавати ідентичність позицій, їм не буде надаватись код, який зробить можливою асоціацію географічної назви з кодом. За запитом цілий файл може бути попередньо закодований в категорії (тобто, проживання в штаті з високим/середнім/низьким рівнем забезпечення Медікейд (американська державна програма медичної допомоги нужденним). Моделі, що використовують географічну зону в якості залежної змінної, не можуть бути видалені із Центру збору даних. Відмінні риси одиниць вибірки, які б могли допомогти в ідентифікації суб'єкту даних, не можуть бути видалені. В загальному, будь-які прямі тотожності або такі, що логічно виводяться, і які не були розголошені по файлах з даними публічного користування, не можуть бути видалені із Центру збору даних.

А.5.б. Національний центр медичної статистики (NCHS)

NCHS це головне федеральне агентство, що розголошує медичну статистику. Вона є частиною Центрів з контролю і профілактики захворювань США (CDC) від Міністерства охорони здоров'я і соціальних служб США. Методи обмеження статистичного розкриття NCHS для CDC представлені в Посібнику з конфіденційності для персоналу NCHS (вересень 2004), Секція 9 «Уникнення ненавмисного розкриття змісту інформації через розголошення мікроданих» і Секція 10 «Уникнення ненавмисного розкриття інформації в табличних даних». Жодні числа порядкових даних не повинні бути основані на менш ніж п'яти випадках, а також використовується правило (n, k). Даючи характеристику більш ранньому виданню Посібника NCHS, Джабін (1993б) стверджує, що «керівні вказівки надають можливість аналітикам брати до уваги конфіденційність і зовнішня доступність даних для публікації, так як і наслідки помилок при неотриманні даних і помилки у відповіді, а також малі частки вибірки для того, щоб зробити ідентифікацію фізичних осіб більш складною». У майже всіх звітах про дослідження не показується жодних географічних даних низького рівня, що значним чином знижує шанс ненавмисного розкриття. Персонал NCHS заявляє, що для таблиць частотних а) «у жодній таблиці не повинні знаходитись всі випадки будь-якого рядка чи колонки не повинні в жодному випадку знаходитись в одній комірці тієї ж таблиці»; і б) «загальне число для рядка чи колонки в комбінаційній таблиці у жодному випадку не повинна бути меншою 5». Один прийнятний спосіб для розв'язання проблеми (або для таблиць частотних даних, або для таблиць порядкових даних) полягає в комбінуванні рядків і колонок, або ж у використанні приховування комірки (плюс додаткове приховування). Інші підходи знаходяться в розробці.

Політикою NCHS є зробити файли з мікроданими доступними для наукової спільноти для того, щоб могли здійснюватись додаткові аналізи на благо країни. Такі файли переглядаються для ухвалення Наглядовою радою NCHS з розкриття відповідно до керівництва і принципів, що містяться у Посібнику для персоналу і Контрольному переліку для Розголошення файлів з мікроданими. Ці керівні вказівки вимагають, щоб детальна інформація, яку можна використати для ідентифікації

фізичних осіб (наприклад, дата народження), не повинна включатись у файли з мікроданими. Тотожність географічних місць і характеристик районів з меншою кількістю населення ніж 100,000 осіб ніколи не зможуть бути ідентифіковані, і може бути необхідним встановити цей мінімум на вище значення, якщо на це вказує дослідження, або з інших міркувань. Інформація по відбору зразка, який може ідентифікувати суб'єктів даних, не повинна включатись.

Всі нові набори мікроданих повинні переглядатись на предмет конфіденційності та ухвалені для розголошення Посадовою особою з конфіденційності NCHS, яка консультується з Наглядовою радою з розкриття NCHS при прийманні рішень в агентстві.

Після успішного подання в Дослідний центр збору інформації NCHS, дослідникам можуть надати доступ спеціальних файлів, що не дозволяють ідентифікації окремих респондентів. Це може відбуватись на об'єкті, в офісах NCHS, або дистанційно через безпечні електронні лінії. В той час як до інформації стосовно названих географічних суб'єктів немає доступу, дані, що замовляються за такими модулями, можуть аналізуватись на такому рівні, який є неможливим із даними публічного користування.

Потенційні дослідники повинні подати пропозицію щодо дослідження, що переглядається та ухвалюється комітетом, судження якого основані на наявності ресурсів RDC, сумісних із завданням NCHS, загальною науковою доцільністю, та обґрунтованістю проекту. Хоча дослідники підписують договори конфіденційності, строгі протоколи конфіденційності вимагають, щоб дослідники з ухваленими проектами завершували свою роботу використовуючи матеріальну базу, що розташована в межах RDC. Дослідники можуть поставляти свої власні дані для об'єднання з наборами даних NCHS. Заповнені персоналом RDC, об'єднані файли доступні тільки для дослідників, які їх зібрали, за винятком якщо надається письмовий дозвіл для забезпечення доступу іншим. Додаткові відомості про Дослідницький центр збору даних NCHS доступні за посиланням http://www.cdc.gov/nchs/r&d/rdc.htm.

Області, що підпадають під поточний розгляд включають в себе програмне забезпечення для балансування якості даних та обмеження статистичного розкриття (SDL) в табличних даних і вдосконалені процедури для SDL, а також оцінювання ризику розкриття в мікроданих.

A.6. Міністерство юстиції: Бюро юридичної статистики (BJS)

Таких вимог, згідно із Розділом 13 Кодексу Сполучених Штатів, що охоплюють Бюро перепису населення, дотримується BJS для тих даних, що збираються BJS через Бюро перепису населення. Що стосується табличних даних, комірки з менш ніж 10 спостереженнями не відображаються в опублікованих таблицях. Опубліковані таблиці можуть в подальшому обмежувати спроможність ідентифікації представляючи кількісно вимірювані класифікаційні змінні (такі як вік і роки навчання) в консолідованих обласних значень. Введені дані в комірці і граничні дані також можуть бути обмежені до ставок, процентних відношень і зважених розрахунків. Стандарти для захисту

мікроданих включено у склад правозастосовного нормативного акту BJS. Окремі ідентифікатори регулярно позбавляються від всіх файлів з мікроданими перед тим, як вони розголошуються для публічного користування.

A.7. Міністерство праці: Бюро трудової статистики (BLS)

Наказ члена комісії 3-04 «Конфіденційний характер записів BLS» від 4 Жовтня 2004 року, містить політику BLS щодо конфіденційних даних, які він збирає. Одна з вимог полягає в тому, що:

> «Публікації повинні готуватись в такий спосіб, щоб не розкривати особу будь-якого конкретного респондента і, як це відомо виконавцю документу, не дозволить щоб інформація стосовно респондента була логічно виведена прямими чи опосередкованими засобами».

Подальше положення дозволяє винятки згідно із умовами інформованої згоди, і вимагає попереднього уповноваження Члена комісії перед тим, як використовується положення про таку інформовану згоду.

Статистичні методи, що використовуються для обмеження розкриття різняться за програмою. Що стосується таблиць то процедура, яка найчастіше використовується, має два кроки – правило граничної величини, за якою слідує правило концентрації. Програми BLS використовують правило відсотка p або правило (n, k) для оцінювання концентрації в залежності від програми. Значення параметрів, що використовуються для граничних величин і різноманітні правила концентрації, що використовуються BLS не розголошуються громадськості. Поточна практика в BLS полягає в заміні користування правила концентрації (n, k) на правило відсотка p.

Наприклад, Щоквартальний перепис зайнятості і заробітних плат (QCEW), перепис щомісячної зайнятості та щоквартальної інформації про заробітну плату із облікових документів Допомоги по безробіттю, використовує правило граничної величини і правило відсотка p для даних календарного року (CY) 2002 і за його межами. До CY 2002, QCEW використовувало правило граничної величини і правило концентрації (n, k). В деяких випадках використовується правило двох кроків - правило (n, k) для окремої організації супроводжується правилом (n, k) для двох організацій. Огляд професійних травм і захворювань використовує правило граничної величини і правило відсотка p для даних CY 2003, замінюючи правило граничної величини, що використовується в поєднанні із «правилом концентрації (n, k)».

Національне опитування щодо компенсації використовує підхід, що об’єднує два правила граничної величини і «правило (n, k)». Правила граничної величини вимагають, щоб кожне оціночне значення складалось із установ хоча б від кількості компаній m (незважених), і що в наявності є кількість чітких професійних відборів t (незважених). Воно також використовує «правило концентрації (n, k)», яке вимагає, щоб зважена зайнятість поміж всіх організацій, що роблять внесок в оціночне значення, і які є частиною компаній n не можуть перевищувати відсоток k від зваженої зайнятості всіх організацій, що роблять внесок до оціночного значення.

Програма з індексу споживчих цін використовує комбінацію правила граничної величини і мінімальне число довідкових цін із відмінних одиниць вибірки. Індекс цін виробників промислової

продукції використовує правило граничної величини на елементах і довідкових цінах в поєднанні з «правилом (n, k)».

BLS розголошує лише декілька файлів з мікроданими публічного користування. Більшість з цих файлів з мікроданими містять дані, що збираються Бюро перепису населення відповідно до міжвідомчої угоди і повноваження Бюро перепису населення згідно з Розділом 13. Для інших оглядів (Поточний огляд населення, Огляд споживчих витрат, і чотири з п'яти оглядів із серії Національних поздовжніх досліджень) Бюро перепису населення визначає процедури обмеження статистичного розкриття, що використовуються. Методи обмеження розкриття, що використовуються для файлів з мікроданими публічного користування, які містять дані з Національного поздовжнього дослідження молоді, що збираються згідно з контрактом між Університетом штату Огайо і Національним центром вивчення громадської думки в Чиказькому університеті, подібні до тих, що використовуються Бюро перепису населення.

Бюро статистики праці (BLS) має в наявності сприятливі можливості на обмежених основах для дослідників із коледжів та університетів, урядових, та уповноважених некомерційних організацій для отримання доступу до конфіденційних файлів даних BLS винятково для статистичних цілей. Ці файли даних виводяться з опитувань BLS та адміністративних баз даних, для яких недоступна жодна версія публічного користування. Ці конфіденційні дані BLS доступні для досліджень, які є винятково статистичними, з відповідними обмеженнями для захисту даних від несанкціонованого розкриття. Файли конфіденційних даних BLS доступні лише для користування в Національному офісі BLS у Вашингтоні, округ Колумбія, по проектах статистичного дослідження, ухвалених BLS. Дослідники, яким надали доступ до конфіденційних даних, підписують угоди, які стверджують що вони несуть відповідальність за дотримання до політики конфіденційності BLS.

BLS розглядає заяви з пропозиціями щодо тематики досліджень чотири рази на рік. Пропозиції щодо тематики досліджень повинні мати обсяг від 5 до 10 сторінок і містити детальну інформацію про дослідницький проект, включаючи огляд літератури і вказівка на те, як запропоноване дослідження робить внесок в наукову літературу, а також вказує на гіпотези, які слід перевірити, набір даних і змінні, які слід використовувати в аналізі, емпіричні методи для застосування, і конкретні вихідні дані, які будуть результатом проекту.

A.8. Міністерство транспорту: Бюро транспортної статистики (BTS)

Бюро транспортної статистики (BTS) збирає дані, що стосуються транспорту. Законодавчі акти BTS з конфіденційності і ряд всеохоплюючих процедур конфіденційності захищають ці дані. *Процедурний посібник з конфіденційності* BTS документує процедури конфіденційності для агентства.

Посадова особа з конфіденційності (CO) BTS несе відповідальність за повсякденні операції програми конфіденційності. CO є також головою наглядової Ради з розкриття (DRB) від BTS, яка несе відповідальність за перегляд мікроданих, табличних даних та іншої інформаційної продукції для ризиків розкриття перед публічним розголошенням. Від персоналу і підрядників BTS вимагається проходити щорічне навчання з конфіденційності, і підписувати угоди про нерозголошення, коли вони

поступають або залишають службу у BTS.

Цілі програми конфіденційності BTS направляють процес перегляду даних на те, чи методи обмеження розкриття повинні застосовуватись. Ці цілі добиваються наступного:

- Захищати конфіденційні дані, і в той же час збільшувати доступ до даних,
- Застосовувати методи обмеження статистичного розкриття (SDL) на індивідуальній основі,
- Брати до уваги погляди користувачів даними щодо застосування методів SDL.

Для більшості мікроданих і результатів досліджень у табличних даних, від керівників програми BTS вимагається завершити контрольний перелік, що ідентифікує потенційні ризики розкриття, та окреслити будь-які кроки, що приймаються для зниження такого ризику. DRB від BTS переглядає результати дослідження даних і контрольний перелік, а також виносить остаточне рішення щодо ризику розкриття. DRB може рекомендувати застосування методів SDL перед широким розповсюдженням.

BTS використовує різноманітні методи мікроданих SDL, основаних на отриманих відомостях з перегляду розкриття та характерних особливостях файлів даних. Деякі процедури SDL, що використовуються, включають в себе приховування і модифікацію даних. Модифікація даних містить перекодування безперервних змінних у категоріальні змінні, стягування категорій, верхнє і нижнє кодування, введення шуму та обмін даними. Керівники програми BTS повинні також ідентифікувати будь-які зовнішні дані, які можна прирівняти до наборів даних BTS і мірив живати заходів щодо мінімізації можливостей для співставлення.

DRB проводить перегляд щодо розкриття результатів досліджень в табличних даних, коли вони розробляються на основі файлів мікроданих, які не розголошуються громадськості. BTS також використовує методи табличних даних SDL, основані на отриманих відомостях з перегляду розкриття і на характеристиках таблиць.

A.9. Міністерство фінансів: Служба внутрішніх зборів, Відділ статистики доходів (IRS, SOI)

Функція Статистики доходів (SOI) у межах більшої організації – Дослідження, Аналіз і Статистика (RAS), полягає у встановленні і введенні в дію керівних правил IRS для публічного розголошення податкових даних в таблицях і файлах мікроданих публічного користування. Ця роль головним чином необхідна у зв’язку із секціями 6108(в) і 6103j(4) Податкового кодексу (IRC), які вимагають, щоб дані у статистичних публікаціях, що створюються IRS та агентствами уповноважених одержувачів, були анонімними.

Адміністративні правила знаходяться у Розділі VI Посібника користувача відділу SOI (Січень, 1985), і вимагають, щоб на державному рівні і вище нього кожна комірка у публічно розголошуваних зведеннях у таблиці були основані хоча б на трьох спостереженнях. Комірки з даними, що не відповідають цим граничним величинам, приховуються і комбінуються з іншими комірками. Комбіновані і видалені дані включаються у відповідні суми для стовпців. Ці правила також застосовуються для другорядного розкриття, в якому характерні риси платника податків можуть

розголошуватись відніманням об'єднаних комірок у межах таблиці або між таблицями, і також непрямо через схожі дані в інших публікаціях.

Процедури розкриття документів SOI знаходяться у своїх власних публікаціях. Наприклад, обмеження розкриття обговорюються в «Обмеженість даних і методологія відбору зразків SOI» в Додатку до щоквартальних Бюлетенів SOI і в інтерактивному режимі за посиланням http://www.irs.gov/taxstats.

SOI створює один щорічний файл мікроданих публічного користування, відомий як «податкова модель» SOI, що містить зразок даних, основаних на Формі серії 1040 окремих податкових декларацій. Процедури захисту від розкриття, що застосовуються до цього файлу, включають: (1) записи вірогідності повторної вибірки за ставкою 33%; (2) усунення певних записів, що мають екстремальні значення; (3) приховування певних полів від всіх записів і географічних полів із записами високого доходу; (4) верхнє кодування і модифікація деяких полів; (5) викривлення деяких полів із записами високого доходу локальним виведенням середніх значень по всіх записах; і (6) округлення полів суми до чотирьох значущих цифр. Для гарантії, що приватність платника податків захищена у файлі податкової моделі SOI, SOI періодично укладала договір із спеціалістами, які застосовують так звані методи «професійного зловмисника», щоб засвідчити захист конфіденційності, і щоб інформувати про методи, які слід застосовувати у майбутніх розголошеннях файлу з моделлю податку SOI. Для додаткових подробиць щодо методів уникнення розкриття, що використовуються для формування файлів публічного користування SOI див.: Сейлер П., Вебер М. і Вонг В., (2001);

Крім своєї ролі у формуванні податкової статистики, SOI також несе відповідальність за координування забезпечення податкових даних для статистичних цілей для уповноважених одержувачів згідно з параграфом 6103j в IRC. Ця функція гарантує, що уповноважені одержувачі податкових даних також дотримувались правил 3/10, описаних вище, або еквівалентної методології, ухваленої SOI, як це обумовлено в Публікації IRS 1075, *Рекомендації по безпеці податкової інформації для федеральних, державних і місцевих агентствах (Червень 2000).* Через значний тягар відповідальності, ця вимога може поширюватись як на SOI, так і на агентства, що використовують альтернативні методології захисту від розкриття. Нещодавно розпочались спроби встановити міжвідомчі договори з досвідченими користувачами, такими як Бюро перепису населення США, в якому відповідальність за альтернативні методології табличного захисту приймається агентством-одержувачем. Договір перепису для цих цілей був введений в дію 2 червня 2003. Через те, що завдання із захисту файлів мікроданих публічного користування вважаються унікальними, і такі дані вважаються більш чутливими до ризику розкриття, файли мікроданих публічного користування виключаються. Тобто, згідно з цими договорами, схвалення IRS все ще буде потрібне перед тим, як стороннє агентство зможе випустити файл мікроданих публічного користування, оснований на податкових даних.

На даний час, Управління з досліджень IRS у межах RAS працює з Бюро перепису населення для гарантії, щоб всі наді в запропонованому Файлі публічного користування в Бюро перепису, основані на податкових даних [прибутках] і приєднані до Огляду доходу та участі в програмі (SIPP), що проводяться Бюро перепису, були анонімними. Запропонована методологія з файлу публічного користування SIPP/прибутку проводить дослідження використовуючи «синтетичні дані» для створення файлів публічного користування, спеціально пристосовані для певних користувачів, на відміну від підходу «без врахування індивідуальних особливостей».

A.10. Національний науковий фонд (NSF)

Національний науковий фонд (NSF), Відділ статистики наукових ресурсів (SRS), задовольняє вимогу оберігати конфіденційність респондентів на противагу бажанню дослідницької спільноти отримувати доступ до зібраних даних з використанням коштів платників податків. NSF застосовує або «правило домінування (n, k)» або «правило p-відсотка», або іноді обидва правила у взаємодії одне з одним в залежності від опитування. Коли можливо створити файл мікроданих, який є корисним для широкої групи дослідників під час захисту конфіденційності респондента, SRS розголошує файли даних публічного користування, які є сумісними з іншими цілями. Розголошуючи файли мікроданих публічного користування, окремі ідентифікатори видаляються зі всіх записів, а інші змінні високого ризику, що містять характерні особливості, вдосконалюються для запобігання ідентифікації учасників опитування та їх відповідей. Верхні коди і нижні коди застосовуються до числових полів для уникнення демонстрації граничних значень полів на записі даних. Значення поза межами верхнього коду і нижнього коду замінюються або середньою величиною значень, що перевищують відповідний верхній код або нижній код, або через застосування різноманітних методологій умовного нарахування.

Коли дослідник демонструє, що доступні файли даних публічного користування SRS не відповідають дослідницьким потребам і дотримуючись місії SRS щодо допомоги з надання статистичної інформації про науково-технічне підприємство США, часом можливо відповісти на запит надавши доступ до файлів з закритими даними. Один метод для доступу представлений нещодавно створеною локальною зоною безпечного аналізу для відвідуючих дослідників. Наступний метод доступу це видача ліцензії за межами об’єкту.

Під керівництвом Дирекції, SRS, Головний статистик координує програму ліцензування даних обмеженого користування. Для отримання файлів обмеженого користування, дослідник та організація дослідника вказують на їх розуміння питань конфіденційності і готовність гарантувати захист даних уклавши офіційний юридичний договір, ліцензійний договір, що пояснює в деталях використання даних, обіцяє запобігти розкриття конфіденційних даних, погоджується на розгляд SRS до прийняття рішення про публікацію, та обумовлює повернення даних до SRS після завершення дії ліцензії. Дослідження, що проводиться ліцензіатами, часто розміщується в наукових журналах, а також часто цитується на форумах політики.

A.11. Адміністрація соціального забезпечення (SSA)

Управління з дослідження, оцінювання і статистики (ORES), статистичне управління Адміністрації соціального забезпечення, переглядає і встановлює методологію процедури для захисту конфіденційності даних. Для розголошення статистичних таблиць, ORES використовує стратегію комбінування як приховування, так і округлення для запобігання розголошення інформації, що легко піддається ідентифікації.

Статистичні таблиці для бенефіціарів Соціального забезпечення і матеріальної допомоги

складаються із підрахунків частотності для бенефіціарів та узагальнених сум матеріальної допомоги. Детальна інформація про бенефіціара приховується, коли граничний підсумок менший ніж малозначима величина, і лише граничне значення показується. Для рядків, у яких показуються лише граничні підрахунки, суми в доларах приховуються коли число випадків, що вносять до підсумку, є меншим ніж малозначима величина. Детальні підрахунки частотності приховуються, коли всі деталі для граничного підсумку знаходяться в одній категорії. Коли приховування вводяться для запобігання розкриття в окремій комірці, додаткові приховування застосовуються для запобігання логічного виведення прихованого значення. Контрольоване округлення використовується як метод уникнення розкриття в статистичних таблицях для підрахунків частотності.

Публікації, про прибутки та інформація щодо працевлаштування, відповідають правилам IRS коли представляють таблиці (Див. параграф A.9 цього розділу). Зокрема, комірки таблиці з менш ніж 3 особами на рівні штату і 10 особами на рівні округу приховуються, а відповідний сумарний дохід також не показується. Які б комірки з даними не було приховано, додаткові приховування вводяться для запобігання логічного виведення прихованого значення. Всі суми в доларах показуються в тисячах доларів. Статистика прибутків і зайнятості випливає швидше із зразка записів IRS, ніж із 100-відсоткового файлу по прибутках та інформації щодо зайнятості.

Розголошуючи файли мікроданих публічного користування, окремі ідентифікатори видаляються зі всіх записів, а інші характерні особливості змінюються для того, щоб запобігати ідентифікації осіб, до яких відноситься запис. Записи сортуються в довільній послідовності для уникнення розголошення інформації у зв'язку з розстановкою записів у файлі. Верхні коди і нижні коди застосовуються до числових полів для того, щоб уникнути відображення екстремальних значень полів на записі даних. Значення за межами верхнього коду і нижнього коду замінюються середньою величиною значень, що перевищують відповідний верхній код або нижній код. Значення верхнього коду і нижнього коду виводяться на національному рівні, а значення заміни виводяться і застосовуються на державному рівні, коли це необхідно. Значення, показані для певних категорійних полів, комбіновані у ширші групування, ніж ті, що присутні на внутрішньому файлі, а суми в доларах округлені. Значенні верхнього коду і нижнього коду, значення заміни, і відповідна інформація надаються користувачам як частина документації файлу.

Наглядова рада з розкриття (DRB) переглядає запропоновані файли мікроданих публічного користування перед їх розголошенням. DRB складається з персоналу із ORES, які знайомі з файлами ключових даних, їх використанням, і вимогами щодо конфіденційності. Крім того, спеціалісти з конфіденційності із інших федеральних агентств можуть служити на DRB для забезпечення подальших перспектив і додаткової експертизи з конфіденційності. Персонал, що несе відповідальність за створення файлу, заповнює *Контрольний перелік з потенціалу розкриття запропонованих випусків даних,* підготовлених Міжвідомчим комітетом з конфіденційності і доступу до даних, а Контрольний перелік включається у перегляд DRB.

Б. Зведення

Більшість із 14 агентств, що охоплюються в цьому розділі, мають стандарти, норми, або офіційні механізми перегляду, які призначення для гарантії того, щоб належні аналізи розкриття здійснювались, а відповідні методи обмеження статистичного розкриття застосовувались до розголошення табличних зведень і мікроданих. Стандарти і норми агентства показують широкий

спектр специфічності: деякі з них містять лише одне чи два простих правила, тоді як інші є більш детальними. Деякі агентства публікують значення параметру, який вони використовують, тоді як інші вважають, що утримання значень забезпечує додатковий захист для даних. Очевидно, існують великі розходження в політиці, процедурах і практиках між Федеральними агентствами щодо належного захисту широких варіацій в змісті і форматі розголошеної інформації.

Б.1. Порядкові і частотні дані

Більшість стандартів і норм передбачають мінімальні розміри комірки і певний тип правила концентрації. Деякі агентства (наприклад, ERS, NASS, і NCHS) публікують значення параметрів, які вони використовують в 2правило концентрації (n, k)», тоді як інші, такі як Бюро перепису населення і BLS, не роблять цього. Мінімальні розміри комірки 3 регулярно використовуються, тому що член комірки розміру 2 міг би вивести питому величину для іншого члена. Деякі агентства привели норми точності як керівні принципи для розголошення певних табличних даних. **Норми точності** мають відношення до спеціальних правил, які агентство застосовує до даних, що стосуються деяких показників якості даних, таких як граничний рівень для відносної стандартної помилки або оціночних показників коефіцієнта варіації.

Більшість агентств, що опублікували свої значення параметру для правил концентрації використовували єдиний набір, із n = 1. Значення k коливались від 0.5 до 0.8. Найбільш продумане правило, що включене у норми або керівні принципи були «правило pq2 EIA і BEA і «правила p-відсотка», що мають відношення до Бюро перепису населення. Всі ці правила мають якість субадитивності. «Правило p-відсотка» і «правило pq» дають аналітику з розкриття гнучкість у визначенні належного розміру вигоди для окремої компанії у формі інформації про своїх конкурентів.

Один можливий метод для того, щоб справитись із комірками даних, в яких переважають один чи два значних респондентів, полягає в тому, щоб запитати в цих респондентів дозволу опублікувати комірки, навіть якщо комірка буде прихована чи замаскована згідно із нормальними процедурами агентства з обмеження статистичного розкриття. Агентства, включаючи NASS, EIA, Бюро перепису населення, і деякі із державних агентств, що співпрацюють з BLS у своїх Статистичних програмах федеральних земель, використовують цей тип процедури для деяких опитувань, щоб дозволити публікацію цих конфіденційних значень комірок. Інший метод обмеження розкриття, що використовується двома агентствами, полягає в застосуванні шуму до ключових мікроданих перед об'єднанням опублікованих значень.

В.2. Мікродані

Агентства, що надають доступ до файлів із мікроданими загального користування розробили статистичні процедури з обмеження розкриття індивідуальних даних під час надання доступу до мікроданих. Деякі агентства відзначали, що процедури з обмеження розкриття індивідуальних даних для організованих ними спостережень встановлювалися Наглядовою радою із розкриття індивідуальних даних Бюро перепису у зв'язку із тим, що такі спостереження здійснювалися за узгодженням із Бюро перепису (Розділ 13). Всі великі установи, які надають доступ до мікроданих загального користування, − Бюро перепису, Національний центр медичної статистики, Національний центр освітньої статистики − розробили формальні процедури щодо сукупностей мікроданих, які пройшли аналіз Наглядової ради, і були нею схвалені. Як писав Джабин (1993b): «В цілому, зазначені процедури не залежать від правил, що враховують конкретні статистичні значення, на зразок тих, що використовуються при табуляції. Замість цього, вони потребують застосування судження рецензента, що враховуватиме наступні фактори: наявність зовнішніх файлів із зіставними даними, засобів,

необхідних «зловмиснику» для ідентифікації індивідуальних даних, чутливість окремих елементів індивідуальних даних, очікуване число унікальних записів у файлі, частка досліджуваної сукупності, що включається до вибірки, очікуваний рівень статистичних помилок даних, а також вік даних".

Географія є важливим фактором. Бюро перепису і Національний центр медичної статистики відзначають, що географічні коди для регіонів із вибірковою сукупністю менше 100000 людей можуть бути включені до сукупностей даних загального користування. При наявності у файлі великої кількості змінних, може застосовуватися більше значення відсікання. Зазначена вимога також обмежує такі характеристики місцевості, як середній дохід, щільність населення і відсоток меншості населення у регіоні перепису оскільки, при включенні достатньої кількості змінних цього типу, місцевість може бути ідентифікована за своїми унікальними ознаками. Цікавим прикладом такої задачі є Спостереження щодо побутового споживання енергії, проведене Адміністрацією з енергетичної інформації, в якому місцеву метеорологічну інформацію, включену до сукупностей мікроданих потрібно було приховати з метою запобігання розкриття індивідуальних даних щодо географічного місце розташування домогосподарств, які брали участь у спостереженні.

Зазвичай для уникнення розкриття індивідуальних даних щодо осіб, або інших елементів у статистичному розподілі із екстремальними значеннями використовується перекодування за верхньою межею. Значення відсікання, виражені у доларах застосовуються для таких груп даних, як прибуток або активи, і при цьому, для елементів, що знаходяться поза межами цих значень відсікання точні значення не наводяться. Спотворення, порожня або умовна перестановка даних, застосування шуму, перекодування, граничні правила та округлення − є іншими методами, що зазвичай використовуються для уникнення розкриття індивідуальних даних.

Огляд методів, що використовуються агентствами

Агентство	**Порядкові дані**	**Частотні дані**	**Мікродані**	**Виключення**	**Дослідникам надається обмежений доступ**
Служба економічних досліджень (ERS)	(n. k), (1..6) 3+	Граничне правило 3+	Ні	Так	Так
Національна сільськогосподарська статистична служба (NASS)	(n, k). р-відсоток Стат. значення Конфіденційні	1 + Не чутлива для визначених спостережень	Ні	Так	Так
Бюро економічного аналізу (BEA)	р-відсоток с=l	1 + Не чутлива для визначених спостережень	Ні	Ні	Так
Бюро перепису	р-відсоток Ст. значення Конфіденційні Додавання шуму	Перестановка даних, Фільтрація даних, Граничне правило	Так − Наглядова рада із розкриття індивідуальних даних	Так	Так
Національний центр освітньої статистики (NCES)	Перестановка даних Укрупнення даних Точність	Перестановка даних Укрупнення даних Точність	Так − Наглядова рада із розкриття	Ні	Так
	Стандарти/Гранич не правило 3+	Стандарти/Гранич не правило 3+	індивідуальних даних		

Адміністрація з енергетичної інформації (EIA)	(n- k), pq. Ст. значення Конфіденційні	Граничне правило Точність Стандарти	Так − Узгодження із відомством	Так	Ні
Національний центр медичної статистики (NCHS)	(n,k), (1,6)	Граничне правило 4+	Так − Наглядова рада із розкриття індивідуальних даних	Ні	Так
Агентство досліджень і оцінки якості медичного обслуговування (AHRQ)	дані відсутні	Граничне правило 4+	Так − Наглядова рада із розкриття індивідуальних даних	Так − Наглядова рада із розкриття індивідуальних даних	Так
Адміністрація соціального забезпечення (SSA)	Граничне правило 3+	Граничне правило. Маргінали 5+, клітинки 3+	Так- Узгодження із агентством	Ні	Ні
Бюро юридичної статистики (BJS)	дані відсутні	Граничне правило 10+, Точність Стандарти	Так- Узгодження із агентством у порядку встановленому законодавством	Ні	Ні
Бюро статистики праці (BLS)	(n, k), «правило р-відсотка». Статистичні значення залежать від спостереження і елементів даних	Мінімальна кількість залежить від спостереження	Відбирається Бюро перепису (BOC) Розділ 13	Так	Так
Федеральна податкова служба (IRS)	Граничне правило 3+	Граничне правило 3+	Так- Контролюється у порядку встановленому законодавством	Ні	Ні
Бюро статистики транспорту (BTS)	Залежить від даних	Граничне правило 3+	Так − Наглядова рада із розкриття індивідуальних даних	Ні	Ні
Національний науковий фонд (NSF)	(n, k) та/або р за необхідності	Залежить від ступеня ризику	Так- Відповідає або перевищує сукупні продукти перепису загального користування	Так	Так

Примітки: Детальну інформацію щодо конкретних використовуваних методик показано у таблиці та

про неї йдеться у тексті в обсязі, наданому відповідними агентствами. Правила, що показані у різних клітинках таблиці (наприклад, p-відсоток, (n, k)) пояснюються в тексті.

На наступній сторінці міститься коротке пояснення основних термінів, що використовуються в таблиці.

Граничне правило: У разі дії граничного правила клітинка у частотній таблиці відповідає визначенню **чутливої**, якщо кількість респондентів менша певного заданого числа. Деякі агентства вимагають наявності як мінімум 5-ти респондентів у клітинці, а інші − 3-х. Іноді граничне правило застосовується до генеральної сукупності всієї таблиці. Наприклад, для публікації значень у всіх клітинках таблиці необхідна нижня межа. Агентство може змінювати таблиці і об'єднувати категорії або використовувати приховування клітинок, а також випадкове або регульоване округлення. Позначка «+» (наприклад, «3 +») означає, що у клітинці мають бути обов'язково наявні декілька спостережень відмінних від нуля. (Див. Розділ II.C.3)

Перестановка даних − це метод, що використовувався Бюро перепису населення США для захисту даних в таблицях починаючи з перепису 2000 року. За допомогою цієї методики можна уникнути розкриття статистичних індивідуальних даних у записах мікроданих до їх використання у таблицях. До відкоригованих файлів із мікроданими доступ не надається − вони використовуються виключно для підготовки таблиць. Із файлу зі всією сукупністю даних і вибіркою було відібрано невелику вибірку домогосподарств і співставлено з домогосподарствами інших географічних регіонів, що мали ідентичні ознаки щодо сукупності обраних ключових змінних. Більшість змінних в узгоджених записах були взаємозамінними. Така методика називається перестановка даних. З метою забезпечення незмінності сукупностей Перепису дозволених законодавством при застосуванні цього методу були обрані ключові змінні для співставлення. Національний центр освітньої статистики рекомендує використовувати перестановку та укрупнення даних для всіх внутрішніх та зовнішніх записів мікроданих. Національний центр освітньої статистики забороняє публікацію будь-яких клітин із кількістю випадків меншою ніж три, а також приховування клітинок у разі невикористання зазначених методів. Таблиці необхідно перекомпоновувати до тих пір, поки всі клітинки із кількістю випадків менше ніж 3 не будуть відсутні.

«Правило P-відсотку»: У випадку, коли користувач може оцінити значення кореспондента, що розкривається досить точно здійснюється приблизне розкриття індивідуальних даних щодо порядкових даних. У разі такого розкриття клітинка таблиці визнається чутливою, якщо верхня чи нижня оцінка значень респондентів ближчі до значення, що розкривається у порівнянні із визначеним заздалегідь відсотком, p. Цей метод передбачає, що до публікації даних користувач може оцінити справжнє значення з точністю плюс-мінус 100%. У Робочому документі статистичної політики №2 дане правило називається «рівень двозначної оцінки p-відсотка», для більш широкого загалу відоме як **«правило p-відсотка».** (Див. Розділ IV.B.1.a).

«Правило pq»: «Правило pq» схоже на «правило p-відсотка», але припускає, що, до публікації даних широка громадськість може оцінити дані компанії в межах відсотка q (де q <100). Таким чином, агентство може визначити ступень попередніх знань шляхом присвоєння значення *q*, яке показує, наскільки точно респонденти можуть оцінити значення іншого відповідача перед публікацією будь-яких даних ($p < q < 100$). (Див. Розділ IV.B.1.b)

«Правило (n, k)»: «Правило (n, k)» або «правило домінування» так описане у Робочому документі №2 статистичної політики. «Незалежно від кількості респондентів у клітинці, якщо невелика кількість (*n* або менше) цих респондентів внесе значний відсоток (*k* відсотків і більше) загального значення клітинки, тоді так званий **«респондент n», «правило k відсотків»** домінування клітинки визначає таку клітинку як чутливу». Багато людей вважають його простим і доступним правилом, тому що, якщо, наприклад, над клітинкою домінує один респондент, тоді опублікований підсумок сам по собі стає природньою верхньою оцінкою найбільшого значення респондента. (Див. Розділ

IV.B.1.c).

РОЗДІЛ IV – Методи для табличних даних

Розділ II представив приклади методів обмеження розкриття, що використовуються для захисту таблиць і мікроданих. Розділ III описав порядки агентства щодо обмеження розкриття. Цей розділ представляє більше деталей стосовно методологічних питань, що мають відношення до захисту конфіденційності в таблицях.

Як було згадано раніше, таблиці класифікуються на дві категорії для цілей аналізу ризику розкриття: таблиці частотних (або рахункових) даних і таблиць порядкових даних. Таблиці, що містять частотні дані, показують відсоток населення, що має певні характеристики, або ж відповідно, число в населенні, яке має певні характеристики. Якщо комірка має лише кілька респондентів і характеристики є достатньо відмінним, то для обізнаного користувача може бути можливим ідентифікувати фізичних осіб серед населення. Що стосується таблиць частотних даних, методи обмеження розкриття застосовуються до комірок з менш ніж визначеним **граничним числом** респондентів для мінімізації ризику, що фізичні особи можуть бути ідентифіковані виходячи з їх даних. Методи обмеження розкриття, що застосовуються після зведення в таблиці, включають довільне округлення, контрольоване округлення, приховування комірок і контрольоване табличне коригування. Методи обмеження розкриття, що застосовувались перед зведенням в таблиці, містять техніку захисту мікроданих, таких як збурення даних та обмін ними.

Таблиці порядкових даних типово представляють результати опитувань організацій або установ, де опубліковані елементи це зведені показники невід'ємних опублікованих значень. Для таких опитувань значення, повідомлені респондентами, можуть коливатись в широких межах, із деякими

надзвичайно великими значеннями і деякими малими значеннями. Проблема з конфіденційністю стосується забезпечення того, щоб особа не змогла використати опублікований підсумок та інші публічно доступні дані для надто точної оцінки значення окремого респондента. Методи обмеження розкриття застосовуються до комірок, для яких **міра лінійної чутливості** вказує, що дані деяких респондентів можуть бути занадто точно оцінені. Для таблиць поточних даних, приховування комірки є найбільш використовуваним методом. Контрольоване табличне коригування пропонує іншу альтернативу. Обидва методи застосовуються після зведення в таблицю. Методи обмеження розкриття, що застосовуються перед зведенням у таблицю, включають техніку захисту мікроданих, такі як додавання шуму.

Таблиці частотних даних обговорюються в Секції А. Головні методологічні сфери інтересу знаходяться в контрольованому округленні, і використання методів мікроданих, таких як обмін даними. Таблиці порядкових даних обговорюються в Секції Б. Ця секція надає деякі деталі стосовно мір лінійної чутливості, здійснення аудиту запропонованих зразків приховування і методологій автоматизованого приховування комірки.

А. Таблиці частотних даних

Таблиці частотних даних можуть мати відношення до людей та установ. Частотні дані для установ в загальному не вважаються чутливими, тому що так багато інформації про установу є публічно доступною. Технології обмеження розкриття в загальному застосовуються до таблиць частотності, основаних на демографічних даних. Як вже обговорювалось раніше, **правило заборони первинного розкриття даних**, що найчастіше використовується для вирішення того, чи комірка у таблиці частотних даних розголошує занадто багато інформації, представляє собою «правило граничної величини». Комірка визначається як чутлива, коли число респондентів менше, ніж деяка попередньо визначена гранична величина. Якщо існують комірки, що визначаються як конфіденційні, необхідно прийняти міри для їх захисту. Методи запобігання розкриття в таблицях підрахунків або частотностей було ілюстровано в II.B.2. Сюди входять: комбінування комірок, довільне округлення, контрольоване округлення, приховування комірки, контрольоване табличне регулювання, і технології мікроданих. Комбінування комірок це в загальному оцінювальна діяльність, що здійснюється керівником з опитування. Не існує жодних методологічних питань для обговорення. Обрання комірок для додаткового приховування є тією ж проблемою для обох таблиць частотностей і таблицями порядкових даних. Додаткове приховування буде обговорюватись в Розділі Б.2 цього Розділу. Контрольоване табличне регулювання є найбільш цінним для даних рівня установи, а також обговорюється в Секції Б2. Технології мікроданих використовувались для публікації даних із перепису, що проводяться з десятирічними інтервалами від 1990. Ці технології було ілюстровано в Розділі II, і технічні питання описуються в Розділі V.

Методи збурені включають довільне округлення і контрольоване округлення, як особливі випадки. Контрольоване округлення це спеціальний випадок довільного округлення. Контрольоване округлення це найбільш актуальний із методів збурення, тому що він встановлює умову, що значення комірки повинні додаватись до підсумкових величин опублікованих рядків і колонок. Це призводить до створення адитивної таблиці (суми введених даних рядка, колонки і шару відповідають опублікованому граничному підсумку). Контрольоване округлення може завжди виконуватись для двовимірних таблиць, і може в загальному виконуватись для тривимірних таблиць. Секція А.1 надає більше деталей щодо методології, що використовується в контрольованому округленні.

А.1. Контрольоване округлення

Контрольоване округлення було розроблене для подолання недоліків звичайного і довільного округлення, і для комбінування їх бажаних особливостей. Приклади довільного округлення і контрольованого округлення було надано в II.B.2. Так як і довільне округлення, контрольоване округлення замінює оригінальну двовимірну таблицю на масив, введені дані якого представлені округленими значеннями, які прилягають до відповідних початкових значень. Однак гарантується, що округлений масив буде адитивним і може бути обраним для мінімізації будь-якого класу стандартних показників відхилення між початковими та округленими таблицями.

Рішення щодо проблеми контрольованого округлення у двовимірних таблицях було знайдено в ранніх 1980-х (Кокс та Ернст, 1982). Згідно із цим рішенням структура таблиці описується як математична мережа, метод лінійного програмування, що користується спеціальними структурами в системі таблиць даних. Мережевий метод можна також використовувати для здійснення контрольованого округлення для наборів двовимірних таблиць, які ієрархічно позв’язані вздовж одного виміру (Кокс і Джордж, 1989).

Для тривимірних таблиць не існує точних мережевих рішень (Кокс та Ернст, 1982). Поточні методи використовують повторне наближене рішення застосовуючи послідовність двовимірних мереж. Точні рішення для двовимірних таблиць і наближені рішення для тривимірних таблиць є як швидкими, так і точними. Проте рішення щодо проблеми контрольованого округлення є в наявності, контрольоване округлення не є звичною практикою між урядовими агентствами США.

Б. Таблиці порядкових даних

Для таблиць даних абсолютного значення, ті значення, що доповідаються респондентами, збираються у комірках таблиці. Приклади даних абсолютного значення представлені доходом для фізичних осіб та об’ємами продажу і виручкою для установ. Особливо для установ, ці опубліковані значення як правило дуже спотворені із кількома дуже великими опублікованими значеннями, які обізнаний користувач може дуже легко співвіднести з певним респондентом. В результаті, більш математичне визначення **конфіденційної комірки** необхідне для таблиць даних абсолютних значень. Що стосується таблиць частотних даних, кожен респондент робить рівні внески у кожну комірку, що призводить до простого граничного визначення конфіденційної комірки.

Математичні визначення конфіденційних комірок обговорюються в Секції Б.1 нижче. Після того, як таблиці було створено і конфіденційні комірки ідентифіковано, необхідно прийняти рішення щодо того, як запобігти випадок розкриття. Що стосується таблиць порядкових даних, можливості включають комбінування комірок і згортання категорій, приховування комірки, і контрольоване табличне регулювання. Всі вони були підсумовані та ілюстровані в Розділі II.

При методі комбінації таблиці реконструюються (категорії згортаються), тому там є менше конфіденційних комірок. Методи реконструювання таблиць це корисні заходи, зокрема із таблицями з нового опитування або де частини таблиці містять багато конфіденційних комірок, тому що населення розсіяне. Проте, в загальному неможливо усунути всі конфіденційні комірки стягуючи таблиці, і ретельні автоматизовані процедури для стягування в загальному все ще потребують розвитку.

Історичний метод для захисту порядкових даних − це приховання комірок. Конфіденційні комірки не публікуються (вони приховуються). Ці приховані конфіденційні комірки називаються **первинними приховуваннями.** Для того, щоб переконатись, що первинні приховування не можна вивести за допомогою віднімання від опублікованих граничних підсумків, додаткові комірки обираються для **додаткового приховування.** Додаткові приховування іноді називаються **другорядними приховуванням.**

Що стосується маленьких таблиць, можливим є вибирати комірки вручну для додаткового приховування, і застосовувати процедури здійснення аудиту (див. Секцію 2.a) для гарантії, що обрані комірки належним чином захищають конфіденційні комірки. Для публікацій великомасштабних досліджень з багатьма спорідненими таблицями, обрання набору комірок додаткового приховування, які є «оптимальними» в певному сенсі є дуже складною проблемою. Додаткове приховування обговорюється в Розділі Б.2.

Контрольоване табличне регулювання також ілюструється в Розділі II. Деякі технічні характеристики обговорюються в IV.Б.3. На закінчення, методи мікроданих все більше використовуються для захисту табличних даних перед зведенням у таблиці. Для даних рівня установи, додавання шуму представляє техніку, яка застосовувалась до сьогоднішнього дня. Це підсумовується в Розділі II, і більш детально обговорюється в IV.Б.4.

Б.1. Визначення чутливих комірок – Правила лінійної чутливості

Визначення і математичні властивості вимірів лінійної чутливості та їх відношення до ідентифікації чутливих комірок в таблицях зроблені офіційними Коксом (1981). Хоча звичайні правила лінійної чутливості були відомі в 1978 і використовувались для ідентифікації чутливих комірок, їх математичні властивості не було формально продемонстровано. Важливі визначення і властивості подаються нижче.

Для даної комірки, X, із респондентами N дані рівня респондента, що робить внесок до цієї комірки може бути організований в порядку від великого до малого: $x_1 \geq x_2 \geq ... \geq x_N \geq 0$. Тоді **вимір верхньої лінійної чутливості,** $S(X)$, є лінійною комбінацією

$$S(X) \quad \sum_{i=1}^{N} w_i x_i$$

визначеною для кожної комірки або об'єднання комірок X і даних його респондента $\{x_i\}$. Послідовність постійних, $\{w_i\}$, називається послідовністю ваг $S(X)$. Ці ваги можуть бути позитивними або негативними. Комірка або об'єднання комірок X є **чутливими** якщо $S(X) > 0$. Візьміть до уваги, що множення міри лінійної чутливості на постійну вимагає іншу (еквівалентну) міру лінійної чутливості. Міри лінійної чутливості, описані в цій секції, всі нормуються таким чином, щоб вага, яка перемножує x_1 дорівнювала 1. Це нормування полегшує їх порівняння. Якщо респондент робить внесок у дві комірки, X і Y, то він залишається єдиним респондентом в об'єднанні X і Y, із значенням яке дорівнює сумі його внесків X і Y.

Одна з властивостей, що допомагає в пошуку за додатковими комірками це **субадитивність,** яка гарантує, що об'єднання розділених комірок, які не є чутливими, також є нечутливим. Кокс показує, що міра лінійної чутливості є субадитивною, якщо послідовність ваг є незростаючою, тобто якщо $w_1 > w_2 >... > w_N$. Субадитивність є важливою властивістю, тому що це означає, що зведені показники комірок, які не є чутливими, теж не є чутливими і не вимагають перевірки. Дійсні додаткові комірки мають таку властивість, що їх об'єднання із чутливою коміркою(ми) в рядку, колонці або шарі, де граничні підсумки публікуються, не є чутливими відповідно до міри лінійної чутливості. Простий результат полягає в тому, що нульові комірки не є варіантами для додаткового приховування в якості об'єднання чутливих комірок, а нульова комірка відповідає чутливій комірці, і таким чином залишається чутливою. Додаткові приховування можуть бути непотрібними, якщо граничні підсумкові значення не публікуються.

Правила первинного приховування, що широко використовуються, описані в Секціях а,б і в нижче. Вони порівнюються в Секції г. Кожне з цих правил включає в себе параметри, які визначають значення, які беруться вагами, $w_1...w_n$. Хоча агентства можуть розголошувати правило первісного приховування, яке вони використовують, вони не повинні розкривати значення параметру, так як ознайомленість з правилом і його параметрам надає респонденту можливість робити кращі логічні висновки стосовно значень, що доповідаються іншими респондентами. Приклад представлений в Розділі 3.

Існує три міри лінійної чутливості, які обговорюються в літературі і застосовуються на практиці. Це «правило p-відсотка», «правило pq» і правило (n, k). Вони описані нижче. Всі вони субадитивні, як це можна побачити з того факту, що ваги в рівняннях які визначають $S(x)$ є незростаючими. «Правило p-відсотка» і «правило pq» класифікують розрахункові дані як чутливі, якщо $n<3$.

Б.l.a. «Правило p-відсотка»

Приблизне розкриття порядкових даних має місце якщо користувач може оцінити опубліковане значення певного респондента занадто точно. Таке розкриття трапляється, а комірка таблиці оголошується чутливою, якщо верхні і нижні оцінки на значення респондента є ближчими до опублікованого значення, ніж попередньо встановлене процентне відношення, p. Це іменується як «рівень неоднозначності оцінки p-відсотка» в Робочому документі статистичної політики 2. В більш загальному розумінні воно іменується як «**правило p-відсотка»,** і має міру лінійної чутливості,

$$S^{p\%}(X) - x_1 - \frac{100}{p}\sum_{i=c+2}^{N} x_i .$$

Тут, *c* це розмір коаліції, груп респондентів, що узагальнюють свої дані з декількох джерел при спробі оцінити найбільше опубліковане значення. Комірка чутлива якщо $S^{p\%}(X) > 0$. Прийміть до уваги, що якщо є менше ніж 3 респонденти ($N < 3$) в комірці *X*, то $S^{p\%}(X) = x_1 > 0$ і комірка чутлива для всіх значень *p* і *c* .

«Правило p-відсотка» виводиться наступним чином. Припустимо, що, виходячи із загальних відомостей, будь-який респондент може оцінити внесок іншого респондента з точністю до 100-відсотків від його значення. Це означає, що респондент, який проводить оцінку, знає що значення іншого респондента є невід'ємними і в два рази менші від фактичного значення. Для «правила p-відсотка», бажано щоб після опублікування даних значення жодного респондента не повинно було піддаватись більш точній оцінці, ніж у межах *p* відсотка (де $p < 100$).

Можна доказати, що коаліція, включаючи другого найбільшого респондента, має можливість для оцінки значення x_1 найбільш точно, і якщо x_1 захищене, то відповідно і всі менші респонденти. Таким чином, достатньо тільки забезпечити захист найбільшому респонденту, і припустити що сторона, яка здійснює оцінку, це коаліція із другого по величині респондента і наступних найбільших респондентів $c - 1$. Так як респонденти з коаліції можуть здійснювати оцінку кожного $x_{c+2},...,x_N$ з точністю до 100 відсотків, вони мають оцінку для суми цих найменших респондентів, *E*, таких що

$$\left| \sum_{i=c+2}^{N} x_i - E \right| \le \sum_{i=c+2}^{N} x_i .$$

Вони можуть оцінювати значення x_1 віднімаючи значення, які вони доповіли для $\left(\sum_{i=2}^{c+1} x_i\right)$ опитування

і їх оцінка підсумку меншого респондента, *E*, із опублікованого підсумку. Похибка в цій оцінці буде рівносильна похибці при оцінюванні *E*, яке є менша або рівна

$$\sum_{i=c+2}^{N} x_i .$$

Вимога, щоб ця оцінка не була точніша ніж «p-відсоток» від значення x_1 ($p < 100$) передбачає, що

$$\sum_{i=c+2}^{N} x_i \ge \frac{p}{100} x_1 .$$

Це можна переписати як правило лінійної чутливості, вказане вище. Простіша версія «правила p-відсотка», яке припускає, що коаліції розміру «c» можуть писатись як вказано далі:

$S = x_1 - 100/p * (T - Tc - x_1)$

Де T це підсумкове значення комірок від всіх респондентів, T_c це сума значень респондента в коаліції, і x_1 це найбільше значення. Використовуючи цю формулу комірка чутлива, якщо S позитивне. У простому випадку, де $T_c = x_2$ (тобто, коаліція має розмір лише одного), то $T - T_c - x_1 = T - x_2 - x_1$, що означає, що інше значення комірки це сума всіх найменших компаній в комірці за винятком двох найбільших. $T - T_c - x_1$ буде дорівнювати нулю лише якщо коаліція (T_c) включає в себе всіх респондентів в комірці крім найбільшої компанії.

Б.І.б. «Правило pq»

При виведенні для «правила p-відсотка», ми припустили, що в наявності були обмежені попередні знання про значення респондента. Деякі люди вважають, що агентства не повинні робити цього припущення. В «правилі pq» агентства можуть визначати те, як багато попередніх знань є в наявності присвоюючи значення *q,* яке представляє як точно респонденти можуть оцінити значення іншого респондента перед тим, як будь-які дані публікуються ($p < q < 100$). Таким чином, в наявності є

уточнена оцінка, E_2, від із $\sum_{i=c-2}^{N} x_i$ такою що

$$\left| \sum_{i=c+2}^{N} x_i - E_2 \right| \le \frac{q}{100} \sum_{i=c+2}^{N} x_i .$$

Це приводить прямо до більш точної оцінки для значення найбільшого респондента, x_1 . Вимога щоб ця оцінка не була більш точною ніж «p-відсоток» від значення x_1 передбачає що

$$\frac{q}{100} \sum_{i=c-2}^{N} x_i \ge \frac{p}{100} x_1 .$$

$$S^{pq}(X) = x_1 - \frac{q}{p} \sum_{i=c+2}^{N} x_i .$$

Це можна переписати як правило лінійної чутливості.

Майте на увазі, що «правило pq» (іноді називається правилом попередньої або подальшої неоднозначності) і «правило p-відсотка» ідентичні, якщо коефіцієнт q / p, «накопичення інформації» рівносильний $100/ p$. У таблиці нижче ми використовуємо коефіцієнт q / p як єдиний параметр для «правила pq». Якщо користувачі фіксують значення для p і значення для $q < 100$, то «правило pq» є більш консервативним (буде приховувати більше комірок) ніж «правило p-відсотка», використовуючи те саме значення p .

Прийміть до уваги, що якщо в наявності є менше ніж 3 респондента ($N < 3$), то $S^{pa} = x_1 > 0$ і комірка X є чутливою для всіх значень c і q / p .

Найчастіше «правило pq» надається із розміром коаліції, що дорівнює 1. У цьому випадку право лінійної чутливості має вигляд

$$S^{pq}(X) - x_l - \frac{q}{p}\sum_{i\ 3}^{N} x_i .$$

Б.l.в. «Правило (n, k)»

«Правило (n, k)», або правило домінування було описано як вказано далі в Робочому документі статистичної політики 2. «Незважаючи на число респондентів в комірці, якщо маленьке число (n або менше) цих респондентів вносять велике процентне відношення (k відсоток або більше) підсумкового значення комірки, тоді так званий **«респондент n», «правило k відсотка»** домінування комірки визначає цю комірку як чутливу». Багато людей вважають це правило таким, що наглядно привертає увагу, тому що, наприклад, якщо в комірці домінує один респондент, то сама опублікована підсумкова величина є натуральною верхньою оцінкою для значення найбільшого респондента. Проте, коаліції конкретно не обговорюються при виведенні «правила (n, k)», агентства вибирають значення *n,* які повинні бути більші, ніж число будь-яких потенційних коаліцій. Багато агентств використовують «правило (n, k)» з *n* =1 або 2.

Міра лінійної чутливості для «правила (n, k)» дається

$$S^{(n,k)}(X) \quad \sum_{i-1}^{n} x_i - \frac{k}{100-k}\sum_{i-n-1}^{N} x_i .$$

Якщо $N \le n$, $S^{(n,k)} = \sum_{i=1}^{N} x_i > 0$ і комірка X чутлива для всіх значень k.

Б.l.г. Відносини між «правилами (n, k)» і «p-відсотка» або «pq»

Таблиця 1 призначена для того, щоб допомагати користувачам вибрати значення параметру p для використання із «правилом p-відсотка» з коаліціями розміру 1 (або для значення коефіцієнта, q / p, для «правила pq» з коаліціями розміру 1) коли вони використовуються для мислення категоріями «правила (n, k)». Для різноманітних значень p% (q / p), таблиця показує значення k_1 і k_2 такі що, якщо правило лінійної чутливості для (1, k_1) або (2, k_2) є позитивним, то правило лінійної чутливості для «правила p-відсотка (pq)» буде позитивним. З цим формулюванням «правило p-відсотка (pq)» є більш консервативним, воно приховає більше комірок, ніж будь-яке із двох «правил (n, k)» окремо, і також більше ніж правило комбінації, основане на двох «правилах (n, k)». Виведення нерівностей, що використовуються в Таблиці 1, представлено в Технічних примітках в кінці цього Розділу. Крім того, райони чутливості для «правил (n, k)», «p-відсотка», і «pq» ілюстровані графічно у Технічних примітках. Див. Робертсон Д.А. (1993) для теоретичного аналізу, що порівнює правила розкриття.

Для ілюстрації використання Таблиці 1, якщо аналітик хоче впевнитись, що комірка, в якій найбільший респондент робить внески, більш ніж 75 відсотки від підсумкової величини

приховується, і що приховується комірка, де два найбільших респондента перевищують 85 відсотки від підсумку, він/вона повинен приблизно завершити це використовуючи «правило p-відсотка» із p, що дорівнює 33.3 відсоткам, або «правило pq» із накопиченням інформації, $q / p = 3$.

«Правило p-відсотка», «pq» і «(n, k)» так як і правило комбінації,

$$S^{comb} = max(S^a(X), S^b(X))$$

є субадитивними правилами лінійної чутливості. (Тут $S^a(X)$ і $S^b(X)$ позначають будь-які дві субадитивні міри лінійної чутливості). Будь-які з цих правил прийнятні з математичної точки зору. Однак, «правилу p-відсотка» або «правилу pq» надається перевага з двох основних причин. По-перше, концепція інтервалу допуску прямо проводить паралель між методами, що на даний час використовуються для додаткового приховування, (див. секцію Б.2.a.iii). По-друге, як представлено в таблиці вище і в прикладі в Технічних примітках, «правило p-відсотка (pq)» забезпечує більш стійкі зони захисту, ніж окрема версія «правила (n, k)».

ТАБЛИЦЯ 1 Відносини між районами приховування для «правила p-Відсотка» або «правила (pq) і (l,k)», «правил (2,k)»

		$S^p\%(X) > 0$ і	Чутлива, коли комірка
P	q / p	x_1 / T перевищує:	$(x_1 + x_2) / T$ перевищує:
50.0%	2	66.7%	80.0%
33.3%	3	75.0%	85.7%
16.7%	6	85.7%	92.3%
11.1%	9	90.0%	94.7%

ПРИМІТКА: $T - \sum_{i-1}^{N} x_i$ це підсумкова величина комірки.

Б.1.д. Інформація у значеннях параметра

Агентства можуть публікувати свої правила приховування, проте, вони повинні зберігати конфіденційність значень параметра, які вони використовують. Обізнаність як з правилом, так і з значеннями параметра надає можливість користувачу зробити кращі логічні висновки стосовно значення прихованих комірок, і можуть зруйнувати мету приховування.

Наприклад, припустимо, що агентство використовує правило p-відсотка із p=20 відсотків, і що таке ж значення p використовується для визначення областей захисту для додаткового приховування. Ми

припускаємо, що підсумкова величина комірки це 100, і що комірка чутлива згідно із «правилом p-відсотка». Ця комірка буде приховуватись разом із іншими підходящими додатковими комірками. Що стосується цієї комірки (як і будь-якої прихованої комірки), будь-який користувач може використовувати пакет лінійного програмування для обчислення верхніх і нижніх меж для підсумкової величини комірки, що основана на опублікованих рівняннях рядка і колонки. Припустимо, що це призводить до наступної нерівності:

80 = нижня межа < підсумкова величина комірки < верхнє значення = 120.

У цьому випадку район захисту, що використовується в обиранні комірок для додаткового приховування запевняє, що підсумкову величину комірки не можна оцінити більш точно ніж плюс або мінус 20 відсотків від значення комірки, або плюс чи мінус 20 у цьому випадку. Обізнаний користувач таким чином, однозначно визначив, що значення підсумкової величини прихованої комірки повинне бути 100. Після того як підсумкову величину для однієї прихованої комірки було однозначно визначено, ймовірно що значення іншої комірки можуть легко виводитись відніманням від опублікованих граничних підсумкових величин.

Б.2. Додаткове приховування

Після того, як чутливі комірки будуть ідентифіковані з допомогою правила первинного приховування, інші нечутливі комірки повинні відбиратись для приховування та гарантії, що дані рівня респондента в чутливих комірках не можна оцінити занадто точно. Це єдина вимога для запропонованого набору додаткових комірок для таблиць порядкових даних і, як в загальному вважається, означає, що дані респондента не можна оцінювати більш точно, ніж плюс чи мінус від якогось процентного відношення.

Існують два способи в які дані рівня респондента можуть бути скомпрометовані. По-перше, об'єднання прихованих комірок, опубліковані в неявній формі, можуть бути чутливими відповідно до міри лінійної чутливості. Це залежить від характеристик даних рівня респондента в об'єднанні комірки, і зазвичай буває проблемою лише коли ті ж самі респонденти роблять внески в обидві комірки. По-друге, рівняння рядка і колонки, представлені опублікованою таблицею можуть розв'язуватись, а значення для прихованої комірки може оцінюватись занадто точно. Автоматизовані методи **здійснення аудиту** зразка запропонованого приховування можуть бути потрібними для гарантування того, щоб первинні приховування були належним чином захищені (див. Секцію Б.2.а).

Будь-який набір комірок, запропонований для додаткового приховування прийнятний за умови, якщо чутливі комірки захищені. Для маленьких таблиць це означає, що обирання додаткових комірок може здійснюватись вручну. Зазвичай аналітик даних знає які комірки представляють найбільший інтерес для користувачів (і не повинні використовуватись для додаткового приховування, якщо це можливо), і які представляють менший інтерес для користувачів (і таким чином ймовірні варіанти частина додаткового приховування). Ручне обирання додаткових комірок є прийнятним за умови, якщо отримана в результаті таблиця забезпечує достатній захист для чутливих комірок. Автоматизований аудит повинен застосовуватись для запевнення, що дані правдиві.

Для великих систем таблиць, наприклад тих, що основані на Економічному переписі, обирання

додаткових комірок є основним завданням. Ручне обирання комірок може означати, що чутлива комірка залишається незахищеною по недбалості, або що сумісність не досягається від однієї таблиці до іншої в публікації. (Кокс, 1980). Несумісність у зразках приховування в публікації підвищує імовірність ненавмисного розкриття. З цієї причини техніки лінійного програмування застосовувались статистичними агентствами до обирання комірок для додаткового приховування протягом багатьох років. (Кокс, 1995). В якості додаткового привілею, агентства очікують, що автоматизоване обрання додаткових комірок призведе до менших втрат інформації через приховування. Приклади теорії і методів для автоматичного обрання комірок для додаткового приховування обговорюються в Секції Б.2.б.

Б.2.а. Аудит запропонованих додаткових приховувань

Б.2.а.і. Об'єднання прихованих комірок, опублікованих в неявній формі, є чутливими

Якщо чутливі комірки захищені приховуванням інших комірок внутрішніх таблиць при публікації граничних підсумкових величин, передбачувані результати полягають в тому, що об'єднання прихованих комірок в рядках, колонках і шарах розголошуються відніманням від підсумкової величини. Таким чином, один спосіб здійснення аудиту захисту, що забезпечується зразком приховування, полягає в застосуванні правила лінійної чутливості до цих об'єднань для гарантії, щоб вони не були чутливими. Тоді як цей тип аудиту є звичайною справою для маленьких таблиць, Кокс (1980) наголошує, що для великих таблиць це може бути нерозв'язним в обчислювальному відношенні, за винятком якщо застосовується систематичний підхід. Цей тип аудиту не включається в стандартне програмне забезпечення з аудиту через його залежність від даних рівня респондента.

Очевидно, що таблиця, для якої були вручну обрані зразки приховування, вимагає щоб аудит гарантував прийнятність зразку. Ранні версії програмного забезпечення з додаткового приховування використовували договори наближеного значення для обирання комірок для додаткового приховування (дані окремого респондента не використовувались.) Ці методи гарантували, що об'єднання прихованих комірок не були чутливими до тих пір, поки різні респонденти робили внески у кожну комірку. Проте, якщо ті ж респонденти робили внески в численні комірки і об'єднання комірок, то аудит був потрібний.

Б.2.а.іі. Рівняння ряду, колонки і/або шару можна розв'язувати для прихованих комірок

Двовимірна таблиця із проміжними підсумками ряду, колонки, і тривимірна таблиця із проміжними підсумками рядку, колонки і шару можуть розглядатись, як велика система лінійних рівнянь. Приховані комірки представляють невідомі значення у рівняннях. Можливим є те, що рівняннями можна маніпулювати, а приховані значення можна оцінювати точно. Аудиторські перевірки для цього типу розкриття вимагають використання методик лінійного програмування. Результатом цього типу аудиторської перевірки є максимальне і мінімальне значення, яке кожна прихована комірка може прийняти за наявності іншої інформації у табличці. Коли максимум і мінімум є рівними, значення комірки точно розкривається. Для певності, що комірки не можуть бути оцінені занадто точно, аналітик переконується, щоб максимальне і мінімальне значення для прихованої комірки не були ближче до правдивого значення, ніж певне встановлений захист процентного відношення.

Добре відомо, що модель мінімального приховування, де представлені граничні підсумкові значення, буде мати принаймні дві приховані комірки в кожному рядку, колонці і шарі, що вимагає приховування. Проте цього не є достатньо, як це було продемонстровано в Розділі 2 Секції В.2.а.

B.2.a.iii. Програмне забезпечення для зразка приховування

Автоматизовані методи здійснення аудиту зразка приховування були доступними від середини 1970-х в Бюро перепису населення США, та у Статистичній службі Канади. Сучасні версії програмного забезпечення поставили завдання лінійного програмування і використовують комерційно доступні пакети лінійного програмування. Всі системи аудиту формують верхні і нижні оцінки для значення кожної прихованої комірки, основаного на лінійних комбінаціях опублікованих комірок. Аудит приховування може віднайти три типи проблем для комірок таблиць: 1) верхні і нижні межі можуть бути однаковими; 2) верхні і нижні межі можуть знаходитись близько один до одного; 3) верхні і/або нижні межі можуть знаходитись занадто близько до значення комірки. Аналітик даних використовує вихідні дані із аудиту для визначення того, чи захист, що забезпечується для чутливих комірок запропонованими додатковими комірками є достатнім. Користувач повинен знати тип і ступінь округлення значень комірки в таблиці, яка проходить аудит для уникнення недостовірних оцінювань захисту даних. Залежно від того, чи приховування застосовувалось до округлених або неокруглених даних, воно може призвести до недостатнього або надмірного приховування комірок в таблиці. Ці методи аудиту застосовуються як до таблиць абсолютних значень, так і до таблиць частотності.

Лінійне програмування це найбільш розповсюджена процедура, що використовується для здійснення аудиту зразків приховування у таблиці, тому що його можна використовувати для більш багатовимірних таблиць. (Заяц 1992a). Мережеві процедури показувались для забезпечення швидких рішень для двовимірних таблиць. Метод потоку в мережі для приховування комірки має можливість самостійного здійснення аудиту лише для двовимірних таблиць, в яких є ієрархія в одному вимірі. Метод потоку в мережі не має можливості самостійного здійснення аудиту для двовимірних таблиць з ієрархічною змінною структурою як у рядку так і в колонці, і він не має функції здійснення самостійного аудиту для тривимірних чи багатовимірних таблиць, що містять ієрархічну структуру. (Массель 2002).

В Бюро перепису населення США обидва типи аудиту включаються в алгоритм, що обирає комірки для додаткового приховування. Внески рівня компанії для комірки використовуються при обиранні рівня захисту або інтервалу допуску для кожної комірки, що забезпечує захист для всіх респондентів в комірці. Алгоритм, який обирає комірки для додаткового приховування передбачає, що первинні комірки не можуть бути оцінені більш точно, ніж визначений інтервал допуску. Додаткові приховування, що обираються застосуванням алгоритму, не потребують додаткових аудитів.

Аудиторське програмне забезпечення було розроблене Комітетом з конфіденційності і доступу до даних, з підтримкою від певної кількості статистичних агентств, і є доступним з документацією за посиланням http://www.fcsm.gov/committees/cdac/resources.html. Це програмне забезпечення написано в SAS® і перевіряє нижні і верхні межі навколо прихованих комірок в таблиці, що містить неадитивні, незалежно округлені комірки. Програма вимагає конкретного формату для вхідного файлу ASCII. Програма також здійснює перевірку на предмет того, чи незалежні округлені комірки існують у таблиці і регулює значення комірок для збереження адитивності у межах ряду і колонок, і у той час вона здійснює функцію імпортування. Користувач має можливість встановлення діапазону методів захисту, основаних на принципі більшого чи меншого відсотка, або абсолютного значення. Програмне забезпечення не обмежене числом вимірів в таблиці, а методологія лінійного програмування передбачає два типи оптимізаторів.

Б.2.б. Автоматизоване обрання комірок для додаткового приховування

Програмне забезпечення, що автоматично обирає додаткові комірки для приховування буди доступні від 1970-х у Статистичній службі Канади та в Бюро перепису населення США. Ці програми зазвичай використовують методи лінійного програмування, що вводиться в дію через здійснення доступу до рутинних операцій лінійного програмування загального призначення, що використовують спеціальні структури в даних. Завдяки вдосконаленням в алгоритмах лінійного програмування, ці рутинні операції тепер виконуються швидше ніж в 1980-х. Методи потоку в мережі можуть розглядатись як спеціальний випадок лінійного програмування. Вони найкраще працюють для двовимірних таблицях, головне з одним рівнем ієрархії (або в рядках чи в колонках). Рутинні операції, основані на методах потоку в мережі типово значно швидші ніж рутинні операції лінійного програмування. Програми приховування комірки можуть використовуватись як для таблиць порядкових даних і таблиць з даними частотності.

В Бюро перепису населення США ці програми використовувались в основному для порядкових даних огляду господарської діяльності. Роберт Джуетт (Джуетт, 1993) написав набір програм з приховування комірок для цих цілей. Вони включають в себе функції поза межами основної моделі приховування комірок. Наприклад, програму можна використовувати для ідентифікації чутливих комірок із наданого файлу вхідних мікроданих, використовуючи «правило p%». Проблема «звичайних респондентів» вирішується визначенням таблиці обчислювальних спроможностей для кожного первинного об’єкту. Він будується якраз перед тим, як додаткові приховування обираються для захисту даного первинного об’єкту. Проблема «звичайних респондентів» часто виникає із даними огляду господарської діяльності так як більшість компаній мають більше ніж одну установу, і часто ці установи є вкладниками в різні комірки тієї ж таблиці. Бюро перепису населення США повинне захищати не лише внесок кожної установи, але й усі суми матеріальної бази установи, включаючи підсумковий внесок компанії. Ці програми також спроможні вести таблиці, які є приєднані і взаємозв’язані з комірками в одній чи більше таблицях. Вона використовує метод пошуку з поверненням для перевірки, що дана прихована комірка має такий самий ступінь невизначеності в кожній комірці, в якій він появляється.

Програмне забезпечення, «tau-Argus», розроблений виходячи із Обчислювальних аспектів статистичної конфіденційності (CASC). Європейський проект пропонує методи для ідентифікації чутливих комірок, вибір алгоритмів для обчислення меж інтервалу прихованих комірок, і модуль для генерування синтетичних значень щоб замінювати первісні приховані значення в публікації. Документація і програмне забезпечення для експлуатації «tau-Argus» доступні за посиланням http://neon.vb.cbs.nl/casc.

При безпосередньому введенні в дію лінійного програмування, чутливі комірки обробляються послідовно, починаючи із найбільш чутливих. На кожній стадії (тобто для кожної чутливої комірки) ідентифікується набір додаткових комірок, що мінімізує функцію витрат (зазвичай це сума прихованих значень). Заяц (1992а) описує формулювання для двовимірних таблиць. Заяц (1992б) надає паралельне формулювання для тривимірних таблиць. Як вказано вище, вони вводяться в дію використовуючи комерційно доступний пакет лінійного програмування. Недолік підходу безпосереднього лінійного програмування полягає в машинному часі, який він вимагає. Для великих задач, час виконання Центрального процесора персонального комп’ютера значним чином зростає із 3 чи більше вимірами.

Інший підхід лінійного програмування базується на описуванні структури таблиці, як математичної мережі, і використанні цієї структури і необхідних інтервалів допуску для кожної комірки для балансування таблиці. Методам мережі надається перевага, тому що вони дають ті ж самі результати як і безпосередні методи лінійного програмування, але рішення вимагає набагато менше машинного часу. Метод мережі прямо застосовується до двовимірних таблиць і двовимірних таблиць з обмеженнями проміжного підсумку в одному вимірі (Кокс, 1995). Обмеження проміжного підсумку мають місце, коли дані в одному вимірі мають ієрархічну адитивну структуру, таку як система кодування Системи статистичної класифікації видів економічної діяльності в Північній Америці (NAICS). Протягом останніх 20 років проводилась значна кількість досліджень в розробці швидших і більш ефективних процедур для як двовимірних, так і тривимірних таблиць. Дослідження охоплювало використання методів, основаних на цілочисельному програмуванні, теорії розподілу потоку і нейронних мереж.

Додаткове приховування і контрольоване округлення можуть розв'язуватись використовуючи теорію мереж. Ернст (1989) продемонстрував неможливість представлення загальних таблиць три- чи більш великої розмірності в якості мережі. З цієї причини, додаткове приховування для тривимірних таблиць, на даний час використовує лінійне програмування як головний підхід. (Заяц, 1992б). Методи безпосереднього лінійного програмування можна використовувати для маленьких тривимірних таблиць. Проте, що стосується великих тривимірних таблиць, використовується повторний наближений метод, оснований на послідовності двовимірних мереж. Зразок додаткового приховування, ідентифікований цим наближеним методом, повинен все ще проходити аудит для гарантії, що окрему чутливу комірку не можна оцінити занадто точно.

Як згадано вище, одна можлива ціль або функція витрат для автоматизованих процедур полягає в мінімізації суми прихованих значень. З цією цільовою функцією, автоматизовані процедури мають тенденцію приховувати багато маленьких комірок, результат який в загальному не вважається «оптимальним». Інші можливі функції витрат містять мінімізацію загального числа прихованих комірок в таблиці або ж мінімізацію приховування для серії конкретних даних в таблиці. Необхідні подальші дослідження ідентифікації функцій витрат для використання при обиранні «оптимальних» додаткових приховувань. Можливості − дослідження функції витрат для використання при одноразовому запуску програмного забезпечення, так само як і функцій витрат для використання при багаторазових запусках програмного забезпечення. Приклад представлено розробкою функції витрат, що використовується протягом другого проходу через програмне забезпечення для усунення надлишкових приховувань. (Заяц, 1992б).

Іншою причиною, з якої додаткові комірки, що обираються автоматизованими методами, не передбачають «оптимальний» набір для таблиці як одного цілого є те, що всі поточні введення в дію захищають чутливі комірки послідовно. Що стосується будь-якої даної чутливої комірки, додаткові комірки обрані для її захисту будуть оптимальним відповідно до цільової функції, обумовленими всіма приховуваннями, обраними для попередньо обговорюваних чутливих комірок. Послідовний характер підходу призводить до надлишкового приховування.

Незважаючи на брак «оптимальності» результату, автоматизовані процедури приховування додаткових комірок ідентифікують корисні набори додаткових приховувань. Однак, інколи потрібно проводити роботу для точного налаштування, зменшення надлишкового приховування, і гарантування, що нематематичне визначення аналітиком «оптимального» рішення реалізується більш точно.

Б.3. Контрольоване математичне регулювання

Контрольоване математичне регулювання є корисною методологією для захисту таблиць порядкових даних, так само як і рахункових даних. Воно обговорюється із прикладом в Розділі 2 Секції Г.3.г. Кожне чутливе початкове значення в таблиці замінюється умовно нарахованим безпечним значенням, що є статистичним відхиленням від істинного чутливого значення. Деякі з інших значень нечутливих комірок регулюються виходячи з їх істинних значень якнайменшу суму, як це можливо для відновлення адитивності опублікованих підсумкових величин. Можна застосувати СТА для розробки рішень, де граничні суми змінюються мінімально. Однак, дозволяючи незначне регулювання граничних значень ви зменшуєте потребу в більших регулюваннях внутрішніх нечутливих комірок в таблиці.

Існує два різних підходи, що застосовують методологію СТА. Оригінальний метод СТА використовує метод лінійного програмування для відновлення адитивності таблиці. Спочатку, процедура **Контрольованого табличного регулювання на основі LP** використовувала обернену величину значень комірки в якості функції витрат для мінімізації сумарного відхилення нечутливих комірок від істинного значення комірки. Інша належна функція оптимізації може полягати в мінімізації суми абсолютних значень регулювань даних. Обернена величина від значення комірки надає можливість для більших змін великих комірок і спричиняє незначні зміни маленьких комірок у порівнянні з іншими функціями витрат. Більшість процедур, основаних на LP, переглядають якість і здійсненність рішення, використовуючи базової структури таблиці. Алгоритм систематично змінює чутливі і нечутливі комірки, спершу намагаючись досягти здійсненного рішення, а потім, після того як здійсненність досягається, він переходить до оптимізації якості регулювання використовуючи попередньо визначену функцію витрат. Програмне забезпечення, що використовує певний тип процесу адаптивної пам'яті для перегляду оптимального регулювання, забезпечує кращі результати в плані мінімальних регулювань значень комірки, ніж ті методи, що застосовують модель «жорсткої пам'яті», таких як метод галузей і меж.

Протягом першого етапу застосування кожного типу методології СТА, чутливі комірки впорядковуються від найбільших до найменших. Використовуючи знакозмінну послідовність, впорядковані значення чутливих комірок потім змінюються, або на нижні, або на верхні межі захисту. Після завершення внесення змін до всіх чутливих комірок в таблиці, нечутливі комірки таблиці вважаються такими, що відновлюють адитивну структуру таблиці.

Другий підхід, що називається **спрощеним Контрольованим табличним врегулюванням,** був розроблений як економна альтернатива оригінального методу СТА, основаному на LP. Спрощене СТА мінімізує процентне відхилення від істинного значення комірки для нечутливих комірок в якості її функції оптимізації. Критерії мінімального процентного відхилення, що використовується у спрощених процедурах контрольованого табличного регулювання, виробляє такі ж результати, як обернена величина функції витрат, основаної на значенні комірки, що використовується в підході на основі LP. (Дандекар, 2004). Спрощену СТА простіше вводити в дію, і вона є більш ефективною в обчислювальному відношенні, ніж процедура СТА на основі LP, хоч подальше дослідження необхідне на різних структурах таблиці для подальшого оцінювання цих двох підходів. СТА на основі LP і спрощені СТА використовують різноманітні підходи для відновлення адитивності структури таблиці. Оригінальний метод СТА використовує метод лінійного програмування для відновлення адитивності таблиці. Спрощений метод СТА, з іншого боку, приймає всі необхідні регулювання в граничних значеннях комірок таблиці для відновлення адитивної структури таблиці.

Б.4. Додавання шуму до мікроданих перед зведенням даних в таблиці

Додавання шуму в ключові мікродані це метод, який використовувався для захисту табличних порядкових даних. Відрізняється від процедур введення шуму, що використовуються для захисту файлів мікроданих публічного використання. Метод додавання шуму врегульовує кожне значення на малу суму (точний відсоток, необхідний для підтримки конфіденційності у межах статистичного агентства). Кожній установі, що доповідає у вибірці чи опитуванні, присвоюється множник, або фактор шуму. Компанія може мати різні матеріальні засоби або установи. В цьому випадку кожній установі може приписуватись трохи інший множник, за умови якщо загальне розповсюдження множників по всіх установах у межах компанії виводить середнє значення визначеного відсотка для регулювання опублікованих значень тієї компанії. (Еванс, 1998).

Наприклад, якщо дані установи регулюється на 10%, то її дані будуть множитись на число, близьке або до 1,1 або 0,9. Будь-який тип розповсюдження може використовуватись для обирання множників для кожної установи. В цьому прикладі, яке б розповсюдження не використовувалось для генерування множника 1.1, важливим є те, що та ж сама форма розповсюдження, або її «дзеркальне відображення», використовувалось для генерування множників, близьких до 0,9 для регулювання даних в протилежному напрямку. Два розподіли множників повинні формувати спільний розподіл множників, що є симетричним і наближений до 1.

Напрямок додавання шуму для кожної опитуваної компанії призначається довільно. Використовуючи приклад 10% в якості основи для збурення, це рівносильно визначенню, чи всі установи в компанії мають множники, близькі до 1.1 або близькі до 0.9. Наступник кроком в процесі є довільне призначення множника кожній установі у межах компанії. Множники будуть генеруватись із тієї половини загального розподілу множників, що відповідає напрямку збурення, призначеного тій компанії. Приклад призначення множників ряду респондентів виглядає наступним чином:

Приклад 1:

Компанія	Установа	Напрямок	Множник
Компанія А		1.1	
	Установа А1		1.12
	Установа А2		1.09
	Установа А3		1.10
	Установа А4		1.11
Компанія Б		0.9	
	Установа Б1		0.89
	Установа Б2		0.93
Компанія В		1.1	
	Установа В1		1.08

У цьому прикладі, очікуване значення суми доданого шуму у значенні будь-якої комірки дорівнює нулю через симетрію розподілу множників і довільне присвоєння як напряму збурення, так і множників у межах кожної компанії. Ймовірність, що установи компанії будуть збурені в позитивному напрямку, дорівнюють ймовірності, що вони будуть збурені в негативному напрямку.

Розподіл множників симетричний відносно 1. Очікуване значення того чи іншого множника дорівнює 1, і таким чином, очікуване значення *рівня* шуму в будь-якій установі дорівнює 0, і рівень шуму у значенні будь-якої комірки це просто сума шуму в установах, що входять до його складу.

Додавання шуму відрізняється від Контрольованого табличного врегулювання, тому що додавання шуму врегульовує опубліковані значення перед будь-якими зведеннями у таблиці. Контрольоване табличне врегулювання коригує комірки після того, як дані було зведено у таблиці по комірці і на її основі. Додавання шуму покладається на довільне умовне нарахування множника, для контролювання наслідків додавання шуму до різних типів комірок.

B. Інтерактивні системи опитування даних

Більшість інтерактивних систем опитування, які було розроблено федеральними агентствами, надають доступ до зведених файлів із матрицями сукупних даних. Ці системи опитування надають можливість користувачам проектувати опитування для генерування налаштованих табличних зведень. Спеціальні методи обмеження розкриття повинні братись до уваги, коли користувачі здійснюють доступ до файлів мікроданих для формування налаштованих табличних зведень.

Один приклад інтерактивної системи опитування, що дозволяє користувачам здійснювати доступ до файлів мікроданих це «Вдосконалена система опитування» (AQS), яка є частиною інтерактивної системи розповсюдження даних «Американський шукач фактів» ("American Fact Finder"), розробленої Бюро перепису населення США. Файли мікроданих в AQS містять інформацію по фізичних особах і домашніх господарствах. Для гарантії того, щоб табличні зведення із цих файлів мікроданих не розголошували особи респондентів, Бюро перепису населення використовує методики запису даних і перестановки даних на додачу до методик мікроданих.

Змінні, такі як географія, детальна інформація про расу, вік, професію, старанність, латиноамериканське походження, і житло групи перекодовуються і/або стягуються. Всі безперервні змінні, такі як дохід, витрати на пальне і комунальні послуги, податки на майно, орендна плата, і виплати за заставою підлягають верхньому кодуванню для маскування аномальних значень в хвостах розподілу кожної безперервної змінної. Перекодовані змінні додаються до файлів, що використовуються AQS. Зовнішній користувач направляється до перекодованих змінних і географічної зони при подачі запиту.

Крім запису, методика обміну також застосовується до записів у файлах мікроданих. Методика складається з обмінних пар записів домашнього господарства, обрані як такі, що мають найвищий ризик розкриття, що базується на попередньо визначеному наборі ключових змінних. У системі AQS, записи обираються для обміну із ймовірністю, обернено пропорційною до розміру блоку.

Будь-який запит, що подається користувачем, проходить через два фільтри; Фільтр опитування і Фільтр статистичних результатів. Ціллю Фільтра опитування є виявлення тих опитувань, що не пройдуть обмеження розкриття перед тим, як опитування подається для виконання, як наприклад

географічна змінна повинна відповідати мінімальній граничній величині. Фільтр статистичних результатів перевіряє остаточні значення в комірках створеної таблиці. Якщо таблиця не проходить через фільтри, вся таблиця приховується і користувач не отримує її. Система AQS не виконує будь-якого приховування комірок. Користувачу надсилається повідомлення, що таблиця приховується в цілях конфіденційності і користувач може потім спробувати подати запит за таблицею із менш детальними даними.

Процедури захисту від розкриття, що застосовуються Опитуванням з управління сільськогосподарськими ресурсами (ARMS), використовують інший підхід ніж система AQS. Інтерактивна система опитування ARMS надає користувачам можливість обирати між наборами даних з опитування і будувати індивідуалізовані звіти. Існує три стадії в процедурах захисту від розкриття, що використовуються у системі ARMS. При першому кроці, шум додається до ваг для ключових мікроданих в унікальний спосіб для того, щоб захистити великі установи, які можуть переважати в комірці. Другим кроком є розробка мінімальних розширених підрахунків по фермі в комірці, і тестування чутливості цієї комірки використовуючи правило «p-відсотка». Третій крок застосовує приховування первісної комірки без будь-якого додаткового приховування. Жодне додаткове приховування не є необхідним, тому що шум, який було початково додано до мікроданих, забезпечує необхідний захист для зведених показників. Комірки у виведених файлах приховуються, якщо вони не відповідають будь-яким із трьох критеріїв: 1) якщо співвідношення значення комірки із шумом до значення комірки без шуму знаходиться поза межами встановленого діапазону, то комірка приховується; 2) якщо зважений підрахунок по фермі для комірки є малим, то комірка також приховується; 3) якщо комірка не відповідає «правилу p-відсотка» і має недостатній рівень шуму для захисту фактичного значення, то комірка також приховується. Підхід, що використовується в ARMS, уникає потребу в додатковому приховуванні і спрощує обчислювальні задачі, що мають відношення до захисту розкриття в інтерактивній системі опитування.

Г. Технічні примітки: співвідношення між загальними мірами лінійної чутливості

Ця секція ілюструє співвідношення між «p-відсотком», «правилами pq» і «(n, k)», описаними в тексті використовуючи ділянки областей чутливості комірки. Для спрощення цієї презентації ми зробимо кілька припущень. По-перше, для «правила p-відсотка» ми припускаємо, що немає жодних коаліцій

(c = 1) і для «правил (n, k)» ми беремо до уваги лише $n = 1$ та $n = 2$. По-друге, замініть $\sum_{i=3}^{N} x_i$ на

$(T - x_1 - x_2)$. По-третє, розділіть кожне правило чутливості на підсумкову величину комірки, *T,* і помножте на 100. І на закінчення, встановіть $z_i = 100x_i / T$, відсоток, який вноситься в підсумкову величину комірки компанією і, Правила чутливості можуть записуватися

$$S^{p\%}(X) - \left(1 \cdot \frac{100}{p}\right) z_1 \mid \frac{100}{p} z_2 - \frac{100}{p} 100,$$

$$S^{pq}(X) - \left(1 + \frac{q}{p}\right) z_1 - \frac{q}{p} z_2 - \frac{q}{p} 100,$$

$$S^{(1,k_1)}(X) - \left(1 + \frac{k_1}{100 - k_1}\right) z_1 - \frac{k_1}{100 - k_1} 100$$

$$S^{(2,k_2)} = \left(1 + \frac{k_2}{100 - k_2}\right) z_1 + \left(1 + \frac{k_2}{100 - k_2}\right) z_2 - \frac{k_2}{100 - k_2} 100$$

Області, де ці правила чутливості є позитивними (тобто, де комірки є чутливими) показані на Малюнку 1. Горизонтальна вісь представляє відсоток, внесений найбільшою одиницею, z_1 а вертикальна вісь представляє відсоток, внесений другою по величині одиницею, z_2. Оскільки $z_1 > z_2$ і $z_1 + z_2 < 1$ (сума двох найбільших менша, або рівна підсумковій величині комірки), єдині можливі значення в комірці таблиці будуть у нижній трикутній області, обмеженій знизу лінією $z_2 = 0$, зверху лінією $z_1 = z_2$ і справа лінією $z_1 + z_2 = 1$.

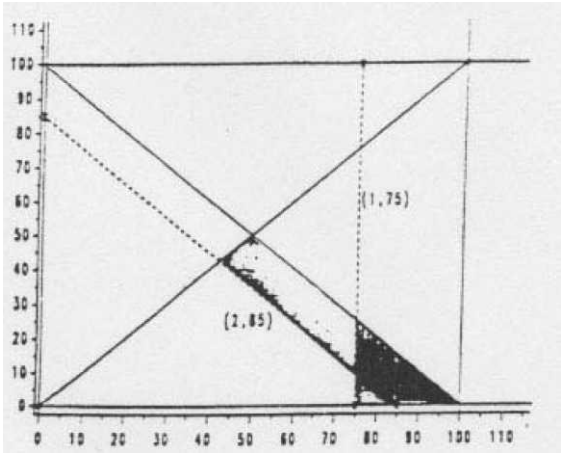
«Правила (l, k_l) і (2, k_2)» є особливо простими для графічної ілюстрації. «Правило нерівності (l, k_l)» спрощує, а комірка класифікується як чутлива якщо $z_1 > k_l$. Лінія розділення між чутливою і нечутливою областю надається вертикальною лінією через точку (0, k_l). Подібним чином, нерівність для «правила (2, k_2)» спрощується і комірка класифікується як чутлива якщо $(z_1 + z_2) > k_2$ (zl + z2) > k2. Розмежувальна лінія між чутливими і нечутливими областями представлена лінією через точки (0, k_2) і (k_2,0). Ця лінія перетинає $z_l = z_2$ в точці $(k_2 / 2, k_2 / 2)$. У всіх випадках чутлива область представлена зоною направо від розмежувальної лінії. Області чутливості для «правил (l,75) і (2,85)» ілюстровані на Малюнку 1А.

Що стосується «правила p-відсотка», вказана вище нерівність виділяє граничну лінію для чутливих комірок як лінія, що поєднується з точками (0,l00) і $\left(\frac{100}{\frac{p}{100} + 1}, 0\right)$. Ця лінія пересікає $z_1 = z_2$ у точці $\left(\frac{100}{\frac{p}{100} - 2}, \frac{100}{\frac{p}{100} - 2}\right)$. «Правило pq» є те ж саме, із q/p = l00/p.

МАЛЮНОК 1А

ПРИКЛАДИ ОБЛАСТЕЙ ПРИХОВУВАННЯ

«ПРАВИЛО (N,K)» ІЗ N=1 І K=75, N=2 І K=85



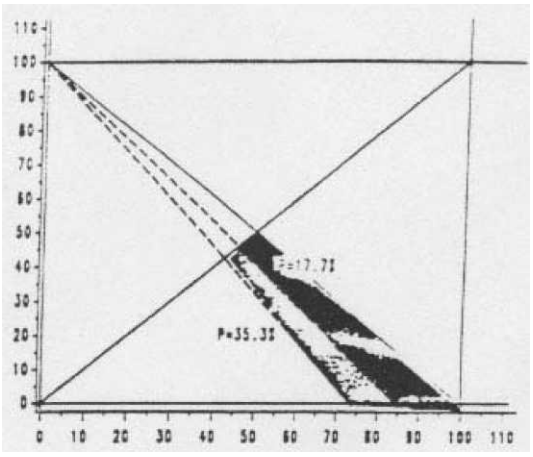
11. ДРУГИЙ ПО ВЕЛИЧИНІ ВІДСОТОК ВІД ПІДСУМКОВОЇ ВЕЛИЧИНИ

21. НАЙБІЛЬШИЙ ОПУБЛІКОВАНИЙ ВІДСОТОК ВІД ПІДСУМКОВОЇ ВЕЛИЧИНИ

МАЛЮНОК 1Б

ПРИКЛАДИ ОБЛАСТЕЙ ПРИХОВУВАННЯ

«ПРАВИЛО Р-ВІДСОТКА» ІЗ P=17,65 ВІДСОТКА, І P=35,3 ВІДСОТКА

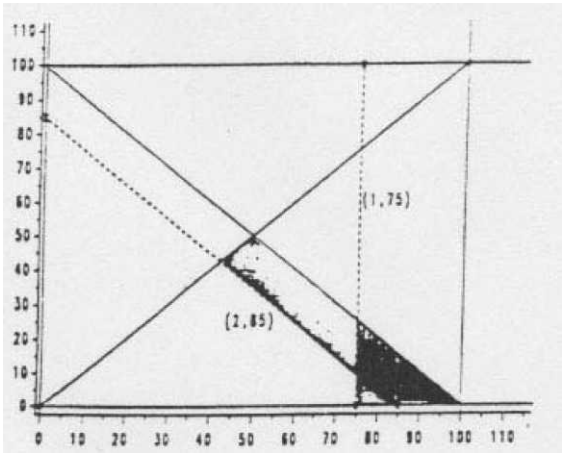


11. ДРУГИЙ ПО ВЕЛИЧИНІ ВІДСОТОК ВІД ПІДСУМКОВОЇ ВЕЛИЧИНИ

21. НАЙБІЛЬШИЙ ОПУБЛІКОВАНИЙ ВІДСОТОК ВІД ПІДСУМКОВОЇ ВЕЛИЧИНИ

МАЛЮНОК 1В

«Р-ВІДСОТОК» МЕНШ КОНСЕРВАТИВНИЙ, НІЖ (2,85), P=17,7 ВІДСОТКА

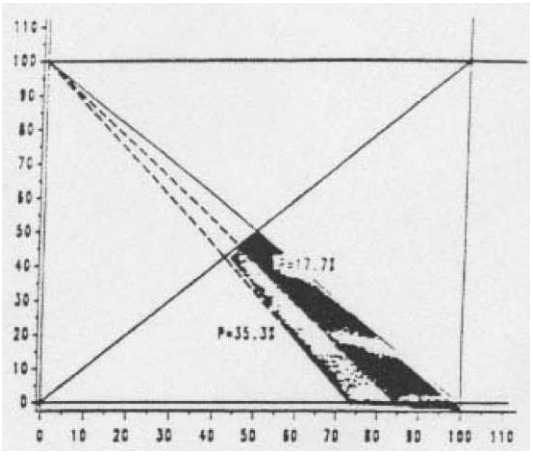


11. ДРУГИЙ ПО ВЕЛИЧИНІ ВІДСОТОК ВІД ПІДСУМКОВОЇ ВЕЛИЧИНИ

21. НАЙБІЛЬШИЙ ОПУБЛІКОВАНИЙ ВІДСОТОК ВІД ПІДСУМКОВОЇ ВЕЛИЧИНИ

МАЛЮНОК 1Г

«Р-ВІДСОТОК» МЕНШ КОНСЕРВАТИВНИЙ, І (1,1), P=35,3 ВІДСОТКА



11. ДРУГИЙ ПО ВЕЛИЧИНІ ВІДСОТОК ВІД ПІДСУМКОВОЇ ВЕЛИЧИНИ

21. НАЙБІЛЬШИЙ ОПУБЛІКОВАНИЙ ВІДСОТОК ВІД ПІДСУМКОВОЇ ВЕЛИЧИНИ

Примітка: Значення, що відповідають коміркам в таблиці, знаходяться в трикутнику, обмеженому лініями z1 = z2, z2 = 0, і z1 + z2 = 1. Значення, які відповідають чутливим коміркам, затемнюються.

Малюнок 1Б показує області чутливості для «правила p-відсотка» з $p = 17.65$ і $p = 35.29$. Обрання цих значень p будуть обговорюватись нижче. Прийміть до уваги, у випадку якщо $p = 0$, лінія чутливості припадає на верх лінії $z_1 + z_2 = 1$. В тій точці немає жодних чутливих комірок. Подібним чином, якщо p є негативним, то немає жодних чутливих комірок.

p% (n,k)

Визначити p так щоб S (X) ≤ S (X) для всіх комірок, X.

Розгляньте випадок, де «правило (n, k)» використовується і також існує вимога, щоб жоден внесок респондента не підлягав оцінюванню з точністю до p-відсотка від його значення. Ми б хотіли знайти значення p для того, щоб правило p-відсотка було найближчим до «правила (n, k)» із $S^{(n,k)}(X) \geq S^{p\%}(X)$. Таким чином, можуть існувати комірки, що класифікуються як чутливі згідно з «правилом (n, k)», які б не були чутливими згідно із «правилом p-відсотка», але всі комірки, класифіковані як чутливі згідно із «правилом p-відсотка» будуть класифіковані як чутливі згідно із «правилом (n, k)». Візьміть до уваги «правило (2, 85)», ілюстроване в Малюнку 1А. «Правило p-відсотка», найближче до «правила (2, 85)», яке відповідатиме вимогам, буде тим правилом, яке пересікає лінію $z_2 = 0$ у тій самій точці, що й «правило (2, 85)». Таким чином, що стосується цього значення k_2, ми повинні мати

$$S^{(n,k)}(X) > S^{p\%}(X)$$

Подібним чином, якщо б нам спочатку було надано значення p для «правила p-відсотка», ми повинні мати

$$S^{(n}\text{»}(X) > s^{p\%}(X)$$

Для «правила (2, 85)», $p/100 = 15/85 = .1765$, так щоб $p = 17.65$ відсотки. Малюнок 1В показує область чутливості (2,85) разом із менш консервативною областю $p = 17.65$ відсотків.

Для «правила (1, k_1)», «правило p-відсотка», яке найближче до «правила (1, 75)» і відповідає цій вимозі, буде тим правилом, що пересікає лінію $z_1 = z_2$ в точці (75, 75). Для даного значення k1 ми повинні мати

$$\frac{p}{100} = \frac{100}{k_1} - 2.$$

Подібним чином, якщо б нам спочатку дали значення p,

$$k_1 - \frac{100}{\frac{p}{100}} \cdot 2.$$

Із $k_1 = 75$, менш консервативне «правило p-відсока» буде мати $p = -66.7$, яке не призведе до жодного приховування комірки. Для $p = 17.65\%$, нам буде потрібно $k_1 = 45.94$, обмежене правило.

Знайти параметр p так щоб $S^{p\%}(X) \geq S^{(n,k)}(X)$ для всіх X.

Ми б хотіли знайти значення p так щоб правило p-відсотка було найближчим до «правила (n, k)» із

$S^{k)(n}(X)^{p\%} < S(X)$. Таким чином, можуть бути комірки, класифіковані як чутливі «правилом p-відсотка», і які не будуть чутливими згідно із «правилом (n, k)», але всі комірки, класифіковані як чутливі згідно із «правилом (n, k)» будуть класифіковані як чутливі згідно із «правилом p-відсотка». Ще раз, ми розглядаємо «правило (2, 85)» яке ілюстровано на Малюнку 1А. В цьому випадку найбільш необхідним консервативним «правилом p-відсотка» буде те правило, що пересікає лінію $z_1 = z_2$ у тій самій точці як і «правило (2, 85)». Беручи до уваги значення k_2, це приводить до

$$\frac{p}{100} = \frac{200}{k_2} - 2.$$

Якщо б нам спочатку було надано значення p, нам було б потрібно

$$k_2 = \frac{200}{\frac{p}{100} + 2}.$$

Для $k_2 = 85$, це складає $p/100 = 200/85 - 2 = .3529$. Малюнок 1Г показує область чутливості (2,85) разом із областю відсотка $p = 35.29$.

Для того, щоб знайти найбільш консервативне «правило p%», необхідне для включення меж чутливості «правила (l, k_1)», нам потрібне «правило p-відсотка», яке пересікає лінію $z_2 = 0$ у тій самій точці, що і «правило (l, k_1)». Беручи до уваги значення k_1 , це приводить до

$$\frac{p}{100} = \frac{100}{k_1} - 1.$$

Якщо б нам спочатку дали значення p, нам би потрібно було

$$k_1 = \frac{100}{\frac{p}{100} + 1}.$$

Для «правила (1,75)», це призводить до $p/100 = 25/75 = .3333$.

Для того, щоб знайти «правило (l,k_l)», яке проходить через таку ж точку як і «правило (2,85)» і «правило p-відсотка» із $p = 35.29\%$, замініть необхідне значення p у вказаному вище рівнянні та знайдіть $k_l = 73.91$.

У цьому випадку, ми розпочали із «правила (2,85)», яке приводить до $p = 35.29$, послідовно менш консервативним «правилом (l, k_1)» є те, що має $k_1 = 73.91$. Таким чином, «правило p-відсотка» із $p = 35.29$ забезпечує більше захисту, ніж «правило (2,85)» або «правило (1,73.91)». Таблиця 1 в тексті підсумовує ці результати для вибраних значень p, або еквівалентно для вибраних значень q/p.

Приклад

Прийміть до уваги три комірки нижче. Нехай x_1^k представляє найбільше значення, опубліковане респондентом в комірці k; а x_2^k представляє друге по величині значення, опубліковане респондентом в комірці k; і так далі. Туту ми припускаємо, що респонденти доповідають лише в одній з комірок 1, 2 або 3. Приналежність комірки позначається через верхній індекс k. Верхній індекс T представляє підсумкову величину.

	Комірка 1	Комірка 2	Комірка 3	Підсумкова величина
	$x_1^1 = 100$	$x_1^2 = 1$	$x_1^3 = 100$	$x_1^T = 100$
		$x_2^2 = 1$		$x_2^T = 100$
		$x_3^2 = 1$		$x_3^T = 100$
		$x_{20}^2 = 1$		$x_{20}^T = 100$
СУМА	100	20	100	220

Припустимо, що ми використовуємо «правило (n, k)» із $n = 2$ і $k = 85$ відсотки. Як описано вище, відповідні правила це «правило p-відсотка» з $p = 17.65$ (більш консервативне), «правило p-відсотка» із $p = 35.29$ (менш консервативне) і «правило (1,73.91)».

Використовуючи будь-яке із цих правил, Комірка 1 і Комірка 3 є виражено чутливими ($N = 1$, отже

S(X) > 0). Також легко засвідчити, що при використанні будь-якого розумного правила Комірка 2 не є чутливою. Ми розглядаємо дві комірки, об'єднання Комірки 1 і Комірки 2, а також Підсумкову величину.

Аналіз чутливості комірки для цих правил здійснюється наступним чином

$S^{(2,85)}$ (Комірка 1 U Комірка 2) = 100 +1 − 5.667 ·19 = -6.67

$S^{(17,6\%)}$ (Комірка 1 U Комірка 2) = 100 − 5.667 ·19 = -7.67

$S^{(1,73..91)}$ (Комірка 1 U Комірка 2) = 100 -2.833 ·20 = -43.34

$S^{(35.29\%)}$ (Комірка 1 U Комірка 2) = 100 − 2.834 ·19 = 46.16

$S^{(2,85)}$ (Підсумок) = 100 +100 − 5.667 ·20 = 86.66

$S^{(17,6o/o)}$ (Підсумок) = 100 − 5.667 ·20 = -13.34

$S^{(1,73.91)}$ (Підсумок) =100 − 2.833·120 = -239.96

$S^{(35,29\%)}$ (Підсумок) =100 − 2.834 ·20 = 43.32

Об'єднання Комірки 1 і Комірки 2 не є чутливим згідно із «правилом (2, 85)» і «правилом 17.65%». Проте, як «правило (1, 75)» а також «правило 33.3%» класифікують комірку як чутливу. Дивлячись на дані рівня респондента, наглядно раціональним є те, що об'єднання Комірки 1 і Комірки 2 є чутливим, навіть якщо правило вибору для цього прикладу, полягало в захисті лише проти домінування 2 найбільших опитуваних. Ця комірка відповідає точці (83.3, .008) на Малюнку 1.

Підсумкова величина чутлива для «правила (2, 85)» і «правила p-відсотка» із p=17.6%. Ця точка відповідає точці (45.5, 45.5) на Малюнку 1.

Візьміть до уваги несумісність у використанні самого «правила (2, 85)». В наведеному вище прикладі, якщо об'єднання комірки 1 і комірки 2 (нечутливі відповідно до «правила (2, 85»,) публікується, то найбільший респондент знає, що значення інших респондентів дорівнюють 20, і кожен з інших респондентів знає, що значення інших респондентів дорівнюють 119. Якщо підсумкова величина (чутлива згідно із «правилом (2, 85)») публікується, то кожен з двох найбільших респондентів знає, що сума значень інших респондентів дорівнює 120, і коженому із малих респондентів відомо, що сума значень іншого дорівнює 219.

Очевидно, буде здаватися, що більше інформації про дані респондента розголошується через публікування нечутливого об'єднання комірки 1 і комірки 2, ніж через публікацію чутливої підсумкової величини. Несумісність можна розв'язати використовуючи комбінацію «правил (n, k)», таких як (1, 73.91) і (2, 85), або використовуючи єдине «правило p-відсотка» із p = 35.29 або «правило pq» з q/p = 2.83. Ці зміни призводять до додаткових, але більш сумісних приховувань.

Прихильники простого «правила (2, 85)» стверджують, що необхідно більше захисту, коли

респонденти мають конкурентів із значеннями, близькими до своїх власних. Прихильники простого «правила (1, 75)» твердять, що необхідно більше захисту, якщо в комірці переважає єдиний респондент. Ці люди стверджують, що використання простого «правила (n, k)» дозволяє їм визначати те, які правила необхідні для їх спеціальних випадків без додаткових приховувань, які були б результатом більш сумісного підходу.

РОЗДІЛ V – Методи для файлів мікроданих публічного користування

Один із методів публікування інформації, що збирається в переписі чи опитуванні полягає в розголошенні файлу мікроданих **публічного користування** (див. Секцію 2.Г). Файл мікроданих складається із записів на рівні респондента, де кожен запис у файлі представляє одного респондента. Кожен запис складається із значень характерних змінних для цього респондента. Типові змінні для файлу з демографічними мікроданими це вік, раса, і стать опитуваної особи. Типові змінні для файлу мікроданих щодо установи представлені кодом Стандартної галузевої класифікації (SIC), кількістю робітників, вартістю поставок, що здійснюється комерційною діяльністю або галуззю, яка підлягає опитуванню. Більшість файлів мікроданих публічного користування містять лише демографічні мікродані. Ризик розкриття для більшості видів мікроданих з установи із значно вищим, ніж для демографічних мікроданих. Причини цього пояснюються в Секції В.4 цього розділу.

Цей розділ стосується файлів мікроданих, які є публічно доступними, і це є файли мікроданих **публічного користування**. На додачу до файлів публічного користування, деякі агентства пропонують файли мікроданих **обмеженого користування**. Доступ до цих файлів обмежено для певних користувачів в певних місцях розташування і регулюється договором обмеженого користування.

Для захисту конфіденційності мікроданих, агентства усувають всі очевидні ідентифікатори респондентів, такі як ім’я та адреса, із файлів мікроданих. Однак, все ще існує стурбованість, що розголошення файлів мікроданих могло б призвести до розкриття. Деякі люди і деяка комерційна діяльність та галузі в країні мають характеристики або комбінації характеристик, які призведуть до того, що вони будуть виділятись серед інших респондентів по файлу мікроданих. Файли мікроданих публічного користування містять деякий ступінь ризику розкриття конфіденційної інформації. Статистичне агентство, що розголошує файл мікроданих, який містить конфіденційних даних, має забезпечити можливість для мінімізації ризику, що сторонній користувач зовнішніх даних може правильно приєднати респондента до запису на файлі. Крім нерозголошення будь-яких мікроданих, не існує жодного способу усунення всіх ризиків розкриття із файлу; однак, агентства повинні

докладати всіх зусиль для мінімізації ризику і надалі розголошувати настільки корисної інформації, як це можливо.

Декілька Федеральних агентств, включаючи Бюро перепису населення, Національний центр статистики освіти, Національний центр медичної статистики, Центри для послуг «Медікер» і «Медікейд», Управління з інформації в області енергетики, Адміністрація соціального забезпечення, Бюро транспортної статистики, і Служба внутрішніх доходів розголошують файли мікроданих. Цей розділ описує ризик розкриття, пов'язаний з файлами мікроданих, математичними основами для вирішення проблеми, а також необхідні і строгі методи обмеження ризику розкриття.

А. Ризик розкриття мікроданих

Статистичні агентства займаються конкретним типом розкриття особистої інформації, що має відношення до респондента, існує декілька факторів, що відіграють важливу роль щодо ризику розкриття файлу мікроданих. Запис знаходиться під ризиком ідентифікації, якщо респондент є унікальним в базі даних по відношенню до набору ідентифікуючих змінних, і якщо зловмисник знає, що респондент перебуває на обліку. Постачальники даних, що підпорядковуються правилу приватності згідно із Актом захисту і права переказу медичної страховки (HIPAA) і/або Законом про захист конфіденційної інформації і статистичну ефективність (CIPSEA) повинні приймати схвальні міри для захисту конфіденційності опублікованих значень перед тим, як база даних розголошується як файл публічного користування.

А.1. Ризик розкриття і зловмисники

Більшість національних статистичних агентств збирають дані згідно із порукою конфіденційності. Будь-яке порушення поруки є розкриттям. Сторонній користувач, який намагається приєднати респондента до запису з мікроданими називається **зловмисником.** Ризик розкриття файлу мікроданих, значним чином, залежить від мотиву зловмисника. Якщо зловмисник шукає записи конкретних фізичних осіб чи фірм, є шанси, що ці фізичні особи, або фірми навіть не представлені у файлі, який містить інформацію про незначну вибірку населення. В цьому випадку, ризик розкриття цього файлу є невеликим, ризик є значно більшим, з іншого боку, якщо зловмисник намагається співставити *будь-якого* респондента з їх записом на зовнішньому файлі. Ми можемо вимірювати ризик розкриття лише в співставленні з конкретною методикою компромісу, яку, як ми припускаємо, використовує зловмисник (Келлер-Макналті, Макналті, і Унгер, 1989).

А.2. Фактори, що сприяють ризику

Існує два головних джерела ризику розкриття файлу мікроданих. Одне джерело ризику це існування записів з високим ступенем ризику. Деякі записи у файлі можуть представляти респондентів з унікальними характеристиками, такими як дуже незвична робота (наприклад, кінозірка, Федеральний суддя) або дуже великі доходи (наприклад, один мільйон доларів). Агентство повинно зменшувати видимість таких записів. Інший тип записів з високим ступенем ризику містить ті випадки, коли численні записи у файлі даних, як відомо, належать до тієї ж групи (наприклад, домашнє господарство або школа). У цьому випадку існує більший ризик, що кожен з них може бути ідентифікований (навіть якщо по суті не надається жодної інформації про групу). Третій тип записів з високим ступенем ризику може мати місце, коли один вимір даних розголошується з надто високим рівнем детальності. В цьому випадку, якщо дані розголошуються для маленьких зон, таких як шкільні райони, то змінні, які не створять проблеми розкриття на вищому рівні накопичення, такому як штат

чи область, можуть призвести до підвищеного ризику розкриття. Прикладом може бути – дохід вчителя на расу/етнічна приналежність і вік.

Друге джерело ризику розкриття, це можливість співставлення файлу мікроданих із зовнішніми файлами. Можуть бути фізичні особи, або фірми серед населення, які володіють унікальним сполученням характерних змінних по файлу мікроданих. Якщо деякі з цих фізичних осіб чи фірм обираються у вибірці населення, представленій в цьому файлі, то існує ризик розкриття. Зловмисники могли б потенційно використовувати зовнішні файли, що містять такі ж характерні змінні та ідентифікатори для приєднання цих унікальних респондентів до їх записів у файлі мікроданих.

Знання того, які фізичні особи брали участь в опитуванні, або навіть які галузі були у зразку, можуть значним чином допомогти зловмиснику ідентифікувати фізичних осіб у файлі мікроданих із опитування. Рекомендація респондентам бути розважливими, коли вони розповідають іншим про їх участь в опитуваннях в минулому, є доречною, але може змусити опитуваних насторожено відноситись до участі в опитуванні. Ризик розкриття файлу мікроданих значно зростає, якщо він містить адміністративні дані, або будь-який інший тип даних із зовнішнього джерела, пов'язаного з даними опитування. Ті, хто надає адміністративні дані, могли б використати ці дані для поєднання респондентів з їх записами у файл. Цим ми не маємо на увазі, що постачальники адміністративних даних будуть намагатись пов'язати файли, проте, існує така ймовірність і необхідно вжити запобіжних заходів. Крім того, у деяких випадках адміністративні дані можуть бути вже розголошені, як файл публічного користування, і тому будь-який зловмисник зможе використати інформацію для того, щоб спробувати ідентифікувати фізичну особу. Потенціал для поєднання файлів (і, таким чином, ризику розкриття) зростає, по мірі того як зростає число змінних, спільних для обох файлів, як зростає точність і вирішення даних, і як зростає число і доступність зовнішніх файлів, не всі з яких можуть бути відомі агентству, що розголошує файл мікроданих.

Поздовжні і панельні дослідження створюють спеціальний випадок ризику розкриття, який може бути пов'язаний з поєднаними файлами. У цьому випадку ризик розкриття файлу мікроданих зростає, якщо деякі записи у файлі розголошуються в іншому файлі з перекодуваннями, що є більш детальними, або які перекриваються (розділення на категорії), тих самих змінних. Таким чином, ризик зростає, якщо деякі записи у файлі розголошуються в іншому файлі, який містить деякі із таких змінних і деякі додаткові змінні.

Як наслідок, існує більший ризик, коли статистичне агентство чітко зв'язує новий файл мікроданих ряду респондентів із опублікованими даними для тих же респондентів у ранній період часу. Це має місце в поздовжніх дослідженнях, таких як Опитування доходу і участі в програмі, що проводиться Бюро перепису населення, де ті ж респонденти опитуються кілька раз, а також поздовжні дослідження середньої школи NCES, де за студентами спостерігають від 10 до 12 років протягом навчання в середній школі, продовженої середньої освіти, і до входження в трудові ресурси, або батьківства. Розмір ризику зростає, коли дані із різних періодів часу можна зв'язати для кожного респондента. Зміни, які зловмисник може, або не може побачити у записі респондента (такі як зміни професії чи сімейного статусу, або значні зміни доходу) поступово може призвести до розкриття особи респондента.

В загальному, ризик розкриття даних зростає в тій мірі, в якій структура даних стає більш складною – незалежно від того, чи це здійснюється через додавання зв'язаних даних із зовнішнього джерела, чи через додавання зв'язаних даних для ряду респондентів протягом певного часу, результат однаковий. Більш складна структура змінної, також призводить до зростання імовірності унікальних потоків відповідей на дані, і відповідно, до зростання імовірності розкриття.

A.3. Фактори, що зазвичай знижують ризик

Вибірка це важливий фактор для зниження ризику розкриття файлів мікроданих. Як ми стверджували раніше, якщо зловмисник володіє таким файлом мікроданих і шукає за записом конкретної фізичної особи або фірми, існує імовірність, що ця фізична особа або фірма навіть не представлені у файлі. Крім того, записи у такому файлі, які є унікальними в порівнянні зі всіма іншими записами у файлі, можуть не представляти респондентів із унікальними характеристиками в населенні. Можуть бути декілька інших фізичних осіб, або фірм в населенні з такими ж характеристиками, яких не обрали у вибірці. Це створює проблему для зловмисника, що намагається зв'язати файли.

Ризик розкриття файлу можна знизити навіть більше, лише якщо підвибірка від обраного населення представлена у файлі. Потім, навіть якщо б зловмисник знав, що фізична особа або фірма брала участь в опитуванні, він чи вона б не знали чи цей респондент був присутнім у файлі. Проте користувачі даними в загальному хочуть цілу вибірку.

Інший фактор, що виникає природним чином, який знижує ризик розкриття, це застарілість даних у файлах мікроданих, і будь-якихпотенційно підходящих зовнішніх файлів. Коли установа публікує файл мікроданих, дані у файлі зазвичай мають принаймні два роки. Характеристики фізичних осіб і фірм, можуть значно змінюватись з перебігом часу. Крім того, давність інформації у потенційно підходящих файлах імовірно відрізняється від давності даних у файлі мікроданих. Одне застереження полягає в тому, що відмінність у давності даних між файлами можуть не ускладнювати завдання щодо зв'язування старіших файлів, якщо зловмисник має доступ до зовнішнього файлу, який за часом відповідає збору даних.

Шум, що виникає природним способом у файлі мікроданих і в потенційно підходящих файлах, зменшує можливість зв'язувати файли. Всі подібні файли даних, будуть відображати мінливість доповідання, неотримання даних, і різноманітні методики редагування та умовного нарахування.

Багато потенційно прийнятних файлів мають декілька спільних змінних. Навіть якщо два файли володіють «такими ж» характерними змінними, часто ці змінні визначаються трохи по-іншому в залежності від мети збирання даних. Іноді змінні у різних файлах записуються по-різному. Визначення будь-яких змінних, що є спільними для обох файлів, повинні перевірятись для засвідчення, що визначення однакові, в іншому випадку змінні можуть фактично вимірювати різну діяльність. Відмінності у визначеннях змінної і перекодуваннях можуть зробити завдання зловмисника важчою.

Остаточними факторами, які знижують ризик є час, зусилля і гроші, необхідні для зв'язування

файлів, хоча із вдосконаленням комп'ютерної технології ці фактори слабшають.

А.4 Ризики розкриття, пов'язані із регресійними моделями

Питання щодо того, чи ризики розкриття існують в моделях регресійного типу стало більш важливим протягом останнього десятиліття, так як федеральні агентства розширюють доступ до своїх мікроданих. Ризики, пов'язані із файлами публічного користування зросли через підвищену обчислювальну потужність в поєднанні з розробкою ускладненого програмного забезпечення із співставлення даних і зростаючу доступність електронних баз даних в Інтернеті. Водночас, попит на доступ до файлів мікроданих зріс так як спільнота дослідників визнала цінність файлів, а підвищена обчислювальна потужність зробила аналіз файлів набагато простішим. У відповідь на ці розробки, агентства розробили декілька режимів обмеженого доступу до даних: Бюро перепису населення США взяло на себе лідерство у заснуванні Дослідницьких центрів зі збору даних (RDC); NCES використало ліцензійні угоди; NCHS розробило системи віддаленого доступу для користувачів та здійснення доступу до файлів мікроданих.

Національний науковий фонд США і NCES спільно фінансували роботу Національних інститутів статистичних наук США (NISS) для вивчення питань розробки «модельних серверів», які дадуть можливість дослідникам оцінити із баз даних конфіденційних мікроданих не маючи прямого доступу до мікроданих. Дослідники NISS дізналися як випускати корисні результати (наприклад, оцінки параметру регресії і діагностика моделі) не компрометуючи конфіденційну інформацію (Гоматам і співавтори, 2005). Вони також дослідили, як оцінювати регресії використовуючи комбінацію конфіденційних даних із декількох джерел; наприклад, із декількох статистичних агентств (Карр і співавтори, 2005).

Ризики розкриття можуть виникати із використання моделей регресії, особливо у стандартній моделі лінійної регресії, яка оцінюється з використанням Звичайного методу найменших квадратів, а також у логіт і пробіт-регресіях (які використовують двійкові (0,1) залежні змінні) та інші Узагальнені лінійні моделі (Резнек 2003, Резнек і Ріггс, 2004). Ризики в моделях регресії, що містять безперервні змінні з правого боку, є маленькими якщо загальна вибірка достатньо велика для проходження аналізу з табличного розкриття. Проте ризики можуть існувати в моделях, що містять фіктивні змінні в якості незалежних змінних. Коефіцієнти моделей, що містять лише повністю взаємопов'язані (насичені) множини фіктивних змінних з правого боку, можуть використовуватись для отримання вхідних даних в комбінаційних таблицях залежної змінної, де категорії комбінаційної таблиці визначаються фіктивними змінними. Ці типи комбінаційних таблиць можуть також виникати із кореляційних і коваріаційних матриць змінних, і з варіаційно-коваріаційних матриць коефіцієнтів моделі, якщо ці матриці включають фіктивні змінні. Ці результати досліджень представляють ризики розкриття, якщо комбінаційні таблиці представляють ризики розкриття.

Б. Математичні методи вирішення проблеми

Хоч було запропоновано декілька математичних ступенів ризику, жоден з них широко не використовувався. Методики, що знижують ризик розкриття мікроданих, охоплюють методи, що зменшують кількість інформації, яка надається користувачам даними, або методи, що трохи спотворюють інформацію яка надається користувачам даних. Декілька математичних мір корисності

наборів даних з обмеженням розкриття було запропоновано для оцінювання порівняльного аналізу між захистом і корисністю. І знову жодна з них широко не застосовувалась. Необхідно більше досліджень для ідентифікації найкращої методології, належної як для користувачів даними, так і для постачальників конфіденційних мікроданих.

Перед тим, як описувати ці математичні методи вирішення проблеми розкриття ризику, ми повинні згадати декілька проблем щодо математики і комп'ютерних наук, які у певний спосіб мають відношення до цієї проблеми. Наприклад, різноманітні методи співставлення файлу мікроданих із зовнішнім файлом можна знайти в літературі, що стосується методології переплетення запису за посиланням http://www.fcsm.gov/working-papers/RLT_1997.html. Методики переплетення запису, 1997 – Міжнародний семінар та експозиція процедур переплетення запису представляє нові видання головних довідкових матеріалів з переплетення запису, так як і обговорення поточної роботи.

Б.1. Запропоновані міри ризику

Вимірювання ризику розкриття файлів мікроданих публічного користування не виключає ймовірності, що зловмисник спроможний ідентифікувати запис. Більша частина дослідження розглянула деякі, або всі із наступних факторів:

- ймовірність, що респондент, за яким шукає зловмисник, представлений як у файлі мікроданих, так в деяких підходящих файлах,
- ймовірність, що змінні, які співставляються, записуються однаково як на файлі мікроданих, так і на відповідному файлі,
- ймовірність, що респондент, за яким шукає зловмисник, є унікальним серед населення на підходящі змінні, і
- ступінь впевненості зловмисника, що він або вона правильно ідентифікував унікального респондента.

Модель для вимірювання ризику розкриття повинна з впевненістю відображати попередні припущення про зловмисника. Рівень ризику варіюється залежно від того, чи зловмисник бажає розголосити опубліковані значення певного респондента, або опубліковані значення будь-якого респондента або групи респондентів. (Див. Стіл, 2004). Дійсність ступенів ризику залежить від точності позначення особою, що підготовила файл, переліку ключових змінних. Це набір змінних із файлу мікроданих, який може використовуватись для ідентифікації унікальних записів у файлі і також існує у даних, що знаходяться в публічній власності (або може утримуватись в приватності від деякого зовнішнього джерела комерційної інформації). Підрахунок частотності записів у файлі мікроданих зазвичай генерується з використанням перелік ключових змінних. Найбільш поширене правило, що застосовується при підготовці публічних файлів з мікроданими – це Правило граничної величини, або імені іменоване як «правило k-анонімності». Це правило вимагає мінімального числа записів, принаймні записів k, (зазвичай k=3), які є ідентичними по відношенню до визначеного набору ключових змінних. Це також використовується як міра ризику в «mu-ARGUS», програмному продукті, розробленому проектом Статистичної служби Нідерландів із Дослідження обчислювальних аспектів статистичної конфіденційності (CASC). (Див. веб-сайти в Додатку Б для подальшої інформації щодо CASC).

Відсоток записів, що представляють респондентів, які є унікальними серед населення, відіграє важливу роль у ризику розкриття файлу мікроданих. Ці записи часто називаються **унікальними записами населення.** Записи, які представляють респондентів, які є унікальними в порівнянні зі

всіма іншими у вибірці називаються **вибірковими унікальними записами.** Кожен унікальний запис населення є вибірковим унікальним записом, проте не кожен вибірковий унікальний запис є унікальним записом населення. Серед населення можуть бути інші особи, яких не було обрано у вибірці, і які мають такі ж характеристики як особа, представлена вибірковим унікальним записом. Робочий документ із статистичної політики 2 наголошує, що «унікальність в населенні є реальним питанням, і це не можна визначити без перепису або адміністративного файлу, що виснажує населення». Цей безпосередній наслідок залишається правильним для кожного окремого запису із вибіркового файлу мікроданих. Було розроблено декілька методів оцінювання відсотка унікальних записів населення у вибірковому файлі мікроданих. Ці методи основані на методиках взяття підвибірок, структурі класу еквівалентності зразка разом із гіпергеометричним розподілом, і моделюванні розподілу розмірів класу еквівалентності (Бетлеем, Келлер, і Паннекук, 1990; Стіл, 2004; Вінклер, 2004).

Ступінь відносного ризику для двох версій того ж файлу мікроданих було розроблено використовуючи класичну функцію ентропії із розподілу розмірів класу еквівалентності (Грінберг і Заяц, 1992). Наприклад, одна версія файлу мікроданих може мати кілька змінних із великою кількістю деталей по цих змінних, тоді як інша версія може мати багато змінних із малою кількістю деталей по цих змінних. Ентропія, що використовується як міра відносного ризику, може виділити які із двох версій файлу мають вищий ризик розкриття.

Б.1.а. MASSC.

Інша міра ризику, що використовується у методі обмеження розкриття «Мікронакопичення, заміна, взяття підвибірок, і калібрування» (MASSC) (обговорюється пізніше в Секції Б.3.г) створює набори ідентифікуючих змінних, що називаються шарами, для пошуку записів, які можуть знаходитись під ризиком розкриття. Унікальний запис у шарі представлений шаром, профіль якого є унікальним для даного набору ідентифікуючих змінних. Запис знаходиться під ризиком розкриття персональної інформації, якщо запис є унікальним поміж набором ідентифікуючих змінних. Після класифікації бази даних на серії шарів, представлених різними наборами ідентифікуючих змінних, міра ризику розкриття обчислюється для кожного шару. Унікальним записам, що відносяться до шару, потім приписується ризик розкриття, що має відношення до шару. MASSC обчислює чотири виміри ризику для генерування міри верхньої межі ризику розкриття для цільового запису, шару, або файлу. Міра ризику розкриття обчислюється на основі того, чи ціль виглядає як однозначний, неоднозначний подвійний, неоднозначний потрійний, і неоднозначний чотирьох-плюс, тобто неоднозначний розмір кластера чотирьох записів чи більше. Загальна міра цілі генерується обиранням середньозваженого значення із чотирьох мір ризику розкриття, де ваги є відносною пропорцією кожного типу запису у врегульованій базі даних. Стягуючи через шари, ризик розкриття можна обчислити як для всієї бази даних, так як і для окремого запису.

Б.1.б. Карта конфіденційності R-U.

Цей підхід намагається виміряти одночасний вплив застосування конкретної методики обмеження на ризик розкриття, і практичність даних, може служити інструментом постачальника даних для обирання відповідного значення параметру. R − це числова міра ризику статистичного розкриття в запропонованому розголошенні файлів даних. Це можна б було виміряти процентним відношенням записів, які можна безпомилково повторно ідентифікувати використовуючи програмне забезпечення

для переплетення записів. U − це числова міра практичності даних із розголошеного файлу. Це можна виміряти порівнюючи середні значення, або матрицю дисперсій і коваріацій оригінальних і збурених даних. Створюючи карту значень R і U на осях Y і X, генерується карта конфіденційності, яка показує порівняльні аналізи між здобутками, якщо такі є, у зменшенні ризику розкриття змінюючи параметри процедури обмеження розкриття, і втратах корисності даних через зміни в аналітичних якостях файлу. Карту конфіденційності R-U можна збудувати для різних методик обмеження розкриття і слугують в якості корисного інструменту при застосуванні конкретної методології обмеження розкриття. (Дункан, Макналті і Строукс, 2001).

Б.2. Методи зменшення ризику зменшуючи розмір розголошеної інформації

Перекодування змінних у категорії − це один із широко розповсюджених способів зменшення ризику розкриття файлу мікроданих (Скіннер, 1992). Кінцева інформація у файлі є не менш правильною, але вона менш точна. Це зменшення точності зменшує можливість зловмисника правильно приєднати респондента до запису, тому що зменшує відсоток унікальних записів населення у файлі. Перекодування змінних може також зменшити високий ризик деяких записів. Наприклад, якщо професія дуже детально описана у файлі, то запис, що показує професію Сенатора Сполучених Штатів в поєднанні з географічним ідентифікатором штату Делавер, вказує на одну із двох людей. Інші змінні по файлу ймовірно призведуть до ідентифікації цього респондента. Професію можна перекодувати у нечисленні, менш дискримінаційні категорії для полегшення цієї проблеми.

Якщо агентство конкретно хвилюється за зовнішній, потенційно підходящий файл, агентство може перекодувати змінні, спільні для обох файлів для того, щоб не було жодних комбінацій унікальної змінної і файлі мікроданих, таким чином запобігаючи однозначним співставленням. Наприклад, замість того щоб розголошувати повну дату народження, агентство може публікувати лише рік народження. Округлення значень, таке як округлення доходу до найближчої тисячі доларів, також є формою перекодування.

Іншим широко розповсюдженим способом зменшення ризику розкриття файлу є через встановлення верхніх кодів і/або нижніх кодів за безперервними змінними (див. Секція II.Г.2). **Верхній код** для змінної це верхня межа щодо всіх опублікованих значень цієї змінної. Будь-яке значення, більше ніж ця верхня межа не публікується у файлі мікроданих. На цьому місці знаходиться певний тип прапорця, що повідомляє користувача про те, який є верхній код, і що це значення перевищує його. Наприклад, замість публікування запису, який показує дохід в $2,000,000, запис може тільки показувати, що дохід становить > $150,000. Подібним чином, **нижній код** представляє собою нижній рівень щодо всіх опублікованих значень для змінної. Верхнє і нижнє кодування знижує високий ризик деяких записів. Прикладами змінних верхнього кодування можуть бути дохід і вік для файлів демографічних даних і значення поставок для файлів мікроданих установи. Якщо агентство опублікувало ці змінні по файлах мікроданих без жодного верхнього кодування, то ймовірно буде мати місце розкриття конфіденційної інформації. Прикладами змінних нижнього кодування можуть бути рік народження, або рік формування для якоїсь конкретної структури.

Перекодування і верхнє колування, безумовно, зменшують корисність даних. Проте агентства можуть приймати міри, медіани і варіації змінних в кожній категорії і значень верхнього кодування для користувачів даними для деякої компенсації втрат інформації. Крім того, перекодування і верхнє кодування можуть спричинити проблеми для користувачів тимчасових рядів даних, коли верхні

кодування та межі інтервалу змінюються з одного періоду в інший.

Б.3. Методи зменшення ризику порушуючи мікродані

Відколи Робочий документ із статистичної політики 2 було опубліковано, дослідники запропонували та оцінили декілька методів для порушення мікроданих для того, щоб обмежити ризик розкриття. Ці методології, описані в Розділі II, трохи змінюють дані в такий спосіб, щоб це заважало зловмиснику, який намагається співставити файли.

Ймовірно найбільш базовою формою порушення безперервних змінних це додавання, або множення на, довільні числа із даним розподілом. Цей **шум** можна додавати до записів даних в їх оригінальній формі, або до деякої трансформації даних, в залежності від цільового застосування файлу. Розподіл ймовірностей може застосовуватись для додавання помилок до малого відсотка категорійних значень. Агентство повинно вирішити, чи публікувати розподіл(и), що використовується для додавання шуму до даних чи ні. Публікація розподілу(ів) може допомогти користувачам даними в їх статистичних аналізах даних, але також можуть підвищити ризик розкриття цих даних. Інший запропонований метод порушення мікроданих полягає в довільному обиранні малого відсотка записів, і видаляти кілька значень у записах (див. Секція II.Г.5). Методики умовного нарахування також використовуються для того, щоб умовно нараховувати значення, які були затемнені.

Б.3.а. Перестановка даних

Перестановка (або **перемикання)** і **перестановка рангу** є двома запропонованими методами порушення мікроданих. Метою будь-якої методології перестановки полягає у запровадженні невпевненості для того, щоб користувач даних не знав чи реальні значення даних відповідають певним записам. Записи з високим ризиком розкриття зазвичай обираються для перестановки. Під час процедури перестановки, малий відсоток записів співставляється з іншими записами в тому ж файлі, можливо, в різних географічних областях, щодо набору попередньо визначених значень, які використовуються як параметри постановки. Значення змінних, що використовуються як параметри перестановки у файлі потім переставляються між двома записами. При процедурі перестановки рангу, значення безперервних змінних сортуються, а значення, які близькі за рангом, потім переставляються між парами записів. В міру того, як зростає процентне відношення переставлених записів, тим більшими є втрати корисності даних у файлі мікроданих. Хоча перестановка не змінює безумовний розподіл будь-якої змінної у файлі, вона все-таки викривлює спільні розподіли, що включають як переставлені, та і не переставлені змінні.

Б.3.б. Перетасування даних

Перетасування даних це інша процедура маскування, яку успішно застосовували до числових даних. Процедура включає в себе два кроки: по-перше, значення конфіденційних змінних модифікуються, і по-друге, процедура перетасування даних застосовується до конфіденційних змінних у файлі. Цей метод зберігає співвідношення рангового порядку між конфіденційними і не конфіденційними атрибутами, таким чином підтримуючи монотонні відносини між атрибутами.

Перед тим, як дані збурюються, неконфіденційні змінні **(S)** і конфіденційні змінні **(X)** у файлі ідентифікуються. Умовний розподіл $f(\mathbf{X}|\mathbf{S} = \mathbf{s_i})$ між конфіденційними і неконфіденційними змінними пізніше виводиться. Що стосується i = 1 до n, генеруйте вектор $\mathbf{y_i}$ із $f(\mathbf{X}|\mathbf{S} = \mathbf{s_i})$. Збурені значення **Y** представляють набір значень $\mathbf{y_i}$ (i = 1, 2, .. .n).

Перетасування записів даних має місце після того, як значення конфіденційної змінної були збурені і класифіковані. Для кожної конфіденційної змінної нехай $Y = (y_1, y_2, \ldots, y_n)$ представляють збурені

значення конфіденційної змінної X = (x_1, $x_{2...}$, x_n). Нехай X^j = (x^1, x^2, x^n) представляє значення впорядковані за рангом X.

Що стосується i = 1 до n: Знайдіть ранг y_i. Нехай цей ранг буде представлений як k. Замініть значення y_i значенням x^k. У вказаному нижче прикладі, візьміть до уваги, що ранг першого збуреного дослідження є 17. Значення 17-го впорядкованого значення X (x^{17}) = 42.79. В результаті, перше збурене спостереження замінюється 42.79. Таким чином, ранг спостереження y_2 це 16 і замінюється x^{16} = 41.74, значення X при 16-му ранзі X. Процес повторюється для кожного збуреного спостереження до тих пір, поки всі збурені значення не будуть замінені оригінальними значеннями із конфіденційної змінної.

Зразок набору даних

Ідентифікаційний номер#	S	X	Ранг X	Збурений Y	Ранг збуреного Y	Перетасований Y
1	41	54.24	27	43.8024	17	42.79
2	53	52.98	25	43.7608	16	41.74
3	40	33.77	4	31.2382	3	32.54
4	51	43.15	18	41.6440	13	40.41
5	37	48.70	22	36.3746	8	36.94
6	41	41.74	16	43.6570	15	40.77
7	24	36.00	7	46.5293	20	46.80
8	57	48.06	21	51.1033	23	48.76
9	52	57.69	29	54.3518	28	55.21
10	27	34.14	5	42.1101	14	40.72
11	39	32.54	3	40.6861	11	38.79
12	54	55.21	28	48.5196	22	48.70
13	52	40.77	15	53.7893	26	53.19
14	47	48.76	23	41.5140	12	39.50
15	41	27.52	1	44.6543	19	45.35
16	52	50.36	24	40.2965	10	38.68
17	20	42.79	17	34.6577	6	35.43

18	42	39.50	12	40.1456	9	38.05
19	52	53.19	26	51.5981	24	50.36
20	45	40.72	14	32.4994	4	33.77
21	52	38.68	10	47.7596	21	48.06
22	42	46.80	20	32.9835	5	34.14
23	50	59.08	30	44.4699	18	43.15
24	48	32.28	2	51.8446	25	52.98
25	33	36.94	8	35.7985	7	36.00
26	50	38.05	9	54.5523	29	57.69
27	46	40.41	13	25.2914	1	27.52
28	43	38.79	11	54.1997	27	54.24
29	56	45.35	19	54.7677	30	59.08
30	41	35.43	6	29.0405	2	32.28

Безумовний розподіл замаскованої (Перетасованої Y) змінної є таким, як і для оригінальної змінної X і кореляція зі змішаних моментів (лінійні відносини) і кореляція порядку за рангом (нелінійні монотонні відносин) не порушуються. В наданому прикладі кореляція між (S і X) є 0.4507 і між (Перетасовані Y і S) є 0.4474. Клореляція порядку за рангом між (S і X) є 0.52, а між (Перетасовані Y і S) є 0.54. Ці оціночні показними наблизяться один до одного як тільки зросте розмір даних.

Б.3.в. Спотворення даних і мікрона копичення

Спотворення охоплює накопичення значень серед маленьких наборів респондентів для обраних змінних і заміни опублікованого значення (або значень) в сукупності. Різні групи респондентів можуть формуватись для різних змінних даних, ставляючи інші змінні, або шукану розділяючи відповідну змінну (див. Секція II.Г.6). Записи розміщуються в групах розміру k, де k, зазвичай, встановлюється між 3 і 10, а оригінальні значення, що мають відношення до чутливих змінних, замінюються зведеним значенням. Дані можуть накопичуватись серед встановленого числа записів, довільно обраного числа записів, або числа, визначеного «правилами типу (n, k)» або «p-відсотка» як вони використовуються для зведених даних. Для визначення «правил (n, k)» і «p-відсотка» див. Розділ IV. Зведене значення, що має відношення до групи, може приписуватись всім членам групи або до «середнього» числа (як із ковзаючим середнім значенням). Накопичення у групах із 3 записів або менше, може бути недостатньо для зменшення ризику розкриття, особливо якщо спотворення виконується лише на одному чи двох змінних у файлі. По мірі того, як розмір групи зростає, шанс повторної ідентифікації зменшується. Якщо угрупування більше ніж 10 записів, у файл мікроданих можуть вводитись більші викривлення, що може призвести до неточно опублікованих даних. **Мікронакопичення –** це форма спотворення даних, де записи групуються на основі міри близькості всіх шуканих змінних, і ті ж групи записів використовуються при обчисленні зведених значень для цих змінних. Спотворення і мікронакопичення можуть здійснюватись в такий спосіб, щоб зберегти міри змінної. Проте спотворення даних окремої змінної, або мікронакопичення можуть призвести до повторної ідентифікації, і таким чином, повинні комбінуватись із іншими методиками обмеження розкриття для забезпечення належного захисту даних.

Інша запропонована методика порушення, функціонує з метою взяття надлишкової вибірки і підвибірки (Кокс і Кім, 1991). Оригінальні дані відбираються із заміною для створення файлу, більшого ніж запланований файл мікроданих. Диференційні ймовірності обирання використовуються для унікальних записів в наборі оригінальних даних, а також налаштовуються ваги запису. Цей більший файл потім підлягає підвибірці для створення кінцевого файлу мікроданих. Ця процедура заплутує ідею унікальності вибірки. Деякі унікальні записи видаляються через необирання, а деякі більше не здаються унікальними через дублювання. Деякі не унікальні записи виявляються унікальними завдяки відмові обирати свої копії (записи із такою ж комбінацією значень). Зміщення, введені цим методом можна обчислити і, можливо, розголосити користувачам в якості доповнення до файлу.

Б.3.г. Мікронагромадження, підстановка, взяття підвибірок, і калібрування (MASSC)

Мікронагромадження, підстановка, взяття підвибірок, і калібрування (MASSC) це методологія обмеження розкриття, що складається з наступних чотирьох головних кроків. Перший крок, мікронагромадження, розділює записи у шари ризику при підготовці для рівня модифікації даних, щоб зменшити ризик розкриття. Деякі перекодування змінних здійснюються протягом цієї фази. Фізичні особи у кожному шарі ризику групуються для того, щоб варіація була маленькою по відношенню до даного ключового набору ідентифікуючих змінних. При другому кроці, Підстановці, значення чутливих змінних переставляються на значення із записів, які є найближчими до них в плані певної міри відстані. При третьому кроці, взятті підвибірок, записи довільно обираються для взяття підвибірок у межах кожного шару. При четвертому кроці, калібруванні, ваги приписуються записам з використанням певних ключових змінних, для збереження доменних підрахунків з оригінального набору даних. Крок калібрування зменшує систематичну похибку завдяки заміні, і він зменшує коливання завдяки кроку здійснення підвибірки. В методології кожен запис в базі даних підлягає модифікації чи перестановці, проте, при застосуванні цієї методології лише маленька довільна кількість записів фактично модифікується. (Ю, Дантеман, Даі, і Вілсон, 2004).

Б.4. Методи зменшення ризику використовуючи симульовані мікродані

Б.4.а. Вибірка латинського гіперкубу.

Вибірка латинського гіперкубу (LHS) це інша методика, що включає в себе створення замінюючого файлу, що містить значення заміни для чутливих змінних у файлі мікроданих. Метод LHS забезпечує, щоб набір синтетичних даних мав приблизно такі ж одномірні статистичні характеристики оригінальних даних, таких як середнє, стандартне відхилення і коефіцієнт асиметрії. LHS можна використовувати для генерування набору синтетичних даних для групи некорельованих змінних. У випадку, коли змінні корельовано, алгоритм обмеженого спарювання, спочатку, застосовується для відтворення структури кореляції рангу реальних даних. Змінні спочатку перетасовуються у файлі, а кумулятивна функція розподілу створюється для обраних змінних і використовується для генерування синтетичних змінних. (Данкедар, Коен, Кіркендол, 2001). Вибірка латинського гіперкубу забезпечує один метод використання багатьох методик умовного нарахування для формування набору псевдо-даних із тими ж визначеними статистичними властивостями як правдиві мікродані.

Б.4.б. Синтетичні дані чинних логічних висновків.

Інша варіація у використанні синтетичних даних для розголошення файлів даних публічного користування здійснюється наведенням прикладів із наступного прогнозованого розподілу врегульованих порядкових даних. У цьому підході, фактична конфіденційна змінна(і) у файлі мікроданих, Y, замінюються використовуючи певний алгоритм обмеження контрольованого

регулювання даних. Початковий крок генерує передбачуване значення для Y, і залишок для кожної змінної Y 10 раз, що називається «заплутує». Статистичні моделі, що використовують дані, потім можуть вивести середнє значення результатів із десяти, спричиняють генерування оцінок середньоквадратичної помилки. В залежності від змінних, що потребують захисту і змінних, в яких зацікавлений дослідник, значення для конфіденційної змінної можна замінити подальшим прогнозованим розподілом для тієї конфіденційної змінної, основаної на даному наборі, або комбінаціях ключових змінних.

Налаштовуючи розподіл передбачуваного Y плюс залишки відповідної конфіденційної змінної, тобто, подальший прогнозований розподіл, можна створити різноманітні набори мікроданих, а статистичні логічні висновки із синтетичних даних є чинними із логічними висновками, що генеруються фактичними опублікованими значеннями. Численні файли публічного користування можна створити із тих самих ключових даних використовуючи цей метод із тим, щоб кожен файл публічного користування налаштовувався для різних груп користувачів. Методологія дійсних синтетичних даних, що логічно виводяться застосовувалась до даних Опитування з доходу та участі в програмі (SIPP) після того, як дані SIPP було зв'язано із даними щодо доходів з Адміністрації соціального забезпечення. (Абовд і Лейн, 2003).

Б.4.в. Алгоритм «FRITZ» для обмеження розкриття.

Система Федерального резерву з Методики логічного виведення «Zeta» (FRITZ) використовується як для умовного нарахування відсутніх даних, так і для обмеження розкриття в Опитуванні із споживчого кредиту (SCF). Модель FRITZ переглядає дані, разом із послідовним, попередньо визначеним шляхом і логічно виводить значення по-одному (інколи по двоє) на раз. Модель також є ітеративною через те, що вона умовно нараховує відсутні значення у файлі даних, і потім використовує цю інформацію як основу для умовного нарахування значень у другому кроці і продовжує процес до тих пір, поки всі значення відсутніх або чутливих оцінок не будуть стабільними і кінцевими. Файл переглядається на ключові змінні, що спричиняють надлишкові ризики розкриття і ці випадки обираються для захисту. Всі значення в доларах у SCF налаштовуються на відсутні значення, а алгоритм FRITZ застосовуються для генерування умовно нарахованих значень. Подальші аналізи цієї методології вказує на те, що логічні виведення забезпечували захист окремими чутливим записам, вони мали мінімальний вплив на розподільчі характеристики файлу (Кенікелл, 1998).

Б.5. Метод аналізування порушених мікроданих для визначення корисності

Існує декілька статистичних випробувань, які можна здійснювати для визначення наслідків порушення на статистичні властивості даних. Вони включають 2-вибірковий Критерій узгодженості Колмогорова-Смірнова, z-трансформацію Фішера із кореляції Пірсона, і Статистику наближення за критерієм хі-квадрат до тесту на критерій відношення правдоподібності для однорідності коваріаційних матриць.

Ці процедури в основному проводяться для того, щоб побачити чи засоби і варіаційно-коваріаційна і кореляційна структури даних залишаються такими ж після порушення. Навіть якщо ці випробування завершаться сприятливо, порушення все ще може мати негативний вплив на статистичні властивості, такі як засоби і кореляційна структура підмножин, а також вчасний серійний аналіз поздовжніх даних. Якщо агентство знає як файл буде використано, воно може порушити дані в такий спосіб, щоб статистичні властивості, доцільні для цього застосування, підтримувались. Проте файли публічного користування доступні для всієї громадськості, і вони використовуються багатьма способами. Рівні

порушення, необхідні для захисту даних від розкриття, можуть обробити кінцеву продукцію, непридатну для багатьох застосувань. З цієї причини, агентства обмежують кількість модифікацій даних у файлі мікроданих, або намагаються обмежити ризик розкриття обмежуючи кількість інформації у файлі мікроданих. Порушення може бути необхідним, проте, коли потенційно здатні до об'єднання файли доступні для користувачів, і спроби перекодування не усувають унікальних значень населення.

B. Необхідні процедури для розголошення файлів мікроданих

Перед тим як публічно розголошувати файли мікроданих, статистичне агентство повинно намагатись зберігати корисність даних, зменшувати видимість респондентів з унікальними характеристиками, і гарантувати, щоб файл не можна було приєднати до будь-яких зовнішніх файлів з ідентифікаторами. В той час коли немає жодного методу для повного усунення ризику розкриття файлів мікроданих, агентства повинні приймати наступні міри перед розголошенням файлу мікроданих для обмеження потенціалу файлу із розкриття. Статистичні агентства використовували більшість із цих методів протягом багатьох років. Вони продовжують бути важливими.

B.1. Видалення ідентифікаторів

Очевидно, агентство повинно очистити файл мікроданих від всіх прямих особистих та інституційних ідентифікаторів, таких як ім'я, адреса, номер Соціального забезпечення, та Ідентифікаційний номер роботодавця. Внутрішній файл з видаленими іменами чи іншими прямими ідентифікаторами все ще може знаходитись під ризиком **непрямого розкриття,** якщо відповідні дані залишаться у файлі, з яким можна співставляти інформацію із зовнішнього джерела, яке *також містить імена та інші прямі ідентифікатори.* У такому випадку особа, так само як і вся інформація у файлі, що має відношення до цієї особи або установи, буде розкриття якщо файл розголошується без подальших модифікацій.

B.2. Обмеження географічних деталей

Співпадіння не повинне бути точним. Зловмисник може зв'язати характеристики всіх респондентів із тією ж одиницею вибірки з подібною інформацією із зовнішнього джерела даних. Інші змінні у файлі можуть спричинити проблему непрямого розкриття, якщо їх можна використати для визначення малого географічного об'єкту на основі певних соціально-економічних характеристик. Коли індивідуальні записи, або записи по установі вже мають відношення до маленької географічної зони, імовірність ідентифікації значно зростає. Географічне місце розташування є характеристикою, що є у більшості файлів з мікроданми. Агентства повинні приділяти особливу увагу географічним деталям перед розголошенням файлу мікроданих, тому що для зловмисника значно простіше зв'язувати респондента із його записом, якщо зловмисник знає місто респондента, наприклад, швидше ніж якщо він чи вона лише знає штат респондента.

На основі цих обговорювань, Бюро перепису населення не ідентифікує будь-якої географічної області з менш ніж 100,000 особами в основі вибірки. Порівняльний аналіз вищого рівня використовується для опитувань із передбачуваним вищим ризиком розкриття. Файли мікроданих із Огляду доходу та участі в програмі, наприклад, все ще мають порівняльний аналіз з 250,000 осіб на визначену область. Агентства, що розголошують файли мікроданих повинні встановлювати географічні порівняльні аналізи, що мають просто нижчі межі щодо розміру відібраного населення кожної географічної області, ідентифікованої у файлах мікроданих. Це простіше сказати, ніж зробити. Рішення цього типу часто основані на прецедентах і простих судженнях. Необхідно більше досліджень для забезпечення наукових підстав для таких рішень (Заяц, 1992a).

Деякі файли мікроданих містять контекстуальні змінні. Контекстуальні змінні – це змінні, що описують область, в якій респондент чи установа перебувають, але не ідентифікують цієї зони. В

загальному описані зони є меншими ніж зони, що зазвичай ідентифікуються у файлах мікроданих. Потрібно проявляти обережність для гарантії, щоб контекстуальні змінні не ідентифікували зони, що не відповідають необхідному географічному порівняльному аналізу. Прикладом контекстуальної змінної, що може призвести до розкриття, є середня температура зони. Управління з інформації в області енергетики додає довільний шум (тому що дані температури є широко доступними) і надає формулу, так щоб користувач міг обчислити приблизні градусо-дні опалювального сезону та охолоджуючі градусо-дні (важливо для регресивного аналізу споживання електроенергії).

B.3. Змінні верхнього кодування високого ризику, що є безперервними

Змінні по файлах мікроданих, що сприяють високому ризику певних респондентів, називаються **змінними високого ризику.** Приклади безперервних змінних високого ризику – це змінні доходу і віку для файлів демографічних мікроданих і значення поставок для файлів мікроданих по установі. Як було вказано раніше, якщо агентство опублікувало ці змінні у файлі мікроданих без верхнього кодування, то ймовірно буде мати місце розкриття конфіденційної інформації. Наприклад, зловмисники можуть імовірно правильно ідентифікувати респондентів, які є віком поза 100 чи які мають доходи більше одного мільйона доларів.

Належні верхні коди (і/або нижні коди у деяких випадках) повинні бути встановлені для всіх безперервних змінних високого ризику у файлі мікроданих. Записи із верхнім кодуванням повинні в такому випадку показувати лише репрезентативне значення для верхнього хвоста кривої розподілу, таке як малозначима величина для хвоста чи середнього або серединного значення для хвоста, в залежності від бажань користувача. Енгл (2003) розробив методологію для оцінювання розподілу значень верхнього кодування, використовуючи більш загальний розподіл ніж традиційний Парето, та ілюструє це використовуючи щорічні трудові доходи. Оцінка моделлю правого хвоста, відсіченого верхнім кодуванням, продемонструвала, що вона має багато динаміки правих хвостів щорічних емпіричних розподілів трудових доходів. Ця методологія використовує модель густини імовірності для генерування правого хвоста розподілу доходу, який було відсічено верхнім кодуванням. Параметри моделі оцінюються за її відповідністю до даних, нижче зрізу верхнього кодування. Правий хвіст моделі використовується в оцінюванні статистики всього розподілу. Модель спроможна генерувати розподіл значень верхнього кодування навіть після зниження граничного рівня для мінімального щорічного трудового доходу, що підлягає верхньому кодуванню, значно нижче 99-го процентиля (Енгл, 2003).

B.4. Запобіжні заходи для певних типів мікроданих

Існують певні типи мікроданих, що можуть підвищити ризик розкриття при перегляді файлу для розголошення.

B.4.a. Мікродані по установі

Більшість файлів мікроданих, що публічно розголошуються, містять демографічні мікродані. Передбачається, що ризик розкриття для мікроданих по установі є вищим ніж для демографічних мікроданих. Дані по установі типово спотворені, розмір генеральної сукупності установи може бути малим, а також існує багато змінних високого ризику файлах мікроданих потенційної установи.

Галузеві публікації і торгівельні асоціації також можуть існувати і функціонувати як зовнішні джерела інформації для користувача даними. Публічно доступні адміністративні бази даних також можуть бути доступними для співставлення із файлами мікроданих по установі і створювати додаткові ризики розкриття. Крім того, існує також велика кількість спеціалістів в конкретній області в багато можливих мотивів для здійснення спроб ідентифікації респондентів по деяких типах файлів мікроданих по установі. Наприклад, може існувати матеріальна зацікавленість, пов'язана із здобуттям інформації про конкуренцію. Агентства повинні взяти до уваги всі із цих факторів обдумуючи розголошення файлів мікроданих по установі.

B.4.б. Поздовжні мікродані

Існує більший ризик, коли мікродані у файлі беруться із поздовжнього опитування, де ті ж респонденти опитуються кілька разів. Ризик збільшується, коли дані з різних періодів часу можна поєднати для кожного респондента, тому що існує набагато більше даних для кожного респондента і тому що зміни, які можуть або не можуть мати місце у записі респондента протягом певного періоду можуть призвести до розкриття особи респондента. Агентства повинні прийняти це до уваги коли обдумують розголошення такого файлу. Порада полягає в тому, щоб ви планували заздалегідь. Розголошення першого перехресного файлу не задумуючись над майбутніми планами для поздовжніх файлів може спричинити непотрібні проблеми, коли приходить до розголошення останніх. Ціла програма збору даних повинна обговорюватись з огляду на винесення рішень щодо розголошення мікроданих публічного користування.

B.4.в. Мікродані, що містять адміністративні дані

Ризик розкриття файлу мікроданих зростає, якщо він містить адміністративні дані або будь-який інший тип даних із зовнішнього джерела, приєднаного до даних опитування. Ті, хто забезпечує адміністративні дані, можуть використати ці дані для приєднання респондентів до їх записів. Це не означає, що постачальники адміністративних даних будуть намагатись приєднати файли, проте, існує така теоретична можливість і необхідно прийняти запобіжні заходи. В крайньому випадку необхідно здійснити деякий тип порушення в адміністративних даних, або ж адміністративні дані повинні бути розділені на категорії для того, щоб не існувало жодних унікальних комбінацій адміністративних змінних. Це зменшує можливість того, що зловмисник зможе знайти зв'язок між файлом мікроданих та адміністративним файлом. Існують занепокоєння, що агентства взагалі не повинні розголошувати такі мікродані або ж повинні розголошувати їх лише згідно із договором обмеженого доступу.

B.4.г. Розгляд потенційно сумісних файлів та унікальних значень населення

Статистичні агентства повинні намагатись ідентифікувати зовнішні файли, які є потенційно сумісними із файлом мікроданих, що розглядається. Співставність подібних файлів із обговорюваним файлом, має бути досліджена. Бюро перепису населення використовує повторну ідентифікацію та експерименти із поєднання записів для визначення, чи їх файли підлягають співставленню із зовнішніми файлами щодо певного набору ключових змінних. Національний центр статистики освіти співставляє файли мікроданих, що обговорюються, для розголошення комерційно доступним навчальним файлам для ідентифікації унікальних співставлень. Повторна ідентифікація мікроданих має відношення до спроможності використання публічно доступної інформації для додавання імен, адрес, та інших частково унікальних ідентифікаторів до індивідуальних записів у файлі публічного

користування. Ідентифікатор частково унікальний, якщо він може використовуватись в поєднанні з іншими змінними, для повторної ідентифікації запису, навіть якщо він не може точно ідентифікувати зв'язок між двома записами сам. Програмне забезпечення для пов'язування записів було розроблено для керування великою розмаїтістю як незначних, так і основних варіацій написання і помилками у змінних, що використовуються в процесі співставлення.

Інша міра ризику повторної ідентифікації для файлу – це число або пропорція унікальних значень населення, де розгляд обмежений до тих змінних, які вважаються доступними в зовнішніх файлах. Було розроблено статистичні моделі, що відносять розподіл унікальних значень за вибіркою у файлі до розподілу унікальних значень по населенню. Однак, ці моделі лише надають оцінку для процентного відношення унікальних значень за відсотком, які є правдивими унікальними значеннями за населенням. Ця оцінка має тенденцію до високої розбіжності, і оцінювання процентного відношення не надає будь-якої вказівки до визначення того, які унікальні значення є артифактами вибірки, а які є унікальними значеннями із населення. Експерименти із поєднання запису можуть також надати міру ризику повторної ідентифікації, але значно це залежать від здобування або моделювання джерел зовнішніх даних (Вінклер 2004). Експеримент із поєднання записів може ідентифікувати деякі унікальні значення із населення, але не повинен вважатись запевненням, що всі ризиковані записи було виявлено.

Г. Строгі методи обмеження ризику розкриття

Існує декілька процедур, які можна виконувати із файлами мікроданих перед розголошенням і які строго обмежують ризик розкриття файлів, такі як перестановка та збільшення даних. Проте необхідно мати на увазі, що корисність отриманих опублікованих даних також буде надзвичайно обмежена. Отримані файли будуть містити або набагато менше інформації, або ж інформацію, яка є неточною до тієї міри, що залежить від файлу і його вмісту.

Г.1. Не розголошуйте мікродані

Один очевидний спосіб усунення ризику розкриття мікроданих полягає в тому, щоб не розголошувати записи з мікроданими. Статистичне агентство змогло розголосити лише дисперсійно-коваріаційну матрицю даних або можливо визначений набір кінцевих моментів даних нижчого порядку. Це значно зменшує корисність даних, тому що користувач отримує набагато менше інформації, а аналізи даних обмежуються.

Г.2. Перекодуйте дані для усунення унікальних значень

Перекодування даних у такий спосіб, щоб жодні унікальні значення вибірки не залишались у файлі мікроданих в загальному вважається належним методом для обмеження ризику розкриття файлу. М'якша процедура, що передбачає ширший розподіл на категорії − перекодування, де немає жодних унікальних значень за населенням – буде достатньою. Перекодування даних для усунення або вибірки, чи унікальних значень за населенням, ймовірно призведе до дуже обмеженої опублікованої інформації.

Г. 3. Порушення даних для запобігання співставлення із зовнішніми файлами

Наголошування на тому, що файли, які містять порушені мікродані, не можуть бути успішно співставлені з оригінальним файлом даних, або з іншим файлом порівняльними змінними, в загальному вважається достатнім свідченням належного захисту. Потрібно використати декілька мір близькості під час спроб поєднання двох файлів. Альтернативна демонстрація належного захисту полягає в тому, щоб жодне точне співставлення не було правильним, або щоб правильне співставлення для кожного запису у порівняльному файлі не знаходилось поміж найточніших співставлень K. Мікронакопичення або перестановку даних можна використовувати для захисту даних, можливо застосовуючи «правила типу (n, k)» або «p-відсотка», як це використовується для таблиць. В такий спосіб, жодних індивідуальних даних не надається, а зловмисникам буде заважати співставляти дані із зовнішніми файлами. Див. Розділ IV для визначення «правил (n, k)» і «p-відсотка». Мікронакопичення, спотворення даних, та інші методи порушення, що ускладнюють співставлення, тим не менше, можуть спричинити викривлення опублікованих даних. Якщо застосувати ці методи до такої степені, що буде абсолютно запобігати співставленню, вони звичайно призведуть до значним чином викривленої опублікованої інформації.

Д. Висновки

Файли мікроданих публічного користування використовуються для різноманітних цілей. Будь-яке розкриття конфіденційних даних у файлах мікроданих може містити порушення закону, або політики агентства і може ускладнити спроможність агентства збирати дані в майбутньому. За винятком випадків повної відсутності випуску інформації, немає жодного способу повністю усунути ризик розкриття. Проте існують методики, які, якщо їх здійснювати на даних до розголошення, повинні належним чином обмежувати ризик розкриття файлу мікроданих. Необхідне дослідження, щоб краще зрозуміти наслідки цих методик для ризику розкриття і для корисності отриманих файлів даних (див. Секція VI.A.2).

РОЗДІЛ VI – Рекомендована практика для федеральних агентств

А. Вступ

На основі свого перегляду поточної практики агентства і відповідного дослідження, Комітет з конфіденційності і доступу до даних (CDAC), підкомітет FCSM, розробили набір рекомендацій для процедур обмеження розкриття. Впровадження цих процедур федеральними агентствами призведе до загального зростання захисту від розкриття і покращить розуміння і простоту використання результатів федеральних досліджень даних, обмежених для розкриття. Іноді методи, що використовуються для зменшення ризику розкриття, роблять дані, що невідповідають для статистичного аналізу (наприклад, як згадано в Розділі V, перекодування може спричинити проблеми для користувачів даними тимчасового ряду, коли верхні кодування змінюються від одного періоду до наступного). При вирішенні які статистичні процедури використовувати, агентствам також необхідно

прийняти до уваги корисність отриманих результатів досліджень даних для користувачів даних.

Перший набір рекомендацій в Секції Б.1 є загальним і відноситься як до таблиць, так і до мікроданих. Секція Б.2 описує рекомендації CDAC для таблиць частотних даних. Рекомендації від 7 до 11 в Секції Б.3 відносяться до таблиць порядкових даних. І на завершення, рекомендації 12 і 13 в Секції Б.4 відносяться до мікроданих.

Б. Рекомендації
Б.1. Загальні рекомендації для таблиць мікроданих

Рекомендація 1: Звертайтесь за порадою від респондентів і користувачів даними. Для того, щоб планувати та оцінювати політику і процедури обмеження розкриття, агентства повинні проконсультуватися як з респондентами, так і користувачами даних. Агентства повинні шукати більшого розуміння того, як респонденти відкликаються про ризики розкриття даних, обмін даними серед агентств, доступності співставлення із зовнішніми адміністративними файлами даних, і захист даних згідно із опитуваннями CIPSEA та інших ніж CIPSEA.

Таким чином, агентства повинні консультувати користувачів даними з питань, що мають відношення до: врівноваження ризику розкриття на противагу втраті корисності даних; підвищуючи доступність файлів мікроданих публічного користування; потреба в процедурах обмеженого доступу до даних, для того, щоб дослідники могли здійснювати доступ до мікроданих в контрольованому і безпечному середовищі, і розробка бази даних публічного користування для інтерактивних систем опитування через Інтернет. Інші питання, що впливають на корисність даних, полягають в тому, чи користувачі швидше дадуть перевагу методам обмеження розкриття, що модифікують, замінюють чи регулюють дані у деякий спосіб, швидше ніж методи, що приховують дані.

Рекомендація 2: Стандартизація і централізація перегляду агентством результатів досліджень даних, обмежених для розкриття. Важливим є те, щоб політика і процедури обмеження розкриття окремих агентств були внутрішньо сумісними. Результати процедур з обмеження розкриття повинні переглядатись. Агентства повинні стандартизувати процес перегляду прийнявши стандарти і/або норми із захисту конфіденційності даних. «Контрольний перелік з потенціалу розкриття запропонованих розголошень даних», доступний за посиланням http://www.fcsm.gov/committees/cdac/ повинен використовуватись як довідник для цього процесу перегляду. Контрольний перелік повинен бути змінений таким чином, щоб співпадати із політикою і процедурами агентства щодо розголошення даних. Агентства повинні також централізувати відповідальність за цей перегляд в організаційній структурі через механізми, такі як наглядові ради із розкриття (постійні чи спеціальні), або посадова особа з конфіденційності, група експертів, або група робочого персоналу, обізнана і досвідчена в області процедур обмеження розкриття і захисту конфіденційності даних.

CDAC рекомендує, щоб агентства ознайомились із зовнішніми базами даних, які є доступні користувачам для співставлення із результатами досліджень даних агентства. Вони повинні оцінювати будь-яке запропоноване розголошення даних, як в плані внутрішніх ризиків розкриття для змінних і значень всередині файлу, так і в плані зовнішніх ризиків розкриття через потенційне співставлення із зовнішніми файлами. У агентствах із маленькими, або єдиними програмами для

розголошення мікроданих, це можна приписати до однієї фізичної особи, обізнаної в методах статистичного обмеження розкриття і політикою конфіденційності агентства. В агентствах з численними чи великими програмами, наглядова рада повинна формуватись з відповідальністю перегляду кожного файлу мікроданих, запропонованого для розгляду, і визначати чи він підходить для розголошення. Наглядова комісія повинна: якнайширше представляти програми агентства, як це тільки можливо; бути обізнаною в методах обмеження розкриття для мікроданих; бути готовою рекомендувати і сприяти використанню методологій обмеження розкриття керівниками проекту, і їх агентство повинно надати їм повноваження для підтвердження, що ці методики обмеження розкриття були належним чином застосовані.

Результати досліджень агентств в табличних даних також повинні переглядатись. Обмеження розкриття і приховування повинні бути підконтрольним і відтворюваним процесом. (Обмеження розкриття для мікроданих на даний час не знаходиться на стадії, де подібний підхід є прийнятним). Існує ефективність адміністративного керівництва для централізації перегляду як файлів мікроданих, так і табличних файлів. В залежності від інституційного розміру, програм, і культури, агентство повинно комбінувати перегляд мікроданих і таблиць окремою фізичною особою, наглядовою радою чи офісом.

Рекомендація 3: Обмінюйтесь програмним забезпеченням і методологією серед керівництва.

Федеральні агентства повинні обмінюватись програмними продуктами для обмеження розкриття і зв'язування записів, так як і методологічними і технічними вдосконаленнями. Конкретно, CDAC повинно продовжувати створювати програмне забезпечення для методологій обмеження розкриття і документацію, доступну на його веб-сайті федеральним агентствам і громадськості для користування. Програмне забезпечення має бути написаним на звичайній мові для обробки даних, яку легко модифікувати за допомогою зрозумілої документації.

Коли розробляються вдосконалення до програмного забезпечення для обмеження статистичного розкриття і зв'язування записів науковими співтовариствами, урядом, і приватними суб'єктами підприємницької діяльності, CDAC повинно оцінювати ці нові методології і програмне забезпечення, і надавати інструкції для федеральних агентств щодо практичних і належних прикладних програм для користування. CDAC має програмне забезпечення, доступне на веб-сайті за посиланням http://www.fcsm.gov/committees/cdac/, яке здійснює першочергове і додаткове приховування, а також програмне забезпечення для аудиторської перевірки приховування, яке переглядає і генерує звіт, що вказує на застосований ступінь захисту, із моделі приховування, яка використовується в таблиці.

Рекомендація 4: Для обміну даними потрібна формальна міжвідомча співпраця. Обмін файлами даних між агентствами вимагає формалізованих договорів між агентствами для того, щоб гарантувати захист конфіденційності даних і відповідати правовим зобов'язанням агентства щодо збору і публікації інформації. Розголошення ідентичних чи подібних даних різними агентствами або групами в межах агентств (або з ідентичних, або з подібних наборів даних), або спроможність співставлення із зовнішніми файлами є іншими факторами, що сприяють потребі у міжвідомчій співпраці. Міжвідомчі ради або команди можуть бути потрібними для планування і перегляду діяльності з обміну даними між агентствами. Міжвідомча співпраця з перегляду наборів даних, що перекриваються, і використання ідентичних процедур з обмеження розкриття заохочується. Агентства повинні розширювати спільне використання дослідницьких центів із збору даних в якості

методу для підвищення доступу дослідників до конфіденційних даних. Агентства можуть також розглянути варіант подачі запиту представникам з інших агентств, що мають більше досвіду із розголошення файлів мікроданих публічного користування, для обслуговування наглядових рад із розкриттям для того, щоб знаннями і досвідом можна було ділитись серед агентств.

Рекомендація 5: Використання належної практики. Агентства повинні прагнути застосовувати методи обмеження розкриття в стандартні способи, і бути послідовними при визначенні категорій в різних результатах досліджень даних протягом певного періоду. Вони повинні стандартизувати визначення змінної внутрішньо до тієї міри, в якій вона відповідає потребам програми агентства, а загальні визначення між агентствами повинні розроблятись там, де це можливо. Такі порядки покращать доступ до даних для громадськості, і вимагатимуть запровадження методологій обмеження розкриття. Приклади складають використання сумісних схем для комбінування категорій, встановлення стандартизованих процедур для подібних даних, таких як розподіл за категоріями або змінні верхнього кодування, такі як вік або дохід, і рух в напрямку стандартизованого застосування мінімальних обмежень географічного розміру для даних з домашнього господарства. Програмне забезпечення потрібно розробити, зробивши його широкодоступним, і використовувати для запровадження цих методів для гарантії як послідовності, так і правильного запровадження.

Б.2. Таблиці даних з підрахунку частотності

Рекомендація 6: Необхідне дослідження для порівняння та оцінки методів.

Значна кількість досліджень здійснювалась щодо методів обмеження розкриття для таблиць частотних даних. Найбільш поширеним методом, що використовується, є приховування. Крім приховування, інші добре розроблені методи, що є доступними, складають контрольоване округлення, контрольоване табличне врегулювання, і застосування методів збурення даних перед зведенням у таблиці. Необхідно провести додаткові дослідження для застосування цих методів до різних типів даних, а також порівняти та оцінити ці різні методи в плані захисту даних і корисності отриманих результатів досліджень даних. Якщо використовується приховування, то керівні вказівки, перелічені в Рекомендаціях 9 і 10, також застосовуються до таблиць частотних даних.

Б.3. Таблиці порядкових даних

Рекомендація 7: Використовуйте лише субадитивні правила розкриття для ідентифікації чутливих комірок.

Агентства повинні розробляти і застосовувати операційні правило лінійної чутливості (Див. Розділ 4) для ідентифікації і потім захисту комірок первинного розкриття. Правила розкриття, що мають математичну властивість **субадитивності** надають гарантію, що комірка, сформована комбінацією двох нечутливих комірок, залишається нечутливою. Агентства повинні застосовувати лише субадитивні правила первинного розкриття. «Правила p-відсотка, pq, N, та (n, k)» є всі субадитивними. **Комірки первинного розкриття** повинні захищатись, використовуючи методики обмеження розкриття.

Рекомендація 8: Надається перевага «правилам p-відсотка і pq-неоднозначності».

Рекомендуються «правила p-відсотка і pq-неоднозначності», тому що використання лише «правила (n, k)» є несумісним за кількістю інформації про респондентів, дозволеної для виведення (див. Розділ IV). «Правила p-відсотка і pq» дійсно надають належний захист всім респондентам. Особливо, «правило pq» потрібно використовувати, якщо агентство може кількісно визначити міру, в якій користувачі даних знають щось про значення респондента. Якщо, тим не менше, агентство вважає, що респонденти потребують додаткового захисту від близьких конкурентів у межах тих самих комірок, вони можуть використовувати «правила p-відсотка» або «pq» в поєднанні з «правилом (n, k)». При використанні лише «правила (n, k)», послідовність «правил (n, k)» краще ніж один набір параметрів. Прикладом послідовності «правил (n, k)» є (1,75) і (2,85). Коли застосовується комбінація «правил (n, k)», комірка чутлива якщо вона порушує будь-яке із правил.

Рекомендація 9: Не розголошуйте параметри приховування. Для сприяння розголошенню так багато інформації, як це тільки можливо на прийнятних рівнях ризику розкриття, агентства заохочують оприлюднювати тип правила, яке вони використовують (наприклад, «правило p-відсотка»), але вони не повинні оприлюднювати конкретне(і) значення правила обмеження розкриття (наприклад, точне значення «p» в «правилі p-відсотка»), оскільки такі знання можуть зменшити захист від розкриття. (Див. Розділ 4 Секцію Б.4 для ілюстрації того, як обізнаність із правилом та значенням параметру може надати можливість користувачеві зробити логічне виведення значення прихованої комірки). Значення параметрів, що використовується для обмеження статистичного розкриття, може залежати від програмних міркувань, таких як чутливість даних, які підлягають розголошенню.

Рекомендація 10: Переконструюйте таблиці, застосуйте приховування комірок, контрольоване табличне врегулювання, або методи збурення до мікроданих перед зведенням в таблиці.

Існує чотири методи обмеження розкриття в таблицях порядкових даних. По-перше, для окремих таблиць, або наборів таблиць, які ієрархічно не пов’язані, агентства можуть обмежити розкриття комбінуючи рядки і/або колонки. По-друге, для більш складних таблиць, приховування комірки може використовуватись для обмеження розкриття. По-третє, контрольоване табличне врегулювання може застосовуватись для захисту чутливих комірок після зведення в таблиці. По-четверте, чутливі комірки можуть захищатись перед зведенням у таблиці застосовуючи певний метод збурення, що додає шуму до ключових мікроданих.

Приховування широко застосовується федеральними агентствами. Приховування комірки видаляє із публікації (приховує) всі комірки, що представляють розкриття, разом із іншими комірками, що не підлягають розкриттю, які можна використати для повторного вираховування або деякої оцінки первинних, чутливих комірок розкриття. Нульові комірки часто легко ідентифікувати, і їх не слід використовувати в якості додаткових приховувань. Моделі приховування повинні пройти аудиторську перевірку для визначення чи алгоритми, які обирають модель додаткового приховування, дозволяють оцінювання значень прихованої комірки у межах «занадто вузького» діапазону. Методи приховування повинні надавати захист із мінімальними втратами даних, як це вимірюється відповідним критерієм, таким як мінімальне число прихованих комірок або мінімальне підсумкове значення, яке є прихованим. Якщо втрата інформації через приховування комірки підриває корисність даних, інші методи можуть бути більш корисними.

Контрольоване табличне врегулювання, що застосовується до таблиць, і методи збурення, що

застосовуються до мікроданих, перед зведенням у таблиці, усувають втрату інформації, пов'язану із приховуванням. Одне попередження говорить, що обидві методології не можуть забезпечити достатнього захисту комірці, що має одного респондента або комірці, в якій переважає один респондент. Може існувати також втрати інформації, виведені логічно, через змінювання даних. Взаємозв'язок між таблицями також необхідно перевірити для мінімізації будь-яких врегулювань комірок в інших таблицях, або наборах таблиць, а також повинен бути переглянутий для перевірки, чи будь-які із аналітичних властивостей таблиці(ць) було спотворено чи обмежено. Ці рекомендовані процедури також застосовуються, якщо приховування використовується для таблиць даних підрахунку частотності.

Рекомендація 11: При застосуванні приховування комірки, необхідно здійснити аудиторську перевірку табличних даних.

Таблиці, де використовується приховування для захисту чутливих комірок, повинні пройти аудиторську перевірку для гарантії, що значення у прихованих комірках не можуть виводитись маніпулюючи рівняннями ряду і колонки. Ця рекомендація застосовується як до таблиць частотних даних, так і до таблиць порядкових даних.

Б.4. Мікродані

Рекомендація 12: Видаліть прямі ідентифікатори та обмежте іншу ідентифікуючу інформацію із файлів мікроданих. Складне завдання із застосування методів статистичного розкриття до мікроданих полягає в тому, щоб запобігати ідентифікації респондента виходячи з даних, що є у записі при дозволі розголошення максимальної кількості даних. Спроможність співставляти змінні із зовнішніх файлів генерує додаткові ризики розкриття, що розширюють перелік змінних у файлі, який необхідно переглянути. Першим кроком до захисту конфіденційності респондента, є видалення із файлу мікроданих всієї **прямої персоніфікованої інформації**, такої як ім'я, номер соціального забезпечення, точна адреса, або дата народження. Певна одномірна інформація, така як професія чи точне географічне місце розташування також може бути персонінкованою. Інша одномірна інформація, така як дуже високий дохід чи наявність рідкісної хвороби, може служити як для ідентифікації респондента, так і для розкриття конфіденційних даних. Ці дані повинні також видалятись або захищатись. Агентства повинні, також, продовжувати ідентифікувати одномірні дані, що мають тенденцію сприяти ідентифікації чи представляти розкриття, і встановлювати обмеження щодо способів, в які ця інформація доповідається. Наприклад, Бюро перепису населення представляє географічну інформацію лише для зон із 100,000 чи більше осіб. Дохід та інша інформація може проходити верхнє кодування до попередньо визначеного значення, такого як 99-й процентиль від розподілу. І під кінець, відповідні розподіли і комбінаційні табличні зведення повинні перевірятись для гарантії, що фізичні особи не ідентифікуються прямо. Обставини можуть коливатись в широких межах між агентствами, або у межах агентства між файлами мікроданих.

Після того, як прямі ідентифікатори було видалено, файл все ще може залишатись простим для ідентифікації, якщо достатня кількість даних залишається у файлі з яким можна співставити інформацію із зовнішнього джерела, що також містить імена або інші прямі ідентифікатори. З цієї причини агентства повинні проводити дослідження повторної ідентифікації і намагатися співставити змінні у розголошених файлах із зовнішніми файлами за межами агентства:

Рекомендація 13: Агентства повинні обмінюватися інформацією із оцінювання ризиків розкриття.

Агентства повинні обмінюватись інформацією щодо того, які зовнішні файли доступні користувачу для співставлення із даними результатів досліджень агентства. Інформація щодо зовнішніх файлів повинна застосовуватись, і поширюватись серед статистичних агентств для того, щоб наглядові ради із розкриття, конфіденційні посадові особи, та інші спеціальні наглядові ради з розкриття, могли належним чином оцінити ризик розкриття із запропонованого випуску даних.

ГЛОСАРІЙ

Розкриття атрибуту – Розкриття, що розголошує чутливу інформацію про суб'єкт даних.

Аудит – Перевірка запропонованої моделі приховування для того щоб переконатись, що чутливі комірки належним чином захищені.

З нижнім кодуванням – Заміна значень, що знаходяться нижче певного числа або класифікації за процентилем, таким же значенням.

Додаткове приховування – Утримання нечутливих комірок від розголошення для того, щоб захистити інші чутливі комірки від розкриття конфіденційної інформації.

Конфіденційна інформація – інформація, що доповідається відповідно до очікування, що вона не буде розголошуватись у спосіб, що робить можливою публічну ідентифікацію респондента чи завдає шкоди респонденту.

Розкриття – розголошення інформації, що має відношення до особи суб'єкту даних, або деякої чутливої інформація про суб'єкта даних, через розголошення або таблиць, або мікроданих.

Дані з підрахунку частотності – Дані, що показують число одиниць аналізу в комірці.

Ієрархія – Серія елементів, організованих чи класифікованих відповідно до рангу або порядку; особливо схема рангової класифікації, що використовується для будови структури таблиці, або файлу мікроданих, таких як коди NAICS.

Високий ризик – Інформація, що має велику ймовірність того, що її використають або для

ідентифікації респондента, або для виявлення конфіденційної інформації про респондента.

Форма, що піддається ідентифікації – Будь-яке представлення інформації, що дозволяє належне логічне виведення особи респондента, до якого застосовується інформація, або прямими, або непрямими засобами.

Логічно виведене розкриття – Розкриття, що робить можливим визначення значення деяких характеристик будь-якої фізичної особи більш точно, ніж це могло бути можливим в іншому випадку.

Розкриття особи – Розкриття, що ідентифікує суб'єкта даних.

Інформована згода – Письмова згода від респондента на публікацію чутливих значень комірок. Вона має ефект дії як відмова від обіцянки захищати чутливі комірки, і спеціальне уповноваження або згода, надана агентству для публічного розголошення конфіденційної інформації.

Зловмисник – Зовнішній користувач, який намагається приєднати пов'язати респондента із записом з мікроданими.

Лінійна міра чутливості – Правило, яке вказує наскільки точно можна оцінити дані респондента, виходячи із опублікованого значення комірки.

Порядкові дані – Дані, що показують зведену величину «кількості інтересу», що застосовується до одиниць аналізу в комірці.

Правила первинного приховування – Лінійна комбінація даних рівня респондента, що використовується для визначення, чи дана комірка таблиці змогла б оприлюднити інформацію про окремого респондента.

Первинне приховування – Утримання від публікації будь-яких комірок, які ідентифікуються як такі, що відповідають правилам первинного приховування.

Публічне користування – Результати дослідження даних, що розголошуються статистичними агентствами будь-кому без обмежень у використанні, або інших умов, крім виплати зборів для купівлі даних в електронній формі.

Обмежені дані − Врегулювання даних у випущених таблицях і файлах мікроданих, або обмеження кількості розголошеної інформації.

Обмежений доступ – Накладання порядків та умов на доступ користувачів до результатів досліджень даних.

Вибірка – Набір записів або елементів даних, взятих із населення, і які використовуються для оцінювання характеристик населення.

Чутливий – Класифікація значення комірки, встановлена використанням правилом первинного приховування.

Приховування – Утримання інформації в обраних комірках таблиці від розголошення.

Субадитивність – Властивість, яка полягає в тому, що об'єднання двох нечутливих комірок є також нечутливим.

Табличні дані – Дані, представлені в таблицях.

Тривимірна таблиця – Таблиця, що містить сумарні значення комірки стосовно трьох змінних.

Верхнє кодування – Заміна значень, вищих певного рангу процентиля таким же значенням.

Двовимірна таблиця – Таблиця, що містить сумарні значення комірки стосовно двох змінних.

ДОДАТОК А – Технічні примітки: Поширення правил первинного приховування на інші звичайні ситуації

Додаток містить містить порядки, які статистичні агентства вважають корисними при застосуванні обмеження розкриття до таблиць в звичайних ситуаціях. Процедури первинного і додаткового приховування для таблиць порядкових даних, що обговорюються в Розділі IV, засновані на припущенні, що опубліковані дані є строго позитивними, і що опубліковане число є простою сумою даних від всіх респондентів. В деяких ситуаціях опубліковані дані не є простими сумами, і незрозумілим є те, як застосовувати методологію первинного і додаткового приховування. Наприклад, в цьому додатку, ми розширюємо дію правил первинного приховування, що використовуються для табличних даних, на таблиці з умовно нарахованими даними.

Крім того, методи, що обговорюються в цьому документі, повинні в наявній формі застосовуватись до кожної опублікованої змінної. На практиці було зроблено припущення, що спрощують розрахунок, для зменшення робочого навантаження, пов'язаного із обмеженням розкриття і для покращення послідовності опублікованих таблиць протягом певного періоду часу.

Секція 2 представляє процедури обмеження розкриття, які використовувались там, де могло виникати певне питання стосовно способу застосування стандартних процедур. Секція 3 представляє припущення, що спрощують підрахунок, і які вважаються федеральними статистичними агентствами корисними. Обидві секції передбачені як довідка для інших агентств, що наштовхуються на схожі ситуації.

1. Підґрунтя

«Правила (n, k), pq-неоднозначності і p-відсотка», описані в Розділі IV, можуть всі бути записані в наступній формі:

$$S(X) - \sum_{i=1}^{n} x_i - c\left(T - \sum_{i=1}^{s} x_i\right)$$

де значення *n, c* і *s* залежать від конкретного правила та обраних параметрів, *T* це підсумкова величина, яка підлягає опублікуванню, x_1 це найбільше опубліковане значення, x_2 це друге за величиною опубліковане значення, і так далі. У цій структурі *xi* є всі невід'ємними.

2. Розширення порядків обмеження розкриття

2.a. Дані вибіркового опитування

Вказане вище рівняння припускає, що всі дані доповідаються (як у переписі). Як це правило можна застосувати до даних із вибіркового опитування? Один спосіб здійснення цього полягає в тому, щоб дозволити значенням найбільших респондентів, і, бути визначеними незваженими опублікованими значеннями, але дозволити Т бути зваженим опублікованим значенням для публікації. (Примітка: це належний спосіб зазначення, що із даними з вибіркового опитування немає жодного розкриття, коли жодні об'єднання не обираються із впевненістю, а долі вибірки є маленькими).

2.б. Таблиці, що містять умовно нараховані дані

Якщо деякі дані нараховуються умовно, потенціал розкриття залежить від методу умовного нарахування.

а) Умовне нарахування для вибіркового опитування здійснюється врегулюванням ваг: В цьому випадку застосовується метод 2.а (врегульовані ваги використовуються для обчислення зваженої підсумкової величини, Т).

б) Умовно нараховані значення можуть бути основані на даних інших респондентів, як у «гарячому покриванні»: В цьому випадку умовно нараховане значення не повинно містити розкриття про особу, що проходить опитування, так щоб умовно нараховане значення (зважене, якщо це доцільно) включається в оцінену підсумкову величину, Т. Умовно нараховане значення підраховується, як окреме опубліковане значення для цілей ідентифікації найбільших респондентів лише для респондента, що подає інформацію.

в) Умовно нараховані значення можуть бути основані на даних за минулий період від особи, що не бере участь в опитуванні: Якщо умовно нараховане значення було розголошено, воно може містити розкриття про особу, що не бере участь в опитуванні (наприклад, якщо умовно нараховане значення засноване на даних, що подаються тим же респондентом в інший період часу). Умовно нараховане значення включається в оцінену підсумкову величину, Т, і також трактується як подані дані для цілей ідентифікації найбільших респондентів.

2.в. Таблиці, що доповідають негативні значення

Якщо всі опубліковані значення є негативними, правила приховування можуть застосовуватись прямо беручи абсолютне значення опублікованих даних.

2.г. Таблиці, де відмінності між позитивними значеннями доповідаються.

Якщо опублікований елемент це різниця між двома позитивними величинами, що доповідаються для того ж періоду часу (наприклад, чиста продукція дорівнює валовій продукції мінус видатки), то застосовуйте правило первинного приховування наступним чином:

а) Якщо отримана різниця є в загальному позитивною, застосовуйте процедуру приховування до першого елементу (валова продукція у прикладі вище).

б) Якщо отримана різниця є в загальному негативною, застосовуйте процедуру приховування до другого елементу (видатки у прикладі вище.)

в) Якщо отримана різниця може бути, або позитивна, або негативна, і не номінується жодною із них, існує два підходи. Один метод полягає в тому, щоб встановити граничне значення для мінімального числа респондентів в комірці. Дуже консервативний підхід полягає в тому, щоб обрати абсолютне значення різниці перед застосуванням правила первинного приховування.

2.д. Таблиці, що доповідають чисті зміни (тобто різниця між значеннями, що доповідається в різний час)

Якщо кожне із значень, що використовується для обчислення чистої зміни, було приховано в оригінальній публікації, то чиста зміна також повинна приховуватись.

2.е. Таблиці, що доповідають середньозважені значення

Якщо опублікований елемент представлений середньозваженим значенням двох позитивних опублікованих кількостей, таких як середньозважена за об'ємом ціна, застосовуйте процедуру приховування до змінної зважування (об'єм є у зразку).

2.є. Результат обчислення із статистичних моделей

Результат обчислення із статистичних моделей, таких як економетричних рівняння, що оцінюються з використанням конфіденційних даних, можуть мати ризик розкриття. Часто кінцеві результати обчислень із статистичних аналізів приймає форму коефіцієнтів параметру у різноманітних типах рівнянь регресії або систем рівнянь. Оскільки, можливо точно відновити вхідні дані із рівняння регресії тільки якщо число коефіцієнтів рівне числу результатів спостережень, результати обчислення регресії в загальному не представляє жодного ризику розкриття. Проте іноді фіктивні (0,1) змінні використовуються в моделі для захоплення певних ефектів, і ці фіктивні змінні можуть приймати значення лише для малої кількості спостережень.

Один спосіб для того, щоб справитись з цією ситуацією, передбачено Центром економічних досліджень від Бюро перепису населення. Вони трактують фіктивні змінні таким чином, ніби вони є комірками в таблиці. Використовуючи «правило (n, k)», аналіз розкриття здійснюється за спостереженнями, для яких фіктивна змінна приймає значення 1.

3. Процедури спрощення

4.3.а. Приховування ключового елементу

В декількох економічних переписах, Бюро перепису населення застосовує приховування ключового елементу: проведення аналізу первинного розкриття і додаткове приховування лише на певних елементах ключових даних, і застосування такої ж моделі приховування до інших відповідних елементів. При приховуванні ключового елементу, менша кількість ресурсів агентства присвячена обмеженню розкриття і результати досліджень даних є більш однорідними серед елементів даних.

Ключові і відповідні елементи ідентифікуються експертною оцінкою. Вони повинні залишатися стабільними протягом певного періоду часу.

3.б. Попередні і кінцеві дані

Що стосується порядкових даних, які розголошуються як у попередній так і в кінцевій формі, модель приховування, що ідентифікується і використовується для попередніх даних повинні зараховуватись до кінцевої публікації. Таблиці кінцевих даних потім підлягають аудиту для гарантії, що немає жодних нових випадків розкриття. Цей консервативний підхід зменшує ризик того, що третя сторона ідентифікує дані респондента виходячи із змін у моделях приховування між попередньою і кінцевою публікацією.

3.в. Дані тимчасового ряду

Для шаблонних щомісячних або щоквартальних публікацій порядкових даних, може розроблятись стандартна модель приховування (первісна або додаткова), основана на щомісячних даних попереднього року. Ця модель приховування, після здійснення аудиту для гарантії відсутності жодних нових випадків розкриття, буде використовуватись в регулярних щомісячних публікаціях.

ДОДАТОК Б − Урядові посилання і веб-сайти

1. Звіт із методик статистичного уникнення розкриття. Робочий документ із статистичної політики 2 (травень 1978). Вашингтон, округ Колумбія: Міністерство торгівлі США, Управління політики і федеральних статистичних норм. Цей звіт доступний в Національній службі технічної інформації (NTIS): Продаж документів NTIS, 5285 Port Royal Road, Springfield, VA 22161; 703-487-4650. Номер документа NTIS це PB86-211539/AS.

2. Інструкція із застосування стандартів Управління з інформації в області енергетики. (Вересень 2002). Управління з інформації в області енергетики, Міністерство енергетики США. Вашингтон, округ Колумбія. http://www.eia.doe.gov/smg/Standard.pdf

3. Федеральна статистика: Звіт Президентської комісії з федеральної статистики, Том 1. Президентська комісія з федеральної статистики. Вашингтон, округ Колумбія: Агентство друку уряду США.

4. Меморандуми політики і стандартів NASS. Національна служба сільськогосподарської статистики, Міністерство сільського господарства США. Вашингтон, округ Колумбія.

5. Статистичні стандарти NCES. (Червень 2003). Національний центр статистики освіти, Міністерство освіти США, Вашингтон, округ Колумбія. http://nces.ed.gov/statprog/2002/stdtoc. asp

6. Стандарт NCES із «Підтримування конфіденційності». Національний центр статистики освіти, Міністерство освіти США. Вашингтон, округ Колумбія, Міністерство освіти США. http://nces.ed.gov/statprog/2002/std4 2.asp

7. Довідник персоналу NCHS із конфіденційності. (Вересень 2004). Національний центр медичної статистики, National Center for Health Statistics, Міністерство охорони здоров'я і соціальних служб США. Вашингтон, округ Колумбія. http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

8. Методики зв'язування записів − 1985, Порядок проведення семінару із методологій точної відповідності. Видання 1299 (Лютий, 1986). Відділ статистики доходу, Служба внутрішніх доходів, Міністерство фінансів США. Вашингтон, округ Колумбія.

9. Посібник з користування підрозділу SOI. (Січень 1985). Відділ статистики доходу, Служба

внутрішніх доходів, Міністерство фінансів США. Вашингтон, округ Колумбія.

ВЕБ-САЙТИ ДЛЯ ДОДАТКОВИХ ДЖЕРЕЛ

1) http://www.fcsm.gov/committees/cdac/ Веб-сайт для Комітету з конфіденційності і доступу до даних. Цей сайт надає корисні посилання на ресурси для методологій уникнення розкриття і відповідних питань щодо доступу до даних.

2) http://www.amstat.org/comm/cmtepc/index.cfm Веб-сайт для Приватності, конфіденційності, і безпеки даних американської статистичної асоціації. Цей сайт надає вичерпну інформацію та посилання з методологічних, правових, етичних і технічних питань, що виникають в процесі захисту і використання статистичних даних

3) www.census.gov/srd/sdc/index.html Цей сайт надає посилання і вичерпні додаткові документи для дослідження, що фінансується Бюро перепису населення США в галузях контролю, конфіденційності, та обмеження статистичного розкриття

4) http://aspe.os.dhhs.gov/datacncl/privcmte.htm Веб-сайт Комітету з приватності Міністерства охорони здоров'я і соціальних служб США.

5) http://neon.vb.cbs.nl/casc/ Веб-сайт для Обчислювальних аспектів статистичної конфіденційності (CASC) (що управляється Бюро статистики Нідерландів). Цей сайт надає посилання для завантаження «Mu-Argus» і «Tau-Argus» та застосування правил уникнення розкриття, до мікроданих, або до табличних даних; В наявності є й інші корисні посилання на книжки, документи і презентації.

ДОДАТОК В – Довідкові документи

Метою цього переліку є оновлення довідкових документів із методології обмеження розкриття, які були цитовані в Робочому документі із статистичного розкриття 2 та оригінальній версії Робочого документу статистичної політики 22. Декілька документів було написано з того часу, як ці обидва Робочі документи із статистичної політики були опубліковані в 1978 і 1994 відповідно.

У Федеральній статистичній системі Бюро перепису населення було провідним агентством із проведення дослідження методів обмеження статистичного розкриття. Персонал Бюро перепису населення було активним в публікуванні результатів через їх веб-сайт, про який йшлося в Додатку Б. З цих причин дослідження статистичного обмеження розкриття, що фінансується Бюро перепису населення детально і належним чином охоплюється в цій бібліографії. Крім того, важливі документи, що або описують нову методологію, або підсумовують важливі питання з дослідження в областях обмеження розкриття для таблиць порядкових даних, таблиць частотних даних і мікроданих також включаються.

«Книги», перелічені нижче в алфавітному порядку, відносяться до традиційних технічних книг, написаних одним автором або кількома співавторами, спеціальних зібрань документів за багатьма різними авторами, спеціальними виданнями журналів, присвячених розкриттю, або різноманітних інтерактивних ресурсів (наприклад довідкові документи, посібники).

Книги

«Конфіденційність, розкриття і доступ до даних: теорія і практичне застосування для статистичних агентств»; за редакцією Пета Дойла, Джулії І. Лейн, Джулс Д.М. Теувес, Лаури В. Заяц. Опубліковано в 2001 видавництвом «Elsevier Science» B.V., Амстердам, Нідерланди.

Том має шістнадцять розділів, написаних провідними дослідниками на широкий спектр тем з розкриття. Опис і перелік статей можна знайти за посиланням:

www.elsevier.com/wps/find/bookdescription.cws home/622129/description#description

Розділ 1 доступний в Інтернеті: www.census.gov/srd/sdc/ConfidentialityCH1.pdf

«Елементи контролю статистичного розкриття» Леона Вілленбурга і Тон де Вала. Опубліковано видавництвом «Springer» в 2001. Конспект лекції із статистики, том 155. Цей том більш теоретичний ніж попередній, написаний цими авторами, і заглиблюється у багато важливих методів. Книга має багато розділів по (I) ризику розкриття (II) втраті інформації (III) методики без збурення (IV) збурюючі методики для мікроданих і потім для табличних даних. В наявності є 119 посилань на довідкову літературу, представлених в кінці тому.

«Для запису, що захищає електронну медичну інформацію», видана Національною академією наук і Національною науково-дослідною радою. Опублікована в 1997 видавництвом «National Academy Press», Вашингтон, округ Колумбія. В 1996 Рада з комп'ютерних наук і телекомунікації (CSTB) сформувала Комітет з підтримання приватності і безпеки прикладних програм охорони здоров'я Національної інформаційної інфраструктури на 15 чоловік. Комітет звертається до загроз інформації охорони здоров'я, відповідності існуючих мір з приватності та безпеки, найкращих процедур. Результати роботи комітету було опубліковано в книзі.

«Покращення доступності і конфіденційність даних дослідження», Комітет національної статистики, Національна науково-дослідна рада, редаговано Крістофером Макі і Норманом Брадбурном, опубліковано Національною науково-дослідною радою, видавництво «National Academy Press», Вашингтон, округ Колумбія, 2000. Стислий виклад семінару, проведеного CNSTAT для сприяння дискусії про методи для досягання часто конфліктних цілей, з використання дослідницького потенціалу мікроданих і підтримання прийнятних рівнів конфіденційності.

«Приватні життя і публічна політика: конфіденційність і доступність урядової статистики», редагована Джорджом Т. Дунканом, Томасом Б. Джабіном, Вірджінією А. де Вулф; опубліковано Комітетом національної статистики і Радою з досліджень в галузі суспільних наук, видавництво «National Academy Press», Вашингтон, округ Колумбія, 1993. Цей короткий (23 сторінки) але важливий том складається із робочого резюме і рекомендацій Ради з конфіденційності і доступу до даних. Цю раду було організовано CNSTAT і Радою з досліджень в області суспільних наук для розробки рекомендацій, що б могли допомогти федеральним статистичним агентствам у їх управлінні даними для політичних рішень і досліджень.

«Зв'язування записів і приватність, питання щодо створення нового федерального дослідження і статистичної інформації», (GAO-01-126SP). Ця книга надає стислий виклад різних методологій і методик співставлення файлу мікроданих із зовнішнім файлом. Вона оновлює попередній стислий виклад математичних методів, що використовуються для співставлення, які знаходяться в «Методиках зв'язування записів − 1985, Порядок проведення семінару із точних методологій співставлення», Міністерство фінансів, IRS, SOI, Publication 1299 (2-86).

«Контроль статистичного розкриття на практиці», редагована Леоном Віленборгом і Тон де Ваалом. Опубліковано видавництвом «Springer» в 1996. Конспект лекції із статистики, том 111. Ця книга має на меті обговорити різноманітні аспекти, що мають відношення до розповсюдження персональних, або ділових даних, що збирається в переписах або опитуваннях чи копіюється із адміністративних ресурсів. Існує два детальних розділи із контролю статистичного розкриття, що обговорюють питання захисту мікроданих і декілька методик, які розроблялись і використовувались в різноманітних агентствах. Існують подібні розділи для табличних даних. В наявності є 79 посилань на літературу, які представлені в кінці тому.

Звіти про конференції і семінари

Семінар із конфіденційності статистичних даних (Скоп'є, Македонія, березень 2001). Фінансовано Європейською економічною комісією Організації Об'єднаних Націй (UNECE).

Порядки проведення доступні за посиланням http://192.91.247.58/stats/documents/2001.03.confidentiality.htm. Цей сайт також надає корисні посилання на документи та інші матеріали із статистичної методології.

«Управління процесом логічного виведення в статистичних базах даних: від теорії до практики» (конференція в Люксембурзі, грудень 2001). За редакцією Джозефа Домінго-Феррера. Опубліковано видавництвом «Springer» в 2002 в серії «Конспектів лекції з комп'ютерних наук», LNCS #2316. Перелік статей із короткими зведеннями доступні за посиланням: http://www.springerlink.com/app/home/search-articles results.asp?wasp=5n5d6ynmwn0vwp8d4gfy&referrer=searchmainxml&backto=journal,1,1;linkingpublicationresults,1:105633,1

Семінари фінансуються Євростат, або проходять спільне фінансування з ним. «Приватність в статистичних базах даних», праці наукової спільноти Барселони, конференція червня 2004). За редакцією Джозе Домінго-Феррера і Віценча Торра. Опубліковано видавництвом «Springer» в 2004 у серії «Конспекти лекції з комп'ютерних наук», #3050. Перелік статей із короткими зведеннями можна знайти за посиланням: http://www.springerlink.com/app/home/search-articles-

results.asp?wasp=3l93gmuvtj7yuk32wmf0&referrer=searchmainxml&backto=journal,1,1;linkingpublicationresults,1:105633,1

«Монографії офіційної статистики: робоча сесія із конфіденційності статистичних даних» (Матеріали

Люксембурзької конференції, квітень 2003). Опубліковано Євростатом в 2004. Наступні три інтерактивні документи у форматі .pdf складають всі матеріали конференції.

http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-1/EN/KS-CR-03-004-1-EN.PDF

http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-2/EN/KS-CR-03-004-2-EN.PDF

http://epp.eurostat.cec.eu.int/cache/ITY_OFFPUB/KS-CR-03-004-3/EN/KS-CR-03-004-3-EN.PDF

Спеціальні випуски журналів

Журнал офіційної статистики: спеціальний випуск щодо методів обмеження розкриття для захисту конфіденційності статистичних даних, том 19, № 4, грудень 1998. За редакцією Стефана Е. Фінберга і Леона С.Р.Дж. Віленборга (Цей журнал опубліковано Статистичним управлінням Швеції). Для перегляду переліку статей дивіться: http://www.jos.nu/Contents/issue.asp?vol=14&no=4

Журнал офіційної статистики: спеціальний випуск із конфіденційності і доступу до даних, том 9, № 2., червень 1993. Для переліку статей дивіться: http://www.jos.nu/Contents/issue.asp?vol=9&no=2

Журнал «Of Significance», опублікований Асоціацією користувачів публічними даними, мав спеціальний випуск по Конфіденційності у 2000. Це том 2, номер 1 і він є доступний в Інтернеті за посиланням: www.apdu.org/resources/docs/OfSignificance_v2n1.pdf

Офіційна статистика Нідерландів: спеціальний випуск із контролю статистичного розкриття, том 14, весна 1999. www.cbs.nl/nl/publicaties/publicaties/algemeen/a-125/1999/nos-99-1.pdf

Посилання в Інтернеті

Анотований перелік посилань міститься у статті Джона М. Абовда і Саймона Д. Вудкока у томі «Конфіденційність, розкриття, і доступ до даних: теоретичне і практичне застосування для статистичних агентств». Цей перелік також доступний в Інтернеті за посиланням http://www.census.gov/srd/sdc/abowd-woodcock2001-appendix-only.pdf

Перелік Посилань із конфіденційності мікроданих, складений Вільямом Е. Вінклером в березні 2004, можна також знайти за посиланням www.census.gov/srd/sdc.

Веб-сайти, присвячені питанням розкриття і/або посилання:

www.fcsm.gov/committees/cdac/cdac.html

www.census.gov/srd/sdc.

Посібник

Контрольний перелік із потенціалу розкриття запропонованих розголошень даних (приготовлений Комітетом з конфіденційності і доступу до даних (CDAC) Федерального комітету із статистичної методології (FCSM). http://www.fcsm.gov/committees/cdac/cdac.html

Звіт Оперативної робочої групи щодо розкриття : методологія серії GSS, № 4, Урядова статистична служба. Грудень 1995, Національна статистична служба, Лондон. Цей звіт доступний в Інтернеті: http://www.statistics.gov.uk/downloads/theme_other/GSSMethodology No 04 v2.pdf

Посібник з конфіденційності для робочого персоналу, виданий Національним центром медичної статистики http://www.cdc.gov/nchs/data/misc/staffmanual2004.pdf

Статті

Абауд, Дж. М. і Лейн Дж. І., «Синтетичні дані і захист конфіденційності», (Вересень, 2003). Технічний документ № TP-2003-10, Бюро перепису населення США. Автори описують метод створення численних файлів публічного користування із єдиної бази даних, де фактичні значення замінюються науково чинними оцінками. Аналітична вартість обраних конфіденційних змінних зберігається під час забезпечення захисту від розкриття для файлу.

Енгл, Джон. (2003). «Імітація саламандри: репродукція відсіченого правого хвоста розподілу доходів». Цей документ пропонує метод оцінювання правого хвоста розподілу доходів, використовуючи відомості про ліву і центральну частину розподілу змінної і забезпечує ознайомлення із застосуванням верхнього кодування до файлу мікроданих. http://www.fcsm.gov/03papers/Angle Final.pdf.

Бетлеєм Дж. Г., Келлер В.Дж., і Панекоек Дж. (1990), «Контроль розкриття мікроданих», журнал Американської статистичної організації, том. 85, с. 38-45. Представляється загальний огляд ризику розкриття при розголошенні мікроданих. Теми, що обговорюються – це унікальність населення, унікальність вибірки, унікальність групи населення і процедури захисту від розкриття, такі як додавання шуму, перестановка даних, мікронакопичення, округлення і стягування. Один висновок, до якого дійшли автори полягає в тому, що дуже важко захищати набір даних від розкриття через можливе використання процедур співставлення. Їх точка зору полягає в тому, що дані необхідно розголошувати користувачам із правовими обмеженнями, які запобігають використанню співставлення.

Сесіль Дж. С. (1993), «Законодавство із конфіденційності і Федеральна статистична система

Сполучених Штатів», Журнал офіційної статистики, том9, № 2, с. 519-535. Доступ до даних, як статистичних так і адміністративних, що здійснюється федеральними агентствами у Сполучених Штатах, управляється складною мережею федеральних статутів. Автор надає деякі деталі стосовно Закону про охорону прав особи від 1974, який застосовується до всіх агентств, і законів, які застосовуються конкретно до Бюро перепису населення США, Національного центру статистики освіти і Національного центру медичної статистики. Автор також описує способи, в які ці агентства надали доступ до даних для дослідників.

Кокс Л. Х., (1980) «Методологія приховування і контроль статистичного розкриття», Журнал Американської статистичної асоціації, Том 75, № 370, с. 377-385. Ця стаття висвітлює співвідношення між процесами визначень розкриття, формуванням приватних задач, приховуванням додаткових комірок, і затвердженням результатів. Вона запроваджує застосування лінійного програмування (теорія транспортування) до аналізу і затвердження додаткового приховування. Вона представляє математичний алгоритм, для мінімізації підсумкового числа додаткових приховувань серед рядків і колонок у двовимірних статистичних таблицях. В переписі або загальному опитуванні, типово велике число комірок таблиці і лінійних відносин між ними робить необхідним розбиття єдиної задачі щодо розкриття добре визначеної послідовності взаємопов'язаних приватних задач. Надмірне приховування можна мінімізувати, а ефективність обробки даних підтримувати, якщо процеси приховування комірки і затвердження спершу виконуються на агрегаціях найвищого рівня, і послідовно на зведених показниках нижчого рівня. Документ надає приклад таблиці із 2 чи більше прихованими комірками в кожному рядку і колонці, де значення чутливої комірки можна точно визначити, як приклад для потреби затвердження.

Кокс Л. Х. (1981), «Міри лінійної чутливості в контролі статистичного розкриття», Журнал статистичного планування і логічного виведення, том 5, с. 153-164. Через аналіз важливих критеріїв чутливості, таких як правила концентрації, міри лінійної чутливості, як це представляється, природним способом виникають із практичних визначень статистичного розкриття. Цей документ надає кількісну умову для визначення того, чи певна міра лінійної чутливості є субадитивною. Це є основа, на якій слід приймати чи заперечувати запропоновані визначення розкриття. Обмеження уваги до субадитивних мір лінійної чутливості приводить до добре визначених методик додаткового приховування. Цей документ представляє математичні основи для уточнення, що будь-яке правило лінійного приховування, що використовується для правила розкриття, повинно бути «субадитивним». Вона надає приклади «правила n-k», «правила pq», і «правило p відсотка» та обговорює питання чутливості об'єднань комірок. А також, надає обмежені аргументи для оцінювання (у спеціальних випадках), чи потенційна додаткова комірка змогла б захистити чутливу комірку.

Кокс Л. Х. і Ернст Л. Р. (1982), «Контрольоване округлення», INFOR, Канадський журнал дослідження операцій та обробки інформації, том 20, № 4, с. 423-432. Перевидано: Деякі найновіші досягнення в теорії, обчисленні і застосуванні методів потоку в мережі, видавництво «University of Toronto Press», 1983, с. 139-148.) Цей документ демонструє, що рішення проблеми контрольованого (обмеженого нулем) округлення у двосторонніх таблицях існує. Рішення основане на проблемі підготовленого переміщення.

Кокс Л. Х., С.К. Макдональд і Д.В. Нельсон, (1986). «Питання конфіденційності в Бюро перепису населення США», Журнал офіційної статистики, том 2, № 2, с. 135 -160. Цей документ описує політику і процедури Бюро перепису населення США після основної програми перегляду і дослідження із захисту конфіденційності даних протягом середини 1980-х.

http://www.jos.nu/Contents/jos_online.asp

Кокс Л. Х. (1987), «Конструктивна процедура для незміщеного контрольованого округлення», Журнал американської статистичної асоціації, том 82, с. 520-524. Незміщене контрольоване округлення в таблиці себе куди входять округлення до основи цілого числа, зберігаючи адитивну структуру, і гарантуючи, що очікуване значення округленого запису дорівнює оригінальному запису. Цей документ надає легкий для запровадження алгоритм для отримання незміщеного контрольованого округлення у 2-вимірній таблиці. Метод також вирішує проблему двосторонньої стратифікації в досліджуваній вибірці і може використовуватись для гарантії підрахунків вибірки цілого числа у незміщений спосіб після, наприклад, ітеративного пропорційного припасування (згрібання).

Кокс Л. Х. і Джордж Дж. А. (1989), «Контрольоване округлення для таблиць із проміжними підсумками», Щорічники з дослідження операцій, 20 (1989) с. 141-157. Контрольоване округлення у двосторонніх таблицях, Кокс і Ернст (1982), розширюється до двосторонніх таблиць із обмежуючими умовами проміжного підсумку. Документ відзначає, що ці методи можна вважати такими, що надають об'єктивні рішення. Метод, що використовується – це підготоване формулювання мережі (перевантаження). Рішення є чітким із проміжними підсумками рядка чи колонки. Показано, що мережеве рішення із обмеженнями проміжного підсумку як рядка так і колонки є адитивним, але що воно може завалити обмеження, що встановлюють нульові значення параметру, і може залишати великі підсумкові суми допоміжних таблиць неконтрольованими для умови суміжності. Надається приклад таблиці, для якої не існує жодного контрольованого округлення із обмеженням, що встановлює нульові значення параметру.

Кокс Л. Х. (1995), «Моделі мережі для приховування додаткових комірок», Журнал американської статистичної асоціації, том 90, № 432, сс. 1453-1462. Додаткове приховування комірок це метод захисту даних, що мають відношення до окремих респондентів із статистичного розкриття, коли дані представляються у статистичних таблицях. Декілька математичних методів для здійснення додаткового приховування комірок було запропоновано у статистичній літературі, деякі з них було запроваджено у широкомасштабних середовищах обробки статистичних даних. Кожен запропонований метод має обмеження або теоретично, або в обчислювальному відношенні. Цей документ представляє рішення проблеми додаткового приховування комірки, основаної на лінійній оптимізації по математичній мережі. Ці методи, як демонструється, є оптимальними для певних проблем і пропонують декілька теоретичних і практичних переваг, включаючи зручність маніпулювання та ефективність обчислення.

Кокс Л. Х. (1996), «Захист конфіденційності у статистиці медицини і навколишнього середовища серед маленького населення», Статистика медицини, том 15, с. 1895-1905. Цей документ обговорює проблеми конфіденційності у малих доменах, і пропонує використовувати взяття підвибірок, а також надлишкових вибірок для обмеження розкриття у файлах мікроданих.

Кокс Л. Х. (2002), «Межі для введених даних у 3-вимірних таблицях спряженості ознак згідно із даними граничними підсумковими величинами», у: Контроль логічного виведення у статистичних базах даних − Від теорії до практики, Конспект лекцій з комп'ютерних наук 2316 (Дж. Домінго-Феррер, редактор), Нью-Йорк: видавництво «Springer», с. 2133. Цей документ вивчає проблему визначення точних меж для прихованих введених даних у 3-вимірних таблицях спряженості ознак із заданими конкретних граничних підсумкових величин і недоліків у попередніх підходах, і порівнює декілька методів аналітично.

Кокс Л. Х. (2003), «Про властивості багатовимірних статистичних таблиць», Журнал статистичного планування і логічного виведення, том 117, 251-273. Цей документ вивчає математичні властивості багатовимірних статистичних таблиць, включаючи проблеми і процедури для гарантії існування придатних таблиць із заданими конкретними таблицями граничних значень, неспроможність лінійного програмування виробляти рішення в цілих числах із заданими обмеженнями цілого числа, і умови, згідно з якими рішення в цілих числах гарантуються, базуючись на мережевій структурі і мережевому лінійному програмуванню.

Кокс Л.Х. і Дандекар Р. А. (2004), «Новий метод обмеження розкриття для табличних даних, що зберігає точність даних і простоту використання», Документи із Семінару статистичної політики FCSM від 2002, Робочий документ із статистичної політики 35, Федеральний комітет із статистичної методології, Вашингтон, округ Колумбія: Адміністративно-бюджетне управління США, с. 15-30. http://www.fcsm.gov/working-papers/spwp35.html

Цей документ представляє контрольоване табличне врегулювання федеральній статистичній спільноті, зосереджуючись на його потенціалі із вдосконалення якості даних.

Кокс Л. Х., Келлі Дж., Патіл Р. (2004). «Балансування якості і конфіденційності для множинних табличних даних. Цей документ пропонує використання певних лінійних і нелінійних моделей згідно із конкретними обмеженнями, що можуть використовуватись для регулювання табличних даних для того, щоб зберігати адитивність, коваріацію, співвідношення, і коефіцієнти регресії, а інші зв'язки даних із оригінальної таблиці зберігаються.

Кокс Л. Х., Джеймс П. Келлі, і Раул Дж. Патіл (2005). «Обчислювальні аспекти контрольованого табличного регулювання: алгоритм та аналіз» у книзі «Наступна хвиля в технологіях обчислення, оптимізації і рішень», за редакцією Б. Голдена, С. Раувана, Е. Васіля, опубліковано видавництвом «Springer». Цей документ представляє алгоритм побудови відсікаючої площини для пришвидшення контрольованого табличного врегулювання.

Дандекар Р., Коен М., і Кіркендол Н. (2002). «Захист чутливих мікроданих з використанням методики взяття вибірки латинського гіперкубу». Конспекти лекцій із комп'ютерних наук, том 2316, сс. 117125, квітень 2002. ISSN 0302-9743. Том «Контроль логічного виведення у статистичних базах даних», за редакцією Джозефа Домінго-Феррера, Берлін: видавництво «Springer-Verlag». Цей документ обговорює методологію для створення синтетичних мікроданих, які можна використовувати замість фактичних опублікованих даних або для створення або адитивного, або мультиплікативного шуму, який при поєднанні з оригінальними даними може забезпечувати захист від розкриття в той самий час, відтворюючи багато необхідних якостей файлу з оригінальними мікроданими. Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique <http://taz/smg/papers/BARCEL.pdf

Дандекар Рамеш А., (2004) «Економне запровадження синтетичного зведення в таблиці (також відоме як Контрольоване табличне врегулювання) в існуючих і нових системах публікації статистичних даних», (2004), с. 428-434, Монографії офіційних статистик, Люксембург: Євростат. Документ описує спрощену процедуру в якості альтернативи лінійному програмуванню, основаному на методології

контрольованого табличного врегулювання (СТА) для генерування синтетичних табличних даних для захисту даних, що містять чутливу інформацію. Спрощена процедура СТА демонструє економний підхід, що дозволяє статистичним агентствам використовувати стандартні загальнодоступні програмні засоби, для генерування синтетичних програмних даних.

Дандекар, Рамеш, (2004). «Проект табличного серверу максимальної корисності і мінімальними втратами інформації для контролю статистичного розкриття табличних даних». Конспекти лекцій з комп'ютерних наук, видавництво «Springer-Verlag Heidelberg», ISSN: 0302-9743, том 3050, с. 121-135. Документ обговорює спрощену версію СТА застосовує її до категорійних і порядкових даних випробувань. Вона також надає порівняльну оцінку цього спрощеного підходу СТА і основаного на LP (лінійному програмуванні) СТА використовуючи порядкові результати випробувань. Що стосується цих даних випробувань, спрощене СТА має можливість захищати таблиці з багатьма меншими регулюваннями значень комірки, ніж цього вимагає основане на LP СТА.

Де Лоера Дж., Он, Шмуель, «Всі раціональні політопи представлені політопами транспортування і всі політопні множини цілих чисел представлені таблицями спряженості ознак». IPCO 2004, LNCS 3064, сс. 338-351. Цей документ показує, що будь-який раціональний політоп є поліноміальним часом, що представляється як «несуттєвий» $r \times c \times 3$ тристоронній політоп транспортування лінійної суми. Ця теорема універсальності має важливі наслідки для лінійного і цілочисельного програмування, а також конфіденційного розкриття статистичних даних. Вона забезпечує вбудування поліноміального часу довільних лінійних програм і цілочисельних програм в таких несуттєвих програм з транспортування і в дводольних програмах подвійного потоку. Вона розв'язує декілька актуальних проблем щодо 3-сторонніх політопів транспортування. Вона також демонструє, що діапазон значень, які може досягти вхідне значення у будь-якій несуттєвій 3-сторонній таблиці спряженості ознак із визначеними 2 межами можуть містити довільні невідповідності, які припускають, що розкриття «k-меж» «d-таблиць» для $2 \leq k < d$ є конфіденційним. http://www.opt.math.tu-graz.ac.at/IPCO/prog.10

Добра, Адріан, Фінберг, Стефен Е., (2000), «Межі для місткості комірок у таблиці спряженості ознак із даними граничними підсумковими величинами і розкладними графами». Протоколи Національної академії наук США, том 97, № 22, с. 1885-1892: Верхні і нижні межі із підрахунків комірки грають важливу роль в обмеженні статистичного розкриття. Цей документ надає теоретичні основи і докази точності меж на розкладних графічних логлінійних моделей. Що стосується таких формул, то прості формули, замість вимогливих в обчислювальному плані програм цілочисельних обчислень, видають точні межі. Деякі з цих моделей є подібними в статистиці, наприклад, моделі повної незалежності, але в загальному цей цілий клас моделей відносно малий.

Добра, Адріан, Фінберг, Стефен Е. (2001) «Межі для вмістимого комірок у таблиці спряженості ознак, наведені встановленими граничними підсумковими величинами із застосуванням до обмеження розкриття». Статистичний журнал Організації Об'єднаних Націй ЕСЕ. Том 18, с. 363-371. Цей документ є більш описовою версією результатів, представлених в «Dobra and Fienberg» (2000) по точних межах з обчислення для розкладних графічних моделей.

Дункан Г. Т., Келлер-МакНатті С. А., і Строукс С. Л. (2001), «Ризик розкриття проти корисності даних: карта конфіденційності R-U», технічний звіт Лос-Аламоської національної лабораторії, LA-UR-01-6428. Методи обговорюються для оцінки ризику розкриття файлу, а компроміси в корисності даних, як параметрів в різноманітних методологіях обмеження розкриття змінюються. Автори описують метод для обчислення окремих силових оцінок ризику розкриття і корисності даних, в той

час передбачаючи різні значення в якості параметрів обмеження розкриття.

Еванс Т., Заяц Л., Сланта Дж., (1998). «Використання шуму для обмеження розкриття табличних даних по установі», Журнал офіційної статистики, том. 14, с. 537-551. Цей документ обговорює метод обмеження розкриття для захисту граничних табличних даних по установі додаючи шум до ключових мікроданих перед зведенням у таблиці.

Ернст Л., (1989), «Подальше застосування лінійного програмування до проблем здійснення вибірки», Протокол секції з методами опитувального дослідження, Американська статистична асоціація, с. 625-630. В попередньому документі, Кокс і Ернст (1982), було продемонстровано, контрольоване округлення існує для кожної двовимірної адитивної таблиці. В цьому документі автор з'ясовує за допомогою контрприкладу, що природне узагальнення їх результату не утримується стосовно трьох вимірів.

Фінберг, Стефен Е. 1997. «Конфіденційність і методологія обмеження розкриття: складні завдання для національної статистики і статистичних досліджень». Технічний звіт Відділу статистики Університету Карнегі-Меллон, робочий документ № 668. Відділ статистики Університету Карнегі-Меллон. Піттсбург, штат Пенсильванія. http://www.stat.cmu.edu/tr/tr668/tr668.html. Цей документ надає огляд статистичних питань, що мають відношення до еволюціонуючої сфери методології статистичного обмеження розкриття.

Фішетті М., і Салазар Дж.Дж. (1999), «Моделі та алгоритми для проблеми приховування 2-вимірної комірки при статистичному контролі розкриття», математичне програмування, том 84, 283-312. Цей документ представляє метод Фішетті-Салазара для вирішення задачі прийняття рішень, що має відношення до додаткового приховування комірок. Незважаючи на попередні методи, він захищає всі чутливі комірки відразу, ніж послідовно, і може виробляти оптимальні результати в середньому до великих задач.

Фішетті М. і Салазар Дж.Дж. (2000), «Моделі та алгоритми для оптимізації приховування комірок в табличних даних із лінійними обмеженнями», журнал Американської статистичної асоціації, том 95, с. 916-928. Алгоритми для додаткового приховування комірок для табличних даних, що, як показується, досягають до оптимально великих, але не величезних, формулювань задачі.

Гоматан С., Карр А. Ф., Саніл А. П. (2005), «Перестановка даних в якості задачі прийняття рішень», журнал офіційної статистики. Цей документ обговорює формулювання практичності ризику щодо перестановки даних для категорійних даних.

Гоматам С., Карр А. Ф., Райтер Дж. П., Саніл А. П. (2005), «Розповсюдження даних та обмеження розкриття у світі без мікроданих: структура практичності ризику для серверів дистанційного доступу», Статистична наука, том 20, с. 163 − 177. Сервери для аналізу з дистанційним доступом дозволяють користувачам подавати запити для остаточного результату із статистичних моделей, підходящих для використання конфіденційних даних. Самим користувачам не дозволяється доступ до даних. Сервери аналізу, однак, не є вільними від ризику розкриття, особливо зважаючи на численні, взаємодіючі запити. В цьому документі автори описують ці ризики і пропонують міри ризику, що піддаються кількісному визначенню, і корисність даних, які можна використовувати для

визначення того, на які запити можна відповісти, і з яким остаточним результатом. Структура корисності ризику ілюстрована для регресійних моделей.

Гонзалес Дж.Ф. і Кокс Л.Г. (2005), «Програмне забезпечення для захисту табличних даних». Статистика в медицині, том 24 (4), с. 659-669. Цей документ описує програмне забезпечення для захисту даних у двосторонніх таблицях, розроблених для Національного центру медичної статистики: додаткове приховування комірок, округлення, збурення і контрольоване табличне врегулювання. Програмне забезпечення доступне безплатно.

Грінберг Б. і Заяц Л. (1992), «Стратегії для вимірювання ризику у файлах з мікроданими публічного користування». Видавництво «Statistica Neerlandica», том 46, № 1, с. 33-48. Описуються методи зменшення ризику розкриття для файлів мікроданих і факторів, що зменшують можливість зв'язувати файли та отримувати правильні співставлення. Пояснюються два методи оцінювання відсотка унікальних значень населення у файлі мікроданих. Представляється міра відносного ризику для файлу мікроданих, основаного на понятті ентропії.

Гріффін Р. А., Наварро, А., і Флорес-Баез Л. (1989), «Уникнення розкриття для Перепису населення 1990», Протоколи Секції із методів оглядового дослідження, Американська статистична асоціація, Александрія, штат Вірджинія, с. 516-521. Цей документ представляє процедури обмеження розкриття Перепису населення 1990 року для 100 відсотків і вибіркових даних та наслідків для даних. Метою бюро перепису населення є максимізувати рівень корисної статистичної інформації, що надається відповідно до умови, що конфіденційність не порушується. Ці типи процедур для 100 відсоткових даних було досліджено: приховування, контрольоване округлення і редагування конфіденційності. Переваги і недоліки кожного методу обговорюються. Редагування конфіденційності основане на обиранні малої вибірки перепису домашніх господарств із файлів даних внутрішнього перепису та обміні їх даних із даними інших домашніх господарств, що мають ідентичні характеристики на наборі обраних ключових змінних. Що стосується даних вибірки перепису населення, здійснення вибірки забезпечує належний захист, проте, у малих блоках. Методології, основані на затемненні та умовному нарахуванні, запропоновані для зменшення ризику розкриття у малих блоках.

Хавала С., Заяц Л., Роуланд С., (2004). «Американський «Шукач фактів»: Бюро перепису населення США працює в напрямку задоволення потреб користувачів при захисті конфіденційності», Журнал офіційної статистики, том 20, с. 115-124. Цей документ обговорює спеціальні методики обмеження розкриття, що застосовуються для захисту конфіденційності табличних зведень, генерованих із інтерактивного запиту файлів мікроданих. http://www.jos.nu/Contents/jos online.asp

Джабін Т. Б. (1993a), «Процедури для обмеженого доступу до даних», Журнал офіційної статистики, том 9, № 2, с. 537-589. Статистичні агентства має дві основні опції для захисту конфіденційності даних, які вони розголошують. Один полягає в обмеженні даних через використання процедур обмеження статистичного розкриття. Інший полягає в накладанні умов щодо того, хто може мати доступ, для яких цілей, в яких місцях розташування і так далі. Для другої опції використовується термін **обмежений доступ**. Цей документ є стислим викладом процедур обмеженого доступу, які використовують статистичні агентства США для надавання доступу до даних іншим статистичним агентствам, а також іншим організаціям і фізичним особам. В наявності є багато включених прикладів, що ілюструють як успішні моделі, так і процедури для надавання доступу, а також невдачі отримати бажаний доступ. http://www.jos.nu/Contents/jos online.asp

Джабін Т. Б. (1993б), «Процедури статистичного обмеження розкриття Статистичних агентств США», Журнал офіційної статистики, том 9, № 2, с. 427-454. Однією з тем, що вивчаються Радою з конфіденційності і доступу до даних Комітету з Національної Статистики Національної академії наук було використання процедур обмеження статистичного розкриття для обмеження ризику розкриття особистої інформації коли дані розголошуються Федеральними статистичними агентствами в табличному форматі, або у форматі мікроданих. Для допомоги Раді у її перегляді, автор приготував стислий виклад процедур обмеження розкриття, що використовувались агентствами в ранньому 1991. Цей документ є адаптованою версією цього стислого викладу.

http://www.jos.nu/Contents/jos_online.asp

Джует Р. (1993), «Аналіз розкриття для Економічного перепису 1992 року», неопублікований рукопис, Підрозділ економічного програмування, Бюро перепису населення, Вашингтон, округ Колумбія. Автор детально описує методологію потоку в мережі, що використовується для приховування комірок для 1992.

Економічні переписи. Програми, що використовуються у системі розкриття та їх вхідні і вихідні дані також описуються. http://www.census.gov/srd/sdc/Jewett.disc.econ.1992.pdf

Карр А. Ф., Лін К., Райтер Дж.П., Саніл А. П. (2005). Безпечна регресія на розподілених базах даних. Журнал про обчислювальні методи графічної обробки статистичних даних, том 14, № (2), с. 263-279. Ця стаття представляє декілька методів для проведення лінійної регресії по об'єднанні розподілених баз даних, що різною мірою зберігають конфіденційність цих баз даних. Такі методи можуть використовуватись федеральними і державними статистичними агентствами, для обміну інформацією з їх особистих баз даних, або ж робити таку інформацію доступною для інших.

Келлер-МакНатті С., МакНатті М. С., та Унгер Е. А. (1989), «Захист конфіденційних даних». Протокол 21-го симпозіуму по інтерфейсу, Американська статистична асоціація, Александрія, штат Вірджинія, сс. 215-219. Широкий огляд аналітичних методів, що були використані, або ж які можна було використати для захисту конфіденційності, надається як для файлів мікроданих, так і для табличних розголошень. Описуються деякі методи, які б можна було використати із мікроданими, наприклад «затемнення», «розподіл на шари». Автори також описують потребу у стандартній мірі «контролю» або захисту.

Кеннікель, Артур Б. (1998). «Метод множинного відновлення в опитуванні споживчого кредитування». Протоколи спільних статистичних зібрань Американської статистичної асоціації від 1998. Цей документ описує систему «FRITZ» методу множинного відновлення, розробленого для Опитування споживчого кредитування. Крім описування застосування системи до звичайних задач умовного нарахування відсутніх даних, документ представляє результати використання системи для ряду експериментів із моделювання даних для уникнення розкриття.
http://www.federalreserve.gov/pubs/oss/oss2/papers/impute98.pdf

Кім Дж. Дж. (1986), «Метод для обмеження розкриття в мікроданих, оснований на довільному шумі і

трансформації», Американська статистична асоціація, Протоколи секції щодо методів оглядового дослідження, с. 370-374. Хоча додавання шуму є ефективним при зменшенні ризику розкриття, воно має несприятливий вплив на будь-який аналіз даних. Якщо особа знає як дані повинні використовуватись, трансформації даних перед і після додавання шуму, можуть підтримувати корисність даних. Автор рекомендує використання лінійних перетворень згідно із обмеженнями, що перший і другий моменти нової змінної є ідентичними до змінних оригіналу. Він представляє властивості перетвореної змінної, коли розбіжність відома, і коли вона оцінюється. Він викладає наслідки маскування за оціночними показниками параметру регресії згідно із різними умовами зберігання першого і другого моментів оригінальних даних.

Кім Дж. Дж., і В.Е. Вінклер (1995). «Маскування файлів мікроданих», Американська статистична асоціація, Протоколи секції із методів оглядового дослідження, с. 114-119. Жодна схема маскування на даний час не відповідає потребам всіх користувачів даних. Ця стаття описує схему маскування, що використовується для конкретного випадку надавання мікроданих двом користувачам, що взяли до уваги їх аналітичні проблеми. Оскільки це було зроблено перед Кімом (1990б), кожна група маскувалась окремо. У цьому прикладі користувач планував збудувати моделі множинної регресії, із залежною змінною двох типів – пропорцій, перетворених у логіти, і медіан. Кім обговорює 1) чи додавати шум перед, або після перетворення, 2) який розподіл шуму використовувати, і 3) чи додавати корельований, або некорельований шум. Він у деталях представляє процес маскування, статистичні властивості маскованих змінних, і як вони задовольнили потреби цього користувача. Чудові результати було отримано для змінних середньої величини і варіації/коваріації, крім випадків, коли значне цензурування супроводжувало логіт-перетворення пропорцій.

Лвмберт Д. (1993), «Міри ризику розкриття і шкоди», журнал офіційної статистики, том 9, № 2, с. 313-331. Визначення розкриття залежить від контексту. Іноді розкриття, як вважається, має місце незважаючи на те, що розголошена інформація є невірною. Розкриття може порушити анонімність респондента та іноді розголошувати чутливу інформацію. Цей документ намагається розплутати питання розкриття диференціюючи між зв'язуванням респондента із записом і вивченням чутливої інформації із зв'язування. Ступінь, до якої розголошений запис може бути зв'язаний із респондентом, визначає ризик розкриття; інформація, розголошена коли респондент зв'язаний із розголошеним записом, визначає шкоду від розкриття. Шкода може бути присутня, навіть якщо ідентифікується невірний запис, або логічно виводиться невірне чутливе значення. В цьому документі розглядаються міри ризику розкриття і шкоди, що відображають те, що дізнаються про респондента, і надаються деякі логічні висновки щодо політики. http://www.jos.nu/Contents/jos online.asp

Лі Дж., Холломан К., Карр А. Ф. і Саніл А. П. (2001), «Аналіз накопичених даних в зборі вибірок опитування із застосуванням до опитувань щодо застосування добрив/пестицидів». Дослідження в офіційній статистиці, том 4, с. 101-116: Цей документ пропонує підхід, оснований на застосуванні Баєсовького моделювання для аналізу даних, накопичених та захисту розкриття.

Массел, Пол Б., (2002). «Моделі оптимізації і програми для приховування комірок у статистичних таблицях», Протоколи Спільних статистичних зборів Американської статистичної асоціації від 2002. Цей документ порівнює різні математичні підходи до застосування приховування комірок та оцінює корисність різних програм, основаних на методі оптимізації, так як і на інших практичних рекомендаціях. Програми з мережевими зв'язками і програми з розширеними мережевими зв'язкамм

порівнюються із лінійним програмуванням, програмами на основі цілого числа і на основі гіперкуба. http://www.census.gov/srd/sdc/Massell.JSM2002.v4.pdf

Массел, Пол Б., (2004). «Порівняння методів контролю статистичного розкриття для таблиць: ідентифікація ключових факторів», Протоколи Спільних статистичних зборів Американської статистичної асоціації від 2004. Цей документ описує ключові фактори, причетні до вирішення того, як обирати метод статистичного розкриття, який є доступним для захисту даного набору таблиць. http://www.census.gov/srd/sdc/Massell.JSM2004.paper.v3.pdf

Міхалевіч, Збігнєв (1991). «Безпека статистичної бази даних», у статистичних і наукових базах даних, у редакції видавництва «Ellis Horwood, Ltd». Ця стаття обговорює безпеку статистичної бази даних, також відому як керування процесом логічного виведення або контроль розкриття. Припускається, що всі дані доступні в інтерактивному режимі, як у файлі мікроданих. Критичний відгук про поточні методи, як обмеження запиту так і збурювання, включений з використанням абстрактної моделі статистичної бази даних. Атаки типу **відслідковування** широко обговорюються. Розробляється баланс між безпекою і придатність для користування, із залежністю придатності для користування методами обмеження запитів від числа, і діапазонів інтервалів обмежених даних. Методи для визначення цих інтервалів порівнюються.

Муралідар К., Саразі R. (Травень, 2002). «Процедура перетасовки даних для маскування даних». Цей документ обговорює методологію і теоретичні основи для застосування двостадійної процедури перестановки даних щодо захисту конфіденційних числових даних. Звіт до Бюро перепису населення, травень, 2002. http://gatton.uky.edu/faculty/muralidhar/maskingpapers.

Паасс Г. (1988), «Ризик розкриття і уникнення розкриття для мікроданих». Журнал ділової та економічної статистики, том 6, с. 487-500. Цей документ дає оціночні значення для частки записів, що піддаються ідентифікації, коли певні типи зовнішньої інформації можуть бути доступними для дослідника, і в той час, ця частка залежить в основному від числа змінних в Паасс потім оцінює показники діяльності мір із запобігання/уникнення розкриття, таких як розділення на шари, мікронакопичення, і рекомбінації. В додатку він представляє технічні деталі запропонованих методів.

Квіан К., Стікел М., Карп П., Лант Т. і Гарві Т., «Виявлення та усунення каналів логічного виведення у багаторівневих системах управління реляційною базою даних», Симпозіум IEEE із дослідження безпеки і приватності, Окленд, штат Каліфорнія, 24-26 травня, 1993. Цей документ береться за вирішення проблеми, де інформація із однієї таблиці може використовуватись для **логічного виведення** інформації, що міститься в іншому файлі. Він приймає інтерактивну систему управління реляційною базою даних їх декількох таблиць. Логічно виведене рішення проблеми полягає у класифікації (і таким чином, запереченні доступу до) відповідних даних. Перевага цього підходу полягає в тому, що такі відкриття здійснюються на етапі **проектування**, а не на етапі виконання. Недоліком є те, що методологія звертається лише до тих ситуацій, де логічні виведення завжди містять, а не ті випадки, коли логічне виведення залежить від конкретних значень даних. Методику необхідно дослідити на застосовність до проблеми обмеження розкриття.

Раунатан Т.Е., Райтер Дж. П., і Рубін Д. Р. (2003), «Метод множинного відновлення для обмеження

статистичного розкриття», Журнал офіційної статистики, том 19, с. 1-16. Ця стаття оцінює використання структури методу множинного відновлення для захисту конфіденційності відповідей респондента у вибіркових опитуваннях. Основна пропозиція полягає в тому, щоб симулювати додаткові копії населення, з яких були відібрані ці респонденти, і щоб розголосити довільну вибірку із кожного з цих синтетичного населення. Користувачі можуть аналізувати набори даних синтетичної вибірки із стандартним програмним забезпеченням повних даних для простих довільних вибірок, потім отримувати чинні логічні висновки комбінуючи точкові оцінки, та оцінки дисперсії використовуючи методи в цій статті. http://www.jos.nu/Contents/jos online.asp

Райтер Дж. П. (2002), «Забезпечення обмежень розкриття з допомогою наборів синтетичних даних», Журнал офіційної статистики, том 18, № 4, с. 531-543. Для уникнення розкриття, Рубін запропонував створити численні, синтетичні набори даних для публічного розголошення для того, щоб (i) жодне об'єднання в розголошених даних не мало чутливих даних із фактичного об'єднання в населенні, і (ii) статистичні процедури, що є чинними для оригінальних даних є чинними для розголошених даних. Цей документ обговорює, через використання імітаційних даних, що чинні логічні виведення можна отримати із синтетичних даних у різноманітних налаштуваннях, включаючи просте довільне здійснення вибірки, відбір з імовірністю, пропорційною розмір, двоетапну гніздову вибірку, і районовану вибірку. http://www.jos.nu/Articles/abstract.asp?article=184531

Резнек А. П., «Ризики розкриття у моделях поперечної регресії», (2003). Цей документ описує ризики розкриття, що мають відношення до певних типів моделей поперечної регресії. Зокрема, він показує через приклади, що моделі лише із фіктивними (0,1) змінними, що повністю взаємодіють з правої сторони надають можливість відновлення введених даних із таблиці засобів змінної лівої сторони, що розбиваються на категорії фіктивних змінних. Протокол Спільних статистичних зборів Американської статистичної асоціації від 2003.

Резнек, Арнольд П. і Т. Лінн Ріггс (2004). «Ризики розкриття і моделях регресії: деякі подальші результати». Протокол Спільних статистичних зборів Американської статистичної асоціації від 2004. Цей документ ілюструє, що кореляційні матриці і варіаційно-коваріаційні матриці змінних, так як і варіаційно-коваріаційні матриці коефіцієнтів моделі, можуть також дозволити відновлення вхідних записів таблиці якщо змінні включають фіктивні змінні.

Робертсон Д. А., (1993), «Приховування комірок у Статистичній службі Канади», Протокол щорічної дослідницької конференції Бюро із перепису населення 1993 року, Бюро перепису населення, Вашингтон, округ Колумбія, сс. 107-131. Статистична служба Канади розробила Комп'ютерне програмне забезпечення (CONFID) для гарантії конфіденційності респондента через приховування комірки. Воно об'єднує комірки табличних зведень із мікроданих, визначає конфіденційні комірки і потім обирає додаткові приховування. Цей документ обговорює проект та алгоритми, що використовуються, і показники його діяльності у Канадському переписі сільського господарства від 1991.

Рубін Д. (1993), «Обговорення, обмеження статистичного розкриття», Журнал офіційної статистики, том 9, № 2, сс. 461-468. Рубін пропонує, що уряд повинен розголошувати лише «синтетичні дані», аніж фактичні мікродані. Синтетичні дані будуть генеруватись використовуючи метод множинного

відновлення. Вони будуть виглядати як окремі опубліковані дані і будуть мати такі ж багатомірні статистичні властивості. Проте, із цією схемою не буде жодної можливості розкриття, так як жодні індивідуальні дані не буде розголошено.

Саалфельд А., Заяц Л., і Хоел Е. (1992), «Контекстуальні змінні через географічне сортування: підхід ковзаючих середніх значень», Протокол секції із Методів оглядового дослідження, Американська статистична асоціація, Олександрія, штат Вірджинія, с. 691-696. Соціологи хотіли б провести просторовий аналіз мікроданих. Вони хочуть знати відносну географічну інформацію про кожен запис, такий як середній дохід сусідніх фізичних осіб. Змінні, що надають цей тип інформації, називаються «контекстуальними змінними». Цей документ представляє методику, яка буде генерувати контекстуальні змінні, які не містять точного місця розташування респондентів. Методика основана на обиранні ковзаючих середніх значень відсортованого набору даних.

Сейлер П., Вебер М., і Вонг В., (2001), «Випробування на розкриття файлу публічного користування із індивідуальною податковою декларацією від 1996», Протокол Американської статистичної асоціації 2001; Цей документ надає огляд методик перевірки на розкриття, що застосовуються до Статистики доходу із файлу публічного користування із індивідуальною податковою декларацією (PUF). Він також обговорює результати двох тестів цих процедур: співставлення публічно доступної маркетингової бази даних до PUF: і співставлення Окремого головного файлу із PUF.

Саніл А. П., Карр А. Ф., Лін К., Райтер Дж. П. (2004), «Регресійне моделювання, що зберігає приватність, через розподілене обчислення», Протокол Десятої міжнародної конференції із виявлення знань та аналізу даних ACM SIGKDD 2004, с. 677-682. Цей документ обговорює безпечну регресію для розподілених, вертикально розбитих даних коли відповіддю обмінюються.

Сінгер Е. і Міллер Е. (1993), «Найновіші дослідження із питань конфіденційності в Бюро перепису населення», Протокол Щорічної дослідницької конференції Бюро перепису населення від 1993, Бюро перепису населення США, Вашингтон, округ Колумбія, с. 99-106. Бюро перепису населення провело дискусії з групами респондентів на конкретну тему стосовно реакцій учасників на використання адміністративних записів для Перепису 2000 року, і їх побоювання стосуються порушень конфіденційності, їх реакцій на ряд мотиваційних пояснень, і способи запевнення, про конфіденційність їх даних. Цей документ виділяє результати цих дискусій і пов’язує отримані відомості із іншого дослідження в цій області.

Стіл, Філіп М. (2004) «Оцінка ризику розкриття для мікроданих». Це вступ до оцінки ризику для мікроданих, для початкового спеціаліста-практика. Він представляє певне підґрунтя щодо правових концепцій можливості бути ідентифікованим, обговорює вимірювання ризику і його застосовність, демонструє як публічні дані і контекст можуть впливати на ризик. Існує також еклектичний набір посилань. http://www.census.gov/srd/sdc/Steel.Disclosure%20Risk%20Assessment%20for%20Microdata.pdf .

Ван Ден Хаут А., і Ван Дер Хейден П. Г. М. (2002), «Рандомізована відповідь, контроль статистичного розкриття, і помилкова класифікація: перегляд». Міжнародний статистичний перегляд, том 70 (2), с. 269-288. Цей документ обговорює аналіз категорійних даних, які були помилково класифіковані і де відомі імовірності помилкової класифікації. Поля, де має місце цей тип помилкової

класифікації, представлені рандомізованою відповіддю, контролем статистичного розкриття, і класифікацією із відомою чутливістю і специфічністю. Надаються оцінки правдивих частотностей, та обговорюються налаштування до відношення шансів. Моментальна оцінка та оцінки максимальної правдоподібності порівнюються і доказується, що вони однакові всередині простору параметрів. http://isi.cbs.nl/ISReview/abst01-13.pdf

Вінклер, Вільям Е. (1998). «Методи повторної ідентифікації для оцінювання конфіденційності аналітично чинних мікроданих», Дослідження в офіційній статистиці, том 1, с. 87-114. Цей документ порівнює декілька методів маскування в плані їх спроможності виробляти аналітично чинні, конфіденційні мікродані. Для того щоб файл мікроданих публічного користування був аналітично чинний, він повинен бути, для невеликої кількості застосувань, давати аналітичні результати, що є приблизно такими ж як і оригінальний, конфіденційний файл, який не порушується. Якщо файл мікроданих містить середню кількість змінних і від нього вимагається відповідати єдиному набору аналітичних потреб, то набагато більше записів мають ймовірність повторної ідентифікації через сучасні методи зв'язування записів ніж через методи повторної ідентифікації, що типово використовуються у літературі з конфіденційності.

Вінклер, Вільям Е., (2004). «Маскування і методи повторної ідентифікації для мікроданих публічного користування: огляд і проблеми дослідження». Цей документ надає огляд різноманітних методів, що застосовуються для маскування мікроданих. Віен також обговорює різноманітні міри для оцінювання ризику розкриття для файлу даних публічного користування.
http://www.census.gov/srd/papers/pdf/rrs2004-03.pdf

Ю, Дантеман, Даі, і Вілсон (2004). «Вимірювання і показники діяльності MASSC використовуючи файл публічного користування NCHS-2000 NHIS». Цей документ обговорює метод обмеження розкриття за допомогою мікро агломерації, заміни, взяття підвибірок, і калібрування. Робоча сесія із конфіденційності даних. Конференція європейських статистів від 2003.
http://www.unece.org/stats/documents/2003.04.confidentiality.htm

Заяц Л. (1992а). «Використання методології лінійного програмування для цілей уникнення розкриття», Серія звітів відділу із статистичних досліджень, Перепис/SRD/RR-92/02, Бюро перепису населення, Відділ статистичних досліджень, Вашингтон, округ Колумбія. Цей документ представляє схему лінійного програмування для знаходження додаткових приховувань для первинного приховування, що застосовується до дво- чи тривимірних таблиць. Метод дає хороші, але неоптимальні результати. Документ обговорює три способи покращення результатів: 1) сортування первинних приховувань з допомогою необхідного їм захисту і знаходження додаткових комірок для кожної первинної комірки починаючи із найбільшої; 2) додавання додаткового короткого перегляду лінійної програми із налаштованою функцією витрат для усунення непотрібних додаткових приховувань, ідентифікованих при першому запуску; і 3) використання різних функцій витрат. Загальне порівняння із методологією потоку в мережі також надається. Документ також надає приклад, використовуючи комерційно доступний пакет лінійного програмування, LINDO.

Заяц Л. В. (1992б), «Методологія лінійного програмування для цілей уникнення розкриття в Бюро перепису населення». Протокол секції із методів оглядового дослідження, Американська статистична асоціація, Олександрія, штат Вірджинія, с. 679-684. Цей документ рекомендує специфічні підходи до знаходження додаткових приховувань для двовимірних таблиць, маленьких тривимірних таблиць і

великих тривимірних таблиць. Процедури потоку в мережі рекомендуються для двовимірних таблиць. Методи лінійного програмування рекомендуються (та описуються) для малих тривимірних таблиць. У випадку великих тривимірних таблиць, рекомендована процедура представлена послідовністю алгоритмів потоку в мережі, що застосовуються до двовимірних допоміжних таблиць. Необхідно потім провести аудит отриманої системи приховувань для гарантії, що чутливі комірки захищені. Описується алгоритм лінійного програмування для затвердження моделей приховувань.

Заяц Л. (2002). «SDC 2000 в Переписі США 2000, що проводиться з десятилітніми інтервалами», Управління процесом логічного виведення у статистичних базах даних: від теорії до практики, видавництво «Springe», с.193-202. Цей документ описує методики статистичного обмеження розкриття, що використовуються для всіх результатів досліджень даних Перепису США 2000. Він описує процедури для таблиць короткої форми, таблиць повної форми, файлів мікроданих публічного користування, та інтерактивна система опитування для таблиць. Процедури, що використовувались, включають перестановку даних, округлення, додавання шуму, стягування категорій, і застосування граничних значень.

Заяц Л., Массел П., і Стіл П. (1999). «Практика обмеження розкриття і дослідження в Бюро перепису населення США», Офіційна статистика Нідерландів, весна, 1999, том 14, с. 26-29. Цей документ обговорює практику обмеження розкриття, що діють в Бюро перепису населення, так як і поточні дослідження Бюро перепису населення по відношенню до альтернативних процедур обмеження розкриття і деяких аналізів цих процедур.

ДОДАТОК Г – Комітет з конфіденційності і доступу до даних

В 1995, Міжвідомча група з конфіденційності і доступу до даних (ICDAG) була сформована для (1) сприяння і реалізації цілей та рекомендацій, окреслених в Розділі 6 Робочого документу із статистичної політики #22 (2) підвищення співпраці та спільного користування методами обмеження статистичного розкриття між федеральними агентствами і (3) забезпечення форуму для спільного користування інформацією та ідеями щодо захисту конфіденційності даних і покращення доступу до даних. Його членами є працівники федеральних агентств Виконавчого органу, що працюють над питаннями конфіденційності даних і доступу до даних, які виражають потребу у форумі для обміну своїми знаннями та обговоренні спільних питань і занепокоєння. Ще в 1995, ICDAG було неформально приєднано до Федерального комітету із статистичної методології (FCSM).

В 1997, FCSM офіційно призначило ICDAG як «Група осіб, об’єднана спільними інтересами» для сприяння комунікації і співпраці між агентствами. В 2000, назву групи було змінено на Комітет з конфіденційності і доступу до даних (CDAC). З 1997 CDAC розробило декілька приладів обробки і передачі даних, щоб допомогти централізувати перегляд агентством приладів обробки і передачі даних, обмежених для розкриття , методології обміну, програмного забезпечення, та інформації по федеральних агентствах із питань і діяльності щодо конфіденційності і доступу до даних. Див. http://www.fcsm.gov/committees/cdac/ Крім того, його члени надають презентації щодо методології статистичного розкриття різноманітним аудиторіям протягом року для того, щоб допомогти розширити робочі знання в цих галузях.

Результати досліджень даних, які розробили CDAC, включають в себе:

Контрольний перелік із потенціалу розкриття запропонованих розголошень даних – Цей документ стандартизує перегляд для ризиків розкриття, пов'язаних із будь-яким запропонованим розголошенням даних.

Брошура на тему «Питання конфіденційності і доступу до даних між федеральними агентствами – Ця брошура описує деякі приклади захисту даних, що використовується федеральними агентствами – правові санкції, видалення персональних ідентифікаторів із наборів даних, застосування статистичних процедур до опублікованої інформації, сертифікатів з конфіденційності, інституційних і наглядових рад з розкриття, та обмеженого доступу до даних (дослідницькі центри зі збору даних, дистанційний доступ, спеціальний статус співробітника і ліцензування даних).

Процедури обмеженого доступу – Цей документ обговорює різноманітні методи, що використовуються п'ятьма федеральними агентствами, для надання доступу до статистичних даних, в той же час обмежуючи ризик розкриття конфіденційної інформації. Методи описують Дослідницькі центри із збору даних (RDC), системи дистанційного доступу та інтерактивного опитування, дослідницькі співтовариства і постдокторстькі програми, і ліцензійні договори.

Схильність до ідентифікації у файлах мікроданих – Цей документ надає розуміння того, які змінні і типи даних роблять окремих респондентів такими, що піддаються ідентифікації у файлі мікроданих.

Програмне забезпечення для аудиту розкриття – Це програмне забезпечення SAS на базі ПК ідентифікує нижні і верхні межі по значеннях утриманої (прихованої) комірки у зведеній статистичній таблиці, і передбачає інші корисні міри для здійснення аудиту моделі приховування в таблиці.

Звіти, доступні у серії робочих документів по статистичній політиці Федерального комітету із статистичної методології

1. ***Звіт із статистики для розподілу грошових коштів,*** 1978 (NTIS PB86-211521/AS)
2. ***Звіт із методик статистичного розкриття та уникнення розкриття,*** 1978 (NTIS PB86-211539/AS)
3. ***Короткий нарис помилок: зайнятість, як виміряний поточним оглядом населення,*** 1978 (NTIS PB86-214269/AS)
4. ***Глосарій термінів систематичних помилок: ілюстрація семантичної проблеми у статистиці,*** 1978 (NTIS PB86-211547/AS)
5. ***Звіт по точних і статистичних методиках співставлення,*** 1980 (NTIS PB86-215829/AS)
6. ***Звіт по статистичному використанню адміністративних записів,*** 1980 (NTIS PB86-214285/AS)
7. ***Міжвідомчий огляд політики виправлення часових рядів,*** 1982 (NTIS PB86-232451/AS)
8. ***Статистичні міжвідомчі договори,*** 1982 (NTIS PB86-230570/AS)
9. ***Укладення договорів для опитувань,*** 1983 (NTIS PB83-233148)
10. ***Підходи до розробки опитувальних листів,*** 1983 (NTIS PB84-105055)
11. ***Перегляд індустріальних систем кодування,*** 1984 (NTIS PB84-135276)
12. ***Роль телефонного збору даних у федеральній статистиці,*** 1984 (NTIS PB85-105971)
13. ***Федеральні поздовжні дослідження,*** 1986 (NTIS PB86-139730)
14. ***Семінар по статистичному використанні мікрокомп'ютерів у Федеральних агентствах,*** 1987 (NTIS PB87-166393)
15. ***Якість зборів статистичних відомостей по підприємствах,*** 1988 (NTIS PB88-232921)
16. ***Порівняльне вивчення звітних одиниць в обраних системах збору даних роботодавця,*** 1990 (NTIS PB90-205238)
17. ***Область спостережень,*** 1990 (NTIS PB90-205246)
18. ***Редагування даних у Федеральних статистичних агентствах,*** 1990 (NTIS PB90-205253)
19. ***Комп'ютерний збір оглядових даних,*** 1990 (NTIS PB90-205261)
20. ***Семінар по якості федеральних даних,*** 1991 (NTIS PB91-142414)
21. ***Непрямі методи оцінювання у федеральних програмах,*** 1993 (NTIS PB93-209294)
22. ***Звіт по методології статистичного обмеження розкриття,*** Версія друга 2005
23. ***Семінар по нових напрямках у статистичній методології,*** 1995 (NTIS PB95-182978)
24. ***Електронне поширення статистичних даних,*** 1995 (NTIS PB96-121629)
25. ***Семінар та експозиція по редагуванні даних,*** 1996 (NTIS PB97-104624)
26. ***Семінар із статистичної методології в державній службі,*** 1997 (NTIS PB97-162580)
27. ***Навчання на майбутнє: звертаючись до завтрашніх завдань з опитування,*** 1998 (NTIS PB99-102576)
28. ***Семінар по міжвідомчій координації і співпраці,*** 1999 (NTIS PB99-132029)
29. ***Конференція Федерального комітету із дослідження статистичної методології (Матеріали конференції),*** 1999 (NTIS PB99-166795)
30. ***Конференція Федерального комітету із дослідження статистичної методології 1999 року: повна процедура проведення,*** 2000 (NTIS PB2000-105886)
31. ***Вимірювання і доповідання джерел помилки в опитуваннях,*** 2001 (NTIS PB2001-104329)
32. ***Семінар по інтегруванні федеральної статистичної інформації і процесів,*** 2001 (NTIS PB2001-104626)
33. ***Семінар по можливості фінансування в оглядовому дослідженні,*** 2001 (NTIS PB2001-108851)
34. ***Конференція Федерального комітету із дослідження статистичного розкриття (матеріали конференції),*** 2001 (NTIS PB2002-100103)
35. ***Семінар по складних завданнях Федеральної статистичної системи із сприяння доступу до статистики.*** 2004.
36. ***Семінар по можливості фінансування в опитуванні і статистичному дослідженні.*** 2004.
37. ***Конференція Федерального комітету із дослідження статистичного розкриття (Матеріали конференції),*** 2003.

38. ***Підсумковий звіт семінару FCSM-GSS по збору даних, основаному на Інтернет-технологіях.*** 2004.
Копії цих робочих документів можна замовляти із «Продажу документів NTIS (Національна служба технічної інформації)», 5285 Port Royal Road, Springfield, VA 22161; телефон: 1-800-553-6847. Серія Робочого документу із статистичної політики є також доступний в електронному вигляді із веб-сайту FCSM <*http://www.fcsm.gov*>.