# Пертурбативные методы для таблиц по итогам переписи 2021 года

**Презентация Эрика Шульте Нордхольта, Женева, сентябрь 2022 года**

Statistics Netherlands

# План

- Перепись 2021 года и конфиденциальность
- Проект «Гармонизированная защита данных переписи в Европейской статистической системе (ЕСС)»
- Определение тестовых сценариев
- Инструменты для ЦЗЗ и МКЯ
- Вопросы

# Перепись 2011 года и конфиденциальность (1)

Данные европейской переписи населения 2011 года представляют собой важный источник статистической информации о естественном движении населения от самого низкого уровня малых районов до национального и международного уровней

Гармонизированные таблицы по итогам переписи в 32 европейских странах доступны на сайте Центра переписей (https://ec.europa.eu/CensusHub2/)

Данные переписей являются подробными и конфиденциальными; государства-члены несут ответственность за защиту данных переписи

# Перепись 2011 года и конфиденциальность (2)

Хотя результаты переписей гармонизированы, сравнение данных из разных стран затруднено из-за того, что в них применяются различные подходы к обеспечению конфиденциальности статистических данных

Два специальных соглашения о выделении грантов (ССГ) посвящены поиску и тестированию эффективных методик для проведения переписи 2021 года:

- ССГ по гармонизированной защите данных переписей в Европейской статистической системе (ЕСС) (2016 -2017)
- ССГ по пертурбативным методам обеспечения конфиденциальности (2018-2019)

# Проект «Гармонизированная защита данных переписи в ЕСС» (1)

- Начало: 1 сентября 2016 г.
- Завершение: 31 августа 2017 г.

- Четыре комплекса работ (КР):
  - КР 1 Управление (7 результатов работы)
  - КР 2 Анкета (2 результата работы)
  - КР 3 Разработка и тестирование рекомендаций; поиск передовых методик (4 результата работы)
  - КР 4 Распространение (5 результатов работы)

# Проект «Гармонизированная защита данных переписи в ЕСС» (2)

- Участвуют шесть стран:
  - Центральное статистическое управление Нидерландов (Эрик Шульте Нордхольт, Петер-Пауль де Вольф),
  - Национальный институт статистических и экономических исследований Франции (Маэль-Люк Бюрон),
  - Федеральное статистическое управление Германии (Сара Гисинг, Тобиас Эндерле),
  - Центральное статистическое управление Венгрии (Ласло Антал, Беата Наги),
  - Статистическое управление Финляндии (Анну Кабрера) и
  - Статистическое управление Республики Словении (Андреа Смукавец, Юнош Лукан)

# Проект «Гармонизированная защита данных переписи в ЕСС» (3)

- Проанализированы регламенты и методы защиты данных в отдельных странах
- Обеспечен гармонизированный подход к защите данных переписи 2021 года (с учетом национальных ограничений)
- Государствам-членам рекомендованы приемлемые методы обеспечения конфиденциальности статистических данных для гиперкубов
- Даны рекомендации об эффективном обращении с конфиденциальными ячейками в квадратах сетки и при разбивке по регионам (риск рассекречивания данных при вычислении разности)

# Определение тестовых сценариев (1)

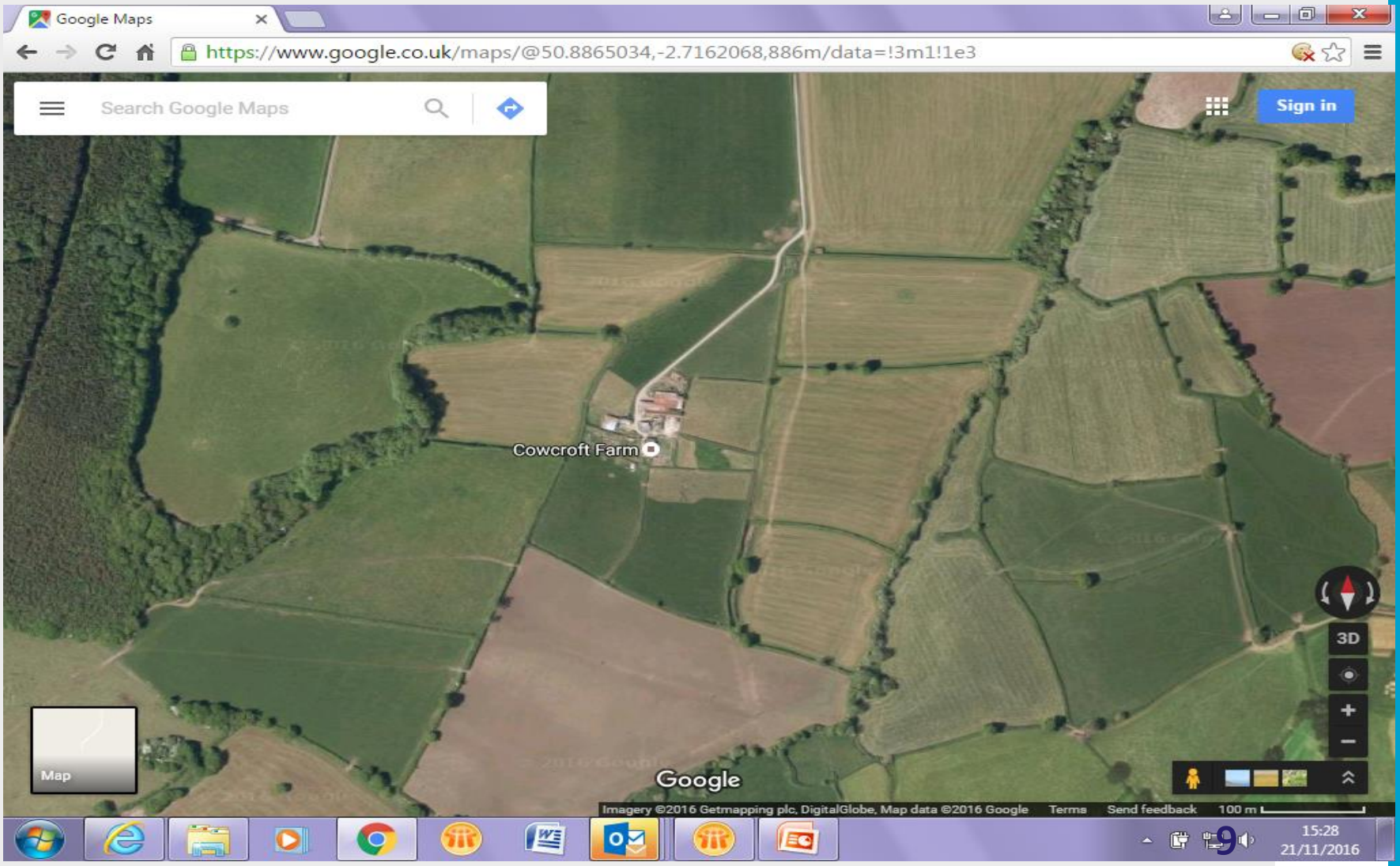
Ограничения:

– Нельзя менять структуру таблиц или осуществлять глобальное перекодирование (структура гиперкубов зафиксирована в исполнительном распоряжении о порядке проведения переписи населения в ЕС)
– Нельзя блокировать ячейки (очень сложно для связанных многомерных гиперкубов, а иначе невозможно рассчитать итоговые значения для Европы)

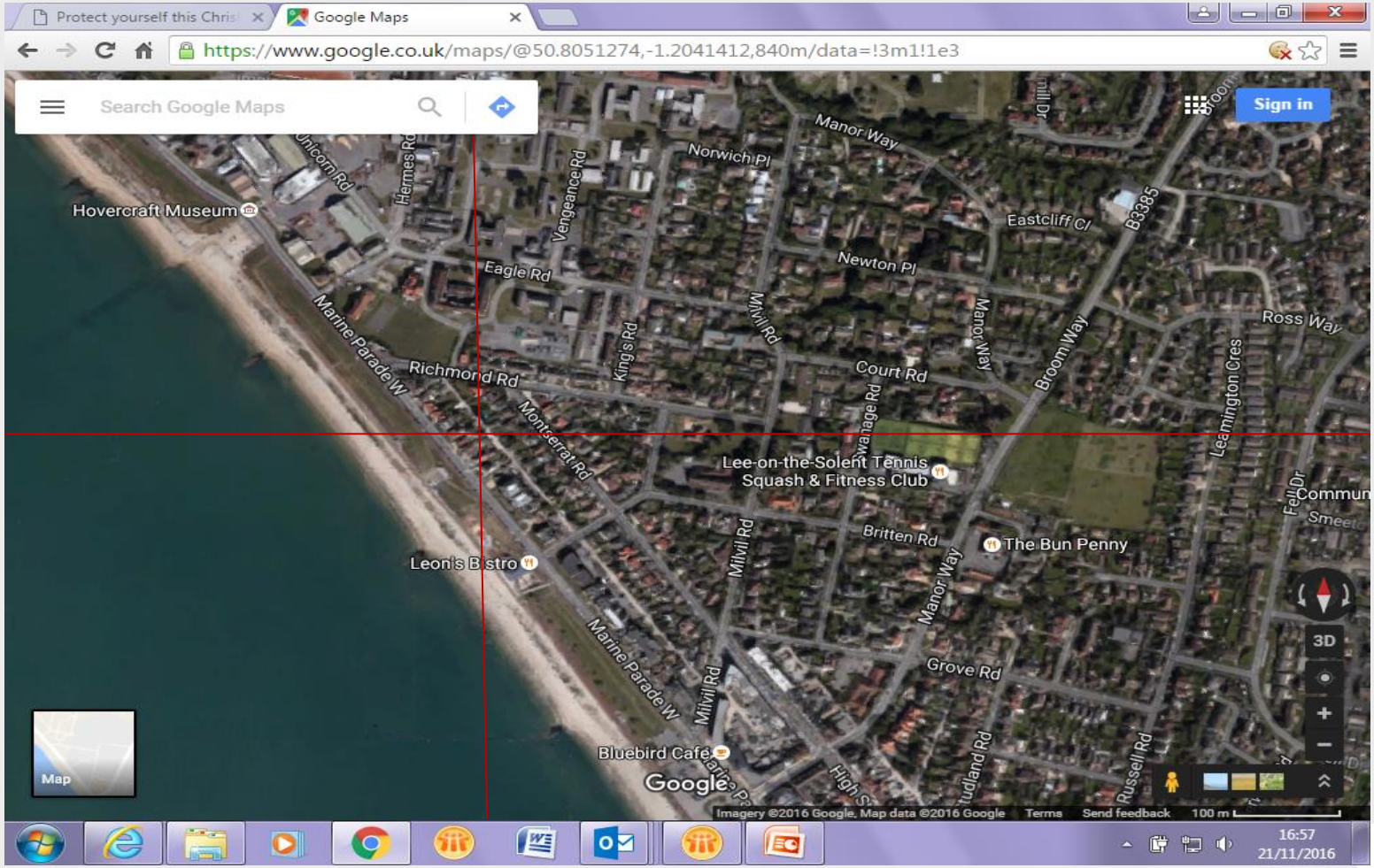
Сложности:

– Ячейки сетки площадью 1 кв.км - множество малых значений в ячейках
– Ячейки сетки площадью 1 кв.км ⇔ административные регионы (риск рассекречивания данных при вычислении разностей)

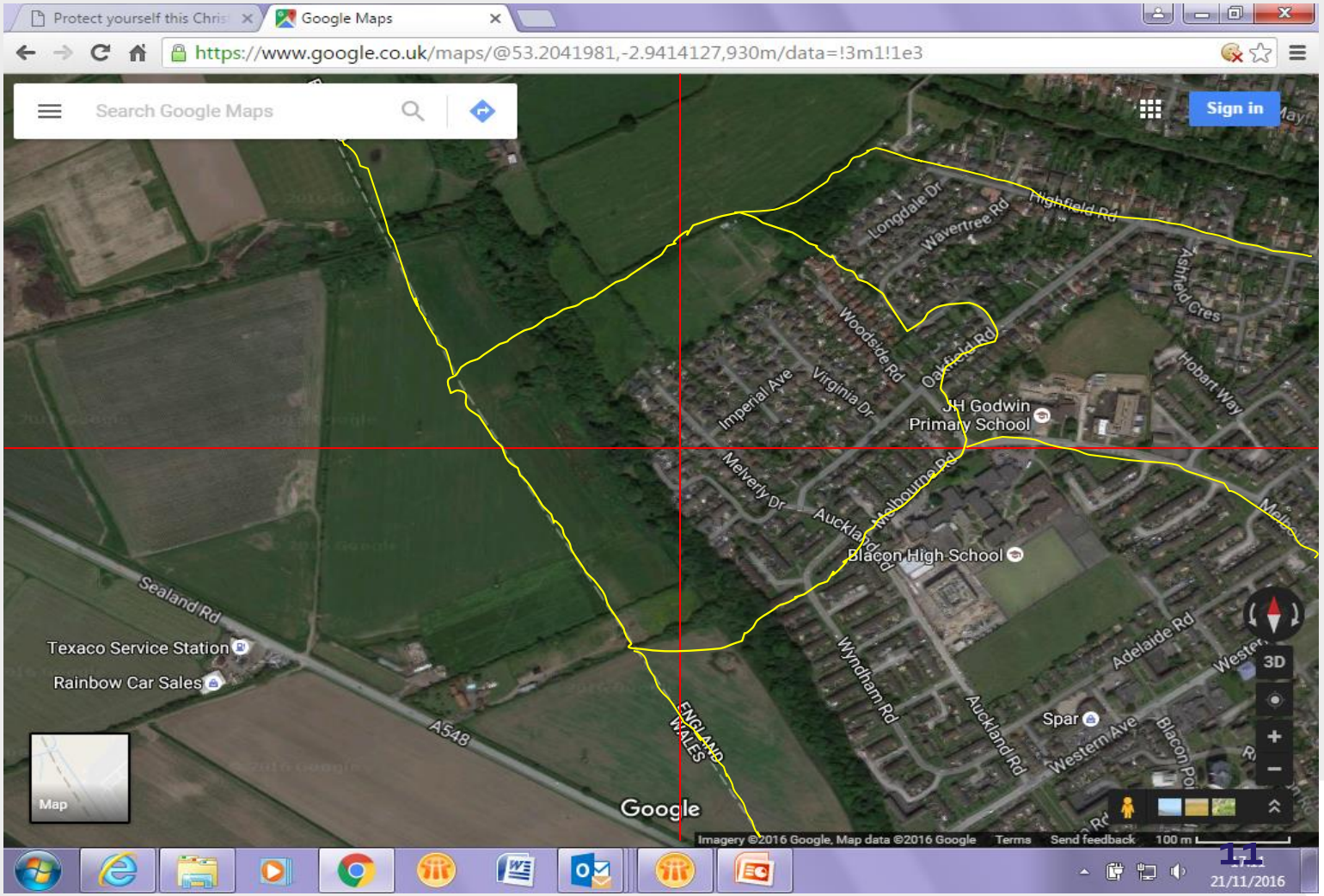
# Определение тестовых сценариев (2)



# Определение тестовых сценариев (3)



# Определение тестовых сценариев (4)



# Определение тестовых сценариев (5)

Решение, позволяющее обеспечить конфиденциальность статистической информации, не должно слишком сильно изменять пространственное распределение данных в сетке:

- Сетки с нулевой частотой не следует слишком часто менять на положительные частоты
- Редко встречающиеся ненулевые частоты в каком-либо районе не следует существенно менять

Обычные риски рассекречивания:

- Малые количества (могут приводить к непосредственной идентификации)
- Рассекречивание атрибута (положительная частота может приводить к рассекречиванию информации из гиперкуба)

# Определение тестовых сценариев (6)

Гибкий метод, который можно адаптировать к национальным потребностям государств-членов:

– Метод замены записей до составления таблицы
– Метод ключа ячейки после составления таблицы

Преимущество: появляются (оценочные значения для) всех ячеек

Недостатки: может быть велика относительная погрешность для малых значений ячеек, а также в таблицах утрачивается суммируемость

# Определение тестовых сценариев (7)

Метод замены записей и ключа ячейки:

- Доработанный вариант метода ключа ячейки, разработанный Австралийским бюро статистики (ABS)
- Предоставлен Национальной статистической службой (НСС) и адаптирован в рамках данного проекта

# Инструменты для ЦЗЗ и МКЯ (1)

- Новое специальное соглашение о выделении гранта (ССГ) начало действовать весной 2018 года: ССГ «Инструменты с открытым кодом для пертубативных методов обеспечения конфиденциальности» - срок действия 15 месяцев (до осени 2019 года)
- Цель этого нового ССГ - интегрировать полученные коды в легкие в использовании пакеты программного обеспечения с открытым исходным кодом
- Семь стран, участвующих в данном ССГ: Австрия, Финляндия, Франция, Германия, Венгрия, Нидерланды и Словения
- Программное обеспечение для тестирования данных переписи (целевая замена записей и метод ключа ячейки) доступно на сайте github (https://github.com/sdcTools/CensusProtection)

# Инструменты для ЦЗЗ и МКЯ (2)

Целевая замена записей — это предтабличный метод (изменения в микроданных)

Подготовка:

- **Указать переменные, определяющие риск (k-анонимность)**
- **Указать переменные, определяющие региональную иерархию**
- **Рассчитать риск для всех домохозяйств на каждом региональном уровне**
- **Указать переменные, определяющие «похожие» домохозяйства**
- **Указать минимальный показатель замены**

# Инструменты для ЦЗЗ и МКЯ (3)

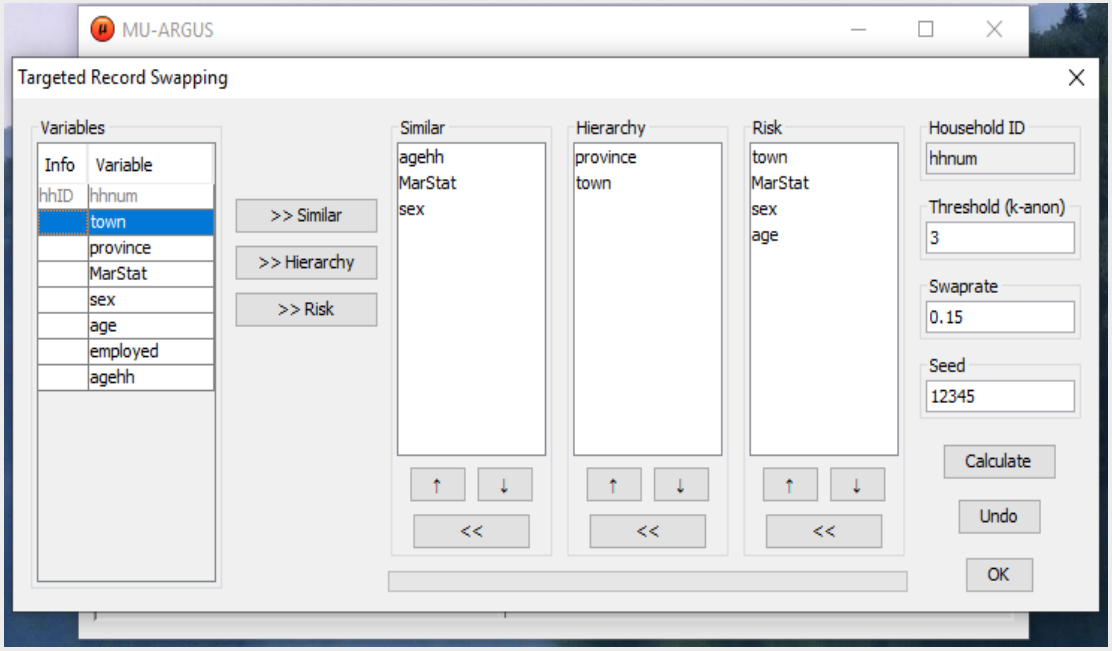
Для реализации ЦЗЗ необходимо идти с самого высокого до самого низкого регионального уровня:

- **Создать донорский набор домохозяйств**
  - «Похожие» домохозяйства как домохозяйства высокого риска
- **Выбрать донорское домохозяйство для домохозяйства высокого риска**
  - Тот же региональный *уровень*, другой *регион*
  - Поменять местами все региональные переменные
- **При недостижении минимального уровня замены необходимо поменять местами дополнительные домохозяйства на самом низком региональном уровне**

# Инструменты для ЦЗЗ и МКЯ (4)

Код C++

Возможность вызова из μ-Argus



# Инструменты для ЦЗЗ и МКЯ (5)

Метод ключа ячейки - это послетабличный метод (шум добавляется к ячейкам таблицы)

1. **Определить таблицу вероятностей**
2. **Определить $\mathcal{U}(0,1)$ значение для каждой записи = *ключ записи***
3. **Суммировать ключи записей для записей в каждой ячейке таблицы, дробная часть этой суммы становится *ключом ячейки* для каждой ячейки таблицы**
4. **Использовать *значение ячейки*, И *ключ ячейки*, И таблицу вероятностей, чтобы получить количество шума для добавления к этой ячейке**

# 1. Определить таблицу вероятностей

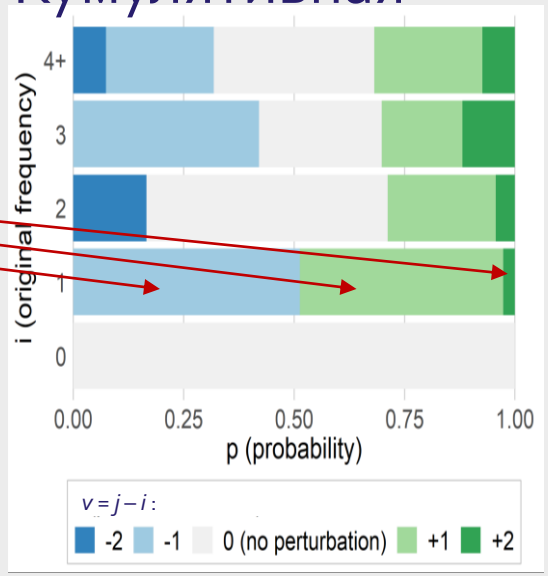
Таблица вероятностей: вероятности перехода (пакет в R `таблица вероятностей`)

$p_{ij}$ = P (значение ячейки i изменяется на значение j)

Например, вероятности

| i \ j | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0,5133 | 0 | 0,4600 | 0,0267 | 0 | 0 | 0 |
| 2 | 0,1656 | 0 | 0,5463 | 0,2449 | 0,0432 | 0 | 0 |
| 3 | 0 | 0 | 0,4208 | 0,2776 | 0,1824 | 0,1192 | 0 |
| 4 | 0 | 0 | 0,0739 | 0,2442 | 0,3637 | 0,2442 | 0,0739 |

Кумулятивная



# 2.–3. Определить ключи записей и получить ключи ячеек

| Идентификатор | Пол | Возраст | Ключ записи |
|---|---|---|---|
| 1 | М | А | 0,34582249 |
| 2 | Ж | В | 0,68438579 |
| 3 | Ж | В | 0,95880618 |
| 4 | Ж | С | 0,62902289 |
| 5 | М | В | 0,86598721 |
| 6 | Ж | С | 0,36307981 |
| 7 | М | А | 0,91420393 |
| 8 | М | А | 0,69629390 |
| 9 | М | В | 0,53460054 |
| 10 | Ж | В | 0,68511663 |
| 11 | Ж | В | 0,03426370 |
| 12 | М | В | 0,33696811 |
| 13 | Ж | В | 0,11181613 |
| 14 | Ж | А | 0,56526973 |
| 15 | М | А | 0,01047942 |

Пол = М

Возраст = В

Сумма=1,73755586

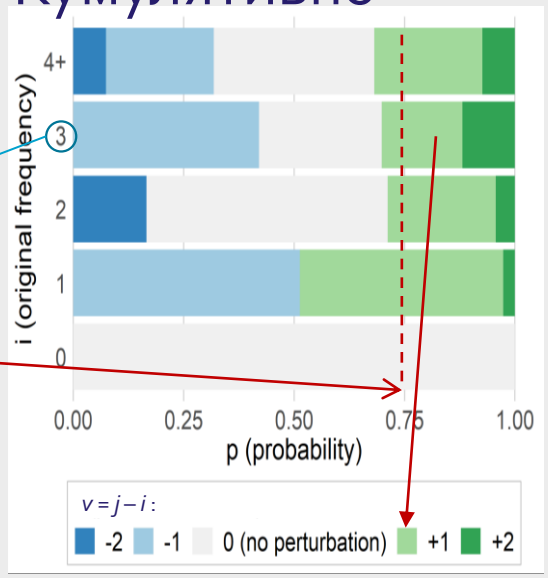
| Пол | Возраст | *i* | Ключ ячейки |
|---|---|---|---|
| Т | Т | 15 | 0,73611646 |
| Т | А | 5 | 0,53206947 |
| Т | В | 8 | 0,21194429 |
| Т | С | 2 | 0,99210270 |
| М | Т | 7 | 0,70435560 |
| М | А | 4 | 0,96679974 |
| М | В | 3 | 0,73755586 |
| М | С | 0 | 0 |
| Ж | Т | 8 | 0,03176086 |
| Ж | А | 1 | 0,56526973 |
| Ж | В | 5 | 0,47438843 |
| Ж | С | 2 | 0,99210270 |

# 4. Определить количество шума для добавления

| Пол | Возраст | *i* | Ключ ячейки |
|---|---|---|---|
| Т | Т | 15 | 0,73611646 |
| Т | А | 5 | 0,53206947 |
| Т | В | 8 | 0,21194429 |
| Т | С | 2 | 0,99210270 |
| М | Т | 7 | 0,70435560 |
| М | А | 4 | 0,96679974 |
| М | В | 3 | 0,73755586 |
| М | С | 0 | 0 |
| Ж | Т | 8 | 0,03176086 |
| Ж | А | 1 | 0,56526973 |
| Ж | В | 5 | 0,47438843 |
| Ж | С | 2 | 0,99210270 |

(M, B): $i = 3$

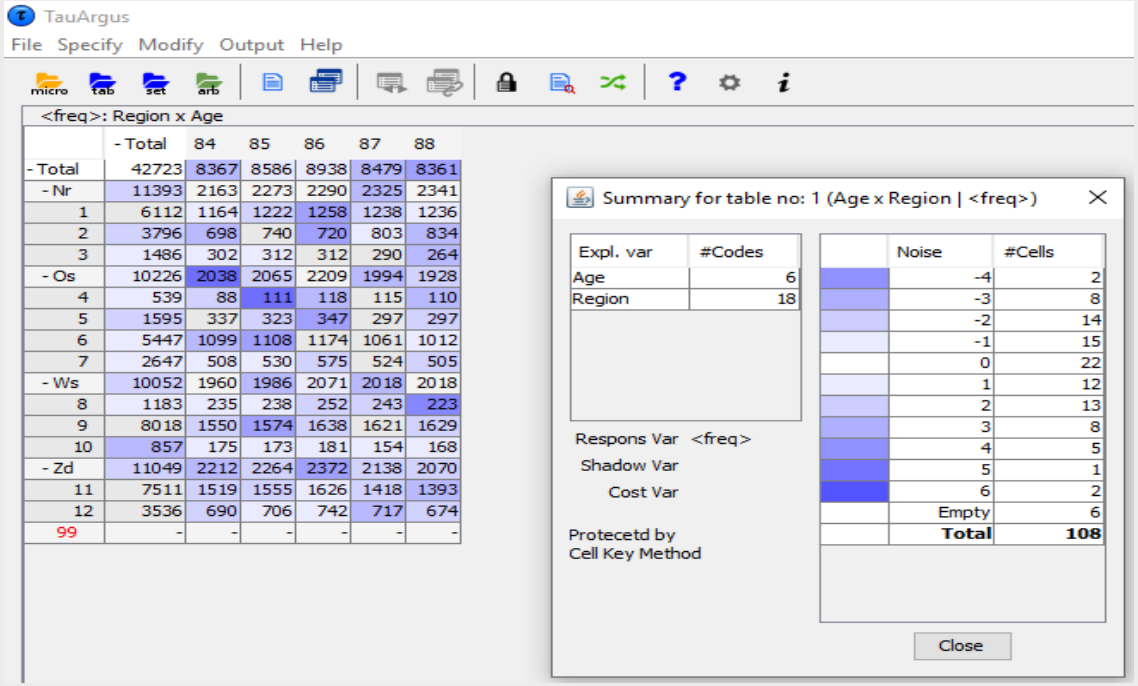
Кумулятивно




(M, B): $j = i + 1 = 4$

# Инструменты для ЦЗЗ и МКЯ (6)

Добавление шума методом ключа ячейки входит в τ-Argus и также вызывается из R




| | - Total | 84 | 85 | 86 | 87 | 88 |
|---|---|---|---|---|---|---|
| - Total | 42723 | 8367 | 8586 | 8938 | 8479 | 8361 |
| - Nr | 11393 | 2163 | 2273 | 2290 | 2325 | 2341 |
| 1 | 6112 | 1164 | 1222 | 1258 | 1238 | 1236 |
| 2 | 3796 | 698 | 740 | 720 | 803 | 834 |
| 3 | 1486 | 302 | 312 | 312 | 290 | 264 |
| - Os | 10226 | 2038 | 2065 | 2209 | 1994 | 1928 |
| 4 | 539 | 88 | 111 | 118 | 115 | 110 |
| 5 | 1595 | 337 | 323 | 347 | 297 | 297 |
| 6 | 5447 | 1099 | 1108 | 1174 | 1061 | 1012 |
| 7 | 2647 | 508 | 530 | 575 | 524 | 505 |
| - Ws | 10052 | 1960 | 1986 | 2071 | 2018 | 2018 |
| 8 | 1183 | 235 | 238 | 252 | 243 | 223 |
| 9 | 8018 | 1550 | 1574 | 1638 | 1621 | 1629 |
| 10 | 857 | 175 | 173 | 181 | 154 | 168 |
| - Zd | 11049 | 2212 | 2264 | 2372 | 2138 | 2070 |
| 11 | 7511 | 1519 | 1555 | 1626 | 1418 | 1393 |
| 12 | 3536 | 690 | 706 | 742 | 717 | 674 |
| 99 | - | - | - | - | - | - |

# Инструменты для ЦЗЗ и МКЯ (7)

Тестирование ЦЗЗ проведено с использованием наборов данных, а тестирование МКЯ — с использованием больших гиперкубов переписи 2011 года с приемлемым временем исполнения

Помощь доступна через github:
https://github.com/sdcTools/UserSupport/wiki

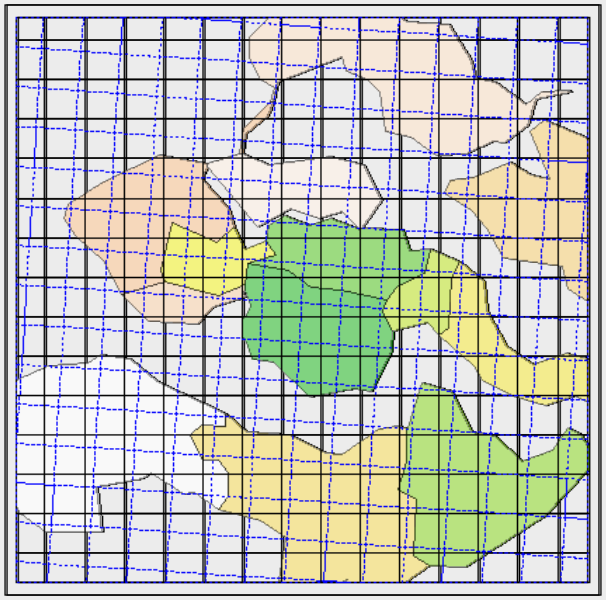


# Инструменты для ЦЗЗ и МКЯ (8)

Дальнейшие действия:

- Предложены меры риска и полезности
- В разных странах продолжается работа над поиском соответствующих значений параметров
- Методы известны, программное обеспечение доступно, тестирование идет
- Важно впоследствии донести информацию о результатах работы!
- Следующий обучающий курс в рамках Европейской статистической программы обучения запланирован на 25-27 января 2023 года в Люксембурге

# Иллюстрации разных сеток



# Вопросы

Вопросы или замечания?