Организация Объединенных Наций ECE/CES/GE.41/2022/3

Экономический и Социальный Совет

Distr.: General
23 June 2022
Russian
Original: English

Европейская экономическая комиссия

Конференция европейских статистиков

Группа экспертов по переписям населения и жилищного фонда

Двадцать четвертое совещание
Женева, 21–23 сентября 2022 года
Пункт 5 предварительной повестки дня
Переход в методологии переписи; планы, опыт и инновации

Пертурбативные методы составления таблиц переписей 2021 года в рамках Европейской рамочной программы

Записка Статистического управления Нидерландов*

Резюме

Переписи составляют неотъемлемую часть статистической программы национальных статистических ведомств на протяжении многих десятилетий. Европейская перспектива стала важным дополнительным аспектом распространения всех различных результатов переписи. Объединение результатов национальных переписей, очевидно, требует определенной координации и согласования. В качестве первого шага была проведена гармонизация плана таблиц выходных данных переписей для европейского измерения. Это явно облегчает объединение таблиц переписей разных стран. Однако, поскольку государства-члены использовали разные методы контроля за раскрытием информации для защиты частной жизни своих жителей, преимущества согласованного плана таблиц оказались не такими большими, как предполагалось. Таким образом, также необходима гармонизация используемых методов контроля раскрытия статистических данных. Чтобы предложить такой гармонизированный подход к контролю за раскрытием таблиц переписей, были начаты два европейских проекта. Государства-члены не обязаны использовать предлагаемые методы, но, если их будут использовать многие страны, это определенно улучшит сопоставимость таблиц европейских переписей. Этот документ частично основан на результатах двух европейских проектов, направленных на разработку и внедрение согласованного подхода. Кроме того, мы выделим некоторые оставшиеся вопросы, которые следует рассмотреть, когда предлагаемые методы будут использоваться для европейской переписи 2021 года.

* Подготовлена Эриком Шульте Нордхольтом и Питером-Полем де Вольфом.
Примечание: Употребляемые обозначения в настоящем документе не означают выражения со стороны Секретариата Организации Объединенных Наций какого бы то ни было мнения относительно правового статуса той или иной страны, территории, города или района или их властей или относительно делимитации их границ.

I. Введение

1. 2021 год — год Европейской переписи населения. Это означает, что все государства — члены Европейского союза (ЕС) должны провести перепись населения и жилого фонда с отсчетной датой в 2021 году (днем переписи). Это важное средство согласования результатов европейской переписи. Более того, все страны ЕС опубликуют набор согласованных таблиц, позволяющих проводить сопоставления. Этот набор связанных многоразмерных таблиц дает точное описание людей, живущих в ЕС, и их жилищной ситуации. Этот набор таблиц называется гиперкубами Европейской переписи 2021 года. Кроме того, впервые набор таблиц с квадратной сеткой станет обязательным для Европейской переписи населения 2021 года.

2. Опыт переписи 2011 года в Европе показал, что требуется дальнейшая гармонизация, позволяющая сделать данные более сопоставимыми на международном уровне. Различные европейские страны применяли существенно различающиеся методы для защиты своих таблиц переписи 2011 года, что в значительной степени затрудняло возможность сопоставления результатов по странам. Юридически невозможно предписать, как именно должны быть защищены гиперкубы Европейской переписи 2021 года. Однако, обменявшись информацией о передовом опыте и рекомендуя методы защиты этих таблиц, можно сделать важный шаг вперед. В данной записке представлены эти рекомендации. Рекомендации основаны на опыте многих отдельных стран ЕС и за его пределами. Большой прогресс был достигнут в двух европейских проектах. Основные результаты этих проектов описаны в данной записке.

3. В разделе II представлена историческая ретроспектива. Предлагаемые методы защиты описаны в разделе III. В разделе III также обсуждается комбинирование этих методов. Раздел IV показано, как эти методы оценивались до сих пор в Нидерландах. Записка завершается заключением в разделе V. Заключение содержит некоторые замечания по вопросам, которые могут возникнуть при использовании предложенных методов на практике.

II. Историческая ретроспектива

4. Переписи служат важными массивами данных. Все государства — члены Европейского союза проводят переписи населения и жилого фонда для получения исчерпывающих данных о своем населении. Государства — члены Европейского союза должны предоставлять данные переписи Евростату, статистическому управлению Европейского союза. Евростат компилирует данные переписи европейского уровня на основе данных, предоставленных государствами-членами. В большинстве государств-членов данные переписи могут быть опубликованы только в том случае, если были приняты меры для предотвращения раскрытия информации об отдельных респондентах. Таким образом, контроль раскрытия статистической информации (КРС) является важным шагом перед публикацией данных переписи. В Hundepool et al. (2012) содержится более подробная общая информация о контроле за раскрытием статистической информации. В этой записке мы рассматриваем конкретный аспект защиты подробных и связанных таблиц переписей.

5. Для Европейской переписи 2001 года отсутствовала правовая основа. Было достигнуто лишь джентльменское соглашение о том, что все государства-члены сделают все возможное для предоставления таблиц переписи в Евростат. Ясно, что это не было достаточно прочной основой для составления всех таблиц, требуемых от всех государств-членов.

6. Ситуация с европейской переписью 2011 года улучшилась благодаря введению в действие Акта о европейской переписи населения (European Commission, 2008). Поскольку передача микроданных переписи в Евростат столкнулась с юридическими препятствиями, в то время была введена концепция гиперкубов переписи. Гиперкубы — это многомерные таблицы. Из этих таблиц можно составить множество более простых таблиц для целей публикации. В то же время формат, в котором должны были предоставляться таблицы, изменился с Excel на SDMX.

7. К таблицам переписи можно применять обычные не связанные с теорией возмущений методы КРС. Эти методы включают изменение плана таблицы, глобальное перекодирование и локальное подавление. Макеты таблиц Европейской переписи были исправлены, чтобы облегчить сопоставление данных разных стран. В переписи 2011 года это было сделано с использованием обязательного нового формата SDMX. Однако эти фиксированные макеты таблиц также подразумевают, что изменение плана таблиц и глобальные перекодировки больше не являются вариантами защиты таблиц от раскрытия индивидуальной информации, в то время как в переписи 2001 года эти два метода широко применялись несколькими странами.

8. Может показаться приемлемой альтернативой локальное подавление для защиты многомерной таблицы. Действительно, также часто можно одновременно защитить несколько связанных таблиц. Однако набор гиперкубов переписи 2011 года был слишком большим и сложным, чтобы его можно было оптимально защитить с помощью локальных подавлений. Под оптимальным мы понимаем получение набора подавлений, при котором ни одна из первичных небезопасных ячеек не может быть (приблизительно) пересчитана, в то время как гиперкубы сохраняют достаточное количество информации, чтобы оставаться полезными для пользователей.

9. Проблема того, как должным образом защитить набор гиперкубов переписи, была признана, и в 2008 году была создана Целевая группа по контролю за раскрытием статистических данных переписей. Работа Целевой группы осложнялась тем, что еще не существовало настоящих гиперкубов, а странам по закону не разрешалось делиться своими старыми микроданными переписей. Кроме того, поскольку страны несут ответственность за защиту своих таблиц переписей, нельзя в обязательном порядке предписать, как следует защищать гиперкубы переписей. В итоге все страны защищали свои гиперкубы переписи 2011 года своими способами. Еще одна сложность была связана с тем, что у представителей разных стран были неодинаковые представления о том, какая информация конфиденциальна и нуждается в защите. Этот результат разочаровал.

10. Гармонизация, намеченная Целевой группой, должна была привести к более сопоставимым результатам по странам, но на практике гиперкубы, доступные в Census Hub (см. URL: https://ec.europa.eu/CensusHub2/), не всегда были сопоставимы между странами из-за большого разнообразия различных методов защиты. Некоторые страны вообще не защищали своих таблиц, многие страны ввели пропущенные значения в чувствительные ячейки и множество других ячеек для защиты чувствительных ячеек, а другие страны по-разному добавляли шум для защиты своих гиперкубов.

11. В гиперкубах переписи 2011 года в Нидерландах были опущены две чувствительные переменные со многими категориями, что имело существенные последствия. Применительно к этим переменным страны рождения и страны гражданства были опубликованы только агрегированные данные, а информация по отдельным странам была скрыта, когда данные были преобразованы в формат SDMX и опубликованы в Census Hub. Разумеется, это защищало конфиденциальную информацию, но потеря информации была огромной и привела к множеству запросов о предоставлении дополнительных таблиц в последующие годы. Такие запросы обычно принимались, если не запрашивались многомерные таблицы, поскольку многомерные таблицы по-прежнему приводили к раскрытию индивидуальной информации. Очевидно, что Статистическое управление Нидерландов не стремилось использовать этот метод и при составлении таблиц переписи 2021 года. Точно так же многие другие страны не были удовлетворены своим методом КРС применительно к таблицам переписи 2011 года.

12. Стало ясно, что процесс подготовки результатов европейской переписи необходимо усовершенствовать. Закон о европейской переписи населения (European Commission, 2008) является правовой основой проведения европейских переписей 2011 и 2021 годов. Для переписи 2021 года дополнительно были приняты четыре исполнительных регламента, точно определяющих, что должны предоставить государства-члены (European Commission, 2017a, 2017b, 2017c, 2018). Более того, чтобы попытаться предотвратить подобную нежелательную ситуацию с защитой гиперкубов для Европейской переписи 2021 года, в последние годы были

подписаны два Соглашения о специальных грантах (ССГ) в рамках Рамочного программного соглашения (РПС) № 11112.2014.005-2014.533. Результаты этих двух ССГ рассматриваются в двух следующих разделах.

III. Предлагаемые методы

A. Введение

13. Реализация первого ССГ, упомянутого в разделе II (№ 11112.2016.005-2016.367), началась в сентябре 2016 года и продолжалась один год. Были привлечены статистические ведомства шести европейских стран (Венгрии, Германии, Нидерландов, Словении, Финляндии и Франции), а Статистическое управление Нидерландов выступило в качестве руководителя проекта. В ходе этого проекта «Гармонизированная защита данных переписей в Европейской статистической системе (ЕСС)» среди стран ЕСС был проведен опрос о защите их таблиц переписей. Очевидно, его анкета включала вопросы по юридическим, методологическим и техническим аспектам. Цель ССГ «Гармонизированная защита данных переписей в ЕСС» состояла в том, чтобы предоставить рекомендации по защите таблиц переписи 2021 года. Такие рекомендации могут быть сделаны должным образом только в том случае, если будут приняты во внимание национальные (правовые) ситуации, и поэтому был задан ряд вопросов об этих ситуациях. Кроме того, были заданы вопросы об оценках странами своих методов защиты гиперкубов переписи 2011 года и об использовании ячеек сетки в (национальных) гиперкубах переписи. В итоге были получены ответы от 33 европейских стран (27 из тогдашних 28 государств-членов и 6 из 7 присоединяющихся стран).

14. Обратите внимание, что рекомендации, сделанные в этом ССГ, не подразумевают юридических обязательств перед странами, участвующими в ЕСС. Цель этого ССГ состояла в том, чтобы предоставить рекомендации по получению хорошо защищенных таблиц переписи, которые легко сопоставлять между странами.

B. Выводы, сделанные в ССГ

15. Законы стран, регулирующие публикацию результатов переписи, часто нечетко определяют, как и что защищать. Гиперкубы переписи 2021 года представляют собой набор связанных многомерных таблиц. Это означает, что многие ячейки таблицы переписи будут иметь очень низкое значение или даже будут равны 0. Таким образом, индивидуальная информация может быть относительно легко раскрыта из гиперкубов переписи 2021 года. Это ясно показывает, что полное отсутствие мер контроля за раскрытием информации было бы незаконным для всех стран ЕСС. Точно так же, несмотря на то, что понятие чувствительности различается между странами, по общему мнению, наиболее проблематичными переменными переписи, по-видимому, являются страна/место рождения (МР) и страна гражданства (СГ). В частности, наиболее подробный уровень этих переменных (отдельные страны) может помочь в раскрытии индивидуальной информации в гиперкубах, в которых они появляются.

16. В ходе опроса многие страны отметили, что посттабличные методы не пользуются популярностью. Однако, на наш взгляд, без посттабличных методов будет практически невозможно должным образом защитить переписные гиперкубы. Это связано с тем, о чем известно большинству стран: при защите гиперкубов переписи следует также обращаться к национальным таблицам переписи. Действительно, даже если и европейские гиперкубы, и национальные таблицы защищены должным образом *самостоятельно*, комбинация публикаций необязательно безопасна. Как следствие, если, например, национальные таблицы публикуются первыми, европейские гиперкубы должны быть защищены условно на основе уже опубликованной информации. Аналогичная ситуация может возникнуть при рассмотрении других (национальных) демографических публикаций.

17. В переписи 2021 года к европейским гиперкубам добавлены новые виды таблиц: таблицы на квадратах сетки 1 км × 1 км. Эти таблицы не детализированы по содержанию (в каждую из этих таблиц включена только одна характеристика), но детализированы по структуре (число квадратов сетки во всех странах намного больше, чем число муниципалитетов). Кроме того, квадраты сетки и региональные распределения не являются вложенными переменными. Это означает, что страны должны проверить, может ли быть раскрыта информация о лицах путем пересечения этих квадратов сетки с муниципалитетами (МАУ), уровнем максимальной детализации региона в европейских гиперкубах. Другие уровни географии (страна, НТЕС1, НТЕС2 и НТЕС3) в гиперкубах представляют собой комбинации МАУ. Кроме того, эти уровни являются вложенными, т. е. имеют иерархическую структуру. Квадраты сетки — единственная географическая переменная, не вложенная в эту иерархическую структуру.

18. Государства-члены готовят и предоставляют гиперкубы переписей для Евростата. Переменные каждого гиперкуба и их категории унифицированы по странам. Таким образом, данные государств-членов могут быть объединены в данные европейского уровня. Однако государства-члены могут применять методы КРС по своему выбору, и различия в используемых методах между странами могут негативно сказаться на качестве данных европейского уровня. Евростат стремится гармонизировать методы КРС в государствах-членах, чтобы повысить качество данных. Чем больше государств-членов будут применять рекомендуемые методы КРС, тем более гармонизированными могут стать данные на европейском уровне.

19. Классические непертурбативные методы, такие как глобальное перекодирование и подавление ячеек, по разным причинам не подходят для защиты европейских таблиц переписи. Чтобы сделать возможным сравнение между странами, форматы таблиц фиксированы и не могут быть изменены. Поэтому глобальная перекодировка не подходит. Оптимальное применение подавления ячеек к такому большому набору многомерных связанных таблиц практически невозможно. Теоретически можно было бы применить подавление ячеек с большим избыточным подавлением, чтобы набор таблиц стал защищенным. Однако это приведет к огромным потерям информации, что неприемлемо с точки зрения пользователя. Еще одна проблема заключается в управлении риском раскрытия путем разграничения данных гиперкуба и данных на уровне сетки, что еще больше усложняет концепции защиты на основе подавления ячеек.

20. Согласованный метод должен обеспечивать некоторую гибкость, чтобы страны могли легко адаптировать его к своим конкретным потребностям и ожиданиям в отношении приемлемого уровня остаточного риска раскрытия информации, с одной стороны, и приемлемого уровня потери информации, с другой стороны. Метод должен быть адаптируемым путем изменения параметров и должен предусматривать возможность использования отдельных модулей в их комбинации. Затем родилась идея включить модули предтабличного возмущения, а также модули посттабличного возмущения.

21. Поэтому команда проекта решила выбрать предтабличный метод целевой замены данных и посттабличный метод ключа ячеек, при котором к ячейкам таблицы добавляется шум. Предлагаемые методы кратко представлены в следующих двух подразделах. Параметры обоих методов не фиксированы; государства-члены могут принимать решения по ним самостоятельно. Оба метода не приводят к подавлению данных, поэтому данные государств-членов, если они обрабатываются этими методами, могут быть объединены в данные европейского уровня.

22. Если многие государства-члены будут использовать один и тот же метод, хотя, возможно, и в разных вариантах, это упростить подготовку данных на европейском уровне. В отличие от подавления ячеек при предложенных пертурбативных методах данные будут доступны для всех ячеек гиперкуба. Это будет большим преимуществом для всех пользователей и значительно повысит сопоставимость данных по странам.

23. Для обеспечения согласованности между европейскими и национальными выпусками данных государствам-членам рекомендуется применять один и тот же

метод КРС ко всем типам выпусков данных. Однако, если для защиты данных о национальных выпусках используется другой метод, государства-члены должны проверить и в конечном итоге разработать варианты, позволяющие избежать остаточных рисков раскрытия информации, которые могут возникнуть, когда пользователи сравнивают европейские данные гиперкуба с национальными выпусками.

C. Целевая перестановка данных

24. Перестановка данных — это предтабличный метод КРС, поэтому он применяется к микроданным до построения гиперкубов переписи. Общая идея перестановки данных заключается в том, что выбираются пары данных и значения определенных переменных переставляются между ними. Выбор данных обычно проводится таким образом, чтобы сохранить некоторые конкретные аналитические свойства и свести к минимуму вносимую погрешность.

25. Целевая перестановка данных (ЦПД), предложенная для гиперкубов переписи 2021 года, основана на методе, разработанном Управлением национальной статистики Великобритании (УНС). Пришлось внести небольшие коррективы, поскольку при реализации этого метода УНС ориентировалось на конкретную ситуацию в Великобритании, а ССГ было нацелено на общую реализацию, применимую ко всем государствам-членам.

26. ЦПД применяется на уровне домохозяйств в том смысле, что перестанавливаются только полные данные по домохозяйствам, а не отдельным лицам. Это один из способов предотвратить слишком сильное изменение распределения характеристик домохозяйств. Кроме того, перестанавливаются только географические переменные. Таким образом, зависимости внутри домохозяйств не будут затронуты слишком сильно.

27. Вообще говоря, ЦПД можно описать следующим образом (упомянутые уровни географии предполагаются вложенными):

a) на каждом уровне географии определяются домохозяйства с риском раскрытия информации выше определенного порога;

b) затем берется самый крупный из имеющихся географический уровень;

c) определяются «похожие» домохозяйства (т. е. домохозяйства, которые имеют одинаковые значения по некоторым указанным характеристикам домохозяйства, но необязательно по другим характеристикам) в других регионах на том же географическом уровне, чтобы получить набор домохозяйств-доноров;

d) случайным образом выбирается одно из домохозяйств-доноров;

e) меняются местами географические переменные домохозяйства, подверженного риску, с данными выбранного домохозяйства-донора (т. е. производится перестановка одинаковых данных для всех членов домохозяйства/данных по домохозяйству);

f) производится переход на следующий подробный уровень географии и повторяются действия с третьего шага, пока не будет достигнут самый подробный уровень;

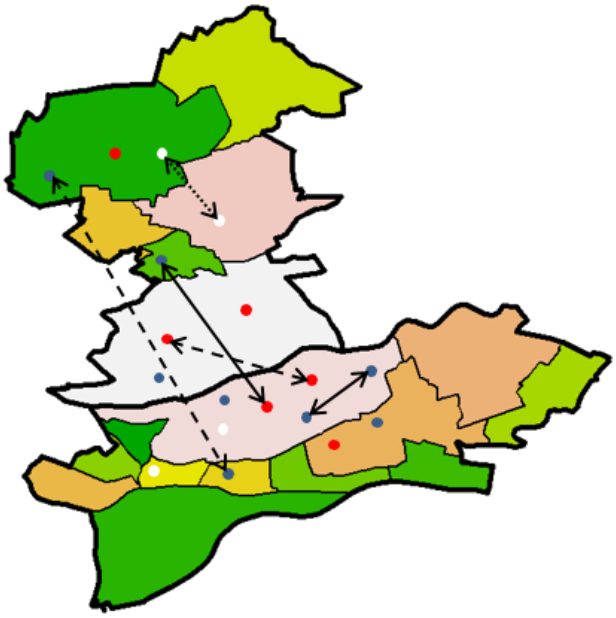
g) если доля переставляемых данных ниже заранее определенного порога после обработки наиболее подробного географического уровня, производится случайная перестановка дополнительных домохозяйств на самом подробном географическом уровне, до тех пор пока не будет достигнута необходимая доля переставленных данных.

28. Еще одно ограничение, которое налагается на этот итерационный процесс, состоит в том, что домохозяйство нельзя менять местами дважды. Кроме того, ясно, что этот процесс приводит к перестановке всех домохозяйств, находящихся в группе

риска. Это означает, что будет намного сложнее идентифицировать индивидуальную информацию после перестановки.

29. Применение ЦПД проиллюстрировано на диаграмме 1. На диаграмме показано два географических уровня, выделенных жирной линией (самый крупный уровень) и окрашенными областями (детальный уровень). Цветные точки обозначают домохозяйства, где похожие домохозяйства имеют одинаковый цвет. Стрелки со сплошной линией показывают недопустимые перестановки: замена непохожих домохозяйств или замена домохозяйствами в пределах одной и той же территории на одном и том же географическом уровне. Стрелки с пунктирными линиями показывают возможные перестановки при работе на самом грубом уровне, стрелки с точечными линиями показывают возможные перестановки при работе на самом детальном уровне.

Диаграмма I
Иллюстрация ЦПД



Примечание: Стрелки со сплошными линиями показывают недопустимые перестановки, стрелки с пунктирными линиями — перестановки на самом крупном географическом уровне, стрелки с точечными линиями — перестановки на самом подробном уровне.

30. После ЦПД можно рассчитать гиперкубы переписи. Это даст гиперкубы, в которых число ячеек может отличаться от «исходного» числа ячеек из-за перестановки домохозяйств и их соответствующих членов.

D. Добавление шума с использованием метода ключах ячеек

31. Добавление шума с использованием метода ключей ячейки (МКЯ) является посттабличным методом, который, соответственно, применяется к построенному набору таблиц переписи. Таким образом, МКЯ не изменяет базовых микроданных, а влияет только на таблицы переписи. МКЯ основан на методе, представленном Австралийским бюро статистики (АБС), см., например, Fraser and Wooton (2006). Их метод основан на так называемых «ключах ячеек», чтобы гарантировать, что случайный шум, добавленный к конкретной ячейке, всегда будет точно таким же, независимо от конкретного гиперкуба переписи, в котором он появляется. Мы несколько скорректировали метод, предложенный АБС, в том смысле, что мы несколько более гибки в приписывании ключей ячеек.

32. Чтобы обеспечить согласованность добавленного шума между различными гиперкубами, процесс приписывания ключей ячеек ячейкам должен быть согласован с

самого начала. С этой целью так называемые «ключей данных» приписываются записям в микроданных, лежащих в основе всех гиперкубов переписи. То есть каждому жителю присваивается случайное число. Всякий раз, когда строится ячейка гиперкуба, вычисляется число записей, попадающих в эту ячейку, и ключи данных этих данных создают ключ ячейки, который будет использоваться для выбора добавляемого шума. Таким образом, случайность ключей данных определяет случайность шума, тогда как детерминированный характер вычисления ключей ячеек обеспечит согласованность между различными гиперкубами.

33. Вообще говоря, МКЯ можно описать следующим образом:

a) присвоение равномерно [0,1) распределенных чисел каждой записи в микроданных переписи;

b) построение p-таблицу, которая определяет распределение шума;

c) при агрегировании микроданных в гиперкуб переписей для каждой ячейки дополнительное вычисление детерминированным способом ключей ячеек с использованием ключей данных для данных, попадающих в эту ячейку;

d) использование ключей ячеек вместе со значением ячеек для определения по p-таблице шума, который нужно добавить в ячейку;

e) добавление шум к значению ячейки.

34. Ключи ячеек вычисляются следующим образом: добавляются ключи записей ко всем данным, которые попадают в эту конкретную ячейку, после чего дробная часть результата берется в качестве ключа ячейки. Таким образом, ключи ячеек также являются равномерно распределенными [0,1) значениями. Затем эти значения можно использовать для извлечения из распределений в p-таблице. По сути, ключи ячеек берутся в качестве аргументов обратного распределения для получения реализации шума.

35. P-таблицы, предлагаемые для использования в таблицах подсчета частот (т. е. для гиперкубов переписи), имеют некоторые специфические параметры, которые можно задать:

a) дисперсия добавленного шума, обозначаемая как *V;*

b) максимальное значение добавленного шума, обозначаемое как D, с учетом распределения шума на $\{-D, -D+1, \dots, -1, 0, 1, \dots, D-1, D\}$;

c) минимальное положительное значение ячеек, разрешенное после добавления шума, обозначается как $j_s + 1$.

36. Кроме того, распределение в p-таблице должно быть таким, чтобы отрицательные значения ячеек были невозможны, а ожидание добавленного шума равнялось нулю для каждой ячейки. Как следствие, нулевые ячейки (ячейки с нулевыми значениями) не могут быть преобразованы в положительные значения. Более того, поскольку иногда известно, что нулевые ячейки не имеют доноров, изменение значений этих ячеек не будет способствовать защите таблиц.

37. Следует обратить внимание, что *js* может, например, использоваться для предотвращения появления значений 1 и 2 в таблицах подсчета частот, как того требуют законы некоторых стран. Тем не менее этот параметр по-прежнему допускает, что положительное значение ячеек может быть изменено на нулевое значение. Если для *js* выбрано значение 2, то набор возможных значений ячеек после применения МКЯ будет $\{0, 3, 4, \dots\}$.

E. Сочетание обоих методов

38. Несмотря на то, что методы ЦПД и МКЯ рекомендуется использовать для защиты гиперкубов переписей согласованным образом, разные государства — члены ЕС по-прежнему имеют некоторую свободу в отношении того, как использовать эти методы. Они могут не только выбрать разные значения параметров, но и решить

использовать только один из методов или комбинацию обоих методов. Действительно, учитывая разные правила конфиденциальности, применяемые в разных европейских странах, а также различия в размерах этих стран, было целесообразно рекомендовать не какой-то один метод. Однако, ограничив число рекомендуемых методов, Евростату, а также другим пользователям будет легче сопоставлять защищенные статистические данные переписей между странами.

39. Преимущество сочетания обоих методов заключается в том, что параметры, используемые для каждого метода, могут задаваться менее строго по сравнению с ситуацией, когда используется только один из методов. Более того, МКЯ специально нацелен на защиту от исчисления разностей, тогда как ЦПД вносит неопределенность в целом, но в основном на уровне данных.

VI. Оценка методов

40. Во время реализации второго ССГ (№ 2018.0108) предварительные версии инструментов для реализации методов были обнародованы, и государствам-членам было предложено опробовать их и предоставить отзывы. К сожалению, лишь ограниченное число государств-членов действительно представили отзывы. Их отзывы в основном касались вопросов установки и концептуальных вопросов, например способов подбора параметров. Дополнительные (исследовательские) работы по выбору адекватных значений параметров в настоящее время проводятся в нескольких европейских странах.

41. Нам известны несколько оценок предлагаемых методов. Здесь мы кратко опишем данные Статистического управления Нидерландов (СУН).

42. Поскольку СУН проводит свою перепись с использованием административных данных, оно смогло составить тестовый массив данных переписи, который был более недавним, чем перепись 2011 года. Массив данных, использованный для оценки обоих реализованных методов, был основан на данных о населении за 2017 год. Намерение СУН состояло в том, чтобы использовать комбинацию ЦПД и МКЯ. По их мнению, основная защита от исчисления разницы создается МКЯ. Однако для того, чтобы МКЯ обеспечивал достаточную защиту как таковой, параметры, вероятно, должны задаваться относительно строго. Это сильно ограничило бы его полезность.

43. ЦПД и МКЯ по-своему влияют на полезность и риск раскрытия информации. Сочетая оба метода, нагрузку на параметры можно разделить между обоими методами. Это позволяет не слишком жестко задавать параметры для каждого отдельного метода. Кроме того, влияние на полезность также может быть подразделено на оба метода. Исходя из этих соображений, СУН оценило только применение сочетание ЦПД и МКЯ.

44. Поскольку публикация гиперкубов на основе ячеек сетки была новой, СУН сосредоточило свою оценку на таблицах этого типа. Оно учитывало различия рисков не только между таблицами, но и между различными типами географических переменных. Например, оно рассмотрело разницу между ячейками сетки и регионами МАУ.

45. Было рассмотрено несколько вариантов комбинации ЦПД и МКЯ с разными значениями параметров. На момент составления этой записки несколько европейских стран решили использовать ЦПД, МКЯ или их комбинацию для защиты своих гиперкубов переписи 2021 года, но окончательных выводов из оценки СУН сделано не было. Однако исследования, проведенные СУН к настоящему времени, показали, что для таблиц переписей небольшой показатель перестановок для ЦПД (например, 1 %), когда, по крайней мере, все данные, подвергающиеся риску, меняются местами, будет в значительной степени способствовать защите в том смысле, что гораздо труднее судить о том, является ли найденное раскрытие реальным раскрытием. Более высокие показатели перестановок приведут к серьезным потерям информации. В случае МКЯ было установлено, что даже при не очень малых дисперсиях (например, 2 или 3) потери информации необязательно должны быть столь высокими. Эти уроки были

использованы для поиска подходящих значений параметров при использовании комбинации ЦПД и МКЯ.

V. Заключение

46. Согласование способов представления таблиц переписи уже давно стоит на повестке дня Евростата. Недавние изменения, связанные с контролем статистического раскрытия гиперкубов европейских переписей, кажутся многообещающими. Несколько стран намерены использовать ЦПД или МКЯ в качестве метода контроля раскрытия информации для гиперкубов переписи 2021 года. Оценки методов, проведенные государствами-членами, показывают, что ЦПД не следует использовать отдельно, поскольку остающийся риск раскрытия информации все еще слишком высок. МКЯ можно использовать отдельно, но представляется, что его совмещение с ЦПД, позволит существенно уменьшить снижение полезности из-за МКЯ.

47. Хотя мы уверены, что для переписи 2021 года многие государства — члены ЕС будут использовать гораздо более гармонизированный подход, все еще есть некоторые вопросы, которые заслуживают внимания при использовании пертурбативных методов.

48. Во-первых, связь с другими выпусками данных. Во многих странах будут опубликованы не только гиперкубы европейских переписей, но и стандартные выходные данные, такие как национальные таблицы переписи и другие демографические таблицы. Эти дополнительные результаты, очевидно, связаны с гиперкубами европейских переписей, если они имеют ту же отсчетную дату, что и в ряде стран. Всякий раз, когда национальные таблицы и другие демографические таблицы защищаются разными методами (т. е. не применяя ЦПД или МКЯ), разные версии могут привести к нежелательной ситуации, когда таблицы, которые *уже сами правильно защищены,* могут быть объединены для раскрытия индивидуальной информации. Прямая идея попытаться обойти эту проблему — использовать те же методы и в других выпусках. Точнее, при использовании ЦПД должен использоваться один и тот же возмущенный набор микроданных (как минимум должны присутствовать те же перестановки), а при использовании МКЯ — одни и те же ключи данных. Это возможно, если дополнительные выходные данные основаны исключительно на тех же микроданных. Однако для некоторых дополнительных выпусков данные переписи объединяются с данными, не относящимися к переписи. Это затруднило бы правильное применение исходных возмущений.

49. В настоящее время среди (аккредитованных) исследователей становится все более популярным анализ наборов микроданных, доступ к которым предоставлен НСУ. Для них было бы трудно защитить свою продукты с помощью ЦПД и МКЯ. Эти методы направлены на защиту таблиц значений частоты. Однако исследователи не обязательно получают такие результаты: они могут рассмотреть более сложную модель оценки или объединить данные переписи с другими данными для составления таблиц порядков величин.

50. Второй вопрос касается сообщения результатов после возмущения. Пользователям данных следует разъяснить, что опубликованные таблицы по-прежнему являются «действительными». Для обычного пользователя следует пояснить публикацию неаддитивных таблиц. Это можно сделать по аналогии с объяснением того, что таблицы с округленными цифрами иногда неаддитивны. Для более опытных пользователей опубликованных данных следует количественно оценить дополнительную неопределенность из-за ЦПД и МКЯ. Однако, как обсуждалось в Enderle et al. (2020), знание максимального возмущения в МКЯ может привести к повышенному риску раскрытия информации.

51. Следует не только объяснить полезность, общественности еще должно быть ясно, что возмущенные таблицы действительно защищают их конфиденциальность. Сами НСУ еще не знают, как выбрать параметры, чтобы оптимально сбалансировать риск и полезность. Это показывает, что обычному пользователю будет еще труднее

понять, что возмущенные таблицы по-прежнему полезны, и в то же время публикация этих таблиц не нарушает их конфиденциальности.

52. Третий вопрос касается выбора параметров метода. Как отмечалось в разделе IV, некоторые НСУ недавно попытались оценить этот вопрос. Евростат также внес свой вклад в это обсуждение (Bach 2021). На наш взгляд, точный выбор будет зависеть от конкретной ситуации НСУ: численность населения государств-членов, культурные различия и т. д. Чтобы содействовать будущим оценкам, мы хотели бы обратить внимание на две недавние публикации. Первая: Giessing et al., "How to select noise parameters for the Cell Key Method?" (2021). Она специально посвящена количественной оценке (оставшегося) риска раскрытия информации. Вторая публикация — Ricciato et al. (2021), в которой рассматривается инструмент с открытым исходным кодом для экспериментов со схемами возмущения на основе шума, специально предназначенный для количественной оценки полезности статистических продуктов.

53. Несмотря на вышеупомянутые проблемы, мы по-прежнему считаем, что предлагаемая гармонизация защиты гиперкубов Европейской переписи 2021 года является большим шагом вперед. Это лучшее, что можно сделать, когда нет возможности предписать обязательное использование тех или иных методов защиты.

Литература

Bach, F., 2021, "Differential Privacy and Noisy Confidentiality Concepts for European Population Statistics", in *Journal of Survey Statistics and Methodology*, 2021 (smab044, https://doi.org/10.1093/jssam/smab044).

Enderle, T., Giessing, S. and Tent, R., 2020, "Calculation of risk probabilities for the cell key method," in J. Domingo-Ferrer and Muralidhar, K. (Eds.), *Privacy in Statistical Databases*, pp. 151−165. New York: Springer-Verlag, LNCS, volume 12276.

European Commission, 2008, Regulation (EC) No 763/2008 of the European Parliament and of the Council of 9 July 2008 on population and housing censuses. *Official Journal of the European Union*, L218, pp. 14−20.

European Commission, 2017a, Commission Implementing Regulation (EU) 2017/543 of 22 March 2017 laying down rules for the application of Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses as regards the technical specifications of the topics and of their breakdowns. *Official Journal of the European Union*, L78, pp. 13−58.

European Commission, 2017b, Commission Regulation (EU) 2017/712 of 20 April 2017 establishing the reference year and the programme of the statistical data and metadata for population and housing censuses provided for by Regulation (EC) No 763/2008 of the European Parliament and of the Council. *Official Journal of the European Union*, L105, pp. 1−11.

European Commission, 2017c, Commission Implementing Regulation (EU) 2017/881 of 23 May 2017 implementing Regulation (EC) No 763/2008 of the European Parliament and of the Council on population and housing censuses, as regards the modalities and structure of the quality reports and the technical format for data transmission, and amending Regulation (EU) No 1151/2010. *Official Journal of the European Union*, L135, pp. 6−14.

European Commission, 2018, Commission Implementing Regulation (EU) 2018/1799 of 21 November 2018 on the establishment of a temporary direct statistical action for the dissemination of selected topics of the 2021 population and housing census geocoded to a 1 km^2 grid. *Official Journal of the European Union*, L296, pp. 19−27.

Fraser, B. and J. Wooton, 2006, "A proposed method for confidentialising tabular output to protect against differencing," in *Monographs of Official Statistics*. Work session on Statistical Data Confidentiality, Eurostat Office for Official Publications of the European Communities, Luxembourg, pp. 299−302.

Giessing, S., Enderle, T. and Tent, R., 2021, *How to Select Noise Parameters for the Cell Key Method?*, presented at NTTS 2021, 9 – 11 March 2021, extended abstract at https://coms.events/NTTS2021/data/x_abstracts/x_abstract_10.docx.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., de Wolf, P.P., 2012, *Statistical Disclosure Control*. Wiley series in Survey Methodology, John Wiley & Sons, Ltd, ISBN: 978-1-119-97815-2.

Ricciato, F., Stocchi, M., Bach, F., Bujnowska, A. and Kloek, W., 2021, *An open-source tool for experimenting with noise-based perturbation schemes*, presented at NTTS 2021, 9 – 11 March 2021, extended abstract at https://coms.events/NTTS2021/data/x_abstracts/x_abstract_105.pdf.