Организация Объединенных Наций ECE/CES/GE.41/2019/9

# Экономический и Социальный Совет

Distr.: General
3 July 2019
Russian
Original: French

## Европейская экономическая комиссия

Конференция европейских статистиков

**Группа экспертов по переписям населения и жилищного фонда**

**Двадцать первое совещание**
Женева, 18–20 сентября 2019 года
Пункт 2 предварительной повестки дня
**Результаты тестирования с точки зрения методологии, технологии, участия и других аспектов**

# Использование имен для усовершенствования измерения однополых пар при проведении переписи

**Записка Национального института статистики и экономических исследований (НИСЭИ), Франция***

*Резюме*

На основе результатов французской переписи трудно составить надежные статистические данные о количестве однополых пар во Франции. В самом деле, значительное число однополых пар (по данным исследования 2011 года – более 40%) учитывается в таком качестве из-за ошибки в кодировке пола одного из партнеров, что приводит к завышению результатов. Предлагаемая процедура корректировки заключается в расчете пропорции, в которой то или иное имя является скорее мужским или скорее женским, и в использовании этой информации с целью исправления переменных половой принадлежности для лиц, которые, по данным переписи населения, проживают в однополой паре. Этот метод представляется эффективным и дает результаты, которые согласуются с данными из других источников.

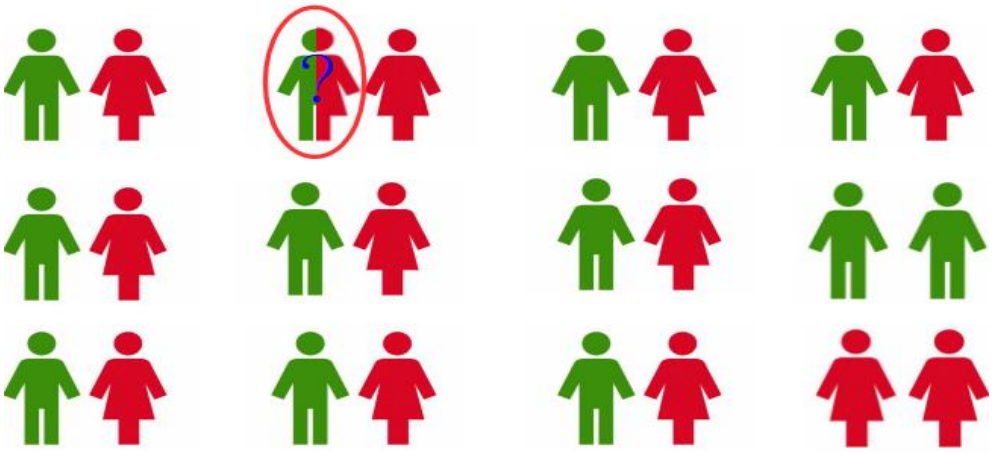
* Документ подготовили Элизабет Альгава и Себастьян Аллепе.

GE.19-10957 (R) 150719 160719
*1910957*

# I. Введение

1. В настоящее время из-за методологических затруднений невозможно на основе результатов переписи получить надежную статистику количества однополых пар (ОПП), проживающих во Франции. Действительно, в разнополой паре ошибка в указании или кодировке половой принадлежности одного из партнеров обычно приводит к учету такой пары в качестве однополой. Хотя подобные ошибки затрагивают лишь небольшую долю лиц, живущих в разнополых парах, этого достаточно для весьма значительного завышения доли лиц, образующих ОПП. Опасность такой ошибки актуальна не только для однополых пар; она появляется всякий раз, когда речь идет об оценке редких групп населения, охватывающей немногочисленную обследуемую выборку. Например, при измерении числа вдовцов моложе 30 лет или состоящих в браке лиц 18-летнего возраста нужно учитывать ошибки, связанные с возрастом или семейным положением, причем частотность таких ошибок может быть выше частотности ранних браков или раннего вдовства. Но особенность данной ситуации состоит в том, что индивидуальная ошибка в половой принадлежности для одного из партнеров приведет к тому, что оба партнера будут считаться имеющими однополого партнера (рис. I), т. е. повлечет за собой двойную ошибку.

Рис. I
**Воздействие ошибки в кодировке на переменную половой принадлежности**




*Примечание:* В эту фиктивную популяцию включены 12 пар и 24 лица – 9 разнополых пар (РПП), 2 однополые пары (ОПП) при наличии сомнения относительно кода половой принадлежности одного из партнеров в последней паре.

Если говорить об ошибке в кодировке, то 1 ошибка на 24 лица приведет к перемещению каждой 12-й пары из РПП в ОПП. Следовательно, такая ошибка является «двойной». Таким образом, количество ОПП искусственно возросло бы на 50%, а количество РПП искусственно уменьшилось бы только на 10%. Можно вновь констатировать факт, что незначительные ошибки в коде оказывают намного более заметное воздействие на редкие группы населения.

2. Эту трудность удалось преодолеть в 2011 году, когда НИСЭИ провел обследование семей и жилищ (ОСЖ). Для анализа выборки учитываемых единиц жилья одновременно с вопросниками переписи был предложен дополнительный четырехстраничный вопросник. Счетчики собрали около 360 000 вопросников; это ОСЖ дало возможность провести новые исследования по различным темам (Bodier et al., 2015; Imbert et al., 2018), в том числе по однополым парам. Была проделана большая работа по проверке ответов путем сопоставления информации, представленной в материалах переписи и ОСЖ (Breuil-Genier et al., 2016). На основе ее результатов было оценено число лиц в однополых парах, составившее 205 000 человек, из которых 173 000 проживают совместно. Из совместно проживающих пар 0,57 % составили однополые пары, а 0,36% – «ложные» однополые пары (Buisson et Lapinte, 2013; Banens et Le Penven, 2016). Таким образом, масштабы исправления данных по ОПП довольно значительны: от 295 000 соответствующих лиц

перед реорганизацией до 173 000 после ее проведения. Следовательно, было исправлено более 40% ситуаций. Этот этап исправления данных позволил провести новые аналитические действия в отношении лиц, находившихся в однополых парах в 2011 году (Rault, 2016, 2017, 2018).

3. Проведение очередного обследования семей, которое в принципе позволит получить новые, контролируемые статистические данные о числе однополых пар, предусмотрено не ранее 2023 года. Важно получить такие данные в более ранний срок, учитывая недавние изменения в законодательстве (в частности, майский закон 2013 года, разрешающий однополые браки) и обязательство Франции предоставить данные об однополых парах для европейской переписи 2021 года[1]. Более точное выявление таких пар в ходе переписи позволит Европейскому статистическому институту добиться качественных результатов. Благодаря проведению переписи и, следовательно, ежегодному сбору информации о миллионах лиц и жилищ (Godinot, 2016) такое усовершенствование также даст возможность вновь проанализировать эту группу населения, ее демографические, семейные и социально-профессиональные характеристики. Конечно, при проведении переписи анализ ограничивается совместно проживающими парами, тогда как супружеские отношения между однополыми партнерами существуют «на расстоянии» чаще, чем среди разнополых пар (Toulemon et al., 2005; Rault, 2018). Тем не менее эта тема весьма интересна, поскольку совместное проживание пар – это форма союза, составляющая абсолютное большинство даже среди ОПП (в 2011 году – 84%, Buisson et al., 2013).

4. Таким образом, эта тема оправдывает применение решения, позволяющего отличать внутри предположительно однополых пар тех лиц, которые действительно являются однополыми, от тех, которые учитываются в этом качестве из-за ошибки в кодировке половой принадлежности. Для этого предусматривается добавление в цепи обработки данных переписи новой индивидуальной производной переменной, которая показывает, в какой пропорции указанное имя является скорее мужским или скорее женским. Затем эта переменная будет использована для исправления переменной пола для лиц, которые по данным переписи живут в однополых парах. Применение предлагаемой процедуры в цепи обработки предусматривается не ранее проведения ежегодного переписного обследования 2020 года. При этом все же запланировано ее экспериментальное внедрение вне системы стандартной обработки данных в ходе ежегодных переписных обследований 2017–2019 годов.

5. Представив экспериментальные разработки и решения, применяемые за границей, и проведение французской переписи, мы переходим к предлагаемой процедуре и ее первому практическому применению.

# II. Решения, применяемые за границей

6. Поскольку статистическое измерение однополых пар вызывает затруднения не только во Франции, целый ряд различных решений был разработан и в других странах. Так, в публикации Banens et Penven (2016) представлены оценочные показатели доли «ложных» однополых пар в общей численности пар, составленные по итогам американских, канадских и британских переписей; эти показатели варьируются от 0,25% до 0,57%. Доля таких «ложных пар» среди общего числа пар, представляющихся в качестве однополых, составляет 27–55%. Таким образом, показатели весьма близки к тем, которые были рассчитаны для Франции, причем с аналогичными трудностями: незначительные ошибки в общей численности, которые приводят к существенным искажениям реальных оценок ОПП. Страны, столкнувшиеся с этой трудностью, в экспериментальном порядке применяли различные стратегии для ее преодоления.

[1] Исполнительный регламент № 1201/2009 Европейской комиссии.

## A. Избыточность и сопоставление информации

7. Первую группу составляют решения, которые заключаются во внесении изменений в вопросник или в протокол обследования или переписи для получения дополнительной подтверждающей информации. Общий исходный принцип состоит в том, что ошибки являются редкостью и что вероятность дублирования данных весьма незначительна.

8. При проведении канадской переписи с 2001 года и американской переписи начиная с 2020 года среди всех категорий, позволяющих квалифицировать отношения двух лиц, брачные отношения описываются в четырех вариантах: супруг/супруга противоположного пола, гражданский партнер противоположного пола, супруг/супруга одного пола, гражданский партнер одного пола. Поэтому есть возможность сопоставить эту информацию с полом обоих партнеров. Если оба однополых партнера отметили вариант «супруг противоположного пола», то, вероятно, была допущена ошибка в кодировке пола одного из этих двух партнеров; если же, напротив, они выбрали вариант «супруг одного пола», то вероятность двух ошибок крайне невелика и речь идет о «подлинной» ОПП. Кроме того, для респондентов, которые будут участвовать в американской переписи 2020 года через Интернет, предусмотрено контрольное окно на случай несоответствия между выбором варианта и указанными полами (например, если женщина указывает себя как супруга противоположного пола другой женщины). Предварительные испытания показывают, что сокращение несоответствий достигает весьма значительного уровня благодаря новому вопросу и автоматическим механизмам проверки при направлении в случае заполнения вопросников в Интернете (Kreider, 2017).

9. Эта методика сопоставления данных, собранных разными способами, весьма схожа с той, которая применялась при обследовании семей и жилищ 2011 года.

## B. Сопоставление с административными данными

10. Эксперимент по подтверждению указанного и кодированного пола путем его сопоставления с административными данными был проведен в Соединенных Штатах (Kreider, 2015). Сопоставляя данные переписи с регистром социального обеспечения, авторы выявляют как минимум для одного из партнеров более частые несоответствия между полом, кодированным при переписи, и полом, указанным в регистре социального обеспечения (Numident), когда пары в ходе переписи предположительно заявлены как однополые: 72,7% пар, состоящих в браке, и 6,4% пар, которые не состоят в браке.

## C. «Статистическое обоснование» с помощью имени

11. При обработке данных переписи 2010 года Бюро переписей США использовало указатель имен (O'Connell 2011), составленный на основе ответов респондентов самой переписи и отражающий долю мужчин с соответствующим именем («маскулинность») по шкале от 0 до 1 000. Для внесения исправлений авторы установили порог в 50 случаев на 1 000, или 5%; по их мнению, такая пороговая величина является «консервативной». Иначе говоря, если согласно указателю имя соответствующего респондента, которое, например, кодировано как мужское, носят лишь 5% мужчин или менее, пол респондента подлежал исправлению; в противном же случае указанный респондентом пол сохранялся. Применение этого порога привело к тому, что в 50% пар респондентов как минимум для одного из партнеров несоответствие между именем и полом было исправлено на «однополый», причем чаще это касалось пар, состоящих в браке (69%), чем живущих в свободном союзе (21%).

12. Впоследствии полученные результаты были сопоставлены с данными регистров социального обеспечения, с тем чтобы проверить, действительно ли внесенные исправления соответствуют ошибкам в кодировке пола (Kreider, 2015). Проверка показала, что в 85% случаев имена лиц были признаны недвусмысленными, и поэтому их ситуация могла подлежать исправлению. В 96% случаев пол, определенный на основе имени, был идентичен полу, указанному в регистре социального обеспечения.

13. Таким образом, метод статистического обоснования на базе имени представляется достаточно надежным, как минимум в случае его применения в ходе американской переписи 2010 года. Такие результаты служат стимулом для проверки возможностей применения данного метода при проведении переписи во Франции.

# III. Французская перепись и недавние изменения в ее проведении

14. С 2004 года французская перепись населения представляет собой исследование на основе выборки. Сбор данных проводится ежегодно путем заполнения вопросника в бумажном или электронном формате. При опубликовании результатов переписи законно проживающего населения в разбивке по коммунам используются данные пяти ежегодных сборов данных. Например, результаты переписи населения 2014 года основаны на использовании данных, которые собирались ежегодно в период 2012–2016 годов. Вместе с тем ежегодное переписное обследование (ЕПО) 2017 года соответствует данным ежегодного сбора, проведенного в 2017 году, т. е. содержит информацию о лицах и жилищах, которые были учтены именно в этом году. Впоследствии мы используем результаты ежегодных обследований, взятые в индивидуальном порядке.

15. По каждой единице жилья, включенной в выборку, счетчик должен собрать заполненный жилищный формуляр, в котором, в частности, описаны родственные связи между проживающими в нем лицами, и личные опросные листы по каждому из жильцов, где указываются их социально-демографические характеристики. В 2015 году форма личного опросного листа была изменена: вопрос о законном семейном положении лица (в браке, разведен, вдовец, холост) был заменен вопросом о его фактическом положении (в браке, в нотариально оформленном гражданском браке – ПАКС, в свободном союзе, разведен, вдовец, холост). Это повышает качество ответов, поскольку респондентам удобнее находить нужные им ответы, в частности тем, которые заключили ПАКС: ранее они нередко указывали себя состоящими в браке, полагая, что этот вариант наиболее близок к их реальной ситуации (Buisson, 2017).

16. Изменения в жилищный формуляр были внесены с момента сбора данных в 2018 году. Его новый формат позволит эффективнее выявлять характер брачных отношений между двумя лицами на основании заявленной ими информации, а не, как ранее, исходя из того факта, что каждый из них указал, что проживает в паре (не уточнив при этом, с кем именно). Такое изменение формата косвенно создает условия для совершенствования измерения ОПП. Действительно, для его практического применения необходимо проводить систематическое сопоставление между, с одной стороны, лицами, которые фигурируют в списке жильцов данной единицы жилья, с указанием их попарных связей (жилищный формуляр), а с другой стороны – между личными вопросниками каждого из этих жильцов. Такое сопоставление проводится по критериям пола и года рождения, а если этих двух переменных недостаточно для сопоставления, то по критериям фамилии и имени. Таким образом, сочетания имен и фамилий будут использоваться при обработке (а затем будут изъяты из распространяемых записей); начиная со сбора данных 2016 года в порядке опережения предусмотрен новый порядок получения информации, позволяющий применение процедуры, которая предлагается далее в настоящем документе.

17. Одним из основных недавних изменений в проведении переписи является сбор данных через Интернет. Такая методика, которой в 2013 году практически не существовало, в 2017 году уже охватывала более половины респондентов. Для них ошибки в кодировке ответов, связанные с оптическим распознаванием и внесением исправлений в бумажные вопросники в ручном режиме, становятся неактуальными. Особое значение в этом случае имеет разница в качестве между электронным и ручным методами сбора данных, связанная с использованием имен респондентов, тем более что для ограничения затрат пропорционально использованию этих имен критерии качества, касающиеся ввода имен, собранных в бумажных вопросниках, являются невысокими. Это различие оказалось весьма значимым и потребовало корректировки предлагаемого метода обработки.

18. Если абстрагироваться от этого методологического аспекта, то сбор данных через Интернет может показаться обеспечивающим больше гарантий с точки зрения конфиденциальности и способствующим большей искренности заявлений со стороны респондентов, в частности внутри однополых пар. Впрочем, по мнению Бюро переписей США, повсеместное распространение электронного метода сбора данных для проведения американской переписи 2020 года является одним из путей совершенствования измерения однополых пар (Kreider, 2017).

# IV. Поиск оптимального решения

19. На американском примере видно, что использование имени – это довольно эффективный способ для выявления ошибок в кодировке пола и, следовательно, для выделения ложных ОПП из числа предполагаемых ОПП. Опираясь на тот факт, что имя респондента будет вводиться для контроля качества сбора данных в ходе переписи с 2016 года, мы проверяем эффективность такой методики «статистического обоснования» с помощью имени.

## A. Промежуточный показатель: пары, предположительно являющиеся однополыми

20. Чтобы оценить численность предположительно однополых пар и ее динамику и иметь возможность проверить, какова доля тех из них, которые действительно являются однополыми, и тех, которые отнесены к этой категории из-за ошибки в кодировке пола, мы используем упрощенный показатель: если именно два человека указывают, что проживают вместе в данном жилище и что они относятся к одному полу, то их классифицируют как предположительно однополую пару. Этот показатель не совершенен, поскольку два человека, которые проживают совместно и оба указывают, что они «живут в паре», поскольку у каждого из них есть партнер, проживающий в другом жилище, будут ошибочно учтены как составляющие отдельную пару. Но сопоставление такого упрощенного показателя с результатами измерения, которые подтверждены данными обследования семей и жилищ (ОСЖ) 2011 года, свидетельствует о том, что этот показатель позволяет проводить правильную оценку предполагаемых ОПП. Как и ожидалось, он охватывает слишком много пар, значительную часть которых составляют гетеросексуальные пары, засчитываемые как ОПП из-за ошибки в кодировке, но при этом пропускается мало «подлинных» ОПП. Благодаря переработке жилищного формуляра в 2018 году можно будет опираться на более точное измерение.

## B. Постоянная демографическая выборка (ПДВ): оптимальный инструмент для проверки эффективности процедуры обнаружения явных ошибок в кодировке пола

21. ПДВ – это масштабная социодемографическая платформа, созданная во Франции для изучения демографических, семейных, профессиональных и географических характеристик (Durier, 2018). Общий ее принцип заключается в том, что по входящим в выборку лицам (около 4% населения) сохраняется информация, полученная из пяти статистических источников, которые «подпитывают» ПДВ (книги записи актов гражданского состояния, переписи, избирательные списки, платформа «все лица наемного труда» с указанием полученных вознаграждений, а с 2011 года – налоговые и социальные данные о доходах).

22. Первоначальная задача использования ПДВ в нашем подходе состоит в том, чтобы обеспечить сопоставление «подлинного пола», который зарегистрирован в национальном регистре физических лиц (НРФЛ) – а данные регистра считаются точными, так как используются в сфере управления номерами социального обеспечения, – с тем полом, который указывается в ежегодных переписных обследованиях. Это позволяет оценить долю ошибок в кодировке пола при проведении

переписи. В исследовательской базе 2016 года из 1,3 млн человек «из ПДВ» (родившихся в период, охватываемый рамками ПДВ), учтенных хотя бы один раз в ходе переписи за период 2010–2016 годов и живущих в паре, частота ошибок в половой принадлежности составляет 0,17%. Следовательно, речь идет о весьма редком явлении. При этом частота ошибок среди лиц, предположительно относящихся к ОПП, намного выше (18%).

23. В рамках ПДВ обращение к ресурсам НРФЛ производится только в связи с каким-либо «лицом из ПДВ», чтобы включить его в выборку и дополнить статистические данные, касающиеся этого лица. Статистическая информация о жильцах, проживающих в его жилище, также включается в выборку, но без идентификации, предполагающей обращение к НРФЛ. Поэтому удостоверить пол других жильцов соответствующего жилища, включая супругов, не представляется возможным. Поскольку ошибка в кодировке пола затрагивает 0,17% лиц из ПДВ, живущих в паре, то, предполагая независимость ошибок по каждому из партнеров, можно считать, что ошибка в кодировке пола в отношении одного из партнеров затронет 0,31% пар, вследствие чего они будут ошибочно учтены как ОПП, а ошибка в кодировке по обоим партнерам коснется 0,03% пар (таблица 1).

Таблица 1
**Прогнозируемые последствия ошибок в кодировке пола на уровне пар**

| *Реальная ситуация* | | | *Предполагаемая ситуация* | | |
|---|---|---|---|---|---|
| *H1 : 0,6% «подлинных» ОПП среди всех пар* | | | *H2 : код половой принадлежности ошибочен для 0,17% опрошенных лиц* | | |
| | | | *Пары с одной ошибкой (0,31%)* | *Пары с двумя ошибками (0,03%)* | *Пары без ошибок (99,66%)* |
| Из 1 000 000 пар | 6 000 ОПП | ➔ | 19 предполагаемых РПП | 2 ОПП | 5 980 ОПП |
| | 994 000 РПП | | 3 078 предполагаемых ОПП | 287 РПП | 990 634 РПП |

*Источник:* Обследование семей и жилищ 2011 года, исследовательская база ПДВ 2016 года, НИСЭИ.
*Выборка:* Совместно проживающие пары.
*Пояснение:* Согласно обследованию семей и жилищ, 0,6% пар являются «подлинными» ОПП, а 99,4% – «подлинными» РПП (H1). Если одна ошибка в коде затрагивает 0,17% опрошенных лиц (H2), то тогда две ошибки в коде касаются 0,03% пар (0,17 x 0,17), а одна ошибка в отношении одного из партнеров затрагивает 0,31% пар (0,17 + 0,17 – 0,03). Таким образом, 0,31% РПП переходят в категорию предполагаемых ОПП (т. е. 3 078 пар на миллион). В свою очередь 0,31% ОПП переходят в категорию предполагаемых РПП (т. е. 19 пар). Наконец, в 0,03% пар ошибки в кодировке половой принадлежности затрагивают обоих партнеров (т. е. 2 ОПП и 287 РПП). Даже если ни у одного из партнеров в паре зачитываемый пол не соответствует реальному, предполагаемая ситуация их пары совпадает с реальной: эта пара остается в категории РПП. Поэтому при доле ошибок, составляющей 0,17%, можно ожидать, что ОПП переходят из реальной ситуации, где они составляют 0,6% пар, в предполагаемую ситуацию в рамках ЕПО, где на их долю приходится 0,9% пар (5 980 + 3 078 + 2 = 9 060 пар на миллион).

## C. Выбор словаря имен

24. Второй аргумент в пользу применения ПДВ состоит в том, что имя, указанное в ходе переписи, вводится в производственную базу для содействия сопоставлению (в то же время оно отсутствует в исследовательской базе, используемой для подготовки статистических данных, чтобы не допускать прямой идентификации лиц). Поэтому есть возможность составить словарь и указатель ошибок таким же образом, как и в системе подготовки будущих ежегодных переписных обследований.

25. В словаре имен для каждого имени указывается процент женщин (в соответствующих случаях – мужчин), носящих это имя. Затем оно сопоставляется с именами респондентов, чтобы сравнить кодированный пол каждого респондента с тем

полом, который чаще всего соответствует данному имени. Конечная цель – выявить самые вероятные ошибки в кодировке. Использование ПДВ дало нам возможность сравнить эффективность разных словарей и выбрать из них тот, который наиболее результативен в обнаружении ошибок кодировки пола лиц, относящихся к ПДВ.

26. Была определена подборка различных словарей с использованием двух источников. Основным источником, касающимся исключительно лиц, родившихся во Франции, является регистр имен, которые давались при регистрации актов гражданского состояния с 1900 года, в разбивке по полу[2]. Чтобы рассчитать долю мужчин и женщин, носящих каждое из имен, необходимо провести огромное число наблюдений. Этот регистр был дополнен случаями присвоения имен лицам, которые были опрошены в 2017 году[3] и родились за границей, поскольку они не охвачены системой регистрации актов гражданского состояния. Зачастую эти лица носят какое-либо имя, которое не значится в словаре, составленном на основе данных этой системы. Выбранный словарь представляет собой еще и комбинацию данных в том смысле, что сначала нужно искать совпадение в максимально подробном словаре (то же имя, тот же год рождения). При отсутствии совпадения используется менее подробный словарь: та же первая часть имени без учета года рождения. Таким образом, для того или иного лица может быть в конечном счете определена пропорция, отражающая долю женщин среди лиц того же года рождения и имеющих то же имя, или долю женщин среди всех лиц, имеющих имя, первая часть которого идентична. Для упрощения задачи будем считать, что этот показатель отражает долю женщин с одним и тем же именем. Затем из этой доли вычитается показатель ошибки: если кодировка в ходе переписи соответствует мужскому полу, то речь идет о доле женщин, имеющих одно и то же имя согласно используемому словарю. Если кодировка соответствует женскому полу, то тогда речь идет о доле мужчин. По договоренности при отсутствии соответствия, а значит, и доли этот показатель считается равным нулю, и никакое исправление не может быть внесено. Чем выше значение показателя, тем выше вероятность ошибки в кодировке.

## D. Метод выявления ошибок в кодировке пола с помощью имен весьма эффективен в отношении лиц из ПДВ

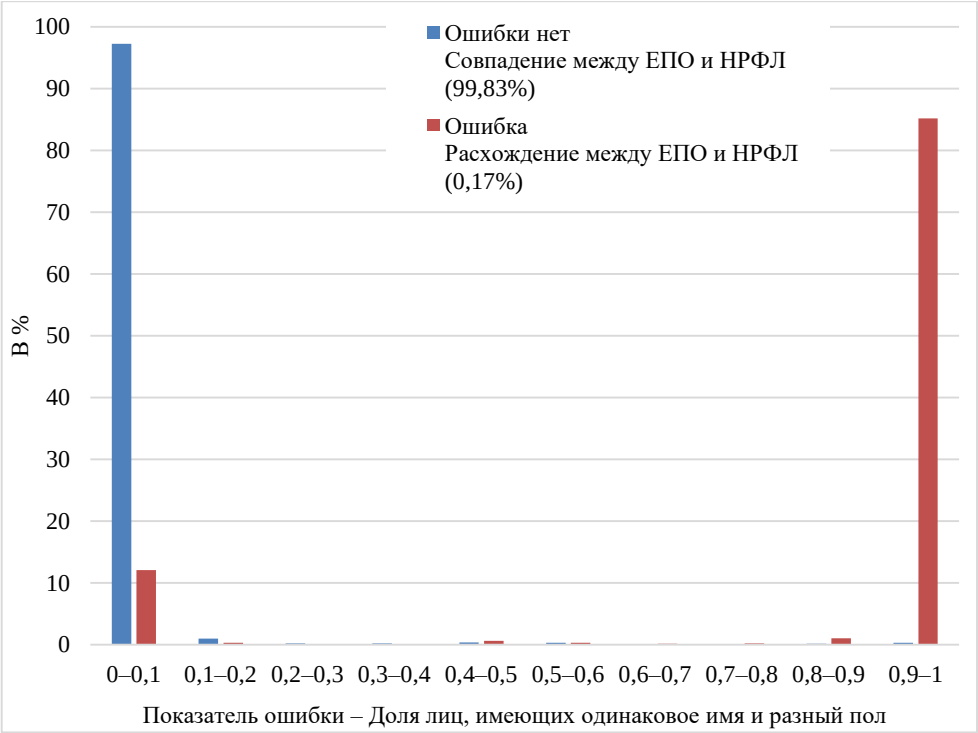
27. Этот показатель очень эффективен в части обнаружения ошибок в кодировке. При отсутствии подтвержденной ошибки почти все значения являются нулевыми или очень близкими к нулю (график 1). В случае подтвержденной ошибки (кодированный пол в рамках ЕПО не совпадает с полом, зарегистрированным в НРФЛ) почти все значения, напротив, близки к 1. Промежуточные же значения редки: немногие люди имеют имена обоюдного рода, такие как Доминик.

[2] Имена, которые являются слишком редкими, были изъяты.

[3] Для организованного проведения переписи эти имена, которые имелись в наличии для респондентов, отвечающих через Интернет, были введены для респондентов, заполняющих бумажный вопросник. Затем, как и было предусмотрено, эти имена были удалены сразу по завершении обработки данных в течение 2017 года.

График I

**Распределение показателя ошибки в зависимости от наличия подтвержденной ошибки**




*Источник:* Исследовательская база ПДВ 2016 года, НИСЭИ, взвешенные данные по всем лицам, живущим в паре.

*Пояснение:* Среди лиц старше 15 лет, опрошенных в период 2010–2016 годов, которые в ходе переписи указали, что проживают в паре и пол которых, указанный при опросе, был подтвержден данными НРФЛ, почти все носят имена, соответствующие указанному ими полу: значения показателя ошибки, построенного с этими именами, сосредоточены вблизи 0. Напротив, для подавляющего большинства лиц, чей пол, указанный в ходе переписи, не соответствовал данным НРФЛ, показатель на базе имени свидетельствует о высокой вероятности ошибки в половой принадлежности, указанной при переписи, с величинами в районе 1. Значения около 0 соответствуют в основном тем лицам, чьи имена не удалось сопоставить со словарем имен (например, когда у респондента слишком редкое имя).

28. Для выбора оптимального порога, с которого принимается решение о внесении исправления, используются те же инструменты, что и при разработке диагностических тестов в эпидемиологии: меры специфичности (эффективно выявлять ошибки) и меры чувствительности (не вносить ошибочных исправлений в правильные ответы), определение оптимального порогового значения и сопоставление тестов с помощью кривой ОРМ (Robin, 2011). Это приводит к установлению порога на уровне 41%: если код лица соответствует мужскому полу, но более 41% лиц с таким же именем являются женщинами, в данные вносится исправление. Может показаться странным, что исправление вносится даже тогда, когда большинство лиц с данным именем составляют мужчины, однако в реальности наблюдается крайне мало случаев, когда значение по словарю является промежуточным: почти все значения близки к нулю или, напротив, к 100%. В столь благоприятных условиях последствия установления пороговой величины носят ограниченный характер. Благодаря выбору словаря и порогу на уровне 41% такой метод позволяет достигнуть 98% чувствительности в 95% специфичности. Если говорить только о предполагаемых ОПП, то число ошибочных исправлений в ПДВ составит около 150 раз примерно на 2 000 выявленных реальных ошибок.

Таблица 2
**Эффективность словаря в выявлении ошибок кодировки пола**

| | | *Лица в предполагаемых ОПП* | | |
|---|---|---|---|---|
| | *Все лица, живущие в паре* | *Всего* | *Бумажные вопросники* | *Электронные вопросники* |
| Размер выборки | 1 343 908 | 11 349 | 9 605 | 1 744 |
| Явные ошибки | 2 336 | 2 059 | 1 852 | 207 |
| Оптимальный порог (в %) | 41 | 41 | 41 | 51 |
| Специфичность (в %) | 99 | 98 | 98 | 99 |
| Чувствительность (в %) | 89 | 95 | 95 | 95 |
| Ошибочные исправления | 16 950 | 147 | 134 | 9 |
| Правомерные исправления | 2 088 | 1 951 | 1 754 | 197 |

*Источник:* Исследовательская база ПДВ 2016 года, НИСЭИ, невзвешенные данные.
*Выборка:* Лица старше 15 лет, живущие в паре и опрошенные в период 2010–2016 годов.

# V. Неожиданные результаты транспонирования процедуры ПДВ в систему ЕПО и необходимость адаптации этой процедуры к методу сбора данных

## A. Прямое транспонирование в систему переписи приводит к чрезмерным исправлениям

29. Таким образом, использование ПДВ дало возможность оценить процедуру внесения исправлений на выборке, содержащей известные ошибки в кодировке. Поскольку данные ПДВ взяты из результатов переписи, этот метод, казалось бы, дает право ожидать идеальной результативности в исправлении ошибок в случае его применения к переписи в целом. К сожалению, прямое транспонирование приносило неожиданные результаты: в 12% респондентов, опрошенных в 2017 году в рамках ЕПО с помощью бумажных вопросников (независимо от их положения с точки зрения проживания в паре), носили имя, почти всегда применимое к противоположному полу (таблица 3); это – чрезмерная доля по сравнению с 0,2% ошибок в кодировке пола, которые мы стремимся обнаружить. При этом выяснилось, что ошибки были связаны скорее с вводом имени, чем с половой принадлежностью: из-за объема затрат уровень качества, требуемый от поставщика услуг при вводе имен респондентов, заполняющих бумажные вопросники, является низким, за исключением тех вопросников, которые подлежат вводу в ПДВ.

Таблица 3
**Сопоставление словаря, используемого в ПДВ и при проведении ЕПО 2017 года в зависимости от способа сбора данных**

| | *Все пары* | | | | | *ОПП* | | |
|---|---|---|---|---|---|---|---|---|
| | | *ЕПО 2017 года – до учета способа сбора данных* | | *ЕПО 2017 года – после учета способа сбора данных* | | | *ЕПО 2017 года – после учета способа сбора данных* | |
| *Носители имени…* | *ПДВ* | *бумажный* | *Интернет* | *бумажный* | *Интернет* | *ПДВ* | *бумажный* | *Интернет* |
| …почти всегда относятся к тому же полу, что и респондент (ПО <5%) | 93 | 72 | 95 | 41 | 97 | 79 | 32 | 84 |
| …в основном относятся к тому же полу, что и респондент (ПО = 5–40%) | 3 | 8 | 3 | 36 | 2 | 2 | 32 | 1 |
| …относятся к обоим полам (ПО = 40–90%) | 1 | 7 | 1 | 21 | 1 | 1 | 26 | 1 |
| …очень редко относятся к тому же полу, что и респондент (ПО = 90–95%) | 0 | 1 | 0 | 1 | 0 | 0 | 3 | 0 |
| …почти никогда не относятся к тому же полу, что и респондент (ПО >95%) | 0 | 12 | 2 | 1 | 0 | 18 | 8 | 13 |
| **Итого** | **100** | **100** | **100** | **100** | **100** | **100** | **100** | **100** |

*ПО:* показатель ошибки.

*Пояснение:* 32% респондентов в предполагаемых ОПП, заполнивших бумажный вопросник, являются носителями (по данным, взятым из их личного опросного листа) имени, которое почти всегда носят лица одного с ними пола: словарь указывает, что доля носителей такого имени среди лиц противоположного пола составляет менее 5%.

*Выборка:* Лица в предполагаемых ОПП по данным ЕПО 2017 года, без учета ФНОЖ и домохозяйств, где одному из партнеров пол был присвоен.

*Источники:* Исследовательская база ПДВ 2016 года, ЕПО 2017 года, НИСЭИ.

## B. Словари, различающиеся в зависимости от способа сбора данных

30. Для преодоления этой трудности было найдено решение, состоящее во включении способа сбора данных в структуру словарей. Благодаря этому доля ситуаций, когда пол более 95% респондентов не соответствует полу, указанному респондентом, становится крайне незначительной даже при сборе данных по бумажным вопросникам. Оборотной стороной медали при этом является высокая доля промежуточных значений, требующая особого учета при обработке результатов. Поскольку уровень качества в постоянной демографической выборке выше даже при сборе данных по бумажным вопросникам, проверить на ПДВ худший показатель качества при сборе данных на бумажных вопросниках переписи не представляется возможным.

31. С другой стороны, при сборе данных через Интернет полученные результаты намного ближе к ПДВ. Следовательно, качество обнаружения ошибок в кодировке пола в ПДВ можно считать репрезентативным для качества обработки в электронном аспекте ЕПО. Поэтому предлагается использовать адаптированный к Интернету словарь и установить пороговую величину в 51%, позволяющую максимизировать в части электронного сбора данных ПДВ уровни чувствительности (до 95%, т. е. в 5% случаев в верные данные вносятся ошибочные исправления) и специфичности (до 99%, т. е. выявляются 99% явных ошибок). Указанный пол лица, относящегося к предполагаемой ОПП, будет исправлен в случае, если доля лиц противоположного ему пола с тем же именем больше или равна 51%.

### C. При сборе данных с помощью бумажного вопросника соответствие пола и имени носит менее явный характер

32. При сборе данных из бумажных вопросников в ходе ЕПО словарь был составлен таким образом, чтобы избегать внесения неправомерных исправлений. Тем не менее показатель ошибки зачастую приобретает промежуточные значения: в отличие от электронного сбора данных, здесь критерий имени функционирует не столь успешно, а результат весьма чувствителен к установленному порогу. Например, если пороговый уровень составляет 95%, то данные 8% респондентов, предположительно относящихся к ОПП, которые были собраны с помощью бумажных вопросников, будут исправлены (при сборе через Интернет – 13%). При пороге в 40% доля исправлений данных в бумажных вопросниках возрастет до 37%, а при электронном опросе составит 14%, т. е. практически не изменится. При заполнении бумажных вопросников промежуточные величины показывают, что имя может быть внесено ошибочно (причем для одних имен такая вероятность выше, чем для других). Тогда для решения вопроса о том, относится ли данное лицо к «подлинной» ОПП, нужно опираться на другие переменные. Эти переменные выбираются из ПДВ, которая позволяет измерить долю реальных ошибок в половой принадлежности лиц, предположительно относящихся к ОПП и заполнявших бумажные вопросники.

33. Наконец, если доля лиц, пол которых противоположен указанному в вопроснике, составляет более 90% среди лиц с тем же именем, данные о половой принадлежности исправляются; если эта доля менее 20%, то исправления не вносятся. Если же эта доля находится в промежутке от 20% до 90%, имя нужно признать двусмысленным и использовать дополнительные переменные. По итогам ЕПО 2017 года было выделено три группы: лица, не состоящие в браке, для которых показатели частоты ошибок являются наиболее низкими (10%); лица, состоящие в браке, в возрасте менее 50 лет (23% ошибок); лица, состоящие в браке, в возрасте 50 лет или старше (32%). По итогам ЕПО 2018 года вследствие изменения жилищного формуляра группы составлены с учетом типа семьи (пара без детей, семья без детей, восстановленная или нет), а обработка данных улучшена благодаря учету положения пары (а не двух партнеров по отдельности).

## VI. Применение: сколько ОПП во Франции и какова динамика их численности

34. Возможность проводить оценку численности предполагаемых ОПП в рамках ЕПО имеется с 2005 года, а в рамках ПДВ – с 2010 года. Их количество постоянно растет. Перерыв в этой динамике, отмеченный в 2018 году, объясняется резким снижением числа ошибок в кодировке пола после изменения формы жилищного формуляра.

35. Что касается «подлинных» ОПП, то различные методы оценки их численности (процедура использования имен в ходе ЕПО в 2017 и 2018 годах, экстраполяция индивидуальных ошибок в парах в 2010–2018 годах в рамках ПДВ) представляются согласованными между собой и с данными ОСЖ. Очевидно, доля лиц в ОПП среди лиц, живущих в паре, с 2010 года показывала регулярный рост и увеличилась с 0,5% до 0,8%. При уточненной классификации семейного положения соответствующие доли в рамках ОСЖ 2011 года и ЕПО 2018 года, по оценочным расчетам, становятся немного выше, указывая на наличие некоторого роста – с 0,6% в 2011 году до 0,9% в 2018 году.

Таблица 4
**Доля предполагаемых ОПП и оценка «подлинных» ОПП в различных источниках за период 2005–2018 годов**

| | *Предполагаемые ОПП* | | | *Оценка «подлинных» ОПП* | | |
|---|---|---|---|---|---|---|
| *%* | *ЕПО* | *ПДВ* | *ОСЖ* | *ЕПО* | *ПДВ* | *ОСЖ* |
| 2005 | 0,6 | | | | | |
| 2006 | 0,7 | | | | | |
| 2007 | 0,7 | | | | | |
| 2008 | 0,7 | | | | | |
| 2009 | 0,8 | | | | | |
| 2010 | 0,8 | 0,8 | | | 0,5 | |
| 2011 | 0,8 | 0,8 | 0,9 | | 0,5 | 0,6 |
| 2012 | 0,9 | 0,9 | | | 0,6 | |
| 2013 | 0,9 | 0,9 | | | 0,6 | |
| 2014 | 1,0 | 1,0 | | | 0,6 | |
| 2015 | 1,0 | 1,0 | | | 0,7 | |
| 2016 | 1,1 | 1,1 | | | 0,7 | |
| 2017 | 1,2 | 1,2 | | 0,8 | 0,8 | |
| 2018 | 1,0/1,0* | 1,0 | | 0,8/0,9* | 0,8 | |

*Источник:* ЕПО 2005–2018 годов, исследовательская база ПДВ 2016 года и дополнительные данные 2017 и 2018 годов, НИСЭИ.

*Выборка:* Лица, живущие в паре. В части результатов ЕПО – без учета ФНОЖ и случаев, когда пол не указан одним из партнеров.

* На основе новых данных, полученных благодаря изменению жилищного формуляра: в изучаемую выборку входят лица, которые по данным нового анализа семейных домохозяйств проживают в паре. Формуляры необследованного жилья возвращаются в систему переписи наряду с отсутствующими ответами о половой принадлежности (такие случаи прекратились при заполнении бумажных вопросников и крайне редки при опросе через Интернет).

# VII. Заключение

36. Проведенный анализ показывает, что, несмотря на ряд ограничений, применение метода на основе имен дает в контексте Франции удовлетворительные результаты. Первое исследование 2018 года, цель которого – проанализировать характеристики лиц, живущих в однополой паре, проводимое с применением такого метода, позволит уточнить имеющиеся результаты. Последующая цель – внедрить этот метод в цепи обработки данных переписи, чтобы повысить качество переменных половой принадлежности, которые фигурируют в распространяемых записях.

# Ссылки


Algava E., Hallepee S., 2018, « Estimer les effectifs de couples de personnes de même sexe au recensement : expérimentation d'une solution de validation du sexe par le prénom », document de travail F1807, Insee

Banens M., Le Penven E., 2016, « Les erreurs de sexe dans le recensement et leurs effets sur l'estimation des couples de même sexe », *Population* 71(1), p. 135–148.

Bodier M. et al. (coord.), 2015, *Couples et familles*, Paris, Insee Références, 190 p.

Breuil-Genier P., Buisson G., Robert-Bobée I., Trabut L., 2016, « Enquête *Famille et Logements* adossée au Recensement de 2011 : comment s'adapter à la nouvelle méthodologie des enquêtes annuelles et quels apports ? », *Économie et statistique*, n°483–485, p. 205–226.

Buisson G., Lapinte A., Le couple dans tous ses états. Non-cohabitation, conjoints de même sexe, pacs... », Insee Première n°1432, Insee, 2013.

Buisson G., La situation matrimoniale dans le recensement : impact de la refonte du questionnaire de 2015, document de travail F1707, Insee, 2017.

Durier S., 2018, *L'échantillon démographique permanent a 50 ans : retours sur un dispositif statistique original*, Présentation aux Journées de méthodologie statistique, Paris, juin.

Godinot A., Durr J.-M., 2016, La rénovation du *Recensement de la population*. In: *Économie et statistique*, n°483–485, p. 7–14.

Imbert C., Lelievre E., Lessault D. (dir.), 2018, *La famille à distance : mobilités, territoires et liens familiaux*, Ined, collections Questions de population, 376 p.

Kreider R., Bates N., Yeris M.-G., 2017, Improving Measurement of Same-Sex Couple Households in Census Bureau Surveys: Results from Recent Tests, SEHSD Working Paper 2017-28.

Kreider R., Lofquist.D, 2014, "Matching Survey Data with Administrative Records to Evaluate Reports of Same-Sex Married Couple Households." SEHSD Working Paper, 2014-36.

Lathe H., Ménard F.-P., Martel L., Hallman S., "Les couples de même sexe au Canada en 2016", Le recensement en bref, Statistiques Canada, No 98-200-X2016007, 2017.

O'Connelll M., Feliz S., 2011, Same-Sex Couple Household Statistics from the 2010 Census**,** SEHSD Working Paper, 2011-26.

Rault W., 2016, « Les mobilités sociales et géographiques des gays et des lesbiennes. Une approche à partir des femmes et des hommes en couple », Sociologie, 7(4), p. 337–360.

Rault W., 2018, « La distance, une composante plus fréquente des relations conjugales et familiales des gays et des lesbiennes ? » in Imbert Christophe, Lelievre Eva, Lessault David (dir.), La famille à distance, Paris, Ined, Questions de populations n° 2.

Rault W., "Secteurs d'activités et professions des gays et des lesbiennes en couple : des positions moins genrées", *Population*, 2017/3 (Vol. 72), p. 399–434.

Toulemon L., Vitrac J., Cassan F., 2005, « Le difficile comptage des couples homosexuels d'après l'enquête EHF », in Lefevre Cécile, Fillon Alexandra (dir.), Histoires de familles, histoires familiales. Les résultats de l'enquête Famille de 1999, Ined, Cahier n° 156, p. 589–602.