Организация Объединенных Наций ECE/CES/GE.41/2022/6

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

Использование методов машинного обучения для определения типа частного домохозяйства в статистике населения и домохозяйств

Записка Федерального статистического управления Швейцарии*

Резюме

Для статистики населения и домохозяйств (STATPOP) была создана переменная «тип домохозяйства». Она обозначает тип домохозяйства для всех частных домохозяйств постоянного населения по основному месту жительства. Она в основном основана на регистровой информации и условном исчислении. Эта переменная используется с 2010 года в Структурном обследовании, ежегодном обследовании, проводимом на выборке из примерно 300 000 человек, которое дает прямые и надежные оценки для районов с населением не менее 15 000 человек.

Эта новая переменная, используемая во всех частных домохозяйствах постоянного населения, позволяет проводить анализ на более детальном географическом уровне, например рассматривать изменения типов домохозяйств с течением времени в муниципалитетах. Ее также можно использовать на индивидуальном уровне в качестве независимой переменной для статистических моделей и для лучшего понимания других тем.

Как и в Структурном обследовании, тип домохозяйства в статистике населения и домохозяйств (STATPOP) рассчитывается по признаку отношений между всеми членами домохозяйства. Однако в регистрах имеются данные только об отношениях между некоторыми членами домохозяйства. Построение типологии домохозяйств STATPOP основано на информации о родственных связях из регистров, дополненной

* Подготовлена Сабриной Браво и Эстель Криппа.
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.

информацией о родственных связях из ежегодных выборок Структурного обследования. В дополнение к этим двум источникам для выявления дополнительных отношений используются детерминистические алгоритмы. В совокупности информация об отношениях из этих источников позволяет нам определить тип домохозяйства для 87 % домохозяйств. Затем применяется алгоритм машинного обучения для условной классификации оставшихся 13 %. Были опробованы различные подходы к условному исчислению отсутствующих данных о родственных связях или отсутствующих данных о типах домохозяйств. На основе проведенных исследований результатов и оценок качества как лучший метод в этом контексте было определено дерево решений.

Результаты в настоящее время публикуются в виде экспериментальной статистики. Переменная «тип домохозяйства» будет интегрирована в текущее производство статистики домохозяйств STATPOP на втором этапе. Ожидается улучшение расчетов и отзывов пользователей.

I. Введение

1. Целью этого проекта является создание переменной «тип домохозяйства» для частных домохозяйств в Швейцарии для постоянного населения, проживающего в основном месте жительства. Тип домохозяйства является важной частью информации не только в области домохозяйств, но и в связи с другими темами. До 2000 года переменная типа домохозяйства выводилась на основе информации, собранной в ходе сплошных переписей населения, которые проводились каждые десять лет. С внедрением новой системы переписей с 2010 года эта переменная теперь используется в Структурном обследовании, ежегодном обследовании, проводимом на выборке из примерно 300 000 человек. Согласно Структурному обследованию, тип домохозяйства определяется по признаку отношений между всеми членами домохозяйства. Обследование дает прямые и надежные оценки для географических районов с населением не менее 15 000 человек. В регистрах населения, которые являются основой статистики населения и домохозяйств, сведения об отношениях между всеми членами домохозяйств отсутствуют. Благодаря новому проекту типологии домохозяйств переменная «тип домохозяйства» теперь доступна в исчерпывающем виде в статистике населения и домохозяйств (STATPOP). Она позволяет проводить анализ на более точном географическом уровне, чем Структурное обследование. Например, можно проследить эволюцию типов домохозяйств с течением времени в муниципалитетах или сопоставить доход от работы по типу домохозяйства на низком географическом уровне. Его также можно использовать на индивидуальном уровне в качестве объясняющей переменной для статистических моделей, чтобы лучше понять другие вопросы, например, связанные с анализом бедности.

2. Переменная типа домохозяйства STATPOP создается в несколько этапов. Во-первых, для установления отношений между отдельными лицами в каждом домохозяйстве используются различные источники данных, затем детерминистический алгоритм присваивает тип домохозяйства по признаку отношений в домохозяйстве для домохозяйств, где все отношения известны. Наконец, алгоритм машинного обучения приписывает тип домохозяйства домохозяйствам, тип которых не указан (примерно 13 %).

3. Категории переменной типа домохозяйства являются следующими[1]:

a) домохозяйство из одного человека;

b) семейная пара без детей;

[1] Вторая типология, основанная на той же процедуре, была рассчитана путем изменения максимального возраста, принятого для детей, на 18 лет вместо 25 лет.

c) сожительствующая пара без детей[2];

d) однополая пара без детей[3];

e) одинокий родитель, по крайней мере, с одним ребенком в возрасте до 25 лет[4];

f) супружеская пара, имеющая как минимум одного ребенка до 25 лет;

g) сожительствующая пара, имеющая как минимум одного ребенка до 25 лет;

h) однополая пара, имеющая как минимум одного ребенка в возрасте до 25 лет;

i) другое домохозяйство из нескольких человек.

II. Источник информации об отношениях

4. Информация об отношениях между разными членами домохозяйства берется из следующих источников:

a) Компьютеризированный реестр актов гражданского состояния (INFOSTAR): для всех лиц, у которых в Швейцарии был зарегистрирован акт гражданского состояния (например, рождение, рождение ребенка, вступление в брак и т. п.), фиксируются отношения «отец−ребенок», «мать−ребенок», «супруга−супруг» и «зарегистрированный партнер». ИНФОСТАР охватывает около 85 % постоянного населения (99,8 % граждан Швейцарии и 45,5 % иностранцев);

b) Реестр дипломатов и сотрудников международных организаций (ORDIPRO): все лица, имеющие разрешение на работу в Швейцарии, выданное Федеральным департаментом иностранных дел, а также члены их семей и помощники по хозяйству вносятся в реестр ORDIPRO. В этом реестре имеется переменная, указывающая на отношения между основным лицом (дипломатом или сотрудником международной организации) и некоторыми взрослыми членами семьи, такими как супруг, партнер и ребенок;

c) Структурное обследование с 2010 года по год n-1: отношения между каждой «парой» членов домохозяйства выводятся на основе ежегодно обследуемой выборки населения. Отношения, указанные в Структурном обследовании, берутся для всех пар людей, проживающих вместе в домашнем хозяйстве в период между временем проведения обследования и составлением типологии, с условиями о семейном положении для совместно проживающих лиц, партнеров, супругов, родственников и неродственников. Охвачены все типы отношений, предусмотренные в Структурном обследовании. Результаты Структурного обследования публикуются после введения переменной типа домохозяйства в STATPOP, поэтому самым последним доступным годом является год n-1;

d) детерминистические алгоритмы: используя демографические переменные, присутствующие в данных статистики населения и домохозяйств на основе регистров (STATPOP), можно определить отношения между двумя людьми с относительно надежным результатом. Например, два человека разного пола, состоящие в браке с одинаковой датой вступления в брак, живущие в одном домашнем хозяйстве, будут определяться как муж−жена. Например, можно найти супружеские пары, иммигрировавшие и не включенные в данные INFOSTAR, поскольку акт их гражданского состояния был зарегистрирован не в Швейцарии.

[2] Сожительствующими парами считаются только пары противоположного пола.

[3] К однополым парам относятся как сожительствующие пары, так и зарегистрированные партнерские отношения.

[4] Ребенок определяется как дочь или сын члена домохозяйства.

III. Подготовка данных

5. Отношения, взятые из указанных выше источников, проверяются на достоверность (например, в отношениях родитель−ребенок между двумя людьми должна быть разница более 12 лет, у ребенка не может быть более двух родителей, человек может состоять в браке только с одним другим лицом) и добавляются возможные симметричные отношения (в случае, когда описывается связь A−B, но не связь B−A).

6. Затем добавляются отдаленные отношения. Для этого на основе проверенных отношений выводятся те отношения, сведений о которых не имеется в данных учета (например, дед, зять, дядя и др.). Цель состоит в том, чтобы выявить как можно больше родственных связей, поскольку тип домохозяйства можно определить только в том случае, если известны отношения между всеми членами домохозяйства.

7. Вклад каждого источника информации об отношениях на 2018 год выглядит следующим образом: из примерно 16 млн «пар» людей в домохозяйстве 84,04 % отношений берутся из источника эмпирических данных или рассчитываются напрямую. Отношения добавляются в следующем порядке: INFOSTAR (75,39 %), ORDIPRO (+0,05 %), Структурное обследование (+6,12 %), детерминистический алгоритм (+1,78 %), отдаленные отношения (+0,7 %).

Таблица 1
Имеющиеся данные об отношениях по источникам данных

	INFOSTAR[5]	*ORDIPRO*	*Структурное обследование*	*Детерминистический алгоритм*	*Отдаленные отношения*
Супруг	X	X	X	X	
Зарегистрированный партнер	X	X	X	X	
Сожитель		X	X	X[6]	
Отец/мать	X	X	X		
Сын/дочь	X	X	X		
Брат или сводный брат/ сестра или сводная сестра[7]	X	X	X		
Отчим/мачеха	X		X		X
Пасынок/падчерица	X		X		X
Дед/бабка	X	X	X		X
Внук/внучка	X	X	X		X
Родственник[8]		X	X		X
Неродственник[9]	X		X		X

[5] Непосредственно доступны только данные об отношениях «супруг», «зарегистрированный партнер», «отец/мать» и «сын/дочь», остальные отношения источника «INFOSTAR» выводятся путем условного определения типа домохозяйства.

[6] Для пар, не состоящих в браке друг с другом, имеющих общего ребенка в одном домохозяйстве.

[7] В классификационную группу сводных братьев/сводных сестер также входят брат и сестра по одному из родителей.

[8] В категорию «родственник» входят дяди, тети, двоюродные братья и т. д.

[9] К категории «неродственников» относятся, в частности, родители жены и мужа, дети сожителя (супруга), дети сожительствующего партнера, сожитель родителя.

8. Алгоритм, который присваивает тип домохозяйства на основе признака отношений между членами домохозяйства, затем применяется к домохозяйствам, где все отношения известны. В 2018 году тип домохозяйства отсутствовал у 13,78% домохозяйств. Другими словами, для 13,78% домохозяйств, по крайней мере, одно из отношений неизвестно при имеющихся данных, что требует вменения.

IV. Условное исчисление

9. Было опробовано несколько методов условного определения типа, в том числе алгоритм случайного леса применительно к отношениям, алгоритм случайного леса применительно к типам домохозяйств, процедура, которая детерминистически условно рассчитывает тип домохозяйства, полиномиальная регрессия и дерево решений. В итоге было выбрано дерево решений. Хотя метод случайного леса дал немного лучшие индивидуальные результаты, метод дерева решений лучше оценивает качество своих условного определения типа и более гибок в плане корректировки.

10. Дерево решений создается с использованием обучающей выборки, содержащей известные типы домохозяйств и выбранную переменную, которая должна иметься как в обучающей выборке, так и в выборочной совокупности домохозяйств (например, число лиц в домохозяйстве, средний возраст членов домохозяйства, число разных фамилий в домохозяйстве, размер муниципалитета и т. д.). С помощью статистических тестов сохраняются переменные, которые лучше всего различают типы домохозяйств и определяют выбранное дерево. Затем это дерево применяется к домохозяйствам, требующим вменения, и, в зависимости от характеристик этих домохозяйств, приписывается тип.

11. В условных расчетах Структурное обследование играет ключевую роль в создании обучающей выборки. Действительно, пересечение между случайно составленной выборкой домохозяйств в Структурном обследовании и подсовокупностью домохозяйств, требующих условной классификации за тот же год, можно рассматривать как область анализа в рамках структурного обследования. Это пересечение рассчитывается для года n-1 и формирует обучающую выборку для условного расчета года n.

V. Полученные результаты

12. Сравнение распределений за 2018 год: распределение типов домохозяйств, полученное после условной классификации численности населения в составе домохозяйств, сопоставляется с распределением типов домохозяйств, оцененным с помощью Структурного обследования. Евклидово расстояние между этими двумя распределениями равно 1,92.

Таблица 2
Распределение типов домохозяйств за 2018 год по данным Структурного обследования и STATPOP
(%)

Тип домохозяйства	*Структурное обследование*	*Исчерпывающая типология (STATPOP)*
Домохозяйство из одного человека	35,69	35,69
Одинокий родитель, по крайней мере, с одним ребенком до 25 лет	4,64	5,24
Другое домохозяйство из нескольких человек	7,59	8,01
Семейная пара без детей	20,11	19,45
Сожительствующая пара без детей	6,53	7,51
Однополая пара без детей	0,61	0,25
Супружеская пара, имеющая как минимум одного ребенка до 25 лет	22,12	20,87
Сожительствующая пара, имеющая как минимум одного ребенка до 25 лет	2,70	2,96
Однополая пара, имеющая как минимум одного ребенка в возрасте до 25 лет	0,03	0,03

13. Оценка погрешности для каждого типа домохозяйства: пересечение между выборкой домохозяйств из Структурного обследования и подгруппой условно определенных домохозяйств за тот же год используется для индивидуального сопоставления каждого условно определенного типа домохозяйств. Это возможно потому, что данные Структурного обследования за этот год не используются для получения исчерпывающего «типа домохозяйств»; поэтому их можно использовать в качестве тестового набора для модели условной определения типа. Веса Структурного обследования учитываются в оценках ошибок.

Таблица 3
Оцененная ошибка условной определения типа по годам

Год	*Оценочная доля ошибок условного определения типа по всем типам домохозяйств (в %)*
2018	1,45
2019	1,48
2020	1,47

14. На данный момент на два типа ошибок приходится около 76 % всех ошибок (т. е. около 1,1 % из 1,45 %). Алгоритм не различает должным образом «гражданскую пару без детей» и «другое домохозяйство, состоящее из нескольких человек» (около 56,5 %), как и «однополую пару без детей» и «другое домохозяйство, состоящее из нескольких человек» (около 19,5 %). Эта путаница в основном затрагивает домохозяйства из двух человек: в случае двух людей противоположного пола алгоритм склонен считать их сожителями, а в случае двух людей одного пола алгоритм склонен рассматривать их как совместно проживающими в одном помещении взрослыми. Этим вызвана недооценка категории «однополая пара без детей». Это один из моментов, который будет пересмотрен и значительно улучшит качество условного определения типа.

VI. Заключение

15. В заключение отметим, что переменную типа домохозяйства STATPOP можно использовать относительно безопасным способом. Так, 87 % типов

домохозяйств получены на основе эмпирических данных, а для остальных 13 % классифицированных в результате условного определения типа домохозяйств оценочная ошибка классификации невелика (около 1,45 % неправильной классификации по всем домохозяйствам). Первоначальные отзывы пользователей подтверждают преимущества исчерпывающего предоставления этой переменной. Прежде чем эту переменную можно будет добавить к текущему производству STATPOP, необходимо доработать несколько моментов. В их числе — уточнение условного определения некоторых категорий типов домохозяйств, добавление проверок достоверности типа домохозяйств во время условного определения типа[10] и, возможно, объединение определенных категорий в одну группу, если переменная по-прежнему соответствует потребностям пользователей. Подробная документация по методике доступна на сайте ФСУ (Typology of households (STATPOP)).

[10] Определение наиболее вероятного типа домохозяйства в соответствии с деревом решений, которое также является достоверным в соответствии с данными о членах домохозяйств. Например, если «супружеская пара, по крайней мере, с одним ребенком младше 25 лет» является наиболее вероятным типом, но ни один член домохозяйства не моложе 25 лет, будет условно приписан второй наиболее вероятный тип, если он достоверен.