Методы машинного обучения для восстановления недостающих данных о типе частного домохозяйства

Встреча экспертов ЕЭК ООН по переписям населения и жилого фонда

21-23 сентября 2022 года

Д-р Жан-Поль Каутен, Федеральное статистическое управление Швейцарии

Авторы

- Сабрина Браво
- Эстелле Криппа

План презентации

1. Контекст
2. Актуальность
3. Производство статистической информации и восстановление недостающих данных
4. Результаты
5. Заключение

1. Контекст

- До 2000 года информация о типе частного домохозяйства собиралась один раз в 10 лет в ходе полной переписи населения.

- После 2010 года информация о типе частного домохозяйства собирается в рамках Структурного обследования, причем ежегодно размер выборки составляет 300 000 человек (позволяет получить достоверные оценки для групп > 15 000 человек).

- Начиная с 2021 года тип домохозяйства определяется как на основе имеющейся в регистрах информации, так и с использованием алгоритмов **машинного обучения.** Публикация на сайте экспериментальной статистики ФСУ.

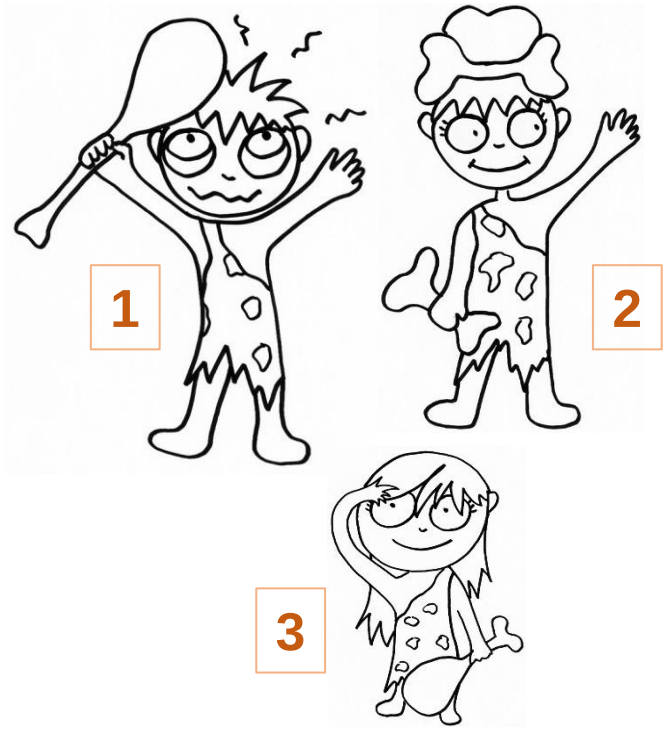
2. Актуальность

Исчерпывающая информация о типе домохозяйства позволяет:

- Проводить анализ на уровне малых географических единиц, например:
  - Распределение типов домохозяйств в разных районах больших городов;
  - Сравнение уровня социальной помощи в разных муниципалитетах с разбивкой по типу домохозяйства;
- Проводить исследования на индивидуальном уровне (как объясняющая переменная для статистических моделей, например при анализе бедности).

3. Производство статистической информации - ход работы

Тип домохозяйства зависит от отношений внутри домохозяйства



Отношения

P1	P2	Отношение
1	2	Муж
1	3	Отец
2	1	Жена
2	3	Мать
3	1	Дочь
3	2	Дочь

Тип домохозяйства

Семейная пара с ребенком

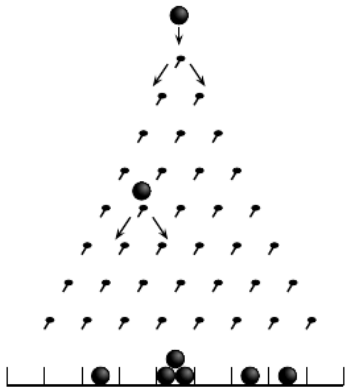
3. Производство статистической информации - источники

Источники данных для определения типа домохозяйства:

	Отношения	Тип домохозяйства
1. Отношения по данным *регистров*	**75,4%**	
2. Отношения по данным *Структурного обследования*	**+6,1%**	
3. Отношения, выведенные логически с использованием *детерминированных алгоритмов*	**+2,5%**	
	84% →	**86,2%**
4. Типы, рассчитанные с помощью ***машинного обучения*** (дерево принятия решений)	**+13,8%**	
	100%	

3. Производство статистической информации - восстановление недостающих данных

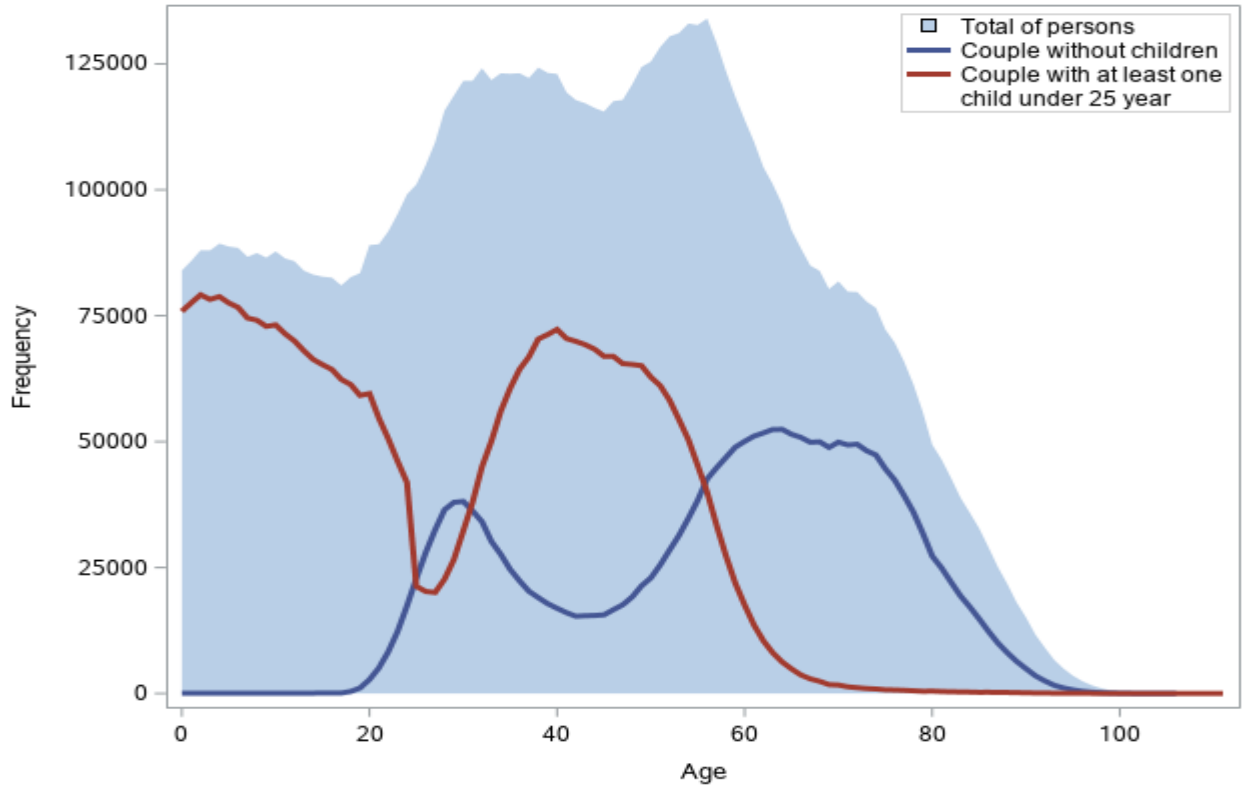
- Для 14% домохозяйств, тип которых невозможно было определить в силу отсутствия информации об отношениях между, по крайней мере, одной парой лиц, были протестированы различные методы восстановления недостающих данных.
- В итоге для восстановления недостающих данных и определения типа остальных домохозяйств был выбран такой метод как **дерево принятия решений**:



1. Использование обучающего множества, в который входят домохозяйства известного типа и выбранные переменные (например, средний возраст членов домохозяйства, размер муниципалитета, ...).
2. С помощью статистических тестов на обучающем множестве сохраняются переменные, которые лучше всего различают типы домохозяйств и определяют выбранное дерево.
3. Затем это дерево применяется к домохозяйствам, тип которых необходимо восстановить, и в зависимости от характеристик этих домохозяйств определяется их тип.

4. Результаты - пример распространения

Постоянное население в частных домохозяйствах с разбивкой по возрасту и типу



5. Заключение

- В 86% случаев тип домохозяйства определяется на основе имеющихся данных и в 14% случаев - с помощью машинного обучения.
- Приблизительно 1,45% неправильных распределений по всем домохозяйствам.
- Первоначальные отзывы пользователей подтверждают, что такая переменная должна быть исчерпывающей, то есть без пропущенных значений, и пригодной для всего постоянного населения.
- Эта переменная будет добавлена в текущую ежегодную работу после доработки некоторых аспектов.