Оценка качества результатов переписи

Eric Schulte Nordholt/Эрик Шульт Нордхолт

Statistics Netherlands

Содержание (1)

- Качество официальной статистики
- Методы сбора данных
- Затраты
- Нагрузка на ответ
- Качество продукта
- Статистика на основе регистров по сравнению со статистическими обследованиями
- Возрастающая роль административных данных в статистическом процессе
- Качество, административные данные и статистический процесс
- Конфиденциальность данных

Содержание (2)

- Комбинированное исследование
- Рассмотрение данных в переписи населения Нидерландов 2011 г.
- Гиперразмер/гипериизмерения источника результатов
- Гиперразмер/гипериизмерения метаданных результатов
- Гиперразмер/гипериизмерения данных результатов
- Выводы комбинированных исследований
- Файл об образовании
- Вменение файла образовательного достижения
- Подведение итогов

Качество в официальной статистике

- Определение качества в статистике согласно европейскому «Своду правил»
- Качество продукта
  - Релевантность
  - Точность
  - Своевременность и пунктуальность
  - Сопоставимость и согласованность
  - Доступность и прозрачность
- Качество процесса
  - Лучшие методы
  - Экономия затрат
  - Низкая нагрузка на ответ

Методы сбора данных

Опции
- Перепись/полное статистическое обследование
- Образец исследования
- Административные регистры

Административные данные собираются для цели
- Статистика на основе регистров является вторичным использованием существующих данных.

Решение о методе сбора данных является компромиссным между:
- Экономией затрат
- Нагрузкой на ответ
- Качеством продукта

Затраты

Текущая ситуация во многих странах
- The NSIs have experienced budget cuts / restrictions
- Users demand new and more detailed statistics
- Must increase efficiency in production of statistics

Административные данные
- Практически никаких затрат на сбор данных (для НСУ)
- Используйте ресурсы для улучшения существующих данных вместо сбора данных для статистических целей
  - Дополнить и исправить существующие данные
  - Большинство ресурсов используется при создании статистических систем на основе регистров
  - НО: системы должны поддерживаться

Статистика на основе регистров не бесплатна, но обычно дешевле выборочных обследований и особенно традиционной переписи

Нагрузка на ответ

Использование административных данных означает отсутствие дополнительного бремени на ответ

- Для компаний
  - “Сообщение властям занимает слишком много времени”
- Для граждан
  - “Власти не должны запрашивать информацию, которую я уже предоставил”
- Для НСУ
  - Увеличение проблем с отсутствием ответов в выборочных обследованиях и переписях

Качество продукта (1)

Релевантность/актуальность

- Данные регистра основаны на административных определениях, которые могут отличаться от статистических определений
  - Единицы измерения, покрытие, переменные, ссылки на время и т. Д.
- “У нас есть правильные ответы, но можем ли мы ответить на правильные вопросы?»
- “Картина властных структур мира?”
- Объединение данных из разных регистров для повышения релевантности
- В некоторых случаях: необходим дополнительный сбор данных

Качество продукта(2)

Точность

- Регистры обычно имеют хорошее качество для административных целей.
- Повышение точности за счет объединения данных из нескольких регистров
  - Редактирование для статистических целей

Качество продукта(3)

Своевременность и пунктуальность

- Время производства иногда больше, чем для статистических обследований
  - Административный процесс может занять некоторое время (пример: налоговые данные)
  - Задержка в обновлении реестров
  - Извлечение данных: необходимо подождать несколько недель или месяцев, а иногда и дольше

Качество продукта(4)

Сопоставимость и согласованность

- Создание согласованной статистической системы на основе регистров
- Согласование со статистикой на основе других источников
  - Опыт Нидерландов

Доступность и прозрачность

- Практически не зависит от используемых источников данных

Статистика на основе регистров по сравнению со статистическими обследованиями(1)

- Затраты (++)
- Нагрузка на ответ(++)
- Актуальность/релевантность (-)
  - Не все переменные включены в регистры
  - Менее прямой контроль над содержанием данных
- Точность (о)
- Своевременность (-)

Статистика на основе регистров по сравнению со статистическими обследованиями(2)

Предложение административных регистров

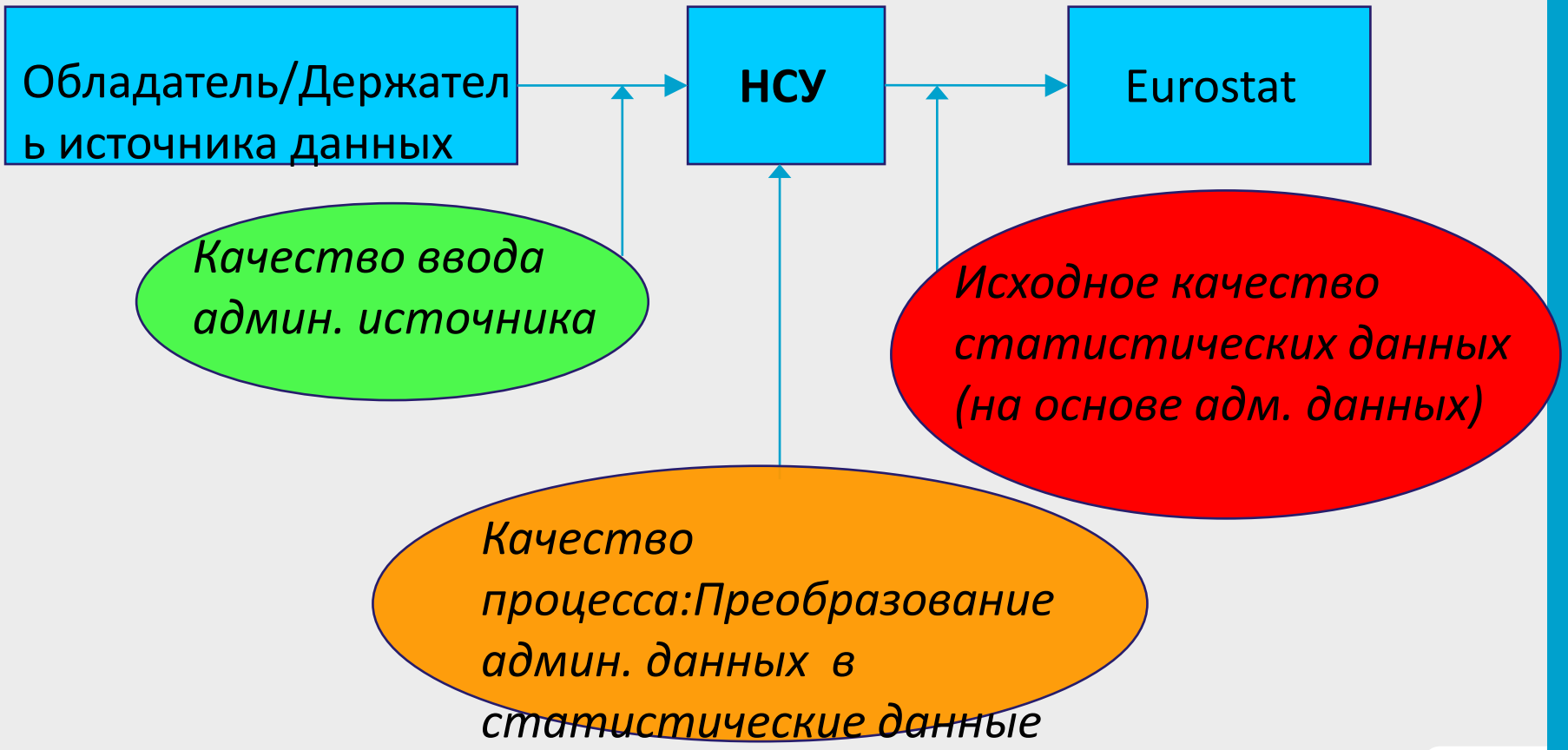
- Полное покрытие по низкой цене
  - Возможна статистика по малым группам (по сравнению с выборочными обследованиями)
- Годовые (или более частые) данные по всем переменным
  - Годовые “переписи”
- Производство статистики на основе административных данных оказалось эффективным
- Статистика на основе регистров должна быть дополнена информацией из выборочных обследований.

Возрастающая роль административных данных в статистическом процессе

- Все больше и больше статистических учреждений используют административные источники для статистических целей.
  - В основном для снижения затрат и нагрузки на ответ

- Однако в результате они :
  1. Become more *dependent* on data sources collected and maintained by *others* Становятся *более зависимыми* от источников данных, собранных и поддерживаемых *другими*
     - Необходимо контролировать *качество* этих источников данных, когда они входят в офис
  2. Необходимо найти новые источники данных, которые содержат необходимую информацию
     - Перед применением необходимо оценить удобство использования этих источников данных.

Качество, административные данные и статистический процесс



Конфиденциальность данных(1)

Законодательство:

– Закон о статистике Нидерландов
– Закон Нидерландов о защите данных → с 2018 г. GDPR

Эти законы:

– разрешают Статистическому управлению Нидерландов использовать персональные данные
– обязывают Статистическое управление Нидерландов принять адекватные меры, направленные на защиту конфиденциальности

Конфиденциальность данных(2)

Меры :

- Ключи привязки анонимны, из данных удалены исходные персональные идентификаторы
- Права доступа к микроданным ограничены

Комбинированное исследование

Разработка системы качества для административных данных

Решения о данных о вторичных источниках в переписи населения Нидерландов 2011 г.

Рассмотрение данных в переписи населения Нидерландов 2011 г.

Регистры:
– Регистр населения (PR), 17 миллионов записей
– Файл вакансий, содержащий всех сотрудников
– Файл самозанятых, содержащий всех самозанятых
– Регистр пособий по безработице(UR/РПБ)
– Регистр социального обеспечения (SR/СР)
– Регистр образования(ER/РО)
– Новый жилищный регистр(HR/НЖР)

Исследования:
– Обследование рабочей силы(LFS/ОРС)

Гиперразмер/гипериизмерение источника результатов

Dimensions	*Data sources*				
	ER	UR	SR	HR	PR
1. Supplier	+	o	o	+	+
2. Relevance	o	+	+	o	+
3. Privacy and security	+	+	+	+	+
4. Delivery	-	+	+	+	+
5. Procedures	o	o	o	+	+

Низкая частота доставки

Серьезно страдает от выборочного недостаточного охвата

Назначение поставщика данных неясно

Отсутствуют важные переменные

Результаты гиперизмерения метаданных

Dimensions	*Data sources*				
	ER	UR	SR	HR	PR
1. Clarity	+	+	+	+	+
2. Comparability	-	o	o	+	+
3. Unique keys	+	+	+	o	+
4. Data treatment	+	+	+	+	+

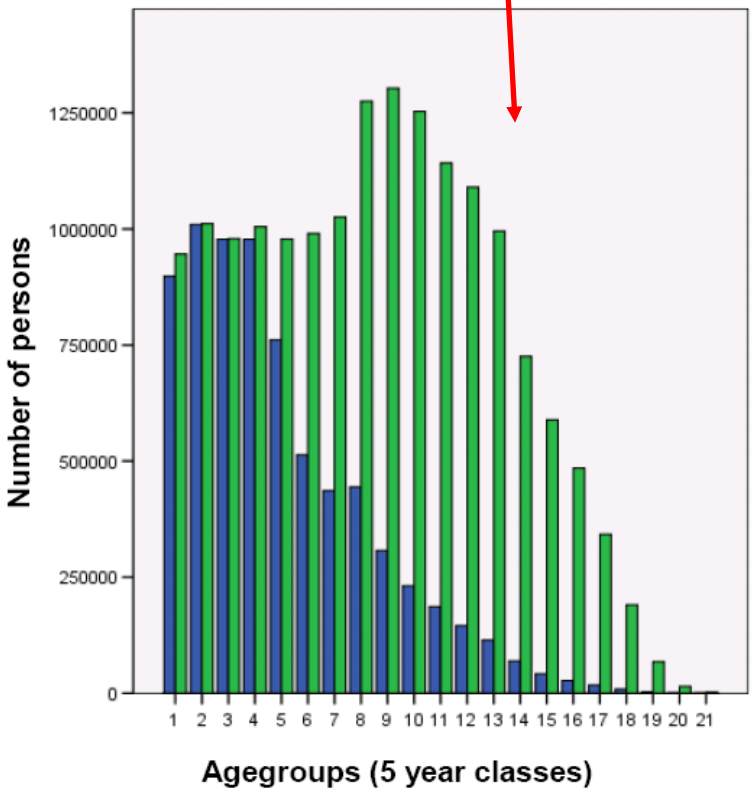
Временной период в источнике не может быть легко перенесен в нужную временную точку

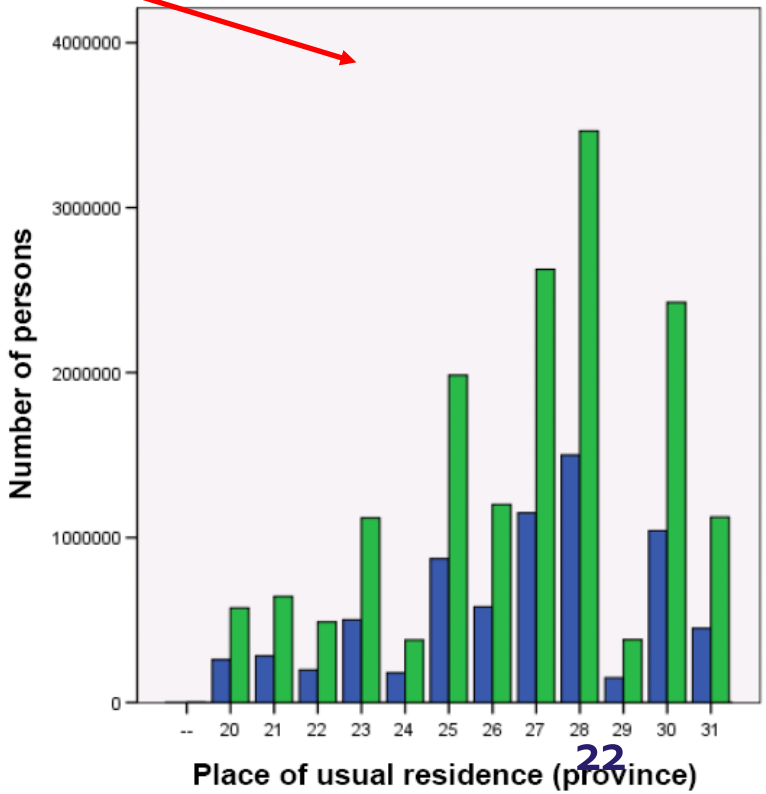
Разница во времени отчетных периодов

Уникальные ключи не могут быть легко использованы для установления связей

Рез-ты гипериизмерения данных- полнота

Variable	*Number of missings*	*Percentage missing (%)*
Educational attainment	9.238.212	56,3
Current activity status	2.140.266	13,0

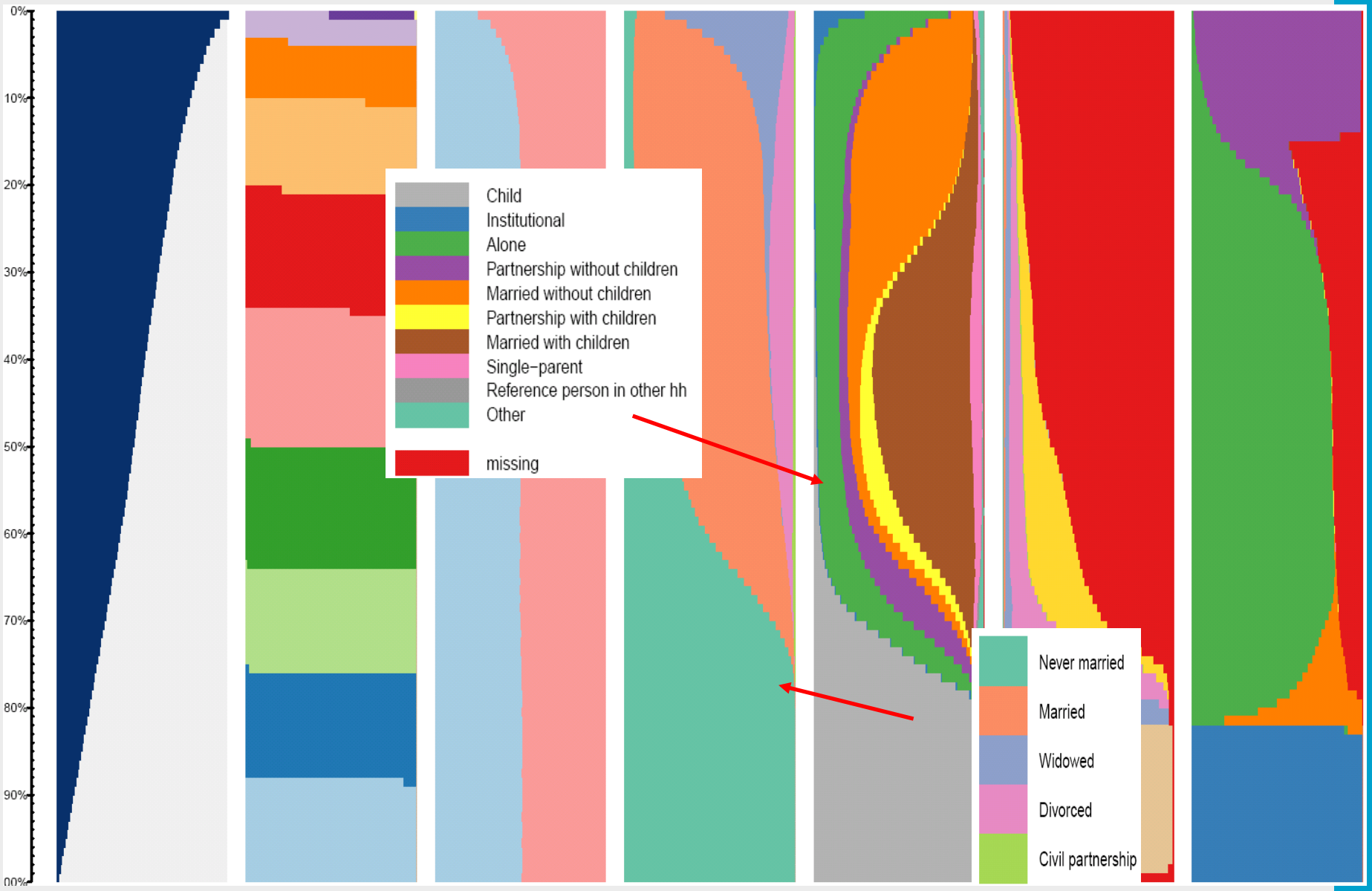




Результаты гипериизмерения данных– точность

Ageclass	Current activity status							
	Missing	0	1	2	3	4	5	6
1: [0, 5)	0	945861	0	0	0	0	0	0
2: [5, 10)	0	1011159	0	0	0	0	0	0
3: [10, 15)	0	978964	0	0	0	0	0	0
4: [15, 20)	34911	0	482180	33	0	487533	11	293
5: [20, 25)	113286	0	716411	106	0	147395	190	711
6: [25, 30)	142149	0	818167	107	0	28396	486	677
7: [30, 35)	163141	0	856030	129	0	4506	744	771
8: [35, 40)	216807	0	1053407	180	0	2418	1138	1056
9: [40, 45)	228634	0	1070204	228	0	1853	1076	1224
10: [45, 50)	236102	0	1013249	242	0	1134	1076	1434
11: [50, 55)	262473	0	875724	253	1	504	1261	1789
12: [55, 60)	330898	0	714959	263	39705	232	1776	2253
13: [60, 65)	390062	0	343089	122	256826	78	2348	2764
14: [65, 70)	8730	0	88209	1	628490	16	3	46
15: [70, 75)	5306	0	35690	1	548059	3	0	22
16: [75, 80)	3822	0	14705	0	466339	2	0	19
17: [80, 85)	2166	0	5897	0	333936	0	0	8
18: [85, 90)	1115	0	2360	0	186690	0	0	8
19: [90, 95)	405	0	662	0	66339	0	0	0
20: [95, 100)	162	0	136	0	14386	0	0	0
21: [100, ∞)	97	0	18	0	1450	0	0	0

[4] Current activity status: (0). Persons below minimum age for economic activity, (1) Employed, (2) Unemployed, (3) Pension or capital income recipients, (4) Students not economically active, (5) Homemakers, (6) Others


0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
00%
Child
Institutional
Alone
Partnership without children
Married without children
Partnership with children
Married with children
Single-parent
Reference person in other hh
Other
missing
Never married
Married
Widowed
Divorced
Civil partnership

Выводы комбинированного исследования

– Качество официальной статистики является важным аспектом, особенно когда используются интегрированные данные.

– Виртуальная перепись населения оказалась успешной концепцией в Нидерландах

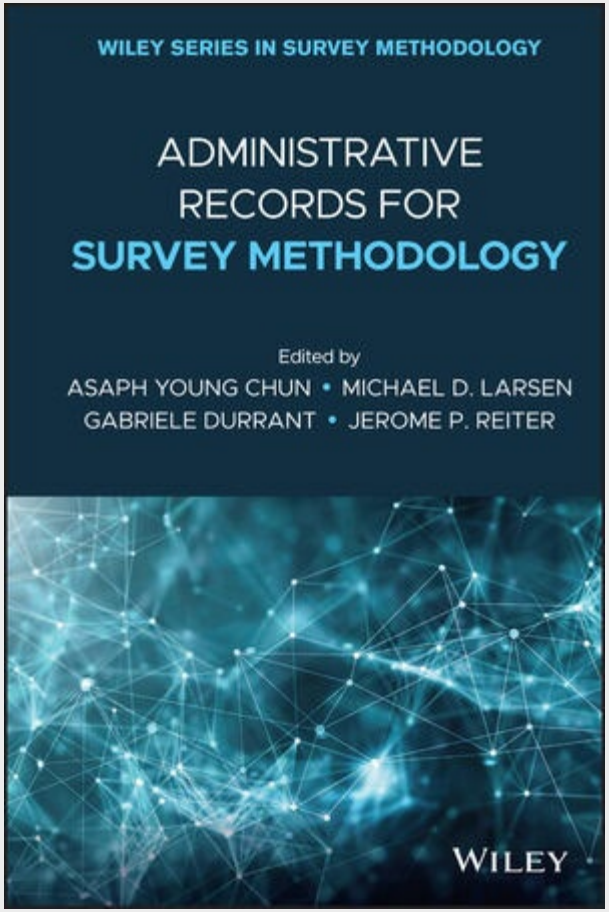
– Рамки качества - полезный инструмент для принятия решений о данных в виртуальной переписи населения

– Последующие действия: регистр образования был усовершенствован (включая новые данные, иностранное и частное образование) и используется в переписи населения 2021 года.

Выводы комбинированного исследования

Больше информации содержится в:
Daas, P., E. Schulte Nordholt, M. Tennekes and S. Ossen, 2021. Evaluation of the Quality of Administrative Data Used in the Dutch Virtual Census. In: Administrative Records for Survey Methodology, Wiley Online Library, 2021, pp. 63-83.

ISBN: 978-1-119-27204-5
April 2021
384 Pages



Файл об образовании

- Сложный процесс интеграции микроданных из OPC и журналов обследований
- Новая версия, содержащая также информацию о частных учебных заведениях, доступных с 2016 года.
- Около 60 % записей содержат информацию о наивысшем достигнутом уровне образования
- Взвешивание известных маргиналов населения для статистики уровня образования
- Лучшее качество и более подробные таблицы образования, чем раньше, когда использовалась только информация LFS/OPC.

Включение файла образовательного достижения

В этом проекте:

- Следует найти хорошую модель включения для файла образовательных достижений (модель логистической регрессии).
- Снова создать набор гиперкубов переписи 2011 г., теперь используя вмененный файл образовательных достижений.
- Разработать набор показателей качества (основа для принятия решения о том, насколько подробной будет будущая публикация переписи)
- План высшего уровня образования, полученного в ходе переписи населения Нидерландов 2021 года (также представляет интерес для других стран)

Подведение итогов

- Темой обсуждения была оценка качества административных данных и конфиденциальность данных
- Многие вещи можно объяснить, некоторые вещи нужно испытать в вашем собственном (национальном) контексте
- Качество в административном контексте сильно отличается от качества в контексте обследования
- Использование административных данных в будущих переписях будет расширяться
- Много интересной работы на будущее!

Большое спасибо за Ваше внимание!