ЕЭК ООН

Руководство по оценке качества административных источников для использования в переписях населения

ОРГАНИЗАЦИЯ ОБЪЕДИНЕННЫХ НАЦИЙ

ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ
ОРГАНИЗАЦИИ ОБЪЕДИНЁННЫХ НАЦИЙ

Руководство по оценке качества административных источников для использования в переписях населения

Подготовлено Целевой группой Конференции европейских статистиков по оценке качества административных источников для использования в переписях населения

Организация Объединённых Наций
Женева, 2021

Данная публикация издана на английском и русском языках. Перевод на русский язык был подготовлен Межгосударственным статистическим комитетом Содружества Независимых Государств в рамках проекта Фонда ООН в области народонаселения CISPop «Качественные данные — Эффективная политика», финансируемого Российской Федерацией.

Публикация Организации Объединённых Наций,
выпущенная Европейской Экономической Комиссией
Организации Объединённых Наций

ECE/CES/STAT/2021/4

Предисловие

Основная цель настоящей публикации заключается в предоставлении организаторам переписей населения и жилищного фонда рекомендаций по оценке качества административных данных для использования их при проведении переписей населения. Руководство охватывает практические этапы оценки, от работы с поставщиком административных данных (или административным органом) для понимания источника, его сильных и слабых сторон, вплоть до получения и анализа фактических данных. Руководство охватывает ключевые параметры качества, по которым проводится оценка, с использованием различных инструментов и показателей. В завершение Руководство также включает информацию об этапах обработки и результатах переписи с использованием административных источников.

Настоящая публикация была подготовлена Целевой группой, учреждённой Конференцией европейских статистиков (КЕС) и состоящей из экспертов национальных статистических служб (НСС), работа которой координировалась Европейской экономической комиссией Организации Объединённых Наций (ЕЭК ООН).

Данный перевод на русский язык был подготовлен Статкомитетом СНГ по согласованию с Европейской экономической комиссией ООН в рамках проекта CISPop «Качественные данные — Эффективная политика», реализуемого Фондом ООН в области народонаселения и финансируемого Российской Федерацией.

Выражение признательности

Данное Руководство было подготовлено Целевой группой ЕЭК ООН при участии следующих членов:

Стивен Данстан (Председатель), Великобритания
Катрин Чонер, Австрия
Кристоф Вальднер, Австрия
Хосе Морель, Канада
Лионель Эспинасс, Франция
Стефан Диттрих, Германия
Тобиас Каленберг, Германия
Томас Кёрнер, Германия
Ингеборг-Ворндран, Германия
Шила Бонэм, Ирландия
Брендан Мерфи, Ирландия
Алаа Атраш, Израиль
Яэль Файнштейн, Израиль
Херардо Галло, Италия
Донателла Зиндато, Италия
Снежана Ремикович, Черногория
Эрик Шульте Нордхольт, Нидерланды
Кристин Бикрофт, Новая Зеландия
Эбби Морган, Новая Зеландия
Януш Дыгашевич, Польша
Кшиштоф Возница, Польша
Жоау Фаррахота, Португалия
Сандра Лагарто, Португалия
Паула Паулино, Португалия
Дмитрий Калинку, Республика Молдова
Валентина Истрати, Республика Молдова
Марина Перес Хулиан, Испания
Альберто Сальседо, Испания
Эбнем Беше-Канполат, Турция
Мухаррем Гюрлейен Гёк, Турция
Мехмет Шабан Укари, Турция
Луиза Блэквелл, Соединённое Королевство
Адриана Кастальдо, Соединённое Королевство
Сара Коррейя, Соединённое Королевство
Сара Хейлок, Соединённое Королевство
Шарлотта Хиллард, Соединённое Королевство
Джек Сим, Соединённое Королевство
Стефан Тиц, Соединённое Королевство
Клэр Уотсон, Соединённое Королевство
Марина Райт, Соединённое Королевство
Том Мул, Соединённые Штаты Америки
Эдуард Джонгстра, ЮНФПА
Диана Белтадзе, Евростат
Сорина Важу, Евростат
Иан Уайт, независимый эксперт
Фиона Уиллис-Нуньес, ЕЭК ООН

Руководство было разработано и согласовано всей Целевой группой ЕЭК ООН. Каждая глава была составлена группой под руководством одного или нескольких человек:
Методология переписи и использование административных данных для переписей населения: Сара Коррейя и Иан Уайт;
Система качества: Сара Коррейя и Сорина Важу;
Этап Источник: Сорина Важу, Жозе Морель, Диана Белтадзе и Стивен Данстан;
Этап Данные: Кристоф Вальднер, Тобиас Каленберг и Сара Коррейя;
Этап Процесс: Сара Хейлок, Эбби Морган, Адриана Кастальдо и Стивен Данстан;
Этап Итоговые материалы: Сара Коррейя, Марина Перес, Сандра Лагарто и Стивен Данстан.
Руководство редактировалось Ароной Пистинер.

Европейская ассоциация свободной торговли (EACT), межправительственная организация, в которую входят Исландия, Лихтенштейн, Норвегия и Швейцария, оказала финансовую поддержку в редактировании настоящего Руководства. Статистическое управление EACT оказывает странам статистическую помощь в сотрудничестве с другими международными организациями.

Целевая группа выражает особую благодарность Управлению национальной статистики Соединённого Королевства (ONS) за неоценимый вклад многих его сотрудников в подготовку этого окончательного Руководства, в первую очередь Стивену Данстану, Саре Коррейя и Адриану Кастальдо.

Данный перевод на русский язык был подготовлен Статкомитетом СНГ по согласованию с Европейской экономической комиссией ООН в рамках проекта CISPop «Качественные данные — Эффективная политика», реализуемого Фондом ООН в области народонаселения и финансируемого Российской Федерацией.

Содержание

Перечень вставок

Перечень диаграмм

Перечень таблиц

Перечень тематических исследований по странам

Аббревиатура

ABPE	Административные данные, базирующиеся на оценках численности
ABS	Бюро статистики Австралии
AIDA	Интегрированная база административных данных, Италия
CAPI	Персональный опрос
CATI	Опрос по телефону
CAWI	Опрос по Интернету
CAxI	Мультимодальный опрос
CDB	Центральная база данных (Австрия)
CES	Конференция европейских статистиков
CIS	Информационная система потребителей
CPR	Центральный регистр населения
CT	Тестирование переписи (Португалия)
DSE	Двойственная оценка системы
DA	Демографический анализ
ESS	Европейская статистическая система
ESSnet	Сеть Европейской статистической системы
FDP	Окончательный набор данных (Австрия)
FPC	Предпереписной файл (Испания)
GSBPM	Общая статистическая модель бизнес-процесса
GMTMM	Обобщенная модель с несколькими признаками и несколькими методами
INSEE	Национальный Институт статистики и экономических исследований (Франция)
ISO	Международная организация стандартизации
LAU	Местная административная единица
LMS	Юридическое брачное состояние
MOU	Меморандум о взаимопонимании
MSE	Среднеквадратичная ошибка
NHS	Национальная служба здравоохранения (Соединенное Королевство)
NIP	Идентификационный номер налогоплательщика (Польша)
NRFU	Последующие меры в связи с отсутствием ответа
NSO	Национальный статистический орган
NUTS	Номенклатура территориальных единиц для статистики
NZ	Новая Зеландия
ONS	Национальная статистическая служба (Соединенное Королевство)
PBR	Базовый регистр населения (Италия)
PE	Оценка численности населения
PES	Постпереписное обследование
PESEL	Универсальная электронная система регистрации населения (Польша)
PHC	Перепись населения и жилищного фонда
PIA	Оценка воздействия на конфиденциальность
PII	Личная, идентифицируемая Информация
PIN	Personal Identification Number
PPHC	Постоянная перепись населения и жилищного фонда (Италия)
PR	Регистр пациентов
QA	Оценка качества
RBI	базовый регистр населения (Италия)
REGON	Идентификационный номер предприятия (Польша)

ROC	Рабочая характеристика приемника
RSBL	Базовый статистический регистр адресов (Италия)
SCD	Набор статистических данных переписи (Польша)
SDC	Контроль за раскрытием статистических данных
SE	Статистика Эстонии
SIR	Интеграционная система статистических регистров (Италия)
SOL	Признаки жизни
SP	Статистика Португалии
SPD	Набор статистических данных о населении (Португалия)
Stats NZ	Статистика Новой Зеландии
TSE	Общая ошибка опроса
UK	Соединенное Королевство
UKSA	Статистическое управление Соединенного Королевства
UNECE	Европейская экономическая комиссия ООН
UPRN	Уникальный идентификационный номер объекта
US	Соединенные Штаты
VOA	Оценка офиса
VQS	Переменная система качества

Краткое содержание

Использование административных данных в переписях продолжает расти во всех странах региона ЕЭК ООН и за его пределами, будь то поддержка традиционной переписи в рамках комбинированной переписи или методологии переписи на основе регистров, при которой производится подсчёт населения и / или заполнение данных переписи с использованием административных данных. Важно, чтобы НСС понимали сильные и слабые стороны административных данных для использования их в переписи населения, чтобы гарантировать принятие правильных решений об использовании таких данных.

Цель Руководства заключается в том, чтобы предоставить организаторам переписей практические рекомендации по оценке качества административных данных с использованием последовательных этапов оценки. В Руководстве используются системы качества и передовые практики, принятые НСС по всему миру, включая широко используемую структуру Статистического управления Нидерландов (см. Daas et al. 2012), модели общей погрешности, принятой Статистическим управлением Новой Зеландии (см. Zhang 2012) и результаты проекта «Методологии статистической сети для комплексного использования административных данных в статистическом процессе» (см. Eurostat ESSnet MIAD 2014).

Руководство основывается на четырёх этапах: Источник, Данные, Процесс и Итоговые материалы, причём первые два из этих этапов являются основным объектом Руководства, обеспечивая оценки качества вводимых данных (т.е. оценка качества административных источников для использования в целях переписи).

Этап Источник. Стадия источника включает оценку административного источника посредством работы с поставщиком данных и анализа соответствующих метаданных. Этот этап включает оценку того, может ли источник удовлетворить потребности переписи по качественным параметрам релевантности, точности, своевременности, согласованности и сопоставимости. Также проводится оценка доступности и интерпретации административного источника, включая любые ограничения доступа и использования, приемлемость использования для общественности. Наконец, проводится оценка того, может ли поставщик данных удовлетворить потребности НСС, с учётом таких факторов, как прочность отношений с поставщиком и статус поставщика.

Этап Данные. Этап данных включает оценку, основанную на анализе фактических данных (переданных поставщиком данных) и путём сравнения с другими источниками. Этот этап включает проверку данных при получении, оценку точности и надёжности (включая ошибки охвата и измерения), своевременности и пунктуальности, а также оценку возможности связывания. Для этапов Источник и Данные оценка проводится по ключевым параметрам качества данных, для которых предусмотрены различные инструменты и индикаторы.

Этап Процесс и этап Итоговые материалы представлены для полноты и дают пользователям информацию об основных процессах и соображениях по преобразованию административных данных для использования в переписи, а также для оценки качества результатов переписи, основанных на административных данных.

Руководство включает опыт нескольких стран с использованием иллюстраций и более подробных практических примеров.

В заключительной главе Руководства также представлены предлагаемые направления для дальнейшей работы и набор ключевых рекомендаций для НСС, которые следует учитывать:

1. Определите административные источники для конкретных вариантов использования, чтобы оценить ожидаемые или требуемые результаты использования источника для определённого варианта применения.
2. Налаживайте и поддерживайте отношения между НСС и поставщиками данных, используя правовую основу для предоставления и использования данных, а также механизмы совместной обратной связи.
3. Используйте отношения с поставщиками для обеспечения всестороннего понимания исходных метаданных.
4. Оцените согласованность и совместимость административного источника с данными переписи, чтобы понять различия, если таковые имеются, между требуемыми совокупностями, концепциями, определениями и временными измерениями.
5. Определите ограничения и проблемы, связанные с получением данных из административного источника и их включения в перепись, сравнивая ценность преодоления этих трудностей с усилиями и рисками, связанными с этим.
6. Оценка и управление рисками, связанными с использованием административных источников.
7. Следует обеспечивать прозрачность в общении с пользователями данных и с общественностью о том, как и почему административные данные используются для переписи, уделяя особое внимание процедурам обеспечения эффективного использования и защиты данных.
8. Провести технико-экономическое обоснование в качестве «доказательства концепции» и провести тестовые прогоны с реальными данными до включения административных данных в переписи населения.
9. Используйте экспертную оценку и проведите сравнения источников с течением времени, чтобы выявить проблемы с качеством в каждом источнике.
10. Записывайте и публикуйте результаты оценки качества на всех этапах.
11. Разрабатывайте структуру и стратегию обеспечения качества для конкретных НСС, подкреплённые чёткой и исчерпывающей документацией и процедурами обучения, с упором на непрерывную оценку и обмен информацией между НСС, пользователями данных и поставщиками данных.

Глава 1 Введение

1.1 Справочная информация

1. В 2017 году Целевая группа ЕЭК ООН по регистровым и комбинированным переписям подготовила Руководство по использованию регистров и административных данных в целях переписей населения и жилищного фонда[1]. Руководство включало раздел, посвящённый «источникам данных и их качеству», с общим обсуждением этой темы. Эксперты, принявшие участие в Совещании экспертов ЕЭК ООН-Евростата по переписям населения и жилищного фонда (Женева, 4–6 октября 2017 года), определили качество административных источников как тему, имеющую первостепенное значение для многих стран. Исходя из этого Совещание экспертов призвало к созданию новой Целевой группы ЕЭК ООН по измерению качества административных источников для использования в целях переписи, которая будет опираться на результаты работы предыдущей Целевой группы.

2. Данная Целевая группа была учреждена в 2018 году и круг её ведения[2] был утверждён на совещании Бюро Конференции европейских статистиков (КЕС), состоявшемся 14–15 февраля 2018 года в Хельсинки. Целевая группа представила доклад Руководящей группе ЕЭК ООН по переписям населения и жилищного фонда, которая, в свою очередь, представила доклад КЕС и её Бюро.

3. Задача Целевой группы заключалась в разработке руководства по измерению качества административных источников для использования в целях переписи[3]. Круг ведения предусматривал разработку руководства, которое будет актуальным для всех стран ЕЭК ООН, с опорой на работу по использованию административных источников в целях формирования официальной статистики, проводимую в рамках проекта Евростата ESS.VIP ADMIN[4].

4. Целевая группа встречалась очно во время совещаний экспертов ЕЭК ООН-Евростата по переписям населения и жилищного фонда 2018 и 2019 годов и провела дополнительное очное совещание 5-6 марта 2020 года в Женеве, Швейцария.

5. Целевая группа представила годовые отчёты о проделанной работе Группе экспертов ЕЭК ООН-Евростата по переписям населения и жилищного фонда в 2018, 2019 и 2020 годах. Полный проект данного Руководства был представлен на рассмотрение до начала онлайн совещания экспертов (30 сентября – 1 октября 2020

[1] URL: http://www.unece.org/index.php?id=50794.
[2] URL: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/bur/2018/February/06Add.1-TF_on_quality_of_admin_data_for_censuses__ToR__apr.pdf.
[3] Впоследствии Целевая группа приняла решение скорректировать свое название и соответствующую задачу на «оценку», а не «измерение» качества административных источников для использования в целях переписи.
[4] Более подробную информацию об этом проекте можно найти на сайте: https://ec.europa.eu/Eurostat/cros/content/ess-vision-2020-admin-administrative-data-sources en

года). Отзывы, полученные от участников, были использованы для доработки Руководства.

6. Настоящее Руководство служит практическим пособием по оценке и измерению качества административных источников для использования в целях переписей населения и жилищного фонда.

1.2 Использование административных данных в целях переписи

7. Источниками административных данных являются массивы данных, содержащие собранную, главным образом, информацию для административных целей[5]. Они содержат данные, которые собираются государственными органами, правительственными ведомствами и другими организациями в целях регистрации, проведения операций и учёта, как правило, в связи с предоставлением каких либо услуг. К ним относятся административные регистры с уникальным идентификатором, такие как регистры населения, предприятий, адресов, образования, здравоохранения, занятости, налогообложения, а также другие административные источники без уникального идентификатора. Административные регистры и/или другие административные источники используются для создания статистических регистров, которые используются исключительно в статистических целях, в том числе для проведения переписи. Административные источники, наиболее часто используемые при переписи, описаны в Главе 2 Руководства.

8. Использование в переписях административных данных в различных странах неодинаково. Такие источники могут быть использованы для расширения или дополнения традиционной переписи, для проведения комбинированной переписи или для проведения полностью регистровой переписи. Наблюдается явная тенденция к расширению использования административных данных в целях переписи. Это было мотивировано теми преимуществами, которые могут принести административные данные, включая сокращение расходов и нагрузки на респондентов, повышение своевременности и периодичности получения результатов, улучшение качества и повышение гибкости для удовлетворения потребностей пользователей (см., например, раздел 4.1 публикации ЕЭК ООН 2018 года). Кроме того, во многих странах изменились условия в плане поддержки и облегчения использования административных данных во всех национальных статистических системах (см. раздел 4.2 публикации ЕЭК ООН 2018 года). Это произошло благодаря изменениям в законодательстве, приемлемости для общественности и заинтересованных сторон, а также благодаря развитию технологий и статистических методологий.

[5] В Руководстве ЕЭК ООН «*Использование административных и вторичных источников данных в официальной статистике*» (ЕЭК ООН 2011, стр. 1-3) обсуждается эволюция понимания того, что означают «административные цели». В руководстве делается вывод о том, что всё более предпочтительным является широкое и всеобъемлющее определение, охватывающее данные частного сектора.

9. Важность административных данных была ещё раз подчёркнута проблемами, с которыми в настоящее время сталкиваются национальные статистические службы (НСС), когда речь идёт о сборе данных непосредственно от населения, будь то из-за нежелания общественности участвовать в переписи или их способности делать это. Эта проблема ярко проявилась в начале кризиса Covid-19 в 2020 году, который существенно затронул как способность общественности взаимодействовать с НСС, так и способность НСС взаимодействовать с общественностью. Различные способы использования административных данных в переписи описаны в Главе 2 настоящего Руководства.

1.3 Ключевые риски для качества

10. При всех преимуществах, которые могут обеспечить административные данные, существует ряд ключевых соображений качества, которые необходимо оценить и изучить, прежде чем включать административный источник в перепись. Во-первых, НСС будет иметь лишь ограниченный контроль над тем, как собираются и обрабатываются данные. Поэтому существует значительная зависимость от органов, являющихся держателями административных данных. Например, если административный орган не в состоянии выполнить требования НСС в отношении предоставления требуемых данных в нужное время, это повлияет на своевременность результатов переписи. Аналогичным образом, если административный орган не будет адекватно взаимодействовать с НСС в отношении любых потенциальных изменений в источнике, это может повлиять на согласованность и сопоставимость.

11. Во-вторых, использование НСС административных данных в целях, отличных от тех, для которых эти данные были изначально собраны, вызывает вопросы конфиденциальности, безопасности и юридические проблемы. Поэтому НСС должно оценивать приемлемость данных для общественности исходя из того, что требуемые гарантии имеются в наличии и что они доводятся до сведения общественности (поставщика данных). Использование также должно быть законным. Без одобрения или согласия как со стороны общественности, так и со стороны поставщика данных, или без внушающей доверие правовой основы для использования административного источника, возникнет значительный риск для репутации НСС и его способности проводить высококачественную перепись. Это может произойти, если общественность изменит своё отношение к взаимодействию с административным органом или НСС, в связи с опасениями в отношении того, как НСС использует их данные.

12. В-третьих, административные данные (в целом) не собираются для статистических целей. Следовательно, источники данных могут использовать концепции, классификации и определения, отличные от тех, которые требуются в целях переписи, они могут относиться к различным отчётным периодам, могут быть подвержены задержкам в обновлении и могут иметь ограниченный охват населения, подлежащего переписи. Кроме того, точность и полнота данных будут в значительной степени зависеть от важности этих данных для функции

административного органа. Административные источники также могут быть подвержены изменениям с течением времени и несоответствиям в способах сбора данных по различным слоям населения. Кроме того, источники данных могут не иметь необходимых идентификаторов или переменных, позволяющих увязать данные, необходимые для переписи.

13. И наконец, сложность административных данных, а также доступность и полнота связанных метаданных будут влиять на способность НСС понимать, получать доступ и использовать административный источник. Например, административные данные могут храниться в больших, сложных структурах данных, что создаёт значительные технические проблемы, которые необходимо учитывать и преодолевать НСС. Сложность административных данных может также влиять на доступность и чёткость измерения качества результатов для пользователя. То есть пользователям переписи может быть трудно понять использование административных данных в переписи и как это использование влияет на качество результатов переписи.

14. Эти ключевые соображения качества послужат основой для принятия решений об использовании административных данных в целях переписи. В Руководстве подробно рассматривается каждое из этих соображений.

1.4 Сфера охвата и структура Руководства

15. Основное внимание в Руководстве уделяется оценке качества источников административных данных для использования в переписи (т.е. качеству вводимых данных). Они не охватывают другие источники как таковые (например, большие данные, коммерческие данные). Тем не менее, большая часть материалов Руководства применима не только к административным данным (руководство по оценке качества больших данных можно найти в ЕЭК ООН 2014).

16. Руководство начинается с предоставления информации о различных методологиях переписи и о том, как административные данные могут использоваться в рамках каждой из этих методологий, включая типы используемых источников данных. Цель состоит в том, чтобы предоставить информацию, которая может быть полезной для НСС, которые хотят включить новые источники административных данных в структуру своих переписей (Глава 2). В этой главе кратко рассматривается влияние пандемии COVID-19 на использование административных данных в переписях населения.

17. В Главе 3 описываются общие рамки качества, на которых основано Руководство. Структура построена вокруг четырёх этапов оценки. Этапы в целом относятся к жизненному циклу использования административных данных в переписи:

    (a) Понимание, оценка и работа над получением источника (этап Источник),

    (b) Получение фактических данных и оценка их качества (этап Данные),

(c) Обработка административных данных для использования в переписи (этап Процесс),

(d) Оценка качества результатов переписи, в которых используются административные данные (этап Итоговые материалы).

18. В главе также описываются параметры качества, оцениваемые на каждом этапе, и связанные с ними ошибки (например, ошибки представления и измерения). Глава завершается кратким описанием важности проведения технико-экономического обоснования использования административных данных с объяснением того, как этапы в рамках Руководства могут быть использованы для этой цели.

19. Глава 4 охватывает первый этап оценки (этап Источник), на котором собирается информация об административном источнике посредством связи с поставщиком данных и путём анализа существующих метаданных. На данном этапе основное внимание уделяется оценке соответствия источника потребностям переписи, охватывающей точность, своевременность, согласованность и сопоставимость, доступность и интерпретация. Также проводится оценка институциональной среды, в том числе того, может ли поставщик данных удовлетворить потребности НСС, с учётом таких факторов, как прочность отношений с поставщиком и статус поставщика.

20. Глава 5 охватывает этап оценки данных, на котором данные поступают от поставщика данных и оцениваются путём анализа данных и сравнения с другими источниками данных. Как на этапе Источника, так и на этапе Данные оценка и измерение качества сопоставляются со многими измерениями качества данных с использованием различных инструментов и показателей. Эти два этапа вместе обеспечивают оценку качества вводимых данных.

21. Информация и понимание, полученные на этапах Источник и Данные, полезны не только для определения того, можно ли использовать конкретный источник в переписи, но и для определения необходимой обработки административных данных для использования в переписи. Как правило, административные данные не могут быть использованы непосредственно в переписи из-за концептуальных различий и различий в определениях. Существуют также ограничения по охвату, полноте и точности. Необходимо преобразовать данные из административных источников (включая регистры), используя информацию, полученную на этапах Источник и Данные. Некоторые из наиболее важных процессов и связанные с ними соображения качества рассматриваются в Главе 6 Руководства.

22. Этапы Источник, Данные и Процесс напрямую связаны с качеством результатов переписи в соответствии с показателями качества результатов Европейской статистической системы. И наоборот, оценка результатов переписи даст ценную информацию о том, где могут быть ограничения или проблемы, связанные с административными данными или обработкой этих данных, которые не были выявлены изначально на этапах Источник, Данные и Процесс. Существует итеративный процесс оценки, который может информировать как о текущих улучшениях административных источников (работа с поставщиком данных для улучшения источника), так и об улучшениях в обработке административных

данных НСС. Оценка качества результатов переписи, в которых используются административные данные, кратко рассматривается в Главе 7 (Этап Итоговые материалы).

23. В Руководстве приводятся различные примеры и тематические исследования по конкретным странам с использованием базовых иллюстраций или более подробных тематических исследований. Главы иллюстрируют применение этапов контроля качества на практике.

24. Наконец, в Главе 8 Выводы и рекомендации содержится краткое изложение рекомендаций, представленных в предыдущих главах. В заключительной главе предлагается дальнейшая скоординированная на международном уровне работа по обеспечению качества административных данных.

Глава 2 Методологии переписи и использование административных данных для проведения переписей населения

25. В этой главе кратко излагается спектр методов переписи и видов использования административных данных в переписях, которые являются общими для всех стран ЕЭК ООН. Это поможет НСС в регионе ЕЭК ООН и за его пределами при использовании административных данных в своих переписях независимо от принятой методологии сбора данных.

2.1 Методологии проведения переписи

26. Как отмечалось в предыдущих публикациях ЕЭК ООН (см. ЕЭК ООН 2015; ЕЭК ООН 2018), существует несколько различных способов осуществления процесса сбора данных при переписи населения и жилищного фонда. В этом разделе представлен обзор методов проведения переписи и тех случаев, когда это Руководство может быть полезно НСС.

27. Для упрощения в настоящем документе кратко описываются только три основных метода сбора данных о населении:

    (a) Традиционная перепись,

    (b) Перепись на основе регистров, и

    (c) Комбинированная перепись.

28. Вики-сайт переписи ЕЭК ООН [6], на котором собрана информация о раунде переписей 2020 года, представленная странами-членами, указывает на то, что тенденция отхода от традиционной переписи продолжается. Из 52 стран ЕЭК ООН, по которым имеется информация, менее половины (23) проводят традиционную перепись в текущем раунде переписи 2020 года, при этом 13 стран планируют провести перепись на основе регистров, а 16 стран — использовать комбинированный подход. Тем не менее, как обсуждается ниже, у НСС, которые проводят традиционную перепись, все ещё есть возможности и преимущества для использования административных данных.

29. Ниже кратко излагаются ключевые особенности трёх упомянутых методов переписи. Более подробно различные методы проведения переписи, включая необходимые преобладающие условия, преимущества и проблемы, излагаются в *Рекомендациях КЕС для переписей населения и жилищного фонда 2020 года* (см. ЕЭК ООН 2015). Подробное описание основных особенностей переписи и как они соотносятся с различными методологиями переписи, см. в Главе 3 ЕЭК ООН (2018).

[6] Доступно на bit.ly/UNECECensusWiki2020

2.1.1 Традиционная перепись

30. Термин «традиционная перепись» в самом широком смысле относится к переписи, основанной на прямом подсчёте всех лиц, домашних хозяйств и жилых единиц и сборе информации об их характеристиках путём заполнения переписных листов на бумаге либо в электронном виде. Данные собираются на местах путём полного учёта по всей стране в течение сравнительно короткого промежутка времени.

31. Информация может быть собрана одним или несколькими из следующих методов:

    (a) Переписные листы заполняются непосредственно самими домохозяйствами (доставка и сбор анкет осуществляется счётчиками, почтовой службой или иными способами),

    (b) В режиме онлайн с использованием электронных вопросников,

    (c) Счётчиками во время личного опроса домохозяйства с использованием бумажных или электронных вопросников.

32. С 2001 года некоторые страны внесли существенные изменения в свои процессы по сбору данных, все ещё подпадая под определение традиционной схемы. Например, в Соединённых Штатах Америки Бюро переписи населения сосредоточено на сборе только кратких данных (10 вопросов, в основном демографических, в сочетании с тремя жилищными вопросами) в полном перечне вопросов переписи, проводимой раз в десять лет (с контрольной датой Дня переписи, 1 апреля). Крупное выборочное обследование домашних хозяйств, Исследование американского сообщества, ежемесячно собирает более подробные данные (демографические, социальные, экономические и жилищные). Новые данные из Исследования американского сообщества публикуются ежегодно в течение десятилетия, заменяя необходимость в "длинной форме переписи", которая ранее рассылалась выборочной совокупности населения.

33. В отличие от этого, Национальный институт статистики и экономических исследований Франции придерживается иного подхода. Скользящая перепись проводится с помощью совокупного непрерывного выборочного обследования, охватывающего всю страну в течение всего десятилетнего периода, а не путём переписи, проводимой одновременно во всех районах на конкретную контрольную дату. Французская скользящая перепись также проводится в крупных муниципалитетах (более 1000 жителей) на основе исчерпывающего реестра жилых помещений. Этот реестр обновляется с использованием административных данных (разрешений на строительство) и проверок муниципалитетами. Файлы налоговых данных также используются для оценки 40 процентов небольших муниципалитетов (менее 10 000 жителей) каждый год. Ежегодное обследование, подобное тому, которое проводится во Франции, может проводиться в течение года, в определённый месяц или в более короткие сроки. При таком подходе можно построить следующую структуру:

    (a) Национальные результаты с помощью одного ежегодного обследования,

(b) Региональные результаты путём обобщения данных нескольких последовательных ежегодных обследований,

(c) Результаты на небольших территориях путём обобщения данных за более значительное число лет.

2.1.2 Перепись на основе регистров

34. Перепись на основе регистров представляет собой совершенно иной подход, первоначально разработанный странами Северной Европы в 1970-х годах, среди которых Дания первой в 1981 году провела перепись, полностью основанную на регистрах. При таком подходе отсутствует прямой сбор данных от населения, а традиционная регистрация заменяется административными данными из различных регистров (регистр населения, регистр зданий или адресов, регистр социального страхования, налоговые записи и т. д.) посредством процесса сопоставления, обычно с использованием личных идентификационных номеров (PIN-кодов). При наличии качественной системы статистических регистров такой подход позволяет получать данные переписи при значительно меньших затратах и с гораздо меньшими человеческими усилиями.

35. Этот методологический подход явно требует максимального использования административных источников и, следовательно, в значительной степени зависит от установления и обеспечения наивысшего уровня качества данных из таких источников.

2.1.3 Комбинированная перепись

36. С 1990-х годов несколько стран региона ЕЭК ООН и другие страны разработали инновационные методы проведения переписи, сочетающие использование административных данных со сбором, часто сокращённого, набора данных в результате учёта населения на местах. Учёт на местах все ещё может быть основным методом сбора данных переписи. Однако административные данные используются там, где это возможно, для уменьшения нагрузки на респондентов и добавления дополнительной информации, не собранной в ходе переписи (например, вопросы, связанные с доходом). Цель учёта населения при опросе направлена на получение конкретных переменных, для которых соответствующие данные недоступны из какого-либо административного источника. При таком комбинированном подходе сбор полевых данных на местах может носить как сплошной, так и выборочный характер.

37. Этот методологический подход был использован некоторыми НСС при переходе от традиционной переписи к переписи, полностью основанной на регистрах. Это Руководство было написано в первую очередь для того, чтобы помочь производителям статистических данных в ходе такого перехода или при проведении комбинированной переписи или переписи на основе регистров. Тем не менее, они также будут способствовать оценке административных данных, используемых в основном в рамках традиционной переписи.

2.1.4 Влияние пандемии COVID-19 на методологию переписи

38. Всё большее число стран КЕС используют при переписях населения раунда 2020 года методологию комбинированных или полностью основанных на регистрах переписей населения. В ходе консультаций в рамках КЕС по проекту этого Руководства многие страны указали, что в результате пандемии этот сдвиг ускорился и/или что они всё шире используют административные данные для поддержки традиционных переписей населения. Основные последствия пандемии для использования административных данных в переписях, выявленные странами в ходе консультаций, кратко изложены во Вставке 1.

Вставка 1: Влияние пандемии COVID-19 на переписи населения и административные источники

Пандемия COVID-19 оказала сильное влияние на переписи населения во всем мире и на использование административных данных (см., например, ЕЭК ООН 2021). Это повлияло на то, как проводились мероприятия по сбору данных переписи, активизирован или ускорен переход к использованию административных данных (особенно там, где сбор данных на местах невозможен) и даже задержало проведение переписи во многих странах. Кризис также продемонстрировал необходимость более частой и своевременной статистики о населении во времена беспрецедентных изменений, поскольку лица, принимающие решения, ищут информацию о том, где и как люди живут, учатся и работают, а также о состоянии здоровья и смертности.

Там, где был ускорен переход к более широкому использованию административных данных, это потребовало быстрых изменений и улучшений в системах сбора и обработки данных как в административных органах, так и в НСС. В частности, потребовались новые процедуры, протоколы и даже законодательство для облегчения сбора, обмена и использования данных. Это потребовало эффективного сотрудничества между административными органами и НСС (см. Раздел 4.1.5).

Для стран, которые отложили перепись после 2021 года, административные данные будут важны для поддержки производства статистики переписи за предыдущие учётные годы. Например, чтобы выполнить требования Евростата и Германия, и Венгрия, которые отложили свою перепись до 2022 года, будут использовать административные источники для получения статистики за отчётный 2021 год на основе своей национальной переписи 2022 года.

Пандемия также оказала значительное влияние на качество и содержание административных источников. Например, некоторые виды взаимодействия со службами здравоохранения могут снизиться (люди избегают медицинских услуг из-за опасений заразиться вирусом), в то время как другие увеличились (люди регистрируются для тестирования, лечения или вакцинации). Эти изменения повлияют на охват регистров здоровья.

Кроме того, необходимость предоставления новых услуг и поддержки общественности привела к разработке новых административных процессов и систем. В Великобритании, например, это включало системы для поддержки тех, кто остался без работы из-за пандемии (посредством увольнения), и для поддержки развёртывания программы тестирования, отслеживания и вакцинации (предоставляя новые источники данных).

Наконец, давление на общественные и административные органы из-за пандемии повлияло на своевременность и точность административных данных. Примеры включают задержки в регистрации рождений и снижение уровня гарантии качества (с временным отвлечением ресурсов в другое место).

2.2 Использование административных данных

39. Степень использования данных из административных источников для целей проведения переписи населения и жилищного фонда явно будут зависеть от типа методологии, используемой при сборе данных.

40. В рамках различных типов методологии переписи, описанных выше, административные данные могут использоваться различными способами. Среди них ключевыми являются следующие варианты использования:

    (a) Построение и оптимизация основы выборки переписи и полевых операций (как принято в США и Канаде),

    (b) Обеспечение качества оценок переписи путём сравнения с административными источниками и внесение корректировок посредством, например, редактирования и импутации (как принято в Эстонии для переписи 2011 года),

    (c) Получение существующих переменных переписи и добавление новых переменных переписи (как принято в Соединённом Королевстве),

    (d) При построении статистических регистров населения [7] и непосредственном использовании для переписи административных данных (как это было в Испании и Новой Зеландии, соответственно), и

    (e) При переписи полностью основанной на административных данных (например, в Нидерландах).

2.2.1 Построение и оптимизация основы выборки переписи и полевых операций

41. Первый вариант применения связан с использованием административных данных для построения и оптимизации выборки адресов жилых зданий (помещений) для полевых операций переписи. Это включает оценку качества основы выборки переписи, построенной на основе административных данных. Сценарий использования также устанавливает, могут ли административные данные определить, будет ли адрес, вероятно, занят и кем, или может ли определённый адрес быть «труднодоступным» [8], тем самым оптимизируя полевые операции переписи (см. Вставку 4 и раздел 6.3).

42. Для тех стран, в которых сохраняется какой-либо элемент сплошного учёта — либо при полностью традиционной переписи, либо в тех странах, где применяется комбинированный подход, — данные из административных источников могут использоваться для поддержки полевых операций. Многие такие страны могут, например, использовать информацию из адресных регистров или регистров зданий

[7] См. Определение статистического регистра в глоссарии.
[8] См. Определение термина «труднодоступный» в глоссарии.

для создания счётных участков одинакового размера, которые содержат примерно одинаковое количество домашних хозяйств или жилищ.

43. В качестве альтернативы, такая информация может использоваться для формирования соответствующих выборок домохозяйств или жилых единиц, когда полный набор данных не собран от всего населения.

44. Качество переписи на основе административных данных выиграет от оценки источников данных на этапах Источник, Данные и Процесс, предложенных в настоящем Руководстве. Однако, учитывая повторяющийся характер полевых операций (т.е. совокупность переписи улучшается на протяжении всего сбора), такая оценка может подчеркнуть аспекты охвата (связанные с актуальностью) по сравнению с измерением точности.

2.2.2 Замена и/или добавление новых переменных переписи

45. Второй вариант использования связан с оценкой качества административных данных, используемых для замены и добавления новых переменных в перепись.

46. Когда страны решают уменьшить размер (и связанные с этим расходы) сплошной регистрации путём применения комбинированного подхода к переписи, данные из соответствующих административных источников могут использоваться для замены информации, собранной из вопросника домохозяйства. Например, достоверно точная информация о брачном состоянии и статусе занятости или годе иммиграции может быть легко доступна из административных регистров, что устраняет необходимость сбора таких данных непосредственно от физических лиц.

47. В качестве альтернативы, пользователи могут привести веские аргументы в пользу того, что НСС собирает в ходе переписи информацию, которая либо оказалась деликатного характера, либо требует такого уровня детализации, что многие люди могут быть не в состоянии точно сообщить об этом в традиционном вопроснике переписи. Например, информация, касающаяся младенческой смертности, может быть основана на культурных особенностях в некоторых странах, в то время как данные о доходах домохозяйства часто могут требовать обмена потенциально конфиденциальной информацией между другими членами домохозяйства. В этих случаях эквивалентные данные, относящиеся к связанному физическому лицу, могут быть получены из административных источников (таких как записи актов гражданского состояния или налоговые отчёты).

48. Оценка качества административных данных на этапе Источника может помочь в принятии решения о выборе административных источников для использования в таких случаях. Кроме того, оценка выбранного источника (источников) на этапах обработки Данных и Процесса может в конечном итоге обеспечить качество Итоговых материалов.

2.2.3 Создание статистических регистров и прямое использование административных данных

49. Третий вариант использования относится к административным источникам для подсчёта населения (см. также раздел 6.3). Все население может быть подсчитано с помощью административного списка (например, реестра населения) или административные данные могут использоваться для подсчёта части населения, например, тех, кто были пропущены при опросе населения на местах [9]. Различаются ситуации, когда НСС могут полагаться на уникальные идентификаторы для интеграции нескольких источников в один реестр или когда идентификаторов не существует (и когда они полагаются на детерминированные / вероятностные методы для персональных решений и привязки источников с такими переменными, как имя, дата рождения, адрес и т. д).

50. Организация Объединённых Наций (2014) отметила, что регистры населения в настоящее время хорошо налажены в ряде страна, особенно в регионе ЕЭК ООН, где они эффективно используются в качестве источника статистических данных на протяжении десятилетий. Регистры можно рассматривать как логический продукт развития системы статистики естественного движения населения. Они стали важным источником информации для различных статистических обследований и, во многих случаях, для переписей населения и жилищного фонда.

51. Основные характеристики, которые могут быть включены в регистр населения, — это дата и место рождения, пол, дата и место смерти, дата прибытия/выбытия, гражданство и семейное положение. Более того, если регистры населения будут полными, они могут предоставить данные как о внутренней, так и о международной миграции через смену места жительства, а также о международных прибытиях и выбытиях. Регистры могут использоваться в качестве непосредственной основы для «административного перечисления», чтобы заменить традиционную операцию сбора данных на местах.

52. Как и в предыдущем варианте использования, оценка качества исходных данных на этапах Источники и Данные будет иметь важное значение при разработке методологии построения статистических регистров населения. В конечном итоге это будет итеративный процесс проектирования, при котором контроль качества на этапе Итоговые материалы может выявить проблемы, которые необходимо решить на более ранних этапах. Предполагается, что при создании регистров НСС должны следовать всем этапам обеспечения качества, предложенным в этом Руководстве.

[9] В тематическом исследовании 6.7.4 в Италии приведён пример использования административных данных для корректировки неполного охвата обследования в рамках сложной системы оценки с использованием различных административных источников и обследований.

2.2.4 Оценка качества и корректировки

53. Четвёртый вариант использования относится к качеству источника данных, который будет использоваться для улучшения существующих переменных переписи. В этом случае административные данные используются для редактирования и условного исчисления существующей переменной переписи, в отличие от прямой / полной замены традиционного набора данных.

54. Даже в тех странах, которые продолжают проводить традиционную перепись, данные из административных источников могут использоваться либо для обеспечения качества информации, полученной от домашних хозяйств, либо для корректировки таких данных, когда можно показать, что имеются ошибки или упущения в собранных на местах данных.

55. Более того, если данные, представленные в традиционной переписи, содержат ошибки по существу или пропуски, неправильные ответы могут быть отредактированы и/или недостающие ответы могут быть рассчитаны с использованием либо информации, полученной в самой переписи от аналогичных домохозяйств или данных, относящихся к рассматриваемой переменной и индивидууму в соответствующем регистре.

56. При использовании административных данных для оценки качества данных переписи, собранных на местах, ключевыми являются этапы Источник, Данные и Процесс. Кроме того, несмотря на то, что это выходит за рамки настоящего Руководства, важно учитывать вопросы цикличности в отношении общей структуры переписи. Например, если источник административных данных использовался для восполнения отсутствующих значений в данных переписи или замены переменной переписи, его также не следует использовать в оценке качества.

2.2.5 Полная перепись на основе регистров

57. Наконец, последний вариант использования касается измерения качества источников, когда вся перепись проводится на основе административного регистра населения, а не по традиционной методологии переписи.

58. Очевидно, что наиболее широкое использование данных из административных источников, по определению, происходит тогда, когда НСС проводят перепись полностью на основе регистров. В контексте полной переписи на основе регистров оценка качества на каждом из предложенных этапов имеет чрезвычайно важное значение.

59. Качество результатов переписи особенно зависит от постоянного улучшения качества на этапах Источник, Данные и Процесс. В зависимости от возможности надлежащей увязки с другими регистрами, к одной записи может быть добавлено много дополнительной информации, хотя и не записанной в самом регистре населения, такой как, язык (языки), этническая принадлежность, уровень образования, статус активности и род занятий. В странах, где проводятся переписи

на основе регистров, качество и стабильность основных административных источников на ранних этапах таковы, что результаты переписи на основе регистров считаются «золотым стандартом». Сбор данных переписи таким способом, однако, не препятствует проведению НСС постпереписного обследования на местах в качестве средства независимой оценки качества охвата или содержания показателей в итоговой базе данных переписи.

2.3 Типы административных источников

60. Как было отмечено КЕС (см. ЕЭК ООН 2015), разработка системы переписи населения на основе регистров (будь то в контексте подхода полностью основанного на регистрах или комбинированного подхода) является длительным процессом, который может потребовать многих лет. Многие страны продолжат использовать элементы традиционного сбора данных, даже когда они начнут использовать административные регистры в качестве альтернативного источника данных.

61. В этом разделе Руководства кратко рассматриваются некоторые типы административных источников, из которых данные чаще всего используются НСС для целей переписи, а также способы использования данных каждого из них. В соответствующих случаях эти виды использования относятся к темам, которые в настоящее время рекомендованы КЕС для включения в перепись (см. ЕЭК ООН 2015).

2.3.1 Использование административных источников для поддержки традиционной переписи

62. Частота, с которой НСС используют данные из административных источников, постепенно увеличивается от переписи к переписи по мере перехода от традиционной сплошной регистрации с использованием комбинированного подхода к полной переписи на основе регистров. Но даже те страны, которые продолжают применять традиционную перепись, вероятно, будут все больше использовать административные данные для поддержки своих переписных работ.

63. В настоящее время НСС обычно используют *адресные регистры* для составления списков жилых помещений и домохозяйств. Регистры могут создавать и отображать счётные участки, что обеспечивает сбалансированную рабочую нагрузку для счётчиков, или обеспечивать стратифицированные схемы выборки для постпереписных или других выборочных обследований. Создание специально созданного списка адресов НСС может включать объединение данных из нескольких отдельных и независимых регистров (которые, возможно, созданы для различных административных целей), чтобы свести к минимуму неполный или чрезмерный учёт.

64. Например, списки зарегистрированных избирателей, используемые для целей голосования на национальном и местном уровнях, или списки жилых домов,

используемые местными властями для оценки стоимости, могут не включать все почтовые адреса, используемые национальными или коммерческими почтовыми перевозчиками. Более того, здания, идентифицированные национальным картографическим агентством для целей создания точных крупномасштабных официальных карт, могут определять местоположение адресов, которые не используются в жилых целях и часто исключаются из адресной базы данных переписи.

65. Те НСС, которые проводят традиционную перепись, могут использовать данные из административных источников для оценки качества данных, собранных в вопроснике для домохозяйства. Например, данные *национальной системы регистрации актов гражданского состояния* могут предоставить точную информацию о количестве рождений и смертей в течение последовательных 12-месячных периодов до переписи. Эти данные о возрасте детей младшего возраста можно сравнить и сопоставить с данными переписи. Аналогичным образом, данные об изменении адреса, которые необходимо сообщать местным властям в целях ведения регистров населения, могут использоваться для проверки информации о миграции с момента предыдущей переписи.

66. Однако следует отметить, что в тех случаях, когда данные используются для *оценки качества* информации, представленной в вопроснике переписи, и для *дополнения* данных переписи для учёта отсутствующих или неправильных ответов, можно считать, что перепись прошла путь от традиционной методологии к методологии комбинированного подхода.

2.3.2 Использование административных источников для получения характеристик населения при переписи

67. Одним из способов использования административных источников для переписей является предоставление данных для получения требуемых выходных переменных без необходимости сбора соответствующей информации непосредственно от общественности. Тип, структура и содержание таких административных источников, конечно, будут варьироваться от страны к стране в зависимости от административных целей, для которых данные используются поставщиками данных. Ниже приводится краткое описание наиболее распространённых универсальных типов регистров, используемых для этой цели.

68. *Регистры населения* — это регистры (часто ведутся национальным правительственным ведомством и/или соответствующими местными органами власти, отвечающими за вопросы внутренней безопасности), которые предоставляют информацию о лицах, обычно проживающих в стране. Эти регистры обычно ведутся для выполнения юридического требования, согласно которому как граждане, так и иностранцы, проживающие в стране, должны зарегистрироваться в местных органах власти. Агрегирование этих местных регистраций приводит к учёту численности населения и перемещений населения на национальном и местном уровнях. Кроме того, они часто фиксируют информацию о некоторых характеристиках отдельных лиц, из которых могут быть получены данные по нескольким основным темам переписи, например дата и

место рождения, пол, дата прибытия / выбытия, гражданство и семейное положение для каждого постоянного жителя с разбивкой по месту обычного проживания (как бы это ни было определено).

69. *Регистры социального обеспечения* — это регистры, которые ведутся официальными органами, как правило, для целей управления национальными программами социального страхования и распределения пенсий и пособий (т.е. безработных, семей, пенсионеров, инвалидов и хронических больных). Данные из таких регистров могут использоваться для получения показателей переписи по таким признакам, как пол, возраст, семейное положение, статус безработного, доход и инвалидность / состояние здоровья.

70. *Налоговые регистры* — это регистры, которые ведутся национальными и местными налоговыми органами для целей администрирования и сбора подоходного налога, налогов на покупку, строительных ставок и других национальных и местных налогов. Данные из таких регистров могут использоваться в первую очередь для получения данных переписи о доходах физических лиц или домохозяйства, которые в противном случае было бы сложно или слишком деликатно собрать непосредственно с помощью вопросника для домохозяйства. Другая информация, содержащаяся в таких регистрах, может включать пол, возраст, семейное положение, статус занятости, род занятий, место работы и место обычного проживания.

71. *Регистры занятости* — это регистры, из которых берутся официальные данные о занятости и безработице в стране. Зарегистрированные данные могут позволить НСС получать данные переписи, относящиеся к ключевым социально-экономическим показателям экономической деятельности, статусу занятости, роду занятий, рабочего времени и места работы (последние два позволяют анализировать схемы поездок на работу).

72. *Регистры предприятий* (бизнес-регистры) — содержат информацию, необходимую для предоставления целого ряда услуг, которые могут варьироваться в зависимости от страны. В основном их целью является регистрирование, мониторинг и хранение корпоративной информации, такой как правовой статус компании, её головной офис, капитал и юридические представители. НСС может использовать эту информацию для получения данных переписи по экономическим темам, в частности по промышленности.

73. *Регистры образования* — ведутся как централизованно, так и отдельными образовательными и академическими учреждениями с целью регистрации приёма и успеваемости студентов, а также трудоустройства преподавательского состава. Имеющиеся данные могут использоваться НСС для получения статистических данных переписи населения об обучении, посещении учебных заведений, грамотности и самом высоком уровне образования, хотя следует признать, что такие данные могут часто относиться только к текущему контингенту учащихся. Данные о лицах, прекративших обучение в образовательном учреждении, должны быть получены из других источников.

74. *Регистры здоровья* — ведутся местными органами здравоохранения для целей оказания медицинских услуг, независимо от того, относятся ли они к национальной службе здравоохранения или предоставляются частными агентствами, основанными на страховании. Исходная информация, которую они содержат, обычно рассматривается как конфиденциальная, но может быть анонимной в достаточной степени, чтобы позволить НСС использовать её для создания данных о состоянии здоровья, географическом районе, уровне инвалидности и паритете.

75. *Регистры зданий и жилых помещений* — это регистры, которые обычно ведутся агентствами по оценке земли и имущества, а также местными органами власти, ответственными за разработку жилищной политики и городское планирование. Они могут включать информацию, касающуюся собственности, размера и физической конструкции отдельных жилых единиц, но не обязательно могут относиться к лицам, проживавших в них. Имеющиеся данные могут позволить НСС получать данные для создания статистических данных переписи, соответствующих потребностям переписи жилищного фонда, таких как тип жилья, площадь жилого помещения, уровень пола, материалы наружных стен и период постройки, а также могут различать жилые и нежилые дома.

76. НСС могут также иметь доступ к данным из других административных источников для предоставления тематически ориентированных результатов переписи. Например:

    (a) *Реестры автотранспортных средств* могут позволять собирать данные о наличии автомобилей,

    (b) *Реестры иностранных граждан* могут предоставлять информацию о мигрантах, году въезда в страну, гражданстве и лицах, ищущих убежища,

    (c) *Списки военнослужащих* (при условии разрешённого доступа для НСС) могут указывать занятость в вооружённых силах,

    (d) *Тюремные реестры* могут предоставить некоторую базовую информацию о членах специальной группы населения, отбывающей наказание, которую особенно трудно зарегистрировать при традиционной переписи, и

    (e) *Реестры поставщиков услуг общественного пользования* могут содержать информацию о наличии бытовых удобств, таких как водопровод, электричество и/или газ, а также канализация и средства удаления отходов.

Глава 3 Система обеспечения качества

77. Качество статистических данных зависит от того, соответствуют ли статистические материалы требованиям своего предполагаемого использования. Например, определение качества Европейской статистической системы разработано на основе семейства стандартов ISO 9000 «Качество — степень, в которой ряд присущих объекту характеристик удовлетворяют требованиям» (см. ISO, 2015). В официальной статистике объект может включать «статистический продукт, услугу, процесс, систему, методологию, организацию, ресурс или ввод данных» (см. Eurostat, 2020, p. 17). В контексте переписи качество используемых административных данных должно рассматриваться с учётом того, как данные собираются и обрабатываются поставщиками и НСС, вплоть до итоговых материалов переписи.

78. На протяжении всего процесса могут возникать ошибки, снижающие качество. Здесь под ошибкой понимается разница между окончательной оценкой и истинным параметром совокупности, который она представляет. Это подчёркивается в Общей модели статистических бизнес-процессов, которая обеспечивает стандартную структуру для описания большинства статистических процессов и включает «качество» как аспект, охватывающий все его этапы (см. Eurostat ESSnet MIAD 2014). Кроме того, Lothian et al. (2019) аргументировали необходимость понимания всего процесса статистического производства при работе с альтернативными источниками данных, такими как административные данные. Оценка качества административных источников требует отображения ошибок, которые могут возникнуть до и после предоставления данных в НСС, и определения того, как любые такие ошибки могут быть устранены (например, путём внесения изменений в сбор, обработку и/или интеграцию с другими источниками). В этом Руководстве определены четыре основных этапа проведения переписи: Источник, Данные, Процесс и Итоговые материалы. Затем они описывают, как можно оценить качество административных данных, определяя ключевые параметры качества на каждом этапе и соответствующие инструменты и индикаторы для контроля качества.

79. Этот подход основан не только на Общей модели статистических бизнес-процессов, но и на Daas et al. (2009), которые определили сквозные области, которые касаются качества или «представлений» на качество, которые они называют «гиперизмерениями», в отношении источника, метаданных и данных (см. Daas et al. 2009, p. 3). Каждое из этих представлений включает несколько параметров качества данных, оцениваемых с помощью показателей качества. В соответствии с этим подходом настоящее Руководство также определяет параметры качества, показатели и методы, используемые при оценке административных данных, с особым акцентом на переписи. В то же время основное внимание уделяется этапам подготовки переписи, которые были бы более понятны для НСС, для которых было написано это Руководство. Сосредоточение внимания на этапах производства подчёркивает, что качество является неотъемлемой частью статистического проектирования и позволяет НСС

сосредоточить внимание на той части (частях) Руководства, которое наиболее актуально для их варианта использования и/или текущего этапа производства.

3.1 Качество и погрешность в рамках переписи

80. В тех случаях, когда официальная статистика формируется с использованием методики выборочного обследования, вопросы анкеты обследования разрабатываются и тестируются таким образом, чтобы уменьшить погрешности измерения, обеспечивая тем самым максимальную точность и надёжность. Таким образом, предполагается, что погрешность полученных оценок вызвана недостаточной выборкой и обычно измеряется и сообщается с использованием среднеквадратической ошибки и/или через доверительные интервалы. Однако такие измерения не фиксируют не связанных с выборкой ошибок, которые особенно важны в контексте переписей, когда цель состоит в том, чтобы охватить всю совокупность населения. Таким образом в более общем плане, как и в случае статистики, сформированной на основе административных данных, основными источниками ошибок в контексте переписей являются не ошибки выборки, а ошибки репрезентативности (охвата) и измерения (см. Zhang, 2012). Общепринятой практикой является корректировка оценок переписи на основе результатов постпереписного обследования (см. Глава 7).

81. В тех случаях, когда в переписях используются административные данные и другие альтернативные источники данных, такие как Большие данные, диапазон возможных ошибок является больше, чем при проведении традиционных переписей, поскольку процессы сбора данных не контролируются НСС. В работе Zhang (2012), опирающейся на работу Groves et al., 2004, проводится различие между двумя широкими типами ошибок в статистике, сформированной с использованием административных данных: ошибки измерения и ошибки репрезентативности. К первому типу относятся ошибки в измерении характеристик (например, возраста, пола и т. д.), а ко второму — ошибки в охвате единиц населения или объектов (например, отдельных лиц или домохозяйств в ходе переписи) [10]. Zhang также проводит различие между качеством отдельных источников, обеспечиваемым поставщиками данных, и качеством преобразованных и/или интегрированных источников, после обработки НСС. Этот подход отражён в Руководстве, которое оценивают качество отдельных

[10] На основе работы Zhang (2012). В случае вводных данных ошибки измерения обусловлены различиями между предоставленными и целевыми характеристиками (например, пол, возраст, этническая принадлежность, род занятий и т. д.) и включают несколько типов ошибок в переменных, включая релевантность (несовпадение определений), связывание (ошибки в переклассифицированных показателях из-за плохой эквивалентности между предоставленными и целевыми классификациями, которые, следовательно, могут потребовать корректировок, например, методом импутации) и ошибки сопоставимости (ошибки между переклассифицированными и скорректированными показателями). Ошибки репрезентативности связаны с разницей между предоставленными и целевыми единицами. К ним относятся ошибки, связанные с избыточным и недостаточным охватом (несогласованность с целевой совокупностью), идентификацией (ошибки в классификации единицы из-за расхождений между несколькими источниками) и ошибками единиц (ошибки при статистическом создании представляющих интерес статистических единиц в случаях, когда они отсутствуют в каком-либо доступном источнике данных).

административных источников (см. этапы Источник и Данные ниже) и интегрированных источников (см. этапы Процесс и Итоговые материалы) с особым акцентом на выявление ошибок измерения и представления.

82. Кроме того, для оценки качества административных данных был также адаптирован подход Общей погрешности наблюдения. В отличие от среднеквадратической ошибки, Общая погрешность наблюдения позволяет выявить более широкий спектр ошибок, включая ошибки достоверности, кадра/охвата, отсутствия ответов, измерения, обработки и моделирования. Таким образом, подходы Общей погрешности наблюдения стремятся отразить то, как различные ошибки накапливаются в результате статистического проектирования и методологии, позволяя определить окончательную погрешность любой заданной оценки. Этот подход был адаптирован для сообщения качества статистики, опирающейся на интеграцию административных данных (см. например, Reid, Zabala and Holmberg, 2017, Rogers and Blackwell, 2020). В то же время качество статистических данных нельзя сводить только к оценке погрешности. В случае интеграции данных из административного источника в проект переписи следует оценивать влияние такой интеграции на качество с точки зрения того, в какой степени она добавляет погрешность или неопределённость в конечные результаты, в сопоставлении с преимуществами интеграции, например, уменьшением нагрузки на респондентов по предоставлению ответов, повышением своевременности, снижением затрат. Исходя из этого в настоящем Руководстве определены дополнительные параметры, которые могут влиять на общее качество результатов переписи, включая институциональную среду и необходимость балансировки параметров качества для удовлетворения потребностей пользователей.

83. Следование принципам настоящего Руководства поможет обеспечить, чтобы оценки переписей основывались на наиболее подходящих источниках и методах и не вводили в заблуждение. В то же время следует также рассмотреть вопрос о том, как предполагается использовать административные источники при планировании переписи (см. Глава 2). Учитывая разнообразие возможных видов использования, эти рамочные принципы следует использовать гибко и адаптировать к уровню качества, требуемому для различных видов использования административных данных НСС и для удовлетворения различных статистических потребностей пользователей данных переписи, включая общественность, организации, местные и национальные органы власти. Поэтому оценка качества неизбежно опирается на квалифицированное профессиональное суждение на протяжении всего процесса статистического производства, от сбора до публикации, для удовлетворения потребностей пользователей данных.

3.2 Измерение качества

84. Качество оценок переписи, полученных с использованием административных источников, особенно трудно оценить и/или измерить из-за сложности и многомерности используемых данных. Как отмечалось выше, многие факторы, влияющие на качество, не поддаются количественному измерению. Боле того, это

то, что представляет собой «соответствие целям», и высококачественная статистика неизменно будет варьироваться от одного пользователя к другому, например, некоторые пользователи могут отдавать предпочтение своевременности, а не точности. Поэтому важно оценивать / измерять качество административных данных по всем ключевым параметрам, которые будут представлять интерес для производителей и пользователей статистических данных. Что подразумевается под оценкой и измерением, нуждается в дополнительном уточнении.

85. В настоящем Руководстве проводится различие между оценкой качества, т. е. качественной оценкой, и измерением качества, т. е. присвоением количественного параметра данной оценке качества. В тех случаях, когда невозможно разработать показатели для количественной оценки или они ещё не разработаны, рекомендуется проводить качественную оценку их влияния на качество. В дополнение к ним существует ряд дополнительных принципов, служащих руководством для формирования официальной статистики (см. ЕЭК ООН, 1992) и которые применимы на протяжении всего статистического процесса и в более широком контексте НСС (например, приверженность принципам качества, независимости, защиты данных, конфиденциальности статистических данных и т. д.). Эти вопросы актуальны для всех статистических процессов и не в полной мере охвачены настоящими Руководством. Однако следует признать, что перепись, в которой используются административные источники, опирается на данные, которые были сформированы за пределами статистической системы, в другой организации, над которой НСС, как правило, не имеет контроля[11]. По этой причине необходимо тщательно рассмотреть влияние использования внешних источников на эти принципы.

3.3 Этапы оценки качества

86. Для обеспечения простоты выполнения этого Руководства контроль качества административных источников рассматривается на четырёх широких этапах жизненного цикла переписи. Они применимы независимо от типа переписи (см. Глава 2). Хотя разработка статистического плана никогда не бывает полностью линейной, аналитическое обоснование того, как проводить оценку качества, позволит производителям статистики оперативно определить ключевые параметры качества, наиболее значимые для их собственного контекста. Этими этапами являются:

(a) **Этап Источник**: основанный на метаданных контроль качества новых или повторно предоставленных административных источников, которые будут использоваться в переписи. Этот этап не требует от НСС обладания

[11] В некоторых случаях НСС имеют некоторый контроль над реестром. В Швейцарии, например, Федеральный регистр зданий и жилых помещений или Регистр предприятий являются частью Федерального статистического управления. В долгосрочной перспективе может оказаться целесообразным интегрировать определённые подходящие реестры в НСС. Последствия / преимущества этого кратко обсуждаются в Разделе 4.2.5.

фактическими данными, но он имеет решающее значение для последующих этапов;

(b) **Этап Данные:** Контроль качества первичных административных данных, предоставляемых НСС административными органами. Это потребует от НСС проверки предоставленных данных на основе выводов этапа Источник. Помимо базовой проверки, этот этап включает в себя любую обработку, необходимую для установления качества предоставленных данных по сравнению с тем, что ожидалось, а также сопоставления с альтернативными источниками;

(c) **Этап Процесс:** Процессы, часто осуществляемые в отношении данных административных источников, в ходе переписей с целью преобразования данных для использования в рамках переписи и/или повышения качества. Выявленные процессы включают:

(i) Связь записей

(ii) Статистические регистры и методология «признаков жизни»;

(iii) Регистрация с использованием административных данных;

(iv) Методы сопоставления качества переменных в разных источниках;

(v) Редактирование и импутация.

(d) **Этап Итоговые материалы:** Общая оценка качества результатов переписи, полученных с использованием административных данных. Хотя с концептуальной точки зрения это не отличается от оценки итогов традиционной переписи, в настоящем Руководстве предпринята попытка определить возможные отличия.

87. На Диаграмме 1 представлены Этапы оценки качества:

Диаграмма 1: Этапы оценки качества

ЭТАП 1: ИСТОЧНИК – основанная на метаданных оценка качества новых или повторно предоставленных административных источников, которые будут использоваться в переписи	**ЭТАП 2: ДАННЫЕ** – оценка качества первичных административных данных, предоставляемых НСС административными органами
ЭТАП 3: ПРОЦЕСС — Процессы, осуществляемые в данных административных источников, в переписях для преобразования данных для использования в переписи и/или для повышения качества	**ЭТАП 4: ИТОГОВЫЕ МАТЕРИАЛЫ** – Общая оценка качества результатов переписи, полученных с использованием административных данных

88. Настоящее Руководство ориентировано в первую очередь на качество исходных административных источников, а также на этапы Источника и Данных. Тем не менее, качество Процесса и Итоговых материалов учтены для полноты. В конечном итоге на вопрос о том, являются ли административные данные достаточно хорошими для целей переписи, можно ответить только на основе их планируемого использования или результатов переписи, которые они генерируют. Эти четыре стадии не могут быть осмысленно разделены. Для первых двух этапов в Руководстве подробно определены ключевые параметры качества данных для оценки, ключевые инструменты, используемые при выполнении оценки, и, где это возможно, изложены критерии, по которым может проводиться оценка. Кроме того, кратко рассматриваются ключевые вопросы обеспечения качества процесса и результатов, когда оценки переписи производятся с использованием административных данных. Ключевые рекомендации представлены по каждому из этапов, которые кратко изложены в Главе 8, вместе с предложениями по направлениям дальнейшей работы.

3.4 Параметры качества

89. Как отмечалось ранее, под качеством статистических и административных данных понимаются многочисленные параметры, которые не сводятся к ошибкам охвата или измерения. Например, статистические данные, которые являются точными, но устаревшими, имеют ограниченное применение. Определённые Единой статистической системой параметры качества включают:

    (a) Актуальность,

    (b) Точность и надёжность,

    (c) Своевременность и пунктуальность,

    (d) Доступность и ясность, и

    (e) Согласованность и сопоставимость[12].

90. Однако для оценки административных данных эти «стандартные параметры качества не всегда применимы» (см. Daas et al., 2008, p. 2). С другой стороны, они охватывают все значимые параметры качества административных данных. В нижеследующих таблицах указаны параметры для оценки административных источников, описанных в настоящем Руководстве, на каждом из вышеупомянутых Этапов.

[12] Альтернативные параметры используются различными НСС (см., например, Statistics Canada (2017), Statistics Austria (2009). В целом, эти альтернативные подходы охватывают примерно одно и то же содержание, хотя и используют различную терминологию или классификации.

Таблица 1: Параметры качества на этапе Источник

	Параметры качества	Определение
ЭТАП ИСТОЧНИК	**Актуальность и точность**	Степень, в которой административный источник данных удовлетворяет потребностям переписи. Это охватывает совпадение генеральной совокупности переписи концепциям и определениям (актуальность) и степенью того, насколько правильно данные описывают явления, для измерения которых они были разработаны (точность)
	Своевременность	Промежуток времени между окончанием отчётного периода, к которому относится информация, и датой, когда информация становится доступной для НСС
	Согласованность и сопоставимость	Степень, в которой административный источник может быть успешно объединён с другими источниками, используемыми в переписи, включая увязки
	Доступность и интерпретируемость	Простота, с которой НСС может получать административные данные, охватывающие влияние любых ограничений, конфиденциальность и безопасность, приемлемость использования для общественности, лёгкость передачи и получения данных, а также доступность метаданных
	Институциональная среда	Организационные факторы, влияющие на способность поставщика данных предоставлять данные с ожидаемым качеством. Охватывает прочность отношений, предыдущий опыт, наличие официальных соглашений, риски, связанные со статусом поставщика и стандартами качества поставщика

Таблица 2: Параметры качества на этапе Данные

	Параметры качества	Определение
ЭТАП ДАННЫЕ	**Проверка и согласование**	Файлы данных, предоставленные в НСС, имеют машиночитаемый формат. После передачи данных в НСС проводятся дополнительные мероприятия по проверке и согласованию данных для подтверждения того, что ожидаемые переменные/единицы измерения/базовый период/форматы были предоставлены для обеспечения согласованности обработки данных НСС на всех этапах использования в переписи
	Точность и надёжность	Точность, полнота (для переменных и охвата населения) и согласованность предоставляемых данных соответствуют требованиям конкретного варианта использования в переписи. Сравнение с альтернативными источниками позволяет определить приемлемые уровни погрешности измерения или репрезентативные ошибки
	Своевременность и пунктуальность	Своевременность и пунктуальность предоставляемых данных соответствуют требованиям конкретного варианта использования в переписи, для которого они будут использоваться
	Связываемость	Доступны адекватные связующие переменные (т. е. либо общие уникальные идентификаторы, либо комбинация переменных, которые позволяют идентификацию) и они имеют достаточное качество для обеспечения связывания данных

Таблица 3: Параметры качества на этапе Процесс

	Параметры качества	Определение
ЭТАП ПРОЦЕСС	**Точность связывания записей**	В тех случаях, когда производится увязка нескольких источников (друг с другом или с ответами на вопросы переписи), это связывание является точным и объективным, повышая тем самым общее качество методологии переписи и/или набора переписных данных
	Охват и согласованность статистических регистров и переписей на основе административных регистров	В тех случаях, когда имеются регистры переписи (или субрегистры) или когда административные данные используется для дополнения сбора данных переписи, они в достаточной степени охватывают целевую совокупность / переменные, повышая тем самым общее качество методологии переписи и/или набора переписных данных
	Точность устранения расхождений	В тех случаях, когда различные источники связаны друг с другом и в них доступны одни и те же атрибуты, методы выбора между источниками повышают общее качество методологии переписи и/или набора переписных данных
	Точность редактирования и импутации	В тех случаях, когда переменные / единицы переписи получены / созданы с помощью методов импутации или моделирования, полученный производный результат является точным и объективным, повышая тем самым общее качество методологии переписи и/или набора переписных данных

Таблица 4: Параметры качества на этапе Итоговые материалы

	Параметры качества	Определение
ЭТАП ИТОГОВЫЕ МАТЕРИАЛЫ	**Актуальность**	Степень, в которой статистические материалы удовлетворяют текущим и потенциальным потребностям пользователей
	Точность и надёжность	Близость оценочного результата к неизвестному истинному значению и насколько они надёжны во времени и географии
	Своевременность и пунктуальность	Промежуток времени между публикацией и периодом, к которому относятся данные, а также временной лаг между фактической и запланированной датами публикации
	Доступность и ясность	Действия, предпринятые для того, чтобы помочь пользователю данных найти и понять данные, которые его интересуют
	Согласованность и сопоставимость	Степень, в которой данные могут быть сопоставимы во времени и темам. Степень сходства данных, полученных из разных источников или с помощью разных методов, но относящихся к одному и тому же явлению

Источник: *Eurostat, 2013 и 2018.*

3.5 Исследования целесообразности

91. Маловероятно, что новые источники административных данных будут интегрированы в подготовку переписи без предварительного технико-экономического обоснования, проведённого НСС. Качество источника данных может быть установлено путём получения тестовых данных и оценки их качества на различных этапах, предложенных в настоящем Руководстве. Это поможет конструктивному мышлению, то есть разработке методологии переписи, которая максимально использует имеющиеся административные данные и учитывает влияние их использования на качество переписи в целом.

92. Во-первых, технико-экономическое обоснование включает в себя детальное понимание процессов сбора данных поставщиком, охватываемого населения и

переменных, включённых в источник, а также доступности данных (см. Этап Источник, Глава 4). Во-вторых, следует опробовать предоставление, сбор и приём тестовых данных, а также подробно изучить тестовые данные, чтобы выявить проблемы с качеством и определить очистку и гармонизацию, наряду с проверками достоверности (этап Данные, Глава 5). В-третьих, при объединении данных из нескольких регистров их можно использовать для проверки качества данных и для выбора наиболее надёжных переменных и значений в соответствии с разработанными методологическими правилами (этап Процесс, Глава 6). Наконец, оценки, полученные с использованием тестовых данных, можно сравнить с оценками предыдущей переписи или другим подобным «золотым стандартом», что способствует оценке общего качества результатов (этап Итоговые материалы, Глава 7).

93. Как правило, характеристики переписи не могут быть получены непосредственно из данных административных источников, поскольку они были разработаны для других, нестатистических целей. Большинство определений и классификаций, используемых административными органами, отличаются от стандартных статистических определений. Данные из нескольких регистров могут использоваться для построения или получения определённых характеристик переписи, в то время как другие характеристики могут охватываться дублирующей информацией в нескольких регистрах. Это делает технико-экономическое обоснование ключевым для разработки методов получения характеристик для переписи.

94. Методологи переписи должны решить следующие основные проблемы при получении характеристик переписи:

(a) Определение международного стандарта переписи (определение, классификация и т. д.), применимого к целевой характеристике переписи,

(b) Сравнение и сопоставление определений и классификаций переписи с определениями и классификациями, используемыми в административном источнике,

(c) Проверка точности административных данных от альтернативных источников и совместная работа с поставщиками данных для устранения / минимизации любых недостатков,

(d) Определение того, какие и сколько источников требуются для получения и обеспечения качества каждой целевой характеристики переписи,

(e) Установление оптимальных правил для получения каждой характеристики переписи и разработка необходимого программного обеспечения для обработки данных, оптимизированного с учётом качества требуемых результатов, и

(f) Принятие мер для обеспечения создания необходимого реестра или части реестра (например, предложение изменений в процедурах регистрации, правовой среде и т.д.), если характеристики не охвачены какими-либо административными источниками.

Вставка 2: Технико-экономическое обоснование в Эстонии

В 2016 году в Эстонии была проведена пилотная перепись населения и жилищного фонда. Данные для обязательной переменной переписи «Год прибытия в страну» имеются в административном регистре населения страны. Однако после анализа распределений переменную в регистре нельзя было напрямую использовать для переписи. В первой половине 1990-х годов (когда регистр был впервые создан) 1994 или 1995 год был отмечен как год прибытия в страну для многих людей. Сравнение распределения года прибытия в регистре с альтернативными источниками данных о миграции показало, что иммиграция в Эстонии в 1990-е годы была слишком высокой. Для решения этой проблемы использовались данные из пилотной переписи населения и жилищного фонда 2011 года и различные переменные регистра населения (например, дата создания записи и страна рождения), чтобы производная переменная переписи могла максимально соответствовать определению в Принципах и рекомендациях ООН по переписям населения и жилищного фонда (см. 2008, Пересмотр 2).

Вставка 3: Технико-экономическое обоснование в Израиле

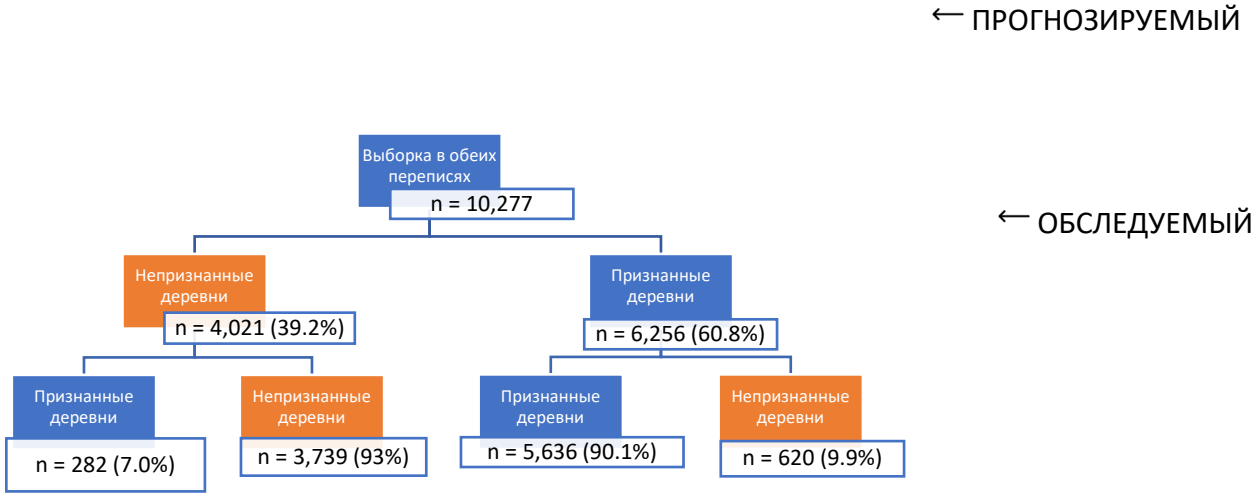
В Израиле было проведено технико-экономическое обоснование для разработки методов выбора наилучшего адреса для труднодоступного бедуинского населения Негева путём сравнения оценок, полученных с использованием административных данных, с оценками, полученными в результате последней традиционной переписи населения в 1995 году. Негевские бедуины — этническая группа, которая включает примерно 283 тысяч арабов-мусульман, проживающих в пустыне Негев. Они представляют собой уникальное население, поскольку традиционно живут как кочевые племена с самобытной культурой (например, 16 процентов мужчин являются полигамными). В ходе традиционной переписи 1995 года были опрошены бедуинские домохозяйства и их места жительства были отмечены на картах. Однако это население считается труднодоступным, поскольку около одной трети этого населения живёт в непризнанных деревнях, которые не подключены к общественной инфраструктуре, такой как электричество, вода или дороги с твёрдым покрытием. Кроме того, у них низкий уровень взаимодействия с государственными учреждениями.

Было проведено исследование по изучению возможностей административных данных для определения географического положения этой группы населения на основе Центрального регистра населения. В нем у каждого человека есть уникальный персональный идентификационный номер (PIN-код), который связан с повседневным взаимодействием отдельных лиц с правительственными учреждениями и службами. Кроме того, каждая запись Центрального регистра населения содержит ссылки на записи об отце, матери, супруге человека и демографические переменные.

Сравнивая данные Центрального регистра населения с данными предыдущей переписи, известно, что этот регистр содержит неотъемлемые ошибки и несовместимость с определениями переписи, включая пропуск резидентов (иностранцев), включение нерезидентов (эмигрантов) и преднамеренно неправильную регистрацию адреса — 20 процентов население не сообщает свой последний адрес. Кроме того, существуют ограничения, характерные для бедуинского населения Негева. Ожидается, что бедуины в пустыне Негев будут зарегистрированы в Центральном регистре населения, их адресная регистрация не позволяет определить точное местонахождение. Это особенно верно для людей, живущих в непризнанных деревнях, зарегистрированных под названиями племён, а не в географическом районе, в котором они живут, потому что племена могут быть разбросаны по всей географической зоне пустыни Негев. Более того, бедуины в непризнанных деревнях намеренно регистрируются в Центральном регистре населения, как если бы они жили в признанных деревнях, чтобы получить услуги, такие как обучение своих детей в школах в признанных деревнях. Наконец, даже бедуины, у которых есть «настоящий адрес» в одной из признанных деревень, могут быть записаны с недостаточной точностью.

В исследовании на первом этапе (начальное местоположение) было использование текущего адреса регистра и их адреса Центрального регистра населения 1995 года , чтобы найти людей в контрольный день. Например, если их адрес не изменился в Центральном регистре населения за период 1995-2019 гг., это означает, что они по-прежнему проживают в той же географической зоне, которая была указана в переписи 1995 г., со своими потомками. На втором этапе использовалась методология «признаков жизни» (см. Глава 6), основанная на других административных источниках (например, записи о браке, изменение адреса, местный налог, водоснабжение, обучении учащихся школы и электроснабжение), чтобы повысить точность данных о местоположении. Затем результаты сравнивались с результатами традиционной переписи 1995 года. Этот метод был протестирован и переоценён путём повторения методологии с данными переписи 2008 года (Диаграмма 2). Было установлено, что с помощью этого метода было предсказано примерно 90 процентов выборки, проживающей в одном и том же географическом районе. Этот результат был дополнительно подтверждён небольшими полевыми испытаниями (n = 110).

Диаграмма 2: Результаты, прогнозируемые с помощью административного метода (уровень 1) по сравнению с методом наблюдения (уровень 2) при переписи 2008 года в Израиле



95. Вышеупомянутые проблемы лучше всего решать с помощью технико-экономического обоснования, как в примерах из Эстонии (Вставка 2) и Израиля (Вставка 3). Второй пример, в частности, подчёркивает как проблемы, так и возможности, которые административные данные могут предоставить при производстве статистических данных о труднодоступных группах населения (см. Вставка 4). Достижение адекватной гармонизации концепций регистров и переписей может быть сложной и трудоёмкой деятельностью, которую не следует недооценивать. Рекомендуется, чтобы включению административных источников данных в подготовку переписи предшествовало проведение технико-экономического обоснования с достаточными ресурсами, которое обеспечивает «подтверждение концепции» для планируемой интеграции административных данных в подготовку переписи. Кроме того, превращение четырёх этапов обеспечения качества в составную часть технико-экономического обоснования позволит НСС напрямую применять уроки, извлечённые из технико-экономического обоснования, в контексте подготовки переписи и лучше информировать пользователей о качестве источников данных.

96. Основываясь на обзоре литературы и опыта НСС, оставшаяся часть настоящего Руководства посвящена инструментам и показателям для оценки качества источников административных данных по каждому из этих параметров. В следующих главах, помимо работы Daas et. al., это Руководство также основано на исчерпывающих наборах показателей качества административных данных, разработанных другими (например, Iwig et. al. 2013; Eurostat ESSnet MIAD 2014; Eurostat ESSnet KOMUSO 2016, 2019).

Вставка 4: Труднодоступные группы населения

Как традиционные переписи, так и административные переписи могут не охватывать определённые группы населения (United States Census Bureau, 2019). И наоборот, административные данные могут включать лиц, которые не были охвачены традиционным сбором данных переписи, например, тех, кто не желает или не может участвовать в переписи, но которые взаимодействовали с административным процессом.

Труднодоступные группы населения часто представляют значительный интерес по политическим соображениям. Это могут быть молодые и подвижные люди; иммигранты, просители убежища и беженцы; принадлежащие к определенным общинам или этническим группам (см. Вставку 3); уязвимые люди, такие как бездомные и люди, живущие в нищете или с плохими условиями жизни. Даже если эти группы были включены в административные данные, их информация может быть значительно устаревшей или плохо записанной.

В Главах 4, 5 и 6 представлены подходы к оценке охвата административных источников и регистров населения, но важно учитывать «скрытые» группы населения и разрабатывать стратегии для охвата этих лиц. Такие стратегии могут включать работу с общественными группами, неправительственными организациями, жилищными ассоциациями и т. д., которые обладают информацией об этих группах или могут посоветовать другие способы охвата интересующего населения (Statistics Canada, 2020). Это также может повлечь за собой изучение новых источников, включая коммерческие источники (где это разрешено законом и приемлемо в соответствии с этическими стандартами), которые могут предоставить информацию о лицах, пропавших без вести в государственных административных системах.

Другой подход состоит в том, чтобы стимулировать труднодоступные группы населения к взаимодействию с административными источниками, что может включать укрепление доверия и устранение опасений по поводу использования их данных; или это может включать стимулирование людей к взаимодействию с административной системой. Реакция на пандемию COVID-19 является примером: образовательные, социальные и медицинские организации предприняли шаги по работе с общественностью в некоторых странах, чтобы взаимодействовать с административными системами, и гарантировать, что они могут адекватно поддерживать все население, в том числе посредством программ тестирования и вакцинации.

Члены труднодоступных групп с большей вероятностью могут быть связаны с противоречивой информацией из административных источников, особенно в отношении их имени и адреса. Например, у представителей высокомобильного студенческого контингента может быть внесён в налоговый регистр адрес родителей, их (устаревший) адрес первого семестра, зарегистрированный в органах образования, и их адрес (текущий) второго семестра, зарегистрированный в органах здравоохранения. В таких случаях для НСС важно выработать глубокое понимание того, как разные группы могут взаимодействовать с различными административными источниками, чтобы иметь возможность принимать решения о том, какая информация может быть актуальной или правильной (см Глава 6).

...

Различия в именах могут быть особенно проблематичными для представителей национальных, языковых или этнических групп и общин меньшинств, где для одного и того же человека в разных источниках могут быть указаны разные имена или разные варианты их перевода. Это может значительно повысить вероятность ложных или пропущенных связей при построении статистического регистра населения из различных административных источников, что может привести к ошибке охвата (см. Главу 6). Чтобы решить эту проблему, важно понимать различные методы присвоения имён, используемые общинами меньшинств и этническими группами, которые затем могут быть включены в методологии увязки.
Охват труднодоступных групп населения и точный сбор их данных остаётся важной задачей для НСС в ЕЭК ООН и за её пределами (Раздел 8.2).

Глава 4 Этап Источник

97. В данной главе приводятся рекомендации по ключевым параметрам качества, процессу оценки и соответствующим инструментам и показателям для оценки качества источников административных данных, которые будут использоваться при проведении переписи, в случае как их первого применения, так и регулярного повторного предоставления НСС. Обычно на этом этапе данные недоступны. Однако, поиск информации об источниках административных данных начинается, скорее всего, в ходе общения и ознакомительных встреч между НСС и поставщиком данных административных источников.

98. Оценка на этом этапе должна привести к выработке рекомендации относительно того, следует ли продолжать осуществление инициативы по получению данных (или продолжить использование этого источника данных). Если будет принято решение о дальнейших действиях, то предоставленные административные данные будут подвергнуты более подробной оценке на этапе Данные.

99. Необходимо оценивать качество источников как при первом получении административных данных, так и в каждом случае, когда они вновь предоставляются НСС. Это связано с тем, что характеристики любого ранее предоставленного набора данных могут претерпевать изменения в концепциях, классификациях, методах сбора и далее.

4.1 Параметры качества источника

100. К параметрам качества данных, которые должны быть изучены на данном этапе оценки качества, относятся:

    (a) Актуальность и точность,

    (b) Своевременность,

    (c) Согласованность и сопоставимость,

    (d) Доступность и интерпретируемость, и

    (e) Институциональная среда.

    Эти параметры описываются ниже, а процессы, инструменты и показатели для оценки представлены в последующих разделах. Следует отметить, что невозможность достичь минимально приемлемого качества по одному из параметров не может быть компенсирована качеством других параметров.

4.1.1 Актуальность и точность

101. Актуальность отражает степень, в которой источник административных данных отвечает потребностям НСС в отношении предполагаемого использования. Для

того чтобы считаться релевантным, источник административных данных должен, например, отвечать целям его использования. Это может быть связано с сокращением затрат или нагрузки на респондентов; повышением качества результатов переписи; подготовкой расширенных или новых материалов переписи. Для этого административный источник должен быть репрезентативным для населения, подлежащего переписи (целевая совокупность), а измерения, полученные от населения, должны соответствовать потребностям переписи. Ключевой частью оценки актуальности является понимание контекста, в котором были собраны административные данные.

102. В рамках оценки релевантности рассматриваются также вопросы точности административных данных. Под точностью понимается степень того, насколько корректно данные описывают явление, которое они призваны измерять. Важно понять, как сбор, обработка и обеспечение качества, осуществляемые административной организацией, могут влиять на точность получаемых данных и их полезность.

4.1.2 Своевременность

103. Своевременность означает период времени между датой, к которой информация относится, и датой, когда информация становится доступной для НСС. Своевременность информации влияет на её актуальность.

4.1.3 Согласованность и сопоставимость

104. Согласованность отражает степень, в которой административные данные могут успешно комбинироваться с данными из других источников, используемых НСС, т. е. с данными переписи, в рамках широкой аналитической основы, с течением времени. Использование стандартных концепций, классификаций и целевых групп населения способствует согласованности как внутри переписи, так и между ними. Таким образом, требуется чёткое понимание оперативных определений, используемых административным поставщиком, цели сбора данных и влияния на сопоставимость изменений в административном источнике во времени, предметным областям, географическим факторам при оценке согласованности.

105. Часто требуется связать административный источник на уровне статистической единицы переписи, чтобы интегрировать данные в план переписи. Поэтому вопрос о сопоставимости идентификаторов, используемых в различных источниках данных, которые должны быть связаны между собой, является одним из параметров согласованности.

4.1.4 Доступность и интерпретируемость

106. Доступность и интерпретируемость означает лёгкость, с которой НСС может получать (и понимать) соответствующие административные данные во всей их полноте. Это включает в себя понимание любых ограничений (юридических и тех,

которые налагаются поставщиком); приемлемость для общественности; конфиденциальность и безопасность; лёгкость передачи и получения данных (пригодность формата или носителя для передачи данных и затрат); доступность и ясность документации и метаданных. Крайне важно, чтобы использование источника административных данных основывалось на правовой базе, которая даёт НСС безусловное право на доступ и использование данных и метаданных для статистических целей.

4.1.5 Институциональная среда

107. Под институциональной средой понимаются организационные или институциональные факторы, которые могут влиять на способность поставщика данных предоставлять данные ожидаемого качества и в соответствии с согласованным графиком (пунктуальность). Это включает в себя:

(a) Прочность отношений с поставщиком данных, эффективность каналов связи и скорость отклика поставщика на запросы НСС,

(b) Наличие (или возможность наличия) официальных соглашений и риски, связанные со статусом и сложностью организации поставщика, и

(c) Стандарты качества и процедуры, принятые административными организациями поставщика.

4.2 Инструменты и показатели

108. Качество административного источника должно оцениваться по параметрам качества, указанным в разделе выше. Ниже приводятся рекомендации по процессу оценки, включая инструменты и показатели для оценки административного источника для использования в переписи.

4.2.1 Актуальность и точность

109. Понимание различий между административной совокупностью населения и требуемой переписной совокупностью, а также между параметрами / переменными в административном источнике и требуемыми переписными характеристиками, имеет важное значение для оценки релевантности и точности. Погрешность, возникающая в результате этих различий, называется, соответственно, погрешностью репрезентативности и измерения (см. Zhang, 2012). На этапе Источник можно получить некоторое представление об этих ошибках и их влиянии на актуальность на основе метаданных о цели и методах сбора данных поставщика. Влияние погрешностей репрезентативности и измерения на точность и надёжность также учитывается на этапах Данные и Процесс (этап Источник — Глава 5 и этап Процесс — Глава 6).

4.2.1.1 Целевая переписная совокупность (репрезентативность)

110. Для оценки релевантности НСС должны определить, соответствует ли набор объектов в источнике административных данных единицам совокупности (населения), представляющей интерес для переписи (целевая совокупность). Объектом является базовый элемент совокупности, по которому собирается информация; это может быть человек, домохозяйство, жилище, событие, операции и т. д. Для установления релевантности в плане репрезентативности предлагаются следующие показатели, по каждому из которых возникает ряд вопросов, которые помогут провести оценку:

(a) Согласование (объектов) с целевыми единицами переписи:

(i) Насколько административные объекты сопоставимы с целевыми единицами переписи?

(ii) Какие определения, методы и процессы используются для идентификации и включения объекта в источник?

(iii) Существуют ли какие-либо законы или нормативные акты, определяющие объекты?

(iv) Проводятся ли поставщиком данных какие-либо проверки для обеспечения актуальности определений?

(v) В случае несовпадения с единицами переписи возможно ли преобразование, способное обеспечить удовлетворение потребностей переписи?

(b) Охват (набора объектов) в отношении целевой совокупности переписи:

(i) Соответствует ли охват объектов потребностям переписи населения?

(ii) Имеются ли доказательства недостаточного охвата (объекты, отсутствующие в источнике, но являющиеся частью целевой совокупности переписи) и чрезмерного охвата (объекты, которые имеются в источнике, но не являются частью целевой совокупности переписи), которые способны повлиять на полезность источника?

(iii) Существуют ли различия между географическими районами, обусловленные различиями в практике поставщика данных или законодательством, которые необходимо учитывать?

(iv) Существуют ли какие-либо правила, законодательные или нормативные требования, включая штрафы за несоблюдение, которые могут повлиять на включение и исключение объектов в источнике?

(v) Какие методы и процессы используются поставщиком данных для включения новых объектов, которые отвечают требуемым критериям / определениям включения (например, процедуры

регистрации), и для исключения объектов, которые больше не соответствуют целевой совокупности для административного источника (например, процедуры снятия с регистрации) [13]?

(vi) В случае наличия ошибок охвата, существуют ли другие источники данных, которые можно использовать в сочетании с источником для преодоления проблемы, например, недостаточного или избыточного охвата в источнике?

4.2.1.1 Переменные / характеристики переписи населения (измерение)

111. Для оценки релевантности НСС должны определить, соответствует ли информация, собранная от объектов административных данных, потребностям переписи в отношении целевых концепций (например, статус занятости, размер домохозяйства, статус пользования жильём и т. д.). Для определения значимости в плане измерения предлагаются следующие показатели:

(a) Наличие целевых переменных / характеристик:

(i) Содержит ли административный источник необходимые переменные исходя из предполагаемого использования данных в целях переписи?

(ii) Охватывают ли переменные / характеристики в целом соответствующий базовый период переписи?

(b) Соответствие концепций, определений и классификаций переменных потребностям переписи:

(i) Сопоставимы ли административные концепции, определения и классификации потребностям переписи?

(ii) Существуют ли различия между идеальными целевыми концепциями поставщика данных и тем, что на самом деле достигается с помощью их оперативного целевого показателя, используемого при сборе данных?

(iii) В случае наличия расхождений в концепциях, определениях и классификациях переписи, возможно ли их преобразование для удовлетворения потребностей переписи?

(c) Погрешность согласования / измерения по сравнению с базовым периодом переписи.

[13] Административный источник или регистр может быть создан поставщиком данных путём увязки нескольких источников. В таких случаях важно понимать качество связи и любую возможность ошибки, включая ошибки охвата (см. Глава 5 и Глава 6 для получения подробной информации о связывании данных и связанной с ней ошибке). Например, колумбийская база данных Sisbén (система идентификации бенефициаров социальных программ) связана с базой данных о смертях Национального реестра, однако есть доказательства пропущенных связей.

(i) Какова частота сбора данных для переменной / характеристике?

(ii) Известны ли задержки между событием или явлением, происходящим или фиксируемым в административном источнике (например, родители могут не регистрировать рождение в течение нескольких недель в национальном регистре рождений)?

(iii) Фиксируются ли временные метки в источнике данных для указания того, к какому периоду относится тот или иной элемент данных?

(iv) Существуют ли какие-либо стимулы или сдерживающие факторы для субъекта данных в плане обновления информации о себе по мере изменения своих обстоятельств и изменения информации в административном источнике (например, льготы или штрафы в связи с выполнением / невыполнением соответствующих требований)?

(d) Качество сбора данных и потенциальная погрешность измерения в сравнении с концепциями переписи:

(i) Какую цель преследует поставщик данных при сборе данных и как это может влиять на качество данных?

(ii) Существуют ли какие-либо юридические обязательства, цели или стимулы (или их отсутствие), которые могут влиять на качество сбора данных?

(iii) Вызывает ли процесс сбора данных поставщиком какие-либо проблемы в плане качества переменных, включая возможность каких-либо искажений? Речь может идти о том, что данные регистрируются опосредованно и, следовательно, не сообщаются непосредственно субъектом данных (что увеличивает вероятность ошибок при регистрации).

(iv) Какие процедуры существуют для подтверждения и проверки данных при вводе поставщиком данных?

(v) Существуют ли какие-либо стимулы или сдерживающие факторы для субъектов данных в плане предоставления поставщику данных полной и точной информации?

(e) Качество обработки данных и вероятность ошибки при обработке поставщиком данных:

(i) Можно ли считать, что проводимая поставщиком данных обработка обеспечивает соответствие качества полученных данных потребностям переписи?

(ii) Какие проверки проводит поставщик данных для обеспечения качества?

(iii) Подвергаются ли данные редактированию или импутации? Если да, то когда и как, и имеется ли в источнике данных индикатор, позволяющий определить, когда произошло редактирование и импутация?

(iv) Существуют ли какие-либо правила, положения или стимулы для поставщика данных, которые могут повлиять на способ обработки данных?

112. На этапе Источник оценка показателей обычно основывается на качественной оценке (например, путём указания того, насколько удовлетворяет требованиям каждый показатель (полностью, частично или не удовлетворяет, с объяснением причин на основе ответов на набор вопросов). Количественная оценка погрешности репрезентативности и измерения осуществляется на этапе Данные (на основе анализа данных) по параметру точности и надёжности (этап Данные, Глава 5).

113. Оценка по показателям должна служить основой для принятия решения, часто с опорой на опыт и экспертное суждение, об использовании (или дальнейшем использования) источника в переписи. При принятии решения следует учитывать, может ли источник данных удовлетворять потребностям переписи (например, сокращению затрат и нагрузки на респондентов, улучшению и усовершенствованию результатов переписи), с учётом любых расходов или рисков (см. ниже в разделе "Институциональная среда и аспекты доступности").

114. В литературе описаны различные системы качества, которые предлагают аналогичные показатели, как в настоящей главе, с использованием параметров качества, а также наборы вопросов и системы оценки для информационного обеспечения оценки (см. например, Daas et. al., 2009; Eurostat ESSnet MIAD 2014; Iwig et. al., 2013; Руководство по оценке административных данных Статистического управления Канады (Statistics Canada Lavigne and Nadeau 2014); Оценка качества административных данных Статистического управления Австрии, Документация по методам (Statistics Austria’s Quality Assessment of Administrative Data, Documentation of Methods Framework, Statistics Austria 2019). Тематическое исследование Новой Зеландии (раздел 4.4.1) представляет собой практический пример структуры, используемой для оценки административных источников для использования в переписи.

4.2.2 Своевременность

115. Административный источник может охватывать соответствующий период времени переписи, но для того, чтобы он был полезен, он также должен быть своевременно доступен для переписи. Для оценки своевременности предлагается следующий показатель:

(a) Своевременность и периодичность предоставления данных в соответствии с потребностями переписи.

(i) Каков интервал между окончанием отчётного периода для административных данных и датой, когда данные доступны НСС?

(ii) Как часто данные могут предоставляться в НСС для удовлетворения потребностей переписи?

(iii) Существуют ли какие-либо требования в отношении способа передачи, требуемых форматов и структур данных, используемых НСС, которые могут повлиять на сроки поставщика данных?

(iv) Достаточно ли времени с момента получения данных, чтобы НСС смогли обработать данные для использования в переписи?

116. В тех случаях, когда данные вряд ли будут доступны вовремя, НСС, возможно, пожелает выяснить, может ли предварительная версия набора данных быть предоставлена досрочно. В таких случаях набор данных может быть неполным и подверженным более высоким уровням погрешности. Таким образом, может быть установлен компромисс между своевременностью данных и их точностью.

117. Как указано ниже в отношении параметра институциональной среды, важно прописать даты передачи данных с контрольными периодами в официальных соглашениях с поставщиком данных. Несмотря на то, что данные могут быть доступны своевременно исходя из требований поставщика данных, они не обязательно будут переданы своевременно НСС, в то время как последние несут официальную ответственность за своевременное проведение переписи.

4.2.3 Согласованность и сопоставимость

118. Важно оценить, в какой степени административный источник может успешно комбинироваться с другими источниками данных для использования в переписи. Информация, собранная по показателям, определенным для оценки релевантности, может быть использована для оценки согласованности. Это включает информацию о различиях между основополагающими концепциями, определениями, классификациями и методами источника административных данных и других источников данных для комбинированного использования в переписи.

119. Для проведения полной переписи на основе регистров важно проанализировать характеристики переписи и источник административных данных; провести сопоставление и определение степени, в которой информация, содержащаяся в источнике административных данных, позволяет получить соответствующие характеристики переписи. В частности, НСС должны установить, соответствуют ли данные в регистрах определениям характеристик в переписи. В случае частичного или полного несоответствия НСС следует изучить причины несоответствия характеристик переписи и информации, имеющейся в административном источнике данных.

4.2.3.1 Сопоставимость

120. Административные данные подвержены изменениям и различиям во времени и по географическому признаку в связи с изменениями в законодательстве, нормативных актах и процедурах, которые могут влиять на концепции, определения, классификации и охват источника. В более общем плане изменения могут влиять на все показатели, касающиеся репрезентативности и измерения, как указано в разделе "Измерение значимости". Это имеет особое значение для переписи, когда стабильность во времени может быть ключевой проблемой. Для оценки сопоставимости предлагается следующий показатель:

(a) Сопоставимость во времени и областях.

(i) Существуют ли какие-либо изменения во времени или различия между географическими районами, влияющие на:

- Определение и охват объекта в административном источнике, имеющем отношение к переписи?
- Понятия, определения и классификации, связанные с переменными административного источника, имеющими отношения к переписи?
- Процедуры сбора, обработки и обеспечения качества данных, которые способны повлиять на качество исходных данных для целей переписи?

4.2.3.2 Возможность увязки

121. Одним из соображений согласованности и сопоставимости, является лёгкость, с которой административный источник данных может быть связан с другими соответствующими наборами данных для переписи. В тематическом исследовании Эстонии (Раздел 4.4.3) приводится пример того, как можно связать различные источники административных данных с несколькими разными уникальными идентификаторами. Для оценки возможности увязки источника предлагаются следующие показатели:

(a) Наличие уникального ключа для связи.

(i) Содержит ли источник уникальный идентификатор, совпадающий с уникальными ключами, необходимыми для связывания в целях переписи?

(ii) Доступен ли идентификатор по всем соответствующим объектам в источнике?

(b) Наличие уникальной комбинации переменных для увязки.

(I) Содержит ли источник уникальную комбинацию переменных (например, имя, возраст и адрес), которые можно было бы использовать для увязки с переписью?

(II) Имеется ли уникальная комбинация переменных для каждого объекта в источнике?

122. Качество связующих переменных также оценивается на этапе Данные (Глава 5) и качество процесса связи рассматривается как часть этапа Процесс (Глава 6).

4.2.4 Доступность и интерпретируемость

123. Для оценки доступности и интерпретируемости предлагаются следующие показатели:

(a) Ограничения на доступ к данным и их использование,

(b) Приемлемость для общественности,

(c) Простота передачи и получения данных, и

(d) Интерпретируемость источника — ясные и полные метаданные

124. В нижеследующих разделах приводится подробная информация, необходимая для оценки по каждому из показателей.

4.2.4.1 Ограничения на доступ к данным и их использование

125. Важно выявить любые ограничения, которые могут влиять на способность НСС получать доступ к административному источнику и использовать его. Например, существующие ограничения на защиту данных, прописанные в правовых актах, могут накладывать определённые ограничения на сбор и обработку данных (особенно, если данные защищены дополнительными мерами безопасности или законами на государственном уровне). Такие правовые акты могут касаться конкретных источников (например, см. тематическое исследование Эстонии в разделе 4.4.3) или могут быть более общими, разрешающими НСС доступ к таким источникам данных по мере необходимости, при условии согласия поставщика данных. Поставщик данных может также наложить дополнительные ограничения на данные и разрешённое использование. Они могут включать:

(a) Исключение записей или переменных,

(b) Процедуры раскрытия (до предоставления), такие как шифрование идентификаторов, помехи, группирование или кодирование по верхней границе предоставленных данных,

(c) Ограничения на использование данных,

(d) Ограничения на хранение данных и правила исключения/удаления,

(e) Правила о методах раскрытия информации, которые должны применяться НСС, влияющие на результаты переписи.

126. НСС следует выявить и описать любые применимые ограничения, по которым можно провести оценку воздействия (и рисков воздействия) ограничений на использование административного источника для переписи. В рамках оценки НСС

следует также рассмотреть вопрос о том, способны ли они обеспечить соблюдение ограничений. Это может включать в себя технические и процедурные меры защиты, которые должны внедрить НСС. Эти меры защиты, как правило, являются частью Меморандума о взаимопонимании или Соглашения о безопасности данных с владельцем/поставщиком данных. В частности, в Меморандуме о взаимопонимании может быть прописано, как будет защищена личная информация.

4.2.4.2 Приемлемость для общественности

127. Предоставление НСС доступа к источнику данных для использования в целях переписи также может зависеть от общественного признания. Общественность должна понимать и поддерживать новые подходы и использование своей информации или, по крайней мере, не относиться к этому враждебно. Если общественность выступает против использования административного источника данных, существует риск для качества. Например, это может отрицательно сказаться на сотрудничестве общественности с переписью или административным источником, используемым для переписи. Поэтому НСС следует обеспечивать прозрачность в отношении использования административных источников для переписи, подчёркивая преимущества для общественности и одновременно обеспечивая гарантии конфиденциальности и безопасности.

128. Для оценки приемлемости для общественности могут использоваться следующие инструменты или процессы:

    (a) Консультации с общественностью или вовлечение общественности в обсуждения,

    (b) Оценка воздействия на конфиденциальность, или

    (c) Оценка этики данных.

129. НСС может проводить консультации или привлекать общественность к использованию административных данных для переписи (или для других статистических исследований или результатов). Это может принимать различные формы, включая официальные консультации, анкетирование (через опросы или приглашения НСС оставить отзыв на веб-сайте), качественное исследование общественного мнения или использование представительных групп граждан. Группы граждан имеют целью собрать вместе представителей общественности (быть репрезентативными для определённой совокупности или отражать различные группы интересов) для оценки их взглядов и мнений. Взаимодействие с ключевыми группами общества, такими как общины коренных народов и этнические меньшинства, имеет важное значение для определения и удовлетворения их конкретных потребностей и проблем, связанных с использованием относящихся к ним данных, особенно в тех случаях, когда предлагаемые виды использования не являются теми, для которых данные были первоначально собраны.

130. Оценка воздействия на конфиденциальность представляет собой формальный процесс, результатом которого является документ, описывающий процессы,

выводы и результаты, который помогает НСС изучить воздействие новой программы или услуги (или предлагаемых программ и планов) на конфиденциальность отдельных лиц. В качестве инструмента управления рисками, используемого на этапе планирования программы или инициативы по оказанию услуг, оценка воздействия на конфиденциальность помогает организациям более полно учитывать последствия заданного предложения для конфиденциальности. Оценка воздействия на конфиденциальность также используются для того, чтобы управляющие данными могли выполнять свои обязательства в соответствии с Общим положением о защите данных (в соответствии с Европейским законодательством). Оценка воздействия на конфиденциальность может применяться к различным видам предполагаемого НСС использования источника данных в переписи. Тематическое исследование Новой Зеландии в разделе 4.4.2 описывает связанные с этим риски конфиденциальности и меры по их устранению или снижению, используемые НСС для каждого из рисков.

131. Оценка этики данных проводится для установления того, является ли доступ к использованию и обмену общедоступными данными в исследовательских и статистических целях этичным и служит общественному благу. НСС могут использовать инструмент самооценки этики (см. например, UKSA, 2020), и/или они могут использовать официальный орган для предоставления экспертных заключений или одобрения, например, такой как Консультативный комитет по этике данных (Консультативный комитет по этике данных Национального статистического комитета Великобритании[14]).

132. Выводы, сделанные по итогам консультаций и обсуждений с общественностью, оценки воздействия на конфиденциальность и оценки этики, могут помочь НСС оценить приемлемость использования административных источников в переписи (и для других статистических данных НСС).

[14] Дополнительную информацию можно получить по адресу https://uksa.statisticsauthority.gov.uk/about-the-authority/committees/national-statisticians-data-ethics-advisory-committee/#:~:text=The%20National%20Statistician%E2%80%99s%20Data%20Ethics%20Advisory%20Committee%20%28NSDEC%29,advise%20the%20National%20Statistician%20on%20the%20ethical%20

Вставка 5: Статистическое управление Канады: Центр доверия

В Статистическом управлении Канады есть Центр доверия, который определяет, как защищается информация, уделяя первостепенное внимание конфиденциальности. Это включает в себя то, как уравновешиваются потребности общества в новых данных и защита конфиденциальности с использованием современной структуры «необходимости и соразмерности». Центр доверия предоставляет чёткую и полную информацию, чтобы убедить общественность в использовании их данных, включая использование инфографики и коротких видеороликов, доступных через веб-сайт. Одно из таких видео «Аноним Джо» объясняет, как используются общедоступные данные, в том числе о важности объединения нескольких источников. В Статистическом управлении Канады особое внимание уделяется работе и культуре защиты данных, в том числе обещание защищать личности людей, их семьи и их бизнес.

Открытость и прозрачность лежат в основе Центра доверия, а информация об административных источниках, которые будут собираться и использоваться Статистическим управлением Канады, публикуется (и регулярно обновляется) на веб-сайте (Available at https://www.statcan.gc.ca/eng/trust).

4.2.4.3 Лёгкость передачи и получения данных

133. Поставщик данных может использовать совершенно другие модели данных, форматы, схемы, программное и аппаратное обеспечение, чем те, с которыми знакомы НСС. Речь в частности идёт о методах хранения и передачи данных (включая меры безопасности при их передаче). Структуры данных также могут быть сложными, а размеры файлов — чрезвычайно большими (особенно в случае данных об операциях). Важно, чтобы НСС понимали такие различия и сложности для того, чтобы оценить, смогут ли системы НСС принять и загружать такие наборы данных. Этот процесс может включать переговоры с поставщиком о разработке процессов и систем для облегчения передачи наборов данных в формате, отвечающем потребностям НСС. Однако это может быть трудоёмким и дорогостоящим процессом.

134. В более общем плане затраты являются ключевым фактором, который следует учитывать при оценке лёгкости доступа. Сюда могут входить затраты поставщика данных или затраты НСС на развитие своего потенциала для получения административного набора данных (покупка нового программного или аппаратного обеспечения). Важно сопоставить любые затраты с ожидаемой стоимостью, которую принесёт новый административный источник.

135. На практике подробная информация об условиях передачи данных в НСС, включая структуру и формат файлов (например, неструктурированные файлы, реляционная база данных; SAS, Excel или текстовые форматы и т.д.), переменные, частоту предоставления, даты передачи, стандарты данных и согласованные расходы, будет включаться в Соглашения об обмене данными или передаче между НСС и

поставщиком. Такие соглашения будут подписываться уполномоченными руководителями каждой из организаций.

4.2.4.4 Интерпретируемость источника — ясные и полные метаданные

136. Оценка интерпретируемости связана с наличием и доступностью всеобъемлющих и чётких метаданных и документации об административном источнике. Без этого невозможно понять и оценить административный источник на предмет предполагаемого использования. Метаданные должны включать в себя подробную информацию о:

    (a) Административной организации,

    (b) Цели сбора,

    (c) Используемых концепциях, определениях, классификациях и протоколах,

    (d) Сборе и обработке данных,

    (e) Методах и процедурах проверки и обеспечения качества,

    (f) Единицах отчётности и переменных, включая словари данных, структуры файлов, форматы и взаимосвязи внутри данных.

137. Эта информация важна для оценки по сравнению с другими параметрам качества, описанными в данной главе. Часто бывает так, что чёткие и полные метаданные не будут существовать по всем аспектам административного источника на начальном этапе изучения источника для использования НСС. Поэтому необходимо работать с поставщиком данных в целях построения соответствующих метаданных. Это зависит от хорошей связи с поставщиком и готовности поставщика работать с НСС (см. ниже «Институциональная среда»). В зависимости от сложности административного источника, НСС может принять решение о прикомандировании сотрудников для работы в административной организации в целях углубления понимания источника. После определения и понимания важно, чтобы метаданные записывались, хранились и поддерживались («база метаданных»), чтобы в будущем их можно было легко найти.

Вставка 6: Шаблоны метаданных для оценки административных источников

Новая Зеландия

Статистическое управление Новой Зеландии имеет Руководство по отчётности о качестве административных данных (см. Stats NZ 2016) с соответствующим шаблоном метаданных для административных данных (доступно на Stats NZ 2020). Шаблон является полезным ресурсом для сбора метаданных об административном источнике, охватывающим общую информацию об административной организации, сборе данных, объектах и переменных совокупности населения, изменениях с течением времени и аспектах доступности.

Нидерланды

Контрольный перечень Статистического управления Нидерландов для оценки качества административных источников (см. Daas et. al., 2009) предоставляет собой полезный шаблон (приложение к документу) для записи информации и метаданных об источнике. Упорядочивание измерений и индикаторов в шаблоне помогает пользователю эффективно выполнять запись и оценку метаданных, обеспечивая выявление проблем на ранней стадии, прежде чем переходить к более поздним этапам.

Статистическая сеть по методологиям комплексного использования административных данных (SN-MAID), проект

Результат B2.3 (Источник) и B2.4 (Метаданные) (SN MAID, 2014) предоставляют контрольные списки, включая индикаторы качества и поля для записи метаданных об административном источнике, которые используются для оценки качества источника для использования в статистике. Контрольные списки основаны на работе Daas et. al. (2009).

4.2.5 Институциональная среда

138. НСС полностью зависят от поставщика данных для сбора, обработки и предоставления административных данных в соответствии с ожидаемым качеством и согласованным графиком. НСС также зависят от качества информации, предоставляемой поставщиком о данных (см. Интерпретируемость, раздел 4.2.4.4 выше) и о любых прогнозируемых изменениях в данных. Поэтому важно оценить уверенность в способности поставщика данных удовлетворять эти требования. Для оценки институциональной среды предлагаются следующие показатели:

(a) Прочность отношений с поставщиком данных,

(b) Предыдущий опыт работы с поставщиком данных,

(c) Наличие официальных соглашений,

(d) Риски, связанные со статусом поставщика, и

(e) Стандарты качества поставщика.

139. **Прочность отношений**. Должны быть созданы процессы для управления отношениями с поставщиком данных, обеспечивающие постоянный диалог. Они должны включать в себя механизмы:

(a) Доведение требований НСС поставщику данных,

(b) Своевременное информирование (поставщиком) о любых изменениях, которые могут повлиять на источник данных (например, об изменениях в законодательной базе в отношении данных, в концепциях и классификациях, а также в процессах и процедурах сбора, управления и передачи данных),

(c) Направление поставщику любых вопросов об источнике данных, и

(d) Обеспечение обратной связи с поставщиком, которая может привести к улучшению качества источника.

Вставка 7: Инструментарий измерения качества: общение с поставщиками данных

Инструментарий обеспечения качества административных данных Статистического управления Великобритании (см. UKSA 2015b) описывает «области практики», связанные с качеством данных, включая область связи с поставщиком данных. Эта область охватывает важность сотрудничества со сборщиками данных, поставщиками, ИТ - специалистами, политиками и должностными лицами. Подчёркивая важность официальных соглашений с подробным описанием договорённостей (см. ниже), а также регулярного взаимодействия со всеми вовлечёнными сторонами. В зависимости от важности предлагается три уровня гарантии: базовый, расширенный и всеобъемлющий.

140. **Предыдущий опыт.** Это включает в себя, насколько быстро поставщик реагировал на запросы НСС, возникали ли какие-либо проблемы с предыдущими поставками данных (например, несвоевременная передача, непредвиденные ошибки), предоставлял ли поставщик точную информацию об источнике данных в прошлом (это могло быть установлено в результате проверок, проведённых НСС на более позднем этапе).

141. **Официальные соглашения.** Это включает в себя наличие или возможность разработки письменных соглашений (юридических или иных), охватывающих:

(a) Роли и обязанности НСС и поставщика. Это включает в себя, играет ли НСС важную роль в утверждении изменений в административном источнике, используемом (или планируемом к использованию) для переписи,

(b) Правовые основания для предоставления данных в соответствии с требованиями безопасности и/или конфиденциальности и техническими условиями,

(c) Спецификация требований в соответствии с Соглашением об обмене/передаче данных, о котором шла речь в разделе 4.2.4.3 выше.

Вставка 8: Система базовых регистров Статистического управления Нидерландов

В Нидерландах принята система административных базовых регистров, включающая 13 регистров населения (резидентов и нерезидентов), адресов и зданий, предприятий, объектов недвижимости (границы, владелец, стоимость и т.д.), топографии (карты: земля, вода, дороги), автомобилей (модель, цвет, владелец и т.д.), облагаемого налогом дохода, рабочей силы (заработная плата, работодатели, социальные пособия и т.д.) и подземной инфраструктуры (канализация, кабели и т.д.). Система базовых регистров опирается на законодательство и поддерживает подготовку статистических данных (включая переписи) Статистическим управлением Нидерландов.
Использование данных из базовых регистров обязательно для государственных органов. Цель состоит в том, чтобы все пользователи системы внесли свой вклад в качество данных. Следовательно, пользователи обязаны уведомлять владельцев базовых регистров, если у них есть альтернативные данные, которые считаются более качественными (за исключением НСС, по юридическим соображениям). Пользователи базовых регистров могут рассчитывать на их достоверность. Статистика, основанная на базовых регистрах, требует лишь ограниченного объёма редактирования данных. Поскольку базовые регистры связаны друг с другом посредством идентификационных номеров, из этого следует, что статистические данные также взаимосвязаны.

Каждый базовый регистр имеет свой проектный офис. В проектных советах представлены все группы заинтересованных сторон. Правления проектных советов действуют в рамках законодательства и следят за тем, чтобы данные регистра соответствовали требованиям закона (качество, полнота и т. д.) И чтобы данные использовались правильно. Советы проектного офиса действуют как консультативные советы ответственных министров кабинета министров и собираются несколько раз в год.

142. **Статус поставщика данных**. Риски, связанные со статусом поставщика, должны оцениваться НСС путём выяснения того, является ли поставщик признанной, стабильной и авторитетной организацией. При этом следует рассмотреть вопрос о том, существует ли какая-либо правовая или нормативная основа для выполнения поставщиком административной функции, которая придала бы уверенность в устойчивости и качестве источника. Следует также учитывать риски, связанные со сложностью организации(ий) — поставщика(ов), участвующего(их) в сборе,

обработке и сообщении данных источника. Например, может быть задействовано несколько органов или организаций, каждый из которых влияет на качество передаваемых окончательных данных.

143. **Стандарты качества поставщика данных.** Следует оценить, сможет ли поставщик соответствовать ожиданиям НСС в отношении качества. При этом следует учитывать информацию о принципах, стандартах и руководствах, используемых поставщиком для обеспечения качества, включая действующую процедуру, охватывающую сбор, обработку и передачу данных НСС. Информация о том, как поставщик проверяет, соблюдаются ли стандарты, является ценной; это могут быть результаты внутреннего или внешнего аудита, проводимого регулирующими органами или профессиональными организациями. Поставщик данных также может составлять отчёты о качестве, которые должны рассматриваться НСС. Более подробная оценка, основанная на ключевых аспектах административного источника, включена в раздел "Измерение качества релевантности" выше.

144. После оценки поставщика данных на основе вышеизложенных критериев, НСС может оценить риски, связанные со своевременной передачей административных данных поставщиком, в соответствии с требуемыми критериями качества.

4.3 Рекомендации для этапа Источник

(a) Определить актуальные и перспективные административные источники для использования в переписи (см. Глава 2).

(b) Чётко изложить требуемую целевую группу населения, переменные и концепции, а также ожидаемые результаты использования административного источника в переписи, на котором будет основываться оценка.

(c) Понимать ограничения и проблемы, связанные с получением и интеграцией административных источников в перепись, в том числе в тех случаях, когда могут потребоваться изменения в методах, процессах и технологических системах НСС.

(d) Создавать и поддерживать чёткие и полные метаданные, содержащие всю соответствующую качественную информацию об источнике (это станет ценным ресурсом для НСС). Важно структурировать метаданные с использованием соответствующего согласованного стандартного формата метаданных (см. Cornell University Research Data Management Service Group, 2020).

(e) Развивать хорошее понимание поставщика данных, контекста и целей сбора данных, а также стандартов качества, которых они придерживаются.

(f) Налаживать прочные отношения с поставщиком данных для обеспечения эффективного обмена информацией — формирование общего понимания потребностей друг друга.

(g) Заключение официальных соглашений, в которых чётко излагаются требования к НСС и поставщикам данных, роли и обязанности.

(h) Тщательно оценить ценность приобретения и использования административного источника с учётом любых рисков и затрат. Это может относиться к стабильности источника с течением времени и риску того, что поставщик данных не сможет предоставить данные вовремя или с ожидаемым качеством.

(i) Обеспечение прочной правовой основы для получения и использования административного источника с эффективными гарантиями защиты конфиденциальности субъектов данных.

(j) Быть чётким и прозрачным в отношении использования административных данных, показывая доказательства того, что преимущества перевешивают любые проблемы конфиденциальности.

(k) Признать, что объекты, определения, концепции и контрольные даты в административном источнике могут не соответствовать целевым показателям переписи. Следовательно, необходимо будет преобразовать данные и сделать выводы о том, какие уровни несовпадения приемлемы.

(l) Оценивать качества на постоянной основе (с использованием описанных процессов и инструментов), реагируя на любые ожидаемые или известные изменения в источнике.

(m) Документировать и публиковать сильные и слабые стороны, связанные с административными источниками, используемыми для переписи, чтобы пользователи данных были уверены в данных и могли учитывать любые ограничения.

(n) Будьте готовы к тому, что потребуется время, чтобы понять и получить источники административных данных для использования в переписи, в частности, когда требуется программа действий по разработке регистров для использования в переписи (согласно тематическому исследованию Эстонии).

4.4 Тематические исследования

4.4.1 Новая Зеландия: оценка источника

145. В марте 2012 года правительство Новой Зеландии согласовало стратегию преобразования переписи. Частью первого этапа этой программы было завершение широкого обзора возможностей административных данных для получения развёрнутой (социальной и экономической) информации, которая в настоящее

время предоставляется переписью (см. O'Byrne et al, 2014). В ходе этого процесса были определены источники административных данных, относящиеся к темам переписи, и использовались меры качества для оценки того, насколько вероятно, что эти источники удовлетворяют информационным потребностям, которые ранее удовлетворялись при традиционной переписи. Исследование не включало подсчёт населения и демографические разбивки, которые были исследованы в других местах.

146. Цель этой работы заключалась в том, чтобы на раннем этапе выявить вероятную способность существующих административных источников данных предоставить развёрнутую информацию переписи в долгосрочной перспективе и решать, куда направлять более углублённый анализ.

147. Этапы этого процесса включали:

    (a) Определение источников данных — достигается за счёт использования существующей статистики, использования административных источников, поиска в Интернете и контактов с правительственными учреждениями.

    (b) Понимание характера и содержания потенциальных источников административных данных — достигается путём анализа общедоступной информации, обсуждений с экспертами из Статистического управления Новой Зеландии и агентствами-источниками.

    (c) Использование пяти важнейших параметров качества для оценки качества.

    (d) Присвоение рейтинга качества — вероятность того, что административные данные могут соответствовать теме переписи.

148. Показатели качества, использованные в оценке, были адаптированы из существующих систем измерения качества, таких как модель качества Статистического управления Новой Зеландии (см. Stats NZ quality model, Eurostat, 2009 and 2011). Пятью критериями, определёнными для этой оценки, были: актуальность, точность охвата, точность увязки, своевременность и доступность. Эти показатели качества были выбраны потому, что они являются строго избирательными в том смысле, что они необходимы для использования административных данных для переписи, а также являются показателями, по которым можно сделать разумные суждения на основе метаданных.

149. Эта оценка проводилась путём совместной оценки как можно большего количества источников административных данных, которые могут потребоваться для удовлетворения этой переменной переписи. Для каждой переменной каждое измерение качества оценивалось как отличное, хорошее или плохое, что определяло общую оценку «вероятно», «возможно» или «маловероятно», которая будет удовлетворена административными источниками данных. Ключевые вопросы, рассмотренные для каждого параметра изложены в Таблице 5.

Таблица 2: Ключевые вопросы по каждому параметру
Источник: Статистическое управление Новой Зеландии (Stats NZ).

Показатели качества	Основные вопросы для оценки
Релевантность	Насколько близки административные данные к статистической концепции? (Тема переписи используется как соответствующая для статистической концепции) Кто/что должно быть включено в эти данные (целевая группа населения)? Кто/что включено в эти данные (обследуемое население)?
Точность охвата	Есть ли неохваченные люди или ответы (недоучёт)? Есть ли повторяющиеся записи или другие люди, которых не следует включать (избыточный учёт)?
Точность связи	Возможно ли связать данные с переписью населения или списками жилых помещений?
Своевременность	Как часто предоставляются данные? Через какое время после контрольной даты данные доступны для Статистического управления Новой Зеландии?
Доступность	Существуют ли проблемы с конфиденциальностью или юридические проблемы, связанные с использованием этих данных? Есть ли другие препятствия для доступа?

150. Исследование показало, какие административные источники будут наиболее важными для предоставления информации для переписи, и в настоящее время завершён подробный анализ большинства переменных, определённых как «возможные» или «вероятные». Один из наиболее важных выводов заключался в том, что большинство текущих переменных переписи вряд ли будет получено из административных источников, и переписной лист все равно будет необходим. Используемые рейтинги качества показаны в Таблице 6.

Таблица 3: Рейтинги качества

Показатели качества	Определение рейтинга качества		
	Отличный	Хороший	Низкий
Релевантность	Данные, собранные в административных источниках, близки к статистической концепции	Данные, собранные в административных источниках, не совпадают со статистической концепцией, но близки или связаны с аналогичной статистической концепцией, которая может быть приемлемой.	Данные, собранные в административных источниках, совершенно не соответствуют интересующей нас статистической концепции

Показатели качества	Определение рейтинга качества		
	Отличный	Хороший	Низкий
Точность охвата	Охват (чистый, ниже и выше) аналогичен переписи	Охватывается большая часть населения, а те, кто не охвачены, «пропадают случайно»	Охват (чистый, неполный и избыточный) очень низкий или имеется смещение в распределении недостающих значений
Точность связи	У данных есть отличные индивидуальные идентификаторы, которые могут связывать единицы в одном наборе данных с другими наборами данных	У данных есть хорошие индивидуальные идентификаторы	У данных нет индивидуальных идентификаторов. Увязки данных невозможны
Своевременность	Данные обновляются не реже одного раза в год и доступны Статистическому управлению Новой Зеландии в течение двух лет	Данные обновляются не реже одного раза в два года и вскоре становятся доступными для Статистического управления Новой Зеландии	Данные обновляются не регулярно или с задержками более чем на два года
Доступность	Никаких проблем с конфиденциальностью или законом не существует. Статистическое управление Новой Зеландии понимает данные и поддерживает хорошие отношения с владельцем административных данных	Некоторые проблемы конфиденциальности или юридические проблемы существуют с одним или несколькими ключевыми наборами данных	Существуют серьёзные проблемы с конфиденциальностью или юридическими проблемами. Нет отношений с административным владельцем или истории использования данных

Источник: Статистическое управление Новой Зеландии (Stats NZ).

4.4.2 Новая Зеландия: Оценка воздействия на конфиденциальность

151. Оценка воздействия на конфиденциальность — полезный инструмент при рассмотрении аспекта качества доступности, в частности юридических последствий использования административных данных и укрепления общественного доверия. В Новой Зеландии Управление уполномоченного по вопросам конфиденциальности предоставляет руководства и шаблоны для поддержки организаций, которые проводят оценку воздействия на конфиденциальность. В этом руководстве изложены 12 принципов конфиденциальности (эти принципы взяты из Закона о конфиденциальности 1993 года и варьируются от сбора данных до использования уникальных идентификаторов), которые следует рассматривать как часть оценки воздействия на конфиденциальность. Он также включает руководство по ключевым вопросам, которые следует задать в ходе процесса, некоторые общие риски, о которых следует знать, а также возможные стратегии смягчения последствий, которые следует учитывать. До переписи населения Новой Зеландии 2018 года Статистическое управление Новой Зеландии привлекло внешнюю организацию для завершения независимой оценки воздействия на конфиденциальность о планируемом использовании административных данных в переписи. Позже Статистическое управление Новой Зеландии завершило и опубликовало дополнительную оценку воздействия на конфиденциальность, в которой говорилось о намерении расширить использование административных данных, чтобы снизить уровень отклика ниже ожидаемого. Общая цель оценки воздействия на конфиденциальность в этом контексте — собрать воедино информацию о том, что, почему и как НСС хочет использовать определённые административные данные, а также оценить потенциальную выгоду с учётом ряда соображений конфиденциальности.

152. Ключевые темы, затронутые во втором издании оценки воздействия на конфиденциальность переписи 2018 года:

(a) Информация о преимуществах использования административных данных в переписи и подробные сведения о том, как обеспечивается безопасность в процессе построения окончательного набора данных переписи,

(b) Краткое изложение соответствующего законодательства,

(c) Резюме оценки конфиденциальности по каждому из 12 принципов конфиденциальности,

(d) Рекомендации и план действий по минимизации ущерба, и

(e) Таблица рисков и мер по снижению рисков, содержащая оценки рисков (последствий и вероятности) для каждого из 12 принципов конфиденциальности, а также некоторые дополнительные принципы, отражающие обязательства в соответствии с Законом о статистике 1975 года.

153. Оценка воздействия на конфиденциальность показала, что использование административных данных при переписи является законным, безопасным и выгодным для жителей Новой Зеландии.

4.4.3 Эстония: совершенствование данных с помощью законодательства

154. Статистическое управление Эстонии в течение 2010-2013 гг. проводило работу в сотрудничестве с поставщиками данных и научными сообществами. Целью было обеспечение качества административных источников, которые будут использоваться при проведении переписи. Были проанализированы требования к тем характеристикам переписи, которые изложены в правилах Совета Европы и Европейского парламента, а также в правилах Европейской комиссии (см. European Commission 2008). Был нанесён на карту охват каждой характеристики переписи и были внесены предложения по формированию характеристик переписи в будущем и для анализа качества.

155. На основании этого анализа был сделан вывод, что для предоставления данных достаточного объёма и качества потребуется до 20 различных административных источников (принадлежащих девяти различным органам или министерствам). Статистическому управлению Эстонии было поручено определить минимальные универсальные критерии для всех тех регистров, которые требовались для предоставления данных для удовлетворения потребностей пользователей.

156. Статистическое управление Эстонии было проинформировано об ограничениях в использовании регистров, основной причиной которых было отсутствие достаточной информации о метаданных, предоставленной владельцами регистров. Существовавшие метаданные были собраны просто для удовлетворения административных целей, для которых собирались данные, и часто не имели отношения к статистическому использованию данных. Часто наблюдались концептуальные несоответствия между определениями и классификацией, принятыми в регистре, и теми, которые необходимы для использования в переписи. Кроме того, охват базовой совокупности или наличие тематических переменных в регистрах не всегда соответствовали требованиям национальной переписи, особенно в тех случаях, когда переменные, относятся к самоопределяемым статусам.

157. Задача на 2014 год заключалась в согласовании пакета правовых и организационных мер по повышению качества, своевременности и охвата набора данных для переписи на основе регистров с учётом узких мест, указанных в методологическом отчёте.

158. Начиная с 2014 года Статистическое управление Эстонии активно участвовало в официальных обсуждениях с соответствующими органами с целью внесения необходимых изменений в правовые акты, регулирующие конкретные источники данных, необходимые для переписи. Национальным властям было предложено указать в своих законодательных предложениях, будет ли создан новый источник административных данных или изменён существующий. Также должен был быть прописан любой режим обмена данными. Положения о возможности для начала

или улучшения процесса сбора данных также были предусмотрены в законодательстве.

159. На Статистическое управление Эстонии была возложена ответственность за улучшение качества данных в регистрах. Соответственно, оно разработало дорожную карту на основе предложений экспертов и подготовило улучшенную бизнес-модель для содействия улучшения сотрудничества между административными регистрами. Статистическое управление Эстонии работало над планом действий до 2020 года, который включал различные задачи для владельцев источников данных. Наиболее актуальным из них было создание законодательной среды для добавления необходимых новых характеристик в регистры (таких как род занятий, отрасль и место работы) и для обновления этих характеристик в регистрах (включая регистры налогового управления, регистр планируемых рабочих мест, регистр предприятий и т. д.)

160. Следующей важной задачей было повышение точности регистрации по месту жительства, чтобы улучшить охват домохозяйств, институциональных групп населения и арендаторов. Статистическое управление Эстонии инициировало национальный проект Министерства внутренних дел по добавлению архивных данных о семьях и отношениях между членами семьи в Регистр населения. Это улучшит некоторые характеристики переписи (например, официальное брачное состояние и взаимоотношения внутри домохозяйства).

161. Поправки к законодательству, касающиеся иностранцев, помогли улучшить сбор данных о населении иностранного происхождения. Это позволило усовершенствовать процедуры регистрации для получения более полной информации о вновь прибывших (включая характеристики образования, семейного положения и отношений между членами семьи).

162. В общей сложности поставщикам источников данных было внесено около 20 различных предложений по повышению качества источников данных с использованием законодательной базы.

163. Для создания связанных данных с 2016 года 16 владельцами регистров были приняты некоторые основные правила, предусмотренные специальными постановлениями правительства:

    (a) Все данные в регистрах для лиц, предприятий и жилищ должны быть идентифицированы (с использованием уникальных кодов),

    (b) Адресные данные должны использоваться во всех регистрах в соответствии с установленным стандартом, и

    Метаданные должны быть доступны и обновлены.

164. Другой важный аспект, связанный с качеством используемых исходных данных, касается передачи данных. Необходимо иметь защищённую от ошибок и надёжную среду для передачи данных из разных регистров в НСС. В Эстонии такая среда под названием X-Road облегчает передачу больших объёмов данных между учреждениями или предоставление отдельным лицам их личных данных. Сбор данных для целей переписи был разрешён, согласно постановлению правительства,

через X-Road. Ранее владельцы данных использовали электронную почту или протокол передачи файлов (FTP) в виде закодированных файлов значений, разделённых запятыми (.csv), или файлов Microsoft Excel (.xlsx).

165. Стандарт качества был подготовлен для оценки источников данных. В стандарте качества числовые значения были зафиксированы для допустимых отклонений в переменных переписи и гиперкубах, когда были приняты во внимание следующие параметры качества данных:

    (a) Фактуальность (охват, концептуальные различия и т.д.),

    (b) Своевременность и периодичность (последняя дата обновления записей, задержки в поставках и т.д.), и

    (c) Точность: особенно переменных связей для оценки связности источника.

166. К 2020 году Статистическое управление Эстонии разработало 38 различных переменных, касающихся населения и жилищ, необходимых для текущей программы переписи ЕС, из 26 различных административных источников (см. Statistics Estonia, 2019).

Глава 5 Этап Данные

167. Эта глава представляет собой руководство по ключевым параметрам, инструментам и процессам качества данных для оценки административных данных на этапе подготовки данных. Это относится к обеспечению качества необработанных административных данных, предоставляемых в НСС, со ссылкой на ожидания и требования, установленные посредством оценки на основе метаданных на этапе Источника. Этапы Источник и Данные вместе обеспечивают общую оценку качества входных данных применительно к источнику административных данных (см. UNECE 2018, Глава 6).

168. Качество административных данных на этапе сбора данных оценивается по нескольким параметрам, включая удобочитаемость и достоверность, точность и надёжность, своевременность и пунктуальность, а также возможность связывания. Эти параметры рассмотрены в разделе 5.1), инструменты и индикаторы для их оценки или измерения — в разделе 5.2.

169. На этапе Данные можно установить базовый уровень качества предоставляемых отдельных наборов административных данных на основе правил редактирования и проверки. Их следует разрабатывать на основе известных свойств данных, собранных на этапе оценки источника, и требований к статистическому дизайну. Они также могут быть улучшены с течением времени. Для обеспечения такой базовой оценки может потребоваться определённый уровень обработки данных, включая связь с другими источниками. Эта обработка ограничивается тем, чтобы сделать данные пригодными для прохождения проверок и установить их качество по сравнению с другими источниками.

170. Результаты контроля качества на этапе Данные информируют НСС о любых необходимых исправлениях (посредством повторной поставки данных поставщиком). Они также информируют о необходимой обработке данных для использования при разработке переписи благодаря пониманию ошибки, которую необходимо учесть или скорректировать (см. Глава 6). Кроме того, они предоставляют информацию, необходимую для понимания последствий любых ошибок в источниках для окончательных результатов переписи (см. Глава 7).

5.1 Параметры качества данных

5.1.1 Гармонизация и проверка

171. Общая оценка доступности данных является частью контроля качества на этапе Источник (см. Глава 4). Однако для НСС крайне важно обеспечить, чтобы передаваемые файлы данных были в требуемом «читаемом» формате; базы данных структурированы таким образом, чтобы их могли принимать и считывать системы НСС. Если это не так, НСС может быть не в состоянии обработать переданные файлы данных.

172. Кроме того, после передачи данных в НСС необходимо обеспечить дальнейшее согласование и проверку данных, чтобы обеспечить их использование во всех сценариях использования переписи. Этап Данные предоставляет возможность проверить предоставленный набор данных по метаданным, собранным на этапе Источник, за базисный период и другие требования к данным для конкретных переменных. Для этого может потребоваться некоторая базовая гармонизация, например обеспечение того, чтобы все пропущенные значения кодировались одинаково. Механизмы согласования и проверки могут быть разработаны на основе предыдущего опыта работы с предварительными данными (см. раздел 3.5 о технико-экономическом обосновании). Со временем они могут быть улучшены, так как НСС регулярно пополняется из того же источника данных.

5.1.2 Точность и надёжность

173. Оценка точности входных данных должна проводиться для выявления ошибок **измерения** и **представления** в наборе административных данных (см. Глава 3), как описано в двухфазной модели жизненного цикла в Zhang's (2012) и принято в литературе по обеспечению качества (см. Stats NZ 2016 and Eurostat ESSnet KOMUSO 2019).

5.1.2.1 Ошибки представления

174. Ошибки представления (ошибки, относящиеся к целевым единицам) могут возникать, если данные неправильно сообщаются поставщику данных, например, из-за отсутствия регистрации или отложенной саморегистрации в административном регистре (например, при рождении, смерти или полном регистре населения). Некоторые записи данных могут не передаваться в НСС из-за технических проблем или передаваться с ошибками, если поставщик данных не обслуживает данные должным образом (что приводит к дублированию). Следует отметить, что ошибки представления могут вызывать ошибки измерения при изменении единицы статистического измерения. Например, отсутствие лица в административном регистре населения может привести к занижению значения переменной «размер домохозяйства». Для общей оценки охвата набора данных необходимо изучить как избыточный, так и недостаточный охват. Неполный охват может иметь особое значение в отношении «труднодоступных» групп населения (см. Вставка 4).

5.1.2.2 Ошибки измерения

175. Недопустимые или отсутствующие значения указывают на ошибки измерения (то есть ошибки в пределах переменных) и могут снизить точность исходных данных. Чтобы оценить, является ли значение недостоверным или пропущенным, важно изучить не только конкретные записи, но и распределения переменных для всех записей. Причины неточности могут быть техническими, например, ошибки в процессе передачи данных. Или неточность может быть систематической. Например, это может быть результатом неадекватного представления или обслуживания со стороны поставщика данных, особенно если переменная не имеет

административного значения для поставщика данных. Следовательно, переменная не регистрируется систематически (например, род занятий человека в Австрийском налоговом регистре, см. Eurostat ESSnet KOMUSO 2019). Отсутствующие значения также могут быть из-за того, что административный источник (или переменные в источнике) был создан совсем недавно[15].

5.1.2.3 Повторно предоставленные данные

176. В целом поставщик данных будет улучшать качество данных посредством регулярного обслуживания и обновления источника данных. Однако многие регистры могут подвергаться изменениям в структуре и/или содержании в результате внутренних административных требований и процессов. Эти изменения, в свою очередь, могут привести к снижению качества, особенно в отношении сопоставимости. Когда данные предоставляются периодически, возникает необходимость в дополнительном продольном контроле качества. Повторно предоставленные данные дают возможность оценить надёжность конкретных переменных, таких как близость первоначально предоставленных значений к повторно предоставленным значениям в наборе данных. Обычно предполагается, что более актуальные значения более точны.

5.1.3 Своевременность и пунктуальность

177. Важно, чтобы разница между исходной датой, к которой относятся данные, и датой их предоставления в НСС, была минимальной. Чем дольше задержка, тем менее актуальными становятся эти данные, даже если они все ещё могут быть точными (см. UNECE 2018, p. 15). Этот разрыв между исходной датой и получением НСС называется **своевременностью.**

178. **Пунктуальность —** это разница между ожидаемой датой доставки и фактической датой доставки. Это важно, поскольку НСС обычно несёт ответственность за подготовку результатов переписи в соответствии с согласованным графиком и не хочет, чтобы задержка в предоставлении данных для переписи повлияла на это.

5.1.4 Возможность связывания

179. Часто для определения качества набора данных требуется его привязка к другому набору данных для сравнения. Кроме того, если НСС использует более одного источника административных данных для своей переписи, необходимо иметь возможность увязывать данные из разных источников на уровне единицы/измерения (см. Глава 6). Степень успеха такой привязки повлияет как на точность, так и на актуальность входных данных.

[15] Регистр высшего образования в Венгрии содержит данные только о лицах, получивших высшее образование после 2005 года. Австрийский центральный регистр населения был создан в 2001 году, но признак официального семейного положения не был введён до 2006 года, что привело к отсутствию значений в регистре.

180. Общий уникальный идентификатор снижает усилия, необходимые для связывания данных, упрощая оценку полноты и точности сопоставления. В отсутствие такого идентификатора надёжно связать данные труднее. В этом случае возможна связь записей с использованием нескольких переменных, общих для единиц измерения в каждом источнике данных (обычно имя, дата рождения, пол и адрес) (см. Глава 6). В этом случае НСС необходимо гарантировать, что такие «сопоставимые» переменные имеют достаточное качество во всех источниках, в противном случае пострадает качество связи записей и, следовательно, надёжность данных. Даже если используются методы вероятностного сопоставления, качество переменных связи в конечном итоге будет влиять на риск ложных совпадений и ложных несовпадений на более поздних этапах производства (см. Eurostat 2014, раздел 3.5.2) (см. также Глава 6). Следует разработать расширенные проверки валидации для переменных, которые будут использоваться при связывании.

5.2 Инструменты и индикаторы

181. Следующие инструменты и индикаторы полезны для НСС при оценке качества исходных данных по параметрам, рассмотренным в разделе 5.1 выше. Это приложение инструментов и индикаторов поддерживает последовательную оценку из разных источников, чтобы решить, подходят ли административные данные для этой цели.

5.2.1 Гармонизация и проверка

182. Для обеспечения читабельности и достоверности передаваемых файлов данных крайне важно проводить технические проверки для установления соответствия файлов данных ожидаемому формату данных. Если эта проверка не удалась, НСС может потребовать повторно направить файлы данных в правильном формате. Перед проведением таких проверок данные должны пройти базовую очистку и/или гармонизацию, чтобы они были сопоставимы с другими источниками и были оптимизированы для использования со статистическим программным обеспечением НСС.

183. Примеры процессов **гармонизации** включают последовательное кодирование пропущенных значений, форматирование типов переменных даты и удаление или иное устранение дублирующихся записей из набора данных. Правила гармонизации данных должны быть согласованы НСС и последовательно применяться к данным, независимо от различных вариантов использования переписи, для которых они предназначены. Согласованные общеорганизационные стандарты гармонизации будут способствовать согласованности и сопоставимости. Процессы гармонизации данных и результаты валидации должны быть документально оформлены.

184. В предыдущих источниках были определены конкретные показатели, которые можно использовать для оценки **достоверности** (например, см. Daas et al. 2009; Eurostat ESSnet MIAD 2014; Cerroni, Di Bella and Galiè 2014). Это включает:

(a) Представленные переменные правильно названы и отформатированы (например, числовые, категориальные, переменные данные и т.д.),

(b) Указан правильный базовый период,

(c) Переменные соответствуют ожидаемому заранее заданному содержанию, определяемому с помощью метаданных, собранных на этапе создания источника (и, по возможности, с помощью технико-экономического обоснования), и

(d) Не обнаружено никаких неожиданных различий между текущими и предыдущими поставками источника данных в отношении количества записей и переменных (более подробно рассматриваются ниже).

185. Связывание записей из предоставленных данных с другим надёжным источником данных на уровне единицы обеспечивает инструмент для определения того, предоставлена ли правильная исходная дата (см. Asamer et al, 2016). Также можно проверить переменные с указанием даты, чтобы определить, совместимы ли они с базовой датой переписи. Правильная исходная дата важна, особенно для изменяемых переменных, таких как, например, текущий статус активности сезонных рабочих. По возможности любые такие несоответствия следует исправлять на этапе Процесс (см. Глава 6).

186. Настоящее Руководство не содержит инструкций относительно того, как следует применять правила согласования и проверки действительности, поскольку они должны разрабатываться с учётом производственных потребностей и конкретных запланированных видов использования административных данных в рамках проекта переписи.

5.2.2 Точность и надёжность

5.2.2.1 Ошибки представления

187. Для измерения точности предоставленных объектов или единиц измерения могут использоваться различные показатели, обеспечивающие оценить ошибку представления в данных (см. Daas et. al. 2009; Eurostat ESSnet MIAD 2014; Cerroni, Di Bella and Galiè 2014)[16]. Основные показатели включают:

(a) Общее количество объектов или статистических единиц (для сравнения с ожидаемым количеством),

[16] Как отмечается в глоссарии, в некоторых источниках (например, Zhang 2012) термин «объект» используется для обозначения единиц в наборе административных данных. Этот термин используется для различия единиц в административных данных и статистических единиц после того, как эти данные были каким-либо образом преобразованы. Это особенно актуально в тех случаях, когда единица (или «объект») в административном регистре отличается от целевой статистической единицы. Например, в налоговом регистре, где единицы годовой налоговой декларации (т. е. одно и то же лицо может сдать несколько деклараций за один год или несколько лет) конвертируются в отдельных «людей».

(b) Процент повторяющихся объектов или статистических единиц[17].

188. Дополнительные индикаторы, предложенные Cerroni, Di Bella, and Galiè (2014, p.128), включают:

(a) Процент «несовместимых» объектов или статистических единиц, то есть вовлечённых в нелогические отношения с другими совокупностями объектов или статистических единиц [18],

(b) Процент «сомнительных» объектов или статистических единиц, то есть рассматриваются как неправдоподобные, но необязательно неправильные связи с другими показателями, объектами [19].

189. Широкую оценку чрезмерного и недостаточного охвата данных можно произвести путём вычисления и сравнения общего количества объектов, а также перекрёстных таблиц частот/процентов по характеристикам (например, полу, возрасту, географии и т. д.) на агрегированном уровне между административным источником и другими/альтернативными источниками, взятыми в качестве справочных, или сравнительным **«золотым стандартом»** (например, полный базовый регистр [20] или традиционная перепись). Индикаторы, предложенные Cerroni, Di Bella, and Galiè (2014, p. 129), включают:

(a) Недостаточный охват:

(i) Процент объектов справочного источника, отсутствующих в предоставленном источнике.

(b) Избыточное покрытие:

(i) Процент объектов в источнике, не включённых в базовую совокупность, и / или

(ii) Процент объектов в источнике, не принадлежащих целевой группе НСС.

190. Вышеуказанные индикаторы основаны на двух ключевых предположениях. Во-первых, должен быть доступен подходящий «золотой стандарт» для расчёта избыточного и недостаточного покрытия. Например, умершие лица могут по-прежнему (неправильно) регистрироваться в центральном регистре населения страны, но могут быть идентифицированы как умершие в центральном регистре социального обеспечения. Во-вторых, должно быть ясно, какие объекты набора

[17] Процент идентифицированных дубликатов может быть только нижней границей из-за необнаруженных дубликатов. Если доля необнаруживаемых дубликатов слишком высока, показатель не будет точным.

[18] Примером такой ошибки является взрослый в наборе данных, который обозначает несколько домохозяйств в качестве основного места жительства (количество несовместимых единиц этого типа, делённое на общее количество единиц, будет рассчитано).

[19] Рекомендательное правило может быть определено для выявления появления сомнительных объектов в административном источнике. Например, количество сотрудников, работающих на более чем четырёх работодателей в течение того же периода, использовались для обнаружения сомнительных объектов в данных социального обеспечения Италии (см. Cerroni, Di Bella, and Galiè 2014, p. 138).

[20] В литературе базовые или базовые административные регистры часто отличаются от дополнительных регистров (например, см. Daas et al. 2009). Базовые или основные регистры предполагают наиболее полный охват целевого постоянного населения.

данных «золотого стандарта» должны быть включены для вычисления в рамках покрытия. Примером этого является то, что дети школьного возраста, включённые в базовый регистр, должны в значительной степени включаться в регистр зачисленных учеников.

191. Наконец, можно провести сравнения, чтобы определить процентную долю объектов, несогласованных в пределах и между источниками. Примером несовместимых объектов может быть ситуация, когда каждая строка в наборе административных данных представляет событие регистрации (например, посещение врача), которое включает в себя имя, адресный код, дату регистрации и, возможно, дату отмены регистрации. Два объекта, относящиеся к одному человеку, несовместимы, если период регистрации объектов по разным адресам совпадает. Процент несовместимых объектов указывает на ошибку. Однако, как отмечается, анализ на уровне объекта имеет свои ограничения, поскольку источники могут различаться на микроуровне, но приводить к аналогичным статистическим показателям, таким как средние значения, медианы и т.д. Анализ на уровне единиц «может не выявить такой статистической эквивалентности» (см. Zhang 2012, p.45). Кроме того, если сравнения на уровне единиц производятся между несколькими источниками, важно отметить возможное влияние смещения избирательности в процессе связывания на любые возникающие различия [21].

5.2.2.2 Ошибки измерения

192. **Статистические методы и показатели**, такие как частотные распределения, могут выявить неожиданные закономерности и резко отклоняющиеся значения. Например, перекрёстная таблица возраста и брачного состояния может привести к выявлению неправдоподобных случаев, таких как 5-летний ребёнок, состоящий в браке. Другие примеры включают сравнение даты рождения с датами других событий в немецком тематическом исследовании в разделе 5.4.1. и анализ согласованности адресных данных в польском тематическом исследовании в разделе 5.4.2. Наблюдаемые закономерности могут указывать на систематические ошибки измерения, как также показано в тематическом исследовании Германии (раздел 5.4.1). Обратите внимание, что если несоответствия выявлены и поставщик данных не может исправить такие проблемы, то могут потребоваться определённые правки (как часть этапа Процесс, Глава 6) .

193. Как и выше, предыдущая литература содержит **основные показатели** для измерения полноты измерений, представленных в наборах административных данных на агрегированном уровне (например, характерные переменные, такие как

[21] Смещение избирательности в связи относится к ситуациям, когда переменные или методы связи более или менее точны для определённых групп, особенно в отношении иерархического и вероятностного сопоставления. Например, иностранные имена могут быть написаны с ошибками с большей частотой, что приведёт к большему количеству пропущенных совпадений с использованием ключей сопоставления, которые включают имя. Кроме того, если в методе связывания используется фонетический алгоритм на родном языке для выявления совпадений между записями, в которых имена отдельных лиц были написаны по-разному (например, Steven and Stephen), это приведёт к менее точным совпадениям для тех, у кого нет совпадений для тех, у кого имена не на родном языке.

возраст, пол, этническая принадлежность и т.д.) (см. Daas et. al., 2009; Eurostat ESSnet MIAD 2014, Cerroni, Di Bella and Galiè 2014). Это включает:

(a) Количество и процент отсутствующих значений в ключевых переменных,

(b) Количество и процент значений, выходящих за пределы диапазона ключевых переменных (например, зарегистрированный возраст 120 лет),

(c) Количество и процент неправдоподобных значений (основанных, например, на перекрёстных таблицах различных переменных),

(d) Преобладание неожиданных частот, шаблонов или резко отклоняющихся значений на основе частотного/распределительного анализа ключевых переменных (совокупные сравнения с внешними источниками, а также экспертные знания также могут использоваться для выявления странностей данных), и

(e) Преобладание округления для основных представляющих интерес переменных (может быть обнаружено путём анализа распределений).

194. Степень согласованности предоставленных данных на агрегированном уровне, а именно то, что взаимосвязи между связанными переменными согласованы и не являются неправдоподобными, обеспечивает меру точности переменных. Однако для оценки согласованности на микроуровне следует **проводить расширенные проверки** достоверности связанных переменных в пределах предоставленного набора данных. Основываясь на предыдущей литературе, ключевые индикаторы включают:

(a) Процент объектов, комбинации значений переменных которых участвуют в нелогических отношениях,

(b) Процент объектов с сомнительными значениями переменных или объектов, чьи комбинации значений переменных входят в неправдоподобные, но не обязательно неверные отношения (т.е. резко отклоняющиеся значения),

(c) Процент объектов с пропущенными значениями ключевых переменных, которые имеют разные характеристики для завершения объектов, и

(d) Процент объектов со значениями, рассчитанными поставщиком данных для основных представляющих интерес переменных.

195. Аналогично оценке ошибки представления, эффективным способом оценки точности переменных, особенно при предварительном анализе данных и самом первом использовании данных, является **сравнение данных**; то есть входные данные проверяются путём сравнения с другими независимыми источниками, содержащими ту же переменную. Подходящие независимые источники для

сравнения могут включать национальное обследование (например, обследование рабочей силы) или предыдущую перепись (см. Asamer et. al., 2016).[22]

196. В литературе описаны более сложные методы оценки точности административных данных, когда административные данные связаны со сравнительным источником (который включает переменную/интересующую концепцию). Bakker (2012) использует модели структурных уравнений для оценки достоверности административных переменных с использованием данных обследований. Модель применяется к данным о возрасте, поле, уровне образования и заработной плате. Scholtus и Bakker (2013) также используют имитационное исследование для проверки устойчивости модели к дополнительным компонентам погрешности измерения, к неправильной спецификации модели измерения и к небольшому размеру выборки. Oberski et. al. (2017) применяют обобщённую модель с несколькими признаками и несколькими методами в рамках общей схемы для одновременной оценки качества административных данных и данных обследований. Структура позволяет как обследованиям, так и административным данным содержать случайные и систематические ошибки и, следовательно, не предполагает, что обследование не содержит ошибок, как при использовании других методов (см. Yucel and Zaslavsky 2005). Их подход учитывает общие черты административных данных, такие как дискретность, нелинейность, ненормальность и может улучшить другие используемые модели (например, модели структурных уравнений).

5.2.2.3 Повторно предоставленные данные

197. Административные данные могут быть предоставлены повторно, чтобы гарантировать, что НСС имеют доступ к самым последним и актуальным данным для использования в переписи. Как и в случае с исходными данными, первым шагом для оценки качества повторно предоставленных данных является выполнение макроуровневого сравнения основных ключевых показателей (таких как общее количество записей, количество пропущенных значений и т. д.) против того, что ожидалось получить. Для повторно поставленных данных сравнение с представленными ранее выявит любые неожиданные различия в наборах данных, которые могут указывать на проблему качества. Кроме того, долгосрочное сравнение данных, представленных в текущем и предыдущем периоде, важно для выявления возможных изменений качества, особенно с точки зрения охвата, полноты и возможности связывания.

[22] Следует отметить, что согласованные значения и перекрёстные таблицы, полученные с помощью различных источников и методологий (например, административных данных и данных обследований), позволяют предположить, что оба источника, вероятно, верны. Несогласованные значения оставляют открытым вопрос, какой результат является наиболее точным, т. е. наиболее близким к истинному значению совокупности. Это зависит от того, как будут даны ответы на вопросы обследования и как будет собираться административный источник, что ещё раз подчёркивает важность этапа Источник. Не всегда верно, что источник административных данных будет менее точным (например, см. литературу о получении государственных пособий). Необходим более сложный анализ, чтобы определить точность как административного, так и внешнего источника, чтобы оценить причину обнаруженных несоответствий.

198. Для ключевых переменных, которые, как ожидается, будут стабильными с течением времени, можно сравнить значения одной и той же единицы (например, человека) с течением времени, чтобы проверить наличие неожиданных изменений. Эти проверки легче выполнять для «инвариантных» переменных, таких как дата рождения или место рождения, а также для данных, для которых доступен уникальный ключ и который стабилен во времени. Даже для изменчивых переменных, таких как официальное семейное положение или высший уровень образования, такие проверки могут быть возможны в ограниченной форме. Продольные сравнения могут служить внутренней мерой **надёжности** данных, предоставляя такие индикаторы, как средние или медианные значения различий или относительных различий между новейшими и предыдущими источниками данных.

199. Если не существует ключевой переменной, которая была бы стабильной во времени, то распределение переменных можно использовать для сравнения периодов времени.

5.2.3 Своевременность и пунктуальность

200. Параметры **своевременности** и **пунктуальности**, как определено в разделе 5.1.3 могут быть определены путём сравнения контрольной даты, согласованной даты доставки и фактической даты доставки данных. Следующие индикаторы предложены Cerroni, Di Bella and Galiè (2014, стр.130):

(a) Своевременность

(i) Разница во времени (дни) = (Дата получения НСС) — (Дата окончания отчётного периода, за который источник данных отчитывается).

(ii) Разница во времени (дни) = (Дата получения пользователем) — (Дата окончания отчётного периода, за который отчитывается источник данных).

(b) Пунктуальность

(i) Разница во времени (дни) = (Дата получения НСС) — (Согласованная дата , как указано в соглашении).

(c) Общий временной лаг

(i) Общая разница во времени (дни) = (Прогнозируемая дата, на которую НСС сообщает, что источник может быть использован) — (Дата окончания отчётного периода, за который источник данных сообщает)[23].

[23] Эти индикаторы учитывают временной интервал между поступлением данных в НСС и их доступностью для производственных групп с учётом необходимости очистки, согласования, проверки, обеспечения аналитиков правильными разрешениями для доступа к данным и т. д.

(d) Задержка

(i) Свяжитесь с владельцем источника данных, чтобы сообщить информацию о задержках регистрации.

(ii) Разница во времени (дни) = (Дата фиксации изменения в источнике владельцем источника данных) — (Дата, когда произошло изменение в населении).

201. Индикатор задержки зависит от информации, которая может быть неизвестна или недоступна НСС. Однако, если доступна некоторая информация о том, когда данные для наблюдения были обновлены в источнике, этот базовый показатель можно рассчитать.

(e) Процент наблюдений, обновлённых в течение прошлого года, считая с даты предоставления в НСС.

202. Использование и интерпретация этого индикатора зависят от контекста, поскольку в некоторых обстоятельствах могут быть веские причины для отсутствия обновления, например, если не было соответствующего события, инициирующего изменение в реестре с момента последнего обновления для данной записи.

5.2.4 Возможность связывания

203. Часто источники административных данных связаны с другими источниками, будь то перепись или другие административные источники. Контроль качества переменных в каждом источнике, используемом в увязке, предоставляет общую информацию для разработки успешного процесса связывания, как описано в Главе 6.

204. Независимо от того, доступен ли уникальный ключ или переменная идентификатора или несколько переменных будут использоваться в комбинации для выявления совпадений в процессе связывания, эти индикаторы должны информировать выбор и оценку качества предоставленных переменных связывания, включая:

(a) Процент уникальных значений, которые могут быть вычислены либо при наличии единственной идентификационной переменной (например, личный идентификационный номер), либо при комбинации переменных, которые будут использоваться при связывании (например, совпадающий ключ, объединяющий возраст, дату рождения и адрес), и

(b) Преобладание смещённых распределений относительно показателей точности (как описано в предыдущих разделах, включая пропущенные значения, неправдоподобные значения и т.д.). Есть ли доказательства ошибок измерения в переменных связи, которые не являются случайными? Например, существует ли значительно большая доля значений, выходящих за пределы диапазона, или отсутствующих значений для ключевой переменной связи, такой как дата рождения, в определённых географических регионах?

205. Наконец, если переменные связи были предоставлены НСС в зашифрованной или «хешированной» форме. Хеширование — это практика, которая часто используется в информатике для защиты конфиденциальности отдельных лиц или объектов в данных. Оно включает в себя применение алгоритма к каждому фрагменту информации в исходных данных (например, имени) для создания строки символов, которая однозначно идентифицирует эту информацию и маскирует исходные данные. НСС должна проверить, что хеширование, выполняемое поставщиком, соответствует алгоритму хеширования, используемому в НСС. В противном случае будет невозможно связать предоставленные данные с другими источниками данных, что подорвёт актуальность данных. В Главе 6 представлены дополнительные сведения о связывании зашифрованных ключей.

5.3 Рекомендации для этапа Данные

(a) Как отмечалось, перед использованием источника административных данных в переписи рекомендуется, если не обязательно, провести по крайней мере хотя бы один тестовый прогон с реальными данными. Такой тест должен проводиться достаточно рано, чтобы позволить перенастроить технические инфраструктуру и процессы, чтобы гарантировать читаемость, гармонизацию и проверку данных.

(b) Убедитесь, что предоставленные данные соответствуют метаданным, собранным на этапе Источник, и что указана правильная исходная дата.

(c) Вычисление и мониторинг основных показателей предоставленных данных для определения возможных ошибок представления и измерения.

(d) Проверка согласованности связанных объектов и переменных в предоставленном наборе данных с помощью расширенных проверок достоверности.

(e) Используйте статистические показатели для выявления неожиданных закономерностей и резко отклоняющихся значений.

(f) Сравните общее количество записей и перекрёстных таблиц с независимыми сопоставимыми источниками, чтобы оценить точность.

(g) Убедитесь, что НСС может уточнить запросы данных у поставщика данных. Когда после предоставления данных возникают вопросы, касающиеся данных, должны быть созданы адекватные механизмы, обеспечивающие их решение.

(h) Для повышения качества вводимых данных и обеспечения согласованности предоставьте поставщику данных обратную связь о любых обнаруженных расхождениях (таких как несоответствия в наборе данных), по крайней мере, на агрегированном уровне, при условии, что соответствующие законы о защите данных позволяют это.

(i) Когда данные предоставляются периодически, возникает необходимость в дополнительных, продольных оценках качества.

(j) Определить своевременность и степень пунктуальности предоставления данных.

(k) Определить качество переменных связи, чтобы гарантировать наилучшие возможные результаты связи (см. Глава 6).

5.4 Тематические исследования

5.4.1 Германия: качество данных, полученных из локальных регистров населения для переписи 2021 года

5.4.1.1 Введение

206. Национальная перепись Германии 2022 года [24] представляет собой комбинированную перепись с использованием данных из нескольких источников. Данные из всех локальных регистров населения примерно 11000 муниципалитетов, которыми управляет около 5100 местных регистрационных бюро, являются основным источником данных, но другая информация (не относящаяся конкретно к данному тематическому исследованию) собирается из множества других официальных источников, таких как Федеральное картографическое агентство, Федеральное министерство обороны, Федеральное министерство иностранных дел и Федеральное министерство внутренних дел, строительства и общественной жизни. Всего для переписи 2022 года запланировано шесть предоставлений данных из местных регистров населения. Поскольку человек в Германии может уведомить регистрационный офис об изменении адреса, постфактум будут две разные даты доставки данных, на которых основан подсчёт населения: одна с контрольной датой, эквивалентной контрольной дате переписи, и одна доставка с контрольной датой через три месяца после контрольной даты переписи.

207. Это тематическое исследование сосредоточено только на качестве данных регистров населения Германии и проблемах, которые возникли во время предоставления этих данных в январе 2019 года. Данные за 2019 год имитировали самый большой набор данных из регистров населения, который должен быть предоставлен в контексте переписи 2022 года. Доставка данных в январе 2019 года представляла собой пробный запуск для оценки качества необработанных данных, тестирования передачи данных, оптимизации существующих методов обработки данных и тестирования передачи исторических записей данных. Обратите внимание, что некоторые критики этого подхода считали достаточным тест с анонимными данными или случайной выборкой.

[24] Перепись первоначально была запланирована на май 2021 года, но была перенесена на май 2022 года из-за пандемии Covid-19.

208. В тематическом исследовании внимание уделяется только проверке качества вводимых данных. Для статистических целей в Германии не существует уникального идентификатора, доступного для человека.

209. В целом набор данных включает всех лиц, которые были зарегистрированы по первому или второму месту жительства на базовую дату 13 января 2019 года. Данные включают исторические записи о недавних изменениях в регистрах, близких к контрольной дате.

210. После предыдущей национальной переписи 2011 года были приняты меры по повышению качества регистров населения Германии. Когда человек переезжает из одного муниципалитета в другой, регистрационные офисы двух муниципалитетов автоматически сообщают об этом изменении. Местные регистрационные офисы сообщают о любых изменениях в своём регистре населения в Центральную налоговую инспекцию, поскольку у каждого человека есть уникальный налоговый идентификатор, весьма вероятно, что количество дублирующих записей о первом месте жительства в данных за 2019 год сократилось с 2011 года. Однако эта тенденция находится все ещё находится в стадии изучения.

5.4.1.2 Удобочитаемость

211. В Германии был определён стандартизированный универсальный формат для передачи и доставки данных из местного регистра населения. Получатель (в данном случае это данные для переписи, Федеральное статистическое управление) принимает данные только в том случае, если они доставляются в этом формате. Это помогает улучшить качество вводимых данных.

212. По крайней мере четыре муниципалитета пытались передать некоторые переменные в формате, нарушающем общие правила форматирования. Следовательно, они не могли доставить затронутые записи данных. Следовательно, это привело к неполному предоставлению данных. Для последующего предоставления данных формат этих переменных был расширен, чтобы эта проблема больше не возникала.

5.4.1.3 Точность

213. Учитывая, что регистры населения ведутся на местном уровне, неудивительно, что точность данных варьируется в зависимости от муниципалитета. Следующие два примера иллюстрируют различия в точности данных.

214. В первом примере для более чем 40 муниципалитетов в одной или нескольких из трёх переменных «дата переезда по адресу», «дата переезда в муниципалитет» или «дата регистрации» около 75 процентов или более всех первых записей о первом месте жительства содержат одну и ту же дату. Можно предположить, что это была ошибка, сделанная во время связывания данных, вызванного консолидацией двух или более муниципалитетов. Такие аномалии данных могут иметь решающее значение для выявления повторяющихся записей о первом месте жительства.

215. Во втором примере лица, зарегистрированные примерно в 120 000 записях данных, имели либо одну, либо все даты для трёх переменных, более ранние, чем дата их рождения. В одном из регистров было 60% этих ошибочных записей.

216. Для повышения качества вводимых данных муниципалитеты получили отзывы об ошибках, обнаруженных в их данных на агрегированном уровне, и о проверках достоверности данных, требующих расширения. Впоследствии состоялся обмен с производителями программного обеспечения для реестра населения.

5.4.1.4 Полнота

217. Во время предоставления данных за 2019 год возник ряд технических проблем, которые также отрицательно повлияли на полноту предоставленных данных.

218. Из-за ошибки в программном обеспечении, которое муниципалитеты использовали для получения данных, примерно 1200 муниципалитетов передали файлы с недостающими записями данных. Эта ошибка была обнаружена только случайно. Для некоторых из этих муниципалитетов поставщик программного обеспечения, а также муниципалитет инициировали доставку данных из-за недопонимания. В некоторых муниципалитетах поставщик программного обеспечения имеет точную копию реестра со всеми своими данными. Сравнение этих двух доставок показало, что поставщик программного обеспечения не смог передать некоторые записи данных. Поставщик программного обеспечения должен был запланировать вторую поставку вместо первой. Данные, предоставленные муниципалитетами, были удалены. Следовательно, техническая инфраструктура должна блокировать предоставление данных, состоящих из данных, объединённых от разных отправителей.

219. Как правило, трудно определить, отсутствуют ли какие-либо записи, поскольку получатель может не иметь информации о точном количестве записей, которые необходимо получить. Получатель может только сравнить количество переданных записей данных о первых местах жительства в муниципалитете со своими собственными прогнозируемыми оценками численности населения. Однако нередко эти две цифры различаются на несколько процентных пунктов.

220. Некоторые муниципалитеты действительно передавали для каждой записи данных недостающие значения некоторых переменных. Это проявилось как неполный поиск данных из локальных баз данных. Например, переменные «самая последняя дата переезда в Германию» и «страна происхождения» (которые должны быть пустыми, если это Германия) были пустыми для всех записей данных примерно в 1200 муниципалитетах. Перед внесением данных в базу данных важно проверить, отсутствуют ли переменные во всех данных для всего муниципалитета из-за технических проблем.

5.4.1.5 Измерение, связанное со временем

221. Некоторые муниципалитеты смогли собрать свои данные только через несколько дней после контрольной даты. Лицо, которое сообщает о последующем уведомлении об отъезде в муниципалитет в течение промежуточного периода, не

учитывается. Чтобы уменьшить возможный ущерб от такой ошибки при доставке данных в будущем, крайне важно, чтобы муниципалитеты развили способность извлекать предыдущие варианты своих регистров.

5.4.1.6 Заключение

222. Технические проблемы во время и до доставки значительно снизили качество данных, полученных из регистрационных бюро местных муниципалитетов. Следовательно, пробный запуск переписи 2022 года в январе 2019 года был важен для оценки процедурных и технических недостатков. Тестовый запуск с анонимными данными или случайная выборка не позволили бы выявить большинство описанных недостатков. Время проведения теста более чем за год до предоставления данных переписи 2022 года дало достаточно времени для анализа и устранения ошибок и оптимизации возможностей обработки данных как на центральном, так и на местном уровне. Кроме того, муниципалитеты были проинформированы об ошибках данных на агрегированном уровне, поскольку Федеральному статистическому управлению запрещено по закону возвращать отдельные записи данных. Мы надеемся, что это поможет улучшить качество входных данных, получаемых из регистров населения.

5.4.2 Польша: Польская система качества переменных

5.4.2.1 Введение

223. Данные переписей населения в Польше собираются из нескольких источников, включая административные. Реестры и системы баз данных характеризуются широким разнообразием содержания и сложностью структуры в результате того, что они создаются для разных целей и управляются разными поставщиками данных. Соответственно, стандарты хранения, точности и методы записи, принятые в каждом случае, также различаются. Отсутствие единообразия существует не только между регистрами, но и с данными в одном регистре.

224. Качество данных из используемых административных источников влияет на качество результатов переписи. Адекватное качество вводимых данных является предпосылкой (хотя и не единственной) для получения правильных результатов переписи. При использовании административных источников (не только в контексте проведения переписи) важными шагами являются выявление и понимание проблем и ошибок в данных и исправление данных. Для обеспечения качества исходных данных особенно важен первый пункт.

225. После оценки возможности использования административных источников процесс управления качеством данных, собранных из административных источников в Польше, разделен на три части: ввод (эквивалентный этапам Источник и Данные в настоящем Руководстве), процесс и качество ввода. Процесс управления качеством находится под постоянным контролем. В Статистическом управлении Польши для этой цели используется Система переменного качества. Система переменного качества — это система для просмотра, анализа и представления данных из административных источников.

226. Сначала Система переменного качества проверяет полученные данные. Процесс включает в себя применение набора правил для оценки набора данных на предмет полноты, согласованности и правильного формата для использования в системе. Ключевым моментом является полнота и точность уникальных идентификаторов, предоставляемых в поставке данных, что крайне важно для обеспечения высококачественной интеграции данных. Отсутствующие или ошибочные значения в поле уникального идентификатора препятствуют эффективной интеграции записей в нескольких источниках данных. Данные, которые не соответствуют допущениям проверки, подлежат исправлению — процессу согласования для приведения данных в соответствие с ожидаемым стандартом.

227. После этапов проверки и исправления создаётся отчёт об улучшении качества данных для принятия решения о том, следует ли обращаться к поставщику данных для улучшения их качества для предоставления или завершения любой дополнительной обработки данных. Это позволяет Статистическому управлению Польши внимательно отслеживать все изменения, которые происходят в источниках административных данных, используемых в официальной статистике, и позволяет автоматизировать расчёт показателей качества как для входных, так и для выходных данных. В этом тематическом исследовании основное внимание уделяется обеспечению качества входных данных.

5.4.2.2 Точность и надёжность

228. Система переменного качества содержит результаты профилирования необработанных данных в Польше. Профилирование данных — это процедура, с помощью которой пользователь, помимо прочего, получает информацию о точности необработанных данных. Она предоставляет ряд статистических показателей:

    (a) Порядковый номер,

    (b) Тип данных,

    (c) Количество (количество записей),

    (d) Ненулевое количество, и

    (e) Длина данных.

229. Для числовых переменных:

    (a) Минимальное значение,

    (b) Максимальное значение,

    (c) Среднее значение, и

    (d) Медиана.

230. Для символьных переменных:

    (a) Количество шаблонов,

(b) Уникальное количество,

(c) Минимальная длина,

(d) Максимальная длина,

(e) Частотное распределение, и

(f) Распределение частот по шаблону.

231. В рамках Системы переменного качества проводится анализ согласованности адресных данных для проверки их точности и согласованности. Адрес состоит из следующих иерархических уровней территориального деления:

(a) Воеводство (или провинция, которых в Польше 16),

(b) Повет,

(c) Гмина,

(d) Населённый пункт, и

(e) Улица.

232. При отдельном рассмотрении значения отдельных полей адреса могут соответствовать стандарту, но не могут образовывать последовательную адресную строку, отображаемую в Национальном официальном реестре территориального деления страны (ТЕРИТА). Чтобы считать адрес действительным, правильных частей самих по себе недостаточно. Также должна быть соблюдена логическая структура: улица должна находиться в населённом пункте, населённый пункт в гмине, гмина в повете и повет в воеводстве. Только адреса, следующие этой структуре, считаются согласованными. Связность с улицей считается полной связностью, связность на уровне населённого пункта (совместимая последовательность воеводства, повета, гмина, населённого пункта) или гмина (совместимая последовательность воеводства, повета, гмина), нуждаются в улучшении или дополнении другими имеющимися данными. Что касается анализа согласованности, Система переменного качества генерирует следующие показатели качества:

(a) Сопоставимость словаря ТЕРИТА (количество),

(b) Изменение сопоставимости словаря ТЕРИТА (в процентах),

(c) Сопоставимость словаря преобразования (количество),

(d) Изменение сопоставимости словарей преобразования в результате различных этапов — ввода, вывода (в процентах), и

(e) Уровень согласованности адресных переменных (флаг).

233. Чтобы проверить полноту переменной, Система переменного качества генерирует следующие показатели качества для каждой переменной:

(a) Полнота (количество), и

(b) Изменение полноты (в процентах).

5.4.2.3 Своевременность и пунктуальность

234. Решающее значение имеет долгосрочное, эффективное и прозрачное сотрудничество с поставщиками административных данных. В Польше сбор данных для целей переписи поддерживается правовой базой, включающей Закон о статистике и Закон о переписи. Система переменного качества регистрирует информацию о контрольной дате данных и дате получения данных Статистическим управлением Польши.

235. Данные обычно собираются в конце года или в соответствии с датой соответствующего обследования. Данные для регулярной десятилетней переписи собираются как можно скорее в ходе её проведения, что даёт необходимое время для обработки данных. Чтобы обеспечить максимальную актуальность данных, сбор должен быть как можно ближе к контрольной дате переписи, или, если получение данных является непрерывным процессом, как можно ближе к исходной дате данных.

5.4.2.4 Возможность связывания

236. Полнота и точность имеют решающее значение для уникальных идентификаторов, таких как:

    (a) Номер PESEL[25] — широко используется в административных регистрах населения; число однозначно идентифицирует человека и позволяет различать много людей с одинаковыми именем и фамилией,

    (b) REGON: бизнес-идентификационный номер, и

    (c) NIP: идентификационный номер налогоплательщика.

237. Идентификаторы должны характеризоваться необходимым количеством знаков и соответствием контрольной цифры. Высокое качество идентификаторов имеет первостепенное значение при интеграции данных. Отсутствующие или ошибочные значения не позволяют идентифицировать одни и те же объекты в разных источниках. Система переменных показателей генерирует следующие показатели качества для идентификаторов:

    (a) Количество правильных идентификаторов (число), и

    (d) Изменение количества правильных идентификаторов (в процентах).

5.4.2.5 Заключение

238. В рамках методологической основы для улучшения качества ввода, обработки и вывода Система переменного качества является важным инструментом для контроля качества данных, обеспечения сопоставимости качества между различными поставщиками и отслеживания изменений качества с течением времени.

[25] Номер PESEL (универсальная электронная система регистрации населения - Powszechny Elektroniczny System Ewidencji Ludności) представляет собой 11-значный постоянный цифровой символ, который однозначно идентифицирует каждого человека, зарегистрированного в базе данных PESEL.

Глава 6 Этап Процесс

239. После получения административных данных и оценки качества НСС данные потребуют обработки, чтобы их можно было использовать в переписи. Административные данные необходимо будет интегрировать в проект переписи и необходимо будет решить любые вопросы качества (например, концептуальное несоответствие определениям и концепциям переписи, охват и ошибки измерения). Этап обработки настоящего Руководства содержит обзор ключевых процессов, используемых для интеграции административных данных в перепись, и связанных с этим проблем качества[26].

240. Обработка административных данных основывается на выводах, полученных на этапах Источник и Данные. Например, оценка связности административного источника показывает, как связаны данные. Понимание ошибки охвата даёт информацию для процессов интеграции данных для достижения охвата, необходимого для переписи. Оценка точности административных данных даст информацию на этапах редактирования и условного исчисления и обеспечит понимание в поддержку решений о том, как источники должны использоваться вместе для построения переменных переписи. Однако обработка может привести к дополнительной ошибке (систематической или случайной), что приведёт к смещению или дисперсии окончательных оценок. По этой причине важно, чтобы процессы были надлежащим образом протестированы и оценены. Ошибки необходимо устранять по всей цепочке статистического производства. В этой главе рассматриваются некоторые из наиболее распространённых процессов, необходимых для использования административных данных при переписи. Это:

    (a) Связь (связывание) записей,

    (b) Оценка ошибки охвата в статистических регистрах или административных данных при подсчёте единиц населения,

    (c) Устранение несоответствий в значениях элементов данных из разных источников, и

    (d) Редактирование и импутация.

241. Каждый из этих процессов более подробно описан в следующих разделах вместе с проблемами, связанными с каждым из них, способами оценки качества обработанных данных на основе доступной литературы и опыта разных стран.

[26] См. KUMUSO, Quality Framework for Multisource Statistics, 2019 WP1 (Система качества для статистики с несколькими источниками, 2019 WP1), где представлены индикаторы качества, меры и методы оценки качества процессов и результатов.

6.1 Связь записей

242. Почти каждый источник административных данных требует той или иной формы связи записей с другими источниками данных для проверки данных или для обеспечения адекватного охвата единиц и переменных переписи населения. Например, может потребоваться объединение двух или более источников данных для достижения лучшего охвата целевой группы, в том числе для корректировки потенциального превышения охвата (см. раздел 6.2). Аналогичным образом, увязка нескольких источников может потребоваться для предоставления полных и точных данных для переменных переписи (см. раздел 6.4).

243. Многие страны объединяют административные данные из нескольких источников для создания административных статистических регистров; они могут включать адресные регистры, регистры населения или регистры предприятий (см. UNECE 2018, Глава 8 и раздел 6.2 ниже). Даже страны, не имеющие статистических регистров, стремятся к максимальному использованию административных данных при производстве своей основной статистики населения, социальной и деловой статистики.

244. Это делает объединение записей одним из наиболее важных процессов использования административных данных в переписи. Таким образом, важно и необходимо оценить качество процесса связывания посредством оценки переменных или ключей связывания (как описано на этапах Источник и Данные) и посредством оценки самого процесса, как описано в следующих разделах настоящего документа в этой Главе.

245. Следует учитывать влияние ошибки связи (ошибки представления и измерения) на общую точность оценок населения и переписи (см. Daan Zult et. al, 2019). Например, пропущенные и неправильные ссылки могут привести к чрезмерному или недостаточному охвату переписной совокупности и могут привести к присвоению неправильного значения переменной переписи для данного лица или домохозяйства. Адресные данные часто требуют особого внимания, поскольку адреса могут использоваться как для увязки данных о человеке (например, в качестве ключа связи в сочетании с именем и датой рождения), так и для объединения отдельных лиц в домохозяйстве. Однако люди не всегда предупреждают поставщика данных о своём переезде. Зарегистрированный адрес также может быть не основным адресом проживания. Следовательно, точность адресных данных в административных источниках может быть низкой. Ошибка связи также может вносить систематическую ошибку в двойную систему оценки и привести к смещению (см. Abbott 2009).

246. Методы увязки данных обычно основываются на существовании общих уникальных ключей (или идентификаторов) во всех источниках, которые необходимо связать. Например, Польша использует уникальный идентификатор, номер PESEL (см. раздел 5.4.2.4.).

247. В отсутствии общих уникальных ключей другие общие идентифицирующие переменные, такие как адрес, имя, пол и дата рождения, могут использоваться для

связывания записей из нескольких источников. Хотя это более сложно и подвержено более высокому уровню ошибок, как указано ниже.

248. В некоторых случаях НСС может иметь доступ только к анонимным или «хешированным» идентификаторам в административных данных (см. раздел 5.2.4). Хеширование имеет некоторые важные последствия для качества связи данных (см. Shipsey and Plachta 2020 для описания методов связывания с анонимными данными, проблем и ограничений).

249. Методы связывания бывают двух основных типов: детерминированные, когда совпадения производятся на основе набора общих идентификаторов, и вероятностные, когда совпадения производятся на основе весов связывания на основе модели (см. Harron, Goldstein and Dibben 2015). Вероятностное сопоставление не требует, чтобы значения записей были идентичными между двумя записями, но полагается на сходство между записями. Ещё один метод связывания, который можно применить к несвязанным записям после применения детерминированных и вероятностных методов, — это канцелярская связь, которая включает в себя ручную проверку несвязанных записей. Офисная привязка невозможна, когда данные хешируются.

250. Ошибка связи может возникнуть из-за несвязанных записей, которые должны были быть связаны (также известные как «ложные отрицательные результаты»), и связанных записей, которые не должны были быть связаны (также известных как «ложные положительные результаты»).

251. Двумя распространёнными методами оценки качества связи являются:

    (a) Оценка количества ложноположительных и ложноотрицательных результатов с использованием офисного анализа выборочных данных связанных и несвязанных записей. Хотя офисная проверка может быть сделана только тогда, когда данные не хешируются. Если данные хешируются, НСС должно попытаться получить доступ к выборкам связанных и несвязанных записей в их исходном состоянии до хеширования, чтобы оценить связь.

    (b) Сравнение распределения характеристик связанных и несвязанных записей (например, возраст, пол и этническая принадлежность). Различия в характеристиках могут указывать на то, что систематическая ошибка была вызвана ошибкой связи. Это означает, что определённые типы записей (например, отдельные лица) не могут быть связаны, потому что их сложнее связать.

252. Оценка ошибки увязки с использованием описанных выше методов представлена в тематических исследованиях Великобритании и Новой Зеландии, см. разделы 6.7.1 и 6.7.3.

Вставка 9: Способы связи данных и оценки качества связи: межправительственный обзор Великобритании

Важность увязки административных данных для общественного блага (в том числе для переписи) широко признана и в результате в Соединённом Королевстве был проведён межправительственный обзор для разработки руководства по методам увязки данных, охватывающих обеспечение качества увязки. Обзор основывался на работе экспертов в правительстве, академических кругах, частном секторе и на международном уровне. Результатом стала серия статей, посвящённых будущему методов связывания данных; оценке качества в связывании данных; продольным связям (принципы проектирования и система общих ошибок); сохранению конфиденциальности; увязке с анонимными данными; процедуре повышения эффективности (см. ONS 2020).

6.2 Статистические регистры и методология «признаков жизни»

253. Как упоминалось в разделе 6.1, объединение данных из разных источников для использования в переписи становится все более распространенным явлением; связь записей играет важную роль в этом процессе. Двумя ключевыми параметрами качества, связанными с интеграцией данных из различных источников, являются охват и согласованность. Интеграция данных выполняется для оценки и, возможно, уменьшения ошибки охвата. Это также позволяет и требует оценки согласованности информации по источникам и с течением времени.

254. Одним из примеров интеграции данных для использования при переписи населения и жилищного фонда является построение статистических регистров населения. Связывая информацию из доступных источников на уровне записей, можно определить отдельных лиц или домохозяйства, которые проживают в стране, и их характеристики. Интегрированные данные из этих источников становятся статистическим регистром, а именно базой данных, которую можно использовать для дальнейшей обработки и анализа для получения результатов переписи (см. UNECE 2018, Глава 8).

255. Некоторые из ключевых процессов, участвующих в создании статистического регистра, включают:

(a) Определение источников данных, которые будут использоваться,

(b) Связывание источников (см. раздел 6.1),

(c) Разработка и применение набора правил для принятия решений о том, какие записи должны быть включены в окончательные оценки (раздел 6.2),

(d) Разрешение противоречивой информации (например, даты рождения или адреса) между связанными источниками (раздел 6.4), и

(e) Редактирование и импутация (раздел 6.5).

256. Соображения и показатели качества предложенные в Главе 4 и Главе 5 помогут определить источники данных, которые будут использоваться в статистическом регистре. В этом разделе основное внимание уделяется применению правил принятия решений и некоторым вопросам качества, связанным с этим процессом. В этом разделе также обсуждаются другие методы оценки охвата, которые могут использоваться в статистических регистрах наряду с правилами принятия решений.

257. Правила принятия решений или правила «деятельности» — это критерии включения, которые часто применяются при построении статистических регистров населения, чтобы гарантировать, что в окончательные оценки включаются только лица, отвечающие некоторым заранее определенным критериям обычного проживания. Этот процесс иногда называют методом «признаков жизни» и является широко используемым инструментом для уменьшения избыточного охвата в статистических регистрах (например, включение записей, которые не являются частью обычно проживающего населения).

258. Испания использует «признаки присутствия» из четырёх типов административных источников: налоговых файлов, базы данных социального обеспечения, источников, связанных с рынком труда, и центрального реестра иностранных граждан. Лица, достигшие порогового уровня признаков присутствия, считаются «активными» и включаются в подсчёт населения, тогда как все остальные, называемые «неактивными», не включаются в подсчёт (см. Vega Valle et al 2020 и тематическое исследование Испании, раздел 6.7.2 для получения более подробной информации).

259. В Соединённом Королевстве аналогичный подход используется для решения, какие записи из выбранных административных источников следует включить в свои оценки численности населения на основе административных данных (см. ONS 2019). В более ранней версии административных оценок численности населения запись включалась в оценки численности населения, если она присутствовала в двух выбранных административных источниках. В последующей версии административных оценок численности населения строгие критерии для включения применялись к каждому источнику отдельно (где записи включались только в том случае, если были признаки активности в течение последних 12 месяцев) и правило включения записей, только если они присутствовали в двух источниках, было удалено (с использованием связывания данных, чтобы исключить дублирование записей, появившихся в нескольких источниках). Последующая версия административных оценок численности населения была направлена на дальнейшее сокращение избыточного охвата, обнаруженного в предыдущей версии, за счёт увеличения неполного охвата (записи, которые отсутствуют в оценках численности населения). Ожидалось, что неполный охват будет решён с помощью обследования охвата в сочетании с двойной системой оценки.

260. Успех метода «признаков жизни» зависит от наличия хороших индикаторов признаков активности в отдельных или комбинированных источниках административных данных. Применение метода обычно включает в себя некоторые предположения, которые определяют, кто считается активным, а кто нет. НСС должны чётко понимать эти предположения и, по возможности, предоставлять соответствующие подтверждающие доказательства. В частности, выбор признаков индикаторов деятельности (или правил принятия решений) должен быть обоснован оценкой качества на этапах Источник и Данные (см. Глава 4 и Глава 5), включая консультации с поставщиками данных, перекрёстную проверку источников с течением времени и экспертные знания.

261. Как уже упоминалось в случае Соединённого Королевства, применение методов «признаков жизни» можно комбинировать с другими методами для оценки и учёта ошибок охвата в статистических регистрах. Один из них — провести независимое от статистического регистра обследование и использовать объединённую информацию из обследования и реестра для оценки количества записей, пропущенных в реестре (или обследовании), чтобы улучшить окончательные оценки. Это аналогичный подход к проведению постпереписного обследования после традиционной переписи и применению методов двойной системы оценки (также известный как метод двойного ввода) для оценки уровня неполного охвата при переписи (см. Abbott et. al, 2020).

262. Избыточный охват статистических регистров также можно оценить, связав регистр с обследованием с помощью подхода, называемого «зависимое интервьюирование», которое направлено на проверку административных записей на местах. Этот подход использовался в Италии и в некоторых других странах (например, в Израиле), которые успешно перешли на переписи на основе административных данных. Однако не все страны могут проводить зависимое интервью из этических соображений и конфиденциальности (см. Глава 4). Brown et. al. (2020) обсуждают двух- и многосистемные методы оценки для устранения ошибок покрытия.

263. В Италии зависимое интервьюирование с выборкой домохозяйств, взятой из базового регистра населения (также известного как База данных регистрации отдельных лиц) и методологии «признаков жизни» (с использованием других административных источников) используется в сочетании для оценки и корректировки ошибки избыточного охвата в базовом регистре населения. Кроме того, для корректировки на неполный охват используется выборочный опрос адресов, взятый из базового статистического реестра адресов. В результате этого процесса оценки совокупности получаются путём применения поправочных коэффициентов для ошибок недостаточного и избыточного охвата к отдельным данным в базовом регистре населения. Тематическое исследование Италии в разделе 6.7.4 содержит подробную информацию о полной методологии.

6.3 Подсчёт единиц населения: модели на основе административных данных

264. Что касается создания статистических регистров, административные данные могут использоваться для подсчёта единиц населения (например, отдельных лиц, домохозяйств или занимаемых адресов), для поддержки или дополнения сбора данных переписи на местах. Этот подход использовался как в Новой Зеландии для решения проблемы недостаточного охвата в переписи населения и жилищного фонда 2018 года, так и в Соединённых Штатах Америки для повышения эффективности их полевых операции в связи с отсутствием ответов на местах (мониторинг участия в переписи).

265. Подход включает привязку интегрированных источников административных данных к набору данных «золотого стандарта» (в данном случае традиционной переписи) для построения моделей для оценки качества административных данных и определения условий, при которых административные данные используются для переписи. Подход позволяет частично использовать информацию административных записей там, где они считаются наиболее надёжными.

Вставка 10: Определение занятости адресов (полевая операция переписи населения США)

Целью переписи населения США было использование административных данных для определения свободных и несуществующих адресов и подсчёта занятых адресов в рамках операции по отслеживанию отсутствия ответов. В тех случаях, когда административные данные предсказывали (на основе определённых ограничений), что адрес не занят, количество полевых контактов могло быть сокращено, что снизило затраты и повысило эффективность. Прогностические модели были разработаны на основе наблюдаемых в 2010 году взаимосвязей между результатами переписи (в качестве «золотого стандарта»), государственными административными записями и данными третьих сторон. Затем эффективность моделей была проверена в рамках переписей 2015 и 2016 годов, а также с помощью ретроспективной оценки с использованием переписи 2010 года. Использовались различные административные источники (правительственные и коммерческие), включая данные о налогах, социальном обеспечении, здравоохранении, жилье и почтовых службах.

Производительность моделей использовалась для определения пороговых значений для защиты от недостаточного охвата (когда адреса неверно классифицируются как свободные по административной модели) с целью минимизировать рабочую нагрузку по отслеживанию отсутствия ответов. Особое внимание было уделено работе модели в разных географических регионах с разной концентрацией групп населения (например, испаноязычного, неиспаноязычного и чернокожего населения). Это привело к дальнейшему развитию стратегии защиты от неправильной классификации адресов как незанятых (см. Section 3 Administrative Records Modelling Update for the US Census Scientific Advisory Committee, 2017), которая содержит подробную информацию о проведённой контроле качества.

Вставка 12: Прямой подсчёт (перепись Новой Зеландии 2018 года)

Перепись населения Новой Зеландии 2018 года использовала административные данные для подсчёта людей, которые были пропущены при полевом сборе данных. Данные переписи (предыдущие и текущие) были связаны с административными записями для построения моделей, которые использовались для оценки качества административных данных и определения того, как и когда они будут использоваться для включения людей, семей и домохозяйств в перепись.

Основной целью административной переписи было выявление неполного охвата при переписи. Метод связывания был разработан для минимизации ложных срабатываний (т. е. для минимизации количества административных записей, ошибочно исключённых из набора данных переписи из-за того, что они были неправильно связаны). Кроме того, в процессе окончательной переписи была сделана корректировка для уменьшения количества ложных отрицательных результатов (т. е. административных записей, которые были неправильно не связаны и, таким образом, ошибочно добавлены в набор данных переписи, что привело к чрезмерному охвату).

Административные записи, которые были отобраны для включения после процесса увязки, были разделены на те, которые должны были быть включены в жилища (с созданными семьями и домохозяйствами), и те, которые были включены только в небольшом географическом районе (без отношения к жилью и без создания семьи или домохозяйства). Это решение было принято с помощью статистических моделей, которые были специально разработаны для прогнозирования надёжности административных данных для представления домашних хозяйств. Модели (в которых использовались данные переписи) были оценены с использованием анализа кривой рабочей характеристики приемника.

Для оценки эффективности этого подхода было проведено определение моделей охвата переписи после включения административных данных. Недавно разработанный эталонный показатель численности населения двойной системы оценки обеспечила наиболее подходящую оценку реальной численности обычного постоянного населения переписи, доступной на этом этапе переписи. Были получены распределения населения по возрасту, полу, этнической принадлежности и географии. Распределение показало, что набор данных переписи 2018 года в значительной степени соответствовал эталонному показателю и в большинстве случаев включение административных записей в файл значительно уменьшило (но не решило все проблемы) неполный охват (см. Stats NZ, 2019a). Эти ориентировочные результаты придали уверенность в новых методах, когда были опубликованы данные переписи. См. Тематическое исследование 6.7.3 для получения более подробной информации об этом подходе.

6.4 Разрешение конфликтов / выбор между источниками

266. Как упоминалось в разделе 6.2, при объединении административных данных для создания статистических регистров могут возникать несоответствия в значениях ключевых переменных из разных источников. Например, после принятия решения о том, какие административные записи (лица) включать в обычно проживающее население, если адрес лица в двух или более источниках отличается (например, из-за задержек в уведомлении об изменении адреса, административной обработке задержки или второго/нескольких домов), тогда НСС может потребоваться решить, по какому адресу следует включить человека. Противоречивая (или множественная) адресная информация и любое связанное с этим решение могут привести к недостаточному охвату в одних областях и чрезмерному охвату в других. На совокупном (например, национальном) уровне это не может быть проблемой, потому что человек может быть посчитан только один раз. Однако на уровне небольшой площади это может иметь значение, если два адреса находятся в двух разных областях, так как это может вызвать чрезмерный охват в одной области и недостаточный охват в другой.

267. Abbott et. al. (2020) описывает три подхода к выбору источников в контексте конфликта адресов:

    (a) Удалить запись из совокупности,

    (b) Разделить запись между разными местоположениями в соответствии с весовыми коэффициентами (например, половина, если есть два местоположения),

    (c) Выбрать, какой источник с наибольшей вероятностью будет обновлён на основе характеристик отдельных лиц или административных переменных. Этот подход может также использовать дополнительные источники данных при появлении одного и того же человека.

268. Первый подход увеличивает недоучёт в оценках численности населения. Два других подхода могут дать приемлемые оценки численности населения на агрегированном уровне, но могут привести к значительным смещениям из-за ошибок охвата и увязки в оценках на более низких уровнях дезагрегирования, таких как возраст и пол. Эти подходы (b) и (c) были протестированы в Великобритании в рамках разработки оценок численности населения на основе административных данных; дальнейшие исследования продолжаются (см. ONS 2016, Раздел 6).

269. Аналогичные подходы к измерению качества атрибутов в статистических регистрах, когда один и тот же показатель доступен в разных источниках, использовались в Австрии и Испании. В австрийской полной переписи на основе регистров комбинированный показатель качества рассчитывается с использованием теории Dempster-Shafer, также известной как теория функций убеждений и обобщение байесовской теории субъективной вероятности (теория

Dempster-Shafer: см. Shafer 1992). Сравнение с внешним источником проводится для оценки соответствующих статистических правил (см. Statistics Austria 2019).

270. В регистре населения Испании отсутствует информация об официальном семейном положении отдельных лиц. Для оценки официального семейного положения используются несколько регистров для получения полной информации (см. Argüeso, 2019), включая данные налогового агентства, регистр актов гражданского состояния, базу данных социального обеспечения и центральный регистр иностранных граждан. Поскольку человек может появляться в нескольких источниках данных с противоречивой информацией, применяются правила принятия решений для определения наиболее правдоподобного значения. Правила принятия решений применяются к каждой записи о человеке, после чего может быть присвоено значение официального семейного положения. Если случаи остаются не назначенными, значение рассчитывается на основе возраста, данных прошлых переписей и количества членов домохозяйства. Результаты, полученные с помощью этого метода, многообещающие; дальнейшие исследования продолжаются.

271. Подводя итог, можно сказать, что методы выбора между источниками, когда одни и те же атрибуты доступны в разных источниках, обычно основываются на правилах принятия решений, как в методах «признаков жизни» (см. раздел 6.2). НСС следует рассмотреть и протестировать различные подходы в соответствии с их конкретными потребностями переписи и на основе качественной информации, полученной на этапах Источник и Данные (см. Главу 4 и Главу 5).

6.5 Редактирование и импутация

272. Контроль качества на этапах Источник и Данные (Глава 4 и Глава 5) проинформирует о том, требуют ли административные данные редактирования (для неправильных/неправдоподобных значений) и/или импутации (для отсутствующих значений). Редактирование и условное исчисление могут потребоваться как для одного источника, так и для интегрированных данных.

273. В австрийской переписи на основе регистров используются семь «базовых регистров» для предоставления базовой информации по соответствующим темам переписи. Эти базовые регистры дополняются восемью «регистрами сравнения», которые используются в основном для целей проверки. То есть выбирается один базовый регистр для предоставления значения определённой переменной переписи, а регистры сравнения используются для подтверждения этих значений (см. Schnetzer et al 2015). Однако в некоторых случаях регистры сравнения также предоставляют данные, которые полностью или частично отсутствуют в базовых регистрах. Комбинированный набор данных из базового и сравнительного регистров, называемый Центральной базой данных, дополняется условными обозначениями для элементов, не отвечающих и неправдоподобных значений, создавая Окончательную базу данных. Качество оценивается повсюду, от метаданных и контактов с поставщиками данных (например, для понимания надёжности данных для предполагаемой цели и того, как поставщики данных

поступали с отсутствующими или неправдоподобными значениями) до проверок итоговых данных на основе регистров путём сравнения с независимым внешним источником (см. Statistics Austria, 2019).

274. В австрийской переписи на основе регистров применялись три метода условного исчисления: детерминированное редактирование, статистическая оценка (включая «горячую» регрессию и логистическую регрессию) и статистическое сопоставление. Например, для внесения официального семейного положения использовалось условное исчисление. Это включает объединение людей в группы (колоды) по атрибутам, которые сильно коррелируют с целевой переменной. Предельное распределение целевой переменной в колоде (с существующими значениями) используется для вменения целевой переменной в соответствующую колоду (с пропущенными значениями). В окончательных данных контроля качества, в окончательной базе данных, рассчитывается показатель качества для применения.

275. Schnetzer et. al. (2015) предлагают использовать коэффициенты классификации для оценки различных моделей импутации. Это включает применение метода импутации к уже существующим данным и сравнение результатов процесса вменения с истинными значениями для каждой единицы. Коэффициент классификации выводится как отношение между совпадающими значениями и числами всех сравниваемых единиц. Уровень классификации похож на коэффициент попадания и может использоваться для категориальных и числовых значений.

276. Chambers (2001, цитируется в Schnetzer et. al. 2015) описывает пять связанных с качеством свойств, которым должны соответствовать условные обозначения:

(a) **Точность прогнозирования —** рассчитанные значения должны быть как можно более "близкими" к истинным значениям,

(b) **Точность ранжирования** — процесс вменения должен сохранять порядок вменённых значений данных (для атрибутов, которые являются как минимум порядковыми),

(c) **Точность распределения** — процедура вменения должна сохранять распределение истинных значений данных,

(d) **Точность оценки** — моменты распределения истинных значений более низкого порядка должны воспроизводиться процессом вменения (для скалярных атрибутов),

(e) **Правдоподобие импутации —** процедура вменения должна приводить к вменённым значениям, которые являются правдоподобными.

6.6 Рекомендации для этапа Процесс

(a) Как упоминалось в Главе 5, точность и полнота переменных связи должны быть оценены до увязки данных из разных источников.

(b) Для метода увязки следует оценивать и сопоставлять общие коэффициенты увязки и показатели ложноположительных/ ложноотрицательных результатов. Следует заранее определить пороговые значения ошибки связи и рассмотреть компромисс между минимизацией ложноположительных или ложноотрицательных связей.

(c) Ошибка охвата в статистическом регистре населения должна быть оценена и учтена. Это может быть достигнуто с помощью сравнений с другими источниками, включая методологию «признаков жизни» и / или с помощью обследований (которые могут быть специально разработаны для корректировки с учётом избыточного или недостаточного охвата).

(d) Выбор показателей «признаков активности» (или правил принятия решений) при построении статистических регистров должен основываться на оценке качества на этапах Источник и Данные, а также должны быть проверены различные методы (и лежащие в их основе предположения).

(e) Модели могут использоваться как для оценки качества административных данных для целей подсчёта единиц населения (на основе набора данных, который принимается за «истинный»), так и для определения того, когда и как использовать административные данные для этой цели.

(f) При выборе между источниками, когда в них доступны одни и те же атрибуты, следует рассмотреть и протестировать различные подходы в соответствии с конкретными потребностями переписи и на основе качественной информации, полученной на этапах Источник и Данные.

(g) Качество редактирования и условного исчисления следует оценивать как по отдельным источникам, так и по интегрированным данным; необходимо оценить различные модели условного исчисления.

6.7 Тематические исследования

6.7.1 Соединённое Королевство: измерение качества связи при замене переменной переписи административными данными

277. Десятилетняя перепись населения Англии и Уэльса проводится НСС для подсчёта населения и регистрации характеристик населения и домохозяйств. НСС планирует заменить вопрос переписи о «количестве комнат» для переписи 2021 года с использованием административных данных. Некоторые элементы этой работы ещё предстоит завершить; однако качество связи было проверено с использованием данных переписи 2011 года (см. ONS 2020b).

278. В ходе переписи 2011 года были заданы два вопроса: «Сколько комнат доступно только для этого домохозяйства?» и «Сколько из них спален?» Ответы используются для определения уровня заполняемости путём сравнения доступных комнат/спален с «необходимыми комнатами / спальнями». Отрицательный рейтинг заполняемости означает, что свободных комнат / спален меньше, чем

требуется домохозяйству (переполненность). Эта информация позволяет центральным и местным органам власти разрабатывать соответствующую жилищную политику и планировать предоставление жилья в будущем. Качество ответов на вопрос о количестве комнат переписи 2011 года, измеренное обследованием качества переписи 2011 года, было значительно ниже (67 процентов), чем на вопрос о количестве спален (91 процент). Это, а также мотивация уменьшить нагрузку на респондентов, побудило НСС рассматривать административные данные, в частности данные Агентства оценки, в качестве альтернативного способа удовлетворения информационных потребностей. Агентство оценки — государственное агентство. Оно отвечало за объединение всей домашней собственности в Англии и Уэльсе для уплаты местного муниципального налога с тех пор, как этот налог был впервые введён в начале 1990-х годов.

6.7.1.1 Качество связи

279. Уникальный справочный номер объекта недвижимости, уникальный буквенно-цифровой идентификатор для каждого пространственного адреса в Великобритании, использовался для связи данных Агентства оценки и данных переписи. Чтобы гарантировать высокое качество связи, уникальность этой переменной была оценена как в данных Агентства оценки, так и в данных переписи до связи двух источников. В данных переписи ответы с неуникальным (дублирующимся) справочным номером объекта недвижимости обрабатывались так, как если бы в них отсутствовали значения количества комнат, так как эти случаи нельзя с уверенностью связать с административными данными. Повторяющийся уникальный справочный номер объекта недвижимости в данных переписи имели место, если двум или более разным адресам переписи было присвоен один и то же уникальный справочный номер объекта недвижимости. Примером этого может быть случай, когда квартире на первом этаже и квартире на втором этаже назначается один и тот же уникальный справочный номер объекта недвижимости, но разные идентификаторы адреса переписи. Вероятно, это связано с ошибкой сопоставления, когда записи адреса переписи связаны с адресным фреймом, поскольку метод включает элемент вероятностного сопоставления.

280. В данных Агентства оценки были исключены записи с неуникальной переменной связи (1 процент). Это похоже на дублирование данных переписи 2011 года. Другие записи данных Агентства оценки «очищаются» перед связыванием данных (3 процента), включая удаление записей, которым GeoPlace не присвоило уникальный справочный номер объекта недвижимости (0,2 процента), и записей с дублирующими регистрационными номерами (0,3 процента).

281. Также измерялась степень увязки ответов переписи 2011 года с административными данными по уникальному справочному номеру объекта недвижимости. Высокий коэффициент связи был важен, потому что несвязанные записи переписи с записями Агентства оценки приводят к отсутствию значений в переменной Агентства оценки «количество комнат». Исключая полностью условно исчисленные домохозяйства (без ответов) и неуникальные записи, 96

процентов домохозяйств переписи 2011 года были увязаны с данными о собственности Агентства оценки.

282. Важное допущение предполагаемого подхода к редактированию и условному исчислению (а именно, метода импутации на основе доноров) заключается в том, что набор доноров является как можно более репрезентативным и представительным для получателей. Поэтому до редактирования и условного исчисления распределения несвязанной и связанной переписи с записями Агентства оценки сравнивались по ключевым переменным домохозяйств, таким как тип жилья и количество обычных жителей. Аналогичное сравнение было выполнено для отсутствующих данных о количестве комнат в связанных и несвязанных наборах данных. Хотя наблюдались некоторые различия в распределении, количество доступных записей доноров, в которых количество комнат не отсутствовало, было достаточным, при разбивке по одной переменной домохозяйства и по местности количество доноров всегда превышало количество доноров с пропущенными значениями. Процессы редактирования и условного исчисления были протестированы на десяти местных органах власти с наибольшим процентом недостающего количества комнат.

6.7.2 Испания: использование административных данных при создании базы данных переписи населения для переписи населения Испании 2021 года: метод «признаков жизни»

283. Перепись населения 2021 года в Испании рассматривается как база данных микроданных, содержащая примерно 47 миллионов записей, по одной на каждого жителя. Для переписи населения, административные записи содержат огромное количество соответствующей информации, несмотря на то, что они собираются властями для целей, не связанных с подсчётом населения. Административные источники объединяются для создания статистического регистра населения для определения того, кто проживает в стране, и для оценок численности населения.

284. Основная структура подсчёта населения основана на Padrón, испанском регистре населения, в котором регистрируются все жители каждого муниципалитета Испании. Физические лица должны зарегистрироваться в муниципалитете, в котором они проживают. Поскольку регистрация даёт много преимуществ, резиденты обычно регистрируются.

6.7.2.1 Метод «признаков жизни»

285. При использовании Padrón для целей переписи должен быть составлен соответствующий статистический регистр. После получения исходной базы данных Padrón, на которую ссылаются 1 января каждого года, проводится статистическая обработка. Некоторые предположения сделаны в отношении присутствия в Испании иностранных граждан, срок регистрации которых истёк или скоро истечёт. Кроме того, данные о численности населения статистически корректируются для обеспечения соответствия определению «обычно проживающего» с применением концепции проживания в течение двенадцати месяцев. Короче говоря, численность населения получается из Padrón, но это не

совсем результат подсчёта зарегистрированного населения, поскольку некоторые лица исключаются, а другие добавляются.

286. Из всего регистра народонаселения примерно 1,7 процента лиц исключены из подсчёта населения, в то время как примерно 0,15 процента добавляются и включаются в подсчёт населения.

287. Для определения лиц, обычно проживающих в стране, применяется метод «признаков жизни». Все люди анализируются в доступных источниках административных данных и их передвижения фиксируются в Padrón в течение месяцев, следующих за контрольной датой (см. Vega Valle et al 2020). Четыре ключевых административных источника, используемых для оценки методом «признака жизни», следующие:

(a) Налоговые органы и местные налоговые файлы,

(b) База данных социального страхования, состоящая из физических лиц со страховкой и бенефициаров (работников и пенсионеров),

(c) Источники, связанные с рынком труда, включая:

(i) База данных Национальной службы по безработице, содержащая досье безработных лиц, ищущих работу,

(ii) Реестры принадлежности к социальному обеспечению с информацией о принадлежности занятого населения,

(iii) База данных государственной помощи, содержащая информацию о получателях пособий.

(d) База данных Центрального реестра иностранных граждан предоставляет дополнительную информацию об иностранных гражданах, проживающих в Испании, включая дату подачи заявления на получение вида на жительство, лицензии или отказа в виде на жительство, даты истечения срока действия и проверки места жительства.

288. С помощью метода «признаков жизни» лица, достигшие порогового значения сигналов присутствия в административных данных, будут идентифицированы как «активные» и будут включены в подсчёт населения. Лица, не соответствующие пороговому значению, будут отображаться как «неактивные» и не будут включены. Эти «признаки жизни» из административных данных также могут быть собраны на индивидуальном уровне и на уровне домохозяйства; доступна информация о том, сколько членов семьи являются «активными».

289. Как для испанских, так и для иностранных граждан изменение адреса в Padrón рассматривается в месяцы, следующие за отчётной датой. Есть определённые изменения адреса, которые требуют прямого вмешательства человека. Также может быть проверка места жительства, сделанная муниципалитетом, которая даёт высокую вероятность того, что человек проживает в Испании на контрольную дату. Другие изменения являются хорошими индикаторами того, что человек не

проживает в Испании на отчётную дату. Их можно использовать для идентификации лиц, которые «обычно проживают» в стране.

290. Для несовершеннолетних признаком присутствия является присутствие взрослого в том же доме. Несовершеннолетние, не отвечающие этому требованию, исключаются из подсчёта населения. Возможность использования информации о студентах, обучающихся на официальных курсах, в настоящее время анализируется.

6.7.3 Новая Зеландия: процесс обеспечения качества при включении административного учёта в перепись населения Новой Зеландии 2018 года

291. Впервые набор данных переписи населения Новой Зеландии 2018 года включал административные записи для прямого подсчёта людей, пропущенных при полевом сборе данных переписи, что заменило использование условно исчисленных записей в предыдущих переписях. Эти административные переписи производятся на основе постоянного населения, полученного на основе административных данных, которые уже были оценены на предмет качества исходных данных, и ограничения качества известны (см. Gibb et al 2016; Stats NZ 2017). Административные переписи добавляются к данным переписи только в том случае, если люди находились в Новой Зеландии в ночь переписи и не ответили на перепись (см. Stats NZ 2019a). В этом тематическом исследовании основное внимание уделяется тому, как измеряется и оценивается точность процессов увязки и статистического моделирования.

292. Методология административного учёта была разработана для получения окончательного набора данных переписи с максимально возможным охватом целевого населения переписи. Мы были больше всего озабочены устранением потенциального избыточного охвата из-за использования административных записей как на национальном, так и на местном уровне, и ожидали, что это приведёт к некоторому остающемуся неполному охвату. Процессы увязки были разработаны для обеспечения того, чтобы административные записи добавлялись только для людей, которые ещё не ответили на перепись. Статистические модели были разработаны для управления известными ограничениями качества административного постоянного населения.

293. На самом высоком уровне процесс включения административных записей в данные переписи 2018 года включал увязку ответов переписи с административными данными, выбор административных записей для включения в жилища (с созданными семьями и домохозяйствами), и эти записи включались только в небольшом географическом районе (без связи с жильём и без создания семьи или домохозяйства). На каждом этапе процесса мы оценивали качество и решали, приемлема ли данная методология.

294. Связь между ответами переписи и административным населением была достигнута с использованием полностью автоматизированного процесса вероятностной привязки, предназначенного для минимизации ложноположительных связей (см. Stats NZ, 2019b). Качество связи оценивается

путём оценки количества ложноположительных и ложноотрицательных результатов. Ложноположительная оценка была получена путём ручной проверки небольшой выборки связанных записей. Ложноотрицательная оценка была основана на подходе, разработанном Choi (2019), в котором мы оценивали пропущенные совпадения из подмножества переписных листов, которые отвечали критериям включения в административные данные с высокой степенью достоверности (поэтому мы должны иметь возможность сопоставить записи). Общий показатель достигнутых ссылок был высоким (97,7%), при этом ложноположительные ссылки оценивались в 0,6 +/- 0,3%, а ложноотрицательные совпадения оценивались в 1,21% (см. Stats NZ, 2019c). Высокий уровень связи в сочетании с низким коэффициентом ошибок вселили уверенность в приемлемом качестве связи. В основном беспокоили ложноотрицательные совпадения и возможность их влияния на точность, увеличивая охват данных переписи 2018 года, поэтому поправку на эти ложноотрицательные совпадения были включены позже в процессе административного подсчёта.

295. Методология, используемая для распределения административных записей по жилищам и последующего перехода к небольшим географическим районам, предназначена для уравновешивания ограничений качества административных данных с требованиями к качеству данных переписи 2018 года (см. Stats NZ, 2019a). Как указывалось ранее, главным критерием качества является точность. Чтобы оценить качество административных данных для включения в перепись, были разработаны статистические модели (с использованием данных текущей и предыдущей переписи) для прогнозирования надёжности административных данных для представления всего домохозяйства (см. Gath & Bycroft 2018; Stats NZ 2019d). Использовались данные переписи для тестирования и оценки моделей (предполагая, что домохозяйства, ответившие на перепись, представляют истинные данные). Для каждого административного домохозяйства была создана модельная оценка, показывающая, насколько надёжны административные данные для всего домохозяйства в конкретном жилом помещении. Граничная оценка модели определяла, какие из не ответивших административных домохозяйств были добавлены к данным переписи. Модель оценивалась с помощью анализа кривой рабочей характеристики получателя и анализа показателей производительности, таких как чувствительность, специфичность и точность (см. Stats NZ 2019d), по ряду пороговых значений оценки модели. Мы увидели средние и высокие баллы по критерию чувствительности (доля включённых нами правильно административных домохозяйств) по всему диапазону отсечённых баллов, что дало нам уверенность в том, что мы смогли правильно идентифицировать большинство высококачественных административных домохозяйств. Напротив, мы увидели большую вариабельность в показателе специфичности (доля исключённых неправильных административных домохозяйств), что указывает на то, что мы также могли включить некоторые административные домохозяйства без правильного состава домохозяйств.

296. С оставшейся административной совокупностью, чтобы гарантировать, что мы не представили людей, которых не следует включать, мы сначала скорректировали потенциальный избыточный охват (с использованием строгого подхода метода

«признаков жизни»), а затем скорректировали с учётом дублирования, вызванного пропущенными связями между возвращёнными формами переписи и административными данными. Модель, подобная той, которая использовалась для включения домохозяйств, была применена для прогнозирования вероятности того, что административный сетчатый блок [27] отражает истинное обычное местожительство человека; были включены люди с баллами выше порогового значения.

297. Оценка качества в значительной степени включала определение того, где установить пороговые баллы модели с учётом актуальности, точности, согласованности, интерпретируемости методологии и полученных данных. Пороговое значение для включения административных записей в жилища было установлено как баланс между строгими критериями получения тех же людей в домохозяйстве, которых мы наблюдаем при переписи, и включением административных домохозяйств, которые отражают аналогичные модели состава домохозяйств, что и перепись, даже если мы не можем гарантировать, что все члены семьи одинаковы. Ограничение включения административных записей в небольших географических районах снова представляет собой компромисс; между максимальным использованием административных данных для улучшения национальных демографических подсчётов и минимизацией числа лиц, неправильно перечисленных в районе.

298. Описанные оценки качества имеют несколько ограничений из-за субъективности суждений, статистических допущений и проблем с базовыми административными данными. Оценка ошибочной связи ложноположительных ссылок зависела от качества суждений, сделанных рецензентами, в то время как оценка ложноотрицательных ссылок основывалась на предположении, что записи, используемые при оценке, являются репрезентативными для тех, кто не соответствует критериям. Оценки моделирования также были ограничены субъективностью в установлении подходящего порогового балла модели, надёжностью исходных предположений, таких как данные ответов переписи, представляли истину (которая простиралась до предположения «нет» в случае отсутствия ответа домохозяйства), а также отсутствием доступной информации для определения неверных административных данных. Дальнейшая работа по обеспечению качества процессов будет включать дальнейшую методологическую разработку, тестирование предположений и изучение альтернативных инструментов обеспечения качества для этих процессов.

[27] Сетчатый блок - это наименьшая географическая единица, по которой Статистическое управление Новой Зеландии сообщает статистические данные. Это определённая географическая область, размер которой варьируется от части городского квартала до больших участков сельской местности. Сетчатые блоки являются смежными, что означает, что каждый сетчатый блок граничит с другим, образуя сеть, охватывающую всю Новую Зеландию (включая побережья и бухты).

6.7.4 Италия: совместное использование данных обследований и регистров для подсчёта численности постоянного населения скользящей переписи населения Италии[28]

6.7.4.1 От переписи от двери до двери до скользящей переписи населения

299. Скользящая перепись населения и жилого фонда была разработана на основе программы модернизации Итальянского национального статистического института (Istat'), в соответствии с которой Интегрированная система статистических регистров размещается внутри статистического производства. Роль полевых исследований в этой системе заключается в поддержке регистров в широком смысле для оценки их качества и в добавлении недостающей, неполной или недостаточно качественной информации.

300. Перепись 2011 года, хотя и проводилась с использованием регистров, по-прежнему была традиционной переписью с исчерпывающим полевым сбором данных[29]. Скользящая перепись населения и жилого фонда основана на обратной связи между перечислением полей и регистрами, где данные регистров дополняются сбором полевых данных.

301. В основе Скользящей переписи населения и жилого фонда лежит Базовый регистр населения, основным административным источником данных которого являются регистры местного населения итальянских муниципалитетов. Вместе со статистическим базовым регистром адресов и тематическими регистрами по образованию и занятости он обеспечивает основу для производства данных переписи населения, в то время как специальные обследования используются для измерения ошибок охвата в Базовом регистре населения и сбора данных для переменных, которые недоступны (или доступны только частично) из регистров.

302. Два отдельных выборочных обследования (территориальное обследование участков и списочное обследование) проводятся ежегодно в саморепрезентативных муниципалитетах[30] и каждые четыре года в соответствии со схемой ротации в несаморепрезентативных муниципалитетах, в общей сложности ежегодно около 1,4 миллиона домашних хозяйств (из которых 450 000 домохозяйств участвуют в территориальном обследовании, а 950 000 домохозяйств — в обследовании по списку).

---

[28] ISTAT (2020) Nota tecnica sulla produzione dei dati del Censimento Permanente: la stima della popolazione residente per sesso, età cittadinanza, grado di istruzione e condizione professionale per gli anni 2018 e 2019: Dalla rilevazione "porta a porta" al Censimento permanente [Техническая записка о производстве данных постоянной переписи: оценка постоянного населения по полу, возрасту, гражданству, уровню образования и профессиональному статусу на 2018 и 2019 годы: от обследования "от двери до двери" до непрерывной переписи]. Рим: ISTAT. Доступна на https://www.istat.it/it/files//2020/12/NOTA-TECNICA-CENSIPOP.pdf

[29] Муниципальные регистры населения использовались для ведения полевой переписи, то есть в качестве переписных списков для рассылки вопросников, в то время как другие административные источники, интегрированные в дополнительный список вспомогательных источников, использовались для исправления неполного охвата списком, то есть для подсчёта людей, обычно проживающих, но ещё не зарегистрированных.

[30] Муниципалитетами саморепрезентативными являются муниципалитеты с населением > 17 800 человек плюс муниципалитеты, которые не подлежат ротации в выборке для Обследования рабочей силы. Все остальные являются несаморепрезентативными.

303. В территориальном обследовании выборка адресов и / или счётных участков (в зависимости от качества адресов в муниципалитете), взятых из регистров местного населения итальянских муниципалитетов, изучается «вслепую» (как в традиционных переписях) для подсчёта каждого домохозяйства.

304. Обследование по списку, основанное на выборке домохозяйств, взятых из Базового регистра населения, проводится с использованием метода смешанного режима (CAWI, CAPI, CATI). Первый этап состоит только из «спонтанных ответов», тогда как на втором этапе счётчиками также проводятся полевые наблюдения за не ответившими домохозяйствами. Для каждого не ответившего домохозяйства предварительно закодированный «результат» регистрируется в системе мониторинга обследования в конце работы на местах.

305. В обоих обследованиях используется один и тот же вопросник (с той лишь разницей, что список членов домохозяйства предварительно заполняется данными Базового регистра населения в обследовании списка) и включает все переменные переписи (включая те, которые имеются в регистрах) для проверки качества и охвата данных, уже имеющихся в регистрах, по сравнению с данными, собранными в ходе обследований.

6.7.4.2 Комбинированное использование данных регистра и обследования для оценки и исправления ошибок охвата Базового регистра населения

306. С целью подсчёта населения данные обследований используются для исправления данных в Базовом регистре населения в рамках модели оценки двойной системы, направленной на оценку ошибок охвата регистра. В традиционной переписи постпереписной источник часто используется для измерения недоучёта (при этом постпереписной источник является второй «фиксацией», а сама перепись является первой «фиксацией»). В Непрерывной переписи Базовый регистр населения представляет собой первую «фиксацию», в то время как ежегодные выборочные обследования и «административный метод «признаков жизни» представляют вторую «фиксацию». Кроме того, в отличие от типичного постпереписного источника, нацеленного на измерение недостаточного охвата, в конструкции Непрерывной переписи населения, вторая «фиксация» направлена на измерение и корректировку как недостаточного, так и чрезмерного охвата в Базовом регистре населения.

307. В полевых условиях вторая «фиксация» является двоякой: обследование участков используется для измерения ошибки недостаточного охвата Базового регистра населения, а обследование по списку используется вместе с информацией об «административных «признаках жизни», полученной из Интегрированной базы административных данных для измерения ошибки избыточного охвата в Базовом регистре населения. В результате этого процесса подсчёт населения, наконец, получается путём применения поправочных коэффициентов для ошибок недостаточного и избыточного охвата к отдельным лицам в Базовом регистре населения.

308. Благодаря связи с Базовым регистром населения, исследование участков позволяет статистической службе оценить количество людей, обычно проживающих в

муниципалитете, которые не включены в регистр населения. Аналогичным образом, благодаря связи с Базовым регистром населения, обзор списка позволяет Итальянскому национальному статистическому институту оценить количество лиц, включённых в регистр, которые больше не проживают в муниципалитете. С этой целью не ответившие домохозяйства классифицируются в соответствии с их «статусом охвата» на основе результатов, зарегистрированных в системе мониторинга обследования.

309. Однако, поскольку на само обследование могут повлиять ошибки неполного охвата или неспособность охватить всех обычно проживающих в муниципалитете лиц, перед расчётом уровня охвата предпринимается следующий шаг. Внутри подмножества лиц с «потенциальным превышением охвата» (лиц, все ещё присутствующих в муниципалитете согласно Базового регистра населения и не обнаруженных на местах), проводится разделение на основе «признаков жизни» в муниципалитете, зарегистрированное в Интегрированном архиве административных данных. Таким образом, домохозяйства, не ответившие на опрос в Списке, «выздоравливают», если они демонстрируют убедительные доказательства (т.е. не менее 8 месяцев) в том же муниципалитете, где они зарегистрированы в Базовом регистре населения. В то время как лица, не имеющие такого в регистре «признаков жизни» в муниципалитете, подтверждаются как превышающие охват регистра. Рассматриваемыми «признаками жизни» являются: государственные служащие, частные служащие или самозанятые; получение пенсии по возрасту; посещение образовательного учреждения (в том числе дошкольного) или университета; получение пособия по безработице или основного дохода; или являясь финансово зависимым членом семьи лица с вескими доказательствами регистра «признаков жизни».

310. Поправочные коэффициенты, применяемые к отдельным лицам в Базовом регистре населения, получаются с помощью следующих шагов.

(a) Расчёт необработанного невзвешенного показателя неполного охвата по каждому профилю[31] как отношение между вновь зарегистрированными лицами (т. е. лицами, которых не ожидают в соответствии с Базовым регистром населения), и общим числом перечисленных лиц:

$$p_{ij,under} = \frac{\text{Вновь зарегестрированные лица}_{ij}}{\text{Общее число перечисленных лиц}_{ij}}$$

(b) Расчёт необработанного невзвешенного показателя избыточного охвата для каждого профиля как отношение между лицами, ожидаемыми согласно Базовому регистру населения и не обнаруженными в обследовании (или не «восстановленными» согласно Интегрированному архиву административных данных), и в знаменателе те же лица плюс лица, ожидаемые согласно Базовому регистру населения и перечисленные в опросе (или «восстановленные» согласно Интегрированного архива административных данных:

[31] Все люди, которые имеют одинаковый профиль в муниципалитете, то есть одно и то же гражданство («итальянское» или «иностранное»), получают одинаковое значение корректора.

$$p_{ij,over} = \frac{\text{Ожидаемые и не обнаруженные}_{ij}}{\text{Ожидаемые и не обнаруженные}_{ij} + \text{Ожидаемые согласно регистра}_{ij}}$$

(c) Расчёт корректора необработанного покрытия:

$$corr_{ij} = \frac{1 - p_{ij,\text{чрезмерное покрытие}}}{1 - p_{ij,\text{недостаточное покрытие}}}$$

(d) Расчёт прямых и косвенных оценок — прямые оценки, откалиброванные для чрезмерного покрытия и неполного охвата для каждого профиля, сначала рассчитываются для выбранных муниципалитетов. Процесс калибровки ограничивает веса выборки обследования известными общими значениями совокупности, полученными из Базового регистра населения для каждого профиля. Затем с использованием моделей оценки малых территорий рассчитываются косвенные оценки, чтобы уменьшить вариативность прямых оценок для выбранных муниципалитетов и получить оценки для муниципальных образований, не вошедших в выборку.

(e) Расчёт среднего корректора на 2018-2019 годы [32] – для каждого муниципалитета и отдельно для избыточного и недостаточного охвата рассчитывается среднее значение корректоров 2018 и 2019 годов и взвешивается с соответствующими демографическими параметрами. Оценка среднего корректора 2018-2019 годов — это соотношение между средневзвешенными оценками корректора избыточного охвата и корректора недостаточного охвата.

6.7.4.3 Подсчёт населения на основе поправок на охват Базового регистра населения

311. В конце процесса каждой записи в Базовом регистре населения присваивается «вес», который «корректирует» ошибки охвата регистра, оценённые для данного муниципалитета. Вес, применяемый к регистровым записям, будет равен единице, если на регистр базового населения для данного муниципалитета не влияют ни чрезмерный, ни недостаточный охват (или если две ошибки компенсируют друг друга).

312. Если предполагаемый неполный охват Базового регистра населения больше, чем предполагаемый избыточный охват, корректор, применяемый к каждой записи в регистре, будет больше единицы, и общая совокупность будет больше, чем количество записей в Базовом регистре населения.

313. И наоборот, если оценочный неполный охват в Базовом регистре населения ниже, чем предполагаемый избыточный охват, корректор, применяемый к каждой записи регистра, будет меньше единицы, и общая совокупность будет меньше, чем количество записей в Базовом регистре населения.

[32] Из-за недостаточной стабильности оценок в период с 2018 по 2019 год для каждого муниципалитета был принят корректор средней численности населения данных за 2018 и 2019 годы.

314. После проверки подсчёта населения данные, собранные как в территориальном обследовании участков, так и в списочном обследовании, используются в сочетании с данными Базового регистра населения и данными из тематических регистров по занятости и образованию с использованием прогнозных статистических моделей для получения данных об образовании, гражданстве и статусе рабочей силы.

Глава 7 Этап Итоговые материалы

315. В этой главе содержится руководство по параметрам качества, некоторым ключевым инструментам и процессам, используемым для оценки измерения качества результатов переписи, когда оценки производятся путём интеграции источников административных данных в проект переписи (см. UNECE 2018, Глава 9). В разделе 7.1 описаны параметры качества выходных данных, по которым должна проводиться оценка, а в разделе 7.2 подробно описаны дополнительные инструменты и процессы, которые можно использовать для оценки качества по параметрам.

316. Хотя измерение качества продукции выходит за рамки качества источников самих источников, целью является получение высококачественных оценок с использованием административных данных. Следовательно, это Руководства не будет полными без учёта качества выпускаемой продукции. В то же время необходимо подчеркнуть, что все предыдущие этапы повышения качества влияют на качество результатов. В случае комбинированной переписи или методологии переписи, полностью основанной на административных данных, проект переписи, основанный на строгой оценке качества на этапах Источник, Данные и Процесс, в конечном итоге приведёт к высококачественным результатам (см. Индикаторы качества, меры и методы оценки качества результатов).

317. Измерение качества итоговых данных не может быть сведено к оценке общей неопределённости оценки (измерение точности); скорее оно должно включать оценку по всем другим параметрам качества продукции. Внедрение административных данных, вероятно, приведёт к выигрышам в одних измерениях и потерям в других. Достижение правильного баланса по параметрам качества — ключ к наилучшему удовлетворению потребностей пользователей.

7.1 Параметры качества результатов

7.1.1 Актуальность

318. Актуальность означает степень, в которой результаты переписи удовлетворяют потребности пользователей с точки зрения охвата и содержания. Данные актуальны, когда они касаются вопросов, которые больше всего волнуют пользователей данных. Это измерение может потребовать от НСС корректировки направления своих программ с течением времени, по мере необходимости. Оценка актуальности субъективна, потому что она часто зависит от различных потребностей пользователей. Таким образом, задача программы переписи состоит в том, чтобы сбалансировать любые противоречащие друг другу требования пользователей и максимально приблизиться к удовлетворению наиболее важных потребностей в рамках ресурсов и других ограничений (см. UNECE 2015). В разделе 7.1.6 представлены подробные сведения об удовлетворении потребностей в использовании и балансировании параметров качества.

319. Для оценки актуальности могут использоваться различные инструменты и подходы, включая опросы потребностей пользователей, консультации и опросы удовлетворённости пользователей, путём встраивания механизмов обратной связи с пользователями в процесс переписи и анализа использования данных переписи (см. UNECE 2018, стр. 28).

7.1.2 Точность и надёжность

320. Точность статистической информации — это степень, в которой информация правильно описывает явления, для измерения которых она предназначена. Проще говоря, **точность** — это близость между оценкой и неизвестным истинным значением. Обычно она характеризуется ошибкой в статистических оценках и традиционно подразделяется на систематическую ошибку и дисперсию. В контексте переписи дисперсия применяется в ситуациях, когда часть вопросника используется для выборки лиц или домохозяйств, когда обрабатывается только выборка записей или может быть введена на этапах обработки (например, вероятностное вменение и увязка — см. Главу 6). Точность также может быть описана с точки зрения погрешности измерения и представления.

321. **Надёжность** — это степень близости первоначальных оценок к последующим оценочным значениям (понятие указано в Единой статистической системе вместе с точностью; однако оно также связано с сопоставимостью — см. ниже). Административные данные, по своей природе, могут подвергаться повышению точности с течением времени (например, охват может улучшиться по мере того, как становятся доступными отложенные регистрации и отмены регистрации, а также может улучшаться качество измерений). Следовательно, НСС может использовать «новые» данные для улучшения своей статистики переписи, пересматривая предыдущие оценки. Однако это должно быть сбалансировано с потребностями пользователей в отношении исправлений. Методы оценки точности результатов переписи представлены в разделах 7.2.1 и 7.2.2.

7.1.3 Своевременность

322. **Своевременность** означает промежуток времени между периодом, к которому относятся данные переписи (например, критической датой (днём) переписи), и датой публикации данных. Комбинированная перепись или перепись на основе регистров часто позволяет производить оценки переписи более своевременно и чаще, чем традиционная перепись каждые десять лет — действительно, это одно из самых приветствуемых преимуществ преобразования переписи. Учитывая это, своевременность оценок, которые могут быть произведены, должна быть ключевым фактором качества и по возможности следует вносить улучшения в этот аспект. Своевременность самих данных является важным фактором, определяющим своевременность вывода, таким образом связываясь с этапами качества, обсуждёнными в предыдущих главах. Часто приходится идти на компромисс между своевременностью и точностью. Может случиться так, что разные пользователи данных будут иметь разные взгляды на баланс между ними и

поэтому они могут не иметь одинакового мнения о влиянии повышения своевременности на точность (см. раздел 7.1.6).

323. В литературе можно найти несколько простых показателей своевременности. Количественные показатели могут применяться для измерения временного лага для конечных результатов, например между сбором данных, увязкой данных и публикацией статистических данных. Например, общая своевременность может быть рассчитана как время от конца отчётного периода до получения административных данных, делённое на время от конца отчётного периода до даты публикации, умноженное на 100% (см. Eurostat ESSnet KOMUSO 2016; Eurostat 2013; Eurostat 2014; UNECE 2018).

7.1.4 Согласованность и сопоставимость

324. Система качества Единой статистической системы (см. Eurostat 2019) определяет **согласованность** и **сопоставимость** как достаточность статистических данных, которые должны быть надёжно объединены различными способами и для различных целей, а также степень, в которой различия между статистическими данными могут быть отнесены к различиям между истинными значениями статистических характеристик. Рамки качества Единой статистической системы и Рекомендации Конференции европейских статистиков для переписей населения и жилищного фонда 2020 года (см. UNECE 2018) расширяют определение, включив в него «степень, в которой информация переписи может быть успешно объединена с другой статистической информацией в рамках широкой аналитической основы». Сопоставимость можно рассматривать как особый случай согласованности, где **согласованность** — это степень, в которой данные, полученные из разных источников или методов, но относящиеся к одной и той же теме, схожи, в то время как **сопоставимость** — это степень, в которой данные могут быть сопоставлены по странам, регионам, субпопуляциям и времени.

325. Измерение степени внутренней и внешней согласованности и сопоставимости оценок, полученных с использованием административных данных, является важнейшим аспектом качества результатов для всех типов переписей, включая те, в которых используются административные данные. Такие оценки должны согласовываться с известными характеристиками населения в долгосрочном плане, по географическим регионам и характеристикам населения (см. раздел 7.2.2). Промежуточные итоги должны правильно суммироваться с общими итогами. Кроме того, важно оценить, насколько интегрированные статистические данные переписи сопоставимы на международном уровне и сообщить об этом пользователям данных.

7.1.5 Доступность и ясность

326. **Доступность** обычно определяется как лёгкость, с которой пользователи данных могут получать доступ к статистическим данным и понимать их. **Ясность** связана с доступностью любой дополнительной информации или метаданных, которые могут потребоваться, чтобы помочь пользователю данных интерпретировать и

понимать сопровождающие опубликованные данные. Понятие «ясность» по сути то же самое, что «интерпретируемость». В разделе 7.2.6 представлены подробные сведения об отчётах о качестве и метаданных, которые должны быть доступны и понятны пользователям данных.

7.1.6 Удовлетворение потребностей пользователей и баланс качества

327. Независимо от того, используются ли административные данные в статистическом производстве, при оценке общего качества получаемых оценок необходимо учитывать все вышеуказанные параметры качества. Это включает не только параметр точности — аспект, о котором чаще всего сообщают в связи с методологией обследования, — но также и остальные параметры качества. В контексте переписи общее качество оценок заключается в установлении баланса по параметрам качества, который наилучшим образом отвечает потребностям пользователей данных переписи. Для этого необходимо не только консультироваться с пользователями данных на протяжении всего процесса разработки переписи, но также предоставлять им доступ к общей информации и конкретным метаданным, которые им необходимы для оценки решений по качеству и обратной связи по оценке качества, проводимые НСС. Качественная отчётность и качественные метаданные имеют важное значение (см. раздел 7.2.6). Кроме того, постоянное улучшение качества исходных данных и процессов обеспечит повышение качества итоговых данных. Первому будет способствовать внедрение необходимых механизмов обратной связи с поставщиками (раздел 7.2.3), а второму — проведение независимой экспертизы методов (разделы 7.2.4 и 7.2.5).

7.2 Дополнительные инструменты и процессы

7.2.1 Оценка точности оценок численности населения (ошибка охвата)

328. В некоторых странах общая точность оценок численности населения при переписи традиционно определялась на основе Двойной системы оценки, которая предполагает проведение традиционной переписи (т. е. проведение «инвентаризационной» переписи в определённый момент времени), за которой следовало крупное обследование охвата после переписи (также однажды) и полагается на использование методов перехвата для оценки недостаточного и избыточного охвата (см. O'Hare 2019). Затем эти оценки могут быть скорректированы на основе административных данных о смертности, рождении и миграционных потоках за каждый год между десятилетними переписями. Наряду с этим, в некоторых случаях (например, обследование качества переписи 2011 года в Великобритании) страны проводили небольшие обследования после переписи, когда данные собираются по всем вопросам переписи, а затем сопоставляются с ответами переписи для измерения ошибки респондентов.

329. Для некоторых типов переписей и вариантов использования, описанных в Главе 2, по-прежнему применимы традиционные методы определения общего охвата и качества. Однако новые или пересмотренные методы необходимы в случае оценок численности населения, производимых главным образом на основе административных записей, как в случае комбинированной или полной переписи на основе административных данных. Эти методы, в том числе использование зависимого интервьюирования и методологии «признаков жизни», были описаны в Главе 6. Это продолжает оставаться областью значительного интереса для всех НСС, поскольку страны постоянно развиваются (краткое описание новых и появляющихся методов см. Brown et al (2020).

7.2.2 Демографический анализ: сравнение с альтернативными источниками

330. Демографический анализ[33] может применяться для оценки точности и понимания согласованности и сопоставимости результатов переписи. Демографический анализ включает систематические сравнения, установление пороговых значений приемлемости и понимание любых существенных расхождений. Его невозможно осуществить без концептуального исследования на этапе Источник или работы по проверке и согласованию на этапе Данные. Также может потребоваться объединение нескольких источников для удовлетворения целевой группы на этапе Процесс.

331. Оценки переписи, которые объединяют административные данные, проверяются по альтернативным источникам — например, данными обследования, данными предыдущей переписи или альтернативными источниками. При использовании демографического анализа важно иметь в виду, что оценки в двух источниках могут отличаться в зависимости от пола, возраста или других характеристик. Эти различия могут быть вызваны разными целевыми группами населения, разными датами отсчёта или изменениями в составе населения (при сравнении с историческими данными переписи), концептуальными различиями и вариациями в классификации между переменными, сравниваемыми из разных источников, и / или различиями в выборке, сборе данных, методах и подходах к обработке данных. Любые такие сравнения должны производиться на основе результатов контроля качества на этапах Источник и Данные.

[33] См. O'Hare (2019) для ознакомления с методом и его ограничениями.

Вставка 12: Демографический анализ в Испании

В Испании файл до переписи составляется на основе регистра населения Испании Padrón с применением методологии «признаков жизни» для подсчёта переписного населения. Численность населения, полученная в файле до переписи, затем сравнивается на минимальном географическом уровне с официальным подсчётом населения. Основная цель - выявить и исправить возможные проблемы недостаточного и избыточного покрытия.

Чтобы обеспечить качество численности населения в файле до переписи, население дезагрегируется по наиболее значимым демографическим переменным и сравнивается для каждого уровня переменных: пола, возраста (год за годом), типа гражданства (испанское / иностранное) и национальности (в разбивке по странам). Эти микросравнения помогают установить согласованность общих переменных.

Анализ конкретных подгрупп проводится для выявления возможных проблем с избыточным охватом. Наиболее существенные различия между файлом до переписи и официальным подсчётом населения обусловлены административным характером Padrón, поскольку это не статистический регистр, а административный регистр, который требует обработки для добавления и удаления единиц по мере необходимости (например, добавление рождения или удаления смертей).

С другой стороны, чтобы избежать возможного неполного охвата в файле до переписи, все люди, перечисленные в каждом из доступных административных источников (например, налоговые файлы, файлы социального обеспечения, файлы безработицы и т.д.), которые не были найдены в Padrón на контрольную дату, проверяются. Если есть веские доказательства того, что человек проживает в Испании (учитывая его присутствие в нескольких административных источниках), но не зарегистрирован в Padrón, это лицо включается в файл до переписи.

Типичный пример такой ситуации - люди, которые были удалены из Padrón за несколько месяцев до контрольной даты переписи, 1 января, и которые затем снова появляются в Padrón через некоторое время, например, в феврале. Это может быть случай с иностранцами, срок регистрации которых истёк, и для продления требуется несколько месяцев.

Рассмотрев оценку отдельных источников данных на этапах Источник и Данные, а также источников, объединённых в статистические регистры на этапе обработки, можно сделать профессиональное суждение о том, находятся ли различия, обнаруженные с помощью демографического анализа, в пределах допустимого диапазона. Это будет варьироваться от страны к стране, поэтому рекомендуется, чтобы такие стандарты разрабатывались на местном уровне.

7.2.3 Механизмы обратной связи с поставщиками и стимулы для повышения качества данных

332. Постоянное совершенствование оценок переписи, которые объединяют административные данные, основывается на постоянном улучшении административных данных, которые предоставляет поставщик данных(включая

различные организации, которые могут предоставлять данные для административного реестра, такие как муниципалитеты). Для этого требуются адекватные механизмы обратной связи между поставщиком данных и НСС, а также наличие правильных стимулов как для поставщика данных, собирающего данные, так и для лиц, чьи данные они собирают.

333. Часто поставщик данных также является пользователем данных и будет заинтересован в качестве результатов переписи, которые поддерживают отношения между НСС и поставщиком данных. Взаимодействие между НСС и различными заинтересованными сторонами подробно обсуждалось в Главе 4. Хорошие механизмы коммуникации будут способствовать сокращению разрыва между оперативным и статистическим качеством, тем самым обеспечивая постоянное улучшение качества данных, используемых в переписи, и производимых оценок.

Вставка 133: Демографический анализ в Канаде

Впервые в 2016 году программой переписи населения Канады предусматривался сбор информации о доходах исключительно из административных источников. Оценки, полученные с использованием этих данных, по возможности сравнивались с другими источниками данных. Сравнительный анализ был сосредоточен на различных темах, включая индивидуальный доход с разбивкой по источникам, вопросы охвата, концептуальные различия и различия в обработке, а также региональные различия. Однако, учитывая чувствительность большинства показателей дохода к таким методологическим различиям, пользователи данных должны проявлять осторожность при сравнении оценок доходов переписи 2016 года с оценками доходов домашних хозяйств,, полученными с использованием других обследований, административных данных или данных более ранних переписей.

334. Для поддержки повышения качества НСС может также работать с поставщиком данных для разработки подходящих инструментов, систем и стандартов (например, онлайн-интерфейсов, чётких определений, согласованных областей передовой практики и т. д.) для улучшения сбора, обработки и передача данных.

7.2.4 Независимый обзор методов

335. Независимый обзор проекта и методов переписи будет способствовать постоянному повышению качества, т.е. достижению наилучшего баланса между параметрами качества для удовлетворения потребностей пользователей. Такие обзоры должны проводиться экспертами в области народонаселения и методологии.

336. В августе 2018 года Статистическое управление Новой Зеландии учредила группу экспертов для предоставления консультаций и рекомендаций для Статистики Новой Зеландии по методам, используемым при создании данных переписи 2018 года, а также для пользователей о качестве получаемых данных. Группа одобрила

статистические подходы, используемые для включения административных переписей в данные, и пришла к выводу, что включение этих записей улучшило охват и точность подсчёта населения по основным демографическим характеристикам: возраст, пол, обычное место жительства и этническая принадлежность (см. Stats NZ 2019g).

337. Аналогичным образом, в Соединённом Королевстве внешняя группа по методологическому обеспечению преследует три цели: 1) предоставить внешние независимые гарантии и рекомендации в отношении статистической методологии, лежащей в основе оценок переписи 2021 года, и методов, основанных на административных источниках, 2) выявить значительные пробелы и риски в методах и вносить предложения по смягчению последствий, и 3) анализировать методы административных данных и способствовать их постоянному совершенствованию (см. UKSA 2018b). Обзор группы будет проводится в период с 2018 по 2023 год.

7.2.5 Анализ чувствительности

338. Наряду с привлечением экспертов в области народонаселения для анализа всего метода, качество будет улучшено за счёт привлечения экспертов к анализу, особенно касающегося тематических областей, или решений по качеству на всех этапах качества, которые мы будем называть анализом чувствительности. Чувствительность направлена на установление степени, в которой используемый метод может «подсчитать население в пределах географического региона или демографической группы», что «может быть использовано для понимания систематической ошибки в данных переписи и планирования следующей переписи путём определения групп, которые наиболее трудно поддаются анализу» (см. Stats NZ 2019e, p. 5).

339. Статистическое управление Новой Зеландии привлекла внешних поставщиков для оценки как методов, используемых для добавления людей в данные переписи 2018 года, так и пригодности данных для трёх важных случаев использования, включая определение границ избирательных участков. Анализ чувствительности методов, используемых для добавления людей в файл переписи 2018 года, показал, что пороговое значение для включения в сетчатые блоки оказало наибольшее влияние на то, кто был включён в файл переписи, и что использованный порог был разумным балансом. Дальнейший анализ чувствительности показал, что данные переписи 2018 года были надёжными с точки зрения определения границ избирательных участков, а на границы электората, проведённые с использованием результатов переписи, вряд ли повлиял выбор порогового значения для добавления административных переписей на уровне сетчатых блоков (см. Stats NZ 2019d; Stats NZ 2019e). Это был важный вывод в поддержку качества данных переписи.

7.2.6 Отчёт о качестве и метаданные

340. На последнем этапе обеспечения качества должен быть составлен отчёт для документирования результатов обеспечения качества и гарантий на протяжении

всего периода проведения переписи. Этот отчёт должен включать информацию по каждому этапу обеспечения качества, а также сообщать пользователям данных, где и как учитывался каждый параметр качества. Это позволит производителям и пользователям данных оценивать и предоставлять обратную связь по решениям о качестве, определять, был ли достигнут правильный баланс по параметрам качества, и достаточно ли метаданных для анализа качества.

Вставка 14: Качественные метаданные в Испании

В ходе работы, предшествующей первой переписи населения на основе административных данных в Испании, разрабатывается дополнительная категориальная переменная, обеспечивающая качество данных на основе происхождения каждого значения. Это предоставит пользователям данных индикатор качества для конкретных переменных (см. Pérez Julián, Casaseca and Argüeso Jiménez 2018). Как отмечалось ранее, в Испании статистический регистр населения создаётся путём связывания административного регистра населения Padrón с несколькими административными источниками. Это можно представить в виде огромной матрицы, в которой переменные переписи считаются столбцами, а каждый человек представлен строкой, поэтому ячейки матрицы будут содержать значения для каждого человека по каждой переменной. Чтобы помочь пользователям понять качество данных переписи, для каждой переменной переписи будет добавлена ещё одна категориальная переменная, чтобы информировать пользователей данных о качестве значения каждой ячейки. Как поясняется ниже, эта категориальная переменная предназначена для информирования пользователей данных о качестве, прямо или косвенно.

Первоначальное предложение по разработке этого показателя качества для каждой ячейки основано на типе методологии или источника, используемого для заполнения значения каждой ячейки (см. Таблицу 7). Как правило, значение ячейки, полученное из обновлённого административного источника, имеет наивысшее качество, а значение, полученное посредством детерминированного вменения, - самое низкое. Таким образом, пользователи данных могут косвенно оценить качество каждого значения ячейки.

Кроме того, мера качества для каждого значения ячейки зависит не только от характера базового источника и методологии, но и от остальных характеристик человека. Например, если у 20-летнего человека отсутствуют значения для переменных об официальном семейном положении и его / её основной отрасли, и они детерминированно вменяются в «холостые» и «услуги по размещению питания», соответственно, вероятность того, что первое вменённое значение надёжнее второго очень велико. Связь между возрастом и официальным семейном положении, вероятно, даст хорошие детерминированные оценки условного исчисления, в то время как при условном исчислении ценности для отрасли это не так. Было разработано несколько таких правил для определения качества вменения на основе известных индивидуальных характеристик.

Другое предложение - более прямой способ, который заключался бы в предоставлении переменной пунктуации качества, например, по шкале от 1 до 4, где 1 будет наивысшим качеством, а 4 - самым низким. Это поможет пользователям данных понять, как можно считать "хорошее" или "плохое" вменённое значение.

Оба механизма, косвенный или прямой, предлагают огромный потенциал в оценке качества выпускаемой продукции в двух измерениях: по переменной и по единицам или подгруппам. Предлагается, чтобы все пользователи данных имели свободный доступ к этим переменным качества в выпуске микроданных переписи за 2021 год (примерно 10% от всего продукта переписи) и имели специальные методологические примечания с пояснениями.

*Таблица 4: Первоначальное предложение категорий с указанием качества источника по типу **

Тип данных	Описание	Качество**
DS	Информация предоставлена прямыми источниками актуальна	Самый высокий
DSN	Информация предоставлена из прямых источников, но не актуальна	
CS	Информация о прошлой переписи	
PI	Вероятностное вменение	
DI	Детерминированное вменение	Самый низкий

** По материалам Pérez Julián, Casaseca and Argüeso Jiménez (2018), стр.4*

***Фактический порядок не является неизменным и будет зависеть от переменной, используемых источников и качества лежащего в основе процесса вменения.*

7.3 Тематические исследования

7.3.1 Португалия: оценка качества регистра населения

7.3.1.1 Справочная информация

341. Проект *Census Admin — Административная перепись* (сокращение от *Перепись с использованием Административных данных*) является частью структуры для развития национальной инфраструктуры данных, которая включает стратегию Статистического управления Португалии по интеграции данных из нескольких источников в ответ на все более усложняющееся общество с новыми ожиданиями в отношении статистики.

342. Центральным элементом проекта является создание набора данных о постоянном населении (набор статистических данных о населении объектного типа или Статистический набор данных о населении), охватывающего набор характеристик (географических, демографических и социально-экономических) постоянного населения Португалии. Целью статистической службы Португалии является представление статистических данных о населении с использованием набора данных по статистике населения, начиная с переписи 2021 года и далее.

343. Прототип Статистического набора данных о населении был построен в 2015 году с учётом численности населения 2011 года. Тем временем были созданы четыре новых ежегодных выпуска с ежегодными контрольными датами с 2015 по 2018 год.

344. Для каждого ежегодного выпуска согласованность результатов Статистического набора данных о населении оценивается путём систематического сравнения их с оценками численности населения и известными характеристиками населения.

Кроме того, сравнения с результатами проверки переписи были рассмотрены для измерения качества результатов Статистического набора данных о населении.

7.3.1.2 Подсчёт населения с помощью набора данных по статистике населения по географическому уровню

7.3.1.2.1 Оценка результатов набора данных по статистике населения 2018 года на национальном и региональном уровнях

345. Постоянное население Португалии, оценённое на основе административных данных по данным Статистического набора данных о населении 2018 года, составляет 10 300 502 человека, что представляет собой относительное отклонение +0,2% по сравнению с оценками численности населения 2018 года, опубликованными Статистическим управлением Португалии. Оценки численности населения предоставляют официальные данные о годовом постоянном населении Португалии с использованием компонентов когорты и концепции переписи населения. Расчёты основаны на естественных и миграционных демографических характеристиках, а также на информации из оценок живорождений, смертей, эмиграции и иммиграции.

346. Результаты национального уровня, полученные в рамках проекта Административная перепись, являются многообещающими, учитывая различные допущения, методологии и различные источники двух типов статистических данных: набор данных по статистике населения и оценка численности населения. Соответственно, во всех годовых выпусках набора данных по статистике населения относительное отклонение между этими двумя источниками составляет менее 0,5% (недостаточный или избыточный охват).

347. На региональном уровне относительное отклонение набора данных по статистике населения и оценка численности населения на 2018 год колеблется от –0,4% (Центр) до 3,5% (Алгарве); Лиссабонском регионе –0,1%.

348. Результаты португальского набора данных по статистике населения также являются многообещающими на уровне муниципалитетов: на 2018 год более 76 процентов из 308 муниципалитетов имеют уровни недостаточного или избыточного охвата в пределах 5 процентов по сравнению с оценкой численности населения. Следует отметить, что в 64 муниципалитетах относительное отклонение набора данных по статистике населения и оценка численности населения составляет менее 1% (недостаточный или избыточный охват). Лишь небольшое количество муниципальных образований, в основном менее населённых (15), демонстрируют относительные различия более 10% (выше или ниже).

349. Наряду с географическим распределением набор данных по статистике населения обычно совпадает с оценкой численности населения с точки зрения основных демографических и социально-экономических аспектов. Например, относительные различия набора данных по статистике населения и оценка численности населения в возрастных структурах оценки численности населения очень малы для большинства возрастных групп и для всех версий набора данных

по статистике населения (самые большие различия наблюдаются у пожилых людей).

7.3.1.2.2 Оценка результатов Статистического набора данных о населении 2015 года на местном уровне

350. Сравнения также проводились на более низком географическом уровне; приход или местные административные единицы — уровень 2 (LAU2). Как подробно описано ниже, тест переписи 2016 года (отчётная дата 26 сентября) способствовал оценке результатов данных по статистике населения 2015 года (контрольная дата 31 декабря) на уровне LAU2.

351. Анализ результатов контрольных исследований 2016 года показал, что в четырёх из пяти приходов в выборке, где можно было гарантировать полноту сбора данных, Статистический набор данных о населении 2015 года оценил больше людей, чем те, которые были перечислены. Относительные отклонения колебались от –14,1% до –5,7%. В целом, подсчёт населения в контрольных исследованиях 2016 года по сравнению с расчётным показателем Статистического набора данных о населении 2015 года имел отклонение в –8,8%.

352. Для дальнейшей оценки Статистического набора данных о населении 2015 года микроданные контрольного исследования 2016 года были увязаны с результатами набора данных по статистике населения 2015 года, а для совпавших лиц (около 80%) сравнивались их характеристики. Например, для места обычного проживания 90 процентов респондентов были обнаружены в том же LAU2, что и зарегистрированный в наборе данных по статистике населения 2015 года (вполне удовлетворительно, учитывая 9-месячный временной лаг между контрольными датами пробной переписи и Статистического набора данных о населении).

353. Если взять за основу для сравнения место обычного проживания на уровне муниципалитета, то в целом показатели равенства составляют около 93 процентов, поскольку 3,2 процента лиц в контрольном исследовании 2016 года, сопоставленных со Статистическим набором данных о населении 2015 года, административно проживали в другом приходе того же муниципалитета.

7.3.1.3 Заключительные наблюдения

354. В центре внимания этой работы — оценка качества португальской статистической базы данных для оценки постоянного населения.

355. Для этой цели показаны результаты нескольких сравнений: с оценками численности населения, с разбивкой по географическому уровню (от национального до регионального) и с тестами переписи (более точный географический уровень).

356. Набор административной информации, которая в настоящее время интегрирована в базу данных, имеет высокий потенциал для перехода к модели переписи на основе регистрации или комбинированной переписи. На национальном и региональном уровне результаты Статистического набора данных о населении очень многообещающие. Однако на более низком географическом уровне

сравнение с переписью показало, что оценки базы данных можно улучшить. Статистическая служба Португалии стремится разработать более надёжные методы оценки и пересмотреть правила «регистра признаков жизни». Тем не менее, хотя подсчёт населения на уровне прихода имеет некоторые различия, структура и характеристика населения прихода, приведённая в статистической базе данных 2015 года, очень согласуется с данными, собранными в 2016 году при пробной переписи населения.

Глава 8 Выводы и рекомендации

357. Административные данные могут использоваться в рамках различных методологий проведения переписи и для поддержки всех этапов процесса переписи, включая:

    (a) Создание генеральной совокупности адресов,

    (b) Поддержка полевых операций,

    (c) Регистрация и учёт населения,

    (d) Сбор данных по переменным переписи,

    (e) Обеспечение контроля качества,

    (f) Редактирование и импутация, и

    (g) Моделирование и оценка.

358. Их использование может обеспечить более частое и своевременное получение статистических данных о населении; повышение точности и надёжности и значительное сокращение расходов и нагрузки на респондентов.

359. Однако до того, как административный источник сможет использоваться в целях переписи, необходимо оценить и преодолеть серьёзные проблемы качества. Наиболее важной из них является то, что административные данные, как правило, не собираются для целей переписи. Исходя из этого НСС может иметь малое влияние на:

    (a) Используемые концепции и определения,

    (b) Целевую совокупность,

    (c) Метод сбора данных,

    (d) Процедуры обработки и обеспечения качества,

    (e) Используемые методы, и

    (f) Структуры и системы данных.

360. В Руководстве определены этапы оценки качества с учётом ряда параметров качества, а также предлагаются соответствующие инструменты и показатели, которыми пользователь может руководствоваться в процессе оценки. Применение Руководства призвано помочь в принятии решений об использовании административных данных в переписи, поддерживая при этом процесс постоянной оценки и совершенствования. В Руководстве сформулирован ряд предложений и рекомендаций, которые кратко излагаются ниже.

8.1 Рекомендации

(a) НСС следует **определить административные источники,** которые могут быть релевантными для целей переписи с учётом различных возможностей использования. Важно **определить ожидаемые или требуемые результаты использования источника**, на основании которых можно будет провести оценку релевантности. Они могут включать в себя повышение эффективности проведения переписи с точки зрения сокращения расходов и нагрузки на респондентов; повышение качества переписи; подготовку новых или усовершенствованных материалов переписи. Центральное место в такой оценке отводится определению того, что административный источник должен представлять целевую совокупность и требуемые характеристики этой совокупности для использования в переписи. В Главе 2 Руководства и тематических исследованиях в других главах содержатся примеры использования административных данных в разных странах.

(b) **Отношения между НСС и поставщиком административных данных** имеют крайне важное значение (Глава 4). Они должны подкрепляться надёжными механизмами связи, письменными соглашениями и отличным пониманием потребностей обеих сторон. Кроме того, должна существовать согласованная правовая основа для передачи и использования данных. Чтобы помочь выстроить такие отношения и обеспечить надёжность передачи данных, НСС следует определить области, которые принесут выгоды поставщику. Это могут быть механизмы обратной связи, помогающие поставщику лучше понять свои данные, путём сотрудничества в областях, представляющих общий интерес, или оказания поставщику помощи (путём использования его данных в ходе переписи) в поддержку расширения обеспечиваемого им общественного блага. Разумеется, обратная связь в отношении возможных проблем, связанных с качеством данных, несёт в себе дополнительное преимущество, заключающееся в содействии постоянному повышению качества.

(c) **НСС следует взаимодействовать с поставщиком, чтобы получить углублённое представление об источнике данных.** Это должно привести к созданию чётких и всеобъемлющих метаданных об административном источнике. Метаданные обеспечат полезную справочную базу как для переписи, так и для любой другой статистики, которая может извлечь выгоду из использования источника. Глава 4 содержит подробные сведения о метаданных, которые должны быть собраны, а также различные ссылки на соответствующую литературу.

(d) Поскольку административные данные, как правило, не собираются для нужд переписи, важно, чтобы НСС **понимали и оценивали различия между целевыми совокупностями, концепциями, определениями и временными параметрами.** В более общем плане необходимо тщательно

оценивать согласованность и сопоставимость административного источника, а также его ограничения по различным параметрам качества. Это включает в себя возможность связывания источника, если это является требованием для использования его в целях переписи. Эта оценка будет служить информационным подспорьем для этапов обработки, включая сопоставительный анализ и расчёт, редактирование и импутацию, а также увязку и интеграцию источников (когда решения принимаются между источниками и по ним на основе их качества) (см. Глава 6).

(e) НСС **должны понимать любые ограничения и проблемы, связанные с получением и интеграцией данных административного источника в перепись.** (Глава 4). Это может касаться ресурсов и расходов; рисков, связанных со способностью поставщика осуществлять передачу данных своевременно и требуемого качества, а также приемлемости использования источника для общественности и пользователей данных переписи. В этом отношении существуют важные компромиссы, которые НСС должны учитывать. В частности, **ценность административного источника должна оцениваться с точки зрения его полезности для проведения переписи, с учётом усилий и рисков, связанных с получением и использованием данных.**

(f) НСС имеет ограниченный контроль над сбором и обработкой административных данных, которые могут быть подвержены изменениям в плане охвата и характеристик населения с течением времени. Это может быть обусловлено, например, правовыми, политическими, процедурными или системными изменениями, влияющими на данные и/или их передачу. (Глава 4). **Поэтому НСС должны оценивать уровень риска и управлять им.** Управление риском должно осуществляться путём работы с поставщиком данных в отношении потенциальных или планируемых изменений; гибкого и чуткого реагирования на изменения; а также уменьшения зависимости от какого-либо одного источника или элемента данных, когда это возможно, как за счёт использования других источников данных, так и за счёт адаптации процессов/методологий. (Глава 6)

(g) Важно**, чтобы общественность и пользователи данных понимали, как и почему в переписи используются административные данные.** (Глава 4). Поэтому НСС следует обеспечивать прозрачность в вопросах использования, предоставляя чёткое обоснование преимуществ, в отношении любых рисков и затрат (т.е. явное превышение выгод над издержками). Этого можно достигнуть с помощью **хорошей коммуникации, включая публикацию действующих процедур и политик, способствующих эффективному использованию и защите данных.**

(h) Включению источников административных данных в процесс формирования итогов переписи должно предшествовать проведение адекватно обеспеченного ресурсами исследования целесообразности, которое **предоставит «апробацию концепции» планируемой**

интеграции административных данных в процесс формирования итогов переписи. Рекомендуется провести ряд пробных тестов (с использованием реальных данных) заблаговременно до начала основной переписи, чтобы убедиться в том, что непредвиденные проблемы выявлены, что даст достаточно времени для исправления или корректировки методов, процессов или систем (как описано в Главе 5 и Главе 6).

(i) **Экспертный обзор** (работа с поставщиками данных и профильными экспертами) и **сопоставления между источниками во времени имеют важное значение для выявления любых проблем с качеством** в источнике или регистре. Использование тщательно спланированных обследований (связанных с административными данными или регистрами) может иметь особенно важное значение для выявления и корректировки погрешностей охвата и измерения (Глава 5, Глава 6 и Глава 7).

(j) НСС следует **фиксировать и публиковать результаты оценки и обеспечения качества на протяжении всего процесса переписи**, включая этапы Данные, Процесс и Итоговые материалы. Это позволит производителям и пользователям переписи оценивать и обеспечивать обратную связь, поддерживая постоянный диалог. Это важно для того, чтобы пользователи понимали сильные и слабые стороны и могли определить, был ли достигнут верный баланс между параметрами качества (Глава 7).

(k) НСС следует **разработать свою собственную систему и стратегию** обеспечения качества, поддерживаемую чёткой и исчерпывающей документацией и процедурами обучения. Настоящее Руководство служит полезной основой для решения этих задач, наряду со справочными материалами и тематическими исследованиями, приводимыми в нем. Стратегия должна опираться на постоянную оценку и совершенствование административных данных в рамках планов и процедур проведения переписи. Это должно включать в себя коммуникационные связи между НСС, пользователями и поставщиками данных.

8.2 Области дальнейших разработок

361. В Руководстве основное внимание уделяется оценке качества административных источников для использования в переписях, при этом в нем содержится некоторая информация о процессах, используемых для интеграции и преобразования данных для повышения их качества. Кратко освещается также качество итогов переписей, в которых используются административные данные. Тем не менее в Руководстве не описывается **более широкая концепция общей погрешности или модели** того, как ошибка из каждого источника преобразуется в ошибку в окончательных оценках переписи с учётом изменений качества в результате обработки (что может уменьшить или увеличить погрешность).

362. Разработка такой модели, которая учитывает все источники погрешности, частично рассматривается в рамках модели общей системы ошибок, принятой Статистическим управлением Новой Зеландии (см. Reid et al 2017). Эта модель основана на предложенном Личунь Чжаном (см. Zhang, 2012) расширении парадигмы Общей погрешности наблюдения (см. Groves and Lyberg, 2010; Biemer, 2010). Она состоит из трёх этапов, охватывающих:

    (a) Оценку отдельных источников,

    (b) Оценку интегрированного набора данных, и

    (c) Расчёт и оценку результата.

363. Работа ESSnet KOMUSO 2019 по качеству многоисточниковой статистики (см. Eurostat ESSnet KOMUSO, 2019) также служит полезной основой для оценки качества статистических результатов, основанных на множественных источниках (обследованиях и административных данных).

364. Это может стать областью для дальнейших разработок и международного сотрудничества с **конкретным акцентом на то, как такая модель может быть применена к переписям**. Это может включать в себя изучение того, как можно разработать концепцию или модель общей погрешности и использовать их для оценки качества итогов переписи, опирающихся на множественные источники. Она также могла бы включать в себя работу по изучению того, каким образом воздействие (и смешанное воздействие) различных погрешностей на всех этапах переписи может способствовать принятию решений о наилучшем общем статистическом проекте переписи.

365. В Руководстве основное внимание уделяется оценке административных данных, однако существуют и **другие источники коммерческих данных, которые открывают возможности для использования в целях совершенствования или повышения качества переписей** (например, геопространственные данные, данные мобильных телефонов). Предлагаемые в настоящем Руководстве этапы оценки, параметры, инструменты и показатели качества в значительной степени применимы к источникам, выходящим за рамки административных данных. Это также может быть областью, требующей проведения дальнейшей международной работы, с особым вниманием на то, отличаются ли, и если да, то каким образом, инструменты и методы оценки качества таких источников для использования в переписи от тех, которые обсуждаются в данном Руководстве.

366. Наконец, в ходе консультаций в рамках КЕС по проекту настоящего Руководства несколько стран высказали мнение, что существует необходимость в дальнейшей работе по изучению опыта стран и рассмотрению того, что составляет передовую практику оценки качества административных источников поскольку они касаются охвата **труднодоступных групп населения**. Это может быть та область, которой международное статистическое сообщество возможно пожелает заняться в будущем. В такой работе может быть задействовано более широкое сообщество экспертов, помимо специалистов в области переписи, поскольку эта данная тема имеет отношение к другим областям статистики.

Справочная литература

Abbott, O. (2009). 2011 UK Census Coverage Assessment and Adjustment Methodology. Population Trends 137, Autumn 2009.

Abbott, O., B. Tinsley, S. Milner, A. C. Taylor, and R. Archer (2020). Population statistics without a Census or register. Statistical Journal of the IAOS, 36(1), 97-105.

Argüeso Jiménez, A. (2019). Population and Housing Census in Spain will be fully register-based.

Asamer E.-M., F. Aztleithner, P. Ćetković, S. Humer, M. Lenk, M. Moser and H. Rechta (2016). Quality Assessment for Register-based Statistics — Results for the Austrian Census 2011. Austrian Journal of Statistics Vol. 45, No. 2, pp. 3-14

Australian Bureau of Statistics (2009). ABS Data Quality Framework, May 2009, cat. no. 1520.0, ABS, Canberra.

Bakker, B. F. M. (2012). Estimating the validity of administrative variables. Statistica Neerlandica (2012). Vol. 66, nr. 1, pp. 8–17.

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation, Public Opinion Quarterly, Volume 74, Issue 5, pp. 817–848

Brown, J., C. Bycroft, D. Di Cecco, J. Elleouet, G. Powell, V. Račinskij, P. Smith, S.-M. Tam, T. Tuoto, and L.-C. Zhang (2020). Exploring developments in population size estimation. The Survey Statistician 82, 27-39.

Cerroni, F., G. Di Bella and L. Galiè (2014). Evaluating administrative data quality as input of the statistical production process. In: Rivista Di Statistica Ufficiale. ISTAT.

Chambers, R. (2001). “Evaluation Criteria for Statistical Editing and Imputation.” National Statistics Methodological Series 28: 1–41.

Chieppa, A., G. Gallo, V. Tomeo, F. Borrelli and S. di Domenico (2018). "Knowledge discovery for inferring the usually resident population from administrative registers" in Mathematical Population Studies, Pages 92-106.

Choi, H. (2019). Adjusting for linkage errors to analyse coverage of the administrative population. Statistical Journal of the IAOS, 35(2), 253-259.

Cornell University Research Data Management Service Group (2020). Guide to writing "readme" style metadata. https://data.research.cornell.edu/content/readme

Crescenzi, F., G. Sindoni and D. Zindato (2014). Lessons learned from the 2011 Italian census and innovations leading towards a continuous census. Note by the National Institute of Statistics of Italy, presented at the UNECE/Eurostat Group of Experts on Population and Housing Censuses, Sixteenth Meeting, Geneva, 23-26 September 2014.

Daan Zult, P., P de Wolf, B. Bakker and P. van der Heijden (2019). A linkage error correction model for population size estimation with multiple sources.

Daas, P.J.H., J. Arends-Tóth, B. Schouten and L. Kuijvenhoven (2008). Quality Framework for the Evaluation of Administrative Data. Paper presented at the European Conference on Quality in Official Statistics.

Daas P., S. Ossen, R. Vis-Visschers and J. Arends- Tóth (2009). Checklist for the Quality evaluation of Administrative Data Sources. The Hague: Statistics Netherlands.

Daas, P., S. Ossen, M. Tennekes, J. Burger and F. Cobben (2012). Input Quality of administrative data (BLUE-ETS WP4). Presented at Quality 2012.

European Commission (2008). 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.

Eurostat BLUE-ETS (2011). List of quality groups and indicators identified for administrative data sources.

Eurostat (2013). Use of administrative and accounts data in business statistics

— (2014). ESS Handbook for quality reports.

— (2017). European Statistics Code of Practice, revised edition 2017.

— (2019). Quality Assurance Framework of the European Statistical System. Version 2.0.

— (2020). European Statistical System Handbook for Quality and Metadata Reports,

Eurostat ESSnet KOMUSO (2016). Checklist for Evaluating the Quality of Input Data https://ec.europa.eu/Eurostat/cros/system/files/essnet_wp1_report_final_version4.pdf

— (2019). Quality Guidelines for Multisource Statistics https://ec.europa.eu/Eurostat/cros/system/files/qgmss-v1.1_1.pdf

Eurostat ESSnet MIAD (2014). MIAD deliverable B2, B3 — Quality check list for the Source phase [Data]. Available from: https://ec.europa.eu/Eurostat/cros/content/miad-deliverable-b2_en

Falorsi, S. (2017). Census and Social Surveys Integrated System. Note by the National Institute of Statistics of Italy, presented at the UNECE/Eurostat Group of Experts on Population and Housing Censuses, Nineteenth Meeting, Geneva, Switzerland, 4–6 October 2017.

Gallo, G., A. Chieppa, V. Tomeo and S. Falorsi (2016). The integration of administrative data sources in Italy to increase Population Census data availability. Note by the National Institute of Statistics of Italy, presented at the UNECE/Eurostat Group of Experts on Population and Housing Censuses, Eighteenth Meeting, Geneva, Switzerland, 28-30 September 2016.

Gath, M., and C. Bycroft (2018). The potential for linked administrative data to provide household and family information.

Gibb, S., C. Bycroft and N. Matheson-Dunning (2016). Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI).

Groves, R. M., F. J. Fowler Jr., M. Couper, J. M. Lepkowski, E. Singer and R. Tourrangeau (2004). Survey methodology, Wiley, New York.

Groves, R. M. and L. Lyberg (2010). Total Survey Error: Past, Present, and Future, Public Opinion Quarterly, Volume 74, Issue 5, 849–879.

Harron, K., H. Goldstein, and C. Dibben (2015). Introduction. In K. Harron, H. Goldstein, & C. Dibben (Eds.), Methodological Developments in Data Linkage. New York: John Wiley & Sons.

International Organization for Standardization (2015). Quality management systems — Fundamentals and vocabulary. ISO 9000:2015(en)

Iwig, B., M. Berning, P. Marck and M. Prell (2013). Data Quality Assessment Tool for Administrative Data

ISTAT (2020). Nota tecnica sulla produzione dei dati del Censimento Permanente: la stima della popolazione residente per sesso, età cittadinanza, grado di istruzione e condizione

professionale per gli anni 2018 e 2019: Dalla rilevazione “porta a porta” al Censimento permanente [Technical note on the production of Permanent Census data: estimating the resident population by sex, age, citizenship, education level and occupational status for the years 2018 and 2019: From door-to-door survey to permanent census]. Rome: ISTAT.

Lavigne, M., and C. Nadeau (2014). A Framework for the Evaluation of Administrative Data. In Proceedings of Statistics Canada Symposium.

Lothian, J., A. Holmberg, and A. Seyb (2019). An evolutionary schema for using “it-is-what-it-is” data in official statistics. Journal of Official Statistics 35, 137-165.

Oberski, D. L. A. Kirchner, S. Eckman & F. Kreuter (2017). Evaluating the Quality of Survey and Administrative Data with Generalized Multitrait-Multimethod Models, Journal of the American Statistical Association, 112:520, 1477-1489

O’Byrne, E., C. Bycroft and S. Gibb (2014). An initial investigation into the potential for administrative data to provide census long-form information: census transformation programme. Wellington: Statistics New Zealand.

O’Hare, W.P. (2019). Methodology Used to Measure Census Coverage. In: Differential Undercounts in the U.S. Census. Briefs in Population Studies. Springer, Cham

Office for National Statistics (ONS) (2016). Methodology of Statistical Population Dataset V2.0.

— (2019). Developing our approach for producing admin-based population estimates, England, and Wales: 2011 and 2016.

— (2020a). Joined up data in government: the future of data linking methods. Data and Analysis Method Reviews.

— (2020b). ONS working paper series no 20 – Feasibility of using donor-based imputation for census outputs on number of rooms using Valuation Office Agency data: Demonstration of using a donor-based imputation method (CANCEIS) to address missing values when replacing the number of rooms question on Census 2021. ONS working paper series 20, 31 July 2020, Newport: ONS.

Pérez Julián, M. P., C. Casaseca and A. A. Argüeso Jiménez (2018). Assessing quality in a register-based census. Paper presented at the European Conference on Quality in Official Statistics, Krakow.

Reid, G., F. Zabala, and A. Holmberg (2017). Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ, Journal of Official Statistics, 33(2), 477-511.

Rogers, N., and L. Blackwell (2020). A statistical quality framework for longitudinally linked administrative data on international migration.

Schnetzer M., F. Astleithner, P. Cetkovic, S. Humer, M. Lenk and M. Moser (2015). Quality Assessment of Imputations in Administrative Data, Journal of Official Statistics, Vol. 31, No. 2, pp. 231–247.

Scholtus, S. and B. Bakker (2013). Estimating the Validity of Administrative and Survey Variables by Means of Structural Equation Models. *New Techniques and Technologies for Statistics* conference 2013.

Shafer, G. (1992). Dempster-Shafer theory. Encyclopaedia of artificial intelligence.

Shipsey, R. and J. Plachta (2020). Linking with anonymised data- how not to make a hash of it.

Statistics Austria (2019). Quality assessment of administrative data — Documentation of Method.

Statistics Canada (2017). Statistics Canada's Quality Assurance Framework

— (2020). 'Counting all Canadians' in Painting a Portrait of Canada: The 2021 Census of Population.

Statistics Estonia (2019). Pilot Census Report.

Statistics New Zealand (2017). Experimental population estimates from linked admin data: 2017.

— (2019a). Overview of statistical methods for adding admin records to the 2018 Census dataset

— (2019b). Linking 2018 Census responses to the Integrated Data Infrastructure

— (2019c). Dual system estimation combining census responses and an admin population

— (2019d). Electoral boundaries sensitivity analysis of 2018 Census data.

— (2019e). Predicting the quality of admin location information for use in the 2018 Census.

— (2019f). Population counts sensitivity analysis of 2018 Census data.

— (2019g). 2018 Census External Data Quality Panel: Assessment of variables.

— (2020). Guide to reporting on administrative data quality.

UK Statistics Authority (UKSA) (2015a). Quality Assurance of Administrative Data — Setting the Standard.

— (2015b). Administrative Data Quality Assurance Toolkit. Version 1 January 2015.

— (2018a). Code of Practice for Official Statistics (Edition 2.0).

— (2018b). Methodological Assurance Review panel – Census.

— (2019). Quality Assurance of Administrative Data (QAAD) toolkit.

— (2020). Ethics Self-Assessment Tool.

UNECE (1992). Fundamental Principles of Official Statistics. Available from https://unece.org/statistics/fundamental-principles-official-statistics

— (2011). Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices.

— (2014). A Suggested Framework for the Quality of Big Data.

— (2015). Conference of European Statisticians Recommendations for the 2020 Censuses of Population and Housing.

— (2017). Quality Indicators for the Generic Statistical Business Process Model (GSBPM) — For Statistics derived from Surveys and Administrative Data Sources. Version 2.0.

— (2018a). Guidelines on the use of registers and administrative data for population and housing censuses.

— (2018b). Annex F – Portugal Case Study in Guidelines on The Use of Registers and Administrative Data for Population and Housing Censuses. pp.64-67.

— (2021). Keeping Count: conducting the 2020 round of population and housing censuses during the Covid-19 pandemic.

United Nations (2009). Handbook on Geospatial Infrastructure in Support of Census Activities Studies in Methods: Series F No. 103, ST/ESA/STAT/SER.F/103 New York: United Nations Department of Economic and Social Affairs, Statistics Division.

United States Census Bureau (2009). History: 2000 Census of Population and Housing.

— (2019). Counting the Hard to Count in a Census. *Select topics in international censuses*.

Vega Valle, J.L., A. Argüeso Jiménez and Pérez Julián, M. (2020). 'Moving Towards a Register Based Census in Spain'. Statistical Journal of the IAOS.1 Jan. 2020: 187 – 192.

Yucel, R.M. and A.M. Zaslavsky (2005). Imputation of binary treatment variables with measurement error in administrative data. Journal of the American Statistical Association, Vol. 100, No. 472, 1123–1132.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. Statistica Neerlandica 66: 41–63.

Глоссарий терминов

Административная регистрация (Новая Зеландия): процесс сбора данных, взятых из административного источника, с целью дополнения данных, записанных в вопросниках, собранных при полевой регистрации.

Административные данные: Данные, хранящиеся в регистрах, реестрах и других административных источниках, относящиеся к информации, собираемой правительством и/или другими организациями в основном для административных (не исследовательских или статистических) целей, таких как регистрация, учёт и проведение операций, обычно при предоставлении каких-либо услуг.

Административный источник: Массивы данных, которые содержат информацию, собранную в основном для административных (не исследовательских или статистических) целей. Такие источники включают административные регистры (с уникальным идентификатором) и, возможно, другие административные данные без уникального идентификатора.

Административное население: Совокупность объектов или единиц (например, людей, жилых домов, предприятий), которые фиксируются административным источником или регистром.

Административный регистр: Систематический сбор данных на уровне единиц, организованный таким образом, чтобы было возможно обновление (где «обновление» — это обработка идентифицируемой информации с целью создания, обновления, исправления или расширения реестра. Такие регистры в основном используются в административной информационной системе, в которой данные используются при производстве товаров и услуг в государственных или частных учреждениях или компаниях. Административные регистры, используемые для статистических целей, обычно управляются государством или совместно местными властями, но некоторые регистры эксплуатируемые частными / коммерческими организациями также могут быть использованы.

Административная единица: Единицы, для которых записываются административные данные. Они могут совпадать или не совпадать с теми, которые требуются для статистических результатов (которые называются статистическими единицами).

Атрибут: Социально-демографическая или экономическая характеристика административной или статистической единицы, информация о которой требуется для переписи.

Базовый регистр: Регистры, от которых зависит вся система статистики на основе регистров. Они включают в себя как административные базовые регистры, так и статистические базовые регистры, причём первые представляют собой ресурсы, предназначенные для государственного управления, а вторые основаны на административном регистре, с ключевыми функциями определения важных групп населения и содержащие ссылки на другие базовые регистры.

Бенчмаркинг: Сравнение данных, метаданных или процессов с признанным стандартом.

Большие данные: Большие, часто неструктурированные наборы данных, которые доступны потенциально в режиме реального времени, но которые сложно как эффективно обрабатывать, так и обеспечивать качество с использованием традиционных методов и технологий. Объем и разнообразие доступных данных быстро растёт, и такие наборы данных доступны во многих форматах, включая аудио, видео, компьютерные журналы, транзакции покупок, датчики и сайты социальных сетей. Некоторые из этих данных находятся в свободном доступе в Интернете, в то время как другие принадлежат частному сектору, к которому может не быть свободного доступа.

Владелец регистра: см. «Держатель регистра».

Возможность связывания: Возможность связывать данные из нескольких различных источников административных данных с одной и той же единицей, обычно с помощью уникального идентификационного номера или кода.

Гипермерности Daas: Высокоуровневые измерения или «представления» качества административного источника для использования в статистических целях, к которым относятся: источник, метаданные и сами данные.

День переписи: Дата контрольного момента переписи, независимо от того, когда были собраны данные.

Доступность: Показатель качества данных, характеризующий условия и способы получения, использования и интерпретации данных пользователями.

Держатель регистра: Орган, ответственный за хранение и ведение административного регистра (также именуемый «хранителем регистра» или «контроллером данных»).

Единица: Наименьший объект, к которому относится любой элемент административных данных. Для целей переписи единицы могут относиться к отдельным лицам, домашним хозяйствам, зданиям или жилым помещениям.

Импутация: Процесс, при котором отсутствующие элементы исходных данных вменяются (заменяются) правдоподобными и согласованными значениями.

Исходные данные: Данные иногда называемые «необработанными данными», полученные из административного источника до любой обработки или проверки НСС.

Итоговые данные: Обработанные данные, используемые в статистических итоговых данных.

Качество источника: Качество административных источников, из которых данные поступают в НСС для целей переписи.

Качество исходных данных: качество источников административных данных в сравнении с их использованием в переписи. Этапы Источник и Данные вместе обеспечивают общую оценку качества исходных данных.

Качество процесса: Влияние изменений качества данных, используемых для целей переписи, во время обработки исходных данных НСС.

Качество результатов: Качество обработанных данных, используемых в статистических итогах.

Классификации: Статистические классификации предоставляют собой набор связанных категорий в значимом, систематизированном и стандартном формате, например, стандарт НСС для классификации профессий. Классификации обычно разрабатываются для поддержки системы разработки и, следовательно, для организации и представления статистических данных.

Комбинированная перепись: Перепись, основанная на комбинации данных, взятых из административных регистров и собранных с помощью вопросников.

Метаданные: Данные, которые описывают или определяют другие данные. В широком смысле это относится ко всему, что необходимо знать пользователям для правильного использования реальных данных с точки зрения доступа, обработки, интерпретации, анализа и представления информации. Метаданные включают, например, описания файлов, кодовые книги, детали обработки, образцы дизайна и отчёты о полевых исследованиях. Метаданные следует отличать от «Параданных», которые обычно относятся к деталям, описывающим процесс сбора данных переписи либо из административных источников, либо из полевых переписей/обследований.

Надёжность: Измерение качества, которое относится к степени близости значений данных к более ранним или последующим данным.

Необработанные данные: См. «Исходные данные».

Объекты: Термин «объект» используется для обозначения единиц в наборе административных данных. Этот термин используется для различения единиц в административных данных и статистических единиц после того, как эти данные были каким-либо образом преобразованы. Это особенно актуально в тех случаях, когда единица (или «объект») в административном регистре отличается от целевой статистической единицы. Например, если налоговый регистр, где единицы годовой налоговой декларации (т.е. одно и то же лицо может сдать несколько налоговых деклараций за один или несколько лет), конвертируется в отдельных «людей».

Оценка воздействия на неприкосновенность частной жизни: Процесс, который помогает организациям выявлять и управлять рисками для конфиденциальности, возникающими в связи с новыми проектами, инициативами, системами, процессами, стратегиями, политиками и деловыми отношениями.

Оценка двойной системы: Статистический метод, основанный на методе захвата — повторного захвата, применяемый для оценки численности населения.

Оценки: Термин используется в настоящем Руководстве для обозначения статистических данных, полученных в результате переписи, и отражает процессы, предпринятые НСС для корректировки исходных данных с учётом недостаточного или избыточного охвата, ошибок, недостающих подсчётов и мер по контролю статистических данных.

Оценки переписи: Термин, используемый некоторыми странами для описания итоговых данных переписи, чтобы отразить тот факт, что опубликованные цифры не

претендуют на то, чтобы быть истинными подсчётами, и что всегда должна существовать некоторая степень неопределённости (даже небольшой) в точности цифр.

Ошибка измерения: Ошибки измерения переменных или характеристик (например, возраста, пола и т. д.). Они включают в себя несколько типов ошибок в переменных, включая релевантность (несоответствие определений), отображение (ошибки в переклассифицированных показателях из-за плохой эквивалентности между предоставленной и целевой классификациями, которые, следовательно, могут потребовать корректировок, например, путём импутации) и ошибки сопоставимости (ошибки между переклассифицированными и скорректированными показателями).

Ошибка репрезентативности: Ошибка в представлении единиц населения или объектов (например, отдельных лиц или домохозяйств в переписи). К ним относятся ошибки, связанные с избыточным и недостаточным охватом (отсутствие согласования с целевой совокупностью), идентификацией (ошибки в классификации единицы на основе несоответствий между несколькими источниками) и ошибки единиц (ошибки при создании статистических единиц, представляющих интерес, там, где они не существуют в любом доступном источнике данных).

Параданные: См. «Метаданные».

Периодичность: В контексте предоставления административных данных это период времени между отчётными датами для последовательных итоговых наборов данных. Для переписи в более общем смысле это время между датами последовательных переписей (периодичность проведения переписи).

Перепись на местах: Процесс сбора информации об отдельных лицах, домохозяйствах и /или жилищной единице, охватывающий все население (или его выборку) с помощью вопросников.

«Признаки жизни»: Показатель, используемый для минимизации избыточного охвата лиц, зарегистрированных в различных административных регистрах, полученный с применением строгих критериев или «правил деятельности», чтобы гарантировать, что в оценки переписи включены только живые люди, которые обычно являются резидентами.

Производная переменная: Новая переменная, сформированная с использованием данных из других переменных.

Пунктуальность: Показатель качества данных, характеризующий временной интервал между запланированными и фактическими датами публикации. В контексте административного источника это относится к временному интервалу между ожидаемой (или оговорённой) датой предоставления данных в НСС и фактической датой предоставления.

Путь к данным: Совокупность исходных процессов от сбора данных до их использования в производстве статистики, во многом аналогично общей статистической модели бизнес-процессов.

Рамки: Любой список, материал или устройство, которые определяют границы, идентифицируют и разрешают доступ к элементам целевой группы. Статистический регистр — конкретный пример.

Регистр: Систематизированный набор данных уровня единиц наблюдения, способ организации которых делает возможным их обновление. Обновление – это обработка поддающейся идентификации информации с целью создания, актуализации, исправления или расширения регистра.

Регистр адресов: Регистр адресов жилых помещений, часто используемый для создания счётных участков, включающих сопоставимое количество жилых единиц. В случае многоквартирных домов под одним адресом проживания может быть более одного жилого помещения.

Регистр населения: Статистический регистр и совокупность лиц, обычно проживающих (независимо от определения) в стране. Кроме того, он часто даёт некоторые демографические характеристики людей.

Регистровая перепись: Перепись, при которой все данные собираются из административных регистров. Перепись, основанная на сочетании данных, взятых из регистров и вопросников, называется «комбинированной переписью».

Редактирование данных: Процесс обнаружения и исправления данных, которые содержат ошибки, логические несоответствия и ложные значения.

Релевантность: Показатель качества данных, который относится к степени актуальности, в которой они удовлетворяют потребности пользователей с точки зрения охвата и содержания. Когда речь идёт конкретно об источниках данных, измерение относится к степени, в которой такие источники содержат данные, которые удовлетворяют потребности НСС в отношении их предполагаемого использования.

Своевременность: Показатель качества данных, характеризующий период между моментом получения информации и событием или явлением, которое она описывает (в случае данных переписи это обычно день переписи), и датой публикации данных. При использовании административных данных своевременность также относится к промежутку времени между датой события, зарегистрированного в источнике данных, и датой предоставления данных в НСС.

Сетчатый блок: Наименьшая географическая единица, по которой статистические данные собираются и обрабатываются Статистическим управлением Новой Зеландии.

Скользящая перепись: Альтернативный подход к традиционной модели проведения переписи посредством кумулятивного непрерывного обследования, охватывающего всю страну за определённое время, а не на определённый день. При скользящей переписи необходимо учитывать два основных параметра: продолжительность периодичности проведения, которая сама связана с частотой требуемого обновления и размер выборки, который зависит от бюджета и уровня географического анализа, необходимого для распространения.

Сопоставимость: Показатель качества данных, характеризующий степень, в которой статистические данные сопоставимы между географическими районами и во времени.

Статистический контроль за раскрытием: Процесс, посредством которого необработанные данные, взятые из административного источника или собранные на местах, модифицируются во время обработки данных, во избежание раскрытия информации об идентифицируемых отдельных лицах или домохозяйствах.

Статистический регистр: Регистр, созданный для статистических целей, путём преобразования данных из регистров и/или других административных источников данных. Статистические регистры также называются «вторичными регистрами».

Согласованность: Измерение качества, которое относится к степени сходства данных, полученных из разных источников или методов, но относящихся к одной и той же теме.

Точность: Измерение качества, которое относится к степени, в которой информация правильно описывает явления, для измерения которых она предназначена. Проще говоря, точность — это показатель качества данных, характеризующий степень близости оценок к неизвестным «истинным» значениям.

Теория Демпстера-Шафера: Обобщение байесовской теории субъективной вероятности.

Тестовые данные: Меньшие объёмы данных из административного источника/ регистра, передаваемые НСС для целей технико-экономического обоснования и тестирования систем.

Труднодоступные группы населения: Группы, которые, как правило, недопредставлены либо потому, что они очень малочисленны, либо их трудно идентифицировать, например, из-за отсутствия стандартизированных определений или из-за отсутствия сбора данных по соответствующим переменным, поскольку они предпочитают не идентифицироваться, например, из-за стигмы, связанной с членством в группе; потому что они систематически исключаются из стандартных методов сбора и основ выборки (например люди, живущие в учреждениях); потому что физически труднодоступны (например живущие в отдалённых районах или без постоянного места жительства); или потому что их трудно перечислить даже после того, как они были идентифицированы и взяты из выборки (например люди, живущие с деменцией, люди, не говорящие на национальном языке).

Управляющие данными: см. «держатель регистра».

Целевая совокупность: Совокупность, для которой требуется информация. Целевая совокупность — это набор статистических единиц.

Ясность: Измерение качества, связанное с доступностью любой дополнительной информации или метаданных, которые могут потребоваться для помощи пользователю в интерпретации и понимании соответствующих данных.