Глава 8. Источники, сбор и обработка данных

8.1. Введение

Хотя главному статистику необязательно знать все детали каждой статистической деятельности, общее понимание статистических концепций и статистических процессов жизненно важно для эффективного принятия решений. Отправной точкой является Общая статистическая модель бизнес-процесса (GSBPM), описанная в главе 5. GSBPM описывает статистический процесс с точки зрения восьми *этапов,* а именно: определение *потребностей, проектирование, построение, сбор, обработка, анализ, распространение* и *оценка* . Глава 6 посвящена определению *потребностей* . В этой главе рассматриваются следующие четыре фазы, т. *Е. Проектирование, сборка, сбор* и обработка , а в главах 7 и 9 обсуждаются *анализ* и *анализ.Оцените* фазы соответственно.

Для целей главы статистические процессы разделены на четыре группы в зависимости от источника входных данных, а именно:

а) обследование, включая перепись как частный случай;

б) административные данные;

в) Геопространственные данные; и

г) Большие данные.

Обследования и переписи являются традиционным источником данных и по-прежнему широко используются. Раздел 8.2, самый длинный раздел в главе, подробно описывает все аспекты этапов проектирования, сборки, сбора и обработки. Он также включает подразделы об отношениях с респондентами и обучении персонала, которые пересекаются между этапами.

Al , хотя хорошо известно и точно ориентированы на адрес , указанный потребностей, обследований и переписей населения занимает много времени, трудоемким и дорогостоящим. В то же время НСУ приходится иметь дело с сокращающимися бюджетами и возникающими запросами на более своевременную и дезагрегированную статистику и показатели, охватывающие новые области, в частности, в отношении мониторинга Повестки дня в области устойчивого развития на период до 2030 года и других региональных и национальных политик развития. Для решения этих проблем НСУ и другие производители официальной статистики начинают использовать огромные объемы данных, которые стали доступны в нашем цифровом обществе, в дополнение к существующим материалам обследований. Действительно, «датафикация» общества вместе с рентабельной емкостью хранения и быстрым увеличением производительности вычислений открыли новые возможности для объединения статистических обследований и переписей, административных записей и нетрадиционных источников данных, таких как геопространственная информация и большие данные. .

Наиболее резко возросло использование административных данных. Действительно, столкнувшись с новым запросом данных, первый шаг, который должна предпринять НСУ, - это проверить, можно ли удовлетворить потребность с помощью существующего источника административных данных. В разделе 8.3 описаны особенности этапов разработки, построения и сбора процессов с использованием административных данных, которые отличают их от обследований и переписей.

Геопространственные данные - это данные, которые имеют географический компонент. Самый обширный источник таких данных - спутниковые снимки. Геопространственные данные можно использовать для значительного обогащения других источников данных. Раздел 8.4. описывает типы геопространственных данных, их использование и проблемы, с которыми сталкиваются при таком использовании.

В разделе 8.5 этой главы рассматриваются большие данные, датчики Интернета вещей (IoT), носимые и мобильные устройства как инструменты, которые могут дополнять традиционные способы сбора данных из-за их потенциально высокого охвата населения и использования в повседневной жизни. Этот раздел основан на концептуальном документе CBS Нидерландов и Статистического управления Канады под названием « Будущий расширенный сбор данных» , представленном на 62-м Всемирном статистическом конгрессе ISI в 2019 году. Для НСУ крайне важно рассмотреть возможность внедрения расширенных возможностей сбора данных, чтобы повысить их ценность для обществ, которым они служат. . Платформы данных и другая инфраструктура, обычно называемая *облаком,* могут быть ценными инструментами для объединения и обработки традиционных и нетрадиционных данных из государственных и частных источников.

Выборочные обследования и переписи

Закон Generic по официальной статистике (Glos), ЕЭК ООН (2018) определяет статистическое исследование как *« основной набор индивидуальных данных от респондентов данной группы населения , проводимой производителем официальной статистики исключительно для статистических целей путем систематического использования статистической методология ».*

Для сбора информации непосредственно от респондентов в различных обстоятельствах используются два подхода: *выборочные обследования* и *переписи* . Выборочное обследование - это деятельность по сбору данных, охватывающая только часть (выборку) всего населения, в то время как перепись, как правило, представляет собой исследование каждой единицы (всех или всего) в генеральной совокупности. Переписи часто называют полной переписью или полным подсчетом. Оба подхода используются, чтобы сделать выводы обо всей популяции. Переписи и выборочные обследования дополняют статистическую систему. Каждый из них считается особым случаем опроса.

Переписи - это старейшая из статистических мероприятий, которые служат для получения моментального снимка всего населения. Благодаря полному охвату переписи населения являются наиболее широко известной деятельностью НСО и обычно являются первой ассоциацией, которая приходит на ум неспециалистам, когда они думают об официальной статистике. Доступно множество материалов, касающихся проведения переписи населения, включая подробные отчеты о реальном опыте проведения переписи. Статистический отдел Организации Объединенных Наций разработал серию справочников и руководств для оказания странам помощи в подготовке к переписи населения. Они включают:

- Руководство по использованию технологий электронного сбора данных при переписи населения и жилищного фонда (2019 г.);
- Справочник по геопространственной инфраструктуре в поддержку переписи (2009 г.);
- Справочник по проведению переписей населения и жилищного фонда, ред. II (2016 г.)
- Справочник по редактированию переписи населения и жилого фонда (2009 г.);

- Измерение экономически активного населения при переписи населения: Справочник и сборник экономических характеристик при переписи населения: технический отчет (2010).
- Принципы и рекомендации по проведению переписей населения и жилого фонда, ред. III (2017 г.)

Из-за наличия обширных высококачественных и новейших материалов ООН, переписи населения и другие переписи населения в данном справочнике не рассматриваются подробно.

Помимо переписей населения, НСУ проводят другие мероприятия по исчерпывающей регистрации для сбора характеристик и данных о размере и структуре жилья, экономических единиц, зданий или ферм. Из-за полного охвата целевого населения эти мероприятия чаще всего называют переписями. Экономические переписи особенно полезны, когда другие надежные статистические данные (особенно о структуре экономики), а также надежные регистры и административная информация недоступны. См. Раздел 11.5 главы 11 для получения дополнительной информации о сельскохозяйственных переписях. а также на веб-странице Всемирной программы сельскохозяйственных переписей ФАО .

С появлением теории выборки у переписей появились дополнительные цели, поскольку они стали источником информации для основ выборки и основой оценок для выборочных обследований. Это привело к снижению как затрат, так и нагрузки на респондентов, что позволило собирать более подробные и частые данные. Данные выборочных обследований обычно более сложны, чем базовые данные, собранные в ходе переписи. Обследования часто используются для расширения характеристик тем переписи (и добавления дополнительных тем) и для измерения изменений между переписями. Выбор правильного подхода может зависеть от характеристик населения и других факторов, которые обсуждаются ниже. Австралийское статистическое бюро перечисляет следующие преимущества и недостатки выборочных обследований по сравнению с переписями :

Плюсы ПЕРЕПИСИ	Минусы ПЕРЕПИСИ
• обеспечивает точную оценку совокупности (без ошибки выборки) • контрольные данные могут быть получены для будущих исследований • более вероятно, что будет доступна подробная информация о небольших подгруппах населения	• может быть сложно пересчитать все единицы населения за отведенное время • более высокие затраты как в кадровом, так и в денежном выражении, чем для выборки • как правило, сбор, обработка и публикация данных занимает больше времени, чем из выборки
Плюсы ОБРАЗЦА	**Минусы ОБРАЗЦА**
• затраты, как правило, будут ниже, чем при переписи • результаты могут быть доступны в более короткие сроки	• данные могут не быть репрезентативными для всего населения, особенно если размер выборки невелик

• если используются хорошие методы выборки, результаты могут быть очень репрезентативными для фактического населения	• часто не подходит для получения контрольных данных • поскольку данные собираются из подмножества единиц и выводы, сделанные в отношении всей совокупности, данные подвержены ошибке «выборки» • уменьшение количества единиц уменьшит доступную подробную информацию о подгруппах в популяции

Подробная информация переписи населения небольших территорий используется для разработки основ выборки и выборки для единиц обследования. Хотя программы обследований могут собирать различную информацию из переписи, несколько тем обычно являются общими для обеих. Следовательно, чтобы максимизировать полезность данных из обоих источников, важно стандартизировать концепции и определения. Стандартизация также позволяет использовать современные подходы, такие как оценка малых площадей, которая позволяет создавать оценки результатов обследования на пространственных уровнях, которые ненадежны при использовании традиционных подходов. Оценка малых территорий - это процедура, при которой данные обследования объединяются с данными переписи или административными записями, а затем результаты обследования моделируются в соответствии с общими характеристиками для каждого респондента во всем населении. Этот подход позволяет создавать эконометрические оценки результатов обследования на ненадежных пространственных уровнях. Примером такого подхода в официальной статистике являются оценки бедности на уровне округа (обычно уровень 3 территориальной классификации) по результатам обследования, которое является репрезентативным на муниципальном уровне (обычно уровень 2 территориальной классификации).

8.2.1 Описание функций опроса

Обследование - это наиболее часто используемый механизм сбора данных в официальной статистике, он универсален, относительно дешев и быстр (по крайней мере, по сравнению с переписью), может ответить на широкий спектр вопросов о различных характеристиках населения и используется почти во всех статистических площадь. Обычно это мотивируется необходимостью изучить характеристики популяции, создать базу данных для аналитических целей или проверить гипотезу. Поэтому опросы обычно используются как метод поддержки принятия решений в частных компаниях и в правительстве, а также как важная часть научного метода в исследовательской деятельности. Важно отметить, что опросы официальной статистики также используются для международных сопоставлений. НСУ обеспечивают международную сопоставимость результатов обследований, применяя общую методологическую основу, используя аналогичные методы и процедуры и контролируя качество процессов и результатов с помощью систем управления качеством и структур обеспечения качества (см. Главу 7 « *Управление качеством»*).

Опросы могут использоваться для получения релевантной информации по большинству вопросов, но статистики должны четко сообщать, что опросы - это инструмент для агрегирования результатов, полученных от населения, и что постановка технических вопросов широкой общественности не всегда может дать соответствующие ответы. Ключом к успеху опроса является четко определенный набор вопросов, на которые может дать реалистичный ответ определенная совокупность. При планировании опроса

важно найти баланс между удовлетворением потребностей пользователей и избежанием чрезмерной нагрузки на респондентов.

8.2.2 Типы опросов

Опросы - это универсальный инструмент, который можно использовать для различных целей. Они составляют основу любой статистической программы и, наряду с переписью и обработанной административной информацией, составляют основу официальной статистики. Обзоры могут быть поперечными и продольными. Поперечные опросы - это обследования, в которых каждый раз опрашивается новая выборка респондентов, и поэтому они лучше всего подходят для измерения распространенности какой-либо характеристики в популяции. Лонгитюдные опросы - это опросы, которые проводятся на одной и той же выборке респондентов с течением времени и поэтому лучше всего подходят для измерения распространенности характеристики. Статистические обследования обычно группируются в статистические области, относящиеся к определенным группам населения. Их также можно разделить по типу единицы наблюдения на две основные категории - обследования домашних хозяйств и обследования предприятий. Некоторые обследования представляют собой комбинацию обеих категорий, например обследования сельского хозяйства, которые могут быть обследованиями как домашних хозяйств, так и заведений.

8.2.2.1 Обследования домашних хозяйств

Обследования домашних хозяйств являются основой социальных исследований и используются для определения основных характеристик населения (людей). Темы охватывают многие социально-экономические области, включая бедность, здоровье, образование, занятость, гендерное равенство, продовольственную безопасность и доступ к услугам.

Домохозяйство - это основная жилая единица, в которой организованы и осуществляются экономическое производство, потребление, наследование, воспитание детей и приют. Классификация «домохозяйство» шире, чем классификация «семьи», поскольку семья относится только к группе людей, связанных кровью или браком, например, только родители и их дети.

«Обследование домохозяйств» - это процесс сбора и анализа данных, которые помогают нам понять общую ситуацию и конкретные характеристики отдельного домохозяйства или всех домохозяйств в населении.

Для обследований домашних хозяйств обычно используются вопросники двух типов: а) списки домашних хозяйств; и б) подробные (или индивидуальные) анкеты. Реестр домохозяйства включает в себя список всех членов домохозяйства и их характеристики, такие как возраст, пол каждого члена, отношение к главе домохозяйства, уровень образования и грамотности, занятость, школьный статус (для населения в возрасте от 5 до 24 лет) и семейное положение. .

Подробный (или индивидуальный) опросник раскрывает основную тему исследования. Этот вопросник обычно заполняют только определенные респонденты, такие как глава семьи, супружеские пары, матери детей до пяти лет, дети, не посещающие школу, или дети из неблагополучных семей и т. Д., Поскольку он должен давать надежные результаты для всего населения. , такое обследование требует значительного числа респондентов и, как правило, является довольно дорогостоящим в администрировании. Он может проводиться так часто, как ежемесячно, или только время от времени, в зависимости от того, насколько быстро собираемые данные могут изменить ценность и, в зависимости от бюджета, доступны для проведения обследования.

8.2.2.2 Бизнес-опросы

Обследования предприятий являются основой экономических исследований и используются для определения основных характеристик предприятий и экономики. Их можно разделить на *краткосрочные* и *структурные* исследования. Краткосрочные опросы бизнеса служат для определения изменения выпуска продукции в отрасли или явлений между периодами измерения и используются для отслеживания бизнес-цикла. Они проводятся ежемесячно или ежеквартально, если позволяют ресурсы, в противном случае - реже, в зависимости от имеющегося бюджета.

Краткосрочные бизнес-показатели обычно включают производство, оборот, отработанное время, количество занятых, заработную плату, экспорт и импорт, а также цены производителей и импортных товаров в различных секторах, таких как промышленность, строительство, торговля и услуги. С другой стороны, структурные обследования предприятий обычно проводятся ежегодно или реже и направлены на предоставление подробной информации о структуре, типе деятельности, конкурентоспособности и эффективности экономической деятельности в рамках экономики или сектора. В то время как структурные обследования предприятий в основном используются для составления годовых национальных счетов, краткосрочные ежемесячные или ежеквартальные обследования предприятий измеряют изменения и, следовательно, предоставляют исходные данные для оценки изменения объема в квартальных национальных счетах.

В идеале каждое обследование предприятий использует статистический регистр предприятий в качестве источника совокупности обследования (как более подробно описано в главе 11, раздел 11.3 « *Статистический регистр предприятий* »). В зависимости от используемой статистической единицы обследование предприятий может быть *обследованием предприятий* или *заведений* .

а) **Опросы предприятий**

Согласно глоссарию статистических терминов Евростата , предприятие - это *организационная единица, производящая товары или услуги, которая имеет определенную степень автономии в принятии решений. Предприятие может вести более одного вида экономической деятельности и находиться в нескольких местах. Предприятие может состоять из одной или нескольких юридических единиц* .

Юридическая единица может быть физическим или юридическим лицом, существование которого признано законом независимо от лиц или организаций, которые могут владеть им или членом которых оно является. Примерами являются полное товарищество, частное товарищество с ограниченной ответственностью, компания с ограниченной ответственностью и зарегистрированная компания. Большинство предприятий состоят из одной юридической единицы. Однако во многих странах несколько предприятий, состоящих из более чем одной юридической единицы, составляют огромную часть экономики с точки зрения занятости или добавленной стоимости.

Юридическая единица может владеть второй юридической единицей, и эта вторая юридическая единица может осуществлять деятельность исключительно для этой первой юридической единицы. Оба подразделения могут даже иметь одинаковое руководство. В этом случае они рассматриваются как единое предприятие. Другим примером может быть то, что юридическая единица C нанимает сотрудников, а юридическая единица D владеет средствами производства, такими как машины и здания. Третья юридическая единица E может владеть этими двумя юридическими единицами и управлять ими. Только единицы C, D и E вместе могут что-то производить и, следовательно, должны считаться одним предприятием. Причины разделения предприятия организационной единицы на более чем одну юридическую единицу могут быть самыми разными: среди них - уход от налогов или обязательств,

разная заработная плата в соответствии с коллективным соглашением о заработной плате или отказ от публикации годовых отчетов.

Глобализация еще больше способствовала созданию более сложных структур предприятий. Чтобы быть активным на рынке в стране, часто требуется, чтобы у предприятия было юридическое лицо в этой стране. Юридические единицы такого предприятия могут управляться централизованно из одной страны, бухгалтерский учет может вестись централизованно из другой страны, исследования и разработки могут проводиться в стране с высокой заработной платой, а части производства - в странах с низким уровнем заработной платы. заработная плата. Многонациональные предприятия часто являются очень крупными предприятиями, оказывающими огромное влияние на статистику с точки зрения занятости и добавленной стоимости. Таким образом, качественные данные о многонациональных предприятиях имеют решающее значение для качественной коммерческой статистики и требуют активизации сотрудничества между различными национальными статистическими органами. Это сотрудничество часто осуществляется с помощью профилирования - процесса, позволяющего разграничить сложные и крупные предприятия.

б) **Обследования предприятий**

Заведение - это предприятие или часть предприятия, которое расположено в одном месте и на котором осуществляется только один производственный вид деятельности или в котором основная производственная деятельность составляет большую часть добавленной стоимости. Обследования предприятий - это любые обследования, в которых данные собираются от местных единиц. Сбор данных по каждой местной единице часто затруднен, поскольку это создает значительную нагрузку как для респондентов, так и для НСУ. В связи с этим количество опросов заведений часто ограничено. Однако, поскольку обследования предприятий предоставляют точную и подробную информацию на самом низком уровне, их проведение важно для качества национальных счетов. Когда обследование заведений недоступно для конкретной статистической области, имеющиеся данные обследования заведений (обычно занятость и заработная плата) могут использоваться для получения оценок других переменных.

8.2.3 Типы статистических единиц, которые могут быть предметом обследования

Общий закон об официальной статистике (GLOS) определяет « *статистическую единицу»* как носитель статистических характеристик. Это основная *единица наблюдения.* Согласно Общей структуре метаданных ЕЭК ООН , статистические единицы - это субъекты, респонденты опроса или предметы, используемые для целей расчета или измерения. Это единицы наблюдения, по которым собираются или выводятся данные. Это могут быть, среди прочего, предприятия, государственные учреждения, отдельные организации, учреждения, люди, группы, географические районы или события. Они образуют совокупность, от которой можно собирать данные или над которой можно проводить наблюдения.

В публикации СООН « Статистические единицы» (2007 г.) проводится различие между *статистическими, собирающими* и *отчетными единицами* . *Единица сбора данных* определяется как единица, из которой получены данные и с помощью которой заполняются анкеты. Согласно этому определению, единица сбора - это больше *контактный адрес,* чем *единица* . Например, анкета может быть заполнена центральным административным офисом или бухгалтерской фирмой, которая предоставляет эту услугу своему клиенту (подразделению наблюдения). Такой объект, предоставляющий информацию, называется единицей сбора.

Отчетности единица является единицей , о которой сообщают данные. Подотчетные единицы - это те организации, информация о которых собирается с помощью анкет или интервью. Отчетные единицы в большинстве случаев совпадают с единицами, по которым составляется статистика, т. Е. Единицами наблюдения.

Согласно Международной стандартной отраслевой классификации всех видов экономической деятельности (МСОК) Ред. 4 статистические единицы в экономической статистике можно разделить на следующие категории:

а) предприятие;
б) группа предприятий;
в) единица вида деятельности (КАУ);
г) местное подразделение;
д) учреждение или местная единица вида деятельности;
е) однородная единица продукции

8.2.4 Дизайн обследования

Дизайн обследования начинается с потребностей пользователей, общее содержание обследования, основные концепции и определения, а также доступность данных были оценены, а цели обследования были определены (как описано в главе 6). Как правило, это итеративный процесс, который начинается с определения целевой группы населения и оценки доступных источников для создания основы обследования. Структура обследования включает список единиц, наиболее точно соответствующих целевой совокупности, вместе с данными, необходимыми, во-первых, для стратификации этих единиц для отбора выборки и, во-вторых, для идентификации выбранных единиц и установления контакта с ними. Таким образом, как более подробно описано в главе 11, фрейм содержит имена и другие данные, необходимые для идентификации (идентификационные данные), адреса, номера телефонов и адреса электронной почты (контактные данные) и другие метаданные, используемые для стратификации и выборки.

8.2.4.1 Определение совокупности и построение основы выборки

Выбор рамки обследования может привести к определению обследуемой совокупности, которая отличается от целевой группы, т. Е. Люди с непостоянным адресом или предприятия с 0 сотрудников исключаются из обследования, но это также может повлиять на методы сбора данных. сбор, выборка и оценка, а также стоимость обследования и качество результатов. Кроме того, обследуемое население может исключить районы с чрезвычайно высокими затратами на сбор (например, удаленный остров), если они охватывают незначительную часть целевой популяции. Есть две основные категории фреймов: список и фреймы области. Рамка списка - это список всех единиц в обследуемой совокупности, а рамка области - это особый вид рамки с иерархией географических областей как единиц. В типичном обследовании домашних хозяйств используются оба типа основ в качестве источников для отбора выборки (многоэтапная выборка), где сначала выбирается выборка регионов из территориальной совокупности, а затем систематическая выборка жилищ (обычно стратифицированная) выбирается из каждой из выбранные регионы.

В официальной статистике основа обследования (которая в выборочных обследованиях называется основанием выборки) обычно выводится из статистических регистров или переписей, но для этой цели все чаще используются различные административные регистры (независимо или как метод улучшения качества основы). Использование согласованной совокупности обследования рекомендуется для обследований с той же целевой группой населения или подмножеством целевой группы населения. Когда одна и та же основа выборки используется в разных обследованиях, она называется эталонной

основой выборки или эталонной выборкой. При таком подходе первый этап выборки (выбор территорий) выполняется один раз для всех проведенных обследований, а окончательный отбор домохозяйств - для каждого обследования. Это позволяет избежать несоответствий между обследованиями и снижает затраты, связанные с обслуживанием и оценкой рамок. Более подробную информацию об эталонных образцах можно найти в главе 11, раздел 11.7 « *Основные образцы для домашнего хозяйства»,* а об их использовании - в разделе 8.7 этой главы.

Наличие надежной совокупности обследований имеет важное значение для любого обследования, и, таким образом, отсутствие надежной совокупности может привести к выбору переписи (на основе территориальной регистрации), а не выборочного обследования для первоначального сбора данных.

8.2.4.2 Типы отбора проб

Двумя основными типами выборки являются вероятностная выборка и не вероятностная выборка. Вероятностная выборка описывает процедуры, при которых каждая единица в генеральной совокупности имеет шанс быть выбранной в выборке, и эта вероятность может быть точно определена. Невероятностная выборка - это любой метод выборки, при котором некоторые элементы генеральной совокупности не имеют шанса на выбор или вероятность выбора не может быть точно определена. Невероятностная выборка имеет ограниченное применение для обследований, проводимых НСУ, поскольку необъективный выбор единиц может привести к ложным выводам об обследуемой совокупности, поскольку результаты не могут быть легко обобщены на совокупность. Тем не менее, есть области статистики, в которых полезна маловероятная выборка - например, краткосрочная бизнес-статистика, где часто используется пороговая выборка. Отсеченная выборка - это метод, при котором обследуются все единицы выше определенного порога. Кроме того, маловероятная выборка может быть полезна для поисковых исследований или на этапе разработки обследования (например, для тестирования анкеты).

Вероятностная выборка должна использоваться, когда выводы о совокупности должны быть сделаны на основе результатов обследования. В вероятностной выборке каждая единица кадра имеет ненулевую вероятность быть выбранной, и единицы выбираются случайным образом. В результате выбор является беспристрастным, и можно рассчитать вероятности включения, вычислить дисперсию оценок выборки и сделать выводы о генеральной совокупности. Основным недостатком вероятностной выборки является то, что она требует больше времени, является более дорогостоящей, чем не вероятностная выборка, и требует высококачественной основы выборки, а также по крайней мере одного специалиста по выборке, что может быть роскошью для небольших НСУ.

8.2.4.3 Планы отбора проб

Существуют различные методы и типы планов статистической выборки. Выбор подходящего и эффективного плана зависит от характера доступной основы выборки, а также от материальных затрат, времени, выделенного на проведение обследования, и характера сложности единиц выборки. Дизайн выборки также может влиять на требуемый размер выборки и, таким образом, оказывает значительное влияние на общую стоимость сбора данных. Это также зависит от степени изменчивости изучаемого свойства среди единиц населения. Простейшие планы вероятностной выборки - это простая случайная выборка и систематическая выборка, которые приводят к равной вероятности включения. Более сложные схемы, которые могут привести к неравным вероятностям включения и большинство из которых требуют вспомогательной информации, включают стратифицированную, пропорциональную размеру, кластерную, многоступенчатую и многофазную выборку. Планы с неравной вероятностью обычно используются для повышения статистической эффективности стратегии выборки или снижения затрат на

выборку. Иногда их использование диктуется рамкой выборки. При выборе между различными возможными вариантами дизайна первое, что нужно определить, - это то, какие варианты осуществимы с учетом структуры обследования, единиц в структуре обследования, областей интереса, нагрузки на респондентов, метода сбора данных, бюджета и международного опыта и т. Д. информацию о структуре выборки и ее отборе можно найти в главе 11, раздел 11.8.2 « *Дизайн и оценка выборки»* .

Ссылки на руководства, передовой опыт и примеры:

- Евростат: Справочные руководящие принципы выборки для обследований .
- Статистическое управление Канады - методы и практика проведения обследований .
- Статистическое управление Нидерландов - Теория выборки: план выборки и методы взвешивания .
- Статистическое управление Нидерландов - История выборки .
- СООН Отбор кадров и мастер - образцы .

8.2.5 Сбор и режимы сбора данных

Выбор подходящего метода сбора данных имеет важное значение для успеха исследования. Следует выбирать метод сбора данных, обеспечивающий широкий охват, высокую скорость ответа и сбор точной информации, в то же время минимизируя бремя сбора и имея разумную стоимость. Все это не может быть достигнуто одновременно, поэтому выбор правильного подхода из множества доступных вариантов обычно зависит от конкретного обследования. В соответствии с методами и практикой обследования (2003 г.) - Статистическим управлением Канады методы сбора данных можно разделить на методы самостоятельного заполнения и методы с помощью интервьюера.

8.2.5.1. Самостоятельное заполнение

При самостоятельном заполнении респондент заполняет анкету без помощи интервьюера. Вопросник может быть доставлен и возвращен респондентом разными способами: например, по почте или факсу, в электронном виде или с помощью счетчика. В бумажном формате этот метод называется собеседованием с использованием бумаги и карандаша (PAPI); при использовании компьютера это называется самоинтервью с помощью компьютера (CASI); Компьютерное или компьютерное веб-интервью (CAWI).

Основное преимущество самостоятельного заполнения перед другими методами заключается в том, что он дешевле, чем методы с помощью интервьюера, и его гораздо проще применять, поскольку нет необходимости в управлении интервьюерами. Самостоятельное заполнение также полезно для деликатных вопросов, так как анкету можно заполнить наедине, без интервьюера.

Недостатком самостоятельного заполнения является то, что процент ответов обычно ниже, чем при использовании методов с помощью интервьюера, поскольку респондента не требуется заполнять анкету. Кроме того, качество может быть хуже, чем при использовании методов с помощью интервьюера, поскольку респондент может неверно истолковать информацию или пропустить пропуски в бумажной форме. По этой причине самозаполнение часто требует последующих действий после сбора для исправления ошибок.

Обычно используются три метода самостоятельного заполнения:

а) **Самостоятельное заполнение бумажной анкеты**

Еще несколько лет назад бумажные анкеты были наиболее широко используемым методом сбора данных. Из-за низкой стоимости самозаполнение с использованием бумажных вопросников в основном использовалось в крупномасштабных мероприятиях по сбору данных, таких как переписи, но также и в повторяющихся ежемесячных опросах предприятий, поскольку этот метод сбора подходит для сбора подробной информации. Этот метод часто является медленным и дает довольно низкую частоту ответов и часто требует дополнительных напоминаний и разъяснений. Несмотря на то, что статистические управления значительно сократили количество бумажных вопросников в последние годы, они все еще используются для крупномасштабных мероприятий по сбору данных (например, переписей); для сбора конфиденциальной информации (например, исследований, связанных с насилием); для ведения дневников (в обследовании использования времени) и для ведения расходов (в бюджетных обследованиях). Анкеты могут быть доставлены и отправлены обратно по почте, или они могут быть доставлены или забраны интервьюерами, или может быть выбрана комбинация вариантов. Когда бумажный вопросник доставляется и возвращается по почте, для самостоятельного заполнения требуется длительный период собеседования, поскольку это самый медленный метод сбора данных, но также и самый дешевый метод сбора данных на бумажном носителе. Самостоятельное заполнение бумажного вопросника также требует ввода данных, который часто необходимо выполнять вручную или, по крайней мере, проверять вручную после использования оптического распознавания символов (OCR). Можно утверждать, что Интернет-собеседование в некоторых случаях может быть дешевле, но это зависит от стоимости подготовительных мероприятий, которые обычно довольно высоки.

б) **Самостоятельное заполнение с помощью электронной анкеты**

Многие статистические управления в последние годы заменили большинство своих бумажных вопросников электронными вопросниками, которые в основном доставляются в режиме онлайн через зашифрованную и защищенную паролем часть веб-страницы статистического управления. Этот метод часто называют CAWI (компьютерное веб-интервью). Для статистических управлений основным преимуществом CAWI перед бумажной формой является отсутствие ввода данных, поскольку это делает респондент. Кроме того, электронные анкеты предоставляют возможность интегрировать логические элементы управления, которые направлены на предотвращение ошибок и обеспечение соблюдения ответов.

Недостатком электронных вопросников является невозможность изначально распространять формы в электронном виде, поскольку страны редко имеют полный список адресов электронной почты или официальных цифровых почтовых ящиков для каждого гражданина и каждого предприятия. Таким образом, наиболее распространенный способ введения электронных вопросников - это рассылка начальной информации для входа в систему и требование к респондентам предоставить контактные адреса электронной почты. Небольшое предостережение: несмотря на то, что сбор происходит намного быстрее и качество собранных данных превосходит бумажные вопросники, не следует ожидать высоких показателей отклика. Последующие напоминания для заполнения анкеты часто необходимы, но их можно автоматизировать, и эту функцию следует запланировать на этапе разработки сбора данных. Самостоятельное заполнение электронной анкеты в настоящее время является наиболее распространенным методом сбора данных при повторных опросах предприятий.

с) **Перенос от машины к машине**

Хотя передачу данных от машины к машине можно классифицировать как новый и отдельный метод сбора данных, он имеет много характеристик самонаблюдения. Межмашинная передача - это автоматизированная передача информации из ИТ-системы респондента в систему сбора данных статистического управления. Первоначальная настройка повторяющейся передачи данных должна быть согласована с респондентом, который предоставляет доступ к заранее определенному набору информации в заранее определенные периоды сбора с помощью заранее определенного метода связи. Процесс обычно выполняется путем написания вручную или автоматически запускаемого сценария, который отправляет данные через интерфейс прикладного программирования (API), открытый статистическим бюро. Существуют примеры, когда поставщики программного обеспечения для бухгалтерского учета автоматизировали статистическую отчетность для предприятий, написав сценарии, которые автоматически подготавливают запрошенные данные, которые затем автоматически загружаются на веб-сайт сбора данных. Основное преимущество этого метода заключается в том, что он устраняет повторяющуюся нагрузку на предприятия и является самым быстрым методом сбора данных. Основным недостатком является его стоимость, поскольку подготовка и документация для API могут быть дорогостоящими, часто требует прямого участия ИТ-персонала с обеих сторон, а после его создания часто требуется убеждение для широкого внедрения. Передача от машины к машине также может использоваться для передачи информации об отчетных единицах от единиц сбора (таких как бухгалтерские фирмы) или для доступа к источникам больших данных.

8.2.5.2 Сбор данных с помощью интервьюера

Основное преимущество методов с участием интервьюера заключается в том, что интервьюер, персонализируя интервью и имея возможность интерпретировать вопросы и концепции опроса, может повысить частоту ответов и общее качество собранной информации. Методы с участием интервьюера особенно полезны для обследуемых групп населения с низким уровнем грамотности или когда концепции или анкета являются сложными или когда само заполнение затруднительно. Особым случаем, когда предпочтение отдается методам с помощью интервьюера, являются опросы, целью которых является сбор информации о деликатных темах, таких как опросы о насилии в отношении женщин или обследования гендерного насилия. В этих случаях разрабатываются и соблюдаются специальные протоколы. Интервьюер может увеличить количество ответов, предъявив официальное удостоверение личности и стимулируя интерес к опросу и заверив респондента в любых опасениях, которые у него могут возникнуть в отношении: конфиденциальности данных, цели опроса, ожиданий от респондента. во время интервью, сколько времени будет длиться и как будут использоваться результаты опроса, и т. д. Интервьюер может предотвратить ошибки, немедленно выявляя и исправляя их в присутствии респондента.

Некоторые недостатки методов с участием интервьюера заключаются в том, что они дороги и сложны в использовании. Некоторые из расходов могут включать в себя заработную плату интервьюеров, обучение интервьюеров, расходы на транспорт и проживание для интервьюеров или офисные помещения и телефоны в случае централизованного телефонного интервью. Основная проблема с личными интервью заключается в том, что может быть трудно найти респондента, поэтому интервьюеру может потребоваться несколько поездок, прежде чем успешно связаться с респондентом. Другими недостатками методов с участием интервьюера являются то, что плохо обученные интервьюеры могут вызывать ошибки в ответах, а респонденты могут неохотно отвечать на вопросы по деликатным темам. Если хорошо подготовленные интервьюеры недоступны, могут быть предпочтительны другие методы интервьюирования.

Чаще всего используются три метода интервьюирования с помощью интервьюера.

а) **Личное интервью с использованием бумажной анкеты**

Личные интервью с использованием бумажных вопросников (PAPI) заменяются компьютерными методами, хотя все еще используются для пилотных исследований. Существенным преимуществом перед самозаполнением с использованием бумаги является то, что, если интервьюеры выбраны для качественного почерка и / или прошли дополнительное обучение в этом отношении, то данные из анкет, которые они заполняют, можно будет легче захватить с помощью оптического распознавания символов. Соответствующее снижение затрат на OCR может перевесить затраты на счетчиков. Кроме того, особенно для стран, которые имеют установленный и чрезвычайно короткий период сбора данных для переписи, PAPI все еще может быть наиболее подходящим способом сбора. Например, в некоторых странах Латинской Америки объявлен *выходной день переписи* , установлены особые правила для населения, а подсчет проводится в один день с большим количеством интервьюеров.

б) **Компьютерное личное интервью**

Компьютерное личное собеседование (CAPI) сочетает в себе преимущества методов интервьюера и компьютерных методов, ускоряет процесс сбора данных, обеспечивает более сложные схемы пропуска, автоматическое редактирование и проверку качества, но также помогает в управлении и контроле персонала, проводящего интервью. Системы CAPI могут быть разработаны для генерации управленческих отчетов о статусе интервью (например, процент ответов, количество завершенных интервью, количество невыполненных, продолжительность интервью и т. Д.), Которые полезны для мониторинга качества и управления обследованием.

Существенным преимуществом методов CAPI является возможность автоматического сбора дополнительных данных (таких как геоинформация - широта и долгота устройства) и метаданных (например, время опроса), которые могут использоваться для связывания данных, а также для управления интервьюеры. Основным недостатком метода CAPI является стоимость оборудования и тот факт, что интервьюеры должны быть обучены и хорошо владеть компьютером и приложением для сбора данных. Следовательно, использование CAPI может оказаться не идеальным решением ни для опросов с большим количеством интервьюеров, ни для опросов, в которых высок процент замены интервьюеров.

Компьютер может быть ноутбуком, планшетом или даже мобильным телефоном. Последние два типа устройств меньше по размеру, потребляют меньше энергии и, следовательно, имеют значительно лучшее время автономной работы, что является важным фактором автономности интервьюера. Однако ввод через сенсорные экраны обычно медленнее и менее точен, чем ввод с клавиатуры на ноутбуке. Выбор технологии более подробно обсуждается в главе 14.

с) **Компьютерное телефонное интервью**

Компьютерное телефонное интервью (CATI) - это метод сбора данных с помощью интервьюера, при котором информация собирается по телефону и напрямую импортируется в компьютерную систему. Телефонные интервью предлагают разумную частоту ответов по разумной цене. Телефонные собеседования проходят быстрее и дешевле личных собеседований, так как отсутствуют командировочные и сопутствующие расходы. На общение с респондентом тратится меньше времени, а контроль качества процесса интервьюирования может быть легко осуществлен, поскольку телефонные интервью можно легко отслеживать. Недостатком телефонных опросов является то, что они ограничены длиной интервью и сложностью

анкеты, поскольку респонденты не терпят длинных и сложных интервью по телефону, чем при личной встрече. Более частое использование мобильных телефонов также повлияло на проведение телефонных интервью .

Это привело к увеличению затрат (поскольку звонки на мобильные телефоны, как правило, дороже, чем звонки на наземные линии), и сделало этап подготовки (в частности, построение кадра обследования с хорошим покрытием) намного более трудным и сложным. Кроме того, это несколько снизило скорость ответа, поскольку у пользователей мобильных телефонов функция идентификатора вызова включена по умолчанию, и они с меньшей вероятностью будут отвечать на звонки от неизвестных абонентов, чем пользователи стационарных телефонов. Частично это можно уменьшить, если заранее отправить письмо с объявлением о телефонном звонке для опроса. Кроме того, CATI отлично подходит для повторных опросов и последующих действий, когда респондент уже предоставил надежную контактную информацию и дал согласие на ответ. Есть много примеров, когда первоначальный контакт осуществляется лично через CAPI, а дальнейший сбор данных осуществляется через CATI (либо интервьюером, который проводил первоначальный сбор, либо через колл-центр).

Технологические разработки, такие как синтез речи и распознавание естественного языка в сочетании с системами искусственного интеллекта, были протестированы для сбора информации от респондентов по телефону (пример можно найти здесь). Такие технологии могут быть использованы в официальной статистике, особенно для простого сбора данных (например, подтверждения контактной информации от предприятий). Это может привести к автоматическому сбору данных по телефону (дополнительную информацию о возможностях использования систем ИИ можно найти в главе 14, раздел 14.2.15 « *Искусственный интеллект»*).

8.2.5.3 Соответствующий выбор режима

В соответствии с методами и практикой обследования (2003 г.) - Статистическим управлением Канады при выборе подходящего метода сбора данных следует учитывать следующие вопросы:

а) Сбор информации, доступной во фрейме обследования - если фрейм не включает почтовые адреса, то вопросники для самостоятельного заполнения не могут быть отправлены респондентам по почте. При отсутствии актуальных номеров телефонов интервью по телефону проводить нельзя.

б) Характеристики целевой группы населения влияют на метод сбора данных - если уровень грамотности населения низкий или если учитывается язык (т. е. существует две или более языковых групп), методы с помощью интервьюера могут быть единственными вариант. Если население и выборка разбросаны по стране, личные интервью могут оказаться слишком дорогими и трудными для проведения.

с) Характер вопросов опроса влияет на сбор данных - если предмет является конфиденциальным, то метод сбора, основанный на анонимности, такой как самостоятельное заполнение или телефонное интервью, может быть наиболее подходящим. Если задаются сложные вопросы, может потребоваться интервьюер, чтобы объяснить вопросы и концепции. Если интервьюеру необходимо провести наблюдения или измерения (например, провести тест на грамотность для детей) или показать материал респондента (например, графики или диаграммы), то могут потребоваться личные интервью.

г) Доступные ресурсы сильно влияют на выбор метода сбора данных - эти ресурсы включают доступный бюджет, персонал, оборудование и время. Чтобы использовать

метод с участием интервьюера, необходимо иметь достаточный бюджет для оплаты обучения, найма и поездок интервьюеров. Статистическому агентству также необходимо найти необходимое количество интервьюеров. Если выбран компьютерный метод, то требуются ИТ-специалисты вместе с необходимым компьютерным оборудованием.

е) При выборе метода сбора данных следует учитывать требования к качеству данных. Интервьюеры, хорошо обученные концепциям, используемым в опросе, могут уменьшить количество ошибок в ответах и отсутствия ответов. Также следует учитывать требования к точности: более крупные выборки обычно дают более точные оценки, но это более дорогой метод сбора данных.

f) Выбор подходящего метода обычно зависит от обследования и, наряду с вышеупомянутыми факторами, часто зависит от окружающей среды и руководителя обследования. Некоторые руководители опросов не хотят вводить новшества, в то время как другие активно ищут новые методы. Поэтому иногда необходимо продвигать новые практики и поощрять модернизацию конкретных видов деятельности (обычно тех, которые представляют наибольшую нагрузку на респондентов). Таким образом, личные интервью часто являются самым дорогим методом, а самоанализ - наименее затратным. Использование компьютерных методов повышает качество и скорость, но также увеличивает затраты. Также может иметь значение способность измерять качество и внедрять процедуры контроля качества. Следить за качеством телефонных интервью легче, чем личных интервью.

ж) Часто наиболее эффективным решением может быть комбинация различных методов. Этот подход часто называют сбором данных в смешанном режиме. Смешанный режим особенно полезен для опросов, которые требуют нескольких интервью с одними и теми же респондентами. Это может привести к первоначальному визиту интервьюера, который сможет ответить на все вопросы и собрать исходную информацию, а более подробная или последующая информация будет собрана с помощью методов CAWI или CATI. Сбор данных в смешанном режиме также можно комбинировать с использованием административных данных, чтобы снизить нагрузку на респондентов.

8.2.5.4 Дизайн анкеты

Анкета должна быть разработана таким образом, чтобы минимизировать возможные ошибки ответа. Это включает стандартизацию вопросов, а также стандартизацию объяснений для респондентов и интервьюеров. Формат анкеты также важен, но часто зависит от метода сбора данных. Введение и последовательность вопросов (рекомендуется начинать с вопросов, которые относятся к теме опроса, но на которые легко ответить) могут улучшить участие респондентов. Следует использовать утверждения, вводящие новые темы, а инструкции для респондента или интервьюера должны быть четкими, краткими и доступными. Общий формат и дизайн анкеты должны быть оценены с точки зрения их воздействия на респондента и интервьюера: включая шрифт, заголовки разделов, цвет анкеты, формат категорий ответов, наглядные пособия и т. Д. Наконец, как анкета должна быть следует учитывать: он должен быть разработан таким образом, чтобы облегчить сбор и сбор данных, что особенно важно для сбора на бумажных носителях.

Предварительный вариант анкеты должен быть протестирован и тщательно отредактирован перед окончательной доработкой анкеты. Тестирование может включать неформальное тестирование, когнитивное тестирование, фокус-группы, опросы интервьюера, кодирование поведения, тесты с разделенной выборкой и пилотное тестирование - методы тестирования подробно описаны в ссылках, приведенных ниже.

Создание хорошей анкеты - это сочетание науки, опыта и иногда немного искусства. Хорошо составленная анкета - это вопросник, который эффективно собирает данные с минимальным количеством ошибок и в то же время прост для ответа и администрирования, не создавая ненужного бремени для респондента и статистического агентства. Достижение хорошего баланса между этими целями может быть достигнуто за счет итеративного процесса разработки анкеты, который включает в себя многочисленные консультации, обзоры, тестирование и пересмотр.

Процесс обычно начинается с изучения всех требований к информации, которые должны быть выполнены, после чего следует индивидуальное рассмотрение каждого вопроса, направленное на поиск явного обоснования для включения в анкету. Должно быть известно, почему задается каждый вопрос и как эту информацию использовать. Формулировка вопроса должна быть ясной. Вопросы должны следовать в логической для респондента последовательности. Вопросы должны быть составлены таким образом, чтобы респонденты могли легко их понять и на них могли точно ответить.

Вопросы могут быть двух типов: открытые и закрытые. Открытые вопросы позволяют самовыражаться, но могут быть обременительными и трудоемкими, а также трудными для анализа. Закрытые вопросы могут быть вопросами с двумя вариантами ответов, вопросами с несколькими вариантами ответов и вопросами ранжирования или рейтинга. Закрытые вопросы обычно менее обременительны для респондента, а сбор и сбор данных дешевле и проще. Однако неправильный выбор категорий ответа может вызвать ошибку ответа.

Согласно методам и практике проведения опросов Статистического управления Канады, при формулировке вопроса обследования следует придерживаться следующих рекомендаций:

а) быть простым;

б) определять акронимы и аббревиатуры;

с) обеспечить применимость вопросов;

г) быть конкретным;

д) избегайте двусмысленных вопросов;

е) избегать наводящих вопросов;

ж) избегать использования двойных отрицаний;

з) смягчить влияние деликатных вопросов;

i) убедитесь, что вопросы читаются хорошо.

Разработка анкеты часто осуществляется в сотрудничестве с различными отделами статистического агентства, и часто бывает целесообразно назначить человека (или подразделение), которое будет отвечать за окончательное утверждение анкет. Более подробную информацию об инструментах для поддержки разработки вопросника можно найти в разделе 11.8.1 главы 11 « *Разработка вопросника»* .

Ссылки на руководства, передовой опыт и примеры:

- Статистическое управление Канады - методы и практика проведения обследований .
- Статистическое управление Нидерландов - Разработка анкеты .
- Статистическое управление Швеции - Правильно составьте вопросы - Как разрабатывать, тестировать, оценивать и улучшать анкеты .

8.2.5.5 Минимизация ошибок ответа

Ошибки ответа представляют собой неточность ответов на вопросы. Их можно отнести к различным факторам, включая вопросник, требующий улучшений, неправильное толкование вопросов интервьюерами или респондентами и ошибки в заявлениях респондентов.

Одна из распространенных стратегий уменьшения количества ошибок в ответах - использование терминологии, понятной респонденту, поскольку язык, используемый статистиками, может быть незнаком респонденту (домашнему хозяйству или бизнесу). Использование электронных анкет также может привести к значительному сокращению количества ошибок в ответах, поскольку в вопросы могут быть включены различные условия (например, возрастное ограничение 0-120 лет для вопроса о возрасте респондента, этот доход должен быть больше, чем прибыль, или что общий должно быть суммой его частей).

Ошибки ответа следует анализировать на этапе обработки, и если выявляется общая ошибка ответа, следует принять меры для минимизации ее при последующем сборе данных.

Ссылки на руководства, передовой опыт и примеры:

- Уменьшение количества ошибок ответа в опросах .

8.2.6. Обработка опроса

После сбора данные обследования требуют дополнительной обработки, прежде чем они будут проанализированы и объединены в статистические результаты, и то же самое верно для административных данных, которые используются в статистических целях. Обработка преобразует ответы на опрос, полученные во время сбора, в форму, удобную для составления таблиц и анализа данных. Он включает в себя все действия по обработке данных - автоматические и ручные - после сбора и до оценки. В соответствии с этапом обработки данных GSBPM делятся на подпроцессы, которые объединяют, классифицируют, проверяют, очищают и преобразуют входные данные. Основное внимание в этой главе будет уделено редактированию, кодированию, условному исчислению и обнаружению выбросов, которые более подробно рассматриваются в главе 11, раздел 11.8.1.2 « *Инструменты разработки вопросников».*

8.2.6.1. Редактирование

Редактирование - это применение проверок для выявления отсутствующих, недопустимых или несогласованных записей, указывающих на записи данных, которые потенциально содержат ошибки. Цель редактирования - лучше понять данные, чтобы гарантировать полноту, согласованность и достоверность окончательных данных. Редактирование может варьироваться от простых ручных проверок, выполняемых интервьюерами на местах или административными клерками в случае административных данных, до сложных проверок, выполняемых компьютерной программой. Объем выполненного редактирования - это компромисс между получением «идеальной» каждой записи и расходованием разумного количества ресурсов (времени и денег) на достижение этой цели. В то время как некоторые ошибки редактирования решаются путем последующей беседы с респондентом или ручного просмотра анкеты, исправить все ошибки таким образом практически невозможно, поэтому в оставшихся случаях часто используется вменение.

а) **Пункт редактирования коллекции**

Редактирование точки сбора используется для обнаружения ошибок, допущенных во время интервью респондентом или интервьюером, и для выявления недостающей информации во время сбора, чтобы уменьшить необходимость в последующих действиях. Редактирование во время сбора значительно проще осуществить, если оно автоматизировано с помощью метода сбора с помощью компьютера. В вопросниках

для самостоятельного заполнения респонденты могут редактировать свои ответы. Почти во всех опросах, проводимых с помощью интервьюера, некоторое редактирование выполняется во время интервью, и интервьюеры проинструктированы и обучены проверять ответы, которые они записывают в анкете, сразу после завершения интервью - либо после выхода из дома, либо после того, как повесили трубку. Таким образом, у них по-прежнему есть возможность обнаруживать и обрабатывать записи, которые не соответствуют правилам редактирования, либо потому, что правильная информация может быть свежа в их памяти, либо потому, что они могут легко и недорого связаться с респондентом, чтобы установить правильные значения.

В случае компьютерных методов сбора правки могут выполняться автоматически с помощью программных приложений. Для бумажных вопросников с ручным сбором данных экономично использовать сбор данных как возможность применить правила для очистки данных в достаточной степени, чтобы сделать последующие этапы обработки более эффективными. Как правило, редактирование во время сбора данных должно быть минимальным, поскольку реакция на сбой редактирования замедляет сбор данных. Правки на этом этапе обработки - это, в основном, правки на предмет достоверности и простые правки согласованности.

При работе с административными данными часто бывает полезно предложить поставщику данных включить правила автоматического редактирования в системы сбора, так как это может привести к значительному повышению качества административных данных. Статистическим управлениям следует использовать это как можно чаще, поскольку они могут прямо или косвенно извлечь выгоду из повышения качества административной информации.

б) **Первичное и вторичное редактирование**

Наиболее полное и сложное редактирование обычно выполняется после завершения сбора данных и когда материалы поступают в офис. В некоторых статистических системах этот процесс выполняется в несколько этапов, обычно называемых первичным и вторичным редактированием. Первый этап обычно выполняется в региональных офисах сразу после сбора данных, когда интервьюер может повторно связаться с респондентом и принять меры после выполнения базовой проверки и выявления ошибки или несоответствия. Более сложные правила редактирования обычно зарезервированы для отдельного этапа редактирования после захвата данных - наряду с правками на предмет достоверности часто выполняются более сложные правки согласованности, а также выборочное редактирование и обнаружение выбросов.

В случае ошибок редактирования после сбора данных обычная процедура состоит в том, чтобы пометить поле, для которого не удалось выполнить редактирование, а затем либо присвоить поле, либо исключить запись из дальнейшей обработки. Большинство ошибок редактирования на этом этапе помечаются для вменения. Значения, которые не удалось отредактировать, должны быть помечены специальным кодом, чтобы указать, что было сообщено о недопустимом значении или недопустимом пробеле. Эти флаги особенно полезны при оценке качества данных обследования. В некоторых случаях запись или анкета могут не соответствовать такому количеству правил редактирования - или небольшому количеству критических изменений - что становятся бесполезными для дальнейшей обработки. В таких случаях запись обычно рассматривается как не отвечающая, удаляется из потока обработки и выполняется корректировка веса неполучения ответа.

8.2.6.2 Кодирование

Кодирование - это процесс применения числовых значений к заданным ответам для облегчения обработки данных. Кодирование также может быть выполнено на этапе разработки обследования, когда готовятся вопросники. Это означает, что каждому возможному ответу присваивается заранее определенное числовое значение перед заполнением вопросника. Это делается как часть дизайна вопросника, и его довольно легко применять и управлять закрытыми вопросами и электронным сбором данных. Применение статистических классификаций к основным данным также называется кодированием. GSBPM предоставляет следующий пример: процедуры автоматического (или канцелярского) кодирования могут присваивать числовые коды текстовым ответам в соответствии с заранее определенной статистической классификацией для облегчения сбора и обработки данных. Некоторые вопросы имеют закодированные категории ответов в анкетах или административном источнике данных, другие закодированы после сбора с использованием автоматизированного процесса (который может применять методы машинного обучения) или интерактивного ручного процесса.

8.2.6.3 Обнаружение и обработка выбросов

Выброс - это наблюдение или подмножество наблюдений, которое, по-видимому, не согласуется с остальной частью набора данных. Обнаружение выбросов выполняется путем анализа всего набора данных и выявления неожиданных или экстремальных значений, обычно путем измерения их относительных расстояний от центра данных.

Выбросы, обнаруженные на этапе редактирования в процессе опроса, можно обрабатывать различными способами. В системе ручного редактирования потенциальные выбросы исследуются или отслеживаются и изменяются, если они на самом деле являются ошибками. В автоматизированной системе редактирования замещающие значения для выбросов часто рассчитываются условно. В некоторых случаях специальная обработка выбросов не проводится, если считается, что они не имеют большого значения.

Для обработки выбросов можно использовать следующие подходы:

а) изменить значение;

б) изменить вес;

с) использовать робастную оценку.

Статистики должны понимать свойства собранных данных, поскольку иногда экстремальные значения, которые могут быть обнаружены как выбросы, на самом деле представляют реальность (одним из таких примеров является сектор, в котором есть одна крупная компания и много мелких).

Дополнительную информацию о методах и инструментах определения выбросов можно найти в главе 11, раздел 11.8.1.2 « *Инструменты разработки вопросников».*

8.2.6.4 Вменение

Вменение - это процесс, используемый для определения и присвоения заменяющих значений для решения проблем с отсутствующими, недействительными или несогласованными данными. Это делается путем изменения некоторых ответов, чтобы гарантировать создание правдоподобной, внутренне непротиворечивой записи. Расчет обычно выполняется с помощью тщательно разработанной автоматизированной системы, которая использует характеристики всего набора данных и дополнительные данные (при их наличии) для предложения значения замены. Хорошее вменение имеет контрольный журнал для целей оценки. Вмененные значения должны быть помечены, а методы и источники вменения должны быть четко определены. Исходные и вмененные значения полей записи следует сохранить, чтобы можно было оценить степень и последствия вменения.

Хотя вменение может улучшить качество окончательных данных, следует проявлять осторожность, чтобы выбрать подходящую методологию вменения. Некоторые методы вменения не сохраняют взаимосвязи между переменными или могут исказить лежащие в основе взаимосвязи в данных, в то время как набор данных, требующий значительного количества вмененных значений, обычно является результатом неудач в плане обследования. Пригодность выбранного метода зависит от типа обследования, его целей и характера ошибки. Более подробную информацию о принципах, методах и инструментах вменения можно найти в разделе 11.8.1.2 главы 11 « *Инструменты разработки вопросника»* .

8.2.6.5 Макроредактирование, а именно редактирование на основе обзора агрегированных данных

Макроредактирование - это набор стратегий, направленных на сокращение количества микроредакторов и ручных проверок, которые должны выполняться клерками. Идея макроредактирования состоит в том, чтобы предоставить предварительные результаты перед этапом редактирования и проверить согласованность результатов перед переходом к этапу редактирования. Макроредактирование обычно выполняется путем выполнения анализа набора данных и определения верхнего и нижнего пределов данных, требующих проверки, а также для дальнейшего выполнения дополнительных проверок в зависимости от важности элемента на общем уровне. Макроредактирование может снизить общую стоимость редактирования, так как его можно использовать для определения приоритетов и сокращения проверки собранных данных.

Обзор методов редактирования макросов можно найти здесь .

8.2.6.6. Предварительный расчет

Оценка - это процесс, посредством которого НСУ вычисляет оценки, применимые к генеральной совокупности на основе данных выборки. Принцип оценки в вероятностном обследовании заключается в том, что каждая единица выборки представляет не только себя, но и ряд других подобных единиц в обследуемой совокупности. Оценка включает в себя присвоение (окончательной оценки) *веса* отклику каждой единицы в выборке, где вес указывает количество единиц, которые эта единица выборки представляет в генеральной совокупности.

Отправной точкой для определения подходящего веса единицы выборки является величина, обратная вероятности выбора единицы. Это зависит от дизайна образца и обычно называется *расчетным весом* . Определение этого веса - важная часть процесса оценки. Сумма проектных весов - это размер генеральной совокупности, из которой была выбрана выборка. В случае многоэтапного плана выборки вероятность отбора учитывается на всех этапах отбора. в многоэтапном дизайне отбора.

Затем расчетный вес корректируется для получения (окончательной оценки) веса. Двумя наиболее частыми причинами для внесения корректировок являются, во-первых, учет неполучения ответов и, во-вторых, повышение надежности оценки за счет использования вспомогательных данных.

После того, как окончательные оценочные веса были рассчитаны, они применяются к выборочным данным для вычисления оценок. Сумма весов - это общая численность населения, оцененная на основе данных выборки. Сводные показатели совокупности, такие как итоговые, средние и пропорции, обычно оцениваются по характеристикам, собранным из единиц выборки. Эти характеристики, которые в статистической теории часто называют *переменными* , могут быть качественными, например, пол или семейное положение, или количественными, например, возрастом или доходом. Существуют различные формулы для оценки суммарных показателей в зависимости от оцениваемых характеристик и плана выборки.

Информацию об инструментах оценки можно найти в главе 11, раздел 11.8.2 « *Дизайн и оценка выборки»* .

Ссылки на руководства, передовой опыт и примеры:

- ABS - Взвешивание и оценка стандартных ошибок для обследований домашних хозяйств .
- Справочник Memobust - Взвешивание и оценка .
- Реинжиниринг ежегодных экономических обследований Бюро переписи населения (2018 г.) - выборка и оценка .
- Статистическое управление Канады - методы и практика проведения обследований .

8.2.7 Отношения с респондентами и коммуникация

Ключевой задачей любого статистического агентства является постоянное повышение актуальности, повышение и сохранение доверия. Поскольку качество статистической продукции во многих случаях зависит от качества исходных данных, очень важно обеспечить сотрудничество респондентов. Методы обеспечения такого сотрудничества обычно делятся на два основных подхода, которые часто используются одновременно: использование юридических инструментов для принуждения к соблюдению или предотвращения неповиновения и активное общение и попытки апеллировать к чувству морали для поощрения сотрудничества.

Статистическое агентство должно заслужить общественное доверие, относясь к респондентам с уважением, а не только как к средству достижения своих статистических целей. Важно помнить, что даже при наличии законов, которые делают ответ на один или несколько сборов данных обязательными, участие общественности в обследованиях статистических агентств является в значительной степени добровольным процессом. Даже если процесс опроса не является добровольным, агентство по-прежнему обязано относиться к респондентам этично: то есть минимизировать нагрузку на их время, уважать их частную жизнь и сохранять конфиденциальность, которую им обещали при предоставлении информации.

Отношения с респондентами часто централизованы в специализированное подразделение, которое обычно организовано в отделе взаимодействия с пользователями, поскольку отношения с респондентами часто похожи на работу с трудными пользователями. Подразделение могло выполнять следующие задачи:

а) Обеспечение связей с общественностью, необходимых для потенциальных респондентов, чтобы понять, почему они были выбраны, что от них просят и какое общественное благо служит результатом их сотрудничества;

б) проявление особой осторожности и принятие всех необходимых мер предосторожности в случаях, когда объявленное обследование является либо необычно длинным (например, обследования семейных расходов), либо необычно назойливым (например, обследования потребления вредных наркотиков и обследования фертильности);

с) ведение реестра респондентов, с которыми связывались респонденты, и проведенных опросов, чтобы можно было выявить непокорных респондентов и убедить их принять участие;

г) Обмен информацией с респондентами, чтобы они чувствовали, что они не только внесли свой вклад в общественное благо, но и что есть некоторая личная выгода;

д) Выполнение этих задач требует такта и дипломатии, а также твердости и решимости. Всегда найдутся люди в домашнем хозяйстве или в деловом секторе, которые откажутся подчиняться, независимо от того, насколько убедительны доводы в пользу сотрудничества.

В следующем разделе будут приведены примеры передовой практики в отношениях с респондентами и даны советы о том, как общаться с респондентами.

8.2.7.1 Использование закона для принудительного реагирования

Подходы к этому вопросу различаются от страны к стране. В некоторых случаях соблюдение статистического сбора обязательно - если респонденты не предоставляют информацию в той форме, в которой она запрашивается, и своевременно, они нарушают закон. В других странах некоторые запросы на определенные классы информации являются обязательными и поддерживаются законодательными требованиями, в то время как другие выполняются на добровольной основе. Это, вероятно, наиболее распространенная ситуация, когда закон признает ограниченный набор обязательных обследований или предусматривает процедуру, посредством которой обследование может быть сделано обязательным. В этом случае статистические агентства обычно объявляют экономические расследования обязательными, а другие - добровольными. Наконец, есть случаи, когда в законе нет четкого определения предмета. Когда это правда, статистическое агентство может опасаться требовать слишком много информации: в случае оспаривания суд может постановить, что никакой информации не требуется в обязательном порядке, и результирующая огласка может отрицательно повлиять на количество ответов. Какой бы ни была правовая основа, все агентства считают, что наиболее важной целью является обеспечение сотрудничества со стороны респондентов, особенно со стороны малых предприятий и домашних хозяйств, поскольку принуждение редко может облегчить проблему реагирования.

Ссылки на руководства, передовой опыт и примеры:

- Общий закон об официальной статистике (Glos), ЕЭК ООН (2016)
- Руководство по модернизации законодательства в области статистики ЕЭК ООН 2018

8.2.7.2 Политика и устав респондента

Как отмечалось ранее, поддержание хороших отношений с респондентами является критическим фактором успеха любого статистического агентства. Как и в других областях, стандартизированный, прозрачный и единообразный подход к респондентам обычно дает наилучшие результаты. Примером передовой практики является наличие четкого и легкодоступного документа, иногда представленного в форме *устава респондента,* который предоставляет респондентам основную информацию, чтобы уведомить их об общей цели официальной статистики, проинформировать их о необходимости сбора данных поскольку запрошенная информация не может быть получена другими способами, и заверить их, что информация будет храниться в безопасности и останется конфиденциальной. Обычно это достигается с помощью специальной веб-страницы, на которую респонденты направляются через письма или электронные сообщения, сопровождающие сбор данных. Уставы респондентов особенно полезны для обследований предприятий, поскольку их можно использовать для обеспечения того, чтобы предприятия использовали соответствующий метод сбора данных и что нагрузка на респондентов не была чрезмерной. Наличие устава респондента - хороший способ обеспечить приверженность, поскольку он может обязать агентство отвечать на любой вопрос респондента в течение определенного периода времени.

В некоторых НСО также есть специальный отдел по вопросам политики в отношении респондентов, цель которого состоит в том, чтобы помочь повысить процент ответов и обеспечить, чтобы респонденты охотно предоставляли информацию. Персонал подразделения должен быть оборудован, чтобы ответить на вопросы об использовании информации, внимательности, с которой она обрабатывается, и общем отношении своего агентства. Они должны избегать появления раздражения и небрежности при цитировании закона и должны быть справедливыми и последовательными в своем отношении к бизнесу и домашнему хозяйству. Если в отношениях с респондентами ощущается кризис, главному статистику рекомендуется заняться этим вопросом на своем уровне. Сообщение непосредственно главному статистику на этом уровне может быть хорошим способом продемонстрировать сопротивляющимся респондентам серьезность, с которой агентство рассматривает этот вопрос.

Ссылки на руководства, передовой опыт и примеры:

- UK ONS - Хартия респондентов для бизнес-опросов .

8.2.7.3 Управление ключевыми респондентами и профилирование бизнеса

а) **Управление ключевыми респондентами**

Некоторые очень крупные предприятия представляют особый интерес для НСУ из-за их большого вклада в оценки экономики в целом или определенного статистического региона, области или сегмента классификации. Точные и быстрые ответы от таких предприятий имеют решающее значение для получения оценок хорошего качества. Назначение единого контактного лица в ОПС для конкретного предприятия для всех опросов может привести к значительному повышению точности и своевременности. По этой причине несколько ОПС создали специальное организационное подразделение, обычно называемое *подразделением* по *крупным делам* или подразделением *крупного бизнеса,* с исключительной функцией управления отношениями с крупным бизнесом.

Такие единицы доказали свою полезность, поскольку они могут собирать более подробную информацию от предприятий, в то же время снижая общую нагрузку на предприятия за счет более эффективного сбора данных. Кроме того, такие подразделения особенно подходят для проведения *профилирования бизнеса* , которое (как описано ниже) включает отслеживание границ и структуры крупных предприятий, а также проверку или изменение их механизмов отчетности по опросам. Поскольку в крупных компаниях довольно часто меняются границы и структура, без таких действий заявленные значения могут со временем стать несопоставимыми. Крупные кейсы также важны для отслеживания практики бухгалтерского учета, в частности, для различения внутренних трансфертных цен и рыночных цен, что особенно актуально при отслеживании транснациональных предприятий и экономических явлений, связанных с глобализацией.

б) **Профилирование бизнеса**

Согласно рекомендациям бизнес-регистра Евростата ручное профилирование - это ` ` метод анализа правовой, операционной и бухгалтерской структуры группы предприятий на национальном и мировом уровне с целью установления статистических единиц внутри этой группы, их связей и наиболее эффективных структуры для сбора статистических данных ». Процесс требует проведения углубленных интервью с высокопоставленными представителями компании для получения всей необходимой финансовой информации, взаимоотношений и структур компании. Профилирование используется для улучшения качества реестра предприятий, а, следовательно, и качества всех обследований, в которых он

используется в качестве основы для обследования и источника информации. Более подробную информацию о бизнес-профилировании можно найти в главе 11, раздел 11.2.4.1 « *Профилирование* ». Основным преимуществом профилирования является четкое разграничение видов деятельности внутри компании, разделенных на «виртуальные» статистические единицы, которые могут предоставить более подробный обзор деятельности компании. После профилирования компании чрезвычайно важно, чтобы данные собирались в соответствии с профилированной структурой или, по крайней мере, чтобы была определена процедура вменения недостающих значений.

Ссылки на руководства, передовой опыт и примеры:

- Руководство по обмену экономическими данными, ЕЭК ООН 2020 (предварительная ссылка)
- LCU в Ирландии .
- LCU в Италии .
- Профилирование в европейской деловой статистике .
- Профилирование в Нидерландах

8.2.7.4. Сведение к минимуму случаев неполучения ответов и последующих действий

а) **Сведение к минимуму отсутствия ответа**

Ключом к минимизации отсутствия ответа является сокращение количества неконтактных контактов и отказов. Причины неконтактности зависят от конкретного дизайна опроса. В очных опросах отказ от контакта может быть результатом неспособности интервьюера связаться с респондентом в пределах заранее установленного числа попыток контакта. Увеличение количества попыток контакта не только увеличивает количество «контактировавших» и, следовательно, скорость ответа, но также увеличивает затраты. Альтернативное изменение дней и времени попытки контакта также увеличивает скорость ответа без значительного влияния на стоимость. Это делается путем определения последующих процедур в случаях отсутствия контакта, обычно, если респондент не удается связаться в обычные рабочие часы, второй контакт предпринимается во второй половине дня или в выходные дни.

б) **Последующие процедуры**

В опросах, проводимых самостоятельно, отсутствие контактов может быть результатом ошибок в рамке опроса. Поскольку большинство статистических обследований используют статистические регистры в качестве основы для обследования, решение состоит в том, чтобы поддерживать статистические регистры в актуальном состоянии либо с помощью административных данных, либо путем проверки информации от самих предприятий, обычно посредством обследования регистров, которое предназначено для обновления. информация, хранящаяся в статистическом регистре (дополнительную информацию об улучшении фреймов и статистическом регистре можно найти в главе 11, раздел 11.3.7 « *Получение статистики непосредственно из СРП»*). Инструменты для уменьшения количества отказов также зависят от используемого режима сбора данных. Например, в опросах с интервью могут использоваться специально обученные интервьюеры для преобразования отказов, в то время как почтовые и интернет-опросы должны полагаться на стимулы или специальные контакты для противодействия явным отказам. Столкнувшись со сбором особенно сложных данных, таких как обследование бюджетов домашних хозяйств или обследование использования времени (оба из

которых требуют подробного аннотирования деятельности в форме дневника), статистические агентства часто прибегают к специальным стимулам, которые предлагаются респондентам. Эти поощрения могут быть денежными суммами, подарками или участием в лотерее со специальными призами. Возможность организации поощрений может зависеть от местных законов и правил.

8.2.7.5 Измерение нагрузки на респондентов, индивидуальной и общей

Многие страны прилагают усилия для измерения и снижения административной нагрузки на бизнес. Несмотря на то, что предоставление данных для официальной статистики составляет небольшую часть общей административной нагрузки, НСУ уделяют значительное внимание цели снижения нагрузки на респондентов. Время, затрачиваемое на ответы на анкету (включая время, необходимое для получения необходимых данных и заполнения анкеты), является показателем качества и эффективности (QPI), который следует контролировать для каждого опроса, как более подробно обсуждается в разделе 7.5.6.1. . Даже если время, затраченное на заполнение анкеты, фактически не измеряется анкетой, руководитель обследования обычно может дать надежную оценку на основе кабинетного исследования или тестирования анкеты.

Измерение нагрузки на респондентов в часах можно перевести в денежное выражение, умножив количество часов на среднюю почасовую заработную плату в секторе, в котором проводится вопросник.

Умножение средней нагрузки на респондентов на количество респондентов в опросе дает общую нагрузку на респондентов для опроса. Суммирование цифр по всем опросам в течение года дает общее годовое бремя ответов, возложенное на НСУ. Наличие списка обследований и их нагрузки на респондентов полезно для определения приоритетов развития НСУ, особенно тех, которые могут привести к модернизации деятельности по сбору данных, с помощью которой можно уменьшить нагрузку на респондентов. Примерами являются (1) модернизация бумажного обследования до сбора данных через Интернет и (2) использование административных данных для уменьшения общего количества вопросов в анкете обследования или полного исключения обследования.

Некоторые НСО также измеряют индивидуальную нагрузку на каждого бизнес-респондента и / или отмечают респондентов каждого обследования в своем статистическом реестре предприятий с целью исключения их (если возможно, в рамках плана выборки) из дальнейших выборок обследований на определенный период. Для обследований домашних хозяйств постепенная ротация выборки, а не ее замена, может быть эффективным методом справедливого распределения нагрузки на респондентов и в то же время иметь продольный компонент в том, что в противном случае было бы поперечным обследованием.

8.2.8. Разработка комплексных исследовательских программ

Хотя во всем мире можно отметить постоянные улучшения в проведении опросов, улучшения в основном были сосредоточены на улучшении отдельных опросов, а не на интеграции различных опросов в модульную и единую систему сбора данных. В результате скорость разработки может быть неравной, могут использоваться разные фреймы и процедуры их обновления, могут быть разные методологии выборки, сбора и обработки, что создает ненужную нагрузку для некоторых респондентов (обычно крупных предприятий). Кроме того, пользователи могут сталкиваться с разными стандартами распространения и часто могут использовать одни и те же инструменты для анализа разных микроданных.

8.2.8.1 Интегрированный набор опросов

Решением вышеупомянутых проблем является разработка интегрированного набора опросов. Интеграция в контексте программ обследований подразумевает связи между различными обследованиями или между раундами одного обследования. Интеграция преследует три основные цели: повышение ценности результатов опроса, снижение затрат и уменьшение нагрузки на респондентов.

Комплексные обследования имеют следующие общие характеристики:

a) согласованные концепции и содержание анкеты;

b) использование общей основы обследования, такой как статистический регистр, основа выборки или эталонная выборка, как более подробно описано в главе 11; и

c) принятие общих методологий выборки, сбора и обработки данных, как также дополнительно описано в главе 11.

Этот подход был опробован в различных статистических областях и оказался особенно полезным, поскольку он приводит к значительному снижению нагрузки на респондентов, в то же время позволяя получить более детальное понимание на более низком территориальном уровне. Примеры комплексного набора обследований можно найти в статистике бизнеса, социальной, сельскохозяйственной и экологической статистики, и они были успешно внедрены во многих странах. Комплексные программы обследований, особенно те, на которые приходится значительная часть общего бюджета НСУ, должны быть включены в стратегические планы и должны тщательно планироваться и оцениваться.

Ссылки на руководства, передовой опыт и примеры:

- ФАО - Программа комплексных сельскохозяйственных обследований - AGRISurvey .
- Филиппины PSA - Комплексное исследование BLES .
- Статистическое управление Канады - Комплексная программа деловой статистики .
- Статистическое управление Новой Зеландии - Комплексные обследования домашних хозяйств .
- На пути к интегрированной системе ежегодных бизнес-обзоров .

8.2.8.2 Основные исследовательские машины и дополнительные модули

Обычный подход к программе комплексного обследования состоит в том, чтобы ввести основной набор вопросов, адресованных всем респондентам, вместе с набором конкретных модулей и / или сменяющихся приложений, содержащих дополнительные вопросы. Для обследований предприятий это позволяет соответствующим образом адаптировать анкеты по отраслям с использованием отраслевых модулей, а также чередующихся дополнений по актуальным вопросам. Используя этот подход, основная информация (например, количество сотрудников или оборот для бизнес-статистики, статус занятости или источники дохода для социальной статистики) может быть собрана согласованным образом из нескольких источников, что обеспечивает репрезентативность на более низком территориальном уровне, чем в противном случае можно было бы достичь без увеличения нагрузки на респондентов.

Аналогичный подход также используется в обследованиях домашних хозяйств, которые часто используются для нескольких целей. Например, обследование рабочей силы часто включает основной модуль и дополнительные (специальные) модули, которые используются для более подробных запросов по подтемам с использованием той же основной выборки. Существуют также примеры опросов, предназначенных для сбора данных для различных целей, содержащих слабо связанные модули.

Использование дополнительных модулей также позволяет вменять недостающие переменные и строить оценки с использованием методов оценки малых площадей. Этот подход особенно полезен в сочетании с административной информацией (как в обследованиях предприятий, так и в обследованиях домашних хозяйств), которая используется для связи записей и для дальнейшего снижения нагрузки на респондентов. Это означает, что всегда следует собирать уникальные идентификационные номера, которые позволяют связывать записи, как более подробно обсуждается в разделе 14.2.11 « *Связанные данные »* .

8.2.8.3 Ответ на срочные запросы

Одним из сравнительных преимуществ НСУ по сравнению с другими поставщиками данных является доверие пользователей к официальной статистике. Это доверие, помимо прочего, основывается на точности выходных данных, которая, в свою очередь, зависит от хорошей разработки и тщательного выполнения статистических процессов. Это, в свою очередь, требует времени, больше времени, чем пользователи могут себе представить. Таким образом, НСУ сталкивается с проблемой, когда сталкивается с новым и острым спросом на данные. Неспособность выполнить поставку быстро имеет потенциально негативные последствия для ОПС. С другой стороны, новый и срочный спрос на данные может дать возможность собрать столь необходимые ресурсы, подключиться к новому источнику данных, повысить видимость и / или авторитет НСО и эффективно конкурировать с альтернативными поставщиками данных.

Громкие и срочные запросы чаще всего поступают от правительства, но могут также поступать и от международных организаций. Национальная статистическая служба должна быстро реагировать и использовать ограниченное окно возможностей для согласования выделения дополнительных ресурсов. Главный статистик должен использовать свои личные полномочия, чтобы объяснить всем сторонам особенности сбора статистических данных.

Национальная статистическая служба не должна автоматически инициировать новый опрос в ответ на каждый новый запрос данных. Скорее, он должен систематически исследовать, можно ли полностью или хотя бы частично удовлетворить спрос с использованием регулярно собираемых данных или административных данных. Ощущение срочности со стороны клиента (если клиентом является правительство) может открыть новые возможности для доступа к новым источникам данных. Следует отметить, что не все запросы должны приниматься официальной статистикой, особенно те, которые могут быть легко получены от исследовательских компаний частного сектора.

Хотя запрос может быть только разовым, НСУ следует помнить, что периодичность является важной характеристикой статистической деятельности, и соответственно планировать и сообщать (тем, кто запрашивает данные). Срочные запросы могут быть распространены с пометкой «экспериментальная статистика», если нет уверенности в том, что усилия будут повторяться, или если результаты не удовлетворяют обычным статистическим стандартам.

8.2.8.4 Возможность гибкой съемки

Чтобы иметь возможность реагировать на срочные потребности в новых данных, некоторые НСУ ввели организационное подразделение, которое может разрабатывать и проводить быстрое обследование либо в качестве первого цикла постоянного нового обследования, либо в качестве разового упражнения. Ответственность за технико-экономическое обоснование может быть возложена на такое подразделение, чтобы его сотрудники привыкли предпринимать быстрые усилия, направленные на решение основных вопросов, до того, как это может быть более обстоятельным обследованием. Развивая такие

возможности и периодически демонстрируя свои возможности и возможности, ОПС может повысить свою значимость.

Комплексные программы опросов также увеличивают способность ОПС быстро реагировать на новые запросы, поскольку модуль или дополнительные вопросы могут быть добавлены к существующему опросу быстрее, чем создание полностью независимого сбора данных.

8.2.9 Обучение и опыт инспекторов

Для успешного планирования, разработки, проведения и оценки обследования требуется сотрудничество и координация широкого круга специалистов с различными техническими навыками. Из-за их специфики эти навыки часто нелегко получить за пределами статистической системы, и поэтому их часто приходится развивать и поддерживать внутри (более подробную информацию о вариантах организации обучения персонала можно найти в главе 12). Поскольку надлежащее обучение персонала обследования является ключом к обеспечению качества статистики, важно инвестировать время и ресурсы в обучение. Обучение также позволяет стандартизировать подходы и должно использоваться для содействия оптимизации процессов и использованию новых методов. Организация эффективного обучения особенно важна для интервьюеров, так как хорошо образованные интервьюеры могут значительно улучшить итоговую статистику. В подразделах ниже перечислены наиболее важные знания и навыки исследовательской группы, которые необходимы для эффективного проведения опроса. Группы, проводящие обследование, обычно являются междисциплинарными и обычно состоят из руководителя обследования, эксперта в области исследования, охватываемой обследованием (специалист по предмету), методиста, аналитика компьютерных систем и эксперта по сбору данных и операциям. Все члены исследовательской группы: планируют, управляют и координируют действия в рамках своей компетенции и ответственности, которые рассматриваются в этой главе.

8.2.9.1. Менеджеры опросов

Менеджер опроса отвечает за управление опросом. Обычно он или она является старшим экспертом с обширным опытом участия в нескольких этапах нескольких опросов. Менеджер обследования обеспечивает соблюдение целей, бюджета и графика. Руководитель обследования обычно отвечает за определение необходимых ресурсов для обследования, разработку предварительного плана и координацию подготовки, подготовку бюджета и мониторинг использования ресурсов и прогресса. Менеджер по опросу председательствует на собраниях команды и должен понимать и представлять интересы пользователей. Руководитель обследования поддерживает связь с высшим руководством и клиентом и отчитывается о ходе их выполнения. Он или она обеспечивает соблюдение юридических и нормативных обязательств, а также ведомственных политик, стандартов, руководств и положений.

Менеджер опроса должен быть прекрасным организатором, обладать значительными личными полномочиями и иметь глубокое понимание основных процессов опроса. Руководители опросов обычно приходят из соответствующих тематических подразделений, но иногда методолог или ИТ-специалист также могут выполнять задачи руководителя опроса. Межличностные и организационные навыки должны быть основными критериями выбора для задачи руководителя обследования, но нельзя упускать из виду знание предмета и понимание процессов, поскольку пробелы в этих областях могут подорвать авторитет и вызвать серьезные проблемы при проведении обследования.

8.2.9.2 Специалисты в предметной области

Специалист в предметной области несет ответственность за содержание опроса. Он или она проводит или координирует подготовку определений и концепций, разработку и тестирование вопросников, подготовку спецификаций по сбору и обработке данных, разработку статистических результатов, разработку и внедрение анализа данных и подготовку аналитических текстов. Он или она также координирует проверку результатов обследования и предоставляет экспертные знания в предметной области для оценки качества данных и подготовки соответствующей документации.

Специалист по предмету должен иметь глубокое понимание предмета и лежащих в основе процессов, которые необходимы для надежного ответа (например, бухгалтерского учета). Он или она также должны полностью понимать процессы и результаты обследования, поскольку он или она несет ответственность за подготовку контента и его преобразование в статистические результаты. Специалист в предметной области должен обладать передовыми аналитическими знаниями, которые позволят ему или ей подготовить и интерпретировать результаты. Кроме того, специалист в предметной области должен понимать базовую психологию, которая требуется для разработки вопросника, обладать знаниями в области обработки данных и понимать основные концепции программирования. Понимание связанных методологических и ИТ вопросов необходимо, так как это упрощает сотрудничество с методологическими и ИТ-экспертами и отделами. Это часто включает в себя знание эконометрики (по крайней мере, базовое) и понимание других методов манипулирования данными. Наконец, навыки языка и визуализации данных также важны, поскольку они могут повысить качество и понимание представленных данных.

8.2.9.3 Методологи

Методолог обследования отвечает за проведение и координацию разработки и разработки статистической методологии, которая будет использоваться для обследования. Он или она отвечает за дизайн выборки, взвешивание и оценку, дизайн контроля качества, дизайн и меры оценки качества данных, разработку механизмов или стратегий редактирования и вменения, а также статистические аспекты распространения и анализа данных. Методолог обследования также выступает в качестве консультанта и советника для всех других членов исследовательской группы по вопросам статистической методологии и обеспечивает соблюдение использования надежных и эффективных статистических методов.

Методолог обследования должен обладать передовыми статистическими знаниями, а также передовыми знаниями в области обработки данных. Он или она должны понимать эконометрику (для вменения и моделирования) и обладать хотя бы промежуточными навыками программирования, поскольку он или она часто вынуждены писать программный код и запросы на изменение кода для всех аспектов обработки обследования.

8.2.9.4 Специалисты по сбору данных и последующему наблюдению

Специалист по сбору данных отвечает за разработку спецификаций и процедур сбора данных. Он или она также несет ответственность за координацию набора, обучения, мониторинга и контроля интервьюеров и супервайзеров. В его или ее обязанности входит разработка, внедрение и управление операциями по сбору платежей, а также подготовка материально-технической поддержки. Он или она также выступает в качестве советника для всех других членов исследовательской группы по оперативным вопросам и обеспечивает надлежащее включение спецификаций и требований, разработанных другими членами группы, в процедуры. Роль специалиста по сбору данных может включать координацию сбора данных на местах через региональные отделения, а также выполнение ручных и автоматизированных оперативных действий, выполняемых в головном офисе.

Специалист по обработке данных отвечает за разработку спецификаций и процедур захвата и кодирования. Он или она также несет ответственность за координацию набора, обучения, мониторинга и контроля ввода, редактирования и обработки данных персонала. В его или ее обязанности входит разработка, внедрение и управление действиями по сбору и кодированию, а также координация логистической поддержки, связанной с обработкой данных.

Для небольших проектов роль специалиста по сбору и обработке данных может быть совмещена.

Важнейшие навыки для обеих ролей - это способность эффективно координировать работу большой команды, иметь организационный авторитет и способность обеспечивать соблюдение сроков. Специалисты по сбору и обработке данных должны иметь глубокое понимание процесса сбора данных, а также базовое понимание всех предыдущих и последующих этапов, чтобы при необходимости они могли предложить улучшения.

8.2.9.5 Интервьюеры

В методах и методах опросов Статистического управления Канады перечислены следующие ключи к эффективному проведению собеседований:

а) **Уверенность** : интервьюер должен быть уверен в своих силах. Это возможно только при хорошем понимании обследования и роли интервьюера.

б) **Навыки** аудирования: интервьюеру следует дождаться, пока респондент закончит говорить, прежде чем он или она перестанут слушать. Интервьюер может указать, что он или она слушает время от времени: «Да, я понимаю». Тем не менее, интервьюер не должен делать предположений о том, что респондент собирается сказать, или пытаться закончить предложение. Лучше задавать вопросы, если интервьюер чувствует, что респондент или интервьюер упустили момент.

с) **Сочувствие** : интервьюер должен быть внимательным к ситуации респондента во время звонка или посещения. Если респондент описывает личный инцидент, интервьюер должен проявить интерес (но никогда не выносить суждения), а затем попытаться переориентировать респондента на интервью.

г) **Речь** : вокальное выражение важно, особенно при телефонном интервью. Интервьюер должен говорить очень четко и стараться говорить с умеренной скоростью. Если интервьюер говорит слишком быстро, респонденты могут пропустить части вопроса. Если говорить слишком медленно, респонденты начинают отвечать до того, как интервьюер закончит вопрос. Опускание головы снижает высоту голоса. Более низкий голос более четкий и лучше слышен, особенно по телефону. Во время тренировки следует продемонстрировать правильную скорость и подачу.

д) **Знать анкету** . Интервьюер должен знать анкету, концепции и терминологию, использованную при обследовании. Во время собеседования не будет времени искать определения или ответы на вопросы в руководстве. Ничто не нарушает взаимопонимание быстрее, чем длинные паузы, особенно в телефонных интервью.

8.2.9.6 Специалисты по вводу и редактированию данных

Несмотря на то, что модернизация статистических процессов снизила потребность в ручном редактировании, этот вид деятельности по-прежнему необходим, поскольку он значительно повышает качество официальной статистики. Ввод данных и ручное редактирование часто выполняются сотрудниками без высшего образования, но значительные преимущества могут быть получены за счет образования и специализации. Общий навык, необходимый как при вводе, так и при редактировании данных, - это высокий уровень концентрации, поскольку задачи повторяются, и работники должны тратить много времени на выполнение

одной и той же задачи. Поэтому такая работа требует очень высокого уровня концентрации и терпения. Отсутствие этого атрибута может привести к низкому качеству результатов. Ожидается, что сотрудники по вводу данных, а также по редактированию будут иметь исключительную скорость набора, поскольку им придется вводить и проверять огромные объемы данных за очень короткое время. Они должны хорошо владеть всеми формами устройств ввода данных и пользоваться мышью, клавиатурой, сканерами и т. Д. Также важно, чтобы они разбирались в использовании базового программного обеспечения, такого как обработка текста и электронные таблицы, но они также должны привыкнуть к использованию специализированного программного обеспечения для ввода и редактирования данных, и поэтому необходимы базовые знания об использовании компьютера.

Административные источники

В «Основополагающих принципах официальной статистики» Организации Объединенных Наций (ЮНФПОС) говорится, что «данные для статистических целей могут быть получены из всех типов источников, будь то статистические обследования или административные записи».

Термин «административные данные» здесь относится к данным, собранным государственным министерством, департаментом или агентством в основном для административных (не исследовательских или статистических) целей. Эти административные цели связаны с соответствующими исполнительными или законными функциями, такими как авторизация, регистрация, разрешения, платежи, санкции, контроль и т. Д. Административные данные могут включать как данные в административных регистрах, так и данные из других административных источников.

Использование административных данных для статистических целей - явление не недавнее - есть примеры составления статистики на основе данных о числе рождений и смертей, датируемых началом 17 века, - но в последние два десятилетия оно стало все более распространенным по мере развития технологий и роста. вычислительные мощности позволили статистическим агентствам преодолеть многие ограничения, ранее связанные с обработкой больших наборов административных данных. Это вместе с достижениями в методах увязки данных предоставило НСУ возможность более эффективно использовать административные данные при производстве официальной статистики, как для замены существующих методов сбора данных , так и для дополнения данных статистических обследований и для создания новых статистических продуктов.

В 2011 году ЕЭК ООН опубликовала справочник « *Использование административных и вторичных источников для официальной статистики - Справочник принципов и практики»* . В нем дается обзор источников административных данных и вопросов сбора данных, и он используется в качестве основы для этого раздела. Быстрое развитие в течение последнего десятилетия расширило общие знания статистиков об административных источниках. Дополнительную информацию о последнем опыте можно найти на веб-сайтах национальных статистических организаций и международных статистических организаций, а также на сайтах совместных проектов и различных статистических конференций.

Использование административных данных и производство статистики на основе регистров рассматриваются в разделе 11.2.

8.3.1 Типы административных данных

Государственные органы во всех странах собирают большой объем данных в рамках своей текущей деятельности. Административные источники очень редко, выборочные обследования по своей природе - они обычно полны с учетом области, полномочий органа и цели сбора данных. Эти данные охватывают широкий спектр видов деятельности, таких как сбор налогов, социальное обеспечение и здравоохранение, политика в области занятости и безработицы, а также системы регистрации гражданских событий (рождений, смертей, браков и т. Д.), Предприятий, собственности и транспортных средств среди прочего. Эти административные данные становятся все более доступными для НСУ, и, как следствие, это оказывает значительное влияние на то, как собираются данные и компилируется официальная статистика.

Административные данные поступают из множества различных источников в зависимости от структуры государственного сектора в стране. Наиболее распространенный способ использования административных данных в производстве статистики - это объединение данных из разных административных источников с данными статистических обследований или без них. В этом процессе интеграции данных некоторые наборы административных данных играют основную роль, и эти данные широко используются многими НСУ. Ниже источники административных данных сгруппированы, чтобы проиллюстрировать их важность и полезность для НСУ, которая определяется исчерпывающим охватом и разнообразным содержанием данных.

а) Большие и сложные системы административных данных содержат как регистрационные данные, так и данные транзакций. Типичными примерами являются данные из систем регистрации населения и предприятий, систем социального обеспечения и здравоохранения, систем налогообложения, таможенных систем и систем регистрации зданий и собственности. Обычно эти данные являются полными, исчерпывающими по своей природе, охватывающими всех граждан, все предприятия или весь фонд недвижимости и зданий. Основная информация в административных регистрах этой группы может содержать идентификационный код, имя, адрес, дату регистрации и другую идентификационную и классификационную информацию. Органы, отвечающие за эти регистры, создали системы для постоянного (часто в режиме онлайн) обновления основного содержания регистров. Органы, ведущие эти большие административные регистры, могут также иметь другие системы данных для выполнения своих основных административных функций. Например, налоговые органы могут иметь базовые реестры лиц и предприятий, подлежащих налогообложению, и отдельные системы для подоходного налога с физических лиц и предприятий, налога на добавленную стоимость и налога на имущество. Обычно данные от производителей этих больших систем данных используются в качестве исходного материала во многих различных статистических системах, от статистики населения и социальной статистики до статистики бизнеса, экономики и окружающей среды.

б) Административные данные с более конкретным охватом обычно включают регистры и данные транспортных и дорожных ведомств, органов юстиции и избирательных органов, а также систем образования и школ. Часто эти данные являются важным или единственным источником, например, при производстве статистики транспорта, статистики правосудия и преступности, статистики всеобщих выборов и статистики образования. Административные данные из этой группы также используются в качестве дополнительного источника для деловой и экономической статистики, а также статистики населения и социальной статистики. Кроме того, к этой группе относится большое количество конкретных административных данных, которые полезны при составлении статистики окружающей среды, энергетики и отходов, а также административных данных для деятельности и финансов государственного

сектора. Некоторые административные данные имеют объем и содержание, которые может быть трудно получить с помощью обследований.

8.3.2 Преимущества сбора административных данных

Порядок сбора данных и потенциальные преимущества варьируются от страны к стране в зависимости от национальных условий и от того, в какой степени НСУ планирует использовать административные регистры и данные. Ниже представлены наиболее часто обсуждаемые преимущества использования административных источников при производстве статистики.

а) **Экономическая эффективность** : важным преимуществом использования административных данных для сбора статистики является то, что стоимость сбора данных относительно невелика по сравнению с затратами, которые понесены при проведении переписей, создании и ведении статистических базовых регистров и проведении прямых обследований как основных расходы на сбор данных уже покрыты самим административным процессом. Однако использование административных данных не является бесплатным для НСУ. Может возникнуть необходимость инвестировать в ИКТ и производственные системы, механизмы координации и в развитие статистических методов и новых компетенций. Кроме того, НСО может быть обязано оплачивать расходы на передачу и передачу данных администрациям, даже если сами данные являются бесплатными.

б) **Снижение нагрузки на** респондентов: респонденты могут отрицательно отреагировать на опрос, если они считают, что уже предоставили аналогичную информацию государственным органам. Растущее нежелание как фирм, так и частных лиц участвовать в статистических обследованиях, не в последнюю очередь в обследованиях малого и среднего бизнеса, может стать угрозой для качества статистики. Использование административных данных приводит к снижению нагрузки на респондентов и в то же время может решить проблему увеличения количества случаев неполучения ответов и повышения качества входящих исходных данных и статистики.

с) **Своевременность и частота** : когда правительства разрабатывают свои системы ИКТ и переходят к сбору данных в режиме онлайн и мобильным услугам, административные регистры постоянно обновляются, а административные данные доступны относительно быстро. Следовательно, статистику, основанную на административных данных, можно подготовить быстрее и публиковать раньше, чем на основе данных, собранных с помощью статистических переписей и обследований. Из-за постоянного использования административных регистров в режиме онлайн или других систем обновления, использование административных данных может также увеличить частоту, с которой статистические данные традиционно собираются и публикуются.

г) **Охват и полнота** : административные источники часто дают полный охват целевой группы, тогда как выборочные обследования часто непосредственно охватывают лишь относительно небольшую ее часть. Использование источников административных данных может уменьшить или устранить ошибки из-за отсутствия ответов и других типичных ошибок выборочных обследований. Административные регистры и данные представляют собой хорошие источники для создания и ведения базовых статистических регистров, лучшего охвата целевых групп населения для выборочных обследований и могут сделать статистику более точной. Использование административных данных вместо обследований также может привести к улучшению данных по регионам и небольшим территориям и более подробной информации.

e) **Актуальность** : НСУ может улучшить свою способность реагировать на новые потребности в данных и повысить актуальность статистики путем поиска и использования новых источников административных данных. Административные регистры и данные в сочетании с данными обследований могут повысить гибкость НСУ для быстрого реагирования на новые статистические требования. Административные данные могут играть важную роль, например, в заполнении пробелов в данных, необходимых для измерения прогресса в достижении Целей в области устойчивого развития (ЦУР) Повестки дня в области устойчивого развития на период до 2030 года, которые невозможно удовлетворить только традиционными методами.

8.3.3 Проблемы и проблемы при сборе административных данных

Хотя есть много веских причин для использования административных данных, существует также ряд проблем и проблем, связанных со сбором данных. Они касаются организации сбора данных, готовности поставщиков данных и НСУ и качества самих данных.

Национальное статистическое законодательство обычно дает НСУ прочную основу для сбора и обработки статистических обследований и для публикации статистических данных. Что касается сбора административных данных, законодательство часто бывает неадекватным и может вызвать множество проблем или даже помешать получению.

Проблемы связаны с административной культурой и традициями. Административные органы могут неохотно предоставлять доступ к «своим данным» для использования в статистических целях. Это может произойти даже в стране с соответствующим статистическим законодательством. Административные органы могут иметь сильную собственность в отношении административных регистров и данных, за которые они несут ответственность. В некоторых случаях могут быть введены правовые ограничения или положения о конфиденциальности, ограничивающие доступ к административным данным. В некоторых странах желание и предложения НСУ о незначительных изменениях в механизмах сбора административных данных, которые улучшили бы их полезность, могут оказаться невозможными на практике.

Чтобы справиться с общими проблемами и быть лучше подготовленным к сбору данных из административных источников, в разделе 8.3.4 резюмируются основные требования, основанные на текущих знаниях и практике стран.

Соответствующие механизмы сбора данных облегчают оценку и решение проблем, связанных с качеством административных данных и использованием этих данных в производстве статистики. В каждой стране качество административных данных зависит от источника. Перед принятием решения о сборе и использовании любого набора административных данных необходимо провести тщательную оценку качества и спланировать корректирующие меры. Часто такая оценка качества требует, чтобы ОПС получила доступ к административным данным на уровне записей. Административные данные низкого качества не должны использоваться при производстве статистики.

Ниже перечислены часто упоминаемые проблемы качества, сомнения и предполагаемое отношение к сбору и использованию административных данных в производстве статистики вместе с некоторыми решениями, используемыми в текущей практике.

a) **Различия в единицах, концепциях и определениях переменных** : единицы, концепции и определения переменных, используемые в административных данных, часто отличаются от статистических. Это особенно большая проблема в области деловой и экономической статистики. Существует ряд конкретных исследований того, как эти проблемы могут быть решены НСУ, а также международными организациями. Иногда могут использоваться хорошие прокси, иногда даже

требуются дополнительные данные опроса. Различия и возможные методы исправления должны публиковаться вместе с опубликованной статистикой.

б) **Различия в классификациях** : часто классификации, используемые в административных источниках, отличаются от статистических стандартов. Это может вызвать замешательство у пользователей статистики. В соответствии с законом о статистике НСУ решает и утверждает классификации, которые будут использоваться в официальной статистике. В некоторых странах национальной классификации видов экономической деятельности, основанной на международном статистическом стандарте и подтвержденной НСУ, присвоен статус национального официального стандарта, который также будет использоваться всеми производителями административных данных. Это не может полностью предотвратить неправильное кодирование, поскольку коды, используемые для единиц, могут быть ошибочными. Создание и использование таблиц ссылок - это часто используемый метод для корректировки различий в классификациях, даже если он увеличивает рабочую нагрузку NSO.

с) **Неадекватное качество данных** : данные, включенные в административные регистры и файлы данных, могут быть неполными или неточными, или данные могут отсутствовать . В некоторых случаях персонал органа, ответственного за административные данные, может быть не обучен должным образом, или методология записи соответствующей информации может быть некорректной. Корректирующие меры могут включать хорошие системы контроля качества и методы исправления входящих данных, обучение персонала и обсуждение вопросов качества с административным органом.

г) **Отсутствие** статистических знаний: органы, предоставляющие административные данные, обычно не обладают такими же уровнями статистических знаний и возможностей, как НСУ. В результате могут возникнуть противоречивые данные. Например, некоторые данные могут быть неточно или тщательно записаны и могут быть неполными или неточными для элементов информации, представляющих второстепенный интерес или ценность для основных административных целей. В этих случаях может помочь тесное сотрудничество между НСУ и производителями данных, а также учебные мероприятия НСУ.

е) **Необходимость дополнительной проверки данных** : в зависимости от источника административные данные требуют всесторонних проверок достоверности, кроме того, может потребоваться процесс согласования между различными источниками данных. Это может потребовать значительных усилий и затрат для НСУ.

8.3.4 Требования к доступу и получению административных данных

Проблемы, связанные со сбором данных, связаны с общей административной и правовой структурой страны. Доступ к административным данным, предварительные условия для сбора данных и возможности использования этих данных варьируются от страны к стране. Для облегчения доступа НСУ к административным данным необходимы различные структуры. Эти рамки включают правовые, политические, организационные и технические аспекты. Все эти параметры следует учитывать, когда НСУ планирует сбор данных из административных источников. Краткое описание основных требований для успешного сбора данных кратко излагается ниже.

8.3.4.1 Правовая база

Использование административных данных при составлении официальной статистики требует прочной правовой основы. Полномочия, обеспечивающие доступ НСУ к административным данным, рассматриваются как один из ключевых вопросов при

модернизации статистического законодательства. В зависимости от структуры национального законодательства важно отметить, что полномочия НСУ по сбору данных из административных источников могут также потребовать внесения изменений в другое законодательство. В разделе 3.4 дается дальнейшее руководство по надежному статистическому законодательству. Основные вопросы освещаются в следующих параграфах.

Законодательство не только должно предоставлять ОПС право на бесплатное получение административных данных, но также должно гарантировать, что ОПС обеспечивает соответствующий уровень защиты данных, например, защиту и конфиденциальность данных. Это жизненно важно для получения и поддержания общественного одобрения не только способности НСУ должным образом управлять административными данными, но и для того, чтобы общественность могла доверять распространяемой официальной статистике.

Просто иметь возможность собирать данные из административного источника или иметь к ним доступ - это довольно расплывчатое / приблизительное положение и требует более точной интерпретации. В идеале НСУ должно иметь доступ к административным данным на уровне записей и соответствующим метаданным, оно должно иметь право объединять административные данные с данными обследований и другими данными, и оно должно быть своевременно информировано об источниках административных данных и изменениях в них.

Доступ к административным данным и их использование при производстве статистики подняли важные вопросы относительно того, увеличивает ли это вероятность нарушения конфиденциальности или коммерческой тайны или нарушения правил конфиденциальности статистических данных. Поэтому, наряду со статистическим законодательством, важно синхронизировать другие законодательные акты с законом о статистике. Законодательство должно обеспечивать надежную защиту данных, защиту частной жизни и статистическую конфиденциальность, и эти правила должны быть доведены до сведения административных органов и общества в целом. Также следует гарантировать, что в момент поступления административных данных в НСУ они обрабатываются строго в соответствии с правилами статистической конфиденциальности. Потоки данных идут только в одном направлении, то есть от административного органа к NSO.

Данные, собранные для статистических целей, являются конфиденциальными независимо от источника данных. Данные, собранные из административных источников, являются конфиденциальными в распоряжении НСУ, даже если эти данные являются общедоступными и принадлежат поставщику административных данных. В противном случае доверие к НСУ может быть подорвано.

8.3.4.2 Общественное одобрение использования административных данных правительством

Даже если НСО является автономной организацией, она в то же время является частью государственного сектора. НСУ не может действовать в одиночку в своих усилиях по использованию административных данных в производстве статистики, вместо этого ему нужна поддержка со стороны других органов власти, политических хозяев, широкой общественности и всего общества.

Общественное мнение относительно обмена данными между различными правительственными ведомствами варьируется от страны к стране. Это может быть в пользу из-за растущей эффективности администрирования и может быть даже враждебным из-за страха ухудшения конфиденциальности и «синдрома большого брата». Важно пояснить, что использование административных данных в производстве статистики не означает их совместного использования в государственном секторе. Ответить на вопросы

широкой общественности о конфиденциальности и контролировать подозрения важно. НСУ должны активно разъяснять защитные меры, принятые в соответствии с законодательством, и открыто информировать общественность о своих принципах и практике работы. Открытое обсуждение и обсуждение с объяснением причин и преимуществ использования административных данных в производстве статистики должны быть ключевым принципом НСУ.

Одним из наиболее важных пользователей официальной статистики является государственный сектор с его растущими потребностями в новой, актуальной и высококачественной статистике. В то же время во многих странах усиливается потребность в сокращении бюджетов. В этой ситуации правительства могут быть готовы поддержать усилия НСУ по развитию надежной инфраструктуры, снизить затраты на статистику и улучшить условия для удовлетворения новых потребностей в данных. Это, в свою очередь, дает НСУ возможность продемонстрировать преимущества административных данных в производстве статистики, а также внести предложения и предпринять инициативы.

8.3.4.3 Технические основы

Технические структуры здесь относятся к механизмам, с помощью которых передаются данные, а также к любым соответствующим стандартам данных или метаданных. Механизмы передачи данных варьируются от страны к стране и внутри страны, от одного административного источника к другому, в зависимости от зрелости и сложности ИКТ-систем административных органов, которые являются поставщиками данных, и НСУ. Поставщики данных могут отправлять файлы данных в NSO, или NSO может извлекать данные непосредственно из административной исходной базы данных. Межмашинный доступ к большим наборам данных и онлайн-доступ становятся все более распространенными. Более широкие инициативы в области открытых данных на правительственном уровне могут позволить и расширить прямой доступ в будущем. Важно, чтобы используемые механизмы учитывали технические возможности как НСУ, так и поставщика административных данных.

С технической точки зрения, НСУ становится все проще использовать данные из административных источников, когда правительственные организации разрабатывают ИТ-системы и переводят процессы в цифровую форму. Поставщики административных данных также могут использовать международные стандарты для передачи данных и метаданных. Кроме того, для передачи данных в НСУ могут быть полезны общие национальные стандарты, применяемые во всем государственном секторе.

8.3.4.4 Сотрудничество с административными органами, которые являются поставщиками данных

Использование административных данных в производстве статистики создает прочную связь между НСУ и поставщиком административных данных. Не только ссылка, но и использование административных данных приводит к зависимости НСУ от административного источника. Поэтому важно следить за развитием административных структур и изменениями в законодательстве, касающемся административных источников, и быстро реагировать на любые изменения.

Необходимо хорошее сотрудничество между НСУ, регистрационными органами и другими поставщиками административных данных для обеспечения того, чтобы административные регистры и данные подходили в качестве исходных данных для статистических сборов. Это сотрудничество, в свою очередь, может влиять на содержание и графики административных источников, обеспечивать и повышать качество административных данных и гарантировать плавный и своевременный процесс передачи административных данных в НСУ.

НСО необходимо установить хорошие рабочие отношения и сотрудничество с поставщиками административных данных. Многие НСУ сформировали рабочие группы для сотрудничества с поставщиками административных данных и проводят с ними регулярные встречи. Может быть полезно подготовить меморандум о взаимопонимании между статистическим управлением и поставщиком административных данных в отношении потоков данных, метаданных и коммуникаций. В рамках любого такого меморандума НСУ будет требовать заблаговременного предупреждения о любых изменениях, внесенных в административный процесс, которые повлияют на полученные данные.

Такая политика и соглашения могут привести к все более позитивным результатам. НСУ могут строить отношения с поставщиками административных данных, предлагая свои знания в области сбора, редактирования и хранения данных, способствуя принятию статистических стандартов и предоставляя рекомендации по вопросам качества. Эти меры, в свою очередь, могут улучшить качество исходных данных для статистики.

Граница между производителем официальной статистики и производителем административных данных должна оставаться четкой, даже несмотря на расширение сотрудничества и взаимодействия между НСУ и поставщиками данных. Это замечание относится к профессиональной независимости в принципах ЮНФПОС, особенно к критерию «Свобода от конфликта интересов», представленному в разделе 4.2.2.

8.3.4.5 Подготовка и возможности НСО

Сбор данных из административных источников необходимо тщательно спланировать, чтобы заранее выявить любые потенциальные препятствия и проблемы и избежать их. Когда административные данные используются впервые, время, необходимое для внесения изменений в существующее законодательство, обычно очень велико. На этапе планирования необходимо тщательно проанализировать цели и приоритеты ОПС, а также внешние и внутренние условия и возможности, а также запланировать меры. Возможно, будет разумным включить эти планы в систему многолетнего и годового планирования НСО.

Средства ИКТ, производственные процессы и статистические системы могут нуждаться в реорганизации из-за сбора административных данных. Также могут быть требования к новым методическим навыкам и обучению персонала. Они обсуждаются более подробно в разделе 11.2.

8.3.5 Обработка административных данных

Общая статистическая модель бизнес-процессов (GSBPM) предназначена для применения независимо от источника данных. Его можно использовать для описания и оценки качества процессов на основе административных данных. GSBPM обсуждается в главе 5, раздел 5.5.5 *«Общая статистическая модель бизнес-процесса».* Перед планированием использования административных данных в статистических целях рекомендуется более внимательно изучить эту модель, которая охватывает все этапы процесса статистического производства.

На общем уровне НСУ должно иметь четкое представление о том, какие конкретные административные данные необходимы и для каких статистических целей. Важно, чтобы для каждого набора административных данных был тщательно проанализирован основной административный процесс и соответствующее законодательство. Не менее важно глубокое понимание содержания административных данных, включая определения единиц, концепций и переменных, а также систем обновления. Более того, перед сбором и обработкой административных данных необходимо определить статистические системы, в которых будет использоваться конкретный набор административных данных, и способы

использования этих данных. Этот тип информации должен собираться и анализироваться на этапах процесса определения потребностей, проектирования и строительства, как описано в GSBPM.

Этап сбора модели включает, среди прочего, важный этап обеспечения того, чтобы административные данные предоставлялись НСУ в соответствии с соглашениями. Независимо от точности и детализации соглашений, важно убедиться, что файлы данных имеют правильный формат и содержат ожидаемые поля.

НСУ с установленным использованием нескольких административных источников часто имеют специальное функциональное подразделение для административных данных в отделе сбора данных, через которое должны проходить все данные, поступающие из административных источников, прежде чем они будут обработаны. Это подразделение отвечает за проверку того, что файлы входящих данных приемлемы, а также может проводить первую проверку и проверки качества административных данных. Подразделение также может нести ответственность за управление системой контрактов на сбор данных между поставщиками данных и НСУ. Цели и развитие такого рода подразделения описаны в документе « *Система сбора административных данных и ее соответствие рекомендациям по качеству,* изложенным в *своде правил и экспертной оценке* », опубликованном в 2017 году.

Существует множество вариантов использования административных данных при производстве статистики. Эти данные могут использоваться, например, для замены или дополнения данных обследований, для построения статистических регистров, для создания и обновления основ выборки, для создания интегрированной статистики, такой как национальные счета, и как часть статистики на основе регистров. Некоторые административные источники, например, административный регистр населения, могут использоваться одновременно для многих статистических целей.

Обработка административных данных не является отдельной частью или отдельным подпроцессом GSBPM, но встроена во все фазы процесса. Несмотря на то, что GSBPM определяет возможные этапы статистического процесса, он не требует какого-либо строгого порядка, в котором эти этапы или подпроцессы должны выполняться. Однако при обработке административных данных в качестве единого источника данных для статистики важно, чтобы все необходимые шаги и подпроцессы были рассмотрены и приняты во внимание при планировании процесса. Наглядный пример обработки административных данных в рамках GSBPM описан в статье *: Методологии комплексного использования административных данных в статистическом процессе* .

Обработка административных данных в контексте статистики на основе регистров более подробно обсуждается в разделе 11.2.

Ссылки на руководства, передовой опыт и примеры:

- Использование административных и вторичных источников для официальной статистики : Справочник принципов и практики, ЕЭК ООН, (2011 г.)
- Система сбора административных данных и то, как она соответствует руководящим принципам кодекса практики и экспертной оценки , Статистический журнал IAOS 33 (2017), Статистическое управление Финляндии
- Евростат - Методологии комплексного использования административных данных в статистическом процессе

- Геопространственные данные

Термин «геопространственные данные» относится к данным, имеющим географический компонент. Это означает, что записи в наборе геопространственных данных содержат неявную информацию о местоположении, такую как адрес, город или почтовый индекс. Географическая информационная система (ГИС) - это система, предназначенная для сбора, хранения, обработки, анализа, управления и представления геопространственных данных.

Данные ГИС поступают из спутниковых изображений, которые можно использовать для создания изображений с большим объемом данных с извлеченными векторными объектами и атрибутивными данными. Их можно использовать в картографических приложениях для получения многослойного результата для многих типов анализа. Спутниковые изображения обладают значительным потенциалом для получения более своевременных статистических данных, уменьшения частоты опросов, уменьшения нагрузки на респондентов и других затрат, а также для предоставления данных на более дезагрегированном уровне для принятия обоснованных решений.

Ценность привязки статистической информации к местоположению давно признана, и НСУ уже много лет делают это. Это началось с развития ГИС Канадским статистическим управлением в 1980-х годах и со временем развивалось, поскольку улучшенные инструменты и навыки способствовали более продвинутому геопространственному анализу. Такие данные могут раскрыть новые идеи и взаимосвязи между данными, которые были бы невозможны при изолированном анализе данных.

Возрастает спрос на информацию о местах, людях, человеческой деятельности, бизнесе, экономическом росте и благополучии. Геопространственные данные могут предоставить это и представляют собой ключевой аспект при рассмотрении равенства доступа к государственным (и другим) услугам и при измерении эффективности и действенности предоставления услуг правительствами.

В настоящее время легко признается, что интеграция статистической и геопространственной информации может сыграть ключевую роль в предоставлении данных для процессов принятия решений на местном, субнациональном, национальном, региональном и глобальном уровнях. В частности, это жизненно важно для измерения и мониторинга целей и глобальной системы показателей для ЦУР. Это также будет жизненно важным элементом для будущих переписей. Такая интегрированная информация также позволит проводить сравнения внутри стран и между странами более согласованным образом.

Перед использованием геопространственных данных НСО необходимо будет рассмотреть стратегические вопросы о том, какую ценность они могут принести, какие риски связаны, как работать вместе с картографическими агентствами, как улучшить кодирование местоположения в статистике и как лучше всего согласовать методы между статистики и картографические агентства.

8.4.1 Типы геопространственных данных

Есть два основных типа или формы геопространственных данных:

а) **Вектор** - в этой форме используются точки, линии и многоугольники для представления пространственных объектов, таких как города, дороги и ручьи. Векторные модели полезны для хранения данных с дискретными границами, такими как границы страны, земельные участки и улицы.

б) **Растр** - в этой форме используются ячейки (компьютер часто использует точки или пиксели) для представления пространственных объектов. Города - это отдельные ячейки, дороги - это линейные последовательности ячеек, а потоки - это совокупности соседних ячеек. Растровые модели полезны для хранения данных, которые постоянно

меняются, например, на аэрофотоснимке, спутниковом снимке, поверхности с химическими концентрациями или поверхности возвышения.

Ссылки на руководства, передовой опыт и примеры:

- Евростат - Геопространственный анализ в Евростате - набор методов и инструментов для изучения пространственно-временных отношений, присущих данным, с использованием примеров, когда пространственный анализ проводился со статистическими данными Европейского Союза (ЕС).
- Руководящие принципы ФАО по использованию продуктов дистанционного зондирования для улучшения статистики прогнозов производства сельскохозяйственных культур в странах Африки к югу от Сахары.
- СООН - Картирование плотности посевов. Статистика земельного покрова и землепользования. Пространственный и статистический анализ исторических климатических данных. Карта распределения плотности населения городских и сельских систем.

8.4.2 Проблемы для НСУ при использовании геопространственных данных

Что касается многих областей технологий, основная проблема для НСУ при использовании геопространственных данных заключается в наборе и удержании сотрудников с необходимыми навыками. Существует высокий спрос на специалистов по ГИС, и НСУ часто не в состоянии конкурировать с другими государственными органами и частным сектором на рынке труда.

Дополнительной проблемой является сотрудничество с картографическими агентствами, что привело к формированию Глобального управления геопространственной информацией ООН (UN-GGIM). Целью UN-GGIM является обеспечение совместной работы национальных картографических и кадастровых органов и национальных статистических служб для содействия более эффективному управлению и доступности геопространственной информации, а также ее интеграции с другой информацией, исходя из потребностей и требований пользователей.

ГИС-специалисты (также называемые ГИС-аналитиком, ГИС-техником и картографом) необходимы для создания и поддержки баз данных ГИС, а также для использования программного обеспечения ГИС для анализа содержащейся в них пространственной и непространственной информации. Они также анализируют данные ГИС для определения пространственных отношений, выполняют геопространственное моделирование или пространственный анализ и создают тематические карты.

Идеальным вариантом для НСО является наличие собственного специализированного подразделения картографии или ГИС, но это выходит за рамки возможностей многих НСО, особенно в развивающихся странах. Однако существует множество доступных программных продуктов ГИС с открытым исходным кодом, а также бесплатные источники данных ГИС.

Чтобы обеспечить возможность взаимодействия источников данных, данные съемки должны, по возможности, иметь географическую привязку и связываться с геопространственными данными.

Примеры программного обеспечения ГИС с открытым исходным кодом:

a) GRASS GIS - программный пакет, используемый для управления и анализа геопространственных данных, обработки изображений, создания графики и карт, пространственного моделирования и визуализации.

б) ILWIS (Интегрированная система информации о земельных и водных ресурсах) - это географическая информационная система и программное обеспечение дистанционного зондирования для векторной и растровой обработки. Его функции включают оцифровку, редактирование, анализ и отображение данных, а также создание карт.

с) OpenJUMP - это ГИС с открытым исходным кодом, которая может читать и записывать файлы карт. Он также может считывать данные из пространственных баз данных и может использоваться в качестве средства просмотра данных ГИС.

г) MapWindow - набор программируемых картографических компонентов для анализа и моделирования.

д) QGIS - это кроссплатформенное настольное приложение географической информационной системы, которое поддерживает просмотр, редактирование и анализ геопространственных данных.

е) SAGA - это географическая информационная система, используемая для редактирования пространственных данных.

Источники данных ГИС :

а) Esri Open Data обеспечивает доступ к более чем 67 тысячам наборов открытых данных от организаций по всему миру. Данные можно искать по теме или местоположению и загружать в различных форматах ГИС.

b) NASA Earth Observation (NEO) может просматривать и загружать изображения спутниковых данных с группировки спутников NASA Earth Observing System.

с) Центр социально-экономических данных и приложений НАСА (SEDAC). SEDAC - это центр данных Системы данных и информации НАСА (EOSDIS), который предоставляет библиотеки загружаемых карт и данных.

г) Natural Earth - это общедоступный набор картографических данных, содержащий интегрированные векторные и растровые данные, а также средства для создания карт с помощью картографии или программного обеспечения ГИС.

д) OpenStreetMap собирает геоданные и предоставляет их бесплатно. Это сообщество картографов, которые ежедневно редактируют географию как OpenStreetMap.

f) Открытая топография обеспечивает портал для топографических данных и инструментов с высоким пространственным разрешением. Open Topography обеспечивает доступ к топографическим данным с высоким разрешением, ориентированным на науки о Земле, а также к соответствующим инструментам и ресурсам.

g) Сторожевые спутниковые данные. Пространственные данные высокого разрешения со спутника Sentinel Европейского космического агентства доступны общественности бесплатно. Центр открытого доступа Copernicus обеспечивает полный, бесплатный и открытый доступ к пользовательским продуктам.

а) Terra Populus (TerraPop) включает и объединяет данные переписей из более чем 160 стран мира, а также данные об окружающей среде, описывающие земной покров, землепользование и климат.

h) Экологический обозреватель данных ЮНЕП является источником наборов данных, используемых ЮНЕП и ее партнерами в отчете «Глобальная экологическая перспектива» (ГЭП) и других комплексных экологических оценках. Его онлайн-база данных содержит более 500 различных переменных, таких как национальная, субрегиональная, региональная и глобальная статистика или наборы

геопространственных данных (карты), охватывающие такие темы, как пресная вода, население, леса, выбросы, климат, бедствия, здоровье и ВВП.

i) USGS Earth Explorer является источником данных географических информационных систем (ГИС). Геологическая служба США собирает, отслеживает, анализирует и предоставляет научные данные о состоянии, проблемах и проблемах природных ресурсов.

Ссылки на руководства, передовой опыт и примеры:

- Тематические карты Центрального статистического управления Латвии .
- Европейский форум по географии и статистике (EFGS) .
- Евростат - Объединение статистики и геопространственной информации в Европейской статистической системе (ESS).
- Объединение статистики и геопространственной информации в Европейской статистической системе (ESS).
- Стандарты и инфраструктура данных для статистических и пространственных структур . Статистическое управление Эстонии - примеры пространственного анализа .
- Портал геоданных Национального института статистики Руанды .
- Статистическое управление Кореи - ГИС, карты и статистика .
- Статистическое управление Польши: Статистический атлас Польши.
- Статистическое управление Швеции - Внедрение статистической геопространственной основы .
- Группа экспертов ООН по интеграции статистической и геопространственной информации - Глобальная статистическая геопространственная структура, связывающая статистику и место .
- Целевая группа ООН по спутниковым изображениям и геопространственным данным обеспечивает стратегическое видение, направление и разработку глобального плана работы по использованию спутниковых изображений и геопространственных данных для официальной статистики и показателей для целей развития на период после 2015 года.
- Руководство ЮНИСЕФ по использованию геопространственных технологий .

5 Большие данные

Термин «большие данные» в целом относится к данным, генерируемым бизнес-транзакциями, социальными сетями, телефонными журналами, устройствами связи, веб-парсингом, датчиками и т. Д. (См. Главу 14, раздел 14.2.7 « *Большие данные»*). Общее введение в понятие «большие данные» можно найти в книге Виктора Майера-Шенбергера и Кеннета Кукьера « *Большие данные* » (2013).

Большие данные вызывают значительный и растущий интерес со стороны НСУ в связи с возможностью дополнения традиционной статистики. Это особенно важно в контексте необходимости измерения и мониторинга прогресса в достижении Целей устойчивого развития (ЦУР) и других задач.

Большие данные могут дополнить, заменить или частично заменить существующие статистические источники, такие как опросы, или предоставить дополнительную статистическую информацию, но с других точек зрения. Его также можно использовать для

улучшения оценок или для генерации совершенно новой статистической информации в данной статистической области или во всех областях.

Большие данные широко используются в коммерческом секторе для бизнес-аналитики.[1] , но пока имеется меньше свидетельств его использования в мире официальной статистики. Несмотря на большие надежды на использование больших данных, реальность такова, что, хотя технологии, необходимые для обработки этих огромных наборов данных, доступны и развиваются, самым большим препятствием для НСО зачастую является получение доступа к данным. Такое отсутствие доступа может быть связано с нежеланием компании раскрывать свои данные, юридическими препятствиями, затратами или опасениями по поводу конфиденциальности. Однако там, где для НСУ доступны большие данные, такие как веб-сайты или датчики, администрируемые государственными органами, такие как датчики дорог, они уже успешно используются для экспериментальной или даже официальной статистики.

8.5.1 Типы больших данных

Есть несколько категорий больших данных.

а) **Структурированные данные:** все данные, полученные от датчиков, веб-журналов и финансовых систем, классифицируются как данные, генерируемые машинами. К ним относятся медицинские устройства, данные GPS, данные статистики использования, полученные серверами и приложениями, а также огромный объем данных, которые обычно передаются через торговые платформы, и это лишь некоторые из них. Структурированные данные, созданные человеком, в основном включают все данные, которые человек вводит в компьютер, например его имя и другие личные данные. Когда человек щелкает ссылку в Интернете или даже делает ход в игре, создаются данные.

б) **Неструктурированные данные: в** то время как структурированные данные находятся в традиционных базах данных «строка-столбец», неструктурированные данные противоположны им - они не имеют четкого формата в хранилище. Остальные созданные данные, около 80% от общего количества, составляют неструктурированные большие данные. До недавнего времени с этим ничего не оставалось, кроме как хранить или анализировать вручную. Неструктурированные данные также классифицируются в зависимости от их источника на генерируемые машинами или людьми. Машинно-сгенерированные данные составляют все спутниковые изображения, научные данные различных экспериментов и данные радаров, полученные с помощью различных аспектов технологий. Неструктурированные данные, созданные человеком, включают данные социальных сетей, мобильные данные и контент веб-сайтов. Это означает, что изображения, которые мы загружаем в Facebook или Instagram, видео, которые мы смотрим на YouTube, и даже текстовые сообщения, которые мы отправляем, вносят свой вклад в массу неструктурированных данных.

с) **Полуструктурированные данные:** информация, которая не представлена в традиционном формате базы данных как структурированные данные, но содержит некоторые организационные свойства, которые упрощают обработку, включается в полуструктурированные данные. Например, документы NoSQL считаются частично структурированными, поскольку они содержат ключевые слова, которые можно использовать для простой обработки документа.

8.5.2 Источники больших данных

Существует четыре основных источника больших данных.

а) **Транзакционные данные** генерируются из всех ежедневных транзакций, которые происходят как онлайн, так и офлайн. Счета, платежные поручения, записи хранения, квитанции о доставке - все это данные о транзакциях. Однако сами по себе данные почти бессмысленны, и большинству организаций сложно понять, какие данные они генерируют и как их можно использовать. Бизнес-операции: данные, полученные в результате бизнес-деятельности, могут быть записаны в структурированные или неструктурированные базы данных. Большой объем информации и периодичность ее производства (поскольку иногда эти данные производятся в очень быстром темпе), тысячи записей могут быть созданы за секунду, когда крупные компании, такие как сети супермаркетов, регистрируют свои продажи.

б) **Датчики / измерители и записи активности с электронных устройств** : качество такого источника зависит в основном от способности датчика проводить точные измерения, как это ожидается. Машинные данные определяются как информация, генерируемая промышленным оборудованием, датчиками, установленными в машинах, и даже веб-журналами, отслеживающими поведение пользователей. Ожидается, что этот тип данных будет расти в геометрической прогрессии по мере того, как Интернет вещей (IoT) становится все более распространенным и распространяется по всему миру. Такие датчики, как медицинские устройства, интеллектуальные счетчики, дорожные камеры, спутники, игры и быстрорастущий Интернет вещей, в самом ближайшем будущем обеспечат высокую скорость, ценность, объем и разнообразие данных.

с) **Социальные взаимодействия** : сюда входят данные, полученные в результате взаимодействия людей через сеть. Наиболее распространены данные, полученные в социальных сетях. Этот тип данных зависит от точности алгоритмов, применяемых для извлечения смысла содержимого, которое обычно встречается в виде неструктурированного текста, написанного на естественном языке. Некоторыми примерами анализа, который делается на основе этих данных, являются анализ настроений, анализ тем тенденций и т. Д. Социальные данные поступают из лайков, твитов и ретвитов, комментариев, загрузок видео и общих медиа, которые загружаются и распространяются через популярные социальные сети в мире. платформы. Такие данные дают бесценную информацию о поведении и настроениях потребителей и могут иметь огромное влияние на маркетинговую аналитику. Общедоступная сеть - еще один хороший источник социальных данных, и такие инструменты, как Google Trends, могут быть использованы для увеличения объема больших данных.

г) **Данные, генерируемые гражданами (CGD)** - это данные, производимые негосударственными субъектами при активном согласии и участии граждан, чтобы в первую очередь отслеживать, требовать или стимулировать изменения по вопросам, которые непосредственно их затрагивают. Данные, полученные гражданами, могут быть инновационным источником данных (вторичным источником данных) для производства официальной статистики и использоваться для поддержки эффективного отслеживания прогресса в достижении целей в области устойчивого развития (ЦУР).

ЕЭК ООН разработала многоуровневую классификацию источников больших данных с 24 категориями на самом низком уровне.

8.5.3 Проблемы с доступом и обработкой больших данных

8.5.3.1 Доступ к большим данным

Согласно руководству GLOS. Статья 6.1 «Производители официальной статистики имеют право на бесплатный доступ и сбор данных из всех государственных и частных источников данных, включая идентификаторы, на уровне детализации, необходимом для статистических целей. В долгосрочной перспективе цель состоит в том, чтобы отрегулировать этот доступ в статистическом законе.

Существует ряд потенциальных препятствий, которые НСУ необходимо преодолеть для получения доступа к источникам больших данных. К ним относятся следующие:

а) опасения частных компаний по поводу потери своего конкурентного преимущества;

б) правовые ограничения, касающиеся конфиденциальности и конфиденциальности информации о клиенте;

с) предприятия осознали ценность своих данных и не готовы их просто раздать:

г) затраты на создание необходимой инфраструктуры и обучение персонала для непрофильной деятельности.

НСУ необходимо преодолеть такие юридические требования и заключить соглашение с предприятиями о получении доступа к частным источникам данных, чтобы облегчить использование больших данных в статистических целях. Если соглашения могут быть достигнуты с предприятиями, которым принадлежат данные, существует ряд бизнес-моделей, которые могут обеспечить обмен данными между частными корпорациями и НСО. В документе PARIS21 « Доступ к новым источникам данных для статистики: бизнес-модели и стимулы для корпоративного сектора » перечислены следующие модели:

а) **Собственное производство статистики** : собственное производство статистической модели во многих отношениях является самой традиционной или стандартной моделью. Сегодня он используется большинством статистических агентств и, как таковой, связан с известным набором рисков и возможностей. С положительной стороны модель позволяет владельцу данных сохранять полный контроль над созданием и использованием необработанных данных. Конфиденциальность пользователей может быть защищена посредством деидентификации, а сгенерированные показатели могут быть достаточно агрегированы, чтобы считаться безопасными для совместного использования. С точки зрения безопасности наиболее предпочтительным вариантом является собственное производство статистики.

б) **Передача наборов данных конечным пользователям:** в этой модели наборы данных перемещаются непосредственно от владельца данных к конечному пользователю. Модель дает конечному пользователю значительно большую гибкость в использовании данных. Как правило, необработанные данные деидентифицируются, отбираются выборки и иногда агрегируются, чтобы избежать возможной повторной идентификации. Усилия по деидентификации должны гарантировать, что данные не могут быть повторно идентифицированы путем сопоставления их с внешними данными. Поскольку деидентификация никогда не бывает абсолютной, даже если используются самые сложные методы анонимизации, данные в этой модели обычно передаются ограниченному числу конечных пользователей в соответствии со строгими соглашениями о неразглашении и использовании данных, которые помогают обеспечить определенный уровень контроля. и конфиденциальность.

с) **Обеспечение удаленного доступа к данным** : в модели удаленного доступа владельцы данных предоставляют полный доступ к данным конечным пользователям, сохраняя при этом строгий контроль над тем, какая информация извлекается из баз данных и наборов данных. В этой модели личная идентификация анонимна, но данные фактически не огрубляются. Данные не покидают территорию владельца

данных; скорее, конечному пользователю предоставляется защищенный доступ для анализа данных и вычисления соответствующих показателей. После этого конечному пользователю разрешается извлекать только окончательные агрегированные метрики после завершения анализа данных. Этот метод часто используется в исследованиях, в определенных партнерских отношениях между владельцем данных и группой исследователей, в соответствии с очень строгими соглашениями о неразглашении и использовании данных. На устройствах хранения данных осуществляется строгий мониторинг входящего и выходного трафика, чтобы гарантировать, что данные не будут удалены. Основным стимулом в этой модели является то, что пользователи получают выгоду от бесплатных ресурсов для исследования своих данных.

d) **Использование доверенных** третьих **сторон** (ТЗР): в модели доверенных третьих сторон (ТЗР) ни владелец данных, ни пользователь данных не поддерживают бремя безопасности, связанное с размещением самих данных. Вместо этого обе стороны полагаются на доверенную третью сторону для размещения данных и обеспечения безопасного доступа к источнику данных. Данные анонимны в том смысле, что личные идентификаторы защищены методами хеширования. Кроме того, конечный пользователь не имеет прямого доступа к необработанным данным. Вместо этого конечные пользователи должны сделать запрос отчетов или других промежуточных результатов в ТЗР, который обеспечивает защиту данных.

д) **Перемещение алгоритмов,** а не данных. В этой модели общие алгоритмы позволяют повторно использовать программное обеспечение несколькими владельцами частных данных, желающими выполнять аналогичные аналитические функции на одном или нескольких наборах данных. Например, такая модель может быть эффективной в случае, когда несколько национальных операторов электросвязи хотят оценить плотность населения (или другие структуры населения) на основе своих коллективных данных. Наборы данных от разных операторов необязательно объединять. Вместо этого, хотя аналитические функции, выполняемые для каждого набора данных, могут быть идентичными, сами наборы данных могут оставаться отдельными и находиться под отдельным контролем. Результаты могут быть получены каждым оператором независимо, а затем агрегированные результаты могут быть объединены для получения всеобъемлющего национального или регионального анализа.

Чтобы получить надежный и устойчивый доступ к источникам больших данных, НСУ необходимо формировать стратегические альянсы с производителями данных, что может оказаться длительным процессом без гарантии успеха. Частные корпорации осознают ценность своих данных, не желают тратить ресурсы на деятельность, не являющуюся критически важной и несущую потенциальные риски нарушения деловой информации и конфиденциальности. Правительствам надлежит принять закон, обязывающий корпорации предоставлять свои данные НСУ для использования в общественных интересах. Однако это может занять много лет.

Вопрос о получении доступа к частным источникам больших данных необходимо решать на наднациональном уровне, особенно потому, что многие компании, производящие большие данные, являются международными концернами. Правовые аспекты сложны и трудно разрешить юридически на уровне страны. Более того, НСО - не единственные организации, заинтересованные в получении доступа к большим данным для общественных целей. В ЕС, например, этим вопросом занимается обширная группа экспертов . ЕС также предоставил некоторые рекомендации , в частности, упомянув статистику.

8.3.5.2 Проблемы при обработке больших данных

а) **Конфиденциальность данных** : при использовании больших данных наибольший риск связан с конфиденциальностью данных. Предприятия по всему миру используют конфиденциальные данные, личную информацию о клиентах и стратегические документы. Инцидент с безопасностью может не только повлиять на критически важные данные и привести к ухудшению репутации, но и привести к судебным искам и финансовым штрафам. Принятие мер по обеспечению конфиденциальности данных имеет жизненно важное значение - как показали недавние громкие дела, в случае недостаточной защиты эту информацию можно использовать для профилирования людей и передать третьим сторонам, что приведет к потере доверия потребителей. Таким образом, НСУ необходимо гарантировать, что используемые источники данных и показатели получены без какого-либо нарушения режима конфиденциальности или конфиденциальности.

б) **Затраты** : НСУ также должны инвестировать в уровни безопасности и адаптировать традиционные методы информационных технологий, такие как криптография, анонимизация и контроль доступа пользователей, к характеристикам больших данных. Несмотря на то, что в идеале доступ к данным для NSO должен быть бесплатным, как и доступ к административным данным, может потребоваться оплатить единовременные расходы на подготовку системы передачи данных, такой как API.

с) **Качество данных** : большие данные часто в значительной степени неструктурированы, что означает, что такие источники данных не имеют заранее определенной модели данных и плохо вписываются в обычные реляционные базы данных. Разнообразные структуры также вызывают проблемы интеграции данных, поскольку данные, необходимые для анализа, поступают из разных источников в самых разных форматах, таких как журналы, call-центры, веб-аналитики и социальные сети. Форматы данных, очевидно, будут отличаться, и сопоставление их может быть проблематичным. Ненадежные данные: большие данные не всегда создаются с использованием строгих методов проверки, что может отрицательно сказаться на качестве. Он не только может быть неточным и содержать неверную информацию, но также может содержать дублирование и другие противоречия.

г) **Методы** : использование больших данных может потребовать новых методов и приемов. Например, могут потребоваться новые методы моделирования, особенно если большие данные используются для получения ранних индикаторов или даже прогнозирования текущей погоды. Искусственный интеллект и методы глубокого обучения могут использоваться для обработки неструктурированных текстовых сообщений или спутниковых изображений.

е) **Непостоянство данных** : НСУ не может гарантировать, что источник данных будет надежным, поскольку он не контролирует или не имеет отношений с владельцем данных, как с традиционными источниками данных. Форматы могут измениться в любое время без предупреждения, что может сделать сбор данных и последующие процессы, которые были внедрены NSO, неработоспособными. Источники данных могут даже полностью исчезнуть, если изменятся бизнес-правила, генерирующие данные.

f) **Пробелы в данных** : ЦУР имеют фундаментальное обязательство не оставлять никого позади. Уязвимые группы населения могут не быть охвачены большими данными, если использование таких источников, как мобильные телефоны, недоступно для беднейших и наиболее маргинализированных групп общества.

В перспективном документе CBS Нидерландов и Статистического управления Канады о будущем расширенном сборе данных , представленном на 62-м Всемирном статистическом конгрессе ISI 2019 в Куала-Лумпуре, обсуждается, как данные датчиков и

данные, поступающие с платформ данных (государственных или частных), размещенных за пределами национальных статистических служб, могут играть более важную роль. в будущем сбора данных и максимально использовать преимущества этих источников данных для создания «умной статистики».

Умную статистику можно рассматривать как расширенную роль официальной статистики в мире, пропитанном умными технологиями. Интеллектуальные технологии включают в себя автоматизированные интерактивные технологии в реальном времени, оптимизирующие физическую работу бытовых и потребительских устройств. Сами статистические данные затем будут преобразованы в интеллектуальную технологию, встроенную в интеллектуальные системы, которая преобразует «данные» в «информацию».

Некоторые из основных проблем и возможностей, изложенных в документе о видении, заключаются в следующем:

а) **Методология** : связывание различных источников данных и проверка данных, которые не были собраны специально для официальных статистических целей (административные данные и данные датчиков), требуют совершенно новых и передовых методологических концепций. Ключевым моментом является переход от методологии обследования к методологии данных.

б) **Качество** : своевременность станет важной характеристикой качества статистической продукции, и, конечно же, своевременность может влиять на точность статистической информации. Точность не является синонимом качества, но это одна из характеристик, определяющих качество статистических продуктов. Пока точность информации известна и указана конечному пользователю, это не должно быть проблемой. Конечно, было бы интересно провести исследования по уменьшению потенциального компромисса между своевременностью и точностью.

с) **Доступ к данным с точки зрения социальной приемлемости и правовых рамок** : социальная приемлемость является ключом к получению доступа к частным данным. Это означает, что ОПС должны быть прозрачными и иметь возможность демонстрировать и объяснять обществу ценностное предложение (общественное благо) и учитывать озабоченности общества по поводу доверия, конфиденциальности и конфиденциальности. В то же время необходимо разработать правовую базу, чтобы гарантировать, что НСУ могут использовать все эти новые источники данных в максимальной степени, и чтобы общество могло извлечь выгоду из дополнительных преимуществ, которые НСУ потенциально могут предоставить.

d) **Доступ к данным в отношении технологии и методологии** : ключевые слова для будущего доступа к данным - сбор, соединение и связывание. Технология получения безопасного доступа к данным в сочетании с соответствующей методологией и алгоритмами, гарантирующими конфиденциальность и конфиденциальность, является одним из основных направлений технологического развития на ближайшее будущее. Многосторонние вычисления, совместное использование данных с сохранением конфиденциальности (PPDS) и привязка записей с сохранением конфиденциальности (PPRL) являются потенциально многообещающими технологическими достижениями, которые необходимо развить в виде надежного набора методов.

Ссылки на руководства, передовой опыт и примеры:

- Большие данные в официальной статистике , CBS (2020).
- Бухарестский меморандум об официальной статистике в информационном обществе , 104-я конференция DGINS, Бухарест (2018)

- Рекомендации по доступу к данным частных организаций для официальной статистики , Глобальная рабочая группа по большим данным для официальной статистики (2016).
- Схевенингенский меморандум об использовании больших данных в официальной статистике , Евростат (2013).
- Глобальная рабочая группа ООН по большим данным

[1] https://alltimestech.com/2019/10/22/big-data-and-business-analytics-market-outline-and-pipeline-review-from-2019-2025-international-business-machines-ibm -корпорация-оракул-Майкрософт-корпорация /