Европейская экономическая комиссия ООН

Использование административных и вторичных источников данных в официальной статистике

Руководство по принципам и практике

Организация Объединенных Наций

Нью-Йорк и Женева, 2011

Примечание

Использованные определения и представление материала в настоящей публикации не подразумевают выражения какого-либо мнения со стороны Секретариата Организации Объединенных Наций в отношении правового статуса той или иной страны, территории, города или района или их полномочных органов, или же делимитации их границ, или установления их пределов.

Выражение признательности

ЕЭК ООН хотела бы выразить признательность за ценные вклады более чем двухсот участников международного тренинга по использованию административных источников для статистических целей, и особенно - за вклады названных ниже презентаторов, каждый из которых существенно помог улучшить настоящие рекомендации:

Г-н Driss Afza, проекты MEDSTAT;
Г-жа Sue Fendall, Национальное статистическое управление Соединенного Королевства;
Г-жа Riitta Harala, Статистика Финляндии;
Г-н John C Hughes, Национальное статистическое управление Соединенного Королевства;
Г-н Ben Humberstone, Национальное статистическое управление Соединенного Королевства;
Г-н Pekka Myrskylä, Статистика Финляндии
Г-жа Kaija Ruotsalainen, Статистика Финляндии

Сведения об авторских правах

Предисловие

Статистические организации всего мира оказываются под возрастающим давлением в направлении совершенствования эффективности статистического производственного процесса, особенно в части экономии затрат и кадровых ресурсов. При этом имеется политическая потребность снизить нагрузку на респондентов статистического наблюдения. Это имеет место, прежде всего, там, где респондентами являются субъекты предпринимательской деятельности, ибо многие правительства видят в снижении бюрократии основное средство поддержки и ускорения развития бизнеса.

Под этим давлением статистики все более вынуждены использовать при сборе данных альтернативы традиционным подходам к обследованию. Очевидное решение вопроса может состоять в выяснении того, не существуют ли уже где-либо пригодные к употреблению данные. Многие нестатистические организации собирают данные в различных формах, и хотя эти данные редко являются непосредственной заменой собираемых при статистических обследованиях данных, они часто предоставляют возможности порой путем комбинации разных источников сделать полную или частичную замену прямого сбора статистических данных.

Степень использования административных источников в статистическом производственном процессе значительно варьирует от страны к стране, от стран, имеющих развитые и в полной мере функционирующие статистические системы на основе регистров, к странам, только лишь приступившим к осуществлению этого подхода.

Хотя существует несколько материалов по предмету, до настоящего времени нет общих международных методологических рекомендаций для помощи тем, кто находится на ранних стадиях использования административных данных. Данное руководство ставит своей целью восполнить этот пробел. Оно основывается на материалах, разработанных за десять лет для международного учебного курса по использованию административных источников для статистических целей. Этот курс уже проводился более десяти раз для официальных статистиков из многих стран Европы, Западной и Центральной Азии, и Северной Африки.

При каждом проведении курс улучшался и углублялся за счет обмена опытом и получения отзывов от участников. Его улучшению очень способствовал вклад многих приглашенных для презентаций экспертов из Статистики Финляндии и британского Национального статистического управления.

Стивен Вейл, ЕЭК ООН, руководитель курса

Содержание

Примечания

1) Примечания к ссылкам

Настоящее руководство содержит много ссылок на другие документы, веб-сайты и публикации. Чтобы помочь желающим обратиться к этим ссылкам, везде, где возможно, приведены Интернет-адреса. Все они проверялись при написании документа, но нет гарантии, что они будут актуальны при его чтении. Если читатель найдет поврежденные ссылки, просим сообщить это по адресу support.stat@unece.org.

2) Примечания к упражнениям

Упражнения в конце глав 6 и 7 взяты из курса, на котором основано настоящее руководство. Они включены в качестве практических примеров для подкрепления теории, представленной в этих главах.

1. Что является административными и вторичными источниками

1.1 Введение

Перед началом рассмотрения практических аспектов использования данных из административных и вторичных источников есть смысл уделить некоторое время тому, чтобы ясно определить, что означают эти термины. В имеющейся в настоящее время литературе существует несколько определений, наиболее подходящие из которых рассматриваются в этой главе. Глава заканчивается предложением относительно простого и широкого определения, которое затем используется как основа последующего текста настоящего руководства.

1.2 Традиционные определения

Административные источники были традиционно определены как массивы данных, находящиеся в распоряжении других органов правительства, собираемые и используемые для целей администрирования налогов, компенсационных выплат или услуг. Возможно, наиболее исчерпывающее из традиционных определений было предложно Гордоном Брэкстоуном [Gordon Brackstone] из Статистики Канады в его работе 1987 года "Статистические аспекты административных данных: проблемы и вызовы" [1] . Брэкстоун идентифицировал четыре отличительные характеристики административных данных:

1. Агенты, которые предоставляют данные статистическому агентству, и единицы, к которым данные относятся, являются различными (в отличие от большинства статистических наблюдений);

2. Данные изначально собираются для определенных, не статистических целей, что может влиять на характер работы с единицей – источником;

3. Задачей является полное покрытие целевой совокупности;

4. Контроль методов, посредством которых административные данные собираются и обрабатываются, остается за административным агентством.

Это определение, в общем, соответствует тому, что предложено Инициативой по Стандарту обмена статистическими данными и метаданными[2].

"Хранимые данные содержат информацию, собираемую и поддерживаемую для целей осуществления одного или более административных регулирований".

В течение 1996-97 гг. целевые рабочие группы внутри Евростата изучали пути

[1] Brackstone G J: "Statistical Issues of Administrative Data: Issues and Challenges", в "Статистическое использование административных данных – Международный симпозиум", организован Статистикой Канады, 23-25 ноября 1987 г. (Материалы опубликованы Статистикой Канады, Оттава, декабрь 1988 г.).
[2] См.: www.sdmx.org

лучшей координации работ, связанных с использованием административных источников в различных областях статистики. Эти целевые группы использовали простую типологию источников данных, чтобы понять, как следует определять административные источники. Во-первых, все источники данных были подразделены на первичные источники (данные, собираемые для статистических целей) и вторичные источники (все другие данные). Традиционное, или "узкое" определение административных источников охватывает лишь нестатистические источники государственного сектора, тогда как более широкое определение должно бы включать также источники частного сектора.

Более широкий подход согласуется с определением административных данных, принятым Конференцией Европейских статистиков в публикации "Терминология по статистическим метаданным"[3]:

> *"Данные собираемые источниками, внешними по отношению к статистическим службам"*.

Узкое и более широкое определения можно показать графически в виде следующего:

Рис. 1.1 – Узкое определение

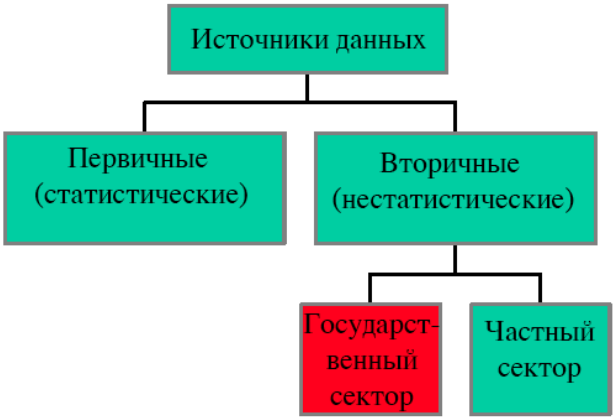
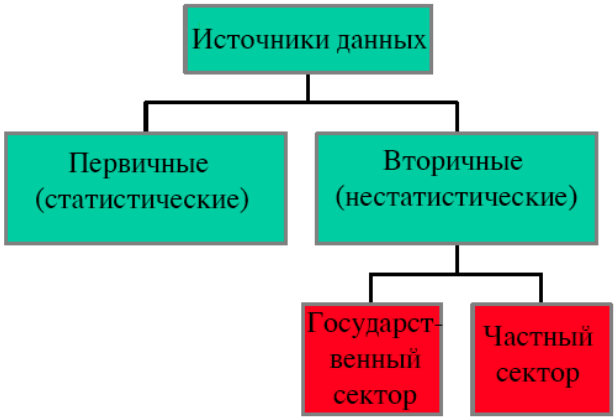


Рис. 1.2 – Широкое определение



Таким образом, при узком определении административные источники являются подмножеством вторичных источников, в то время как при широком определении эти термины синонимичны.

Налицо увеличение числа аргументов в пользу широкого определения, включая:

•Нарастающая приватизация функций правительства:
В некоторых странах регулятивные функции, которые обычно выполнялись департаментами или агентствами правительства, передаются частным, либо полу-частным организациям. Типичные примеры обычно находятся в секторах

[3]См.: www1.unece.org/stat/platform/download/attachments/9110092/Metadata+terminology+2000.pdf?version=1

здравоохранения, образования или коммунального обслуживания[public utilities], где бывшие государственные монополии все более замещаются частными компаниями или некоммерческими институтами.

Функции по регистрации, включая деятельность административных регистров от имени правительственных департаментов, также рассматриваются на предмет приватизации в нескольких странах. Это означает, что традиционные различия между функциями государственного и частного секторов становятся все более расплывчатыми, и что традиционное "узкое" определение административных источников становится слишком ограниченным.

•Развитие информации в частном секторе и "добавляющих стоимость реселлеров":

Количество цифровой информации в мире нарастает экспоненциально, увеличиваясь на порядок приблизительно каждые 5 лет. Даже если лишь малая доля этой "лавины данных" заинтересует официальных статистиков, объем данных и охват покрываемых ими тем окажется все равно громадным.

При этом коммерческая ценность данных начинает становиться очевидной, и в частном секторе быстро возрастает спрос на данные. Это началось с создания и продажи адресных списков для целей маркетинга, расширилось до предоставления сведений о кредитных рейтингах и информации по бизнес-аналитике, а теперь распространилось фактически на все типы данных. Поскольку масштаб этого рынка увеличился, также возросло число бизнесов, стремящихся извлечь прибыль из этого. Частный сектор осознает, что данные являются очень ценным продуктом. Сравнительно недавнее развитие процесса привело к появлению на рынке данных частного сектора "добавляющих стоимость реселлеров" [value added re-sellers]. Эти бизнесы берут существующие данные из различных источников государственного и частного секторов, комбинируют и очищают их, иногда подтверждают правильность, а затем перепродают их другим организациям. Примерами являются продавцы бизнес-данных, как-то: Дан энд Брэдстрит, Бюро Ван Дейк и Hoppenstedt Bonnier.

Этот вид источников данных может интересовать поставщиков официальной статистики, ибо может случиться, что эти поставщики данных из частного сектора смогут, вообще говоря, дешевле обрабатывать и предоставлять данные, чем статистические организации – часто просто потому, что они могут распределять затраты между рядом потребителей. Проект "Eurogroups" по разработке европейского статистического регистра групп предприятий пользуется такими источниками именно по этой причине.

Альтернативой непосредственному использованию микро-данных из таких источников может быть использование агрегированных данных для целей бенчмаркинга, сопоставления степени покрытия целевых совокупностей между

частными источниками и официальными статистическими регистрами. Опыт сопоставления степени покрытия статистическим бизнес-регистром в Соединенном Королевстве с покрытием источниками из частного сектора выявил недостаточный охват предпринимательской деятельности в статистике по проблемным кварталам [inner-city] и курортным зонам, что иллюстрирует трудности, связанные с покрытием маргинальной и сезонной деятельности в официальной статистике, а также дает хороший индикатор масштабов недопокрытия такого рода[4].

•Заинтересованность пользователей в новых типах данных

Пользователи официальной статистики постоянно запрашивают новые типы данных. В силу требований по снижению затрат и нагрузки на респондентов статистических обследований становится непросто предпринимать новые, отвечающие этим требованиям обследования, поэтому статистики все более вынуждены искать альтернативные решения. Поскольку объем, информационное содержание и покрытие у источников частного сектора возрастает, постольку возрастает их привлекательность в качестве альтернативы статистическим наблюдениям.

1.3 Типы административных источников

Как было рассмотрено в предыдущих параграфах, потенциальная сфера применения административных источников, которые могут использоваться для статистических целей, является значительной и расширяющейся. Следующий список не претендует быть исчерпывающим, но он ориентирован на показ сфер и типов потенциальных источников данных, что является решающим шагом, приближающим к рабочему определению административных источников.

• Налоговые данные

- Налог на личный доход
- Налог на добавленную стоимость (VAT)
- Налог на предпринимательство / прибыль
- Налоги на собственность
- Импортные / экспортные пошлины

•Данные органов социального обеспечения

- Отчисления
- Выплаты
- Пенсии

[4]Результаты этого опыта показаны в виде карты покрытия в материале “Развитие бизнес-статистики малых областей в Соединенном Королевстве”, который доступен по: http://live.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf

•Данные систем здравоохранения / образования

•Регистрационные системы личностей / бизнесов / собственности / транспортных средств

•Удостоверения личности / паспорта / водительские удостоверения

•Списки избирателей

•Регистры фермерских хозяйств

•Регистры местных органов самоуправления

•Разрешения на строительство

•Система лицензирования, например телевидения, товаров ограниченного производства и обращения

•Публикуемые счета бизнесов

•Данные внутреннего учета бизнесов

•Частные бизнесы, обладающие данными:

- Кредитные агентства
- Бизнес-аналитики
- Компании по коммунальному обслуживанию [utility companies]
- Списки абонентов телефонной сети
- Компании розничной торговли с магазинными кредитными картами пр.

1.4 Резюме

В заключение в этой главе обосновывается аргументация в пользу широкого определения административных и вторичных источников. Также подчеркивается потребность в креативном прогнозе потенциальной ценности новых типов источников данных. В связи с этим определение административных и вторичных источников не должно накладывать каких-либо ограничений на статистиков и должно быть широким сколь возможно. Поскольку, таким образом, термины “административные источники” и “вторичные источники” рассматриваются как синонимичные, в настоящем руководстве будет впредь использоваться лишь термин “административные источники”, покрывающий оба понятия.

Поэтому предлагается определение:

> **Административные источники есть фонды данных, содержащие информацию, которая изначально не собиралась для статистических целей.**

Это определение используется как базис последующего материала настоящего

руководства.

Вставка 1.1 – Взгляд в будущее: магазинные карты – потенциальный источник данных?

Магазинные карты являются типичным примером нового типа источников данных частного сектора. В обмен на выгоды, такие как скидки и эксклюзивные предложения, пользователи магазинных карт предоставляет магазину множество данных всякий раз, когда они используют ее. Если у Вас есть магазинная карта, магазин знает или может логически вывести следующие данные о Вас:

- Имя, адрес, пол, возраст
- Семейные обстоятельства (например, если Вы регулярно покупаете детские товары, игрушки, корм для животных, либо такие продукты, как пища в определенном количестве или определенного объема, то легко оценить потребление Вашего домохозяйства)
- Индикаторы формы занятости и дохода (например, время совершения ваших покупок может показывать, работаете ли Вы или нет, а типы покупаемых товаров могут указывать на величину располагаемого дохода)
- Другие показатели домохозяйства, такие как владение автомобилем (покупки моторного топлива или средств по уходу за автомобилем), религиозная принадлежность (покупка товаров, связанных с определенной религией, например, халяльное или кошерное мясо), и пр.

Это может казаться скорее крайним случаем потенциального источника, который вряд ли будет рассматриваться в ближайшем будущем применительно к целям официальное статистики. Однако некоторые страны прорабатывали вопрос об использовании данных лент кассовых аппаратов крупнейших ритейлеров в качестве источника данных о розничных продажах и ценах, а Статистика Новой Зеландии произвела экспериментальные ряды данных, используя данные трансакций по электронным картам[5].

[5]http://www.stats.govt.nz/browse_for_stats/Corporate/Corporate/nzae-2007/~/media/Statistics/Publications/NZAE/The%20development%20of%20electronic%20card%20transaction%20statistics/development-of-ect-statistics.ashx

Использование магазинных карт могло бы рассматриваться как следующий логический шаг, особенно при улучшении охвата путем объединения данных из различных систем магазинных карт, а также данных из других коммерческих источников. Если этот вид административных источников данных будет игнорироваться официальными статистиками, то много ли времени пройдет до тех пор, когда частные бизнесы с возможностью доступа к этим данным начнут предлагать приемлемые и более затрато-эффективные альтернативы ключевым официальным статистическим продуктам, таким как данные переписи населения?

2. Преимущества использования административных источников

2.1 Введение

В предыдущей главе были определены суть и границы административных источников, но по-настоящему не рассматривалось, почему эти источники интересны статистикам. В этой главе рассматриваются многие потенциальные преимущества использования административных источников в официальной статистике, направленные на дополнение или замену статистических источников. Конечно, не все является сплошным позитивом, наряду с преимуществами здесь также обычно присутствует ряд проблем, которые надо преодолеть. Эти проблемы и то, как они могут решаться, рассматривается в главе 4.

2.2 Затраты

Статистические наблюдения являются дорогим путем сбора данных. Вопросники должны быть разработаны, выборки спланированы (что может даже потребовать создания специальной основы выборки), с респондентами должен быть установлен контакт, и возможно повторно, чтобы побудить их к ответу, ответы должны быть обработаны и проконтролированы, а результаты – подсчитаны. Хотя компьютеры могут принять на себя большую часть нагрузки по обработке, весь порядок выполнения работы пока еще является трудозатратным, особенно стадия обеспечения получения ответов, которая, возможно, не может быть когда-либо полностью автоматизирована.

С традиционными переписями дело обстоит даже хуже, ибо их проведение значительно более масштабно. Национальные статистические организации по-прежнему проводят традиционные переписи населения, бизнесов, сельскохозяйственных производств и прочие, часто запрашивают специальное финансирование для таких мероприятий, поскольку они являются слишком затратными, чтобы быть покрытыми из обычного бюджета организаций. В силу этого традиционные переписи всегда на виду у политиков и потому уязвимы к изменениям политических приоритетов.

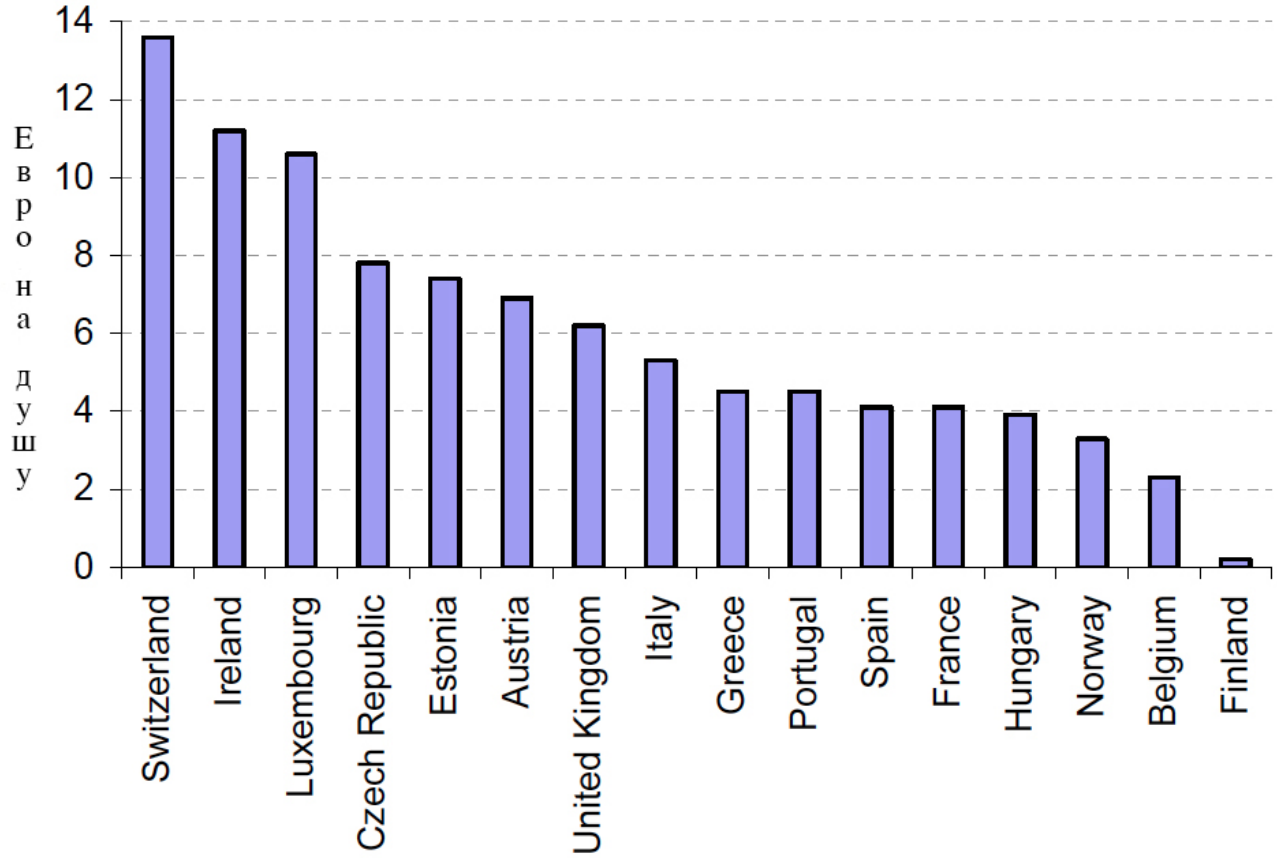
Хотя затраты на подготовку к использованию административных источников при производстве статистических продуктов могут легко оказаться столь же высокими, как и затраты на подготовку статистических наблюдений, затраты на эксплуатацию обычно являются значительно меньшими. Приведенные ниже таблица 2.1 и рисунок 2.1 показывают затраты на проведение переписей населения в странах Европейского Союза в 2000-2001 гг. Огромное различие в затратах в расчете на душу населения между Финляндией, где перепись целиком основывалась на административных источниках, и другими странами –

такими как Соединенное Королевство и Австрия, где использовались традиционные бумажные вопросники, – возможно является самым сильным имеющимся аргументом в пользу большего использования административных источников.

Таблица 2.1 – Затраты на перепись населения в некоторых странах Европейского союза

Страна	Всего затрат (миллионов евро)	Затраты на душу (евро)
Бельгия	24	2.3
Греция	50	4.5
Испания	167	4.1
Франция	248	4.1
Ирландия	44	11.2
Италия	298	5.3
Люксембург	5	10.6
Австрия	56	6.9
Португалия	46	4.5
Финляндия	0.8	0.2
Соединенное Королевство	367	6.2
Норвегия	15	3.3
Швейцария	99	13.6
Чешская Республика	80	7.8
Эстония	10	7.4
Венгрия	40	3.9

Рисунок 2.1 – Сравнительные затраты на перепись населения в расчете на душу



Источник: таблица 22 публикации Евростата “Документация по раунду переписи населения и жилья 2000 г. в странах ЕС, ЕАСТ [EFTA] и странах – кандидатах”[6].

Доступ к административным источникам часто является безвозмездным, особенно если данные берутся из государственного сектора. Но даже если требуется оплата, например, с целью компенсации затрат по извлечению данных и их передаче из государственного сектора, или для покупки данных из частного источника, обычно оказывается все же дешевле использовать административные данные, чем собирать такую же информацию путем обследований.

Там, где по-прежнему используются статистические наблюдения, требуется эффективная и правильно определенная основа выборки. Статистические регистры, используемые для создания таких основ выборки, часто являются столь большими и сложными, что становится очень трудно и дорого пополнять и поддерживать их удовлетворительным образом, используя наблюдения или данные переписи. Поэтому даже если административные данные не заменяют

[6]Доступно на web-сайте Евростата:http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-CC-04-002/EN/KS-CC-04-002-EN.PDF

статистические наблюдения, они все же могут использоваться для наполнения и поддержания статистических регистров, помогая таким образом снизить совокупные расходы.

2.3 Нагрузка на респондентов

Использование данных из административных источников помогает снизить нагрузку на поставщиков данных, порожденную запросами таковых. Таковы устойчивые политические представления во многих странах, особенно если респондентами являются бизнесы. Политика поддержки развития и роста бизнеса обычно включает снижение административного бремени. В этих условиях статистические обследования часто воспринимаются как объекты, легко подводимые под сокращение.

Со своей стороны, предприниматели относятся с пониманием к причинам предоставления данных для целей регистрации и налогообложения, даже если это не нравится им. Однако запрашивание статистических данных они обычно воспринимают как лишнюю, менее необходимую, нагрузку. Если они уже представили информацию другим правительственным учреждениям, то получение похожего запроса от национальной статистической организации может вызвать раздражение. Поэтому когда политические деятели и респонденты объединяются в требовании снижения статистической нагрузки на респондентов, статистическим организациям бывает чрезвычайно трудно противостоять этому давлению, так что повторное использование собранных другими данных является логичным решением проблемы.

2.4 Частота

Со снижением стоимости и нагрузки на респондентов также связано следующее преимущество использования административных источников, заключающееся в возможности в ряде случаев производить статистику с большей частотой, без дополнительной нагрузки на респондентов и с малыми дополнительными затратами. Такое имеет место в Финляндии, где существует возможность производить данные переписи населения на основе административных источников на ежегодной основе, тогда как использующие более традиционные методы страны могут позволить себе делать эти данные лишь каждые пять или десять лет.

Главным ограничением частоты статистики, производимой из административных источников, обычно является частота, с которой административные источники актуализируются. Так, было бы трудно производить ежемесячную статистику на основе административных данных, актуализируемых раз в год, если только эти данные не обновляются на скользящей основе без сезонных смещений (либо при наличии информации, достаточной для удаления сезонных смещений).

Значительную свободу действий предоставляют административные источники, которые не привязаны к какому-либо определенному периоду времени, типа тех, которые регистрируют события (например, рождения, смерти, предоставление разрешений на проведение работ). Это связано с тем, что коль скоро данные о событии регистрируются правильно, они позволяют производить статистику для любого заданного периода вплоть до ежедневного.

2.5 Охват

Административные источники обычно дают полное или почти полное покрытие их целевой совокупности, тогда как выборочные наблюдения обычно могут прямо покрыть лишь небольшую совокупность. Поэтому использование административных источников устраняет ошибки наблюдения, исключает (или значительно снижает) неответы и дает более точные и детализированные оценки для различных подсовокупностей, например, охватывающих респондентов на малых географических территориях или с другими специфическими характеристиками.

2.6 Своевременность

Использование административных источников может повысить своевременность статистических продуктов за счет обеспечения доступа к более близкой к текущему моменту информации, относящейся к определенным показателям. Это связано с тем, что для планирования статистических наблюдений, разработки и апробации форм, анализа совокупности, оптимизации выборки и т.д. обычно требуется время. Особенно это характерно для ежегодных и несистематических сборов данных. Поэтому обращение к подходящим административным источникам может стать более эффективным решением. Следует, однако, отметить, что возможны также случаи, когда использование административных источников приводит к снижению своевременности, особенно в случае краткосрочных индикаторов. Есть область, где административные источники могут иметь особенно благоприятное воздействие на своевременность – это ведение статистических регистров и основ выборки. Административная информация об изменениях целевой совокупности (например, рождения и смерти людей или бизнесов) обычно является намного более свежей, чем информация каких бы то ни было наблюдений – просто в силу упомянутых выше преимуществ в охвате.

2.7 Общественное мнение

Общественное мнение касательно обмена данными, в частности, между различными правительственными ведомствами значительно варьирует от страны к стране. Там где общественное мнение в целом принимает обмен данных или благосклонно к нему, возросшее использование существующих

источников данных может помочь повысить репутацию национального статистического института, делая его более эффективным и затрато-эффективным.

Хотя обмен данными часто вызывает понятное беспокойство широкой публики, существует также давление, направленное противоположно – на повышение эффективности правительства, особенно если его результатом являются более низкие налоги или большее финансирование актуальных для избирателей областей, таких как здравоохранение или образование. Политические лозунги типа “объединенного правительства” часто адресуются публике и могут помочь противостоять страхом утраты конфиденциальности. Таким образом, степень, в которой улучшение публичного имиджа может считаться плюсом использования административных источников, сильно зависит от того, как это использование представляется широкой публике и воспринимается ею.

3. Основания доступа к административным источникам

3.1 Введение

Наличие права доступа к данным из административных источников представляет собой один из ключевых барьеров к более широкому использованию таких данных для статистических целей. В настоящей главе с использованием примеров и опыта ряда стран охарактеризованы разного рода основания необходимые для облегчения доступа к административным источникам. Такие основания обычно имеют несколько измерений: правовое, политическое, организационное и техническое, каждое из которых рассматривается ниже. Необходимость достижения договоренностей во всех этих областях предшествует возможности реализации преимуществ использования административных данных.

3.2 Правовая база

Правовая база обычно создается на национальном уровне и является специфичной применительно к национальным источникам и обстоятельствам. В некоторых случаях может также иметься соответствующее законодательство на суб-национальном (например, субъект федерации) уровне или международном уровне. Примером последнего является статистическое законодательство Европейского Союза, которое обязательно для стран – членов. В таких случаях возможно существование двух или более альтернативных законных путей доступа к административным данным.

Большинство национальных статистических организаций имеют правовые документы, определяющие их роли и ответственность, обычно в форме законов о статистике. Во многих странах эти правовые документы включают конкретные нормы касательно доступа к административным данным. Примерами являются статистические законы Ирландии[7] и Норвегии[8].

Вставка 3.1 – Извлечения из ирландского Закона о статистике 1993

Статья 30. (1) С целью содействия [статистической] Службе в осуществлении ее функций согласно данному Закону Генеральный директор может путем направления уведомления запросить любой орган государственной власти:

(a) разрешить должностным лицам статистической службы в любое разумное время иметь доступ, просматривать и получать копии или делать выдержки из любых данных, находящихся в зоне его ответственности, и

(b) предоставлять Службе, если об этом попросит любой из этих должностных лиц, копии или извлечения из любой из таких записей, и орган

[7] www.irishstatutebook.ie/1993/en/act/pub/0021/index.html (See sections 30 and 31)

[8] www.ssb.no/english/about_ssb/statlaw/statlov_en.html (See chapter 3-2)

государственной власти обязан в соответствии с подстатьей (2) настоящей статьи удовлетворить любой такой запрос безвозмездно.

.....

Статья 31. (1) Генеральный директор может просить любой орган государственной власти о консультации и сотрудничестве с ним с целью оценивания потенциала данных органа власти в качестве источника статистической информации, а также развития (где это уместно и осуществимо) его методов регистрации и систем применительно к статистическим целям, и орган государственной власти должен удовлетворить любую такую просьбу постольку, поскольку позволяют ресурсы.

(2) Если какой-либо орган государственной власти предлагает внедрить, пересмотреть или расширить какую-либо систему хранения и извлечения информации или осуществлять статистическое наблюдение, он должен проконсультироваться с Генеральным директором и принять любые рекомендации, которые тот может обоснованно дать в связи с предложением.

Одни национальные правовые основы предоставляют больше полномочий для доступа к административным данным для статистических целей, чем другие. Это имеет место в силу того, что национальные исторические, политические и культурные особенности имеют сильное влияние на эти основы. Культурные факторы могут быть особенно важны, ибо некоторые культуры намного благосклоннее других к идее обмена данных между правительственными ведомствами и агентствами. В результате этих национальных различий правовые основы не являются особенно гармонизированными или даже согласующимися между странами.

Чтобы решить эти проблемы согласованности, Европейский Союз включил положения по доступу к административным данным в регламент 223/2009 по Европейской статистике, общеизвестный как "Статистический закон"[9]. Этот Регламент дает национальным статистическим организациям стран – членов право доступа к административным данным, необходимым для исполнения их обязанностей, в соответствии с европейским статистическим законодательством, но устанавливает, что такой доступ, тем не менее, подвержен национальным нормам и условиям.

Регулирование Европейского Союза в конкретных областях статистики продвинулось дальше и сняло эту зависимость от национальных ограничений и условий. Примером этого является регулирование бизнес-регистра, которое

[9]Статья 24 Регламента [regulation]ECNo 223/2009 Европейского парламента и Совета от 11 марта 2009г. по Европейской статистике: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:EN:PDF.

предоставляет неограниченный доступ к любым административным источникам, когда данные из этих источников необходимы для исполнения требований регулирования[10].

Наряду с предоставлением доступа к данным из административных источников, правовые основы также устанавливают пределы такого доступа и использования административных данных. Обычно ограничения состоят в том, что данные могут использоваться только для именно статистических целей, и что должна соблюдаться конфиденциальность индивидуальных данных.

Например, могут быть специальные ограничения на использование данных по некорпорированным бизнесам, особенно индивидуальным предпринимателям, где бизнес-данные могут рассматриваться как персональные данные, относящиеся к владельцу бизнеса. В таких случаях прибыль от хозяйственной деятельности может считаться равной персональному доходу. Многие страны имеют законодательство по защите данных, покрывающее информацию об отдельных гражданах, поэтому важно проводить четкое различие между тем, что является бизнес-данными и персональными данными в таких случаях.

Законодательный процесс может занимать много времени, а статистика часто воспринимается законодателями как сравнительно низкий приоритет, поэтому может потребоваться длительный период лоббирования и демонстрации преимуществ использования административных данных. Учитывая все усилия, необходимые для внедрения или изменения статистического законодательства, приходиться извлекать максимум пользы их представляющихся удобных случаев. В частности, очень важно избегать ошибок, связанных с внесением законодательных предложений, лишь отвечающих текущим потребностям. Может пройти десять или более лет до следующей благоприятной для изменения законодательства возможности, поэтому надо иметь долгосрочную стратегию использования административных данных, и заботиться о том, чтобы законодательные предложения отвечали всем нуждам, предусматриваемым на обозримое будущее. Таким образом, законодательство может рассматриваться как барьер в краткосрочном плане, но как благоприятная возможность – в долгосрочном.

Даже пока законодательство остается барьером, это не обязательно означает невозможность какого-либо использования административных данных. Один пример: в период ожидания правовой базы, необходимой для доступа к налоговым данным корпораций, сотрудник Национальной статистической службы Соединенного Королевства был направлен в налоговую службу для изучения возможности использования этих данных для статистических целей.

[10]Статья 4 Регламента [regulation]ECNo 177/2008 Европейского парламента и Совета от 20 February 2008, устанавливающая общие основания для бизнес-регистра для статистических целей: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:061:0006:0016:EN:PDF

Это лицо имело доступ к микро-данным во время прикомандирования и физически выполняло работу в помещении налоговой службы, но могло взять в статистическую службу лишь не приводящие к раскрытию этих данных агрегированные материалы. Такой подход предполагал, что различные касающиеся данных проблемы, включая адекватную оценку реальной ценности налоговых данных, могли прорабатываться при одновременном изучении возможных правовых режимов получения доступа [к данным].

Следует также отметить, что правовые ограничения обычно относятся к использованию микро-данных, то есть информации по отдельным людям и бизнесам. Хотя статистики обычно привыкли работать с данными этого уровня, иногда может оказаться целесообразным вместо них работать с агрегатами низкого уровня, не приводящими к раскрытию самих данных. В некоторых случаях это можно сделать просто путем использования в качестве статистической единицы не индивида, а малой группы имеющих определенные характеристики индивидов, возможно с весом, равным численности членов этой группы.

3.3 Политическая база

Во многих странах сложились общепризнанные политические принципы по совместному использованию данных внутри правительства, которые будут влиять на права доступа к административным данным для статистических целей. Однако обычно легче изменить политические компоненты, чем законодательство, кроме того политика склонна со временем изменяться. Поэтому важно, чтобы национальные статистические организации в полной мере участвовали в разработке политики, и играли активную роль в любых обсуждениях в правительстве, могущих привести к изменениям в политике. При этом любые изменения должны формулироваться так, чтобы предоставить максимум возможной пользы статистической системе.

Политическая база также включает рекомендуемые кодексы деятельности[codes of practice], важнейшими из которых для статистических целей в Организации Объединенных Наций являются Фундаментальные принципы официальной статистики [11] . Принцип 5 касается эффективности затрат и в этой связи рекомендует использовать данные из административных источников:

“Данные для статистических целей могут собираться из всех видов источников, будь то статистические обследования или административная отчетность. Статистические ведомства должны выбирать источник с учетом качества, своевременности, затрат и нагрузки, которая ложится на респондентов”.

В пояснительном примечании к принципу 5 также подчеркивается

[11]http://unstats.un.org/unsd/methods/statorg/FP-English.htm

эффективность затрат, и далее говорится:

“Статистические службы должны быть эффективными по затратам, наилучшим образом выбирая концепции, источники и методы путем достижения баланса качества, своевременности, затратности и отчетной нагрузки на респондентов.… На совокупную затрато-эффективность службы влияют организационное планирование и управление, обоснованное применение статистической методологии, использование информационной и коммуникационной технологии, а также доступ к административным источникам”.

Кодекс практики Европейской статистической системы[12] содержит аналогичные положения, однако рекомендация по использованию данных из административных источников дается в несколько отличающемся контексте. В принципе 2 говорится о полномочиях по сбору данных и указывается:

“Статистические органы должны иметь четко выраженные законные полномочия по сбору информации для целей европейской статистики. Административные органы, предприятия и домохозяйства, а также какой-либо круг лиц могут быть законом принуждены предоставить доступ к данным либо данные для целей европейской статистики по запросу статистических органов”.

Принцип 9 касается обеспечения того, чтобы нагрузка на респондентов статистических наблюдений не была избыточной. В нем отмечается:

“Нагрузка по составлению отчетности должна быть пропорциональна нуждам пользователей и не должна быть избыточной для респондентов. Статистические органы мониторируют нагрузку по составлению отчетности и определяют цели по ее постепенному снижению”.

Одним из индикаторов, рекомендуемых для оценивания применения этого принципа, является следующее:

“Административные источники используются всегда, когда возможно для избежания дублирования запросов на информацию”.

Кодексы практики могут существовать также на национальном уровне, что является весьма важным способом убеждения общественности в том, что данные будут использованы для ограниченных и необходимых целей. Для получения желаемого результата важно, чтобы эти кодексы практики были доступны для общественности, обычно посредством интернет-сайтов национальных статистических организаций.

[12]http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice

3.4 Организационная база

Коль скоро есть правовые и политические основания, дающие возможность использования административных данных, необходимо проработать организационные механизмы, обеспечивающие потоки информации. Обычно это осуществляется в форме письменных соглашений. Это может быть договор, особенно если вовлечены организации частного сектора, однако в случае соглашения между правительственными департаментами или агентствами оно по всей вероятности будет договоренностью об объеме и составе услуг, протоколом или соглашением. Различие состоит в том, что договоры обычно бывают юридически обязывающими, тогда как другие формы договоренностей – нет.

В любой такой договоренности присутствуют определенные ключевые элементы. Они следующие:

- Правовые основания: ссылка на законодательство, разрешающее доступ к административным источникам для статистических целей, и на любые законодательные акты, налагающие ограничения на такой доступ.
- Идентификация лиц, передающих / получающих данные: должны быть указаны имена, контактная информация и должности ключевых лиц, задействованных в передаче данных, как в административной, так и в статистической организациях. В некоторых случаях это может включать всех лиц в статистической организации, кто уполномочен использовать или просматривать данные.
- Детальное описание охватываемых данных: оно должно включать идентификацию массивов данных и содержащихся в них показателей.
- Частота предоставления данных: должно быть точно определено, когда и как часто административная организация будет поставлять данные.
- Стандарты качества: здесь определяются параметры качества поставляемых данных. Примерами могут служить требования по соответствию записей определенным стандартам, либо по максимально допустимой доле пропущенных или ошибочных значений показателей, что обеспечит соответствие получаемых данных целевому назначению. Степени приоритетности различных показателей и, следовательно, усилия по обеспечению качества зачастую различаются между административными и статистическими организациями, поэтому установление единых стандартов может быть затруднительно.
- Правила конфиденциальности: важно подробно указать, для чего данные могут быть использованы, какие правила и процедуры должны применяться для предотвращения раскрытия информации, при каких обстоятельствах данные могут быть переданы пользователям статистической организации.

- Технические стандарты: более детально они рассмотрены ниже в разделе о технической базе.
- Предоставление метаданных: важно, чтобы потоки данных сопровождались относящимся к ним метаданными, которые могут включать даты, описания всех используемых кодов, информацию по используемым единицам и т.д.
- Условия оплаты поставляемых данных: передача данных между правительственными ведомствами и агентствами обычно является безвозмездной, хотя в некоторых случаях от статистической организации может потребоваться внесение средств на покрытие затрат по извлечению и передаче данных. Данные организаций частного сектора обычно оцениваются по рыночной стоимости, хотя может оказаться возможным договориться о скидках, особенно когда внутри правительства есть несколько пользователей источника данных частного сектора. В некоторых случаях может иметься возможность предложить статистический анализ или консультационные услуги в качестве формы оплаты за полученные данные.
- Время договоренности: соглашения могут распространяться на установленный период времени, но при этом они должны включать условия возобновления или продления срока по мере надобности. Альтернативный подход – иметь соглашение, которое действует до тех пор, пока одна из сторон не захочет внести изменение.
- Случаи изменения обстоятельств: для статистических организаций важно иметь заблаговременное оповещение об изменениях, затрагивающих административные источники. Договоренность должна устанавливать, что о любых предполагаемых изменениях сообщается статистической организации в возможно короткие сроки, что позволит минимизировать воздействие изменений на выходные статистические данные.
- Процедура разрешения разногласий: договоренность должна определять метод разрешения любых разногласий между статистической и административной организацией, которая обычно состоит в передаче проблем по инстанции вышестоящим менеджерам, а может быть даже соответствующим министрам.

3.5 Техническая база

Техническая база – это механизмы, посредством которых передаются данные, а также относящиеся к этому стандарты данных и метаданных. Механизмы передачи данных могут принимать любую форму - от посылаемых по почте записей на бумаге до обновления данных в реальном времени по защищенным электронным каналам связи. Используемый механизм должен учитывать технические возможности, доступные и для отправляющей, и для получающей

организаций, поэтому обычно он является компромиссом, отражающим не вполне оптимальное решение, по крайней мере, для одной из этих организаций.

Существует ряд международных стандартов на передачу данных и метаданных, включая XML, SDMX и DDI, если назвать лишь некоторые. Некоторые страны также имеют национальные версии, особенно для передачи данных внутри правительства. Поэтому важно прийти к соглашению, какие стандарты надо использовать.

3.6 Резюме

Важно иметь правовую базу, дающую основания для использования административных данных для статистических целей. Все другие описанные выше основания не являются жизненно важными, однако они очень ценны для обеспечения беспрепятственного потока данных, а также минимизации проблематичности или отсутствия взаимопонимания между поставщиком данных и статистической организацией. По этой причине целесообразно, чтобы они были отражены в письменных документах, согласованных всеми сторонами.

Сравнение опыта по странам может быть полезным для выверки целей, однако следует помнить, что специфичные национальные ситуации и проблемы часто требуют специфичных решений. Международные стандарты могут оказать помощь в виде предоставления методологических рекомендаций (и соответствие им может рассматриваться как политическая цель), таким образом, на них следует ссылаться везде, где возможно в дискуссиях с административными департаментами.

Вставка 3.2 – Анализ опыта: Основания доступа в Соединенном Королевстве

• ***Правовая база***

Закон о статистике и службе регистрации 2007 г. обеспечивает основания для доступа к административным данным, но не дает, как во многих других странах, безоговорочного права доступа. Условия доступа и использования административных данных обычно определяются специальным для источника законодательством, как, например, Закон о налоге на добавленную стоимость 1994 г. Доступ к новым административным источникам является предметом одобрения парламентом. Как член Европейского Союза, Соединенное Королевство также подчиняется положениям европейского законодательства, касающегося использования административных источников.

•***Политическая база***

В дополнение к применению Фундаментальных принципов официальной статистики Организации Объединенных Наций и Кодексу практики Европейской статистической системы имеется национальный кодекс практики[13] для работников статистической службы Правительства. Ключевыми положениями, относящимися к использованию административных данных для статистических целей, являются:

- "5(f) К данным, извлеченным из административных источников, должны применяться те же самые стандарты конфиденциальности, которые применяются к данным, собранным специально для статистических целей".
- "7(c) Следует признать ценность административных данных для производства национальной статистики, и значение статистических целей при проектировании административных систем должно повышаться".
- "7(d) Статистические системы будут разрабатываться так, чтобы максимизировать потенциал увеличения их ценности за счет интеграции данных".

Кодекс практики поддерживается различными протоколами, включая Протокол по регулированию нагрузки на респондентов[14], который содержит следующие положения:

- "2. Новые статистические обследования не будут дублировать существующие источники… Производители национальной статистики должны рассмотреть возможность использования данных имеющихся обследований, административных данных и других, не являющих обследованиями источников, прежде чем вводить новое обследование…

[13]http://www.statistics.gov.uk/about/national_statistics/cop/default.asp
[14]http://www.statistics.gov.uk/about/national_statistics/cop/downloads/respondentload.pdf

Обследование будет проводиться только в тех случаях, когда отсутствует подходящий альтернативный источник данных".

- "4. Следует признать ценность административных данных для производства национальной статистики, и значение статистических целей при проектировании административных систем должно повышаться. Национальная статистика будет, где это целесообразно, извлекаться из информации, представляемой для управления деятельностью правительства и государственными услугами. Это будет достигаться везде, где это возможно, путем прямого извлечения соответствующих данных из систем, поддерживающих управленческий аппарат. Производители национальной статистики будут стремиться оказывать влияние на тех, кто отвечает за разработку административных систем, так чтобы эти системы смогли собирать также данные для статистических целей экономичным способом".

• Организационная база

Организационная база передачи данных между ведомствами и агентствами правительства развивается в направлении ее инкорпорирования в "соглашения об обслуживании". Таковые подписываются на уровне руководства, но не являются юридически обязывающими. В основной части соглашения содержатся общие положения, а в приложениях дается детализация требований по конкретным данным и технические характеристики. Оплата обычно отсутствует, но в некоторых случаях в обмен предоставляются статистические анализ или инструменты.

Ведомства и агентства правительства, которые предоставляют административные данные для статистического бизнес-регистра, представлены в управляющем комитете этого регистра, который включает также пользователей. Это помогает им лучше понять, как используются их данные, а также значение качества данных. Агентство по регистрации компаний (Дом компаний) работает обычно на коммерческой основе, поэтому основания для передачи данных из этого агентства имеют форму договора, предусматривающего оплату. Данные по правам собственности на бизнес и связям по контролирующему участию также покупаются от коммерческих поставщиков данных частного сектора.

• Техническая база

Большинство данных передается посредством текстовых файлов с полями фиксированной длины или со стандартными разделителями. Формат адресов бизнесов является сферой, в которой стала возможна определенная стандартизация. Она стала проще в силу наличия программного обеспечения по привязке адресов, основанного на стандартах Почтовой службы.

Большинство данных в настоящее время передается посредством посылаемых почтой дисков, либо (для меньших наборов данных) – посредством защищенных каналов электронной почты. Однако, по данным о налоге на добавленную стоимость, используемым в статистическом бизнес-регистре, система ежедневной актуализации организована с использованием файлов сообщений, посылаемых по защищенному интранету правительства.

Метаданные обычно передаются в виде таблиц справочной информации, которые либо сопровождают данные, либо передаются отдельно с меньшей частотой. Метаданные, определяющие коды, хранятся в виде просмотровых таблиц соответствия, тогда как более общая информация записывается в базу данных стандартов и руководств.

4. Типичные проблемы и решения

4.1. Введение

Хотя в главе 2 приведены многие убедительные доводы в пользу использования административных источников, существует также ряд проблем, связанных с их использованием. Некоторые из этих проблем специфичны для какого-то конкретного источника или применения, но многие из них по своей природе являются более общими. Данная глава очерчивает некоторые из более общих проблем и предлагает методы их решения или, по крайней мере, минимизации их воздействия. Специфические проблемы получения доступа к административным источникам и связывания данных рассматриваются отдельно в главах 3 и 6 соответственно.

4.2 Общественное мнение

В главе 2 рассматривалось, как общественное мнение может благоприятствовать обмену данными в некоторых странах. Однако в других странах может существовать беспокойство широкой публики в связи с намерением обмениваться данными внутри правительства. Такую озабоченность очень трудно уменьшить, и возможные подходы могут включать публикацию четких границ и правил использования данных и принятие мер к тому, чтобы люди и бизнесы понимали, что уязвимые данные, используемые или собираемые для статистических целей, не будут передаваться другим органам правительства (в частности, налоговой службе и агентству по выплатам).

Это соответствует Фундаментальным принципам официальной статистики Организации Объединенных Наций, где принцип 5 (“Данные для статистических целей могут собираться из всех видов источников, будь то статистические обследования или административная отчетность. Статистические ведомства должны выбирать источник с учетом качества, своевременности, затрат и нагрузки, которая ложится на респондентов”) поддерживает использование административных данных. Вместе с принципом 6 (“Личные данные, собираемые статистическими ведомствами для подготовки статистической информации, независимо от того, относятся ли они к физическим или юридическим лицам, должны носить строго конфиденциальный характер и использоваться исключительно для статистических целей”) это формирует принцип одностороннего потока данных.

Другие способы, помогающие преодолеть неблагоприятное общественное мнение, включают публикацию анализа затрат и выгод для правительства и респондентов от использования различных источников. Кроме того, возможно можно утверждать, что микро-данные являются более защищенными, когда

используются административные источники. Нет посылаемых по почте вопросников, данные не хранятся интервьюерами в бумажном или электронном виде, меньший канцелярский персонал нужен для процесса статистического производства, следовательно, меньше людей имеет доступ к уязвимым данным.

4.3 Общественный статус

Прямой контакт с публикой при обследованиях помогает повысить престиж статистической организации. Использование административных данных может уменьшить этот контакт и поэтому также уменьшить осведомленность публики о работе статистической организации. Если это становится проблемой, то наиболее очевидным решением является улучшение “маркетинга” статистической организации и выпускаемых ею данных. Для этого может потребоваться, чтобы небольшая доля экономии за счет использования административных источников направлялась в маркетинговый бюджет.

Возможно наиболее эффективным путем промотирования деятельности и продукции национальной статистической организации, особенно в средне- и долгосрочной перспективе, является обеспечение большего сотрудничества с образовательными институтами, деловыми группами и другими целевыми потребителями. В этом отношении также очень важны объединения пользователей, они должны активно поддерживаться.

4.4 Менеджмент изменений

Административные источники государственного сектора обычно создаются для целей сбора налогов или мониторинга политики правительства. Это означает, что они чувствительны к политическим изменениям. Когда политика изменяется, административные источники могут затрагиваться в части покрытия, определений, порогов и т.п., а возможно могут даже полностью упраздняться. Изменения в компьютерных системах, используемых для хранения и обработки административных данных, могут также оказывать воздействие на поставку данных для статистических целей. Даже источники частного сектора не являются нейтральными к изменениям такого рода, хотя в этом случае изменения по всей вероятности будут возникать в силу изменений рыночных факторов.

Такие изменения могут происходить внезапно, при минимальной заблаговременности предупреждения, времена особенно высоких рисков наступают, как правило, сразу после смены правительства, смены министра, или изменения в законодательстве. О подобном случае несколько лет тому назад сообщалось из Словении, где предоставление административных данных по занятости было на некоторое время прекращено вслед за сменой министра, который оставил статистическое ведомство с серьезными проблемами в

производстве статистики занятости. Поэтому в практику были внедрены поддерживаемые законодательством процедуры, нацеленные на минимизацию вероятности и влияние изменений этого рода.

Поэтому зависимость от какого-то определенного источника всегда будет сопряжена с некоторым уровнем риска. В какой-то степени эти риски могут быть управляемыми с помощью установленных законом или договорами положений. На практике лучший путь избежать подобные проблемы обычно заключается в регулярных контактах с теми, кто отвечает за административные источники, чтобы следить за их осведомленностью в потребностях статистики, пытаться влиять на них, а также получать раннее предупреждение о любых возможных изменениях. Когда имеется сильная зависимость от какого-то определенного источника, всегда стоит подготовить план действий в непредвиденных обстоятельствах, предусматривающий, что можно сделать, если этот источник станет недоступным. Однозначно лучше работать с заблаговременным упреждением событий, чем реагировать после их наступления.

4.5 Единицы

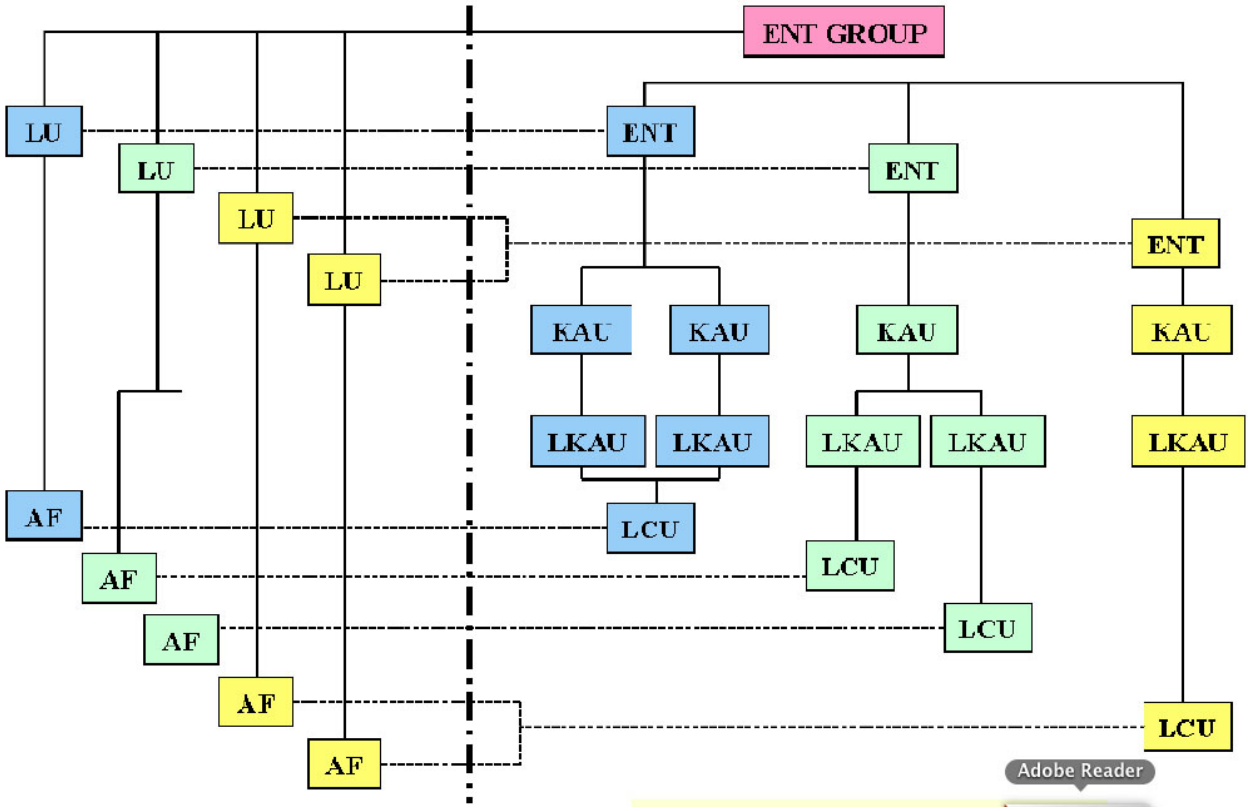
Одна существенная проблема, часто встречающаяся при использовании административных источников, состоит в том, что используемые в этих источниках единицы прямо не соответствуют определениям требуемых статистических единиц. Процесс конвертирования из административных единиц (юридические лица, объекты налогообложения, заявители и пр.) в статистические единицы (предприятия, люди, домохозяйства и пр.) может быть довольно сложным концептуально и часто включает какую-то форму моделирования. В бизнес-статистике этот процесс известен как профилирование, он обычно является функцией статистических бизнес-регистров. Евростат опубликовал руководство по этому процессу в главе 19 Руководства с рекомендациями по бизнес-регистру[15], где профилирование определяется как “метод анализа юридической, организационной и учетной структуры группы предприятий на национальном и мировом уровнях с целью установления статистических единиц внутри этой группы, их связей и наиболее эффективных структур для сбора статистических данных”.

Рисунок 4.1 показывает, что структурированный набор взаимосвязанных бизнес-единиц может выглядеть совершенно по-разному с правовой / административной точки зрения в сравнении со статистической точкой зрения. Профилирование, как оно определено выше, может рассматриваться как процесс создания статистической структуры и установления ее соответствия с

[15]Доступно на web-сайте Евростата: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-BG-03-001/EN/KS-BG-03-001-EN.PDF

правовой / административной структурой.

Рисунок 4.1 Различные точки зрения на группу бизнес-единиц



Правовая / администра-тивная сфера	**Экономическая / статистическая сфера**
LU=юридическое лицо AF=филиал	ENT=предприятие KAU=единица вида деятельности LCU=местная единица LKAU=местная единица вида деятельности

Хотя профилирование дает лучшее понимание сложных структур единиц, оно является дорогим и время-затратным, и требует обученного персонала. Поэтому практически абсолютно нецелесообразно осуществлять детальное профилирование для всех бизнес-единиц в экономике, необходимо фокусироваться на тех случаях, которые приносят наибольшую пользу. Профилирование можно рассматривать как компромиссный выбор между тремя факторами:

- количество профилируемых бизнес-структур;
- качество или глубина профилирования и
- располагаемые ресурсы (определяемые затратами и адекватностью персонала).

Во вставке 4.1 даны четыре примера бизнес-структур, которые были раздельно

профилированы в трех различных странах (Дания, Нидерланды и Соединенное Королевство) в рамках изучения совместимости применения статистического определения предприятия внутри Европейского Союза [16] . Это хорошо показывает, что профилирование в некоторой степени является искусством, и "верный" ответ имеется не всегда, однако эта специфичная задача вылилась в значительную методологическую работу по гармонизации правил профилирования, что частично изложено в главе 19 названного выше Руководства с рекомендациями по бизнес-регистру.

Хотя традиционное профилирование не является практически применяемым для всех единиц большой совокупности, некоторое автоматизированное на основе правил профилирование может таковым быть. Стандартные правила, основанные на признаках или характере связи между единицами, могут помочь преодолеть различия между административными и статистическими единицами во многих областях статистики. Например, статистические домохозяйства могут быть получены на основе отношений между индивидами, проживающими в строении. Этот подход успешно используется в методологии переписи населения на основе регистра, применяемой в странах Северной Европы.

Альтернатива профилированию, которая может быть возможна в некоторых случаях, состоит во внесении поправок на разницы в определениях единиц путем статистических "корректировок". Грубым примером этого подхода может быть случай, когда статистической единицей являются люди, а административной единицей – работы. Принимая во внимание, что по данным обследования известно, что работающие люди имеют в среднем 1,15 работ, этот корректировочный коэффициент может использоваться для оценки числа занятых лиц на основе числа работ.

Вставка 4.1 – Упражнение по профилированию: Сколько предприятий?

Следующие примеры взяты из исследования "Влияние расхождений в интерпретации концепции предприятия", подготовленного для Евростата Статистикой Нидерландов при участии Дании и Соединенного Королевства. За каждым примером следует ответ, даваемый каждой из трех участвующих стран, а также суть их аргументации. Примеры базируются на следующем определении предприятия:

> "наименьшая комбинация юридических лиц, являющаяся организационной единицей, производящей товары и услуги, которая с выгодой для себя пользуется определенной степени автономией в принятии решений....

[16] "Влияние расхождений в интерпретации концепции предприятия" – исследование подготовлено для Евростата Статистикой Нидерландов при участии Дании и Соединенного Королевства.

Предприятие может быть одним юридическим лицом”.

Источник: Регламент [Regulation] ЕС по статистическим единицам.

Пример 1 – Два юридических лица в группе предприятий имеют различные 4-значные коды NACE; оба осуществляют продажи преимущественно третьим лицам вовне группы. Они имеют совместно используемые здания, администрацию, покупаемые ресурсы и работников.

Ответы

- Нидерланды и Соединенное Королевство: объединяются в одно предприятие, принимая во внимание интенсивность совместного использования факторов производства.
- Дания: два отдельных предприятия, поскольку продажи обоих вовне группы составляют более 50% .

Пример 2 – Четыре юридических лица: A и B имеют различные виды деятельности, нет общих покупок, но совместно используются здания. C и D совместно используют здания, персонал и покупаемые ресурсы. Все четыре позиционируют себя в качестве одной фирмы.

Ответы

- Нидерланды и Дания: A и B являются отдельными предприятиями, C и D объединяются в одно предприятие, ибо A и B действуют по рыночным условиям, тогда как C и D совместно используют факторы производства.
- Соединенное Королевство: все четыре – в одном предприятии, поскольку они позиционируют себя как одна фирма.

Пример 3 – Три юридических лица: все производят преимущественно для внешних потребителей, они имеют совместно используемые администрацию и покупаемые ресурсы и позиционируют себя как одна фирма. A и B совместно используют здание. B и C имеют одинаковую деятельность, совместно используют персонал и капитальные товары, не могут предоставлять раздельные данные.

Ответы

- Нидерланды: объединяются в одно предприятие, ибо все имеют совместно используемые администрацию и покупки и позиционируют себя как одна фирма.
- Соединенное Королевство и Дания: B и C объединяются в одно предприятие, поскольку они являются горизонтально интегрированными, и данные доступны только для обоих вместе. A является отдельным предприятием.

Пример 4 – Двенадцать юридических лиц образуют группу предприятий.

Только одно действительно действует, другие не имеют работников.

Ответы

- Нидерланды: одно предприятие, которое состоит только из действующей единицы, поскольку неактивные единицы не являются частью предприятия.
- Соединенное Королевство: одно предприятие, которое включает все единицы, поскольку нет смысла иметь отдельные предприятия по неактивным единицам.
- Дания: каждая единица является отдельным предприятием, ибо между единицами нет крепких связей.

4.6 Определения показателей

Наряду с различиями в определениях единиц, между административными и статистическими системами, очевидно, существуют различия в определениях показателей. Данные административных источников обычно собраны для специфически административных целей, а связанные с этими целями нужды и приоритеты, скорее всего, отличаются от таковых в статистической системе. Например, оборот для целей налогообложения добавленной стоимости (VAT) может не включать оборот, относящийся к продажам товаров и услуг, освобожденных от VAT, тогда как статистической системе, как правило, требуется общий оборот.

Другой типичный пример – определение безработицы. Стандартным статистическим определением является[17]:

> *“"Безработные" включают всех лиц старше определенного возраста, которые в течение отчетного периода были:*
> *(a) "без работы", то есть не находились в состоянии оплачиваемой занятости или самозанятости,*
> *(b) "готовы к работе", то есть готовы находиться в состоянии оплачиваемой занятости или самозанятости в течение отчетного периода, и*
> *(c) "ищущими работу", то есть предпринимали конкретные шаги в определенном предшествующем периоде по поиску оплачиваемой занятости или самозанятости”.*

Однако определения безработицы в административных источниках чаще

[17]См. Резолюцию о статистических данных в отношении экономически активного населения, занятости, безработицы и неполной занятости, принятую 13-й Международной конференцией статистиков по труду (октябрь 1982 г.): http://www.ilo.org/global/statistics-and-databases/standards-andguidelines/resolutions-adopted-by-international-conferences-of-labourstatisticians/WCMS_087481/lang--en/index.htm

базируется на численности людей, претендующих на пособия по безработице, либо зарегистрированных как ищущие работу. Некоторые люди, не имеющие работу, могут не регистрироваться как безработные, если они рассчитывают найти работу быстро, кроме того, в некоторых культурах претендующим на пособие по безработице могут навешиваться негативные стигматизирующие ярлыки. С другой стороны, некоторые люди, претендующие на пособия по безработице, могут быть не в состоянии работать или активно искать работу, поэтому в статистике они не должны учитываться как безработные.

Первый шаг на пути решения проблемы различий в определениях заключается в том, чтобы попытаться понять различия и количественно определить их влияние. Некоторые различия на практике могут не иметь реального влияния, поэтому они смело могут игнорироваться, другие могут быть систематическими, поэтому они могут быть устранены путем корректировок данных. Иногда может оказаться возможным установить или оценить влияние различия путем комбинирования показателей из различных источников, особенно применительно к показателям финансовой отчетности, как, например, упомянутый выше оборот. В некоторых случаях может даже оказаться возможным повлиять на административные определения.

4.7 Системы классификации

Как и показатели, системы классификации, используемые внутри административных источников, могут отличаться от таковых, используемых в статистической сфере. Но даже если они те же самые, они могут применяться по-другому в зависимости от основного назначения административного источника, вполне возможно, фокусируясь на специфических признаках единицы. Например, административный источник, имеющий отношение к лицензированию, охране труда и технике безопасности, либо к защите окружающей среды, может больше интересоваться тем видом экономической деятельности предприятия, который связан с задачами этого источника, а не основным видом деятельности предприятия, который нужен для статистических целей.

В иных случаях классификации внутри административных источников могут не применяться на требуемом для статистических целей уровне детализации, либо классификация может попросту не быть приоритетом для административного источника, способным стать причиной ненадлежащего качества.

Когда классификационные системы или версии являются различными, решение обычно заключается в построении матриц перехода, позволяющих поставить коды административной классификации в соответствие кодам статистической классификации. Такое соответствие может быть один к одному, несколько к одному, один к нескольким, либо несколько к нескольким. В последних двух

случаях может потребоваться какой-либо способ вероятностного распределения по кодам.

Вставка 4.2 – Использование простой матрицы перехода

Код 1	Код 2	Вес	
0100	01300	100	Соответствие 1 к 1
0101	01210	26	Соответствие 1 к нескольким
0101	01221	14	
0101	01222	29	
0101	25730	11	
0101	74332	20	
0102	03200	100	
0103	01300	36	
0103	74332	64	

Это извлечение из матрицы перехода иллюстрирует основные проблемы, имеющие место при конвертации от одной классификационной системы к другой. В данном случае используемые в административном источнике коды (Код 1) поставлены в соответствие таковым, используемым в статистической системе (Код 2), с применением вероятностных весов.

Поэтому первая проблема состоит в том, как определить веса. Можно сделать оценку, однако предпочтительный метод состоит, где это возможно, в том, чтобы получить их в результате анализа единиц, проклассифицированных по обеим системам, опираясь на пропорции между значениями, имеющими определенное сочетание кодов. Быть может, надо ограничить этот анализ лишь рассмотрением сочетаний кодов, которые признаются весомыми, либо оправданными с точки зрения снижения влияния ошибок кодирования.

Выше в первой строке показано соответствие один к одному, данное с весом 100%. Это значит, что все единицы с административным кодом 0100 должны быть отнесены к статистическому коду 01300. Следующие 5 строк показаны с соответствием один к нескольким. Если единица имеет административный код 0101, то имеется 5 возможных статистических кодов. Для каждого из этих статистических кодов вероятность того, что он является верным кодом единицы, отражается весом, таким образом, есть 26-процентная вероятность,

что 01210 является верным статистическим кодом.

В этом случае вероятность того, что единице с административным кодом 0101 будет дан верный статистический код, может быть вычислена путем суммирования квадратов вероятностей каждой комбинации, в примере:

$0.26^2 + 0.14^2 + 0.29^2 + 0.11^2 + 0.2^2 = 0.2234$

Это значит, что есть 77,66% вероятности, что единице с административным кодом 0101 будет дан неверный статистический код. Хотя эта вероятность может казаться неприемлемо высокой, следует помнить, что даже если коды могут быть неверными на уровне единиц, то при условии правильности весов распределение единиц по кодам должно быть верным на агрегированном уровне, и коль скоро нет систематических ошибок в применении матрицы перехода, то не должно быть получаемых в результате смещений в статистических данных по кодируемым таким образом единицам.

Следует также отметить, что такие, как в вышеприведенном примере, матрицы перехода являются однонаправленными. Для преобразования из статистических кодов в административные коды потребуется другая матрица, с другими весами. Например, соответствие один к одному в одном направлении, в противоположном направлении может оказаться соответствием один к нескольким. Это проиллюстрировано выше в таблице, где между кодами 0100 и 01300 имеется соответствие один к одному, но при преобразовании из статистических кодов в административные коду 01300 может соответствовать 0100 или 0103.

Когда точность требуется на уровне микро-данных, метод с использованием матриц перехода имеет серьезные ограничения – это ясно из вставки 4.2. В зависимости от ресурсов и доступности данных могут оказаться возможными различные другие методы, но целесообразный первый шаг всегда состоит в достижении детального понимания, как классификационные данные образуются и обрабатываются административным источником и каков характер административных функций, для которых они используются.

В ряде случаев внутри административного источника могут быть другие показатели, которые могут использоваться для повышения вероятности выбора верного статистического кода. Одним из них может быть текстовое описание, из которого вытекает административный код. Его наличие потенциально более полезно для статистика, чем сам административный код, поскольку статистик может использовать справочник или автоматическую процедуру для получения верного статистического кода напрямую из описания. Этот метод может использоваться совместно с применением матриц перехода, так что текстовое описание подлежит кодированию только в случаях, когда нет соответствия один

к одному между административным и статистическим кодами, хотя есть риск потенциального смещения, когда качество кодирования различается между административной и статистической системами.

Подход, который успешно применен в некоторых странах, состоит в разработке автоматического инструмента кодирования для использования как в статистической, так и в административной системах. Это обеспечивает высокую степень единообразия кодирования и сильно побуждает (но не обязательно принуждает) к использованию общепринятых систем классификации.

Наряду с использованием общепринятых инструментов кодирования, повышению качества кодирования может помочь предоставление поставщикам административных данных практических знаний и обучения. При этом статистикам всегда полезно подчеркивать преимущества использования общепринятых систем классификации. Также помогает заблаговременное извещение о любых пересмотрах системы классификации и предоставление поставщикам административных данных максимально возможной помощи при внедрении внесенных изменений.

4.8 Своевременность

Есть три касающиеся своевременности самостоятельные проблемы, влияющие на практическую ценность административных данных для статистических целей:

- административные данные могут не быть доступными для статистических нужд вовремя,
- административные данные могут относиться к периоду, который не совпадает со статистическим отчетным периодом,
- административные данные могут измеряться за период, тогда как по статистическим требованиям должны быть на определенный момент времени (или наоборот).

Касательно первой проблемы: обычно бывает некий лаг между событием в реальном мире и его регистрацией административным источником, затем следующий лаг имеет место прежде, чем данные делаются доступными национальной статистической организации. Нижеприведенный рисунок 4.2 отображает общий лаг в днях между началом деятельности бизнесов и регистрацией в статистическом бизнес-регистре в Соединенном Королевстве. Лаги, относящиеся к рождениям и смертям предприятий, являются главной причиной ошибок покрытия в бизнес-регистре. Если эти лаги измеряются, на них может быть сделана поправка в любой статистике, основанной на данных регистра.

Анализируя лаги таким образом, можно получить сводную статистику для оценки их влияния. Например, в вышеприведенном случае две трети бизнесов появляются в статистическом бизнес-регистре в пределах 2 месяцев от начала деятельности. Средний лаг – около 120 дней, но этот показатель не особенно удобен, поскольку на него оказывают влияние выбросы, находящиеся в довольно длинном хвосте распределения (усеченном на рисунке 4.2, поскольку наиболее предельные случаи имеют доходящие до десяти лет лаги). Видимо, более полезным измерителем среднего в этом случае является медиана, которая составляет около 40 дней. Другой интересной характеристикой этого анализа является малое количество отрицательных лагов, которые имеют место, когда бизнесы завершают регистрационные формальности значительно раньше начала деятельности.

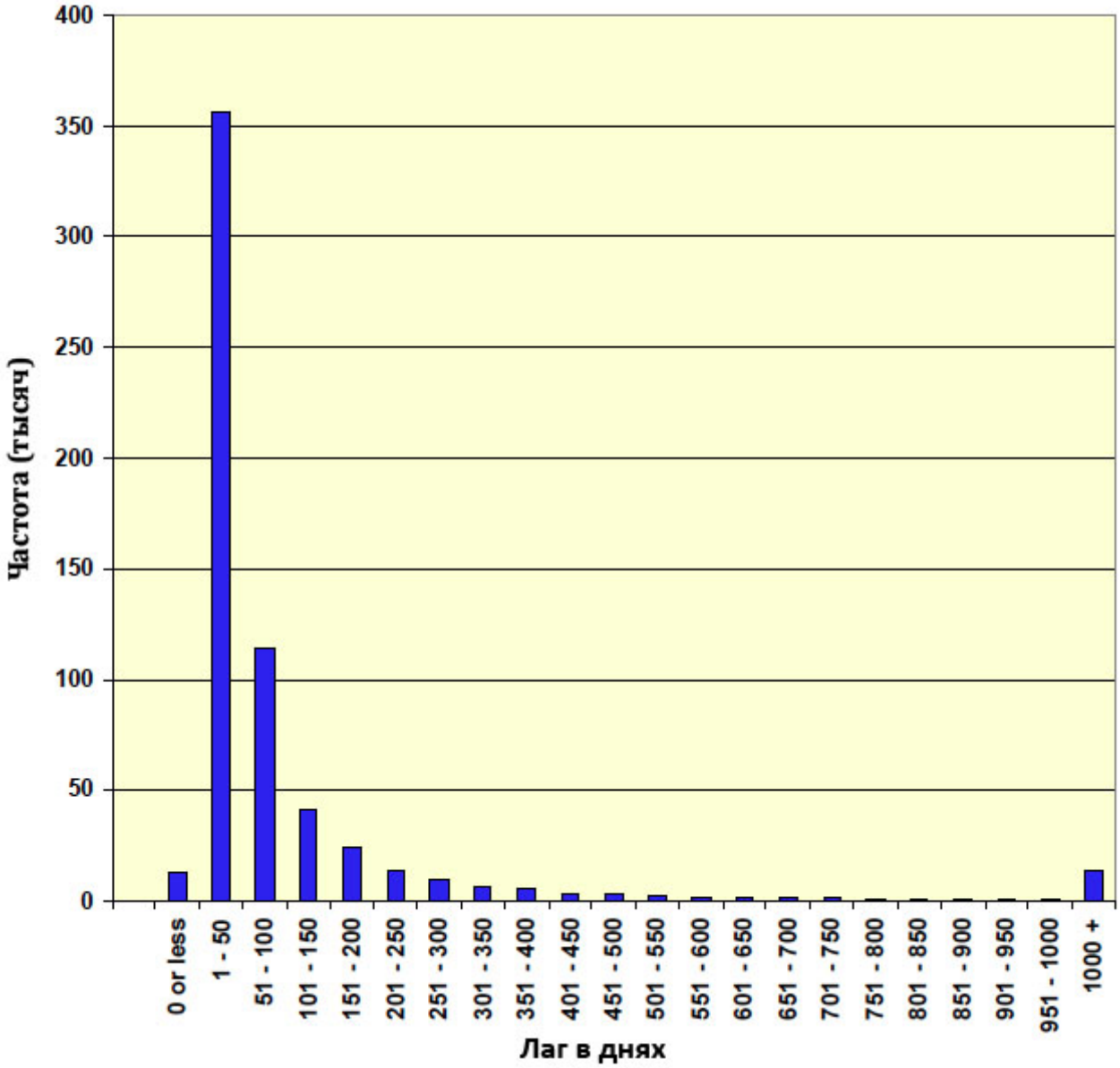
Этот короткий анализ, несомненно, важен, чтобы помочь статистикам понять характер и последствия лагов в источниках, используемых для производства статистики. Это также дает информацию, которая может использоваться для формирования корректировок с целью улучшения качества выходной статистики.

Наличие и длина лагов может сделать затруднительным использование административных источников для краткосрочной статистики, например шестимесячный лаг, видимо, неприемлем для ключевых рядов ежемесячных экономических данных, но не является большой проблемой для годовой статистики.

Первый шаг в решении проблемы лагов – понять их влияние, проведя анализ, подобный вышеизложенному. Коль скоро это сделано, может оказаться возможной разработка моделей, вносящих поправку на их воздействие [18]. Применительно к некоторым сравнительно устойчивым рядам данных возможен также случай, когда противоположные лаги уничтожают друг друга, например, для целей формировании данных по составу бизнес-единиц лаги регистрации бизнесов на рисунке 4.2 вероятно скомпенсируются лагами дерегистрации. Однако предполагать, что это так, рискованно без эмпирических свидетельств.

[18]Пример, соответствующий рисунку 4.2, см. в Приложении Впо вопросу возникновения и закрытия бизнесов: регистрация и дерегистрация по НДС в 2005 г. – Руководство и методология. http://stats.berr.gov.uk/smes/vat/VATGuidance2005.pdf

Рисунок 4.2 Лаги регистрации бизнеса в Соединенном Королевстве[19]



Когда характер и последствия лагов определены, полезно попытаться понять, что является их причиной. В некоторых случаях может оказаться возможным предложить внести изменения в административный источник, которые уменьшили бы лаг. Это может быть целесообразно как для статистика, так и для административного источника.

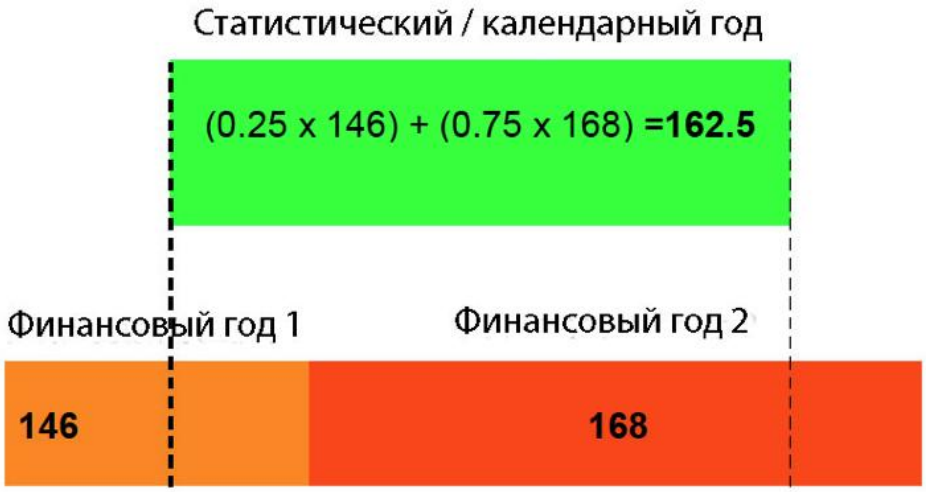
Вторая проблема, касающаяся своевременности, – различие в периодах, например, данные из годовых налоговых деклараций обычно бывают доступны лишь спустя несколько месяцев после окончания налогового года, так что они

[19]Источник: Модельный доклад о качестве в бизнес-статистике, Том III, Евростат http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/MODEL%20QUALITY%20REPORT%20VOL%203.pdf

наверняка не пригодны для месячной или квартальной статистики. Однако в некоторых случаях годовые административные данные могут использоваться для краткосрочной статистики, особенно если они собираются на скользящей годовой основе. Это может быть, если существует требование рассредоточить по всему году нагрузку по сбору и обработке этих данных административным источником. Если распределение единиц, для которых собираются данные в течение года, является в достаточной степени случайным, оказывается возможным извлекать представительные месячные или квартальные статистические трендовые данные из таких источников.

На рисунке 4.3 показан случай, когда административные данные базируются на финансовом годе, длящемся с 1 апреля по 31 марта, тогда как в статистике требуются данные за календарный год. Простейший способ преобразования этих данных – добавить 25% значения показателя в первом финансовом году к 75% значения во втором году. Этот метод должен дать приемлемую аппроксимацию, если долгосрочный тренд данных достаточно стабилен, хотя для более волатильных рядов могут потребоваться другие, более сложные методы оценки.

Рисунок 4.3 Работа с различающимися временными периодами



Третья проблема касается разницы между данными, относящимися к определенному моменту времени, и данными, относящимися к периоду (например, годовое или месячное среднее). Например, в статистике данные по занятости могут требоваться на определенную отчетную дату, тогда как административные данные могут продуцировать лишь месячные средние.

Как в предыдущем примере, первый шаг состоит в том, чтобы проанализировать влияние различия и определить, настолько ли оно значимо, чтобы потребовались дальнейшие действия. Одно возможное решение – математическая корректировка на основе модели, например, если

статистическая отчетная дата находится вблизи начала месяца, то может оказаться приемлемой модель, которая принимает в расчет среднее значение за предыдущий период. Альтернативный подход может состоять в использовании результатов относительно малого обследования для корректировки административных данных.

4.9 Несовместимость между источниками

Проблема несовместимости между источниками характерна для использования нескольких источников. Данные из одного источника могут противоречить таковым из другого источника. Это может быть в силу различающихся определений или классификаций, различий в сроках или просто ошибок в одном источнике. Это может выявиться при сравнении административных данных со статистическими данными, либо при сравнении двух административных (или двух статистических) источников.

Для разрешения противоречий такого рода необходимо установить правила выбора приоритетов, принимая решение, какой источник более надежен применительно к данному показателю. Коль скоро определена степень приоритетности источников применительно к показателю, становится возможным принять меры к тому, чтобы данные из высокоприоритетного источника не замещались источником с более низким приоритетом. Этот процесс становится намного легче, если коды источников хранятся вместе с показателями, для которых имеется несколько источников. Использование и хранение дат может также быть полезным, поскольку даже в том случае, когда один источник считается надежнее другого, данные десятилетней давности из этого источника не могут быть выше качеством, чем данные наиболее недавнего периода из менее надежного источника. Простым методом, подходящим в ряде случаев, является загрузка данных в порядке, обратном приоритетности, что позволяет данным более высокого качества замещать таковые, имеющие более низкое качество.

В большинстве случаев предметом интереса являются несколько показателей, и возможно, что приоритеты будут изменяться от показателя к показателю. Например, административные источники, касающиеся занятости работников, скорее всего, дают адекватные оценки (легальной) занятости, поскольку этот показатель тесно связан с основной функцией источника. Однако он не будет столь хорош для определения вида экономической деятельности работодателя, поскольку это имеет второстепенную важность для целей источника. Таким образом, если несколько источников используются для получения данных о занятости, необходимо проанализировать сравнительное качество каждого показателя в каждом источнике с целью извлечения оптимального для статистики набора данных.

Чем больше источников данных используется, тем более сложным становится процесс такого сравнения, однако наличие нескольких источников обычно помогает обеспечивать качество данных. В ряде случаев определенные источники могут использоваться не непосредственно в статистическом производстве, а лишь для целей бенчмаркинга как часть процесса обеспечения качества[20]. Получаемые в результате сведения о качестве различных источников могут возвращаться в источник (обычно лучше в агрегированном виде, чем пообъектно, с целью защиты статистической конфиденциальности) и служить материалом для рассмотрения вопросов улучшения качества этого источника.

Вставка 4.3 – Данные из различных источников

	Источник 1: Регистр образования	**Источник 2: Регистр населения**
Имя	Steve Vale	Stephen Vale
Адрес 1	5 St Peter's St	5 Saint Peters Street
Адрес 2	Machen	Machen
Адрес 3	Newport	Caerphilly
Адрес 4	Gwent	South Wales
Почтовый код	NP1 8QB	CF83 8QB
Дата рождения	28/12/1967	28/12/1997
Занятие	Статистик	Государственный служащий
Работодатель	CSO	Национальное статистическое управление
Почтовый код места работы	NP10 9XX	NP10 8XG

Этот пример отражает две записи, содержащие вымышленные данные об авторе (регистры населения и образования пока не существуют в Соединенном Королевстве). Он предназначен для иллюстрации нескольких общих проблем при попытках увязать данные из различных источников:

- Ошибки – простая проверка правдоподобия выявляет ошибку в регистре

[20]Пример бенчмаркинга, использующего карты при сравнении степени охвата в статистическом бизнес-регистре с таковой в коммерческом телефонном справочнике, можно найти в работе "Развитие бизнес-статистики малых областей в Соединенном Королевстве" по адресу: http://live.unece.org/fileadmin/DAM/stats/documents/ces/sem.53/wp.7.e.pdf

населения: лица, рожденные в 1997 г. должны быть еще в школе, поэтому они не могут иметь профессию или работодателя. Поскольку регистр образования дает в качестве даты рождения 1967 год, это похоже на простую ошибку ввода. Автоматическая проверка обычно обнаруживает столь очевидные ошибки, хотя ее надо использовать с осторожностью; например, в Финляндии было обнаружено несколько подлинных случаев, когда дети были старше родителей – в связи с усыновлением!

- Временная привязка – приведенные адреса и почтовые коды могут в действительности относиться к одному и тому же зданию, но в разные моменты времени. Различия могут быть обусловлены изменением границ между районами обслуживания почтовых отделений. Это можно установить, справившись по файлам с историей адресов, либо найдя на карте текущий и исторический адреса с использованием географических информационных систем.
- Аббревиатура “St” в конце адресной строки 1 в регистре образования – обычная аббревиатура для “Street”, поэтому эти текстовые элементы следует трактовать как синонимичные, когда они появляются в конце текстовой строки. Заметим, однако, что “St” в начале адресной строки 1 использовано как аббревиатура для “Saint”, так что и здесь нужно некоторое внимание. Похожие примеры можно найти и в других языках.
- Временная привязка и аббревиатуры – в Соединенном Королевстве “CSO” является аббревиатурой для “Central Statistical Office”, прежнего названия Национального статистического управления, которое может все еще использоваться теми, кто не знаком с изменением.
- Различное написание – “Steve” и “Stephen” являются различными вариантами одного имени и должны трактоваться таким образом.
- Проблемы классификации – занятия “статистик” and “государственный служащий” не являются взаимно исключающими. “Статистик” говорит о профессии, тогда как “государственный служащий” больше относится к характеру занятости.
- Значения по умолчанию – иногда, когда значения отсутствуют, либо присутствуют лишь частично, используется какой-либо вид значений по умолчанию. Обычно значениями по умолчанию являются “Z” или “9999999”. В Соединенном Королевстве в случае, когда неизвестна вторая часть почтового кода, обычно используется значение “9XX”, как это видно в поле “Почтовый код места работы”. К сожалению, от использования этого кода пришлось отказаться, когда Почтовая служба начала присваивать реальные почтовые коды, оканчивающиеся “9XX”!

4.10 Пропущенные данные

Проблема пропуска данных не является присущей только административным источникам. Оно может также иметь место вследствие полного или частичного неответа при статистических обследованиях и даже в результате удаления значений данных в процессе редактирования. Однако в случае административных источников последствия могут быть подчас иными, в частности потому, что проблема пропуска данных может зачастую быть более систематической.

Причины этого состоят в том, что тот или иной показатель может вовсе не собираться административным источником, либо собираться только по определенным категориям единиц, когда имеются специфические административные требования. Показатель может просто быть малозначимым для административных целей, поэтому держатели этого источника не видят проблемы в отсутствии данных.

Некоторые стандартные подходы к обработке неответов при статистических обследованиях могут также использоваться для решения проблемы пропуска данных в административных источниках. Различные методы восстановления данных, такие как дедуктивная, “hot-deck” и “cold-deck” импутация, часто пригодны в случаях, когда проблема касается только некоторых единиц. Когда затрагивается большая часть единиц, более подходящим может оказаться метод моделирования.

Вставка 4.4 – Анализ проблемы: работа с пропусками административных данных – Оброт в расчете на одного работника

Для определения размера бизнеса обычно применяются два показателя – число работников и общий объем продаж (оборот). Однако по новым бизнесам, особенно малым, один или оба этих показателя нередко отсутствуют или не заслуживают доверия.

Относительные величины оборота в расчете на одного работника могут использоваться для оценки пропущенных значений, что упрощает решение проблемы. Эти величины формируются с использованием информации по аналогичным бизнесам, для которых оба показателя имеются и признаются надежными, путем вычисления среднего оборота в расчете на одного работника для различных категорий единиц, сгруппированных по видам деятельности и институциональным секторам.

Например, ниже приведены расчетные значения оборота на одного работника, вычисленные для различных классов Международной стандартной отраслевой

классификации (ISIC):

ISIC class	**Оборот в расчете на одного работника**
…….	
45.11	95
45.12	68
45.21	149
…….	

Если бизнес имеет по ISIC класс 45.12, а его оборот равен 200, но данные о занятости отсутствуют, то вмененное значение занятости составит:

200 / 68 = 2.94 (округленно 3)

При вычислении значений оборота в расчете на одного работника часто возникают проблемы, связанные с выбросами, тогда обычно используются такие методы, как отсечение x% наибольших или наименьших значений, а также расчет средней величины интерквартильного размаха.

Относительные показатели этого типа могут также иметь более широкое применение для контроля новой информации, состыковки записей из различных источников, выявления ошибок. Например, графически отображая и исследуя распределение значений оборота в расчете на одного работника, зачастую можно получить полезную информацию о рассматриваемой совокупности единиц. Последующие графики являются примерами такого исследования:

1. Нормальное распределение

В этом случае значения оборота в расчете на одного работника одинаково распределены вокруг средней, свидетельствуя об относительной однородности внутри совокупности единиц и очень ограниченном влиянии выбросов.

2. Несимметричное распределение

В этом случае налицо отчетливое сосредоточение единиц в области сравнительно низких значений, однако выбросы в направлении правого края распределения будут, безусловно, влиять на его среднюю. Это довольно обычно для распределения данных об обороте в расчете на одного работника, что подчеркивает необходимость принятия мер по снижению влияния этих выбросов.

3. Бимодальное распределение

Здесь показан случай, когда рассматриваемая совокупность является в значительной степени гетерогенной, поэтому стоило бы разбить ее на две подсовокупности с целью получения более представительных величин оборота в расчете на одного работника

4.11 Сопротивление переменам

Один из основных барьеров для более эффективного использования административных источников в официальной статистике (и один из наименее признаваемых) коренится в организации. Статистики могут сопротивляться использованию административных данных, поскольку они не доверяют данным, которые они не собирали сами. Они обычно обращают внимание на негативные аспекты качества административных данных, но имеют чрезмерно оптимистичный взгляд на качество данных наблюдений, который зачастую основан на во многом не проверенном предположении того, что данные наблюдений соответствуют статистическим нормам.

Решение очевидно состоит в улучшении образовательной подготовки статистиков в области предоставляемых административными источниками возможностей, побуждении их к более широкому видению всех измерений качества и влиянию на поставщиков и пользователей данных. В этом контексте важно реально определить относительное качество данных обследований и административных данных. Например, часто считается, что данные из административных источников не отвечают требованиям статистических определений, тогда как данные официальных наблюдений им отвечают. Хотя практически здесь может не быть какого-либо реального различия, особенно когда респонденты статистических наблюдений просто копируют значения из последней административной отчетности, не читая зачастую очень длинные пояснения, как тот или иной показатель должен определяться для статистических целей.

Еще одним путем, способствующим разрушению барьеров внутреннего сопротивления, является аргументация того, что снижение затрат за счет использования административных источников не обязательно означает сокращение персонала. Сэкономленные ресурсы могут, хотя бы частично,

использоваться для повышения качества или периодичности выходной статистики.

4.12 Резюме

Эта глава отчетливо показывает, что есть много проблем, которые должны преодолеваться при использовании административных источников. Она ставит своей целью показать, что читатели также оказались перед этими проблемами, и что в большинстве случаев можно найти полные или частичные решения. Она не может покрыть всех проблем, потенциально возникающих перед читателем, особенно привязанных к конкретным источникам, но цель состоит в предоставлении концепции, которая может быть приспособлена к конкретным обстоятельствам.

В целом, будет правильным сказать, что большинство проблем, имеющих место при использовании административных данных для статистических целей в увязке с другими аспектами статистики, могут быть преодолены или, по крайней мере, умерены за счет эффективного планирования и управления, хорошего знания источников данных, креативного мышления, а также желания делиться опытом и учиться у других.

Для административных данных обычно требуется другая технология обработки, чем для статистических источников. Простая подстановка административных данных вместо статистических данных без изменения процесса статистического производства редко будет срабатывать на практике.

Главное, что следует иметь в виду: несмотря на все проблемы, преимущества использования административных данных все же являются намного значительнее затрат.

5. Качество и административные данные

5.1 Введение

Как отмечалось в главе 4, опасения по поводу качества административных данных обычно являются одним из основных барьеров для их возрастающего применения для статистических целей. Эти опасения могут быть оправданными или нет, и они часто основаны только на определенных аспектах качества, таких как своевременность. Чтобы правильно решать вопросы, связанные с этими опасениями, необходима целевая система управления качеством, которая бы анализировала все значимые аспекты качества и обеспечивала принятие компетентных решений. Многие статистические организации уже создали какие-либо системы обеспечения качества данных, собираемых путем традиционных методов наблюдения, однако сравнительно немногие распространили этот подход на данные из административных источников.[21]

5.2 Определение качества

Исходной точкой такой системы является определение качества как такового. Надо сказать, что в этой области национальными и международными статистическими организациями проведена большая работа, большая часть которой основана на международном стандарте 9000/200522 [22] , который определяет качество как

“степень, в которой совокупность внутренне присущих характеристик соответствуют требованиям”.

К сожалению, это определение не является легким для предметного понимания и требует некоторых дальнейших пояснений. Чтобы облегчить интерпретацию, его можно подразделить на следующие части:

1) “требования”

Это, как правило, означает требования пользователей конкретных товаров или услуг, однако это можно также рассматривать и с точки зрения, что требования производителя и даже общества в целом должны также быть приняты во внимание. Например, скоростной автомобиль с большим двигателем может вполне отвечать требованиям индивида, но не соответствовать требованиям общества касаемо загрязнения или дорожной безопасности. Однако готовые продукты официальных статистических агентств, собственно статистика, обычно производятся в государственном секторе как “общественное благо”, так

[21]Одним из примеров является подход, разработанный Статистикой Нидерландов (http://isi2011.congressplanner.eu/pdfs/950481.pdf) и Статистикой Швеции (http://www.scb.se/statistik/_publikationer/OV9999_2011A01_BR_X103BR1102.pdf)
[22]См.: http://www.iso.org/iso/catalogue_detail?csnumber=42180

что в этом случае эти разные группы требований в значительной степени совпадают.

Хотя, если мы рассмотрим массив административных данных сам по себе, то может быть заметное расхождение между требованиями производителя (например, административного агентства) и пользователя (статистическая организация). Более того, поскольку “трансакция” обычно не является рыночной, производителя могут иметь малую мотивацию принимать во внимание требования пользователя. Это может вызывать напряженность по поводу качества, что обосновывает необходимость надежной организационной основы, описанной в главе 3.

2) “совокупность внутренне присущих характеристик”

Пользователи любых товаров или услуг судят о качестве по набору критериев, касающихся различных характеристик этих товаров или услуг. Часто это делается подсознательно, как, например, в отношении обеда в ресторане: индивид будет судить о его качестве по тому, как еда готовилась и подавалась, по ее количеству, обслуживанию, интерьеру и окружению в ресторане, и возможно по нескольким другим критериям (стоимость пока не включается, мы вернемся к ней позже в этой главе). Качество статистических данных может оцениваться аналогичным образом – по набору внутренне присущих характеристик или критериев.

Некоторые статистические агентства разработали перечни критериев для оценки качества статистических данных, в то же время основные международные агентства в настоящее время достигли согласия по следующему перечню:

• Релевантность – степень, в которой статистические данные соответствуют требованиям имеющихся и потенциальных пользователей. Релевантность, таким образом, означает, производятся ли те статистические данные, которые нужны, и нужны ли те статистические данные, которые производятся. Она также характеризует степень, в которой используемые методологии (определения, классификации и т.д.) отражают нужды пользователей.

• Точность – близость статистических оценок к действительным значениям.

• Своевременность – она отражает разницу между временем, когда данные становятся доступными, и временем, когда имеет место событие или явление, которые они описывают.

• Пунктуальность – временной лаг между датой, когда данные фактически опубликованы, и намеченной (обычно предварительно объявленной) датой опубликования.

• Доступность – физические условия, при которых пользователи могут получить данные: куда обращаться, как заказывать, время доставки, прозрачная ценовая политика, удобные условия для клиентов (авторские права и пр.), доступность микро- и макро-данных, разнообразие форматов (на бумаге, файлы, CD-ROM, Интернет…) и т.д.

• Ясность / интерпретируемость – сопровождаются ли данные необходимыми и достаточными метаданными, повышается ли эффективность представления информации за счет иллюстраций, в частности, графиков и карт, имеется ли информация о качестве данных.

• Взаимосогласованность и совместимость данных – данные из различных источников, в частности – из статистических обследований различного характера и/или частоты, могут не быть полностью взаимно согласованы в том смысле, что они могут опираться на различные подходы, классификации и методологии. Поэтому они не могут дать пользователю полностью увязанную информацию, например, пользователь может оказаться в недоумении, когда два различных измерения одной и той же величины публикуются с различными значениями.

• Сопоставимость – степень, в которой разница в статистических оценках обусловлена разницей в действительных значениях статистической характеристики, а не методологическими различиями. Сопоставимость охватывает:

  ° сопоставимость во времени – степень, в которой могут быть сравнены данные за различные моменты времени;

  ° сопоставимость в пространстве – степень, в которой могут быть сравнены данные по различным странам и/или регионам;

  ° сопоставимость между предметными областями – степень, в которой могут быть сопоставлены данные из разных областей статистики.

Применительно к административным данным перечень критериев может использоваться двумя способами. Во-первых, он может использоваться для оценки качества выходной статистики и сравнения данных, основанных на административных источниках, с данными, основанными на обследованиях. Во-вторых, перечень может использоваться при оценке качества различных административных источников самих по себе[23]. Например, если статистику повезло иметь выбор из двух или более административных источников, перечень может помочь определить, у какого источника более высокое качество.

[23] Для применения и распространения этого подхода см. дискуссионный материал Статистики Нидерландов “Перечень вопросов для оценки качества административных источников данных”: http://www.cbs.nl/NR/rdonlyres/0DBC2574-CDAE-4A6D-A68A-88458CF05FB2/0/200942x10pub.pdf39

Однако когда перечень используется для оценки качества административных данных, следует обратить внимание, что истинную точность сложно определить в отсутствии достаточной поддерживающей информации о самой совокупности и процессе сбора данных. В этом случае должны рассматриваться два фактора – надежность источника и правдоподобие данных, то есть, есть ли доверие к источнику и выглядят ли данные адекватными при сравнении с другими источниками и ожидаемыми статистиком значениями. Для более объективной оценки может потребоваться обследование качества каким-либо способом с тем, чтобы определить точные значения некоторых величин.

Близость административных единиц и показателей к единицам и показателям, требуемым для статистических целей, – важный фактор при определении качества административного источника. Чем меньшее преобразование требуется, тем меньше риск ошибок и смещения. Этот аспект может рассматриваться как компонент критерия совместимости.

5.3 Затратные ограничения

Затраты сознательно исключаются из большинства перечней критериев качества в статистике, поскольку они считаются скорее ограничением. Коль скоро качество определено, в уравнение добавляются затраты, что обеспечивает принятие эффективных по затратам практических решений.

Тем не менее затраты особенно важны в случае административных источников, ибо когда оказывается, что административные источники дают более низкий абсолютный уровень качества, чем данные наблюдений, они все равно могут иметь достаточные преимущества с точки зрения затрат, что может их делать наиболее эффективным решением. К тому же может иметься возможность направлять некоторую экономию затрат на улучшение качества, таким образом снижая или элиминируя пробелы в качестве.

5.4 Измерение качества на практике

Чтобы в полной мере разобраться в качестве административных источников и их воздействии на качество статистики, нам надо рассмотреть три компонента:

1) Качество входных данных

Входные данные, будь то данные административных источников или обследований, могут оцениваться по комплексу критериев, подобным тем, которые перечислены выше. Наиболее важными критериями, видимо, являются своевременность, а также релевантность, понимаемая как степень, в которой полнота охвата и методология источника соответствуют требованиям. Сопоставимость с другими источниками может также быть важна, и время от времени могут требоваться те или иные мероприятия по обеспечению

сопоставимости данных из различных источников, дабы иметь ясную картину качества. Обследования на предмет проверки качества иногда используются для этой цели.

Стоит иметь в виду один момент – степень, в которой субъект данных заинтересован в качестве данных. Размер усилий и внимания, вкладываемых в подготовку данных, изменяется в зависимости от ощущения ценности и важности сбора данных, таким образом, субъекты данных в ряде случаев могут предоставлять для административных целей данные лучшего качества, чем они это делают для статистических целей.

2) Качество обработки данных

Даже если входные данные безупречны, их качество, тем не менее, может ухудшиться при различных процессах, которым они подвергаются, прежде чем будут использованы для получения выходных статистических данных. В идеале, обработка должна улучшать качество, однако, к сожалению, это не всегда происходит. Примеры того, как обработка данных может влиять на качество, включают:

- Установление соответствия и связывание данных – слишком много ошибочных соответствий приведет к ошибкам в данных; слишком много ошибочных несоответствий приведет к дублированию, которое вызовет завышение размера целевой совокупности и, возможно, – появление смещения.

- Выявление и обработка выбросов – использование методов выявления выбросов для обнаружения ошибок может помочь улучшению качества данных; вообще, чем экстремальнее выброс, тем вероятнее, что это ошибка. Однако чрезмерно рьяная обработка выбросов приведет к искажению истинных значений данных, и важные тренды данных могут быть не замечены.

- Качество редактирования данных – как и в случае выявления и обработки выбросов, редактирование данных должно улучшать качество, однако, не будучи выполнено с осторожностью, оно может внести ошибки и смещения[24].

- Качество восстановления [импутации] данных – когда восстановление используется для заполнения пропущенных значений или записей, оно может способствовать улучшению охвата, однако и в этом случае использование метода требует тщательного анализа, чтобы избежать внесения смещений.

[24]Полный комплект материалов по различным проблемам редактирования данных см. в документах рабочей сессии по редактированию статистических данных, организованной ЕЭК ООН – http://www1.unece.org/stat/platform/display/kbase/UNECE+Work+Sessions+on+Statistical+Data+Editing

Есть один очень важный принцип, которого следует придерживаться, особенно при обработке данных из административных источников, – это хранение копии сырых данных (и любых ассоциированных метаданных), чтобы при необходимости вернуться назад. Сравнение данных до и после обработки может помочь оценить качество этой обработки и выявить любые характерные проблемы.

3) Качество выходных статистических данных

Распространенная интерпретация статистическими агентствами определения качества ISO состоит в том, что качество всецело определяется соответствием требованиям пользователей. Поэтому качество выходной статистики определяется в этом контексте. Это означает, что важно определить эти требования, обсуждать их с пользователями и получать обратную реакцию, например, посредством обследования степени удовлетворенности пользователей.

Переход от обследований к административным источникам, несомненно, окажет воздействие на качество выходных данных. Как правило, это воздействие может быть позитивным по некоторым критериям качества, но негативным по другим. Во всех случаях необходимо получать обобщенное представление о воздействии, придавая больший вес тем критериям, которые пользователи считают наиболее важными. Например, пользователи могут считать, что повышение своевременности более чем компенсирует снижение точности, особенно для краткосрочных экономических данных. Другим аспектом рассмотрения должно быть воздействие на данные временных рядов, и можно ли сформировать, следуя за изменениями, непротиворечивые ряды достаточной длины.

Особенно важным может являться придание мнению пользователей, по крайней мере, такого же веса, как и мнению статистиков, которые иногда бывают слишком фокусированы на традиционных представлениях о точности. Вообще, является жизненно важным, чтобы любые оценки воздействия на выходные статистические данные базировались на объективных фактах, а не на предположениях, поскольку это является единственным путем противодействия возможному сопротивлению переменам, описанному в главе 4. Одним из способов обеспечения этого является использование докладов о качестве согласно стандартным схемам[25] для документирования воздействия изменений в источниках данных информирования об этом.

[25]Например, предложенных Евростатом: http://epp.eurostat.ec.europa.eu/portal/page/portal/ver-1/quality/documents/ESQR_FINAL.pdf

5.5 Роль метаданных

Метаданные[26] жизненно важны для информирования как производителей, так и пользователей о качестве данных. Они должны иметься на всех трех стадиях, указанных в предыдущем разделе. Входные данные должны сопровождаться метаданными, достаточными для их понимания и обеспечения того, чтобы значения правильно привязывались к соответствующим показателям. Также важна детальная документация по методологиям, определениям и источнику данных, равно как и по используемым методам сбора и обработки данных. Это обеспечит лучшее понимание потенциальных проблем качества, и формирование базы для правил редактирования на стадии обработки.

В ходе обработки данных важно фиксировать, что и с какими записями и значениями было сделано. Этим не только формируется жизненно важная информация для оценки качества обработки, но также механизм изучения любых потенциальных проблем в этом процессе и исправления ошибок.

Выходные статистические данные должны сопровождаться метаданными, достаточными, чтобы дать пользователям возможность извлекать их, правильно интерпретировать и формировать мнение об их качестве. Регулярным и интенсивным потребителям выходных данных информация, необходимая им для формирования на основе данных верных выводов, предоставляется в виде полной документации по всем трем стадиям, желательно в стандартном формате. Часто бывает трудно понять информацию о качестве, ибо некоторые пользователи желают иметь все подробности, тогда как других устраивают самые обобщенные индикаторы. Возможно, наиболее подходящей является модель метаданных, позволяющая пользователям иметь разные уровни информации, начиная с обобщенного уровня, но с правом выбора большей детализации.

5.6 Резюме

Наилучший путь оценки качества административного источника лежит через накопление доскональных знаний об этом источнике, в том числе о главной цели источника и способах сбора и обработки данных. Глубокое понимание источника делает возможным более точную оценку его сильных и слабых сторон.

Чтобы оценить влияние использования различных источников, необходимо сочетать знания об источниках и процессах, используемых для их преобразования в выходные статистические данные, с мнением пользователей об этих выходных данных. Это обеспечит объективную комплексную оценку

[26]Данные, которые определяют и характеризуют другие данные (Источник: ISO/IEC FDIS 11179-1 "Информационные технологии – Реестры метаданных – Часть 1: Структура", Март 2004)

влияния использования административных данных в сравнении с данными статистического наблюдения.

6. Связывание и стыковка данных

6.1 Введение

Есть два основных подхода к использованию административных данных в статистическом производственном процессе:

- в качестве непосредственного источника для статистики,
- в качестве дополнительного источника, в сочетании с другими источниками.

Если данные из нескольких административных источников используются в дополнение к данным наблюдений или для наполнения статистического регистра, статистической организации будет необходимо найти какой-либо способ связывания этих данных. Обычно это приобретает форму стыковки данных[matching], которую можно определить как соединение данных из различных источников на основе общих признаков, имеющихся в этих источниках.

6.2 Общие идентификаторы?

Если эти общие признаки включают какой-либо общий регистрационный или идентификационный номер (впредь именуемый общим идентификатором), процесс называется состыковкой с точным соответствием [exact matching] и он сравнительно легок. При состыковке с точным соответствием существуют два возможных результата – две записи из различных источников либо точно стыкуются с использованием общего идентификатора, либо не стыкуются. Другими словами, запись с идентификатором 123456 будет поставлена в соответствие записи с таким же идентификатором в другом источнике (в предположении, что источники охватывают те же единицы!), однако, она не будет соответствовать записи с идентификатором 123457.

Стыковка с точным соответствием сильно зависит от качества стыкуемых признаков, используемых в каждом источнике. Если есть ошибки в общих идентификаторах хотя бы в одном источнике, есть высокий риск как состыковки не соответствующих друг другу единиц, так и нестыковки единиц, которые должны быть состыкованы. Поэтому даже когда во всех подлежащих стыковке файлах имеются общие идентификаторы, может оказаться недостаточным полагаться только на состыковку с точным соответствием.

Иногда идентификаторы могут включать контрольные числа, то есть один или более символов, формируемых по стандартному алгоритму на основе других символов идентификатора. Когда есть контрольные числа, они помогают гарантировать определенный уровень качества путем устранения большинства ошибок набора и считывания.

6.3 Стыковочные ключи и понятие различительной способности

Когда отсутствуют общие идентификаторы, либо их качество недостаточно для требуемого уровня точности состыковки, надо рассмотреть возможность использования других характеристик, общих для данных источников. Выбранные характеристики обычно именуются “стыковочными ключами”. Заметьте: не всегда обязательно присутствие этих характеристик в обоих источниках, ибо в ряде случаев они могут быть вычислены (например, смотри комментарии об обороте в расчете на одного работника во вставке 4.4). Когда используются показатели, иные чем идентификаторы, техника стыковки основывается на вероятностном определении соответствующих друг другу единиц.

Характеристиками, обычно используемыми для вероятностной состыковки такого рода, являются имя, адрес, дата рождения, код вида занятий или экономической деятельности. Выбор характеристики, используемой для состыковки, должен учитывать “различительную способность” каждой характеристики. Различительная способность характеризует неповторимость значений стыковочного ключа. Некоторые характеристики имеют более высокую различительную способность, чем другие:

- высокая различительная способность: регистрационный номер, полное имя, полный адрес;
- низкая различительная способность: пол, возраст, город, национальность.

Внутри характеристики типа “полное имя” некоторые значения также могут иметь более высокую различительную способность, чем другие. Уникальные имена будут иметь более высокую различительную способность, тогда как более часто используемые (например, Джон Смит во многих англоговорящих странах) – намного меньшую различительную способность.

Различительная способность может также зависеть от степени детализации, например:

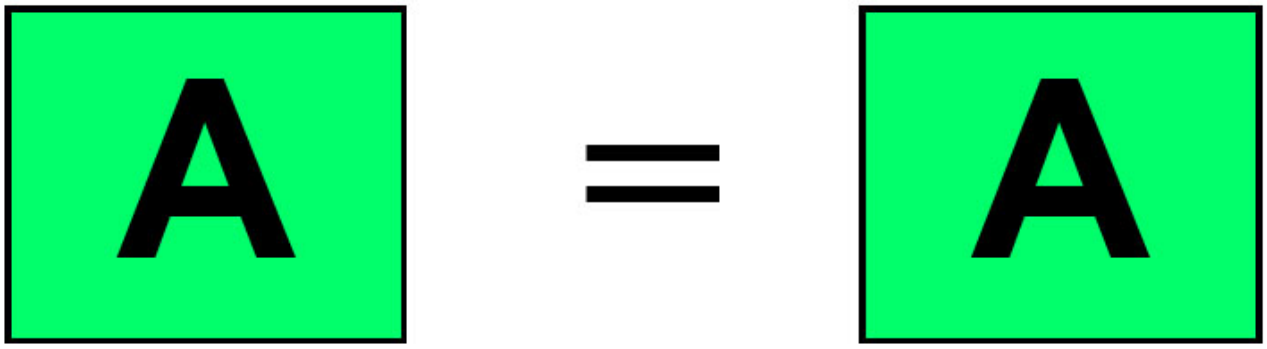
- “Родился в 1960 г., Париж” = низкая различительная способность;
- “Родился 23 июля 1960 г., улица l’Eglise, Монмартр, Париж” = высокая различительная способность.

Поэтому тщательный выбор стыковочных ключей, учитывающих концепцию различительной способности, может значительно повлиять на успешность операции состыковки.

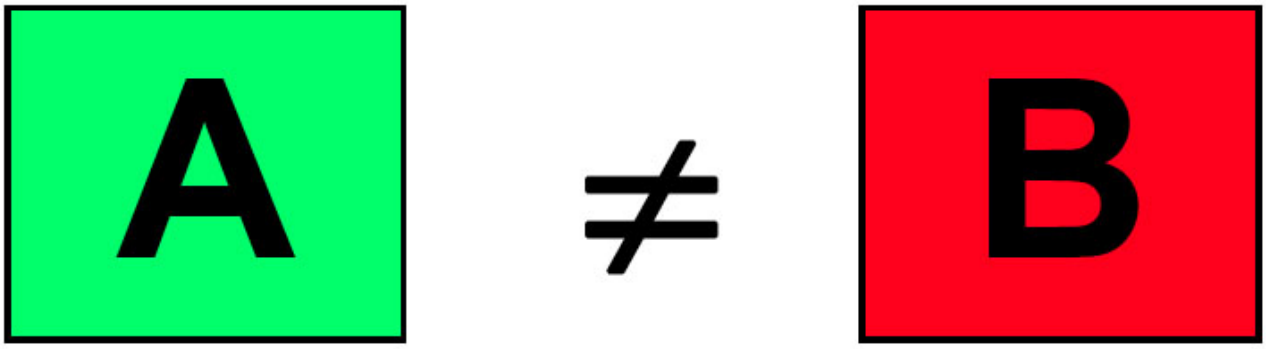
6.4 Некоторые базовые понятия процесса стыковки

Когда сопоставляются две записи, их можно называть “парой”. Следующие сценарии иллюстрируют основные потенциальные исходы применения техники стыковки к этой паре записей.

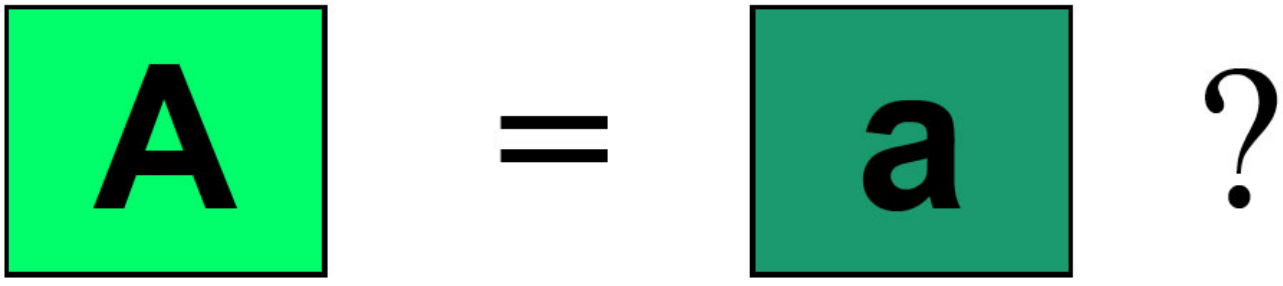
1) Соответствие – пара олицетворяет один и тот же объект действительности.



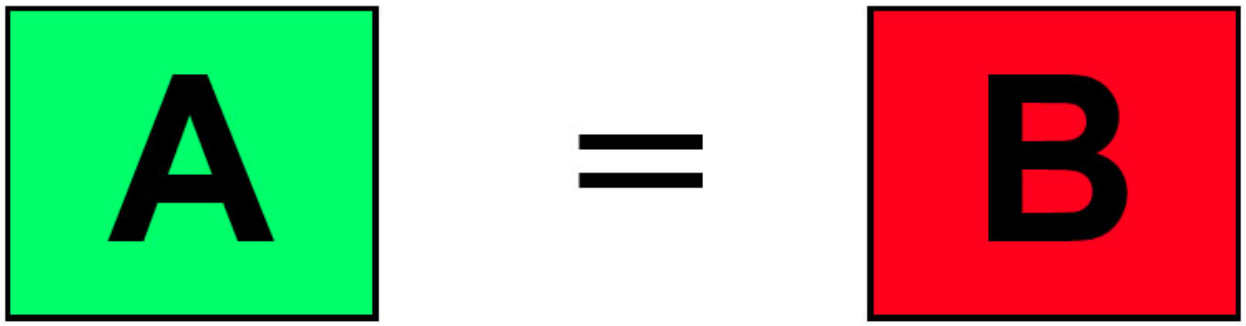
2) Несоответствие – пара олицетворяет два различных объекта действительности.



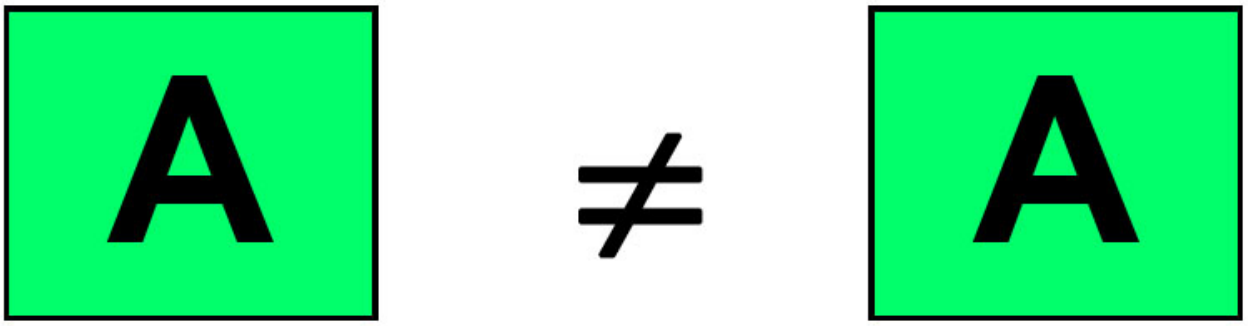
3) Возможное соответствие – пара, по которой нет достаточной информации, чтобы определить, имеет ли место соответствие или несоответствие.



4) Ложное соответствие – пара, которая в процессе состыковки ошибочно определена как соответствие (ложнопозитивный результат).



5) Ложное несоответствие – пара, в которой в действительности имеется соответствие, но в процессе состыковки она ошибочно определена как несоответствие (ложнонегативный результат).



Чтобы достичь лучшего понимания принципов и практических аспектов стыковки данных, пожалуйста, ознакомьтесь с примером по стыковке в конце данной главы. В нем использованы вымышленные, но реалистичные данные, иллюстрирующие, что состыковка без общих идентификаторов требует определенных логических рассуждений, и что состыковка может часто быть скорее искусством, чем точной наукой. Любая форма вероятностной состыковки, скорее всего, приведет к некоторой доле ложных соответствий и ложных несоответствий, а также к необходимости дальнейшего изучения возможных соответствий.

6.5 Методы стыковки

Методы стыковки могут быть подразделены на две основные категории:

1) Ручная стыковка – по определению требует значительной работы человека с информацией, поэтому она, скорее всего, является:

- **дорогостоящей,**
- **нестабильной,**
- **медленной,**
- **зато интеллектуальной.**

2) Автоматическая стыковка – будучи в эксплуатации (то есть без учета единовременных затрат на создание), этот метод минимизирует участие человека, поэтому он способен быть:

- **дешевым,**
- **стабильным,**
- **быстрым,**
- **однако, ограниченно интеллектуальным.**

Поэтому наилучшее решение состоит в использовании инструментов автоматической стыковки для нахождения очевидных соответствий и несоответствий и в передаче возможных соответствий специалистам. Для обеспечения эффективности по затратам необходимо максимизировать применение автоматической стыковки при минимизации участия специалистов. В оставшейся части данной главы рассматриваются основные черты автоматической стыковки, а также возможности ее практического использования и улучшения.

6.6 Как работает автоматическая стыковка

Инструменты автоматической стыковки обычно следуют однотипной последовательности шагов, хотя в зависимости от конкретного приложения некоторые шаги могут быть опущены, а другие – добавлены. Наиболее применимы следующие шаги.

1) Стандартизация

Этот шаг применяется, главным образом, к текстовым показателям, либо к показателям, которые должны соответствовать специфическим форматам. Примерами процессов стандартизации являются:

- Аббревиатуры и общие термины заменяются стандартным текстом, например, текстовая последовательность“ltd” преобразуется в “limited”, а “mfg” – в “manufacturing”.
- Общепринятые видоизменения имен стандартизируются, примерами могут служить различные версии названий городов (“Brussel” / “Bruxelles” в Бельгии, “Derry” / “Londonderry” в Северной Ирландии). Аналогичный процесс необходим применительно к личным именам, когда имеются различные написания одного и того же имени (“Jane” / “Jayne”), либо общепринятые краткие версии имени (“Bill” / “William”). Этот процесс аналогичен стандартизации аббревиатур и, вероятно, может быть объединен с ней.
- “Зашумляющие” слова удаляются – обычно это слова или фразы с очень

низкой различительной способностью, примерами могут быть “шоссе” или “улица” в адресах.

• Почтовым кодам, датам рождения и пр. придаются общие форматы, например, “3 января 1985 г.” должно быть преобразовано в “030185”.

Процесс стандартизации сильно зависит от языка и может изменяться в зависимости от типа стыкуемых записей, поэтому вышеприведенные примеры лишь иллюстрируют процесс. Каждый случай состыковки требует предварительной работы, которая обычно базируется на изучении данных с целью выяснения, какие правила стандартизации следует применять.

Стандартизация может также рассматриваться как форма вычищения данных, и в качестве таковой нести риски искажения и снижения качества данных, а в предельных случаях – снижения вероятности нахождения правильного соответствия. Такие риски обычно очень малы и обычно вызваны неоднозначностью стандартизуемых записей. В английском языке примерами являются аббревиатура “St.”, которая может относиться как к “street”, так и к “saint”, а также имя “Chris”, которое может быть краткой формой “Christopher” (мужское) или “Christine” (женское).

Другим типом стандартизации, в ряде случаев используемым в качестве начального шага в процессе состыковки, является проверка адресов по точному перечню, обычно перечню национальной почтовой службы. Она может варьировать от проверки правильности сочетания почтового кода и наименования поселения / города / региона до полной проверки всего адреса. Успешность такой проверки, очевидно, будет сильно зависеть от качества справочных файлов используемых адресов.

Если решено использовать “очищенные” адреса, будет правильной практика хранения также копий сырых данных. В ряде случаев (включая стыковку бизнес-данных в Соединенном Королевстве) было обнаружено, что использование очищенных адресов повышает вероятность состыковки для одних записей, но понижает – для других. Совмещение результатов двух параллельных процессов стыковки – один с использованием очищенных адресов, а другой с использованием сырых версий – обычно дает наилучший результат.

Следует также отметить два других потенциальных последствия использования очищенных адресов, хотя они прямо не связаны со стыковкой. Первое состоит в том, что в некоторых станах почтовые службы могут предоставлять скидку на оптовую рассылку, если используемые адреса соответствуют определенным стандартам, так что это может способствовать возмещению затрат на процессы очистки и состыковки. С другой стороны, подстановка очищенных адресов вместо тех, которые предоставлены респондентами, в некоторых случаях может

вызвать раздражение респондентов. Использование очищенных адресов при посылке статистических вопросников может повлиять на уровень отвечаемости. Это дает дополнительные аргументы в пользу хранения как очищенных, так и сырых данных всегда, когда это возможно.

2) Парсинг

До некоторой степени парсинг может рассматриваться как развитие стандартизации. На этом шаге текст преобразуется из формы, легко распознаваемой людьми, в форму, которая является более логичной для компьютерной обработки, и потому способствует правильной состыковке. Получаемые текстовые строки часто рассматриваются в качестве стыковочных ключей. Первые подходы к парсингу в английском языке часто использовали “алгоритм саундекса”, впервые запатентованный в 1918 г. Этот алгоритм или его производные составляет основу многих программ для состыковки. Однако правила парсинга для разных языков значительно различаются и для получения наилучших результатов должны перестраиваться применительно к данным.

Примеры правил парсинга могут включать следующее:

- преобразование букв или групп букв со сходным звучанием в общее для них обозначение, например, “f”, “v” и “ph” в “f”;
- удаление непроизносимых букв, например, “h” в имени “Thomas”;
- преобразование всех букв либо в прописные, либо в строчные;
- преобразование гласных букв в один-единственный символ;
- удаление гласных букв в конце имени или слова;
- замена сдвоенных букв одной буквой, например, “Ann” становится “An”.

Например, в процессе парсинга с использованием всех вышеперечисленных правил строка “Steven Thomas Vale” преобразуется в “stafan tamas fal”. Строка “Stephen Tomos Vael” будет давать такой же результат, иллюстрируя, как парсинг может способствовать повышению степени стыкуемости путем снижения влияния различий в написании имени и ошибок написания. Следует также отметить, что изменение порядка, в котором применяются правила парсинга, может повлиять на результат.

Однако, как и в случае стандартизации, если парсинг недостаточно хорошо подстроен под стыкуемые данные, есть риск, что он принесет больше вреда, чем пользы. По крайней мере, на начальных стадиях воздействие методик парсинга должно тщательно анализироваться, и во всех случаях копия сырых данных должна сохраняться для целей сравнения.

3) Разбиение на блоки

Когда файл, с которым осуществляется состыковка, очень велик, может оказаться необходимым подразделить его на более малые блоки, чтобы сэкономить время обработки. Есть несколько способов сделать это, например, если подлежащая стыковке запись имеет адрес в определенном городе, может оказаться возможным работать при состыковке с блоком, содержащим другие записи из этого города, а не со всеми записями в целом по стране.

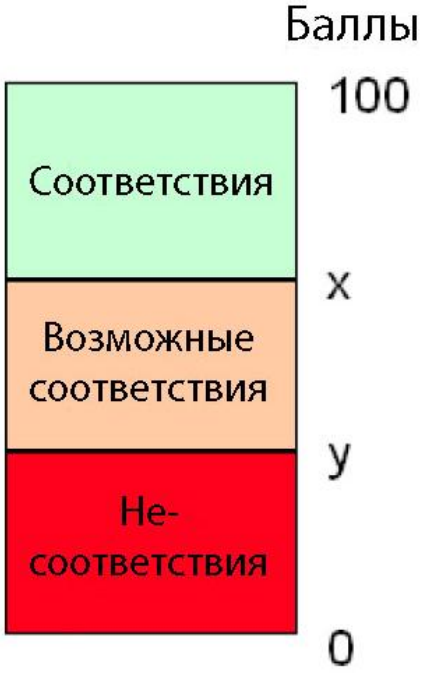
Разбиение на блоки должно применяться с осторожностью, и часто оно может приводить к снижению итоговой степени стыкуемости. Однако если такое снижение минимально, а выигрыш от более быстрой обработки существенен, то разбиение на блоки может повысить эффективность по затратам процесса состыковки. В некоторых случаях может даже оказаться уместным проведение двух или более процедур состыковки, используя различные критерии разбиения на блоки. Например, после применения сильно ограничительного критерия состыковки к полному массиву данных может оказаться полезным еще раз обработать подмассив первоначально не состыковавшихся записей, последовательно используя менее ограничительные критерии разбиения на блоки.

Разбиение на блоки наиболее целесообразно при очень больших массивах данных, такие как индивидуальные записи переписи населения, однако его полезность для процедуры состыковки может снижаться по мере возрастания мощности и быстродействия компьютера.

4) Скоринг

Большинство автоматических процедур состыковки используют какие-либо формы балльной оценки (скоринга) для определения вероятности соответствия между двумя записями. Баллы начисляются на основе того, насколько близко соответствуют друг другу стыкуемые признаки. Эти баллы могут использоваться, чтобы определить, признается ли пара записей точным соответствием, возможным соответствием или несоответствием. Рисунок 5.1 показывает, какая из категорий [соответствия] может быть присвоена на основе пороговых величин баллов x и y по 100-балльной шкале.

Рисунок 5.1 Использование пороговых баллов для определения категории соответствия



Следующий закономерный вопрос состоит в том, как определить значения x и y. Один из вариантов состоит в использовании метода, основанного на моделировании, как предложили Fellegiand Sunter [27] , однако на практике равновозможно использование метода проб и ошибок.

Качество данных изменяется с течением времени, поэтому при многократном выполнении операций состыковки требуется периодическая переоценка значений x и y. Сходным образом, к пересмотру порогов может вести изменение требований к стыкуемым данным, либо изменение доступных для состыковки ресурсов. Пороги могут также значительно различаться для различных массивов данных.

При установлении значений x и y необходимо также учитывать влияние ошибок состыковки различного рода. Если ложное соответствие может приводить к раскрытию статистической информации об одной единице другой единице, то значение x должно быть установлено достаточно высоким, чтобы риск ложного соответствия был приемлемо низким. Однако если риск раскрытия отсутствует, а результаты состыковки будут использоваться в исследовании, где определенная доля ложных соответствий вряд ли будет значительно влиять на результаты, то значение x может быть ниже.

На практике возможности привлечения человеческих ресурсов для анализа возможных соответствий обычно накладывают некоторые ограничения на расстояние между x и y. Во всех таких случаях для работы экспертов должны устанавливаться приоритеты. Это может делаться на основе баллов, так что те возможные соответствия, которые имеют самые высокие баллы, проверяются первыми, поскольку от этого ожидается большая польза. Кроме того, некоторые другие характеристики обрабатываемых единиц (например, число работников предприятий) также могут задавать приоритеты для работы экспертов, дабы минимизировать влияние потенциального дублирования.

6.7 Программные средства для стыковки на практике

[27]См.: Теория сцепления записей, Ivan P. Fellegi и Alan B. Sunter,http://www.jstor.org/stable/view/2286061

Наиболее известными многим людям программными приложениями, предназначенными для состыковки данных, являются поисковые системы в интернете, хотя они работают зачастую по-иному, чем было описано выше. Они берут набранную пользователем текстовую строку, осуществляют поиск соответствующих этой строке web-страниц, обычно проводя скоринг результатов, и возращают их [пользователю] в порядке оцененной ими значимости. Те или иные формы парсинга могут проявляться в результатах, либо рекомендациях по альтернативному написанию.

Кроме того, поисковые системы интернета хорошо демонстрируют концепцию различительной способности. Например, на момент написания данной работы поиск по текстовой строке "соответствие [matching]" в www.google.com дал около 700 миллионов результатов, тогда как по "статистическое соответствие [statistical matching]" – около 30 миллионов результатов, а по "способы парсинга при установлении статистического соответствия [parsing techniques in statistical matching]" – около 1,6 миллионов результатов. Большая детализация, очевидно, помогает сфокусировать поиск.

В сфере официальной статистики сложились два основных подхода к разработке программных приложений для стыковки данных:

- Использование готовых коммерческих программных средств, например, Informatica Identity Resolution (включающего SSAName3)[28]. Следует, однако, отметить: прежде чем коммерческий продукт сможет использоваться в полной мере, зачастую требуется та или иная адаптация к требованиям пользователя.

- Разработка технологии стыковки собственными силами, примерами являются программные средства, разработанные Бюро цензов США [29], Статистикой Канады [30] и ИСТАТом – итальянской статистической службой[31].

Альтернативным подходом к состыковке является метод "триграмм", который работает путем разбиения текстовых строк на группы по три символа, а затем вычисления доли групп, идентичных для двух строк.

Например, состыковка строки "Steven Vale"

Ste/tev/eve/ven/en /n V/ Va/Val/ale

со строкой "Stephen Vale"

Ste/tep/eph/phe/hen/**en /n V/ Va/Val/ale**

[28] http://www.informatica.com/products_services/identity_resolution/Pages/index.aspx
[29] http://www.census.gov/srd/papers/pdf/rr2001-03.pdf
[30] http://www1.unece.org/stat/platform/display/msis/G-Link
[31] http://forge.osor.eu/projects/relais/

имеет результатом 6 совпадающих триграмм (выделены жирным шрифтом) из тринадцати уникальных триграмм в обеих строках, что дает балл 6/13 или 0.46. Парсинг строк может способствовать повышению балла, однако, как рассматривалось выше, может внести ошибки[32].

Вставка 6.1 – Анализ проблемы – Извлечение из "Стыковка записей без общего идентификатора – опыт Соединенного Королевства", авторы Steven Vale и Mike Villars.

Данный текст является извлечением из документа, который можно найти в: http://www1.unece.org/stat/platform/download/attachments/56230020/matching+paper.pdf?version=1

Бизнес-регистр Соединенного Королевства использует данные из нескольких административных и статистических источников, наиболее важными из которых являются данные по налогу на добавленную стоимость и налогу на доходы. Области охвата этими двумя источниками в значительной степени пересекаются, поэтому для минимизации дублирования важно проверять, чтобы новые единицы из каждого источника были действительно новыми, и не являлись уже присоединенными из другого источника. Каждый источник имеет собственную систему идентификаторов единиц, имея в виду, что состыковка на основе имен и адресов является наилучшим решением.

Входные файлы обрабатываются в четыре этапа:

- очистка – эта процедура редактирует строку имени, удаляет специальные символы и заменяет строчные символы прописными;
- форматирование – эта процедура представляет строку имени в виде отдельных слов, удаляет игнорируемые слова, заменяет некоторые слова и присоединяет префикс-слова;
- стандартизация – эта процедура "стандартизует" имена, например, удаляя сдвоенные символы;
- генерация ключа – эта процедура генерирует коды на основе входного текста, например, если на входе имеется "Steven Vale", то сформированный ключ будет

  STEVEN □ STAFAN □ XJXM$$$; and VALE □ VAL □ YLVO$$$$

YLVO$$$$ является ключом для последней части имени, который используется в качестве главного ключа. Он проверяется по таблице ключей имен, сформированных на основе имен каждой хранимой в регистре записи, с целью поиска потенциальных соответствий. Имеющееся на входе имя, адрес и

[32]Практическое приложение этого метода, запрограммированное в SAS, демонстрировалось Статистикой Финляндии в рамках проекта Евростата по развитию статистики демографии предприятий.

почтовый код сравниваются с именем, адресом и почтовым кодом каждого из этих потенциальных соответствий и получает балл по сто балльной шкале. Если этот балл >79, то пара считается явным соответствием. Если этот балл находится между 60 и 79, то это возможное соответствие. Более низкий бал рассматривается как несоответствие.

Двойники удаляются из списка явных соответствий, равно как и записи в списке возможных соответствий, которые также есть в списке явных соответствий. Затем записи из списка явных соответствий автоматически привязываются к соответствующим им единицам в регистре. Записи из списка возможных соответствий и значимые непристыкованные записи передаются для проверки специалистами. При типичном случае апдейтирования около 37% записей определяются как соответствие, и 35% – как возможное соответствие (из которых приблизительно 80% могут быть пристыкованы вручную). Одна из встретившихся проблем – использование "Tradingas" либо "T/A" в наименованиях, например, “Mike Villars T/A Mike’s Coffee Bar”. В этом случае “Bar” могло бы использоваться в роли главного ключа, однако оно имеет низкую различительную способность, ибо в Соединенном Королевстве много баров. Решением стало разбиение наименования так, чтобы последнее слово перед "T/A", то есть Villars, являлось главным ключом.

Приложение к главе 6 – Упражнение по состыковке

Это упражнение содержит пять примеров, в которых новые записи были автоматически сопоставлены с имеющимся массивом записей. Явные соответствия не были обнаружены, однако для проверки специалистам переданы пять возможных соответствий с самыми высокими баллами. Данные являются реалистичными, но в действительности не существуют. Пожалуйста, выберите наилучшее соответствие для новой записи. Наоборот, если ни одно из возможных соответствий не кажется достаточно близким, Вы можете принять решение об отсутствии соответствия. Ответы приведены после примера 5.

Пример 1

Новая запись		Возможные соответствия	
Название:	Bob the Butcher	1	Bob Daley Butchers 17 Barwick Green Sidcup Kent DA15 8HP
Адрес:	16 "Lawrence Street Southfleet Gravesend	2	Brian Dunn Brians Family Butchers 16 Pembroke Close Pembroke Street Dover Kent DA6 1FB
Почт. код:	DA11 7ZP	3	Mr B Dunn and Mrs V Dunn Brian's Family Butcher Pembroke Street Gravesend Kent DA6 1AA
		4	B & B Butchers Mr B Jones 3 Clive Road Dartford Kent DA1 5RH
		5	B Washbrook Bob the Butcher 16 Lawrence Drive Castle Lane Southfleet Gravesend Kent DA11 7ZF

Пример 2

Новая запись		Возможные соответствия	
Название:	Cars of Southfleet	1	Fleet Motors 31-35 Old Dover Road Dartford Kent DA15 7JF
Адрес:	3-5 Old Hill Southfleet Dartford	2	Southwold Cars 1A Southwold Close Greenhithe Kent DA23 9BC
Почт. код:	DA1 9KT	3	Mr D Crane T/A Southeast Cars 12A Old South Road Greenhithe Gravesend Kent DA2 9BN
		4	Mr C James & Mr G Smith Fleet Motors 29-35 Old Dover Road Fleet Kent DA15 9XX
		5	Southfleet Cars 33 Old Hill Southfleet Dartford Kent DA1 9XT

Пример 3

Новая запись		Возможные соответствия	
Название:	Retail Co-operative Limited	1	Mr A Cooper Paintcraft Unit 132 Greenway Estate Lower Station Lane Welling Kent DA18 6GT
Адрес:	35, Station Parade Station Road Dartford	2	Retail Co-op Ltd 030001 35 Station Street Dartford DA1 7DH
Почт. код:	DA1 7ED	3	Co-operative Funeral Services 362 Longfield Street Dartford DA1 1HD
		4	Co-operative Funeral Services Ltd, CFS (No14) Ltd & CFS Pension Fund 29 Station Street Bexleyheath Kent DA32 4RH
		5	Arts Co-operative 62 Highfield Street Dartford DA21 8JD

Пример 4

Новая запись		Возможные соответствия	
Имя:	Dr James Johnson	1	Mr James John Cunningham
Адрес:	Griffons Penny Lane Eynsford Dartford		35 Griffin Drive Darenth Dartford Kent
Почт. код:	DA46 8FF		DA4 6FF
		2	Mr John Jameson 56 Whinfell Road Gravesend Kent DA21 8GF
		3	Mr James Johnson 123 Penny Lane Aynsford Kent DA46 3JF
		4	John James 23 Perry Lane Dartford Kent DA28 3PF
		5	Mr James John Smith 18 Cornfield Lane Eynsford Dartford Kent DA46 8FF

Пример 5

Новая запись		Возможные соответствия	
Название:	Redipure Ltd	1	Redipure Limited Perseverence House 36A Cross Road Howley Dartford Kent DA27 8RR
Адрес:	26A Queens Rd Welling	2	Eradicure Ltd Perseverence House Cross Rd Howley Dartford Kent DA27 8RT
Почт. код:	DA13 8RS	3	Redpull Ltd 152 Lower Wickham Lane Wellington Kent DA13 8ED
		4	Redpull Ltd 12 Lower Wickham Welling Kent DA13 3ED
		5	Redipure Holdings Ltd Crossroads Howley Dartford DA12 3LF

Ответы

Это упражнение показывает, что 100-процентная определенность при установлении соответствия бывает редко. В нижеприведенных ответах приводится соответствие, которое является наиболее вероятным, по мнению экспертов по стыковке.

Пример 1 – наиболее вероятным является соответствие с существующей записью под номером 5. Коммерческое обозначение в этой записи соответствует названию в нашей новой записи, а адреса являются достаточно близкими. В почтовом коде различается один символ: “P” вместо “F”, что легко может быть ошибкой написания в одной из записей.

Пример 2 – наиболее вероятно соответствие снова с существующей записью номер 5. Названия и адреса достаточно близки, и, как и в примере 1, в почтовом коде имеется только один различающийся символ. Этот пример демонстрирует интересный проблемный момент: здесь существующие записи 1 и 4 также могут ставиться в соответствие. Это может означать наличие дублирования в существующих записях и говорит о важности периодического сопоставления массива данных с самим собой с целью снижения риска такого дублирования.

Пример 3 – ближайшим является соответствие с существующей записью номер 2. Главное отличие касается использования аббревиатур в названии в существующей записи (Ltd – Limited; Co-op – Co-operative). Это означает, что автоматический метод сопоставления недостаточно хорош при привязывании аббревиатур к их полным версиям. Такие как здесь аббревиатуры обычно являются специфическими для языков и даже для массивов данных, что свидетельствует о важности иметь возможность настраивать инструменты автоматической стыковки в соответствии с типами стыкуемых данных.

Пример 4 – ближайшим является соответствие с существующей записью номер 3. Запись номер 5 представляет собой точное соответствие в большей части адреса, а также в почтовом коде и при автоматической стыковке могло бы получить больший балл. Это иллюстрирует риски установления слишком низкого порога позитивного соответствия.

Пример 5 – на основе лишь имеющихся данных не похоже, что здесь имеется соответствие. Однако этот случай иллюстрирует важность использования дополнительной информации при проведении стыковки вручную. Аббревиатура “Ltd” в названии новой записи показывает, что это компания с ограниченной ответственностью. Во многих странах общества с ограниченной ответственностью по закону должны иметь уникальные наименования. Это означает, что если задача состоит в сцеплении единиц одного предприятия, то новая запись должна быть привязана к существующей записи номер 1. Разные адреса могут относиться просто к различным местоположениям (местным

единицам или заведениям), в которых осуществляет деятельность данная компания. Поэтому повышению процента автоматической пристыковки может помочь улучшение технологии автоматической стыковки в части распознавания корпоративных бизнесов и большего акцентирования на названиях в таких случаях. Эта стратегия успешно внедрялась в технологию стыковки, используемую в статистическом бизнес-регистре Соединенного Королевства.

7. Использование административных данных в статистических регистрах

7.1 Введение

В предыдущих главах рассматривались различные аспекты получения доступа к административным данным и обеспечения их пригодности для использования в статистических целях. Многие их этих вопросов относятся к повседневному ведению статистического регистра, однако они не будут повторяться здесь. Вместо этого данная глава рассматривает пути вовлечения административных данных в процесс статистического производства посредством их интеграции в статистические регистры. Сначала она дает определение статистических регистров, а затем рассматривает различные модели, которые использовались для интеграции административных данных.

7.2. Определение Статистического Регистра

Существуют различные определения регистров, впрочем, часто имеющие общие черты. Одним из наиболее широко используемых является следующее:

«Регистр – фиксируемый в записях и полный учет, охватывающий регулярно вводимые данные и характеристики по определенной совокупности объектов»[33].

Обычно регистр представляет собой какой-либо вид структурированного перечня единиц, содержащий ряд характеристик для каждой из этих единиц и имеющий какой-либо механизм регулярного апдейтирования. Таким образом, многие файлы административных данных могут считаться регистрами, но результаты единовременного сбора данных – нет.

Можно было бы утверждать, что там, где статистические данные производятся непосредственно из одного административного источника, этот источник не должен рассматриваться как регистр, также как и результаты обследования или даже переписи обычно не рассматриваются как регистры. Этот аргумент является более весомым в том случае, когда административные данные используются в форме агрегированных данных, а не индивидуальных пообъектных данных.

Статистический регистр является регистром, который сформирован и поддерживается для статистических целей, в соответствии со статистическими концепциями и определениями и под контролем статистиков. Поэтому

[33]"Терминология по статистическим метаданным", ЕЭК ООН / Конференция Европейских статистиков. Статистические стандарты и исследования, №. 53, Женева, 2000:http://www.unece.org/stats/publications/53metadaterminology.pdf.

административные регистры могут использоваться как источники для статистических регистров, но обратный случай, как правило, должен рассматриваться бы как противоречащий принципу «одностороннего потока» данных[34].

Как правило, статистический регистр играет роль инструмента координации данных, интегрирования данных из нескольких источников, как статистических, так и административных. Это может осуществляться путем связывания данных с использованием общих идентификаторов, либо используя что-то вроде методов сцепления, описанных в главе 6. Иногда может быть легче использовать данные из единого источника, но в таких случаях часто бывает трудно проверить точность этого источника. Когда используется несколько источников, и они интегрированы в рамках статистического регистра, возможно иметь лучшее представление о точности данных. К сожалению, отрицательная сторона – это то, что становится необходимым иметь стратегию работы с конфликтующими данными из различных источников. Однако если показатели в статистических регистрах хранятся с кодами источников и датами, можно использовать автоматизированные алгоритмы для определения приоритета источников и разрешить большинство противоречий в данных.

Наряду с интегрированием данных из различных источников, статистический регистр может также дать возможность выведения новых показателей. Одним из примеров является использование в некоторых странах [35] данных по организационно-правовой форме, виду экономической деятельности и иностранной собственности из их статистических бизнес-регистров для формирования институциональных секторов[36], используемых в национальных счетах.

Традиционно статистические регистры использовались как основа выборки для обследований, но все больше и больше они рассматриваются как источники статистических данных сами по себе, в частности, применительно к данным по малым географическим областям или малым подгруппам населения. Статистические регистры могут также обеспечить основу для связывания данных из различных источников в динамике по времени, делая возможным изучение явлений во времени. Этот подход уже использовался в нескольких странах для обеспечения изучения когорт людей или предприятий.

---

[34] [34]Смотри Фундаментальные принципы официальной статистики (принцип 6):http://www.unece.org/stats/archive/docs.fp.e.htm

[35] Примером является Австрия – 'Bericht über die Einführung der Sektorklassifikation im Unternehmensregister der Statistik Austria', авторы Karl Schwarz, Roland Schaumann и Thomas Karner.Эта работа содержит резюме на английском языке и доступна в Интернете на 'BR-Net' сайте Евростата с ограниченным доступом.

[36]Смотри: "Система национальных счетов 2008", глава 4:http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf

7.3 Модели создания и ведения статистических регистров с использованием административных данных

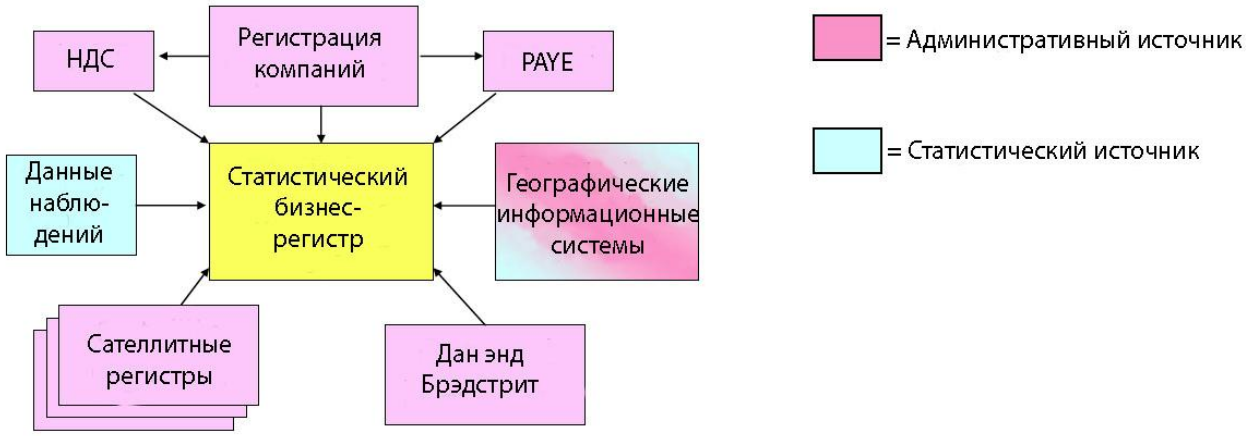
Как отмечено выше, статистические регистры играют важную роль в координировании данных из различных источников. Существует много способов, в рамках которых эти источники могут быть использованы или скомбинированы для формирования выборок и производства статистических данных. В данном разделе рассматриваются некоторые из подходов, использованных в разных странах или для различных областей статистики.

Так как доступные источники значительно различаются от страны к стране, зачастую бывает трудно экспортировать модель или сформулировать международные стандарты. Поэтому, различные приведенные ниже модели не следует рассматривать как рекомендации, которые должны быть реализованы во всех странах, а больше как примеры, показывающие, как другие страны использовали административные данные в статистических регистрах. Замысел заключается в том, чтобы предоставить не готовые решения, а идеи, которые можно адаптировать к конкретным национальным обстоятельствам.

1) Объединение нескольких источников

Рисунок 7.1. внизу является упрощенной моделью источников, используемых для поддержания статистического бизнес-регистра в Соединенном Королевстве. Он преднамеренно показывает статистический регистр в центре как инструмент для объединения и согласования данных из различных источников. Он также представляет концепцию сателлитных регистров, которая будет подробно обсуждаться далее в этой главе, а также идею того, что источники могут быть комбинацией административных и статистических данных. Так, географическая информационная система (GIS) уже содержит комбинацию административных данных (главным образом полученных из почтовой службы) с некоторым статистическим моделированием с использованием данных переписи населения для создания более статистически однородных областей.

Рисунок 7.1 – Упрощенная модель источников статистического бизнес-регистра в Соединенном Королевстве



2) Использование централизованных административных регистров

Централизованные административные регистры часто создаются для повышения эффективности внутри правительства, и во многих случаях они обеспечивают единый интерфейс, посредством которого субъекты регистра могут взаимодействовать с различными правительственными организациями, уменьшая, таким образом, дублирование, и, следовательно, нагрузку в связи с выполнением требований административных процедур. Например, если человек или предприятие меняет адрес, то при существовании такого регистра им нужно предоставить данные об их новом адресе только однажды, и затем эти данные будут сообщены всем соответствующим организациям.

Этот вид административного регистра может быть очень полезен для статистических целей, так как он снимает, по крайней мере, до некоторой степени нагрузку по стыковке и согласованию данных из различных источников. Однако для максимизации этой полезности статистической организации важно иметь некоторое влияние при разработке и управлении административным регистром, чтобы обеспечить его максимально возможное соответствие статистическим потребностям в части единиц, классификаций, определений и процедур.

Хорошим примером того, как этот подход сработал на практике, является использование (административного) Австралийского бизнес-регистра (ABR)[37] Австралийским бюро статистики. ABR был разработан Австралийской налоговой службой для администрирования различных налогов с предприятий, однако поддерживается в тесном сотрудничестве с Австралийским бюро

[37] См.: http://www.abr.gov.au

статистики, которое вносит свой вклад и знания в определенных областях, таких как классификация экономической деятельности.

В результате ABR является удобной основой для статистического бизнес-регистра по всем, за исключением самых крупных и наиболее комплексных, предприятий. Фактически статистический регистр предприятий отчетливо имеет двух-уровневый подход[38]. Большинство записей являются копиями из ABR, и они поддерживаются из этого источника, оставляя статистическим ресурсам свободу сконцентрироваться на ведении структур самых больших и комплексных предприятий.

3) Создание хабов совместного пользования

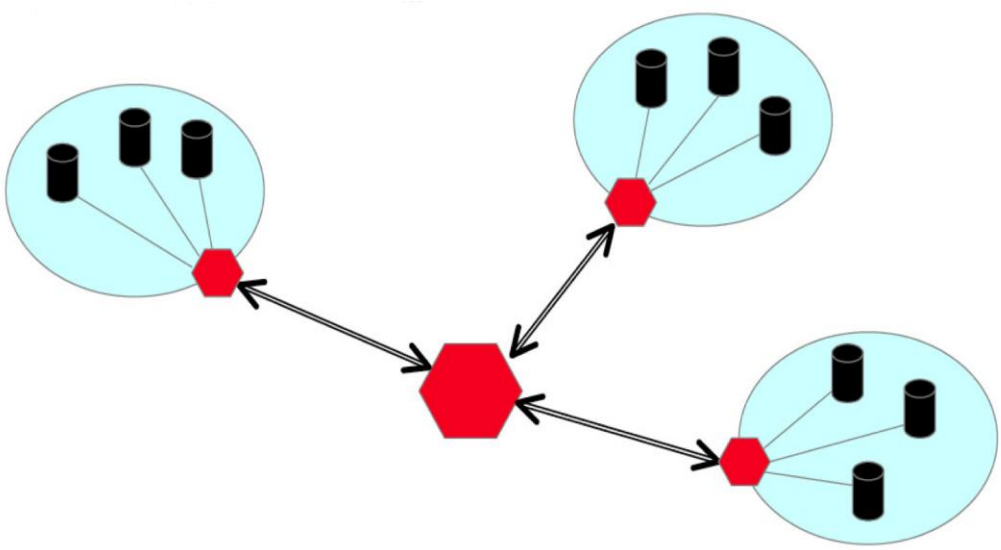
Разновидностью идеи о едином централизованном регистре является концепция хабов совместного пользования. В этой модели центром является не полностью сформированный регистр, а скорее инструмент для нахождения и подбора подходящих данных, хранимых в различных организациях. Он может содержать некоторые самые базовые идентификационные данные, но его главным предназначением является предоставить межсетевой интерфейс, посредством которого данные из различных организаций могут совместно использоваться внутри правительственного сектора.

Рисунок 7.2. взят из исследования осуществимости такого подхода в Соединенном Королевстве[39]. Этот подход не был реализован, но модель остается допустимым вариантом для совместного пользования административными данными. Голубые круги представляют различные правительственные организации, каждая с некоторым количеством фондов данных (черные цилиндры). Каждый из этих фондов данных привязан к порталу, который жестко контролирует, что и кому может пройти через него. Эти порталы в свою очередь привязаны к центральному хабу, содержащему достаточно метаданных, чтобы обеспечить поиск и установление соответствия по привязанным фондам данных. При этом подходе пользователь в одной из участвующих организаций может посылать запрос через центральный хаб и может получать данные из всех имеющихся фондов данных в других организациях, к которым этот субъект имеет права доступа.

[38]Для большей информации см.: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/8165.0Explanatory%20Notes1Jun%202007%20to%20Jun%202009?OpenDocument

[39]Для большей информации см.: http://www.unece.org/stats/documents/ces/sem.46/5.e.pdf

Рисунок 7.2 – Хаб совместного пользования данными



4) Использование административных данных посредством сателлитных регистров

Несколько отличной моделью использования административных данных на практике является организация их в специализированные применительно к источникам регистры, связанные со статистическим регистром. Если эти специализированные применительно к источникам регистры отвечают определенным критериям, они могут быть названы «сателлитными регистрами»[40]. Сателлитные регистры можно определить как регистры, которые доступны для национальной статистической системы, содержат информацию о представляющих интерес единицах и показателях и соответствуют следующим условиям.

- Они не являются неотъемлемой составной частью статистического регистра, но способны к установлению связи с ним.
- Их сфера более ограничена, чем у статистического регистра, но в пределах их сферы они могут иметь более исчерпывающее покрытие единиц и / или показателей.
- Они содержат один или более показателей, которых нет в статистическом регистре. Такие показатели обычно могут использоваться для целей стратификации.
- Базы данных, в которые обычно записываются результаты обследований, не являются сателлитными регистрами.

[40]Их также называют “ассоциированными регистрами”.

Поэтому сателлитные регистры являются средствами для включения административных данных, которые относятся только к разновидности единиц в статистическом регистре. Они могут содержать дополнительные единицы, либо показатели, либо и то и другое. Они могут быть созданы с использованием информации из административных источников, статистических обследований или комбинации того и другого. В некоторых случаях они могут добавить, сочетать или иным образом трансформировать показатели, тогда как в других они могут быть более или менее идентичны тому или иному источнику. Для обеспечения достаточной взаимосвязи сателлитных регистров со статистическими регистрами полезно продумать дополнительные критерии, например, общие идентификаторы единиц, общие определения и классификации. Чем сильнее взаимосвязь, тем, вероятно, более полезен будет сателлитный регистр.

Рисунок 7.3 – Связь между сателлитным регистром и статистическим регистром

Рисунок 7.3 показывает, как сателлитный регистр соотносится со статистический регистром. Эту диаграмму можно интерпретировать как в терминах охваченных единиц, так и содержащихся показателей. В обоих случаях есть некоторая степень перекрытия, но сателлитный регистр также дает дополнительную информацию, либо дополнительные единицы, либо дополнительные показатели для подмассива существующих единиц.

Наиболее современные примеры сателлитных регистров относятся к бизнес-данным, где охват сателлитного регистра можно определить через:

• экономическую деятельность – сателлитный регистр может содержать предприятия определенного вида деятельности, например, розничная торговля, гостиницы, грузовые перевозки, и т.д.;

• размер – сателлитный регистр может содержать единицы с определенным числом работников или оборотом выше определенного уровня, например, подмножество «больших предприятий»;

• характеристики – сателлитный регистр может содержать единицы с общими характеристиками, например, те, которые вовлечены во внешнюю торговлю.

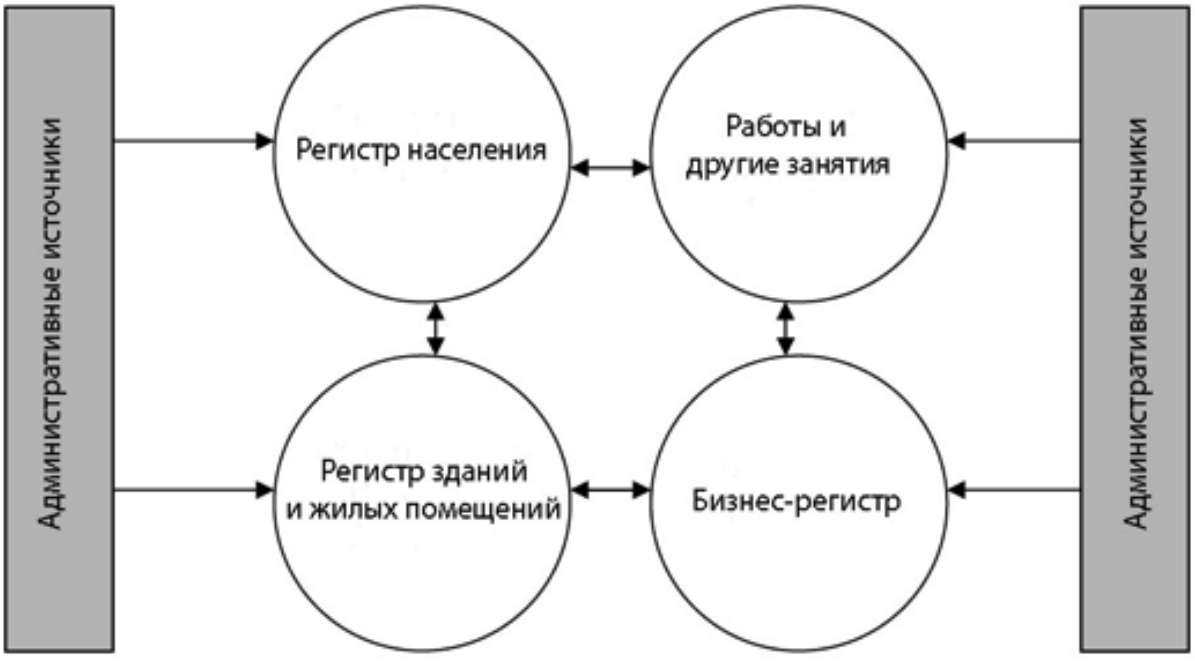
Примеры показателей, характерных для подмножества единиц, включенных в сателлитный регистр, могут включать «категорию» или «количество спальных мест» для гостиниц, или «торговую площадь» для предприятий розничной торговли.

Сателлитные регистры могут усиливать важность статистических регистров путем увеличения диапазона показателей, доступных для целей стратификации или анализа, и повышать эффективность выборки путем повышения качества стратифицирующих показателей. Они также могут увеличивать покрытие целевой совокупности, и в некоторых случаях могут уменьшить количество информации, которую необходимо собрать путем статистических обследований, уменьшая, таким образом, нагрузку на респондентов.

5) Статистические системы, основанные на регистрах

Статистические системы, основанные на регистрах, обсуждаются далее в главе 9, но упомянуты здесь постольку, поскольку они представляют собой модель использования административных данных в статистических регистрах. Основным отличием, по сравнению с описанными выше моделями, является то, что создаются несколько связанных регистров, использующих широкий диапазон административных данных. Эта модель разрабатывалась, главным образом, в скандинавских странах, используя либо три, либо четыре основных статистических регистра. Рисунок 7.4 показывает упрощенную версию модели, принятой в Швеции.

Рисунок 7.4 - Североевропейские статистические системы, основанные на регистрах



Статистический регистр населения привязан к земельному реестру или регистру недвижимости, а также к статистическому бизнес-регистру, используя уникальные идентификаторы населения, земельной собственности и предприятий. В Швеции был введен четвертый регистр, содержащий информацию о профессиях и других видах деятельности. Этот регистр привязывает население к их источникам доходов, включая зарплаты, пенсии и платежи по социальному страхованию, и поэтому отражает взаимоотношения между населением и рынком труда.

Приложение к Главе 7 – Упражнение: создание статистического регистра предпринимателей

Ваше правительство решает, что ему нужно больше данных о предпринимателях, и о факторах, определяющих, являются ли они успешными или нет. Ваша служба решает производить новые ряды данных для предоставления этой информации. Вас попросили создать статистический регистр предпринимателей на основе административных источников для использования в качестве основы выборки.

Ваш годовой бюджет составляет 16000 евро. Стоимость обработки каждого источника данных, которым вы пользуетесь, составляет 2000 евро. Помимо этого, имеют место затраты на покупку данных, которые варьируют от источника к источнику.

Вам доступны следующие административные источники:

1. Данные налоговой службы по людям, которые декларируют доход от самозанятости

- Содержание: идентификационный номер физического лица, имя, адрес, пол, сумма декларированного дохода, наименование бизнеса, вид бизнеса (классифицированный на уровне двух знаков по Международной стандартной отраслевой классификации (ISIC)).
- Доступность: налоговая служба будет предоставлять эти данные ежегодно при условии, что Вы платите за услугу 2500 евро в год, чтобы покрыть ее расходы по извлечению и посылке данных. Они будут посылать данные на CD-ROM.
- Качество: данные имеют точность 95%, исключая «вид бизнеса», где точность составляет только 50%. Когда Вы получите данные, они будут устаревшими на 6-18 месяцев. Покрытие составляет 100% всех людей, занимающихся легальным бизнесом. По оценке, деятельность порядка 20% бизнесов ведется неправомерно (т.е. людьми, которые не декларируют свой доход).

2. Данные налоговой службы по бизнесам с наемными работниками

- Содержание: идентификационный номер бизнеса, наименование и адрес бизнеса, количество наемных работников, вид бизнеса (классифицированный по ISIC на уровне 4-х знаков), год, когда бизнес впервые зарегистрирован как работодатель.
- Доступность: налоговая служба будет предоставлять эти данные при условии, что вы платите ежегодно 3000 евро на покрытие ее расходов по извлечению и посылке данных.

- Качество: данные точны на 90% и обычно являются устаревшими на 2-3 месяца. Данные будут посылаться ежемесячно на CD-ROMs. Они покрывают все бизнесы, которые нанимают работников легально. По оценке, 50% бизнесов имеют работников, и 95% из них ведут деятельность легально.

3. Административный регистр населения

- Содержание: идентификационный номер физического лица, имя и адрес, возраст, пол, уровень образования, занятие, национальность, страна рождения.
- Доступность: эти данные уже используются статистической службой при годовой стоимости 3000 евро. Если Вы будете использовать их, ожидается, что Ваши расходы составят половину этой суммы. Данные доступны ежегодно, и Вы можете получать их в форме электронного файла от Ваших коллег из отдела по статистике населения.
- Качество: данные точны на 95% и являются устаревшими на 1-2 года. Они покрывают 99% легального населения, но оценивается, что порядка 5% всего населения – нелегальные иммигранты, так что они не покрыты данными.

4. Телефонный справочник предприятий («Желтые страницы»)

- Содержание: наименование и адрес предприятия, телефонный номер, вид бизнеса (классифицированный в соответствии с их собственным перечнем из 300 позиций).
- Доступность: эти данные продаются компанией частного сектора. Они доступны ежемесячно на CD-ROM. Стоимость годовой подписки составляет 7000 евро, однако поставщик готов предоставить статистической службе 15%-ю скидку.
- Качество: поставщиком утверждается, что данные точны на 99%, он говорит, что в интересах предприятий следить за тем, чтобы информация была корректной. Данные обычно 1-2 месячной давности. Они покрывают порядка 85% всех бизнесов (легальных и нелегальных).

5. Список лиц, подающих на грант на создание нового бизнеса

- Содержание: идентификационный номер физического лица, имя и адрес, идентификационный номер бизнеса, наименование и адрес, вид бизнеса (классифицированный по ISIC на уровне 2-х знаков).
- Доступность: 500 евро за таблицу, посылаемую по электронной почте каждый март, отражающую заявления на грант за предшествующий год.
- Качество: данные точны, по крайней мере, на 95%, хотя некоторые адреса бывают устаревшими. Примерно 40% людей, начиная новое дело, подают заявление на грант для становления дела, однако обычно это наиболее

успешные предприниматели. Это составляет 6% всей совокупности бизнесов за любой данный год.

5. Список членов «Национального общества предпринимателей»

- Содержание: имя и адрес физического лица, наименование бизнеса, адрес и номер телефона, дата вступления в Общество.
- Доступность: 100 евро за бумажный справочник, публикуемый ежегодно.
- Качество: данные точны, по крайней мере, на 90%, хотя некоторые адреса могут быть устаревшими. Членские взносы довольно высокие, так что лишь примерно 10% предпринимателей являются членами. Это, в большинстве своем, люди с успешным бизнесом, действующим, по крайней мере, 5 лет.

Вопросы:

1. Принимая в расчет ограниченный бюджет (16000 евро), какие источники Вы бы выбрали?
2. Почему Вы бы выбрали эти источники?
3. Как бы Вы состыковывали данные из различных источников?
4. Какой тип обследования Вы бы порекомендовали – личное интервью, телефонное интервью или почтовый вопросник?
5. Какие показатели Вы бы использовали для стратификации выборки для обследования?

Ответы:

На самом деле не существует правильных или неправильных ответов для этого упражнения, однако, факторы, которые следует рассмотреть, включают:

- Источники 1-3 – типичные административные источники государственного сектора, они имеют хорошее покрытие, но только легально зарегистрированных хозяйственных единиц.
- Источник 4 – это характерный пример разновидности источника административных данных частного сектора, который все более и более пытаются применять для статистических целей во многих странах. Учитывая возможность договариваться о цене, существует вероятность дальнейшего снижения стоимости и может пригодиться опыт ведения переговоров по коммерческим контрактам!
- Источники 5 и 6 могли бы рассматриваться как типичные сателлитные регистры, поскольку они имеют ограниченное покрытие, но фокусируются на отдельной группе населения, которая может иметь отличные характеристики от населения в целом.
- Покрытие, своевременность, точность и стоимость обработки должны рассматриваться как часть анализа затрат-результатов для каждого источника.

- Было бы полезно иметь больше информации о требованиях пользователей к итоговым статистическим данным, так как это может повлиять на выбор источников. Опытные статистики будут учитывать, что требования обычно бывают весьма нечеткими, по крайней мере, первоначально, поэтому дальнейший диалог с пользователями бывает полезен. Вопросы для прояснения могли бы включать:
  - Следует ли фокусироваться на бизнесах, которые создают рабочие места, или на количестве предпринимателей?
  - Каков требуемый баланс между своевременностью и точностью?
  - Есть ли какая-либо заинтересованность пользователей в усилиях дать оценки для предпринимателей, ведущих деятельность в неформальной экономике? Если это так, может быть нужен источник 4, возможно, в комбинации с источником 1.

Вопросы 4 и 5 являются до некоторой степени каверзными, ибо прежде следует ответить, нужно ли действительно проводить обследование, или требуемые данные можно произвести непосредственно из статистического регистра, созданного путем объединения выбранных источников.

8. Использование административных данных в дополнение к статистическим обследованиям

8.1 Введение

В этой главе представлен обзор различных моделей использования административных данных в дополнение к данным, собранным в ходе статистических обследований. Она показывает, как подход с использованием смешанных источников может использоваться для производства статистики с меньшими затратами, лучшим качеством, либо и тем и другим.

Многие вопросы, относящиеся к использованию и взаимной привязке статистических и административных данных, уже были освещены в главах 4 и 6, поэтому они здесь не повторяются. Вместо этого данная глава фокусируется на различных моделях использования данных, полученных в результате сочетания административных и статистических источников, для производства выходных статистических данных.

8.2 Модели смешанных источников

1) Метод разбиения генеральной совокупности

В этой модели статистическая совокупность для целей сбора данных разбивается на две или более частей. Этот подход весьма близок к тому, который используется для поддержания австралийского статистического бизнес-регистра, как описано в главе 7.3. Данные из административного источника используются для единиц, по которым эти данные имеют достаточное качество, а статистические источники используются для остального массива единиц.

Типичным сценарием обследования бизнесов является следующий: данные по относительно небольшим бизнесам с простой структурой берутся или выводятся по налоговой декларации, тогда как обследования применяются для сбора данных по ключевым единицам (обычно самым большим и/или имеющим наиболее сложную структуру). Для той части генеральной совокупности, для которой используются налоговые данные, статистические и административные единицы, по всей вероятности, идентичны или весьма близки, и влияние различий между статистическими методологиями и классификациями и их административными аналогами, вероятнее всего, является минимальным, или может быть легко смоделировано.

Оставшуюся часть бизнесов обычно составляют те, индивидуальное влияние которых на качество статистики наиболее велико, и поэтому они являются теми, для которых наиболее важно иметь точные данные. Также вероятно, что эти единицы имеют наиболее сложные структуры, часто требующие

профилирования (как описывается в главе 4.5) для того, чтобы правильно определить статистические единицы, по которым нужны данные. Эти статистические единицы часто являются объединениями административных единиц, либо их частями; и хотя некоторые показатели – такие как занятость – обычно могут просто суммироваться для получения корректного итога, другие – такие как объем продаж и некоторые иные финансовые показатели – не могут, поскольку результат будет включать какой-то объем продаж внутри единицы, поэтому простое суммирование приведет к завышению показателя.

Практическим примером применения метода разбиения генеральной совокупности при обследовании бизнесов является Унифицированное обследование предприятий, проводимое Статистикой Канады. В нем унифицированы требования к годовым бизнес-данным и объединены несколько имевшихся ранее обследований. Административные данные используются вместо данных, собираемых посредством статистических вопросников, по более чем половине покрытых обследованием предприятий, – тем, которые имеют простую структуру, – что приводит к снижению почти на 40% нагрузки, связанной с представлением ответов на статистические вопросники[41].

В случае, когда статистическая совокупность представлена людьми или домашними хозяйствами, может потребоваться обследование определенных групп – таких как студенты, работники-мигранты, либо имеющие два или более местожительств. Все они – потенциальные примеры единиц, по которым административные данные могут не быть достаточно современными, либо точными, особенно в части местонахождения.

Как несколько раз отмечалось в предыдущих главах, в учет должны также приниматься единицы, не покрытые административными регистрами, такие как нелегальные мигранты или бизнесы, действующие в неформальной экономике.

Статистические обследования таких групп могут использоваться лишь ограниченно, поэтому может потребоваться оценка, что означает подключение третьего источника, используемого в производстве требуемой статистики. Эта модель иллюстрируется ниже на рисунке 8.1.

[41]Для большей информации см. работу Marie Brodeur из Статистики Канады “Использование налоговых данных в Унифицированном обследовании предприятий (UES)”: http://unstats.un.org/unsd/economic_stat/Moscow_workshop/Canada%20-%20Use%20of%20tax%20data%20in%20the%20UES-E.pdf

Рисунок 8.1 – Модель разбиения генеральной совокупности

2) Метод разделения данных

При этом подходе определены генеральная совокупность статистических единиц и требования к данным, например, генеральной совокупностью могут быть все физические лица, проживающие в отдельной стране, а требование к данным может состоять в обычном наборе показателей, требующихся при переписи населения. Вместо предоставления всех показателей по части совокупности, как в вышеупомянутой модели разбиения генеральной совокупности, при методе разделения данных административные источники используются для предоставления некоторых показателей по всей совокупности (третий подход также возможен там, где административные источники предоставляют некоторые показатели по некоторой части совокупности).

Поэтому при методе разделения данных не уменьшается количество вопросников или интервью, требуемых для сбора данных, но уменьшается объем данных, которые необходимо получить в каждом вопроснике или интервью. Это обычно относится к большим и сложным наборам данных, когда требуется много показателей, примером чего является перепись населения. Чтобы сформировать массив данных, используемый для получения выходного статистического продукта, для каждой индивидуальной единицы необходимо интегрировать административные данные и данные обследования.

Метод разделения данных часто используется как переходный к некому виду статистической системы, основанной на регистре, описываемой в следующей главе. Как правило, показатели, получаемые путем статистического сбора данных, замещаются их эквивалентами из административных источников в течение ряда отчетных периодов. Нижеприведенная таблица 8.1 иллюстрирует этот процесс, показывая источники данных финской переписи населения и жилищ.

Таблица 8.1 – Метод разделения данных при переписи населения и жилищ в Финляндии в 1960-2000 гг.

	1960	1970	1980	1990	2000
Демографические данные	O	O/P	P/O	P	P
Экономические данные	O	O/P	O/P	P/O	P/O
Образовательные данные	O	O	P	P	P
Данные по домашним хозяйства и семьям	O	O	P	P	P
Данные по проживанию	O	O	O	P	P
Данные по производственным помещениям	O	O	P	P	н/с
Данные по зданиям	O	O	O	P	P
Данные по летним коттеджам	O	O/P	O/P	P	P

Обозначения: O – статистический опросник

O/P – статистический опросник, дополненный административным регистром

P/O – административный регистр, дополненный статистическим опросником

P – административный регистр

н/с – не собраны

Источник: эта таблица является сжатой версией приложения 2 к документу “Использование регистров и источников административных данных для статистических целей – передовой опыт Статистики Финляндии”: http://unstats.un.org/unsd/EconStatKB/KnowledgebaseArticle10169.aspx

3) Предварительно заполненный вопросник

Этот метод на самом деле является частным случаем модели разделения данных, в которой для сбора данных о статистических единицах все еще

используются статистические вопросники, но эти вопросники предварительно заполнены с использованием административных данных насколько это возможно, а респондентов просят просто проверить и поправить эти данные, где это необходимо. На рисунке 8.2 показан фрагмент предварительно заполненного вопросника; этот пример взят из британского статистического обследования предприятий.

Рисунок 8.2 – Фрагмент предварительно заполненного вопросника

2. **Вид деятельности Вашего предприятия** (см. указание 5)
По нашим данным видом Вашей деятельности является:

Производство деревянной мебели

2а. Если Ваша текущая деятельность является другой, пожалуйста, подробно опишите ее в нижеприведенном шаблоне (одна буква в квадратике с пробелами между словами).
Приведите основные товара или услуги. Включите информацию об использованных продуктах и материалах (например, производство окон из НПВХ). Если Вы предоставляете услуги, опишите их подробно (например, лицензированный отель и ресторан).

Предварительно заполненные опросники имеют три главных достоинства:

- они помогают уменьшить нагрузку на тех, кто предоставляет данные, путем экономии их времени (предполагается, что проверка и корректировка осуществляются быстрее, чем отыскание и ввод данных);
- они обеспечивают проверку качества административных данных;
- они являются очень гибкими, в частности, когда покрытие административными данными не настолько полное, как это необходимо для других моделей.

Главный их недостаток – риск внесения смещения, так как некоторые респонденты могут просто соглашаться с предварительно заполненными данными без их проверки, либо могут решить не тратить время на коррекцию ошибок.

4) Использование административных данных при неответах

Этот метод может рассматриваться как вариант обеих моделей – как разделения источников, так и разделения данных. В этом случае статистическое обследование остается приоритетным средством сбора данных. Однако на статистические обследования негативно влияют различной степени неответы, которые воздействуют на эффективность процесса выборочных обследований и качество получаемых статистических результатов. Неответ обычно принимает одну из двух форм: «неответ по единице», при котором отсутствуют данные по интересующей единице, либо «неответ по показателю», при котором частичный ответ представлен, но некоторые пункты не заполнены.

Работа с неответами может оказаться очень затратной для статистического агентства, поскольку она обычно включает повторные попытки связаться по почте или телефону с целью получения отсутствующей информации. Этот процесс, обычно называемый «охотой за ответами», как правило, является очень трудоемким.

Более дешевой альтернативой может стать принятие решения о том, что при непредставлении данных к определенной дате, особенно по единицам, которые не являются критически важными для результатов обследования (например, по малым предприятиям в обследовании предприятий), данные берутся или формируются из административных источников. Это дает возможность сконцентрировать все ресурсы, предназначенные для охоты за ответами, на наиболее важных единицах, что делается для минимизации всяких смещений, обусловленных использованием административных данных вместо данных обследований. Как и по всем касающимся качества вопросам, неизбежен компромисс между затратами и различными аспектами качества (смотри главу 5).

Кроме того, административные данные могут иногда использоваться как основа для восстановления пропущенных данных обследований в привязанных к ним файлах данных[42].

5) Использование административных данных для оценивания

Когда для сбора статистических данных используется выборочное обследование, часто необходимо применять алгоритмы оценивания, особенно если требуется получить значения суммарных итогов по генеральной совокупности (а не значения долей). Поэтому нужна какая-то основа для оценивания значений по не включенной в выборку части генеральной совокупности. Иногда для этого применяются показатели из основы выборки, используемой для извлечения выборки, однако в некоторых случаях точность можно повысить, используя данные из административных источников в

[42]Для примера см. подход Бюро цензов США в главе 3 публикации: Модернизация обследования доходов и программы участия, http://www.nap.edu/catalog/12715.html

качестве вспомогательных показателей в процессе оценивания[43]. На практике многие примеры этого подхода касаются использования административных данных для улучшения оценок по малым областям[44].

8.3 Дополнительные соображения

В любой сложной системе статистической обработки, использующей многие источники, крайне важно учитывать роль метаданных, в частности, метаданных, относящихся к источнику каждого элемента данных. Это дает возможность обрабатывать элементы данных разными способами и посредством различных процессов (включая непредвиденные будущие процессы) в зависимости от путей их получения. Информация об источнике данных часто является эффективным показателем качества и может помочь при определении уровня качества выходного статистического результата.

Использование комбинированных – статистических и административных – данных может рассматриваться, с одной стороны, как результат сам по себе, особенно когда покрытие или качество административных данных не видятся достаточно высокими, чтобы вовсе прекратить сбор статистических данных. Его также можно рассматривать как ступень в постепенном переходе к статистической системе, основанной на регистрах, как это показано в Таблице 8.1.

В любом случае это дает возможность реализовать по крайней мере некоторые преимущества использования административных данных (включая экономию затрат) и в то же время избежать некоторые недостатки, такие как полная зависимость от внешнего источника и потеря контакта с общественностью. Это дает возможность сравнить качество статистических и административных данных, способствует освоению статистиками процессов использования административных данных и разработке новых методов повышения качества обработки.

В силу этих причин подходы с использованием комбинированных источников в настоящее время намного более распространены, чем целиком основанные на регистрах статистические системы, однако со временем доверие к административным данным, по-видимому, будет возрастать, способствуя распространению их использования и дальнейшей реализации преимуществ. Так как баланс все более сдвигается в сторону административных данных, в конечном итоге станет необходимо рассмотреть вопрос о целесообразности

[43]Для примера см.: Использование административных источников данных для [получения] литовских годовых данных о доходах, http://home.lu.lv/~pm90015/workshop2006/papers/Workshop2006_22_Slickute_Sestokiene.pdf

[44]Для примера см.: Использование административных записей для оценивания по малым областям при Обследовании американского общества, http://www.fcsm.gov/99papers/mcf.html

перехода на вариант модели, основанной на регистре и описанной в следующей главе.

9. На пути к статистической системе, основанной на регистре

9.1 Введение

Как отмечено в конце главы 8, если административные данные используются для разработки и поддержания статистических регистров, а также дополняют статистические обследования, следующим логическим шагом является изучение возможностей связать эти регистры и обследования и, таким образом, двигаться к основанной на регистре статистической системе. Этот подход в основном был разработан статистическими агентствами североевропейских стран и, как правило, первоначально предназначался для применения при основанной на регистре переписи населения.

Статистическая система, целиком основанная на регистре, может быть определена как система, в которой все статистические данные (по отдельной области статистики или группе областей) производятся исключительно из административных источников, которые объединены в два или более связанных статистических регистра. На практике такая статистическая система, целиком основанная на регистре, встречается сравнительно редко, поскольку небольшие статистические обследования зачастую необходимы для оценки качества или решения проблем покрытия применительно к специфическим показателям и секторам генеральной совокупности. Поэтому более прагматичным подходом является использование термина «статистическая система, основанная на регистре» применительно к системе, основанной преимущественно на административных данных, которые организованы в виде связанных статистических регистров.

В этой главе кратко рассматриваются некоторые вопросы, касающиеся перехода к статистической системе, основанной на регистрах. Целью является дополнить, но не повторить намного более детальное исследование этой темы в публикации Европейской экономической комиссии Организации Объединенных Наций “Основанная на регистрах статистика в странах Северной Европы” [45] . Эта публикация рассматривает передовой опыт с упором на статистику населения и социальную статистику, она была подготовлена специалистами из нескольких североевропейских стран, так что ее следует считать авторитетной работой по данному вопросу.

[45] См.: http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf

9.2 Реализуемость

Статистические системы, основанные на регистре, не являются реализуемыми во всех странах, или даже всех областях статистики, по крайней мере, в ближайшей перспективе. Это имеет место в силу того, что осуществимость разработки и внедрения такой системы зависит от нескольких непременных условий, имеющих отношение к политике и инфраструктуре, некоторые из которых в разных контекстах упоминались в предыдущих главах. Ключевыми предпосылками успешной статистической системы, основанной на регистре, являются следующие:

- Существование подходящих административных источников: необходимы полные административные регистры целевых совокупностей. Существование большого числа незарегистрированных единиц, например, нелегальных иммигрантов или действующих в неформальной экономике бизнесов, делает чрезвычайно трудным производство полноценной статистики на основе регистров.

- Легкость доступа – административные источники должны быть без затруднений доступны статистикам благодаря наличию оснований, описанных в главе 3. Это включает требование, чтобы они поддерживались в формате, облегчающем передачу данных.

- Общие идентификаторы – хотя глава 6 и показывает, что числовые общие идентификаторы для единиц, которые присутствуют в различных источниках, не являются абсолютно обязательными, они значительно облегчают объединение этих источников и потому существенно повышают эффективность производства статистики на основе регистров.

- Одобрение общественностью – как обсуждалось в главе 4.2, отношение широкой общественности к вопросу связывания и совместного использования данных внутри правительственного сектора является ключевым фактором, определяющим масштабы возможного использования административных данных для статистических целей. Баланс между эффективностью совместного использования данных и заботой о защите индивидуальных данных часто является причиной острой дискуссии с различными последствиями, зависящими от национальной культуры и традиций. В некоторых странах концепция основанной на регистре статистической системы в настоящее время признается неприемлемой для больших групп населения.

Если эти предпосылки отсутствуют, очевидна нереалистичность попыток использования основанной на реестре статистической системы в ближайшем будущем. Эта модель, тем не менее, может быть полезной в качестве

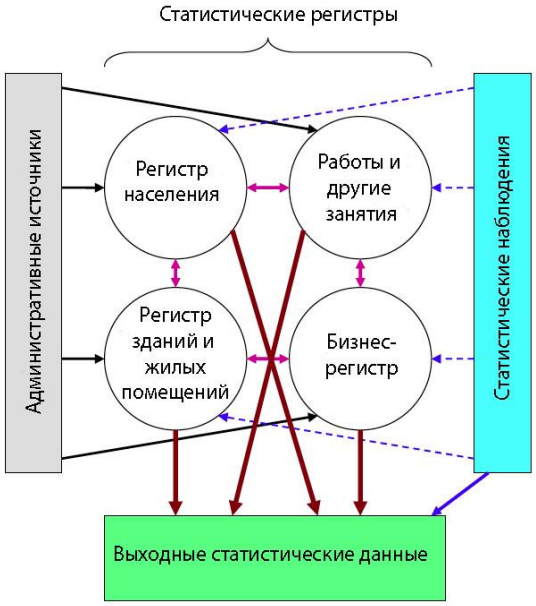
долгосрочной цели, достижимой путем осуществления поэтапной программы преобразований, направленных на формирование необходимых предпосылок. Опыт североевропейских стран подчеркивает важность долгосрочного планирования, ибо внедрение в этих странах переписей населения на основе регистра заняло около двадцати лет.

9.3 Общая модель

Глава 7.3 включает обсуждение основанных на регистре статистических систем в той степени, в какой они представляют собой модель использования административных данных в статистических регистрах. Рисунок 7.4 в той главе показывает общую модель статистической системы, основанной на регистрах, однако фокусируется только на административных входных данных. На нижеприведенном рисунке 9.1 эта модель модифицирована с целью включения статистические входных и выходных данных. Двумя ключевыми характеристиками являются следующие:

- Связи между основными статистическими регистрами (здесь могут быть и другие, более специализированные статистические регистры, но они не показаны на диаграмме для большей ясности).
- Баланс между административными источниками и данными обследования в статистических выходных данных (не существует четкого правила, определяющего каков этот баланс должен быть, но разумно ожидать, что административные источники являются основной входной информацией).

Рисунок 9.1 – Основанная на регистре статистическая система – общая модель



Замечание: статистический регистр работ и других занятий не всегда присутствует в национальных версиях этой модели.

9. 4 Резюме

Очевидно, что основанная на регистрах статистическая система является конечной целью при рассмотрении возможностей лучшего использования административных данных для статистических целей. Во многих странах эта цель видится как очень удаленная, возможно, недосягаемая в течение многих лет. Вместе с тем, приняв стратегический план, основанный на поэтапных улучшениях в направлении создания необходимых предпосылок, можно постепенно приближаться к этой цели.