Предлагаемая читателю работа известного специалиста в области статистики труда и международной миграции Эйвинда Хоффманна посвящена теме, чрезвычайно актуальной для России. Проблемы государственной статистики международной миграции в России, обострившиеся за последние годы, усугубляются отсутствием традиции открытой статистической отчетности административных структур, осуществляющих регистрацию мигрантов. Между тем обработка ведомственных баз данных для получения статистики, публикация результатов, критическое осмысление ошибок и поиск путей их устранения направлены на то, чтобы предоставить властям необходимую для принятия решений информацию, а обществу - возможность оценить результаты работы ведомства.*

Представленный в статье Эйвинда Хоффмана опыт Директората иммиграции Норвегии показателен в том смысле, что административные процедуры не являются единственной задачей работы этого института: обобщение данных, разработка статистических показателей и, главное, постоянное совершенствование этих процессов являются одной из его неотъемлемых функций. Несмотря на очевидные трудности использования записей о регистрации в качестве источника статистической информации, которые связаны и с конструктивными особенностями базы данных, и с противоречиями между интересами чиновников, осуществляющих регистрацию, и специалистов отдела статистики Директората иммиграции, ситуация представляется вполне оптимистичной. В работе хорошо обозначена мысль о том, что результаты обобщения и обработки административных процедур, то есть статистика, востребованы и властями, и обществом в целом. Это в свою очередь является стимулом к развитию методологии и инструментария сбора и обработки первичной информации.

Отметим, что открытая публикация подробных статистических отчетов иммиграционными ведомствами развитых стран является своего рода культурной традицией. Директорат иммиграции Норвегии ежегодно публикует[1] на норвежском и английском языках сборники «Факты и цифры», в которых содержится не только статистическая информация, но и аналитический обзор итогов работы Директората по основным направлениям.

В последние годы в Федеральной миграционной службе России ведется работа по созданию Центрального банка данных по учету иностранных граждан и лиц без гражданства (ЦБДУИГ). В новом Законе о миграционном учете иностранных граждан, который вступил в силу в январе 2007 г., говорится о создании государственной информационной системы миграционного учета иностранных граждан, основанной, видимо, на ЦБДУИГ. Станет ли эта информационная система (или Центральный банк данных) надежным источником статистической информации о миграционной ситуации в России, позволит ли получать разнообразные, актуальные и корректные данные о ежегодных потоках мигрантов и об иностранном населении в нашей стране - покажет время. Начиная такую сложную работу, важно ознакомиться с опытом тех государств, которые накопили к настоящему времени большой практический опыт разработки административной статистики и, кроме того, сформировали традицию открытой публикации основных (но достаточно разнообразных) данных, нужных не только политикам, исследователям, но и всему обществу.

О.С. Чудиновских, *канд. экон. наук,*
МГУ им. М.В. Ломоносова

НАБЛЮДЕНИЕ И ОПИСАНИЕ МЕЖДУНАРОДНОЙ МИГРАЦИИ: ПРОБЛЕМЫ КАЧЕСТВА ДАННЫХ ПРИ ИСПОЛЬЗОВАНИИ РЕГИСТРАЦИОННЫХ ЗАПИСЕЙ ГОСУДАРСТВЕННОЙ ИММИГРАЦИОННОЙ СЛУЖБЫ КАК ИСТОЧНИКА СТАТИСТИЧЕСКОЙ ИНФОРМАЦИИ

Эйвинд Хоффманн (Eivind Hoffmann),
руководитель отдела статистики и анализа Директората иммиграции Норвегии

Введение

Известно, что не существует какого-либо одного источника статистических данных о миграции в страну и за ее пределы, из которого можно получить более или менее полную, достоверную и своевременную информацию, обладающую необходимым описательным и аналитическим потенциалом[2]. Тем не менее работникам статистики и исследователям приходится иметь дело с данными из доступных, в том числе с точки зрения «приемле-

* Доклад «Observing and describing international migrants: quality issues when using registrations of a national regulatory agency as basis for statistics» сделан на конгрессе Международного статистического института (Сидней, 2006), печатается с разрешения МСИ и автора. Перевод статьи канд. экон. наук О.С. Чудиновских.

[1] См.: http://www.udi.no/templates/OversiktssideType1.aspx?id=7530.

[2] По вопросам сильных и слабых сторон различных источников данных см., в частности: [1; 2; 3].

мых» затрат, источников. В свою очередь это часто подразумевает определенную систему записи сведений о поданных заявлениях и результатах их рассмотрения, применяющуюся при реализации национальной иммиграционной или эмиграционной политики. В фокусе данной статьи - методологические проблемы, с которыми сталкиваются пользователи, работающие с данными такого типа. В качестве примера используется опыт эксплуатации Системы регистрации и обработки информации по обращениям Директората иммиграции Норвегии[3]. (Далее для упрощения изложения мы используем термин «Система», *прим. перев.*)

В работе сначала приводится описание основных статистических функций Системы и формулируются некоторые проблемы из числа тех, которые возникают при использовании базы данных для получения статистики. Эти проблемы иллюстрируются рядом примеров, и дается краткое описание методов их решения.

Краткий обзор основных функций Системы регистрации и обработки информации по обращениям иностранных граждан

Эксплуатация Системы регистрации и обработки данных по обращениям в Норвегии началась в 1999 г., что ознаменовало замену систем, применявшихся ранее[4]. Это система коллективного пользования, предназначенная для работы иммиграционных властей Норвегии[5]. Норвежский директорат иммиграции является «владельцем» Системы и отвечает за ее работу и дальнейшее развитие.

Около 3200 пользователей по всех районах Норвегии имеют доступ к просмотру информации в Базе данных иностранцев, и примерно 2200 из них могут вносить в нее изменения через Систему обработки обращений (DUF). Эта Система является основным модулем Базы данных иностранцев, посредством которого записи по обращениям обновляются и используются. В 92 консульствах Норвегии применяется система NORVIS, предназначенная для обработки обращений за визами; пользователями системы SESAM являются около сотни центров приема лиц, ходатайствующих о предоставлении убежища. Приблизительно 1350 человек из числа тех, кому разрешено редактировать данные в Системе обработки (DUF), работают в центральном или местных управлениях полиции. В середине 2004 г. Система регистрации содержала записи приблизительно об 1 млн. человек и 2,2 млн. обращений, и ежедневно производилось более 20 тыс. транзакций. (По информации, полученной от автора статьи, к концу 2006 г. эти цифры могли увеличиться примерно на 20%, *прим. перев.*)

Главной целью развития Системы регистрации было обеспечение иммиграционных властей единым инструментом для корректной и эффективной обработки информации по обращениям. Кроме того, разработчики хотели создать такую программу, которая могла бы гарантировать расширение функциональных возможностей Системы для использования ее в качестве источника агрегированной статистики обращений, а также для получения индивидуальной информации о заявителях и обстоятельствах, имеющих к ним отношение.

Для этой цели была создана специальная и ежедневно обновляемая база данных; были пересмотрены юридические коды[6] для обращений, вынесенных по ним решений[7] и предпринятых действий; была создана ссылка между обращением и последующим решением.

Система регистрации позволяет дифференцировать первичные и повторные обращения, а также содержит поля для сведений, характеризующих мигранта, например таких, как язык, этно-культурная группа, уровень образования и занятие. Также имеется возможность создания прямой ссылки между сведениями о человеке, обратившемся с ходатайством, и связанном с ним лицом, в случае воссоединения или создания семьи. Записи о вынесенных решениях и предпринятых действиях, например о передаче ответственности (за рассмотрение или обработку обращения) от одного ведомства или сотрудника другому, производятся незамедлительно, так как это является неотъемлемой частью процесса принятия решения. Когда использовались прежние системы регистрации, окончательное оформление решений по обращениям могло быть отложено на несколько недель, то есть требовался определенный период ожидания, когда решение будет должным образом зарегистрировано.

Также был установлен порядок ежедневного обмена информацией с Центральным регистром населения Норвегии (ЦРН), преимущественно для того, чтобы обеспечить сотрудникам ЦРН простой доступ к информации, хранящейся в Системе регистрации, в случае принятия решения о регистрации мигранта в качестве резидента[8].

[3] Прим. перев.: Оригинальное название - Datasystem for utlendings - og flyktningsaker . Дословно перевести это название можно приблизительно так: Система обработки данных об иностранцах и лицах, ищущих убежища. В документах и публикациях норвежских специалистов для обозначения этой базы данных применяется аббревиатура DUF. В настоящее время Система (DUF) является интегрированным модулем Базы данных иностранцев (UDB). Помимо Системы обработки обращений (DUF), в эту Базу данных входит модуль NORVIS - Норвежская система обработки обращений за визой (совместимая с системой VIS, применяющейся в странах Шенгенской зоны), а также модуль SESAM - Система управления центрами приема лиц, обратившихся за предоставлением убежища.

[4] С 1988 до 2004 г. учет иностранцев в Норвегии велся в рамках двух регистров - Регистра беженцев и Регистра иностранцев, которые находились в ведении разных институтов - соответственно Директората по иммиграции и Вычислительного центра.

[5] См.: [4], в работе дается обзор различных функций и обязанностей в отношении иммиграционной политики, возложенных на политические и административные власти Норвегии.

[6] Юридические коды один к одному соответствуют конкретным положениям в законодательных актах и нормативных документах.

[7] Юридические коды были привязаны по смыслу к типологии «причин иммиграции», которая используется Норвежской статистической службой и в международных рекомендациях по статистике миграции, с целью дифференцировать разрешения на жительство в Норвегии по таким категориям, как «защита» (убежище), «работа», «учеба», «воссоединение или создание семьи» и «другие причины».

[8] Иностранцы, получившие статус резидента в Норвегии, обязаны зарегистрироваться в Центральном регистре населения Норвегии, поэтому сотрудники ЦРН должны иметь доступ к Системе регистрации иностранцев для контроля обоснованности и точности записей. *Прим. перев.*

Некоторые проблемы, возникающие при использовании системы обработки обращений в качестве источника статистических данных

Хотя создатели Системы заранее позаботились о том, чтобы ее можно было использовать для разработки статистики, в основном и в первую очередь она предназначена для *обработки отдельных обращений*. Как бы то ни было, при использовании регистрационных записей для получения своевременной, полной, актуальной и точной статистической информации возникают серьезные проблемы, которые можно проиллюстрировать приводимыми ниже примерами.

Единицы наблюдения. Регистрация обращений в Системе не отражает реальной миграции людей. Система может дать достаточно точные сведения о числе лиц, получивших *разрешение на пребывание* в Норвегии, но эта информация не показывает, сколько из них на самом деле *прибыли* в Норвегию[9]. Таким образом, отчеты о выданных видах на жительство и разрешениях на работу могут служить лишь в качестве (важной) основы для оценки миграции в Норвегию тех лиц, которые обращаются за индивидуальными разрешениями в качестве основания для переезда.

Полнота охвата. Существуют многочисленные группы иммигрантов, равно как и краткосрочных мигрантов, которые не включаются в Систему. Самая многочисленная группа иммигрантов - иностранных граждан[10] в Норвегии представлена примерно 56505 гражданами Дании, Финляндии, Исландии и Швеции, которым не нужно получать разрешения на проживание или работу для того, чтобы стать резидентом Норвегии. Они составляют 25% от 222277 иностранцев, которые были зарегистрированы в стране в качестве резидентов на 1 января 2006 г.[11]. Другая группа иностранцев, не учтенных в Системе, состоит из граждан стран Европейской экономической зоны[12], ищущих работу и имеющих право находиться в Норвегии в течение шести месяцев без разрешения на работу или на проживание, пока они ищут работу[13]. Незаконные иммигранты и иммигранты с неурегулированным статусом, конечно, не учитываются, также как почти 95% всех лиц, прибывших в Норвегию по краткосрочным визам до мая 2006 г.[14], и не учитываются краткосрочные мигранты, которым виза не требуется.

Качество и полнота записей. Возможность разработки статистики на основе полей для ввода данных и функций Системы определяется тем, введена ли информация и корректно ли это сделано. В иммиграционных службах постоянно присутствует давление, связанное с необходимостью корректно провести обработку обращений в соответствии с действующим законодательством и правилами и без необоснованных проволочек. Поэтому ни сотрудники, ни сами организации не склонны к тому, чтобы записывать сведения, которые непосредственно не связаны с процессом принятия решения по обращениям.

Сотрудники иммиграционных служб, находящиеся под давлением своих руководителей, которые требуют вынесения правильных решений в максимально сжатые сроки, имеют обыкновение пропускать поля для ввода данных, не обязательных для регистрации принятых решений и подготовки соответствующих документов. То обстоятельство, что должностным лицам не хватает навыков для правильного и аккуратного ввода «второстепенных» описательных характеристик, также является ограничением для удовлетворительного качества записи этих данных. Как следствие - невозможно в полной мере использовать потенциальные преимущества всех функций Системы для практического административного контроля и разработки статистики.

Приведем ряд примеров такой ситуации:

1. *Отсутствует запись сведений о лице, связанном с заявителем, в случае воссоединения или создания семьи.* В Системе имеется возможность создать прямую ссылку между записью о лице, обратившемся с ходатайством, и лице, с которым заявитель собирается проживать. Это осуществляется путем внесения в запись о заявителе идентификационного номера этого (рекомендуемого) лица, присвоенного ему в Системе, или его национального идентификационного номера, а также указания степени родства или отношения к заявителю. Такая информация призвана упростить получение часто запрашиваемой статистики о типе отношения к заявителю и статусе лиц, связанных с заявителем, в Норвегии. К сожалению, такая связывающая информация не всегда вводится в Систему, и в этом случае не представляется возможным получить статистические данные, в том числе для того, чтобы проконтролировать, есть ли среди лиц, имеющих отношение к заявителю, постоянные нарушители установленного порядка[15].

[9] Имеется в виду - в текущем году. *Прим. перев.*

[10] Речь идет об иностранных гражданах, проживающих в Норвегии в статусе резидентов. В статистических разработках, выполняемых Центральным статистическим бюро Норвегии на основе данных Центрального регистра населения, кроме того, выделяются категории иммигрантов по месту рождения вне Норвегии (мигранты первого поколения), месту рождения обоих или одного из родителей мигранта (лица с иммиграционным происхождением) и пр. При внесении записи о жителе Норвегии в Центральный регистр населения указываются также персональные идентификационные номера родителей, что позволяет получить информацию о происхождении человека. *Прим. перев.*

[11] См.: http://www.ssb.no/english/subjects/02/01/10/innvbef_en/tab-2006-05-11-03-en.html

[12] Соглашение по Европейской экономической зоне охватывает страны - члены Евросоюза, Исландию, Лихтенштейн и Норвегию.

[13] Они могут быть зарегистрированы службой занятости, если обратятся туда.

[14] До мая 2006 г. в Базу данных иностранцев через Систему обработки обращений попадали сведения только о тех обращениях за визой, которые были рассмотрены Директоратом иммиграции. В 2005 г. Директорат иммиграции выдал 2500 виз , а представительства Норвежской зарубежной службы выдали еще более 82500 виз. С мая 2006 г. информация обо всех выданных визах для краткосрочного посещения Норвегии попадает в Базу данных иностранцев посредством системы NORVIS из 92 консульств Норвегии за рубежом.

[15] В предыдущей системе такая статистика могла быть получена на основе специальных кодов, присваиваемых соответствующим решением, которые (коды) указывали на тип отношений между ходатайствующим лицом и лицом, связанным с ним. С 27 сентября 2004 г. регистрация связанного с заявителем лица и отношения его к заявителю стали обязательными в случаях воссоединения или создания семьи. Ниже будут рассмотрены вопросы о том, возможно ли установить такие связи в ретроспективе для выборочной совокупности обращений, в отношении которых эти данные отсутствуют.

2. *Несопровождаемые несовершеннолетние лица, ищущие убежище.* Когда ходатайство о предоставлении убежища получено, в соответствующей ячейке должна быть сделана запись о статусе несовершеннолетнего лица, ищущего убежище. Сотрудникам рекомендуется выполнять эту процедуру, но она не является обязательной. Как следствие, эта информация оказалась отсутствующей в немалой части записей по обращениям, зарегистрированным в 2003 г., и пришлось вручную выполнять проверку всех обратившихся с ходатайством для определения их статуса как Несопровождаемого несовершеннолетнего лица, ищущего убежище. Спустя некоторое время ситуация значительно улучшилась благодаря кампании по информатизации, и сегодня это не является большой проблемой[16].

3. *Продление.* Когда бланк ходатайства заполнен, сотрудники, производящие регистрацию, должны начать эту процедуру с выяснения, является ли это ходатайство первичным или касается продления действия предыдущего разрешения на пребывание. Это новая функция Системы регистрации, и, возможно, потребуется дополнительная подготовка персонала, прежде чем ее начнут выполнять должным образом. Тем временем мы получаем эту информацию, как это было при предыдущей системе регистрации, то есть путем загрузки информации по каждому предыдущему обращению, которое могло быть зарегистрировано в отношении одного и того же лица.

4. *Прибытие и продолжительность пребывания.* Статистика продолжительности пребывания может дать властям очень полезную информацию, которую не всегда содержат данные о юридическом основании пребывания иностранца в стране. В том числе становится возможным выделить долгосрочных и краткосрочных иммигрантов. Методика ввода таких сведений была усовершенствована после создания Системы регистрации, но статистическую информацию пока получить не так просто.

В отношении многих обращений эти сведения должны быть основаны на дате фактического прибытия в Норвегию. И такая информация не может быть получена и введена в Систему до обращения заявителя с ходатайством в местное отделение полиции и/ или в местное отделение Центрального регистра населения. И если он (она) сообщит о себе эти сведения, информация все-таки может быть не введена в Систему, даже если сотрудник полиции проставит в паспорте заявителя специальный штамп (о регистрации - *прим. перев.*). Лицо будет зарегистрировано в Центральном регистре населения только в том случае, если там сочтут, что выполнено требование пребывания в стране в течение как минимум шести месяцев[17].

5. *Описательная информация: уровень образования, занятие, этническое и культурное происхождение и пр.* Система регистрации и обработки обращений предусматривает возможность внесения записей об этих характеристиках заявителей. Многие из них не имеют прямого отношения к результату рассмотрения заявления, поэтому такие сведения не регистрируются в объеме, достаточном для разработки статистики или принятия последующих решений (если таковые будут). Частично это происходит из-за того, что качество информации, предоставленной заявителем, невысоко, и нет причин улучшать ее с точки зрения перспективы вынесения первого решения. Кроме того, чиновники могут быть не заинтересованы в том, чтобы вносить эти сведения в Систему, особенно в тех случаях, если недостаточными были инструкции, разъясняющие, для чего это должно быть сделано, и как это нужно сделать, чтобы записи были правильными. Наиболее важно обеспечить работников иммиграционных служб адекватными инструкциями по корректному вводу данных, относящихся к таким характеристикам, как уровень образования (в тех случаях когда образование, полученное в другой стране, должно быть «переведено» в соответствующий уровень образования в норвежской образовательной системе), занятие (так как в наличии должны быть списки кодов и поддерживаемая компьютером система кодирования), и этническое и культурное происхождение[18]. Для получения сведений такого типа важно правильно сформулировать вопросы как в бланках заявлений, так и при проведении личных интервью, и обеспечить наличие эффективных инструментов для кодирования ответов[19].

Заключение

С появлением Системы обработки обращений стало возможным выполнить ряд статистических разработок, характеризующих деятельность иммиграционных властей, самих заявителей и результаты рассмотрения обращений. Это шаг в правильном направлении к такой организации дел, при которой мы сможем получить максимум важных и актуальных статистических данных, необходимых сегодня властям и обществу в целом. Тем не менее трудности в обеспечении того, чтобы вся необходимая и доступная информация записывалась правильно, являются проблемой, которую нужно рассматривать с позиций детального знания того, каким образом функционируют все сопряженные процедуры. Также необходимо иметь представление о том, какое давление оказывается на должностные лица, и какие стимулы могут побу-

[16] Предыдущая система делала эту информацию доступной посредством специального кода, присваиваемого «типу обращения за статусом беженца». Такой подход также допускал возможность ошибок, но при переходе к новой Системе понимание важности правильного кодирования было на время утрачено.

[17] Были предприняты шаги для того, чтобы убедиться, что в Системе будут зарегистрированы нерезиденты, которым был присвоен ИНН (также определенный ведомством, проводящим регистрацию), и разработать более удобные процедуры для регистрации прибывших в местных отделениях полиции.

[18] Вопросы об этническом и культурном происхождении могут быть уместны только в отношении отдельных лиц и ситуаций, в частности чтобы избежать совместного расселения представителей конфликтующих этнических групп во временных (транзитных) гостиницах.

[19] Инструменты для кодирования таких переменных, как «образование» и «занятие», были созданы в процессе совместной работы Директората иммиграции и Статистической службы Норвегии.

дить их записывать информацию правильно и в полном объеме. Без таких постоянно обновляемых знаний будет невозможно судить о том, в какой степени получаемая из административного источника статистика является актуальной и достоверной.

Система регистрации и обработки была спроектирована с минимальным числом полей для ввода обязательных данных, так как было решено, что большее число переменных станет препятствием для эффективной обработки обращений разнородной совокупности заявителей, данные о которых будут внесены в базу.

При оценке данного решения была доказана его контрпродуктивность в ряде случаев, так как не была принята во внимание потребность в надежной информации при рассмотрениях последующих заявлений одних и тех же лиц, (например, продлений действия разрешения), кроме того, было замечено: чем лучше и понятнее производимая статистическая информация, тем в большей степени возрастает потребность в новой статистической информации.

Специалисты, наиболее заинтересованные в отсутствующей или дополнительной статистике, часто являются теми же самыми чиновниками и должностными лицами, которые сопротивляются тому, чтобы найти время для ввода необходимых исходных данных. Проблема состоит в донесении до понимания этими людьми простой мысли: «Что вы даете, то вы и получаете», после чего на самом деле станет понятно, что можно получить в итоге.

Проблема, с которой сталкиваются те, кто использует Систему в качестве источника статистики международной миграции, присущи не только этой Системе. Они, судя по всему, в большей или меньшей степени имеют место везде, где административный учет является источником первичных данных для статистики такого рода. Единственная стратегия, которая позволит получать более или менее достоверную статистику из административных систем регистрации, состоит в том, чтобы конструировать эти системы и их операционную инфраструктуру, имея перед собой четкую цель: обеспечить и эффективную обработку заявлений, и статистику хорошего качества. Реализация такой стратегии является задачей, которая требует одновременно юридической и прикладной статистической экспертизы, а также глубокого понимания проблем миграции.

Литература

1. **Bilsborrow, R. et al.** (1996): International Migration: Guidelines for improving data collection systems. International Labour Office, Geneva.

2. **Hoffmann, E.** (1997): «Administrative records and surveys as basis for statistics on international labour migration», in International Statistical Review, 65, 2.

3. **Hoffmann, E. & Lawrence, S.,** (1996): Statistics on International Labour Migration: A review of sources and methodological issues. Working paper from the Interdepartmental Project on Migrant Workers, 1994-95. International Labour Office, Geneva.

4. **Hoffmann, E. & Svensbraaten, A.E.** (2004): «Discussing the challenges in using a case processing system for statistical purposes». Presentation to Workshop F39: Migration Data Sources: Exploring the data programs, at Metropolis 2004, Geneva, 27 September - 1 October.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ВОССТАНОВЛЕНИЯ ОТСУТСТВУЮЩИХ СВЕДЕНИЙ ПРИ ОБРАБОТКЕ МАТЕРИАЛОВ ПЕРЕПИСЕЙ И ОБСЛЕДОВАНИЙ НАСЕЛЕНИЯ

Т.М. Чернышева, *канд. экон. наук,*
Научно-исследовательский институт проблем социально-экономической статистики
Федеральной службы государственной статистики,
Г.Е. Шевердова,
Федеральная служба государственной статистики

Методы восстановления отсутствующих сведений (методы импутации) предназначены для определения пропущенных или ошибочных значений в материале, собранном в результате проведения статистического наблюдения. Использование методов импутации направлено на повышение качества статистических данных.

При обработке материалов обследований и переписей населения часто приходится сталкиваться с проблемой или неответов на отдельные вопросы переписных листов, анкет, опросных бланков и т. п., или с наличием в них некорректных записей. Указанное снижает полноту и достоверность заполненных бланков обследований, а в целом приводит к недостаточному качеству статистической информации. Причем чем больше пропущенных и ошибочных записей, тем ниже качество первичной информации, полученной при проведении переписей и обследований населения.

К наиболее распространенным типам ошибок, встречающихся при обработке материалов переписей и обследований населения, относятся:

- *недействительные данные.* Для этого типа ошибок характерными являются или недействительные ответы, то есть не соответствующие допустимым значениям, определенным для каждого поля, выделенного в переписном

листе (или анкете) для заполнения ответов на его вопросы, или же наличие нескольких ответов на один и тот же вопрос при требовании однозначного ответа;

- *несовместимые данные.* К этому типу ошибок относятся или ответы, не имеющие смысла, или же ответы, противоречащие ответам на другие взаимосвязанные вопросы анкеты или переписного листа;

- *отсутствующие данные (записи)*. Этот тип ошибок характеризует неответы на отдельные вопросы анкеты или переписного листа и в наибольшей степени определяет качество первичной информации.

В основном указанные типы ошибок в первичных данных могут возникать из-за сложности ряда вопросов анкеты или переписного листа, или из-за неоднозначного их истолкования как интервьюером, так и респондентом, или нежеланием людей отвечать на неудобные вопросы, или по другим причинам объективного и субъективного характера (например, вопросы переписного листа или анкеты заполняются со слов одного члена домохозяйства, который не полностью владеет информацией об остальных членах домохозяйства).

Для устранения отдельного типа ошибок широко используются как процедуры формальных и логических проверок, так и процедуры редактирования, которые, как правило, реализуются на этапе обработки материалов переписей и обследований населения. В частности, данные процедуры хорошо себя зарекомендовали при исправлении недействительных и несовместимых данных, а также при восстановлении небольшого процента отсутствующих данных, если правильную запись можно вменить на этапе редактирования с учетом взаимосвязанных ответов на другие вопросы переписного листа или анкеты.

Однако восстановить наибольший процент отсутствующих сведений и изменить ряд записей, признанных несостоятельными при проведении формально-логических проверок и не поддающихся исправлению в рамках реализованных систем редактирования, можно только на основе методов импутации.

Методы импутации включают широкий класс простых и сложных математических, математико-статистических и экономико-математических моделей, предназначенных для определения нового значения переменной вместо пропущенных или несостоятельных полей в анкете или переписном листе. Новое значение переменной должно полностью удовлетворять всем установленным правилам редактирования[1]. При выполнении этого требования восстановленная запись считается корректной. Это и определяет неразрывную связь импутации с процедурой редактирования статистических данных. Поэтому в основу реализации указанных процедур должны быть заложены одни и те же правила, методы и технологии.

На современном этапе разработано множество подходов к реализации процедур редактирования и импутации. Характерным для этих подходов является широкое использование в них системы Филледжи-Хольта[2], отражающей основные требования к редактированию статистических данных (представлена в кратком изложении):

- каждое значение переменной (поля) должно удовлетворять всем правилам редактирования. Это должно достигаться за счет импутации минимального количества полей;

- «импутированные» записи должны полностью удовлетворять правилам редактирования. Это означает, что правила импутации должны автоматически генерироваться из правил редактирования;

- конечным результатом процедуры импутации должно являться ненарушение совместного распределения переменных.

В целом следует отметить, что в современных технологиях достаточно трудно отделить редактирование от импутации и оценить раздельное их влияние на качество результирующей статистики. Характерным для методов импутации, реализованных в современных технологиях, является их разнообразие, что обусловлено и назначением проводимых обследований, и различным представлением правил редактирования, и использованием разных процедур для определения подставляемых значений соответствующего типа, то есть или метрических (количественных) переменных, или категориальных (качественных) переменных (в переписях населения и в обследованиях населения социального и социально-демографического характера большая часть переменных - категориальные).

К наиболее широко применяемым методам импутации с использованием технологии Филледжи-Хольта можно отнести: методы донора, методы оценок, детерминистские[3] методы.

---

[1] Редактирование статистических данных - это деятельность, направленная на обнаружение ошибок в данных и их обработку. Правила редактирования - это специально заданные соотношения, которые связывают между собой переменные (поля) и используются для контроля корректности данных переписей и обследований. Эти соотношения могут представляться в самых разнообразных формах (например, в виде легко модифицируемых таблиц) и содержать различные линейные и нелинейные функции. Достаточно часто процедуру редактирования отождествляют с процедурой формально-логического контроля статистических данных, которая в целом и направлена на обнаружение ошибок. Современное понимание процесса редактирования несколько шире, так как при реализации этой процедуры предусматривается импутация (подстановка) данных (первый этап редактирования), не удовлетворяющих правилам редактирования. Все, что не удается «импутировать» на первом этапе редактирования, восстанавливается или исправляется на втором этапе редактирования, где и используются специальные методы импутации. Однако любое восстановленное значение переменной на втором этапе должно по-прежнему удовлетворять всем ранее установленным правилам редактирования.

[2] Указанная система реализована в большинстве программных продуктов, предназначенных для автоматического редактирования и импутации данных. Основная часть системы Филледжи-Хольта реализуется в виде набора программ, который может быть достаточно легко адаптирован для применения при обработке данных других обследований.

[3] Детерминистские методы импутации - это методы, в результате применения которых подставляемое значение определяется однозначно. К ним относятся: *логическое заполнение*, *историческое заполнение* (используется при повторяющихся экономических обследованиях, эффек-

Методы донора - это методы, в которых пропущенное или ошибочное значение восстанавливается из другой записи без каких-либо изменений. В соответствии с методом донора значения для импутации берутся из записей, удовлетворяющих правилам редактирования. В зависимости от способа выбора записи-донора для восстановления пропущенных значений различают две основные разновидности метода донора: случайный выбор; выбор ближайшего в некоторой метрике соседа. Случайный выбор записи-донора осуществляется в рамках специально выделенных (построенных) стратах (классах, группах), однородных по структуре ряда заданных параметров. Если требуется восстановить несколько полей, то все они берутся из одной отобранной записи-донора.

Выбор записи-донора методом ближайшего соседа основан на определении метрики, характеризующей наименьшее расстояние от импутируемой записи. В качестве метрики для определения расстояния могут быть выбраны различные функции. При использовании метода ближайшего соседа, кроме метрики, определяются и поля, по которым будет оцениваться расстояние. При реализации данного метода для устранения возможного преобладания в выборе одного из параметров (в частности, из-за различных единиц измерения) нередко осуществляются различные преобразования, например ранжирование, нормирование, вычисление максимального числа однотипных пересечений и др.

Методы оценок - это методы, основанные на определении пропущенных или несостоятельных записей в данных путем вычисления некоторой оценки. Эта оценка может основываться на данных текущего обследования или данных предыдущих обследований (методы средних, трендов, регрессии, множественной импутации и др.).

Детерминистские методы достаточно часто используются в системах редактирования и импутации, где реализована технология Филледжи-Хольта (в этих системах возможность применения детерминистских методов проверяется в первую очередь).

Особенно широкое развитие и внедрение методов редактирования и импутации характерно для зарубежной практики.

В современных системах редактирования и импутации заложен многообразный и достаточно сложный математический аппарат, реализация которого требует универсального программного обеспечения. Известно множество программных пакетов[4], предназначенных для редактирования и импутации статистических данных. Однако большинство из них используются в экономических обследованиях, где приходится оперировать преимущественно с количественными (метрическими) переменными.

Из стандартных продуктов, используемых для восстановления отсутствующих сведений в переписной информации, наиболее известной является система NIM, разработанная в 1996 г. в Канаде для обработки данных переписей населения. Эта система в основном оперирует с категориальными переменными. Единственный применяемый в NIM метод импутации - метод донора. Главный критерий этого метода - минимизация количества изменений. Импутация переменной всегда выполняется на основе единственной записи-донора. Система осуществляет поиск донора для домохозяйства-реципиента в целом, а не индивидуально для каждого члена домохозяйства. Подходящая запись-донор определяется по критерию минимума расстояния до записи, подлежащей импутации (то есть в системе реализован метод ближайшего соседа).

В ряде стран (Великобритания, США, Италия) для восстановления отсутствующих записей в переписной информации разработаны программные пакеты для внутреннего пользования, в которых для импутации данных также применяются методы донора и методы оценок (текущего среднего, множественной регрессии для восстановления пропущенного возраста человека).

В целом в этих пакетах заложена донорская импутационная система (ДИС), которая по большинству позиций адекватна Канадской системе NIM.

Обе системы основаны на поиске донорского домашнего хозяйства, близко соприкасаемого с домохозяйством-реципиентом. В указанных системах поиск донора основывается на концепции получения минимального статистического отклонения между двумя домашними хозяйствами.

Анализ мирового опыта показывает, что методы и технологии редактирования и импутации постоянно совершенствуются в направлении разработки универсальных систем автоматического редактирования и импутации, пригодных для широкого применения в обследованиях экономического, социально-экономического и социально-демографического характера.

В нашей стране наибольшее развитие получили методы и технологии редактирования статистических данных,

---

тивно для переменных, имеющих тенденцию к стабильности); *заполнение средними*, *заполнение с подбором* (hot-deck-imputation), то есть подстановкой из подготовленной колоды (случайный или последовательный подбор), *заполнение без подбора* (использование внешнего источника), *заполнение по регрессии*, *заполнение подбором ближайшего соседа* (основано на введении метрики *d* для расстояния между объектами).

[4] К наиболее известным программным пакетам, используемым в экономических исследованиях, относятся: STEPS - разработан в Бюро цензов США - в пакете реализовано два модуля импутации: простой (выполняет импутацию детерминистским методом) и обобщенный (технология импутации основывается на методах оценок); GEIS - система разработана статистическим агентством Канады - для импутации данных используется: детерминистский метод, методы донора (ближайший сосед), методы оценок (предшествующих значений, текущих средних, трендов, множественной регрессии); AGGITS - система разработана Департаментом сельскохозяйственной статистики США - в этой системе, так же как и в GEIS, реализована обобщенная технология редактирования Филледжи-Хольта, для импутации данных применяется метод донора (метод ближайшего соседа) и методы оценок (в систему заложен также модуль, нацеленный на обнаружение выбросов); SOLAS - разработана ирландской компанией Statistical Solutions Ltd, данная система в основном предназначена для импутации: в ней используется метод множественной импутации, а также стандартный метод донора (Random Hot-Deck со случайным подбором) и два типа метода оценок (текущего среднего и исторического) и другие программные пакеты.

являющиеся в целом первым шагом импутации. Причем процедуры формально-логического контроля, направленные на обнаружение ошибок и пропусков в статистических данных и дальнейшее их редактирование, широко используются при обработке и данных переписей населения, и материалов как экономических обследований, так и крупномасштабных обследований населения. Отдельные приемы импутации нашли применение в социологических обследованиях, а также в обследовании предприятий, особенно при обработке выборочных данных с пропусками, обусловленными неответами респондентов.

В последние годы развитию и разработке методов импутации для восстановления записей пропущенных и несостоятельных полей в статистических данных уделяется в нашей стране более пристальное внимание: импутация применяется при обработке данных Всероссийской сельскохозяйственной переписи 2006 г. и планируется к использованию при Всероссийской переписи населения 2010 г.

В настоящее время проводятся исследования по совершенствованию методов импутации, разработанных на основе анализа материалов Всероссийской переписи населения 2002 г., и подготовка их к реализации для восстановления отсутствующих сведений в записях переписных листов при проведении следующей переписи населения. В рамках данного направления:

- разработана «Методология восстановления отсутствующих сведений», в которой максимально учтены специфика обработки переписной информации, типы ошибок, зафиксированные в переписных листах, частота появления неответов на отдельные вопросы переписного листа форм К(Д), а также и взаимосвязь между ответами на вопросы переписного листа форм К(Д);

- разработаны рекомендации по созданию общей базы данных и в ее рамках - двух базовых подсовокупностей: домохозяйств-доноров и домохозяйств-реципиентов;

- определены наиболее эффективные методы импутации пропущенных и несостоятельных полей в записях переписного листа и разработаны процедуры и последовательность их реализации;

- выделена в рамках переписного листа форм К(Д) система основных и вспомогательных признаков, предназначенных для установления схожести домохозяйств-доноров и домохозяйств-реципиентов, а также система признаков переписного листа, по которым импутация пропущенных записей невозможна;

- разработаны рекомендации по построению стратифицированных подмножеств и специальному кодированию структурных признаков переписного листа для эффективного поиска домохозяйства-донора;

- разработаны рекомендации по проведению экспериментальных исследований по использованию методов импутации на базе данных ВПН-2002 трех субъектов РФ (Рязанской, Костромской и Камчатской областей), разработано программное обеспечение и получены первые результаты экспериментальных работ по восстановлению отсутствующих сведений в записях переписного листа.

На основании результатов проведенных исследований, а также анализа зарубежного опыта по использованию методов импутации в переписях и обследованиях населения было установлено, что для восстановления отсутствующих сведений в записях переписного листа при обработке данных Всероссийской переписи населения 2010 г. целесообразно использовать метод донора (по большинству вопросов переписного листа с неответами) и метод оценок (в основном для восстановления неответа на вопрос о возрасте респондента).

Основной функцией метода донора при импутации пропущенных или несостоятельных полей в переписном листе является поиск домохозяйства-донора, схожего по ряду признаков с домохозяйством, у которого в отдельных ответах переписного листа имеются пропущенные сведения у некоторых или всех членах домохозяйства (для подобных домохозяйств используется термин или «реципиент», или «получатель»). Поиск донорского домохозяйства основывается на концепции получения максимальной близости между донором и реципиентом. При применении данного метода поиск донора для соответствующего реципиента осуществляется в целом для домохозяйства, а не индивидуально для каждого его члена.

Все пропущенные записи у членов домохозяйства-реципиента условно восстанавливаются из одной донорской записи. Эти записи должны: во-первых, удовлетворять всем правилам редактирования и, во-вторых, содержать только ответы респондентов и не иметь импутированных значений.

Для реализации метода донора необходимо построить общую базу данных, а в ее рамках выделить две подсовокупности: домохозяйств-доноров и домохозяйств-реципиентов.

Информационной основой для построения общей базы данных в рамках субъектов РФ является совокупность переписных листов, прошедших процедуру редактирования ошибочных записей по правилам формального и логического контроля.

Для выделения двух подсовокупностей необходимо в составленной базе данных каждому домохозяйству, в зависимости от наличия в переписных листах полных ответов или частичных неответов, присвоить соответственно или признак донора, или признак реципиента.

Признак донора присваивается только тем домохозяйствам, у которых записи в переписных листах удовлетворяют всем правилам редактирования.

Признак реципиента присваивается домохозяйствам, у которых в переписных листах, составленных на его членов, обнаружены частичные неответы на отдельные вопросы переписного листа.

Подсовокупность домохозяйств-реципиентов подразделяется на два подмножества:

- в первое подмножество включаются домохозяйства-реципиенты, содержащие совокупность переписных листов, информация по которым не подлежит восстановлению из-за максимального количества неответов;

- во второе подмножество включаются домохозяйства-реципиенты, по которым в переписных листах имеются отсутствующие записи: во-первых, на вопросы перепис-

ного листа, подлежащие восстановлению, и, во-вторых, на вопросы переписного листа как не подлежащие восстановлению[5] из-за высокой вероятности вменения ошибочного значения, так и подлежащие восстановлению. В указанной совокупности переписных листов признак невозможности импутации отсутствующих записей присваивается не домашнему хозяйству, а соответствующему вопросу переписного листа, составленного на отдельных членов домохозяйства-реципиента.

В рамках второго подмножества домохозяйств-реципиентов необходимо выявить домохозяйства (если они имеются), для переписных листов которых характерно отсутствие только записей на вопросы, не подлежащие восстановлению, с целью их исключения из базовой совокупности домохозяйств-реципиентов (однако исключенная подсовокупность домохозяйств никогда не может быть использована в качестве донорской базовой основы).

Таким образом, восстановление отсутствующих сведений в записях переписного листа на основе совокупности домохозяйств-доноров осуществляется только по тем вопросам переписного листа домохозяйств-реципиентов, по которым импутация возможна.

Создание общей базы данных, равно как и выделение в ней двух подсовокупностей (доноров и реципиентов) проводится отдельно по городскому и сельскому населению. Для сокращения временных затрат на поиск донорского домохозяйства объекты наблюдения в образованных двух подсовокупностях стратифицируются по четырем признакам: территориальному (то есть по принадлежности к соответствующему административному району, населенному пункту и, по возможности, - к переписному, инструкторскому и счетному участку), типу домохозяйств (выделяется 111 типов), укрупненному блоку типов домохозяйств (выделяется шесть блоков), размеру домохозяйства (выделяется семь групп).

Построение страт по сочетанию указанных четырех признаков осуществляется отдельно по двум формам переписного листа (форме К и форме Д).

В результате многомерной стратификации объектов наблюдения общей базы данных по четырем признакам будут получены: во-первых, две основные непересекающиеся подсовокупности домохозяйств (доноров и реципиентов), в каждой из которых образован ряд качественно однородных подмножеств (страт) относительно принадлежности объектов наблюдения (домохозяйств) к соответствующим территориальным образованиям, типам домохозяйств, размеру домохозяйств и укрупненным блокам типов домохозяйств; во-вторых, стратифицированные подмножества, готовые для поиска подходящего донора для соответствующего реципиента вследствие их качественной адекватности по выделенным признакам.

Выделенные качественно однородные подмножества представляют те страты, где требуется осуществлять поиск домохозяйства-донора в первую очередь. Выбор же подходящего донора основан на дальнейшей стратификации образованных подмножеств с учетом комбинации ответов на ряд структурных признаков переписного листа, определяющих схожесть донора и реципиента.

В рамках данных признаков выделяются две группы, в одну из которых входят основные признаки переписного листа, а в другую - вспомогательные признаки.

К основным признакам, определяющим адекватность донора и реципиента, относятся следующие признаки (вопросы) переписного листа: «Ваш пол», «Дата Вашего рождения», «Учитесь ли Вы в образовательном учреждении?», «Посещает ли ребенок дошкольное учреждение?», «Ваше образование», «Умеете ли Вы читать и писать?», «Окончили ли Вы профессиональное или профессионально-техническое училище?», «Укажите все имеющиеся у Вас источники средств к существованию», «Имели ли Вы какую-нибудь работу, приносящую заработок или доход, за неделю до начала переписи населения?» и «Кем Вы являлись на основной работе?».

К вспомогательным признакам относятся вопросы: «В какой отрасли экономики Вы заняты?», «Какую основную продукцию или услуги производит (оказывает) предприятие (организация), на котором Вы заняты (включая индивидуальных предпринимателей)?», «Ваше занятие или выполняемая работа», «В случае отсутствия работы, искали ли Вы ее в течение последнего месяца?» и «Ваше состояние в браке» (последний признак может быть использован в качестве основного при восстановлении отсутствующей записи на вопрос о возрасте).

Структура ответов на основные и вспомогательные вопросы переписного листа является основой: во-первых, для определения их различных комбинаций в зависимости от восстанавливаемого вопроса и, во-вторых, для выделения в рамках качественно однородных подмножеств ряда донорских групп, адекватных группам домохозяйств-реципиентов, в переписных листах которых имеются неответы на отдельные вопросы.

Для построения однотипных страт доноров и реципиентов каждому члену домохозяйства присваивается специальная кодовая запись, означающая его принадлежность к определенной страте по комбинации структурных признаков.

С учетом изложенных построений для поиска подходящего донора, близко соприкасаемого с реципиентом, были разработаны два основных метода его поиска: метод многомерной стратификации и метод «прочесывания». В этих двух методах близость домохозяйства-донора и домохозяйства-реципиента определяется по максимальному числу пересечений между ответами на однотипные

[5] К вопросам переписного листа, по которым вменение ответов нецелесообразно, относятся вопросы: «Ваше родственное отношение с проживающими совместно лицами (по отношению к тому, кто записан первым в этой учетной единице)»; «Ваш пол»; «Место Вашего рождения»; «Ваше гражданство»; «Ваша национальная принадлежность»; «Владение языками»; «Ваша работа находится на территории Вашего города (района)?»; «В этом городе (городском поселении или сельской местности района) Вы проживаете непрерывно с рождения?»; «Сколько детей Вы родили?». К указанной совокупности вопросов может быть отнесен и вопрос «Ваше состояние в браке».

вопросы переписного листа. При обнаружении нескольких доноров с одинаковым числом максимальных пересечений один из них выбирается случайно.

Пропущенные или несостоятельные поля в переписных листах (кроме пропуска в ответе на вопрос о возрасте), составленных на членов домохозяйств-реципиентов, восстанавливаются полностью из донорской записи. При восстановлении отсутствующих сведений по вопросу о возрасте используется метод текущих средних (то есть вычисляется среднее арифметическое значение по имеющемуся ряду возрастов в донорских группах, составленному по лицам, адекватным получателю). Этот метод применяется при частичном несовпадении ответов на основные вопросы в группах доноров и реципиентов.

.Однако если у донора и реципиента наблюдается полное совпадение ответов на все основные и вспомогательные вопросы переписного листа, взаимосвязанные с вопросом о возрасте, то возраст может быть восстановлен из донорской записи.

Следует также подчеркнуть, что при восстановлении отсутствующих сведений в записях переписного листа очень важное значение имеет соблюдение последовательности восстановления отсутствующих сведений.

В соответствии с анализом структуры неответов[6] на отдельные вопросы переписного листа, частоты и частости их появления, а также распределения неответов в группах по их количеству было установлено, что восстановление пропущенных сведений необходимо начинать в группах переписных листов с наименьшим числом неответов, то есть практически с группы с одним неответом у членов домохозяйств-реципиентов (группы с наименьшим числом неответов имеют самый высокий удельный вес) и т. д. Для определения этих групп применяется процедура систематизации подсовокупности домохозяйств-реципиентов по количеству неответов на вопросы переписного листа. Данная процедура реализуется в рамках сформированных качественно адекватных подмножеств.

Восстановление отсутствующих сведений в переписных листах с наименьшим числом неответов позволяет использовать домохозяйство-донора несколько раз (один и тот же донор может быть использован не более трех раз и желательно для вменения ответов на разные вопросы).

При поиске подходящего донора может возникнуть ситуация, когда в рамках качественно однородных страт донор не находится. В этом случае осуществляется переход в сопоставимые страты, выделенные или в других типах домохозяйств в рамках одного и того же укрупненного блока, или в другом укрупненном блоке типов домохозяйств (данные переходы осуществляются на основе разработанных блок-схем поиска подходящего донора, в которых предусмотрен и переход в сопоставимые страты географически близких территориальных единиц). Кроме того, следует отметить, что при определении схожести донора и реципиента основное внимание обращается на их адекватность по структуре ответов на вопросы, взаимосвязанные с импутируемой записью. Поэтому при выделении донорских групп не требуется стремиться к идентичности донора и реципиента по всем выделенным признакам. Таким образом, при поиске подходящего донора допускается процедура выведения из группы признаков, по которым определяется адекватность донора и реципиента, сначала признаков, не влияющих на результат импутации, а затем (при необнаружении донора) - и одного из основных признаков при его слабой связи с импутируемой переменной.

Основные функции разработанных методов импутации сначала были апробированы на небольших условных совокупностях домохозяйств разного размера, в которых были выделены две подсовокупности домохозяйств: доноров и реципиентов.

На базе этих подсовокупностей были реализованы: во-первых, методы построения стратифицированных подмножеств и формирования в них специальных страт с учетом вариантов значений ряда структурных признаков анкеты и, во-вторых, методы поиска подходящего донора для соответствующего реципиента и вменение записей вместо пропущенных ответов. Результаты апробации показали, что реализованные методы импутации относятся к эффективным и позволяют достаточно точно восстанавливать отсутствующие записи в ответах на вопросы анкеты.

Дальнейшие исследования в этом направлении были проведены на реальных совокупностях, то есть на части базы данных ВПН-2002 трех субъектов РФ: Рязанской, Костромской и Камчатской областей.

Основными задачами экспериментальных исследований по использованию методов импутации в переписной информации являлись:

- определение как целесообразности использования разработанных методов импутации (включая проверку работы их основных модулей) при обработке данных Всероссийской переписи населения, так и выявления ряда значимых проблем, с которыми можно сталкиваться при реализации данных методов на крупных информационных массивах;

- оценка эффективности реализованных методов импутации и надежности результатов восстановления отсутствующих записей в переписных листах.

Для достижения целей и задач эксперимента на базе данных указанных субъектов РФ было принято решение использовать специальную совокупность переписных листов, на которой было бы возможно не только реализовать методы импутации, но и сравнить полученные результаты после восстановления отсутствующих сведений в записях переписного листа с известными записями в исходной (гипотетической) совокупности.

В соответствии с поставленной задачей за гипотети-

[6] Анализ структуры неответов на вопросы переписного листа форм К(Д) проводился по данным ВПН-2002 в целом по России и указанным трем субъектам РФ (отдельно по городскому и сельскому населению).

ческую совокупность была принята база домохозяйств-доноров, созданная на основе общей базы данных ВПН-2002 трех субъектов РФ, отдельно по городскому и сельскому населению. Характерным для гипотетической совокупности является наличие признака «донор» у всех домохозяйств, входящих в ее состав, отсутствие в этой базе переписных листов с частичными неответами.

В рамках гипотетической совокупности были смоделированы экспериментальные совокупности - специальный информационный массив (СИМ) домохозяйств-доноров и СИМ домохозяйств-реципиентов. Количественный состав СИМ определялся с учетом состава гипотетической совокупности и удельной структуры по двум показателям (количество лиц и число домохозяйств) двух базовых подсовокупностей домохозяйств (доноров и реципиентов), определенной в рамках общей базы данных с учетом территориального признака, принадлежности домохозяйств разного размера к одному из 111 типов домохозяйств и шести укрупненным блокам.

К основным задачам моделирования СИМ доноров и реципиентов в рамках гипотетической совокупности относятся:

- определение абсолютного числа домохозяйств-реципиентов, которое должно быть искусственно образовано в данной совокупности в соответствующих группах по размеру домохозяйств, типам домохозяйств и шести укрупненным блокам, в том числе и по сочетанию указанных качественных характеристик (для формирования СИМ реципиентов используется или случайный, или систематический отбор в зависимости от размера страт в гипотетической совокупности и СИМ реципиентов);

- определение абсолютного числа домохозяйств-доноров, которое должно быть использовано для восстановления искусственно удаленных ответов на отдельные вопросы переписного листа по СИМ реципиентов;

- проведение процедуры искусственного вменения неответов в записи отдельных вопросов переписного листа. Эта процедура проводится на основании множества взаимосвязанных аналитических таблиц, характеризующих по базовой основе реципиентов как общий характер неответов на вопросы переписного листа и их распределение в группах по количеству неответов, так и детализированную структуру неответов по четырем качественным характеристикам, включая частоту и частость появления неответов на конкретные вопросы переписного листа, а также их различные сочетания при одновременных неответах у нескольких лиц в однотипных по размеру домохозяйствах.

Созданный в рамках гипотетической совокупности специальный информационный массив реципиентов[7] с искусственно вмененными неответами на вопросы переписного листа являлся основой для дальнейшего проведения эксперимента в направлении реализации методов импутации для восстановления пропущенных полей.

При проведении экспериментальных исследований по трем субъектам РФ для восстановления искусственно вмененных неответов на вопросы переписного листа были использованы методы донора и методы оценок. Для реализации данных методов и множества подготовительных работ, связанных с формированием базы данных первичных показателей ВПН-2002, подготовкой отчетов о структуре показателей этих баз, формированием отчетов о структуре неответов на вопросы переписного листа, составлением аналитических таблиц с детализированной структурой неответов на вопросы переписного листа, моделированием специальных информационных массивов доноров и реципиентов, проведением процедур по искусственному вменению неответов на вопросы переписного листа, определением последовательности восстановления пропущенных полей, было создано программное обеспечение (его разработка была осуществлена ЗАО «Крокинкорпорейтед», а подготовка отчетов с результатами экспериментальных исследований по использованию методов импутации выполнялась при техническом содействии ГМЦ Росстата).

В рамках данного программного продукта для поиска подходящего донора реализуются методы многомерной стратификации (на этапе построения качественно однородных подмножеств) и метод «прочесывания», аккумулирующего в себе основные функции метода донора и одновременно являющегося наиболее простым для реализации. Кроме того в его рамках предусмотрена и процедура поиска подходящего донора в сопоставимой страте другой территориальной единицы на основе установленной географической[8] близости территориальных единиц по трем субъектам РФ (в частности, эта процедура была использована при поиске подходящего донора в Рязанской и Костромской областях).

Для реализации отмеченных выше двух методов импутации при проведении экспериментальных исследований поиск домохозяйства-донора осуществлялся с использованием метода «прочесывания».

При реализации метода «прочесывания» широко применялась процедура стратификации как на стадии специальной подготовки базовой основы, когда каждый из ответов переписного листа представляется соответствующей кодовой записью, так и на этапе систематизации этих кодовых записей по количеству пересечений (совпадений) между однотипными ответами у донора и реципиента.

[7] СИМ реципиентов был создан в рамках каждого из трех субъектов РФ, отдельно по городскому и сельскому населению, а также отдельно по двум формам переписного листа.

[8] Знание географической близости территориальных единиц особенно эффективно в тех случаях, когда поиск донора не дает положительных результатов ни в рамках выделенной страты, ни в рамках сопоставимой страты укрупненного блока, который включает соответствующий тип домохозяйства, ни в рамках другого блока, содержащего сопоставимый тип домохозяйства. В данной ситуации необходимо или присвоить домохозяйству-реципиенту признак «не восстанавливается», или же попытаться найти подходящего донора в сопоставимой страте другой территориальной единицы, которая является географически достаточно близкой к той территориальной единице, по которой не удалось найти подходящее для реципиента домохозяйство-донор.

Поиск подходящего донора с использованием метода «прочесывания» основан на сквозном просмотре совокупности домохозяйств-доноров и домохозяйств-реципиентов. При сквозном просмотре каждому домохозяйству-донору, имеющему совпадения по условиям с домохозяйством-реципиентом, приписывается количество совпадающих характеристик. Из всех таких домохозяйств выбирается домохозяйство-донор с максимальным количеством совпадений. Если таких домохозяйств-доноров несколько, то выбор последнего осуществляется случайно. Записи выбранного домохозяйства-донора, как отмечалось выше, являются основой для восстановления пропущенных полей в переписном листе у членов домохозяйств-реципиентов (метод текущих средних применяется для импутации отсутствующего возраста).

В целом реализация метода «прочесывания» осуществлялась в следующем порядке:

- сначала домохозяйства-реципиенты из конкретной совокупности страт упорядочивались по количеству неответов в возрастающем порядке;

- затем для каждого реципиента из совокупности страт осуществлялась проверка заполнения всех основных вопросов. Если все основные вопросы не заполнены или заполнены только первые два вопроса переписного листа (соответственно «Ваше родственное отношение» и «Ваш пол») или не заполнены только невосстанавливаемые вопросы, то реципиенту соответственно присваивался признак «Не восстанавливается». Во всех остальных случаях реципиенту присваивался признак «Восстанавливается», и для каждого из них осуществлялся поиск донора.

На начальном этапе поиск донора осуществлялся в страте, образованной по соответствующему типу домохозяйства в сочетании с группой по его размеру.

Если донор не находился при реализации всех процедур по удалению незначащих вопросов, то на втором этапе осуществлялся поиск донора в сопоставимых стратах укрупненного блока, которому принадлежит данный тип домохозяйств.

При необнаружении донора на втором этапе осуществлялся поиск сопоставимого укрупненного блока, в состав которого включены другие типы домохозяйств, но с частично адекватной структурой начальному типу домохозяйств.

При необнаружении донора на третьем этапе осуществлялся переход в сопоставимые страты других территориальных единиц, географически близких к территориальной единице, где осуществлялся начальный поиск донора.

Если подходящего донора не удавалось найти на всех этапах, то домохозяйству-реципиенту присваивался признак неудачного восстановления. Данная ситуация наблюдалась и при проведении экспериментальных исследований по трем субъектам РФ. Однако в целом количество домохозяйств-реципиентов с невосстановленными ответами является приемлемой величиной и составляет менее 1% (примерно 0,84%).

Эффективность реализованных методов восстановления отсутствующих сведений в записях переписного листа при проведении экспериментальных исследований определялась по трем критериям, характеризующим: *во-первых*, надежность результатов, полученных после восстановления отсутствующих записей; *во-вторых*, общее количество групп по выделенным признакам переписного листа в экспериментальной совокупности с адекватными гипотетической совокупности результатами восстановления искусственно вмененных неответов; *в-третьих*, временные затраты как на поиск подходящего донора для соответствующего реципиента (включая переход в сопоставимые страты и укрупненные блоки), так и на восстановление отсутствующих записей в переписном листе.

Надежность реализованных методов импутации проверялась с помощью двух статистических критериев: критерия хи-квадрат и критерия *Z*. С помощью критерия хи-квадрат проверялась адекватность гипотетического и экспериментального (то есть полученного после импутации) распределений числа лиц в структурных группах основных и ряда вспомогательных признаков переписного листа. С помощью критерия *Z* определялась существенность или несущественность отклонений между относительными частотами в ряде групп основных и вспомогательных признаков переписного листа. Фактические значения полученных критериев сравнивались с их табличной величиной на принятом уровне значимости (q = = 0,05). Гипотеза об адекватности распределений в гипотетической и экспериментальной совокупностях принималась, если фактическое значение критерия хи-квадрат было меньше его табличной величины. Гипотеза о значимости расхождений между i-относительными частотами в гипотетической и экспериментальной совокупностях отклонялась, если табличное значение вероятности, соответствующее фактической величине критерия *Z*, было больше принятого уровня значимости (q = 0,05).

Каждый из указанных критериев взаимно дополняет друг друга, и их совместное использование позволяет более точно оценить надежность реализованных методов импутации.

Второй критерий, характеризующий общее число групп с адекватно восстановленной информацией, определяется на основании анализа структурных составляющих критерия хи-квадрат и оценки результатов сравнения двух относительных частот с помощью критерия *Z*.

В рамках проведенных экспериментальных исследований общее число групп системы анализируемых признаков переписного листа с надежными результатами восстановления искусственно вмененных неответов составляет более 99%.

В таблице представлена по субъектам РФ относительная структура общего числа групп, образованных по признакам, по которым на основании вычисленных критериев было принято решение о признании результатов восстановления отсутствующих сведений в записях переписного листа надежными.

Таблица

Относительное число групп признаков переписного листа форм К(Д), по которым восстановление отсутствующих сведений в записях переписного листа может быть принято надежным
(в процентах)

Наименование субъекта РФ	Относительное число групп с надежными результатами импутации				Итого
	городское население		сельское население		
	форма К	форма Д	форма К	форма Д	
Камчатская область	-	99,9	-	100,0	99,9
Костромская область	99,8	100,0	99,4	99,7	99,7
Рязанская область	99,8	99,8	99,1	99,8	99,7
Итого	99,8	99,9	99,4	99,8	99,8

Как видно из представленной таблицы, на основании первых двух критериев выявлены отдельные группы признаков переписного листа с неадекватно восстановленной информацией. Указанное обусловлено прежде всего недостаточностью информации в донорской базе СИМ для успешного восстановления искусственно вмененных неответов на ряд вариантов ответа на вопросы переписного листа и в целом не может повлиять на общую положительную оценку надежности полученных результатов.

Третий критерий (временные затраты на импутацию данных) оказывает достаточно значимое влияние на выбор эффективного метода импутации. При использовании методов импутации для восстановления искусственно вмененных неответов временные затраты определялись в рамках каждой двумерной страты (с учетом количества домохозяйств в них), отдельно по формам переписного листа, отдельно по трем субъектам РФ. Временные затраты указывались на импутацию как в целом по домохозяйствам-реципиентам, так и в среднем на одно домохозяйство-реципиент. Анализ отчета с данными о временных затратах на реализацию алгоритма импутации показал, что:

- в среднем временные затраты на реализацию алгоритма импутации измеряются в секундах или долях секунд;

- в большинстве случаев подходящего донора для соответствующего домохозяйства-реципиента удалось найти в рамках выделенной двумерной страты;

- переход в сопоставимые страты характерен для ситуаций, когда в двумерной страте не удалось найти подходящего донора. Это явление наблюдается в рамках всех трех субъектов РФ;

- наибольшие затраты времени на реализацию алгоритма импутации наблюдаются в проблемных стратах[9], для которых характерна ситуация, когда не удается найти подходящего донора в рамках сопоставимых страт ни в соответствующем административном районе, ни в географически близком ему другом районе. В стратах, в которых вообще не удалось найти подходящего донора для всех домохозяйств-реципиентов, среднее время его поиска составило 9,7 сек. (указанное характерно лишь для 10 страт);

- в целом по трем субъектам РФ общие временные затраты на реализацию алгоритма импутации составили приблизительно 24 часа.

Таким образом, временные затраты на поиск подходящего домохозяйства-донора для соответствующего реципиента и восстановление в переписных листах последнего отсутствующих записей относятся к приемлемым.

Результаты экспериментальных исследований по использованию методов восстановления пропущенных и несостоятельных полей в переписной информации подтвердили, что реализованные методы импутации в сочетании с методом «прочесывания» для поиска донора являются достаточно надежными и эффективными и их можно применять при обработке первичных данных переписей и обследований населения.

В то же время при реализации разработанных методов импутации на специальных информационных массивах удалось выявить ряд значимых методологических и технологических проблем, решение которых потребует введения в алгоритм новых модулей, их экспериментальной апробации, оценки надежности и эффективности всех процедур, направленных на реализацию методов импутации при обработке данных Всероссийской переписи населения 2010 г.

[9] Проблемная двумерная страта - это страта, в которой возникли трудности при поиске подходящего донора для соответствующего реципиента, что привело к переходу в сопоставимые страты (в том числе и страты географически близких районов) и более высоким временным затратам на реализацию алгоритма импутации, а в отдельных случаях - и к отрицательному результату (то есть поиск донора не увенчался успехом).