ОРГАНИЗАЦИЯ ОБЪЕДИНЕННЫХ НАЦИЙ

E

ЭКОНОМИЧЕСКИЙ И СОЦИАЛЬНЫЙ СОВЕТ

Distr.
GENERAL

ECE/CES/AC.6/2008/9
4 March 2008

RUSSIAN
Original: ENGLISH

ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ

КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ

Совместное совещание ЕЭК ООН/Евростата
по переписям населения и жилищного фонда

Одиннадцатое совещание
Женева, 13-15 мая 2008 года
Пункт 4 предварительной повестки дня

РЕДАКТИРОВАНИЕ И ПРОВЕРКА ДАННЫХ

Обзор методов редактирования и условного расчета для следующих итальянских переписей

Записка Национального статистического института Италии

Резюме

Национальный статистический институт Италии готовит следующие переписи населения и жилищного фонда, которые будут проведены в 2011 году. В настоящем документе приводится краткое описание основных процедур редактирования для данных переписи 2011 года. Кроме того, рассматривается воздействие некоторых из вероятных новшеств в переписи на редактирование и проверку данных.

GE.08-21611 (R) 050508 060508

I. ВВЕДЕНИЕ

1. Характеристики данных переписей превращают этап редактирования и проверки данных в чрезвычайно сложную задачу. Некоторые из этих характеристик сходны с особенностями данных других обследований домашних хозяйств, например: иерархическая структура единиц сбора и анализа данных (домашние хозяйства и физические лица, строения и жилища), а также совокупность многих качественных и количественных переменных. Другие характеристики присущи именно данным переписей, например: колоссальный объем данных, для подсчета которых требуются высокоэффективные решения, и присутствие некоторых мелких, но важных групп населения, с трудом поддающихся переписи (пожилые люди, младенцы, жители приграничных районов, работающие в другой стране, иностранцы и т.д.) и требующих применения методов условного расчета, способных обеспечить учет их фактического распределения. Для надлежащего учета характеристик данных переписей необходимо разработать четко продуманную стратегию редактирования и проверки данных.

2. Для стратегии редактирования и проверки данных, принятой для обработки данных итальянской переписи 2001 года, были характерными методологические и технические новшества, направленные на решение специфических задач редактирования. В частности, в целях повышения качества конечных результатов были разработаны, внедрены и успешно использованы новые методы и программные средства.

3. В настоящее время полностью описать детали стратегии редактирования и проверки данных переписи 2011 года все еще не представляется возможным, поскольку на данном этапе обсуждаются некоторые новшества, затрагивающие этапы редактирования и проверки данных, на предмет их возможного внедрения при проведении следующей переписи.

4. Фактически утверждается, что стратегия редактирования и проверки данных переписи 2011 года будет основана на опыте проведения переписи 2001 года: будет обеспечен комплексный процесс редактирования и проверки данных на базе иерархического сочетания ряда процедур, ориентированных на решение конкретных задач. В частности, некоторые процедуры редактирования, использовавшиеся в 2001 году, будут адаптированы к опросному листу 2011 года и будут вновь использованы, а некоторые другие будут усовершенствованы и/или модифицированы с учетом новшеств, которые будут внедрены. Использование универсального программного обеспечения позволит легко согласовывать данные и оперативно корректировать процедуры с учетом особенностей новой стратегии.

5. Стратегия редактирования и проверки данных 2011 года будет направлена также на уменьшение времени обработки для содействия улучшению своевременности при распространении результатов (EUROSTAT, 2008). В этом вопросе ключевую роль будут играть современные с технологической точки зрения аппаратные средства, высокоэффективные алгоритмы, а также надлежащее планирование, внедрение и организация применения процедур редактирования и проверки данных.

6. В настоящем документе высвечиваются важнейшие компоненты процесса редактирования и проверки данных переписи 2011 года с указанием некоторых особенностей, которые будут учитываться на подготовительном этапе. В частности, в разделе II описываются основные процедуры редактирования данных, которые будут использоваться в ходе переписи 2011 года, а в разделе III анализируется воздействие обсуждаемых новшеств в схеме обследования на стратегию редактирования и проверки данных. И наконец, в разделе IV приводятся некоторые соображения, касающиеся анализируемых в настоящее время вопросов и будущей работы.

II. ОСНОВНЫЕ ПРОЦЕДУРЫ РЕДАКТИРОВАНИЯ ДАННЫХ ПЕРЕПИСИ 2011 ГОДА

7. Стратегия редактирования данных переписи 2001 года была нацелена на осуществление приемлемых условных расчетов в целях сохранения максимального объема собранной информации. Стратегия, принятая на вооружение в целях решения этой задачи, заключается в выделении подзадач редактирования данных и нахождении соответствующего решения для каждой из них. Вследствие этого процесс редактирования данных состоял из ряда (связанных) процедур, нацеленных на решение конкретных подзадач, и были внедрены соответствующие методы (Bianchi *et al*., 2004 и 2005).

8. Исходя из вычислительных и оперативных соображений, редактирование переменных из формуляра для частных домохозяйств[1] в рамках переписи населения 2001 года производилось в два последовательных этапа. Редактирование демографических переменных (год рождения, пол, отношение к основному лицу в домохозяйстве, семейное положение и количество лет в браке) было произведено до редактирования остальных переменных, именуемых индивидуальными переменными (гражданство, присутствие и проживание, уровень образования и подготовки, профессиональный и непрофессиональный статус, трудовая деятельность, место учебы или работы). В результате этого многие индивидуальные ответы могли быть ошибочно

[1] Переписной лист для сбора информации о лицах, обычно проживающих в данном жилище.

изъяты (или неправильно условно рассчитаны) в зависимости от условно рассчитанных значений демографических переменных. Эта проблема была решена путем использования специальной вспомогательной переменной (см. подраздел A), основанной на определении кривой ответов респондентов.

9. Важную роль в обработке демографических переменных сыграли методы оптимизации. Фактически на основе методов оптимизации было разработано новое программное обеспечение для применения основанных на данных и (теоретических) минимальных изменениях подходов, использовавшихся для обработки демографических переменных переписи 2001 года (см. подразделы B и C).

10. Важным этапом в реализации стратегии редактирования индивидуальных переменных стало также определение кривой ответов респондентов. Эти кривые определяют классы респондентов, имеющие общие характеристики (см. подразделы A и D). Эти классы использовались в качестве страт для локализации ошибок и условных расчетов. Для локализации ошибок в индивидуальных переменных по существу использовался детерминистский подход, а новые значения присваивались посредством основанных на модели условных расчетов (см. подраздел D).

11. Особое внимание уделялось редактированию данных по некоторым мелким группам населения, таким, как долгожители, младенцы, жители приграничных районов, работающие в другой стране. Для этих групп были предусмотрены особые процедуры, подразумевающие также увязку полученных данных с данными переписи 1991 года (см. подраздел E).

12. Стратегия редактирования данных переписи 2011 года будет основана на стратегии, принятой в 2001 году. В частности, будет предусмотрен тот же самый порядок обработки данных (сначала демографические переменные, а затем индивидуальные переменные), и будут адаптированы и повторно использованы основные процедуры расчета демографических и индивидуальных переменных. В следующих подразделах будет дано краткое описание методологических аспектов этих процедур.

A. Определение кривой ответов респондентов

13. Согласно графической теории (Picard, 1980), опросный лист можно представить в виде связанного графика, в котором вершины - это переменные, а ответы определяют полюса. Когда ответ на вопрос, представляющий собой переменную, требуется только в отношении некоторых значений предыдущей переменной, первая переменная является фильтром, а последующая - зависимой переменной. Например, семейное положение

представляет собой фильтрующую переменную для такой зависимой переменной, как количество лет в браке. Фильтрующие переменные представлены вершинами с более чем одним полюсом. Каждый из этих полюсов переходит в последующую вершину, представляющую зависимую переменную. Две вершины являются смежными, если они связаны полюсом. Кривая представляет собой последовательно соединенные смежные вершины. Кривая определяет класс респондентов, имеющих общие характеристики. Таким образом, кривая может считаться обобщающей переменной, используемой для определения подсовокупностей респондентов.

14. Отсутствующие ответы и/или ошибочные значения могут затруднить определение правильной кривой для некоторых респондентов и тем самым определение должных классов респондентов. Фактически не прошедшая редактирование запись может дать ошибочную кривую (недопустимую кривую), т.е. кривую, не соответствующую правилам систематизации ответов на опросный лист. В подобных случаях необходимо найти наиболее вероятную (допустимую) кривую (кривые).

15. При проведении переписи 2001 года была использована автоматическая процедура определения наиболее вероятной кривой для каждого респондента. Эта процедура позволяла выявлять наиболее вероятную кривую из множества допустимых на основе анализа ответов, данных на вопросы по фильтрующим и зависимым переменным. Если не прошедшая редактирование запись допускала наличие двух или более допустимых кривых (множественные решения), то на основе наблюдавшегося распределения частотности наиболее вероятных кривых ответов респондентов производилась случайная выборка одной из них.

16. Кривая ответов респондентов использовалась:

a) для расчета новой вспомогательной переменной - подмножества допустимых значений (ПДЗ) года рождения (Manzari et al., 2002). ПДЗ года рождения (далее просто ПДЗ) рассчитывалось для каждого лица в домохозяйстве и определяло подобласть года рождения, соответствующую наибольшему числу данных ответов. ПДЗ использовалось для определения страты источников данных при обработке демографических переменных: лицо в прошедшем редактирование домохозяйстве являлось подходящим источником данных для лица в не прошедшем редактирование домохозяйстве (объект экстраполяции данных) только в том случае, если год его рождения попадал в границы ПДЗ лица - объекта экстраполяции. Поскольку индивидуальные переменные обрабатывались на втором этапе редактирования, их значения могли быть обусловлены условно рассчитанным значением переменной года рождения, и в

этом случае могло произойти неоправданное отбрасывание данных. Использование ПДЗ позволяло условно рассчитывать год рождения, соответствующий наибольшему числу индивидуальных переменных, и тем самым значительно ограничивалась утеря информации, обусловленная неоправданными отбрасываниями данных;

b) для условных расчетов неполученных ответов или несоответствий в данных по переменным, касающимся профессионального статуса, трудовой деятельности и места учебы или работы. В частности, по каждой не прошедшей редактирование записи сначала выбиралась наиболее вероятная кривая, а затем подбиралось значение для условного исчисления с учетом кривой ответов респондентов, а также таких переменных, как место проживания, возрастная группа, пол и другие страты.

B. Совместное использование основанных на данных и на минимальных изменениях подходов в рамках ДИЕСИС

17. Данные переписи населения собираются на уровне домашнего хозяйства с получением информации по каждому члену домохозяйства. Демографические переменные соотносятся между различными лицами в домохозяйстве на основе так называемого редактирования сопоставительных данных, а также на уровне конкретного лица на основе редактирования индивидуальных данных. Таким образом, сохранение взаимосвязей между демографическими переменными можно гарантировать в том случае, если переменные обрабатываются в одно и то же время по всем лицам, принадлежащим к одному и тому же домохозяйству, т.е. на уровне домохозяйства. Для решения этой задачи необходимо использовать систему редактирования, позволяющую обеспечивать редактирование данных на основе сопоставлений между респондентами и на уровне одного респондента. Поскольку некоторые редактируемые демографические переменные выражаются через линейные неравенства, система редактирования должна быть также в состоянии обеспечить одновременную обработку качественных и количественных переменных.

18. При подготовке переписи 2001 года Национальный статистический институт Италии (ИСТАТ) и академические исследователи (факультет информатики и систем Римского университета "Ла Сапьенца") совместно разработали новое программное обеспечение - итальянскую систему условного расчета и редактирования данных (ДИЕСИС) (Bruni *et al.*, 2001). Система ДИЕСИС позволяет одновременно обрабатывать качественные и количественные переменные на уровне домохозяйства и на индивидуальном уровне. После тщательной статистической оценки ее эффективности (Manzari and Reale, 2001)

система ДИЕСИС была успешно использована для условного расчета отсутствующих ответов и согласования расхождений в ответах по демографическим переменным при переписи 2001 года.

19. В системе ДИЕСИС применяются два подхода к редактированию данных - основанный на данных и на "теоретических" минимальных изменениях - в рамках алгоритмов, подразумевающих отбор сначала источников данных, а затем подмножеств переменных и, наоборот, отбор сначала подмножеств переменных, а затем источников данных.

20. При алгоритме, подразумевающем отбор сначала источников данных, а затем подмножеств переменных, сначала определяется подмножество потенциальных источников данных, а затем минимальный набор переменных для условного расчета на основе этих источников. Потенциальными источниками данных выступают прошедшие редактирование домохозяйства, максимально приближенные к не прошедшему редактирование домохозяйству. Сходство между каждым не прошедшим редактирование домохозяйством и каждым прошедшим редактирование домохозяйством калькулируется на основе функции, определяемой в качестве взвешенной суммы отклонений (по количественным переменным) или сходных характеристик (по качественным переменным) применительно к каждой переменной по домохозяйству с охватом всех лиц. При этом алгоритме из потенциальных источников данных выделяется минимальное (взвешенное) множество значений для условных расчетов, с тем чтобы новое скорректированное домохозяйство могло пройти все процедуры редактирования (минимальное изменение с учетом потенциальных источников данных). При использовании этого алгоритма условно рассчитанные значения по домохозяйству получаются на основе одного домохозяйства, выступающего источником данных.

21. При алгоритме, подразумевающем отбор сначала подмножеств переменных, а затем источников данных, определяется минимальное (взвешенное) число переменных для условного расчета, а затем потенциальные источники данных (как это описано выше). Затем по каждому лицу - объекту экстраполяции данных при этом алгоритме для условного расчета берутся значения по лицу - источнику данных, максимально приближенному к лицу - объекту экстраполяции данных. При этом алгоритме поочередно рассчитываются переменные по одному лицу. По возможности одновременно производится условный расчет переменных, относящихся к конкретному лицу. Следует отметить, что условно рассчитанные значения по домохозяйству могут получаться на основе двух или более домохозяйств - источников данных.

22. Оба этих алгоритма использовались совместно для обработки демографических переменных в целях обеспечения баланса между надежностью условных расчетов и сохранением собранной информации. В качестве базового был выбран алгоритм, подразумевающий отбор сначала источников данных, а затем подмножеств переменных и была предусмотрена возможность перехода на алгоритм, подразумевающий отбор сначала подмножеств переменных, а затем источников данных, в тех случаях, когда по конкретному не прошедшему редактирование домохозяйству число изменений, предлагаемых при первом алгоритме, оказывалось слишком большим по сравнению с числом изменений, предлагаемых при втором алгоритме (границы были установлены на основе размеров домохозяйства).

23. Алгоритм, подразумевающий отбор сначала источников данных, а затем подмножеств переменных, использовался в основном для обработки данных по домохозяйствам, имеющим обычную структуру, т.е., как правило, имеющим меньшие размеры. Для этих домашних хозяйств обычно можно было найти достаточно потенциальных источников данных. Напротив, при обработке данных по домохозяйствам, имеющим необычную структуру, т.е. наиболее крупные размеры, как правило, имелось очень мало источников данных и при этом они не были очень сходными с не прошедшими редактирование домохозяйствами. В этих случаях условные расчеты, основанные на данных, требовали бы внесения множества изменений для получения скорректированного домохозяйства, способного пройти все операции по редактированию данных, и предпочтение отдавалось подходу, основанному на минимальных изменениях.

C. Определение основного лица в домохозяйстве и потенциальных пар

24. Перед обработкой демографических переменных были задействованы две важные процедуры. Первая была нацелена на проверку основного лица в домохозяйстве, а вторая - на выявление потенциальных пар в домохозяйстве.

25. Одной из важнейших демографических переменных является отношение к основному лицу в домохозяйстве. Она является базисной переменной для редактирования всех данных на основе сопоставления между респондентами и большинства данных на уровне отдельного лица. Кроме того, необходимо определить ядро семьи и, следовательно, структуру семьи. По каждому домашнему хозяйству необходимо выявить (до редактирования демографических переменных) основное лицо, с тем чтобы позволить всем остальным лицам указать отношение к нему. Применительно к основному лицу могут возникнуть три возможные ошибочные ситуации:

a) одно лицо объявляет себя основным лицом в домохозяйстве, однако его ПДЗ (17 лет или моложе) не согласуется с такой ролью;

b) более чем один человек объявляет себя основным лицом в домохозяйстве;

c) никто в домохозяйстве не объявляет себя основным лицом.

26. Процедура, использовавшаяся для проверки основного лица в домохозяйстве при обработке данных переписи 2001 года, основывалась на методах оптимизации и применялась на основе адаптации алгоритма, подразумевающего отбор сначала подмножеств переменных, а затем источников данных, в рамках системы ДИЕСИС к решению конкретной задачи. Эта процедура состояла из двух основных этапов. На первом этапе выявлялись потенциальные основные лица в домохозяйстве. В частности:

i) в случаях a) и c) потенциальными основными лицами являлись все лица в домохозяйстве, имеющие ПДЗ, сообразующееся с ролью основного лица, т.е. лица в возрасте 18 лет и старше;

ii) в случае b) потенциальными основными лицами в домохозяйстве являлись лица, объявившие себя основным лицом, если их ПДЗ сообразовывалось с ролью основного лица; в противном случае потенциальными основными лицами являлись все лица в домохозяйстве, имевшие ПДЗ, сообразующееся с ролью основного лица.

27. Если ни одно лицо в домохозяйстве не имело соответствующего *ПДЗ*, потенциальными основными лицами являлись все лица вне зависимости от соблюдения соответствующего требования к *ПДЗ*.

28. На втором этапе в рамках этой процедуры производился отбор потенциального основного лица в домохозяйстве, позволявший восстановить согласованность редактирования данных по домохозяйству на основе изменения минимального (взвешенного) числа значений демографических переменных.

29. Определение ядра семьи и, следовательно, структуры семьи основывается на анализе пар в домашнем хозяйстве. Одни пары имеют уникальное отношение к основному лицу в домохозяйстве (основное лицо, жена/муж; отец, мать; тесть/свекор, теща/свекровь). Другие пары имеют неуникальное отношение к основному лицу в домохозяйстве (например, сын/дочь, зять/невестка). Лица, формирующие уникальные пары, если они существуют, хорошо поддаются определению в домохозяйстве. В 2001 году

редактирование данных по ним производилось посредством набора специальных операций по редактированию в сопоставлении между лицами (редактирование данных по парам). Более сложная проблема сопряжена с редактированием данных по неуникальным парам, поскольку сначала необходимо выявить формирующих их лиц.

30. В целях максимально возможного сохранения данных по отслеженным неуникальным парам в рамках переписи 2001 года была применена двухэтапная стратегия, основанная на канадской стратегии (Bankier *et al*., 1997; Bankier, 1999). На первом этапе до редактирования выявляются неуникальные пары (потенциальные пары). В частности, выявление потенциальных пар осуществлялось путем присвоения баллов каждой возможной паре лиц в домохозяйстве и отбора пар с наивысшим числом баллов. Присвоение баллов производилось на основе ответов на вопросы по демографическим переменам и отражало вероятность того, что два лица составляют пару.

31. На втором этапе производилось редактирование данных по парам в отношении только лиц, формирующих потенциальные пары (компоненты). С этой целью использовалась специальная производная вспомогательная переменная. Потенциальная пара могла быть сохранена или исключена посредством условных расчетов. В первом случае компоненты имели соответствующие значения по таким переменным, как отношение к основному лицу в домохозяйстве, пол, возраст, семейное положение и количество лет в браке, как и в случае компонентов уникальных пар. Во втором случае ограничивающих критериев компонентов для пар не существовало.

D. Редактирование индивидуальных переменных

32. Индивидуальные переменные относятся только к одному лицу и проходят проверку при редактировании данных на уровне одного лица, вследствие чего они обрабатываются на индивидуальном уровне (за исключением переменной гражданства). Процесс редактирования состоит из различных процедур, адаптируемых к анализируемой единице (частное домохозяйство, институциональное домохозяйство, временно проживающее лицо) и к конкретной измеряемой позиции (место рождения, гражданство, присутствие и проживание, уровень образования и подготовки, профессиональный и непрофессиональный статус, трудовая деятельность, место учебы или работы).

33. Сначала проводился анализ ошибочных ситуаций для классификации типов ошибок (например, систематические ошибки кодирования или компиляции, случайные ошибки). Затем с использованием различных методов производился условный расчет непредставленных ответов и несоответствий. В частности, в случае систематических ошибок кодирования применялись детерминистские условные расчеты, а для условного

расчета непредставленных ответов и несоответствий, обусловленных случайными ошибками, применялся стандартный метод отбраковки. Условно рассчитанные значения получались на основе функций распределения, рассчитанных по прошедшим редактирование записям, сгруппированным в страты, определяемые также кривыми ответов респондентов (см. подраздел A).

E. Проверка данных по долгожителям

34. Долгожители составляют относительно небольшую, но важную группу в общем населении. Стратегия, которая использовалась для проверки данных по долгожителям, охваченным переписью населения 2001 года, представляет собой комбинацию нескольких методологий, непосредственно применяемых к необработанным микроданным, установленным путем оптической считки (Nuccitelli *et al.*, 2006). Главными этапами осуществления этой стратегии являются следующие:

1) автоматическое совмещение данных, касающихся лиц, родившихся в период 1888-1901 годов и охваченных переписью 1991 года, с данными, касающимися лиц, родившихся в тот же период и охваченных переписью 2001 года, с использованием точной (т.е. детерминистской) привязки;

2) автоматическая проверка на предмет внутреннего соответствия данных, не увязанных на этапе 1;

3) ручная проверка на предмет соответствия соответствующим изображениям опросных листов, которые сдаются на хранение в процессе оптической считки, некоторых неоднозначных случаев, выделенных на этапе 2.

35. Что касается этапа привязки, то две записи считались привязанными, если одновременно выполнялись следующие условия:

a) эти записи стопроцентно согласуются по областям, общим для обоих наборов данных (переменные привязки);

b) использованный ключ привязки определяет уникальные связи и не содержит отсутствующих значений.

36. Лица (охваченные переписью 2001 года), соответствующие увязанным записям, считались прошедшими проверку. Неувязанное лицо, представившее ответы по

соответствующим позициям при автоматической проверке на этапе 2, считалось прошедшим проверку.

III. ПРОЦЕСС РЕДАКТИРОВАНИЯ И ПРОВЕРКИ ДАННЫХ С УЧЕТОМ НОВШЕСТВ, ВНЕСЕННЫХ В СХЕМУ ОБСЛЕДОВАНИЯ

37. В настоящее время в ИСТАТ обсуждаются некоторые новшества, вносимые в схему проведения переписи 2011 года (Crescenzi and Fortini, 2007).

38. В числе тех новшеств, которые скорее всего будут внедрены, наиболее значительное воздействие на этапы редактирования и проверки данных окажут следующие:

a) перепись с использованием коротких и развернутых формуляров опросных листов;

b) наличие регистров:

i) данные по отдельным лицам, обычно проживающим в муниципалитете, и их организации в частных и институциональных домохозяйствах - получаются на основе местных регистров населения;

ii) обобщающие данные по отдельным лицам, обычно не проживающим, но временно находящимся в муниципалитете, и их организации в частных и институциональных домохозяйствах - получаются на основе данных из местных или центральных вспомогательных источников;

iii) местные адресные книги;

c) использование множественных способов сбора данных: опросные листы, собираемые счетчиками, передача с помощью почты или вебтехнологий или сбор информации с помощью телефонных опросов с использованием компьютерных систем (ТОКС).

39. В рамках переписи населения 2011 года можно было бы задать ограниченное число вопросов (главным образом по демографическим характеристикам), касающихся каждого лица и единицы жилья, а дополнительные вопросы (по социально-экономическим характеристикам) могли бы касаться выборки лиц или единиц жилья. В этом случае необходимы два формуляра переписных листов. Короткая форма опросного листа могла бы содержать информацию по демографическим переменным (пол, год рождения,

семейное положение, отношение к основному лицу в домохозяйстве, гражданство, присутствие и проживание, уровень образования). В то же время развернутая форма опросного листа могла бы содержать информацию по переменным, включенным в короткую форму, а также по таким переменным, как обучение в школе, уровень образования и профессиональная подготовка, профессиональный статус, трудовая деятельность, место учебы или работы и время проезда до места учебы или работы. Каждое домашнее хозяйство должно получить либо короткую, либо развернутую форму опросного листа.

40. Выбор в пользу обработки демографических переменных (все они содержатся в короткой форме) перед обработкой индивидуальных переменных вполне вписывается в новую стратегию переписи. Вместе с тем может возникнуть необходимость определения *ПДЗ* (см. раздел II.A) с использованием двух разных процедур: одной для короткой формы, а другой для развернутой формы. Следует отметить, что для лиц, получающих короткую форму, *ПДЗ* может оказаться менее эффективным, поскольку при этом значительно ограничивается доступная информация, используемая для расчета *ПДЗ*.

41. Использование развернутой формы на выборочной основе может оказать негативное воздействие на доступность (и следовательно выбор) источников данных по мелким группам совокупного населения. Это может снизить точность методов условного расчета (основанных на источниках данных и основанных на распределении), поскольку это в значительной степени зависит от наличия большого числа потенциальных источников данных. Высокие показатели отсутствия ответов на те или иные вопросы в сочетании с ограниченной доступностью источников данных могут значительно снизить качество окончательных оценок. Эту проблему можно было бы снять посредством особенно тщательного управления этапами сбора данных и отбора круга лиц - источников данных.

42. Кроме того, в результате сбора данных на выборочной основе усложнится этап проверки индивидуальных переменных. Необходимо иметь оперативный доступ к более высоким весовым коэффициентам для испытания и настройки процедур редактирования и сопоставления окончательных данных с макроинформацией, поступающей из внешних источников или из административных файлов. В этом отношении эффективность этапа проверки данных можно было бы повысить, если бы имелся подробный набор показателей (контрольная система) для подкрепления углубленного анализа несоответствий между данными переписи и данными из других имеющихся источников. Поэтому следует тщательно спланировать и внедрить контрольную систему.

43. Значительный вклад в повышение качества окончательных данных можно было бы внести посредством обеспечения доступности некоторых видов информации (например,

пол и дата рождения), содержащейся в местных регистрах, которые ведутся муниципалитетами (пункты b1-b3). Информация из этих регистров обычно является точной и поэтому может быть использована:

a) для улучшения количественного контроля за формами, т.е. для выверки соответствия между числом возвращенных заполненных опросных листов и "ожидаемым" числом опросных листов в целях уменьшения недоучета;

b) для заполнения переписных файлов посредством условного расчета отсутствующих или противоречивых значений данных переписи.

44. В этом случае необходимо провести определенные исследования, нацеленные на определение эффективных стратегий использования макро- и микроданных регистров.

45. В частности, увязка данных переписи и регистровых данных могла бы обеспечить возможность для точного условного расчета отсутствующих или противоречивых значений данных переписи. Трудности, связанные с увязкой данных переписи и регистровых данных, могут возникать главным образом из-за отсутствия надежной идентифицирующей информации (в ходе предыдущих переписей не была собрана информация об индивидуальных номерах каждого налогоплательщика и у ИСТАТ отсутствовала информация об их полных именах), а также из-за фактора времени, необходимого для их обработки. Если информация об индивидуальном номере налогоплательщика будет получена в формуляре переписи 2011 года, то его можно было бы использовать в качестве уникального идентификатора (перменной привязки) для увязки с регистровыми данными. В этом случае, поскольку опросные листы будут заполняться в основном самими респондентами, для проверки правильности переменной привязки необходимо получить также информацию о полном имени. В противном случае на этапе привязки могут возникнуть ошибки, которые могут привести к утрате полезности или неправильному использованию увязанных данных в целях проведения условных расчетов.

46. В качестве альтернативной стратегии регистровые данные могли бы добавляться в записи источников информации и могли бы использоваться условные расчеты, основанные на таких источниках данных. Эта стратегия нацелена на увеличение сходства между данными по объектам экстраполяции данных и источникам информации (данные по лицам - источникам информации могут довольно сильно отличаться от некоторых данных по объектам экстраполяции данных) и, следовательно, на проведение условных расчетов. На основе специального исследования следует ценить точность и эффективность характеристик этих двух альтернативных видов использования

регистровых данных (условные расчеты на основе привязки и условные расчеты на основе расширенных источников данных).

47. И наконец, отнюдь не ничтожное воздействие на количественный контроль за формами оказывает множественность способов сбора данных. В целях недопущения переучета необходимо предусмотреть процедуры, нацеленные на проверку опросных листов-дубликатов в результате использования различных способов сбора данных.

IV. ВЫВОДЫ

48. Процессы редактирования и проверки данных при проведении следующей переписи населения будут основываться в основном на опыте проведения переписи 2001 года, но при этом необходимо предусмотреть соответствующие стратегии с учетом тех новшеств, которые, вероятнее всего, будут внесены в схему обследования. Особого внимания заслуживают следующие аспекты:

a) воздействие методики использования коротких форм на выявление кривых ответов респондентов, на отбор лиц - источников информации и на этап проверки данных;

b) возможность использования информации из местных регистров населения или других интегрированных источников на макро- и микроуровне;

c) наличие местных адресных книг;

d) использование множественных способов сбора данных.

49. В частности, возможности, обеспечиваемые в связи с вышеуказанными пунктами b)-d), позволяют производить редактирование в процессе сбора данных. Это требует внедрения подходящих систем поддержки.

50. Эти новшества будут иметь соответствующие последствия для всего процесса редактирования и проверки данных. На данном этапе некоторые вопросы уже были проанализированы и были предложены решения, но при этом необходимо урегулировать ряд других вопросов. Кроме того, в настоящее время проводятся некоторые исследования в целях сокращения расчетного времени в рамках процесса количественного контроля и повышения эффективности выявления систематических ошибок. Поскольку требуется улучшить своевременность распространения результатов, задача редактирования и проверки данных, вне всякого сомнения, является амбициозной.

СПРАВОЧНАЯ ЛИТЕРАТУРА

Bankier M., Houle A. and Luc M. (1997) 1996 Canadian Census Demographic Variables Imputation, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Czech Republic (Prague).

Bankier M. (1999) Experience with the New Imputation Methodology used in the 1996 Canadian Census with Extension for future Censuses, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Italy (Rome).

Bianchi G., Pezone A., Reale A., Saporito G. (2004) Metodi e Procedure per il Controllo e la Correzione delle Variabili Demografiche Familiari del Censimento della Popolazione 2001, *Internal document* (*in italian),* ISTAT.

Bianchi G., Manzari A., Pezone A., Reale A., Saporito G. (2005) New procedures for editing and imputation of demographic variables, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Canada (Ottawa).

Bruni R., Reale A., Torelli R. (2001) Optimization Techniques for Edit Validation and Data Imputation, *presented at the Statistics Canada Symposium 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective" XVIIIth International Symposium on Methodological Issues.*

Crescenzi F. and Fortini M. (2007) Due strategie per l'uso censuario di dati anagrafici, *Internal document* (*in italian),* ISTAT.

EUROSTAT (2008) Proposal for a Regulation of the European Parliament of the Council on population and housing censuses. Brussels, 11 January 2008.

Manzari A. and Reale A. (2001) Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology, In *IASS Proceedings 53rd Session of The International Statistical Institute, August 22-29, 2001, Seoul,* pp. 634-655.

Manzari A., Pezone A., Reale A. (2002) Evaluation of a new approach for edit and imputation of social and demographical data with hierarchical structure, *Atti della XLI Riunione Scientifica SIS*, Milano, 5-7 Giugno 2002, Sessioni spontanee, pp 689-692.

Nuccitelli A., Pezone A., Reale A., (2006) The Validation of the Census Micro-Data on the Oldest Old Living in Italy, *Proceedings of Q2006 European Conference on Quality in Survey Statistics*, United Kingdom (Cardiff).

Picard C. F. (1980) *Graphs and questionnaires*, North-Holland, Netherlands.