Неофициальный перевод Статкомитета СНГ

ОРГАНИЗАЦИЯ ОБЪЕДИНЕННЫХ НАЦИЙ ECE/CES/2012/56
ЭКОНОМИЧЕСКИЙ И СОЦИАЛЬНЫЙ СОВЕТ

Распространение: Общее
15 мая 2012 г.
Оригинал: английский

Европейская экономическая комиссия ООН

Конференция европейских статистиков

60-е пленарное заседание
Париж, 6-8 июня 2012
2010 раунд переписи – инновации и уроки

Новые технологии при проведении переписи населения и жилого фонда в Албании

Записка Национального института статистики Албании

Резюме

В этом документе говорится о новых технологиях сбора и обработки данных переписи населения и жилого фонда, примененных Национальном институтом статистики Албании. В результате использования новых технологий сканирования эффективность ввода данных заметно повысилась. Эти снизило затраты на проведение переписи а также ускорит публикацию первых результатов переписи. Кроме того, были использованы новые методы управления полевыми операциями, такие как SMS сообщения и система мониторинга на базе GIS, которые хорошо себя зарекомендовали. Приобретенный опыт и вложения в технологию будут использоваться Институтом статистики для проведения будущих переписей и других обследований.

I. Исходная информация

1. Национальный институт статистики Албании (INSTAT) в течение многих лет использует передовые технологии на всех этапах проведения переписи населения. Хотя некоторые их этих технологий использовались и в прошлом, они претерпели дальнейшее развитие в ходе последней переписи населения и жилого фонда.

2. Большое влияние на качество собираемых данных оказала хорошо спланированная структура управления полевыми операциями. Это также способствовало осуществлению мониторинга переписи и сокращению затрат, поскольку дало возможность 1) тщательно планировать материальные ресурсы и кадры для каждой области; 2) ограничить непредвиденные события в ходе общей переписи. Поскольку перепись населения это крупное мероприятие, где фактор времени играет очень большую роль, использование Системы управления переписью очень важно.

3. До 2010 г. INSTAT использовал ручной ввод данных с переписных листов в систему баз данных. В 2009 г. были установлены системы сканирования для поддержки проведения будущей переписи несельскохозяйственных предприятий. Работа была посвящена установке и внедрению системы сканирования для пилотной переписи в апреле 2010 года и затем для Переписи несельскохозяйственных предприятий. Конфигурация системы позволяет использовать ее для всех обследований и переписей, проводимых INSTAT. Цель состояла в том, чтобы обеспечить высокое качество и сократить время на ввод данных. Эта деятельность была поддержана разными донорами и международными экспертами.

4. В Албании этот подход использовался INSTAT для Переписи населения 2011 г. Эта система позволила отсканировать и обработать 20 миллионов переписных листов за 5,5 месяцев.

II. Система управления и мониторинга переписи

5. Поскольку перепись населения – это крупное мероприятие, где фактор времени играет очень большую роль и которое состоит из многих взаимосвязанных операций, использование современной Системы управления переписью очень важно. Эта система была разработана для печати и доставки описаний и этикеток для коробок, содержащих переписные листы, для формирования коробок для отправки в поле, для отслеживания транспортировки коробок из INSTAT в разные офисы переписи и обратно в INSTAT в конце полевых работ, и, наконец, для проверки получения всех материалов. Инструмент для управления переписью был также использован для подготовки коробок и материалов для обследования по почте. Кроме того, этот инструментарий и данные были использованы для правильной организации и управления операциями по обработке данных.

6. Для мониторинга охвата при полевых работах впервые использовались SMS сообщения.

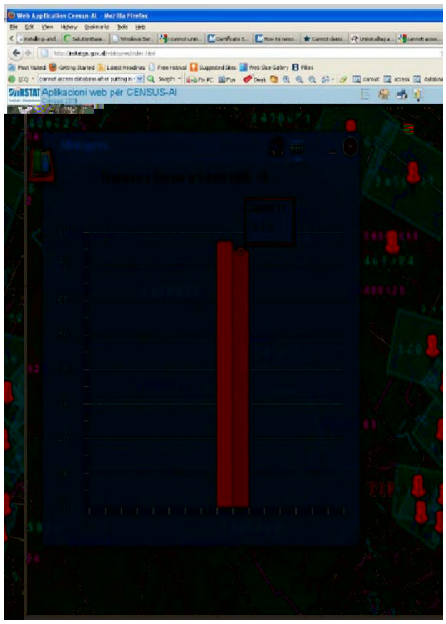
7. Каждому участнику полевых работ была выдана SIM карта для бесплатной связи между сотрудниками. Это было полезно для оперативного решения проблем, возникающих в ходе опроса. Кроме того, переписчики и инструкторы посылали SMS каждое утро на короткий номер с отчетом об общем количестве опрошенных в предыдущий день лиц и домохозяйств. Веб-приложение Географическая информационная система (GIS) была разработана и ежедневно обновлялась на основе данных, полученных по SMS.

8. Во время проведения полевых работ в октябре INSTAT проводил мониторинг охвата переписи с использованием систем отчетов на базе SMS и интерактивной системы мониторинга на базе GIS.

9. Система мониторинга на базе GIS являлась важным инструментом для контроля хода переписи со стороны супервайзеров. Все супервайзеры имели свои учетные записи для входа в систему мониторинга на базе GIS, что давало им возможность наблюдать за работой их персонала на участках. Информация, сообщавшаяся каждым переписчиком по SMS, автоматически сохранялась в центральной базе данных, а в системе мониторинга на базе GIS ежедневно обновлялись результаты по каждому переписчику.

Рисунок 1
Система мониторинга на базе GIS


HARTA E ZONËS SË CENSUSIT
Censusi i Popullsisë dhe Banesave 2011
INSTAT
Instituti i Statistikave
TIRANË
TIRANË
Shkalla 1:630
CENSUS-AL


Aplikacioni web për CENSUS-AL

10. Кроме того, в течение переписного периода INSTAT ежедневно выпускал бюллетени с отчетами на основе данных, полученных накануне по SMS (см. рис.1). В отчетах были показаны кумулятивные данные по району и по муниципалитету. Система мониторинга на базе GIS показывала данные по малым областям, что давало возможность определить географические участки, где процесс шел неудовлетворительно.

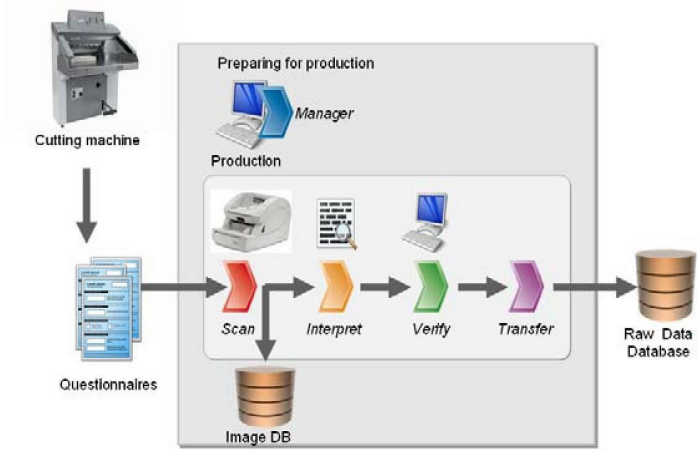
11. Данные, сообщенные по SMS и загруженные в приложение GIS, каждый день сопоставлялись с данными, сообщаемыми региональными офисами INSTAT. На первом этапе переписи наблюдались некоторые различия в данных, сообщаемых обеими системами. Эти различия были вызваны неполучением отчетов из областей, где не было сотовой связи, ошибками сообщений в обеих системах и расхождениями в оценке количества жилых помещений в двух системах. Эти различия практически исчезли на следующих этапах переписи.

III. Технология сканирования

12. Перепись населения и жилого фонда является масштабным проектом по сбору данных, который охватывает всю страну. Для сокращения времени на ввод данных необходимо сделать выбор между использованием большого количества операторов по вводу данных и развертыванием новой технологии.

13. Решения о применении ручного или автоматического ввода данных были основаны на требованиях графика и на соображениях о затратах на персонал и на оборудование. Также рассматривалось, было ли разумно или возможно применить более сложные технологии.

Рисунок 2
Процесс ввода данных с использованием интеллектуальной системы распознавания символов.



14. Технология сканирования изображений использовалась для ввода данных с переписных листов с минимальным участием человека: после сканирования, изображения сохраняются и передаются в Интеллектуальную систему распознавания образов (IMR/ICR), которая пытается определить содержание ответа. В зависимости от степени уверенности процесса распознавания, система IMR/ICR либо принимает подразумеваемый результат, либо отклоняет его (см. Рис.2).

15. Даже при использовании сложных программ интерпретации, необходимо некоторое количество операторов для распознавания изображений, и по понятным причинам они все не могут быть высококвалифицированными. Следовательно, на этом этапе имеется два важных требования: проверка должна быть простой и быстрой, и операторы не должны вносить дополнительных ошибок. Решение, принятое INSTAT, состояло в так называемой «массовой верификации»: изображения всех интерпретируемых символов были представлены в группе в соответствии с их значением. Если символ появляется в неправильной группе, верификатор отбирает его, и он будет автоматически помещен в поле, где возник этот символ, для исправления. Сначала цифры подвергаются массовой верификации, затем буквы и, наконец, поля с отметками (см. Рис.3).

16. В дополнение к массовой верификации, протекает параллельный процесс для обеспечения качества: выбранная группа операторов проверяла второй раз все индивидуальные переписные листы, где было

отмечено хотя бы четыре несогласованности. Специальное приложение было разработано для того, чтобы показывать изображения переписных листов и значения, сохраненные в разных полях, с выделением значений, которые образовывали несоответствия. Здесь задача состояла не в том, чтобы исправить ошибки переписчика, а в том, чтобы убедиться, что несоответствия не были вызваны неправильной интерпретацией системы IMR/ICR.

Рисунок 3.

Сканирование переписных листов, заполненных вручную с операторами для распознавания изображений

17. На основе требований INSTAT компания “ReadSoft” разработала программное обеспечение для контроля и управления крупномасштабными операциями по сканированию. Основной задачей нескольких сотрудников INSTAT, начиная с апреля месяца, состояла в том, чтобы тестировать, исправлять неполадки и сообщать о проблемах разработчикам из ReadSoft. Несмотря на некоторые проблемы вначале, когда в ноябре началось сканирование переписных листов переписи населения и жилого фонда, программное обеспечение работало достаточно стабильно.

18. Правильный дизайн переписного листа выполняет две задачи: облегчает респондентам заполнение форм и ведет к минимизации ручного руда при считывании информации. Поэтому хорошо, когда в разработке дизайна форм участвуют специалисты по сканированию, которые понимают, какие технические аспекты следует принимать во внимание. Переписные листы имели четкую планировку, достаточное пространство для ответов, большой размер шрифта и соответствующие разрывы страниц. Вопросы-фильтры, предназначенные для разных подгрупп и для переходов, были выделены цветом.

19. Номер серии был напечатан на всех страницах переписного листа. Уникальный идентификатор переписных листов помогал при вводе данных, особенно в том случае, если лист был по ошибке отсканирован более одного раза.

20. Во время печатания переписных листов была введена процедура, в соответствии с которой случайная выборка ежедневного тиража проверялась системой сканирования. Несмотря на то, что каждый переписной лист проходил этот прагматический тест, в конце операций было обнаружено несколько неувязок в ограниченном количестве напечатанных вопросников. В системе сканирования был предусмотрен обходной путь для решения таких вопросов. Всего было напечатано около 1,200,000 буклетов (6 отдельных вопросников).

21. Когда используется система работы с изображениями, необходимо обучать переписчиков как правильно заполнять формы, для того чтобы система ICR могла правильно распознавать текст, написанный от руки. INSTAT проводил специальное занятие во время обучения переписчиков о том, как правильно записывать ответы в полях формы, какой ручкой пользоваться и пр. Было потрачено много времени и усилий, чтобы обеспечить как можно более аккуратное заполнение форм и их возврат в хорошем состоянии.

22. В сентябре 2011 г. был организован новый Центр обработки данных, где было установлено 60

персональных компьютеров, 10 компьютеров с улучшенной производительностью и 7 оптических сканеров. Центр ввода данных состоял из пяти подсистем. Каждая подсистема имела один сканер и 12 компьютеров для верификации данных, соединенных в базой данных, так что в случае возникновения проблем потеря производительности не превысила бы одну пятую часть от максимальной. Это оказалось правильным подходом: если были проблемы в одной подсистеме, другие продолжали работать нормально.

23. Успех операций по вводу данных до определенной степени зависит от квалификации персонала, участвующего в процессе. INSTAT организовал обучение для повышения осведомленности персонала об их обязанностях и важности четкого выполнения работы. Отдельные занятия были проведены для операторов сканеров по эксплуатации оборудования и использованию программного обеспечения, для операторов ручного ввода, обучая их процедуре массовой верификации и другим процедурам ПО, для верификаторов, знакомя их с процедурой выделения несогласованностей и общей стратегией валидации.

IV. Первая оценка качества данных

24. К концу апреля INSTAT завершил ввод информации с переписных листов, и началась работа по очистке данных. Процедура очистки данных состояла из 1) локализации ошибок в собранных данных, 2) применения детерминированных правил в специальных программах для коррекции систематических ошибок, 3) импутации для исправления оставшихся ошибок и пропущенных значений.

25. По результатам такого подхода будет проведен анализ влияния процедур редактирования и импутации путем сравнения первоначального и финального распределения.

26. На этом этапе процедура очистки была разработана и апробирована на случайной выборке в 51 000 человек (13 000 домохозяйств, около 2% от итога). Был разработан детерминированный шаг, состоящий из 19 детерминированных правил.

27. Вероятностный шаг был основан на определении 131 явных логических противоречий для редактирования, которые составили полный набор (явные и неявные) из 439 случаев для редактирования. Применение этого набора к случайной выборке выявило следующие результаты: 1) 27,306 точных записей (53.02%), и 24,189 записей с ошибками (46.97%).

28. Результаты подтверждают общее хорошее качество данных переписи и обработки данных в целом.

V. Выводы

29. Опыт Албании в проведении переписи населения и жилого фонда показывает, что сканирование является относительно дешевым способом существенного ускорения обработки данных. При численности населения приблизительно 2,850,000 человек мы может ожидать получение полного массива очищенных данных на несколько месяцев раньше, чем при ручном вводе, при привлечении небольшого количества сотрудников. Это приведет к гораздо более быстрому получению результатов переписи: ожидается, что окончательные таблицы будут сформированы через 9-10 месяцев после завершения работы переписчиков.

30. Опыт использования SMS сообщений и Системы мониторинга на основе GIS был очень ценным и полезным для выявления проблем, требующих оперативного решения.

31. Наконец, вложения в смысле аппаратной базы, программного обеспечения и знаний, будут использованы при проведении будущих переписей и других обследований. INSTAT планирует проведение сельскохозяйственной переписи в октябре 2012 г. с использованием тех же технологий, как в переписи населения и жилого фонда.