Distr.
GENERAL

CES/SEM.40/9
14 September 1998

RUSSIAN
Original: ENGLISH

СТАТИСТИЧЕСКАЯ КОМИССИЯ и ЕВРОПЕЙСКАЯ ЭКОНОМИЧЕСКАЯ КОМИССИЯ

СТАТИСТИЧЕСКОЕ УПРАВЛЕНИЕ ЕВРОПЕЙСКИХ СООБЩЕСТВ (ЕВРОСТАТ)

КОНФЕРЕНЦИЯ ЕВРОПЕЙСКИХ СТАТИСТИКОВ

Совместная рабочая сессия ЕЭК-Евростата
по переписям населения и жилищ 1/
(Дублин, Ирландия, 9-11 ноября 1998 года)

Тема 2

КОНТРОЛЬ КАЧЕСТВА И ПРОВЕРКА ДОСТОВЕРНОСТИ РЕЗУЛЬТАТОВ

Специальный документ, представленный Управлением национальной статистики Соединенного Королевства 2/

I. Введение

1. Существует большое число источников потенциального возникновения ошибок в ходе обработки результатов переписи. В настоящем документе основное внимание уделяется роли, которую призваны играть процедуры контроля качества и проверки достоверности результатов в уменьшении влияния этих источников потенциальных ошибок на результаты переписи населения 2000 года Соединенного Королевства.

1/ Обработка документации для этой рабочей сессии будет вестись на уровне семинара.

2/ Автор: Ник Парсонз, Группа качества данных Отдела переписей.

2. Контроль качества и управление качеством представляют собой фундаментально различные подходы к управлению качеством того или иного процесса. Внедрение контроля качества призвано обеспечить соблюдение минимально приемлемых стандартов качества по каждой единице выпуска того или иного процесса. Управление качеством предусматривает оценку качества выборки продуктов того или иного процесса. Его целью является непрерывное повышение среднего качества продуктов соответствующего процесса во времени благодаря внесению усовершенствований в данный процесс. При обработке результатов переписи наиболее оптимальным решением является использование комбинации двух этих подходов.

3. Управление качеством, как правило, не предусматривает повторной обработки с целью исправления ошибок, в то время как целью контроля качества является недопущение присутствия определенных ошибок в любой единице выпуска. Хотя управление качеством является важным компонентом нашей стратегии организации переписи 2001 года, основное внимание в настоящем документе уделяется контролю качества.

4. Цель проверки достоверности результатов заключается в выявлении систематических погрешностей в данных, которые не были выявлены в результате использования процедур контроля качества или управления качеством.

5. Наша цель заключается в сведении до минимума уровня значительной систематической погрешности в результатах переписи 2001 года за счет использования комбинации методов контроля качества, управления качеством и проверки достоверности результатов.

6. В разделе 2 описываются основные процессы обработки в качестве введения к остальной части документа. В разделе 3 анализируются потребности в процедурах контроля качества результатов обработки и некоторые меры контроля качества, которые мы планируем использовать в 2001 году. В разделе 4 приводится краткое описание подхода, который мы будем использовать для управления качеством процессов ручной обработки, и те этапы, на которых мы будем использовать методы управления качеством в рамках автоматизированных процессов обработки. В разделе 5 описывается роль, которую призвана сыграть проверка достоверности данных в обеспечении качества результатов переписи 2001 года.

II. Справочная информация: краткое описание процессов обработки

7. Перепись 2001 года будет проводиться путем раздачи опросных листов респондентам, которые те должны возвратить после заполнения в переписные органы по почте. За организацию работы на каждом переписном участке (ПУ), в состав которого входит около 200 адресов, будет отвечать отдельный счетчик. В обязанности счетчика входит определение всех домохозяйств в рамках своего переписного участка и раздачи им опросных листов. Каждое домохозяйство, входящее в состав переписного участка, должно идентифицироваться исключительно с помощью регистрационного номера (РН).

8. После критического момента переписи счетчики будут производить опрос тех домашних хозяйств, которые не прислали заполненные опросные листы по почте.

9. Полученные по почте опросные листы не будут объединяться с листами, заполненными методом личного опроса. Единицей отчетности на большинстве этапов обработки будет являться опросный лист (или комплекты опросных листов), а не переписной участок. Данные по тому или иному участку будут обобщаться в предварительные итоги до этапа редактирования и условных расчетов.

Получение

10. Целью данного процесса является регистрация опросных листов, полученных центром обработки. Опросные листы будут поступать в центр обработки из двух источников: от респондентов по почте и счетчиков, собравших опросные листы, не отосланные по почте.

Сканирование

11. Сканирование и оптическое распознавание являются двумя процессами, которые мы будем использовать для ввода ответов с опросных листов. Сканирование представляет собой процесс, в ходе которого создается изображение каждой страницы опросного листа. Эти изображения используются затем для автоматического и ручного кодирования ответов, заменяя собой, таким образом, физические опросные листы.

Оптическое распознавание

12. Оптическое распознавание представляет собой процесс ввода помеченных клеток ответов и письменных ответов, созданных с помощью сканирования изображений, результаты которого используются для создания первого набора данных переписи в электронном формате.

13. Вопросы, включенные в опросный лист переписи 2001 года, делятся на три типа: вопросы, предусматривающие единственный ответ путем пометки соответствующей клетки, комбинированные вопросы, предусматривающие пометку соответствующей клетки/ предоставление письменного ответа, и ответы, предусматривающие только письменный ответ. Ввод ответов респондентов на "клеточные" и "комбинированные" ответы производится с помощью метода оптического распознавания меток (ОРМ). Ввод письменных ответов, в том числе письменных ответов на комбинированные вопросы, производится с помощью метода оптического распознавания символов (ОРС).

Автоматическое кодирование

14. Письменные ответы, преобразованные в электронный формат с помощью ОРС в ходе процесса распознавания, сопоставляются с перечнем ответов. В случае совпадения ответу, созданному в процессе распознавания, присваивается соответствующий этому ответу код перечня.

Полуавтоматическое и экспертное кодирование

15. Ответы, которые не поддаются автоматическому кодированию, будут кодироваться методом полуавтоматического кодирования. Соответствующий сотрудник будет осуществлять кодирование ответа с помощью интерфейса, включающего в себя изображение ответа и автоматизированный пакет кодирования. Ответы, которые не поддаются полуавтоматическому кодированию, будут кодироваться экспертами с использованием различных процедур или справочных материалов.

Редактирование и условные расчеты

16. Процесс редактирования будет в значительной степени автоматизирован. Целью процесса редактирования является решение проблем, возникающих в случаях, когда респондент пометил несколько клеток ответов на один вопрос, и определение для проведения условного расчета элементов данных, не удовлетворяющих требованиям проверки на непротиворечивость.

17. Целью условных расчетов является присвоение величин всем элементам данных в случаях непредоставления ответов на вопросы переписи. Условные расчеты также будут использоваться в отношении данных, отбракованных в ходе проверок на непротиворечивость. Речь идет о замене одного или нескольких противоречивых элементов данных непротиворечивыми величинами.

Процессы, осуществляемые после этапа редактирования и условных расчетов

18. В настоящее время ведется изучение практических возможностей корректировки базы данных переписи 2001 года с использованием информации об охвате, полученной в рамках контрольного обследования, которое будет проведено после завершения переписи. Этот проект носит название "Единая оценка".

19. Процедуры обезличивания данных, целью которых является недопущение идентификации частных лиц на основе данных переписи, будут осуществлены до того, как данные переписи 2001 года будут переданы в Отдел материалов переписи для распространения.

Ручная обработка

20. Наши сотрудники, как и все люди, обладают различным опытом и знаниями, в связи с чем они могут по-разному интерпретировать или понимать один и тот же ответ. Это особо касается письменных ответов на вопросы переписи, в которых используются концепции, являющиеся сложными или не знакомыми для сотрудников, занимающихся кодированием ответов. Следовательно, ручная обработка может также служить источником ввода в данные определенной случайной ошибки.

21. Эти два фактора означают, что результаты будут страдать определенным уровнем дисперсии, введенной различными лицами, работающими на этапе ручной обработки. Проблема данной вариации может решаться с помощью методов управления качеством, но в конечном итоге она в определенной степени будет всегда присутствовать в данных.

22. Систематические погрешности или отклонения также могут быть введены в рамках систем и процедур ручной обработки. Эти систематические погрешности являются главной проблемой для наших пользователей. Одним из преимуществ ручной обработки является то, что персонал, участвующий в этом процессе, способен потенциально выявить эти систематические погрешности в ходе своей работы.

Автоматизированная обработка

23. Обработка результатов переписи населения 2001 года в Соединенном Королевстве будет в значительной степени автоматизирована по сравнению с переписью 1991 года. Внедрение технологии OPM позволит осуществлять около 90% "простого" кодирования (например, данные по признаку "страна рождения") в автоматическом режиме. Около 60% "сложного" кодирования (например, данных по признаку "занятие") также будет осуществляться в автоматическом режиме с использованием технологий OPC и автоматического кодирования.

24. Повышение уровня автоматизации процессов обработки сопряжено с очевидными и существенными выгодами с точки зрения затрат. Однако оно имеет также определенные последствия для управления качеством данных.

25. Автоматизация ведет к значительному повышению непротиворечивости обработки. Один и тот же ответ при автоматической обработке будет всегда кодироваться одинаково. Хотя данная непротиворечивость носит позитивный характер в тех случаях, когда автоматическая процедура обеспечивает правильное кодирование ответов, она также может являться источником систематических погрешностей, если процедура автоматической обработки неправильно кодирует ответ. Эти систематические погрешности будут являться весьма очевидными и более значимыми для пользователей результатов переписей, чем дисперсия, вводимая в ходе ручной обработки.

26. Для сведения до минимума риска возникновения систематических погрешностей в данных, представляемых нашим пользователям, мы будет использовать комбинацию методов управления качеством, контроля качества и проверки достоверности результатов.

III. Контроль качества

27. Одним из основных требований к данным переписи является обеспечение точности данных на самом низком географическом уровне. В Соединенном Королевстве итоги переписи на уровне местных административных единиц используются центральным

правительством в качестве основы для распределения ресурсов и определения границ выборных участков. Многие другие пользователи также используют данные переписи на уровне мелких географических единиц.

28. Контроль качества имеет важное значение для разработки материалов переписи, поскольку они носят дифференцированный характер, т.е. каждый вид материалов является отличным. Если бы дело обстояло иным образом, мы могли бы полностью положиться на процедуры управления качеством и просто заменять любые производимые нами материалы, страдающие дефектами. Обработка результатов переписи во многом напоминает производственный процесс, однако в фундаментальном смысле она представляет собой совершенно иную операцию.

29. Результаты переписи, используемые для подготовки каждого отдельного вида материалов, должны обрабатываться обособленно с целью достижения, по меньшей мере, минимально приемлемого стандарта качества. Для обеспечения соблюдения этих стандартов мы будем использовать процедуры контроля качества.

Сканирование и оптическое распознавание

30. Одной из основных областей потенциального риска в рамках любой переписи является ввод данных, целью которого является преобразование сведений, содержащихся в бумажных формулярах, в первый набор результатов переписи в электронном формате. Для получения качественных изображений опросных листов с помощью сканирования и обеспечения эффективного ввода ответов с этих изображений с помощью оптического распознавания мы будем широко использовать процедуры контроля качества.

31. Нашим пользователям по сути безразлично, является ли причиной отсутствия данных непредоставление их домохозяйством или отдельным лицом, или же ошибки в ходе обработки результатов переписи. Оба случая имеют один и тот конечный результат для пользователя, а именно отсутствие данных о домохозяйствах или частных лицах в материалах переписи.

32. На этапе ввода данных с опросных листов, поступивших в центр обработки, существуют различные источники возникновения ошибок. К ним, в частности, относятся ошибки в процессе сканирования (например, пропуск некоторых формуляров) и ошибки в процессе оптического распознавания (например, ошибки в распознавании идентификаторов ПУ в опросных листах).

33. Особое значение приобретает обеспечение целостности данных, введенных с опросных листов. Для этой цели мы будем использовать процедуры контроля качества.

Данные по каждому переписному участку (ПУ)

34. Проверка наличия опросных листов по каждому переписному участку является важной задачей управления процессом регистрации.

35. Нам также необходимо обеспечить поступление данных по каждому переписному участку в соответствии со схемой переписного районирования, которая определяет переписные участки, по которым будут разрабатываться материалы переписи для распространения.

Данные по каждому домохозяйству

36. Счетчикам будут розданы книги регистрации счетчика (КРС), которые они будут использовать для записи информации о домохозяйствах, которым они выдают опросные листы. Эта информация будет содержать регистрационный номер (РН), который присваивается каждому домохозяйству с целью его идентификации в рамках переписного участка.

37. Некоторые опросные листы будут непосредственно отсылаться респондентами обратно по почте. Не отосланные по почте опросные листы будут собираться счетчиками и передаваться ими в центр обработки.

38. Необходимо будет провести проверку соответствия между регистрационными номерами, по которым должны быть получены опросные листы, и регистрационными номерами, содержащимися в данных, представленных по соответствующему переписному участку. Данный контроль соответствия будет осуществляться после завершения процессов сканирования и оптического распознавания. Регистрационные номера по розданным опросным листам (из книги регистрации счетчика) будут использоваться в качестве контрольной информации.

39. Будут разработаны процедуры для решения проблем, связанных с отсутствием в данных ожидаемых по переписному участку регистрационных номеров, присутствии неожиданных регистрационных номеров в данных и использованием одного и того же регистрационного номера в отношении нескольких опросных листов. Данные процедуры позволят производить корректировку контрольной информации, а также добавление, исключение или модификацию введенных данных.

Данные по каждому лицу

40. По всей видимости, значительная доля домохозяйств не будет вступать в контакт со счетчиками ни на одном из этапов. Типичными случаями могут являться отсутствие респондентов по своему домашнему адресу во время раздачи опросных листов счетчиком, а также оперативная отсылка домохозяйством опросных листов по почте. В отношении таких домохозяйств счетчик не будет располагать информацией о том, какое число лиц будет проходить под одним регистрационным номером.

41. Это означает, что счетчик не сможет представить контрольную информацию об ожидаемых регистрационных номерах. С учетом этого для обеспечения того, чтобы каждое лицо, представившее ответы на опросный лист по домохозяйствам, было включено в данные, необходимо было использовать другую форму контроля качества.

42. В настоящее время мы изучаем возможности использования других видов информации, содержащихся в опросном листе, для проверки числа лиц, указанных в данных о домохозяйстве.

43. Опросный лист переписи 2001 года содержит таблицу, в которой членам домохозяйства предлагается перечислить имена всех членов, обычно проживающих в данном домохозяйстве. Число имен, указанных в данной таблице, рассматривается нами в качестве одной из возможных контрольных цифр.

44. Будут разработаны процедуры для решения проблем, возникающих в том случае, когда число лиц, указанных в данных, отличается от контрольной цифры. Эти процедуры позволят включать или исключать индивидуальные записи по домохозяйствам.

45. Если мы примем решение об использовании процедуры ручной обработки, то эта процедура может быть включена в вышеописанный процесс проверки согласованности. Потенциальные преимущества ручной обработки, в частности, предусматривают использование имен для интуитивного определения пола лица, данные о котором отсутствуют.

46. Если же мы остановим свой выбор на автоматической процедуре, опирающейся на определенные правила, то лица будут автоматически включаться в состав домохозяйств или исключаться из него.

Полнота изображений и данных по каждому опросному листу

47. Процесс сканирования требует обеспечения создания изображений всех необходимых страниц каждого опросного листа.

48. Будут разработаны процедуры, обеспечивающие повторное сканирование опросных листов, по страницам которых не было создано изображений.

49. После завершения процесса оптического распознавания мы проведем проверку полноты данных каждой созданной записи. Будут разработаны процедуры, позволяющие распознавать опросные листы, по которым были созданы все требуемые изображения, однако не все данные с них были введены в ходе процесса оптического распознавания.

Правильность записей по каждому опросному листу

50. Вышеприведенные соображения относятся к записям по домохозяйствам и частным лицам. Они носят несколько упрощенный характер. На практике нам придется работать с несколько иного рода видами записей, в частности с записями по "институциональным заведениям" (к которым относятся гостиницы и больницы).

51. Существуют также другие виды записей, которые содержат информацию о переписном участке, а также информацию о семьях, входящих в состав домохозяйств.

52. Существуют правила, которые регулируют, какие виды записей регистрируются по домохозяйству или институциональному заведению. Подробное содержание этих правил не представляет большого интереса для темы настоящего документа. Главной проблемой для нас является обеспечение правильности записей по каждому регистрационному номеру после завершения процедуры ввода данных.

53. С этой целью после завершения ввода данных информация будет подвергаться проверке для выявления случаев, которые не удовлетворяют этим правилам. Мы также разработаем процедуры, позволяющие включение, исключение и модификацию записей во введенных данных для обеспечения удовлетворения этих правил.

54. Соответствующие процедуры корректировки регистрационных номеров, не прошедших данную проверку, могут быть включены в процесс проверки согласованности.

Этапы обработки после ввода данных

55. Целостность структуры данных после завершения этапов сканирования и оптического распознавания будет обеспечиваться с помощью вышеописанных процедур проверок и корректировок.

56. Кроме того, необходимо предотвратить нарушение целостности данных на последующих этапах обработки.

57. По завершении этих проверок мы будем сохранять информацию о структуре данных по переписным участкам. Данная информация будет затем использоваться в качестве контрольных данных, с помощью которых мы будем проверять структуру данных после завершения последующих этапов обработки.

Редактирование

58. Процесс редактирования может сыграть важную роль в контроле качества данных переписи за счет использования правил, обеспечивающих внутреннюю непротиворечивость информации. При определении концепции внутренней непротиворечивости мы учли требования наших пользователей.

59. При разработке правил редактирования мы устанавливаем требования соблюдения одних условий и недопустимости других. Одни правила направлены на проверку непротиворечивости элементов данных в рамках одной записи. Другие правила обеспечивают проверку непротиворечивости элементов данных в рамках определенного числа связанных между собой записей.

60. Примером редактирования данных в рамках одной записи является проверка того, чтобы в отношении лиц в возрасте до 16 лет не имелось данных по признакам рабочей силы, таким, как занятие. Применение данного правила редактирования обусловлено тем, что наше определение экономически активных лиц устанавливает порог в виде возраста старше 15 лет. Мы обеспечиваем соблюдение этого условия в ходе редактирования.

61. Примером редактирования элементов данных, относящихся к нескольким записям, может являться проверка непротиворечивости данных о возрасте, статуса в домохозяйстве, брачном состоянии и поле лиц в рамках одного домохозяйства. Мы будем применять правила, обеспечивающие соблюдение минимального 13-летнего разрыва в возрасте детей и их родителей. Хотя на практике существуют случаи, когда родителями являлись подростки в возрасте младше 13 лет, мы приняли решение применять данное правило для обеспечения непротиворечивости с данными предыдущих переписей и других обследований.

62. Процедуры редактирования также направлены на решение проблем, связанных с предоставлением нескольких ответов на один вопрос переписи. Мы изучим общие схемы предоставления нескольких ответов с целью их учета в разработке всех редакционных проверок. Наша цель заключается в снижении до минимума влияния допущенных респондентами ошибок, связанных с предоставлением нескольких ответов на "клеточные" вопросы или на комбинированные "клеточные/письменные" вопросы.

63. Мы будем использовать комбинацию "жестких" и "мягких" методов редактирования. Под жесткой процедурой редактирования понимается внесение изменений в данные с целью соблюдения правил. Под мягким редактированием понимается регистрация числа случаев несоответствия данных тому или иному правилу редактирования без внесения изменений.

64. Данные о случаях несоответствия правилам редактирования являются ценным источником информации для управления качеством. Неожиданно высокое или низкое число отклонений от того или иного конкретного правила редактирования может свидетельствовать о наличии проблем, связанных с ответами или с предыдущими этапами обработки.

Проверка соответствия интервалам величин

65. В отношении распространяемых результатов переписи установлен четко определенный набор величин по каждому элементу данных материалов переписи. Этот набор величин, как правило (но не всегда), соответствует классификации результатов по каждому элементу

данных. Целью проверки соответствия интервалу величин является выявление любых случаев, когда величина элемента данных не соответствует набору допустимых величин соответствующего элемента. Эти случаи квалифицируются в качестве ошибок, которые должны корректироваться до передачи данных в отдел по подготовке материалов переписи.

66. Так, например, набором допустимых величин по признаку "пол" являются мужской и женский или более конкретно цифры, используемые в наборе данных для обозначения мужского и женского полов. Допустим, что мужчины регистрируются в данных с помощью "1", а женщины - "2". В случае переменной "пол" целью нашей проверки соответствия интервалу значений будет являться выявление любых случаев, когда то или иное лицо описывается с помощью величины, иной чем "1" или "2".

67. Проверка соответствия интервалам значений будет осуществляться после завершения последнего этапа обработки, в ходе которого в данные могут вноситься изменения, до передачи результатов в отдел по подготовке материалов переписи.

68. По завершении более ранних этапов обработки к данным также может применяться модифицированная процедура проверки соответствия интервалам значений. Для проверки элементов данных, некоторые величины которых могут рассматриваться в качестве допустимых на каком-то этапе обработки, но не отделом по подготовке материалов, необходимо будет использовать несколько наборов допустимых величин. Преимущество применения проверки соответствия интервалам значений на более ранних этапах обработки данных заключается в возможности заблаговременной идентификации введенных погрешностей и, следовательно, наличии большего объема времени для оценки и корректировки этих ошибок. Поскольку повторная обработка больших массивов данных сопряжена со значительными расходами, раннее обнаружение ошибок может в конечном итоге дать значительную экономию средств.

Передача данных в отдел по подготовке материалов переписи

69. Еще одним важным фактором риска, который мы должны учитывать в процессе контроля качества, является этап передачи данных в отдел по подготовке материалов переписи. Данный этап включает в себя подготовку центром обработки файла или серии файлов, которые отдел по подготовке материалов загружает затем в систему табулирования для разработки материалов переписи.

70. Как представляется, между способом организации данных для целей обработки и способом их организации в пакете табулирования для эффективной разработки таблиц существуют значительные различия. Процесс преобразования данных в формат, требуемый пакетом табулирования, может потенциально вводить значительные систематические ошибки в информацию.

71. Мы разработаем процедуру контроля качества, в рамках которой будет осуществляться сопоставление файлов, подготовленных для разработки материалов переписи, с базой

данных центра обработки, на основе которой были подготовлены первые файлы. Эта процедура позволит выявлять любые несоответствия между двумя наборами данных.

72. Отдел по подготовке материалов переписи будет отвечать за обеспечение целостности информации в ходе загрузки файлов из центра обработки в пакет табулирования.

IV. Управление качеством

73. Стратегия контроля качества будет дополняться процедурами управления качеством для решения других проблем в этой области. В настоящем разделе описываются некоторые ключевые проблемы качества, которые мы будем решать с помощью процедур управления качеством, а также излагается наш подход к управлению качеством процессов ручной обработки.

Связь между изображениями и прикладными средствами обработки

74. Связь между изображениями и нашими прикладными средствами имеет чрезвычайно важное значение для обеспечения целостности данных, поскольку после завершения процесса сканирования изображения используются в качестве заменителей физических опросных листов. Результатом процесса сканирования является изображение каждой страницы опросного листа. В некоторых случаях речь может идти об изображениях конкретных разделов опросного листа, а не об изображении всей страницы, однако нижеописываемые принципы применимы ко всем видам изображений.

75. Существует риск того, что ввод или актуализация данных переписи будет производится на основе ошибочных изображений. Если это будет происходить систематически в рамках той или иной процедуры обработки, результаты переписи будут страдать фундаментальными искажениями.

76. Связь между изображениями и данными не требует постоянной проверки в ходе обработки. Она будет тщательно протестирована в ходе проектирования наших средств обработки изображений, предназначенных как для автоматизированных, так и ручных процессов.

77. Для обеспечения считывания нашими средствами обработки требуемых изображений мы проведем сканирование проверочных опросных листов и проверку того, что изображения, используемые средствами обработки, соответствуют надлежащему ответу на соответствующий проверочный опросный лист.

Связь между средствами обработки и данными

78. Важное значение также придается обеспечению надлежащей актуализации средствами обработки результатов переписи. Как и в случае связи между изображениями и нашими средствами обработки, введение систематических ошибок тем или иным средством обработки в ходе актуализации данных создает фундаментальные искажения в информации.

79. Эта связь будет тщательно протестирована в ходе проектирования прикладных средств с целью обеспечения актуализации надлежащих элементов данных в результатах переписи.

Оптическое распознавание меток (ОРМ)

80. Оптическое распознавание меток (ОРМ) в "помеченных" ответах и комбинированных "помеченных"/письменных ответах будет использоваться для кодирования большинства результатов переписи 2001 года. Распознавание "несуществующих" меток или нераспознавание проставленных меток может служить источником существенных систематических погрешностей в данных.

81. Хотя мы не можем полностью предотвратить возникновения ошибок в процессе распознавания меток, мы можем использовать процедуры управления качеством для снижения риска ввода систематических погрешностей.

82. Сканнеры и компьютеры, которые мы будем использовать для создания изображений и распознавания меток, как и любое электрическое оборудование, подвержены сбоям.

83. Для оптимизации качества создаваемых изображений будет вестись регулярное профилактическое обслуживание сканнеров. Качество каждого изображения будет подвергаться процедуре автоматического контроля для обеспечения соответствия изображений требованиям оптического распознавания.

84. Мы также будем проводить на регулярной основе сканирование и оптическое распознавание проверочных опросных листов с целью выявления проблем в области сканирования и оптического распознавания меток.

Оптическое распознавание текста (ОРС)

85. Результатом оптического распознавания символов (ОРС) является текстовая последовательность. Текстовые последовательности являются исходным материалом для процесса автоматического кодирования. Для обеспечения полноты текстовых последовательностей, предназначенных для автоматического кодирования, может использоваться своего рода процедура "исправления" символов, которые не были распознаны ОРС. Этот дополнительный процесс не сказывается на значимости нижеописываемых процедур.

86. Основным фактором риска, связанным с процессом ОРС, является ввод систематических погрешностей "замещения". Речь идет о систематически неправильном распознавании символа в качестве другого. Определенные ошибки замещения неизбежно будут возникать в ходе оптического распознавания письменных ответов. Как и в случае оптического распознавания меток, это может служить источником погрешностей, который мы попытаемся контролировать.

87. Программное обеспечение ОРС позволяет нам установить уровни доверия, которым должен удовлетворять процесс распознавания для того, чтобы символ рассматривался в качестве правильно распознанного. Повышение этих уровней доверия позволит снизить число ошибок замещения в распознанных текстовых последовательностях. Существует очевидный компромисс между эффективностью (распознавание максимального числа символов в ходе ОРС) и качеством (риск ввода ошибок в результате неправильного распознавания символов).

88. Ошибки в распознавании текстовых ответов могут идентифицироваться различными способами. Один из них заключается в проверке результатов процесса ОРС путем сопоставления изображения ответа и текстовой последовательности, полученной в результате оптического распознавания данного ответа. Результаты этих проверок позволят нам определить, какие буквы или цифры систематически распознаются неправильно программой ОРС.

89. Еще одним способом выявления ошибок замещения является включение данной процедуры в процесс управления качеством автоматического кодирования. Мы будем производить проверку результатов автоматического кодирования путем повторной обработки выборки автоматически кодированных ответов с помощью процедуры полуавтоматической обработки с последующим сопоставлением результатов. В этот процесс мы можем включить процедуру проверки распознанных ОРС текстовых последовательностей. В случае несоответствия между кодом, присвоенным в результате автоматического кодирования, и кодом, присвоенным в результате полуавтоматической обработки, мы можем сравнить изображение и распознанную текстовую последовательность ответа. После этого мы можем дать указание нашему персоналу произвести повторную обработку с целью идентификации любых ошибок в текстовой последовательности.

90. Данный подход позволит нам дифференцировать ошибки в результатах автоматического кодирования, вызванные дефектами ошибок в распознавании, от других случаев несоответствия между автоматическим кодированием и полуавтоматической обработкой. Благодаря этому в процессе могут быть выявлены значительные ошибки замещения.

Автоматическое кодирование

91. Факторы риска, связанные с автоматическим кодированием, заключаются в возможности систематически неправильного кодирования конкретных ответов.

92. Полуавтоматическая обработка идеально подходит для проверки качества результатов автоматического кодирования, поскольку ее задачей является кодирование ответов на одни и те же вопросы переписи. С этой целью будет сформирована выборка ответов, обработанных с помощью автоматического кодирования, для повторного кодирования методом полуавтоматической обработки. Затем будет проведено сопоставление кодов, полученных в результате двух процессов по одному и тому же ответу, с целью определения

того, какой код является правильным. Анализ всех случаев, когда автоматически присвоенный код является неправильным, позволит нам определить систематические погрешности.

93. Еще один изучаемый нами в настоящее время подход заключается в целенаправленной проверке ответов. В нашей базе данных будут храниться все ответы. Таким образом, мы можем идентифицировать все индивидуальные ответы и ранжировать их по частоте. Поскольку процедура автоматического кодирования всегда присваивает один и тот же код индивидуальному ответу, нам потребуется только проверить код, присвоенный каждому индивидуальному ответу. Составление перечня наиболее часто встречающихся ответов позволит нам эффективно управлять качеством автоматического кодирования.

Управление качеством процессов ручной обработки

94. Полуавтоматическое и экспертное кодирование являются основными процессами ручной обработки. Мы будем использовать стандартный подход, заключающийся в повторной обработке выборки результатов для оценки их качества. Методика использования этой информации будет опираться на общую философию управления качеством, которая весьма хорошо документирована по вопросам управления процессами ручной обработки.

95. Краеугольным камнем этой философии является то, что причиной большинства ошибок в результатах являются недостатки самого процесса. Поскольку мы сами проводим обучение наших сотрудников и разрабатываем системы и процедуры, используемые в этом процессе, то за качество результатов также отвечаем мы, а не наш персонал. Совершенствование процесса является ключевым условием повышения качества результатов.

96. Наши сотрудники являются наиболее ценным ресурсом, поскольку именно они могут дать нам наиболее полезные советы по совершенствованию процесса, в котором они участвуют.

97. Мы будем постоянно заниматься повышением качества на основе следующего цикла:

i) измерение качества
ii) выявление значительных ошибок
iii) выявление основных причин этих ошибок
iv) внесение корректировок и повторное измерение качества.

98. Для выявления и решения проблем качества мы будем использовать группы, в состав которых войдут специалисты в области обработки данных, эксперты по вопросам кодирования и администраторы.

99. Наш подход будет сосредоточен на выявлении и анализе основных причин возникновения значительных систематических погрешностей в процессах ручной обработки.

Контроль процессов

100. Процедуры контроля процессов будут внедряться для обеспечения прохождения данных через различные процессы в правильной логической последовательности.

101. В рамках каждого процесса также будет разрабатываться информация об изменениях, внесенных в данные в ходе этого процесса. Данная информация послужит эффективной проверочной основой, позволяющей нам производить мониторинг изменений, внесенных в данные в рамках каждого процесса, а также выявлять, на каких этапах возникают проблемы с качеством данных.

V. Проверка достоверности результатов

102. Когда мы обратились к нашим пользователям с просьбой дать оценку результатам нашей деятельности в 1991 году, то в качестве двух из наиболее важных причин их недовольства были указаны значительные ошибки в результатах и несоблюдение нами намеченного графика публикации данных.

103. В 1991 году мы не проводили проверку достоверности данных перед их передачей в отдел по подготовке материалов переписи. Выявление ошибок и их последующая корректировка привели к задержкам с публикацией материалов. В данных также содержался ряд значительных ошибок, которые были выявлены только после публикации материалов.

104. В настоящее время мы намерены решить эти две проблемы путем мониторинга данных в процессе обработки и осуществления проверки достоверности результатов. Специально для этой цели будет создана группа по проверке достоверности результатов. Она будет заниматься осуществлением стандартных проверок, а также выявлением и анализом потенциальных проблем.

105. Мы будем осуществлять мониторинг данных с самого начала процесса обработки, с тем чтобы выявить значительные систематические погрешности на как можно более ранней стадии. Невыявление ошибок уже на ранних этапах обработки позволит нам посвятить больше времени нахождению источников этих ошибок и их устранению. Выявление и устранение источников систематических погрешностей на ранних этапах может также содействовать значительной экономии ресурсов за счет снижения масштабов повторной обработки и связанных с ней трудозатрат.

106. В некоторых случаях исправить ошибки будет невозможно (например, если источником ошибок являются респонденты), однако это позволит нам заблаговременно уведомить о них пользователей.

Стандартные проверки

107. Группа по проверке достоверности данных будет осуществлять агрегирование и табулирование данных до проведения проверок. Эти проверки будут осуществляться на различных географических уровнях и на различных этапах обработки.

Проверка 1 - Итоги

108. Проверка итогов направлена на обеспечение согласованности итоговой суммы таблицы с конкретными элементами данных. Так, например, если элемент данных применим к частным лицам, мы будем проверять соответствие итога таблицы общей численности населения в заданной географической зоне.

Проверка 2 - Неприменимость

109. Целью проверки неприменимости является обеспечение того, чтобы все записи по конкретному элементу данных являлись применимыми. Так, например, если признак "занятие" применим только к экономически активным лицам, то мы должны проверить наличие данных о занятии по всем экономически активным лицам и отсутствие таковых по лицам, являющимся экономически неактивными.

Проверка 3 - Изменения в период между проведением переписи

110. Целью проверки изменений в период между проведением переписей является сопоставление распределений в таблице переписи 2001 года с распределением данных переписи 1991 года по одной и той же географической зоне. В тех случаях, когда результаты переписи 1991 года являются несопоставимыми или искомые данные не собирались в 1991 году, мы будем использовать, по мере возможности, альтернативные контрольные данные. Эти альтернативные контрольные данные будут включать в себя оценки численности населения и определенные виды административных данных.

111. Мы будем осуществлять сопоставление распределений (процентов от применимого итога), а не числовых итогов.

112. Мы установим допустимые уровни для разницы в распределениях 1991 и 2001 годов. В тех случаях, когда распределение 2001 года будет превышать допустимые уровни по той или иной заданной категории, Группа по проверке достоверности данных проведет дополнительные исследования.

113. Так, например, при проверке достоверности данных по переменной "пол" мы будем изучать распределение мужчин и женщин в общей совокупности населения. Так, например, если распределение в 1991 году по какой-то заданной географической зоне составило 49% мужчин и 51% женщин, то допустимый уровень изменения в период между переписями будет установлен в размере 1%. Если распределение в 2001 году составит

49,5% мужчин/50,5% женщин, данные будут признаны достоверными, однако в случае процентного соотношения 50,5% мужчин/49,5% женщин данные будут признаны недостоверными.

114. При сопоставлении данных за десятилетний период невозможно установить единого интервала изменений в период между переписями, который бы подходил для всех элементов данных и всех географических районов. В качестве общего правила мы будем заниматься отслеживанием любых изменений в период между переписями, превышающих 5% на уровне округов (в Великобритании в 1991 году насчитывалось примерно 70 округов). На более низких географических уровнях данная вариация, по всей видимости, может быть намного более значительной, в связи с чем допуск будет установлен в размере 10%. При определении уровней изменений в период между переписями по различным переменным мы в значительной степени будем опираться на результаты генеральной репетиции переписи, которая будет проведена в 1999 году.

115. Проверка изменений в период между проведением переписей будет также осуществляться по ограниченному сбору простых комбинационных таблиц, в особенности тех, которые включаются в стандартные материалы переписи и описывают конкретные подгруппы населения.

116. В настоящее время мы занимаемся изучением потенциальных возможностей автоматизации процедуры проверки изменений в период между проведением переписей. Это позволит сэкономить ресурсы, требуемые для табулирования данных и сопоставления распределений, а также выделить больше времени для изучения потенциальных проблем качества данных.

Последующий анализ

117. В случае отбраковки того или иного элемента данных в ходе стандартных проверок Группа по проверке достоверности данных будет проводить дополнительные исследования.

118. Отбраковка данных в ходе проверок итогов или неприменимости свидетельствует о наличии проблемы. Мы будем использовать информацию об изменениях, внесенных в данные на каждом этапе, для выяснения причин, по которым кодирование элемента данных не производилось по соответствующей совокупности населения или же некоторые группы населения включались (исключались) ошибочно из охвата признака по причине неприменимости. После этого мы будем осуществлять корректировку ошибок в отработанных данных и заниматься устранением источника ошибок во избежание его дальнейшего влияния на обработанные данные.

119. Более сложную проблему представляют собой случаи отбраковки данных в ходе проверки изменений в период между проведением переписей. В данном случае речь может идти о проблеме качества данных. Мы будем производить проверку того, чтобы все данные точно отражали ответы, представленные в опросном листе. Это будет

осуществляться путем формирования выборки записей и отслеживания процесса обработки элементов данных начиная с этапа оптического распознавания до текущей стадии обработки с использованием контрольной информации об изменениях в данных. В тех случаях, когда выявленный неожиданный сдвиг в распределении значений по какому-то элементу данных представляет собой ошибку, внесенную на одном или нескольких наших этапов обработки, мы будем исследовать варианты корректировки самих данных и устранения источника ошибок в процессах. Если же данные точно отражают ответы, представленные в опросном листе, то мы проанализируем изменение и подготовим пояснения для наших пользователей в отношении этой характеристики данных.

120. Так, например, опросный лист переписи 2001 года содержит вопрос о хронических заболеваниях, который служит для получения информации о том, не испытывают ли респонденты какие-либо ограничения в своей повседневной жизни и работе в связи с хроническим заболеванием, проблемами со здоровьем или инвалидностью. Респондентам предлагается пометить клетки "Да" или "Нет". Представим, что при сопоставлении распределения ответов "Да"/"Нет", представленных в ходе переписи 2001 года, с аналогичным показателем переписи 1991 года, выясняется, что изменение распределения между ответами "Да" и "Нет" превышает допустимый уровень, установленный в отношении признака "хроническое заболевание".

121. Для выяснения причины такого сдвига мы сформируем выборку записей, кодированных по ответу "Нет". В отношении этих записей мы произведем проверку изображений ответов на вопрос "хроническое заболевание" с тем, чтобы удостовериться в том, что коды данных, присвоенные после оптического распознавания, соответствуют помеченным ответам в опросных листах. Если помеченные ответы были распознаны неправильно, мы внесем корректировки в процесс распознавания и повторим его в отношении всех соответствующих ответов.

122. Если мы установим, что пометки были распознаны правильно, следующим шагом будет являться проверка других процессов, в ходе которых в данные вносились изменения.

123. В случае вопроса "хроническое заболевание" другими процессами, в ходе которых в этот элемент данных могут вноситься изменения, являются редактирование и условные расчеты. Редактирование призвано решать проблемы, связанные с пометкой респондентом более одной клетки ответов. Целью условных расчетов является присвоение величин данных тем респондентам, которые не представили ответы на вопрос "хроническое заболевание". Мы будем производить проверку того, имело ли место существенное изменение в распределении ответов "Да" и "Нет" в ходе редактирования или условных расчетов. Для этого мы составим таблицу данных после завершения этапов оптического распознавания и редактирования. Затем мы сопоставим распределение по признаку "хроническое заболевание" с распределением после проверки изменений в период между переписями. Если будет выявлено значительное расхождение между распределением после оптического распознавания и распределением после редактирования или условных расчетов,

мы изучим варианты повторной обработки или корректировки данных, которые уже прошли через эти процессы, и попытаемся устранить источник проблем, возникших в процессе редактирования или условных расчетов.

124. Если редактирование или условные расчеты не ведут к изменению распределения и ввод ответов на вопрос "хроническое заболевание" был произведен без ошибок с опросных листов, то это означает, что наша система обработки не нуждается в корректировке. Мы могли бы назвать этот результат "чистым карантинным свидетельством" для обработки.

125. Ответ на вопрос о хроническом заболевании в значительной степени зависит от личной оценки состояния своего здоровья респондентом. Лучшая осведомленность общественности о вопросах состояния здоровья и соответствующая политика, направленная на расширение доступа к рабочим местам лиц, страдающих хроническими заболеваниями, могут оказать влияние на оценку респондентами того, в какой мере хроническое заболевание оказывает ограничительное воздействие на их повседневную жизнь или доступ к занятости. Данное изменение в восприятии может носить частично "реальный" характер (облегчение повседневной деятельности и увеличения числа рабочих мест, доступных людям, страдающим хроническими заболеваниями) и частично субъективный характер (респондент, наблюдая, как другие люди занимаются той или иной деятельностью или работой, рассматривает свое заболевание или инвалидность, в качестве менее ограничительного фактора). Наши пользователи должны быть проинформированы о необходимости учета этих факторов при принятии решений на основе данных о хронических заболеваниях.

Система мониторинга качества данных

126. В настоящее время ведется разработка системы мониторинга качества данных (СМКД), которая призвана дать группе по проверке достоверности результатов возможность производить просмотр и анализ результатов переписи и информации об изменениях, вносимых в данные в ходе обработки. Мы стремимся "по мере возможности" создать среду "реального времени".

127. Во избежание снижения производительности системы обработки для СМКД создается отдельная база данных, в которую будут регулярно загружаться результаты переписи, информация об изменениях в данных и другая информация из систем обработки.

128. Члены группы по проверке достоверности данных пройдут обучение по методам анализа данных и будут использовать эти методы с помощью программного обеспечения анализа и составления таблиц для СМКД. В настоящее время мы также занимаемся оценкой потенциальных выгод географической визуализации данных.

Пользователи-эксперты

129. Наша стратегия проверки достоверности данных предусматривает использование услуг пользователей-экспертов для оказания нам содействия в проверке данных. Данные о месте работы (которые мы используем для географического кодирования мест работы респондентов) являются примером информации, для проверки которой могут привлекаться пользователи-эксперты.

130. Данные о местах работы весьма подвержены ошибкам, основной причиной которых является неполнота и неправильность получаемых на этот вопрос ответов. К сожалению, эти ошибки носят систематический и локализуемый характер, что объясняется большим числом случаев ошибочного кодирования респондентов по неправильной "зоне" мест работы. Ошибки могут носить весьма очевидный характер в тех случаях, когда они обладают легко поддающимися идентификации характеристиками: например, определенная больница физически расположена в одной "зоне", однако весь больничный персонал кодируется в качестве работающего в другой "зоне".

131. Задействуя услуги пользователей-экспертов для оказания нам содействия в выявлении систематических ошибок в данных о рабочих местах, мы одновременно можем воспользоваться их знаниями местных условий или же конкретной экспертизой (например, в области моделирования поездок). С учетом их рекомендаций мы будем осуществлять корректировку данных в целях снижения или устранения воздействия этих систематических ошибок.

Генеральная репетиция переписи

132. Генеральная репетиция переписи будет проведена в 1999 году с целью проверки наших процедур и систем регистрации и обработки данных. Она послужит прекрасной возможностью для полного апробирования нашей стратегии проверки достоверности данных, включая привлечение к проверкам пользователей-экспертов. Мы планируем провести оценку качества результатов, полученных в рамках генеральной репетиции, выявить возможные проблемы с качеством данных, а также осуществить доработку наших систем и процедур, направленных на устранение этих проблем.

VI. Выводы

133. Снижение до минимума уровня значительных систематических погрешностей в результатах является одной из основных целей переписи населения 2001 года Соединенного Королевства. Для достижения этой цели мы планируем использовать гармоничное сочетание процедур контроля качества, управления качеством и методов проверки достоверности данных до распространения результатов среди наших пользователей.
