2001 Census: Assessment of downstream processing, output production, and data quality management system

Contents

Introduction and summary

This document brings together a number of strands of 2001 Census evaluation, particularly in the areas of downstream processing, output production, and data quality management system. Recommendations relating to all areas covered are brought together at the end of the document.

Downstream processing

1. Once data has been captured from Census forms and coded, it must be prepared for an output database that is complete and consistent. Given this aim, changes to captured data should be minimal and remove any bias caused by non-response.

2. Data for Scotland was delivered from the data capture and coding contractor in eight lots called Estimation Areas (EAs). Scotland was divided into EAs for the process of estimating under-enumeration known as the One Number Census (ONC). Each EA was a grouping of council areas, except for some processes where boundaries were shifted to recognise health board areas. Estimates of the numbers of persons and households missed were done and assessed separately for each EA and areas within it. A final assessment of under-enumeration at the Scotland level was made before the estimates were finally signed off. (Further details will be published in a separate report on the ONC.) The data capture and coding contractor also supplied for each EA a file containing images of the Census forms they had scanned and a 'tick and text' file containing the content ('ticked' or 'not ticked') of each tick box and any text captured in write-in boxes on the form. Other items delivered by the contractor (but not considered directly in this report) were data and images from the Census Coverage Survey (CCS). The CCS was used to estimate under-enumeration in the main Census.

3. Various pieces of analysis, and data correction, had to be curtailed because the contractor delivered the various items above later than scheduled, halving the effective time for downstream processing.

4. Data for each EA was processed separately to the point where extracts were taken and combined into a single Scottish database for output production. Processing after data capture and coding up to output production was denoted 'downstream processing'. A diagram showing an outline of the processes within downstream processing is

available from the website of the Office of National Statistics (ONS).

5. The basic strategy was one of automation, with minimal if any clerical intervention. In the event clerical processes were required to prevent the clustering of certain errors within small areas or within small populations. There was far more clerical intervention than originally intended. 'Data file amendments' (DFAs) were applied to the database for each EA at the most appropriate point. Ideally the earlier in downstream processing an amendment was made the better chance it had of not creating inconsistencies with other data. In particular, if done before the data was edited, amendments would get the benefit of the edit process to sort out any errors in the supposed corrections. DFAs done after edit would not be checked other than undergoing a simple range check. Inconsistencies between data items may therefore be introduced in DFAs after the edit process (and to a small extent were). The DFA procedure itself did not lend itself to amendments being easily checked. Given that it is impossible to guarantee that automatic processing can do all of the data cleaning needed, it is recommended that there should be an easy-to-use clerical edit process that incorporates the checking of the edited records.

6. The main stages of downstream processing were:

- Preparation of captured data for loading into input databases
- Load
- Edit and imputation, first application
- One Number Census
- Edit and imputation, second application
- Disclosure control (i.e. record swapping); other measures to control disclosure are taken as part of output processing
- Sundry other processes

7. There was a 'process control' system to which GROS staff had read-only access. This enabled us to find out which stage had been reached by any of the Scottish EAs (or UK EAs for that matter). This proved to be a very useful tool and it is recommended that something on the same lines be set up for next Census no matter how it is processed.

Output production

8. Once the data had completed its journey through downstream processing (including the addition of records for persons and households by the ONC process), extracts from the several EA databases were

taken and loaded into bought-in SuperSTAR software for output production. Largely because of the different approaches taken within the UK to modifying tables after production for the control of disclosure, each Census Office had an output database for its own part of the UK.

Data quality

9. Investigation of data quality and identification of records for correction was by means of

- A Data Quality Management System (DQMS) incorporating a data interrogation tool called EasyAsk. The DQMS was originally conceived as one producing pre-planned comparisons of univariate and bi-variate distributions against non-Census sources, the 1991 Census, and 2001 data at previous points of processing. Further ad hoc investigations were carried out depending on what the pre-planned analyses showed. Increasingly the ad hoc element of the DQMS pre-dominated as unforeseen data issues arose. The data analysed in the DQMS using EasyAsk was taken from 'dumps' of the database for each EA as it passed through key stages of processing.

- An Image Viewer that allowed sight of the images of pre-selected pages of Census forms.

Role of ONS

10. All of the pre-planned downstream systems described (including the output extract programs) were developed by the Office of National Statistics (ONS) for all 3 UK Census Offices to a joint specification. ONS also provided a number of other services, chief of which was a team of 3 responsible for the 'coherence' between the various systems. ONS's review and evaluation of the activities covered in this report may be found on their website. GROS had to devise a few of its own systems to deal with unforeseen problems such as 'black lines', excess numbers of same-sex couples, etc.

Preparation of systems for live running

11. Finally, by way of introduction, it was not possible to test all of the processes of downstream processing and output production in the time between the Census Rehearsal (1999) and the full Census. It is recommended that a rehearsal is held in good time for all downstream processing and output production to be rehearsed before live production. Besides, data from the Rehearsal was not exactly in the same format as that from the full Census. Some testing for full processing was possible

with an early small consignment of 2001 data from the data capture contractor.

Successes and recommendations

12. Generally the systems worked well as specified. The GROS team got through a lot of work caused by the finding of unexpected errors in the data. But these errors (e.g. 'black lines' corrupting images of forms or entries of form-filler generating same-sex couples) were actually spotted and corrected. The time needed for the work was slotted into pauses while the data in error was being processed by other teams (chiefly ONC at the stage of matching of Census forms with CCS or the stage of examining the results of imputation of extra records).

13. Because most contingency time was used up there are inevitably several loose ends still ideally to be tied up. These appear below usually in the form of a recommendation that the outstanding work is completed. Whether or not any of these items does actually get completed will be, as usual, a matter of resources at a time when work has already started on the 2011 Census.

14. Output production in great part met its targets with most output released on time. The main exception was that the tables of the topic of migration and travel were left out of the main release as arrangements for the handling of data for these topics were not ready on time.

Data capture and coding and preparation for Load

Scope of capture and coding

15. The data capture contractor had the job of capturing the ticks and text on the forms and converting them to data in the form of codes for each variable. Write-in answers were converted into codes by a mix of automatic and clerical processing. Variables with such coding were: occupation, industry, country of birth, ethnicity, address 'one year ago' (i.e. one year before the Census), address of place of work or study, and addresses of enumeration not in pre-enumeration lists. Unlike for elsewhere in the UK, write-in answers were not coded to the two questions on religion (current, and of upbringing).

16. In order to contain costs within budget, it was decided not to code occupation and industry for all persons for whom answers had been written on the Census form. The questions were for all persons aged between 16 and 74 inclusive who were either in work the week before the Census or had ever worked. Coding was restricted to those in work and, for those not in work but who had been in employment at some

previous time or another, to those aged 16 to 64 who had worked in 1996 or later. The process agreed with the contractor that selected records for the coding of occupation and industry was called ‘Filter X’.

17. When a form was captured with a form number for a pre-listed address, the contractor ignored the address of enumeration on the form and instead took this information from the geography database it had been given. Only when a form had a number that did not correspond to a pre-listed address was the address information taken from the form itself. An extra record for the additional address was included in the data from the contractor. Address records were also created for addresses on the form that are *remote* from place of enumeration: address one year ago, and travel destination.

18. GROS splits postcodes whose addresses belong to more than one council area, adding an extra character (A, B, etc) to denote each part of the split. When coding *remote* addresses on a form for any part of the UK, the contractor may come across one whose postcode requires the extra character for a split. The address would be referred to GROS, ONS or NISRA. If referred to GROS, then we would decide from the address on the form to which part of the split postcode it belonged. If referred to another Census office, then they may have passed the query to GROS or they may decide arbitrarily to assign the address the A part of the postcode – usually the largest and hence most likely part. Other quirks of Scottish address coding for the contractor to deal with included the practice of calling secondary schools ‘academies’ (perhaps the use of this term will have spread to England by 2011). It is recommended that given inaccuracies with these variables the A split be used invariably – unless there are some key B splits containing large employing establishments.

19. The data from the forms were supplied in four types of record: for households, communal establishments, persons, and addresses. An address record was created for each household or communal establishment not at a pre-listed address in the enumerator’s record; also for each address of travel destination or address one year ago for which the contractor had to find a postcode.

20. Also, as stated in paragraph 2, the contractor supplied, in EA batches, electronic images of the forms that had been scanned as part of the capture process, and ‘tick and text’ files.

Edits done as part of capture and coding

21. Rules to resolve ‘multi-ticking’ were devised as part of the capture of responses to questions where one tick was required. Responses that were not resolved by these rules were given a null ‘multi-tick’ value to be dealt with in subsequent downstream processing. Similarly, responses

that were out of range were dealt with by the contractor if possible; otherwise they were given a null 'out of range' code.

Edits done as part of Load

22. The data captured and coded by the contractor for each EA was loaded by ONS into a database. The load process included a number of checks and processes including:

- **The '2 of 4' rule**. There was a risk of creating a record for a non-existent person when one or more tick boxes in a Person section of the Census form contained spurious marks e.g. because the form-filler had scored through unused pages. The capture process would create a near empty record and the '2 of 4' rule would reject the record unless at least 2 of the first 4 data items were complete. These items were Name, Sex, Date of Birth and Marital Status. At least one of Name and Date of Birth had to be complete – so there had to be at least one non-tick box response.

- **'Duplicates' rule**. It was discovered before full processing got under way that a person might enter his or her details in more than one Person section. The rule would compare records for persons over 70 within a household and where two or more matched on Date of Birth, Name (both surname and forename will be analysed using Soundex), and Sex. The Load process would create a single record taking items from the duplicated records.

- **Filter rules**. The answers to some of the questions for a household or person on the Census form determined what further questions were to be answered. Checks on each of these 'filter' questions and its dependent questions were included in the load process. Each record was thus given a structure with, for given values of filter questions, dependent questions set as NCR (for 'no code required').

- The data item 'activity last week' was derived from 5 questions on a person's economic activity in the week before the Census.

- In certain circumstances, a small number of households apparently containing no adults were deleted. They were often spuriously created because a household continuation form (used when a household had more than 5 residents) or an individual form apparently had no corresponding 'parent' household form. Misreading of a form identity can cause this 'separation' of a continuation or individual form from its 'parent'. It was considered that the loss of genuine person records deleted in this way would be made good in the adjustments of the ONC.

- Any household or CE record for a non-pre-listed address had to have

a companion address record, otherwise it would be rejected.

Re-ordering of processes and additional preparation

23. The original planned sequence of downstream processing was based on the simple idea that the data should be loaded into the database, subjected to a range of checks within the load process, and then move on to the edit and imputation stages (more below). However, during examination of prints of errors from the Load process, the loaded data records and corresponding images of forms, a number of major defects were noticed. It was decided that data for Scottish EAs would be loaded twice. The first time would simply be to use the loaded data and images to generate pre-load corrections to the 'raw' data that would then be loaded again in a 'substantive' load.

24. In the pre-load corrections, it was decided to concentrate on variables featuring in filter rules as they were crucial in establishing the structure of each record. By leaving corrections to other variables till after load – and correction via a DFA – delay to the substantive load would be minimised. The post-load DFAs would be applied while the loaded data was being matched against the Census Coverage Survey – one of the preliminary operations of the ONC.

25. The problems dealt with in pre-load corrections included:

- **ED-postcode mismatch** Despite the use of geography control files, the postcode of enumeration on a household record did not always belong to the Enumeration District (ED). Checks were done and records with mismatched postcode and ED were corrected – with, if the correction was to the ED, a possible transfer of the affected records to another EA. In all, these errors turned out not to be too frequent, numbering around 100.

- **Record for non-pre-listed Household and communal establishment (CE) but no address record** A few records (households and CEs) were rejected for this reason. The household or CE was checked and, if genuine, a missing address record was created and added to the raw data. Around 200 address records were added.

- **Dispersal of CE persons** Some individuals enumerated in communal establishments (CE) had mistakenly been given form numbers by the enumerator that, in effect, removed the individuals from the CE added them to surrounding households. The enumerator had invented an individual sequence number and written it in the space on the individual form for the household/CE form number. To sort this out entailed checking clusters of households apparently with additional persons enumerated with individual forms. 3466 person identifiers were changed for persons enumerated with individual

forms so as to re-import them into the correct CE. It is recommended that there is a space on the individual form for an individual sequence number so that enumerators aren't tempted to use the form number for that purpose. An individual number would also be very useful in data quality work (see paragraph 212).

- **Wrong format** The format of certain values of a few variables in the raw data was not as specified for a variety of reasons and so the load program rejected the records concerned. Examples were offshore travel destinations beginning with the letter D, and type of client of CE. The latter item had been captured according to the format for England and Wales. (There were a few cases of unavoidable errors caused by an England form being returned from a Scottish household.) Error reports from the Load program were checked for records rejected for format errors and 8115 records corrected.

- **Coding and other errors** Various systematic errors were noticed in (e.g.) occupation. For example, machinists in the textile trade had been coded as typists. Discrepancies were noticed between age and travel destination where it appeared that school age children were travelling to offshore destinations. This was often caused by an error in the person's date of birth. 4383 corrections were under this heading. This excludes errors for occupations requiring qualifications picked up by ONS checkers. Coding routines changed in time for the last of the 8 Scottish EAs to be captured. The other 7 had DFAs applied by ONS during downstream processing (see paragraph 52).

- **Black lines and white bands** The scanning process had added horizontal lines to the images of runs of forms – usually successive even-numbered pages. These lines could pass through tick boxes. They varied in vertical position on the form but the position was fixed for any given run of forms. The result was that sequences of records contained data that was corrupted in one way or another not necessarily put right by the 'multi-tick' rules in data capture. This quickly became known as the '*black lines*' problem. A similar defect was that there were sequences of forms where a horizontal band of a page (again usually successive even-numbered pages) was blanked out. This was called the '*white band*' problem and particularly affected the question on travel destination. Other scanning problems were fuzzy images where an empty tick box contained enough displaced black image to appear as though it had been ticked, and a form-filler's tick in one box straying near enough to another box (i.e. within a set tolerance) to cause the capture process to record the latter to have been ticked as well. Hunting black lines entailed checking for sequences of person records that failed the '2 of 4' rule. These sequences were caused by black lines appearing on unused Person sections of household forms. Runs of records affected by black lines would have the

same value of a given variable as perhaps the only content of the record. Images of pages for records that had not been rejected in the same run of forms were inspected to decide on corrections to captured data. Occasionally fuzzy boxes were detected this way. The process also pinpointed sequences of forms with 'white bands' because the bands often had black borders. Gaps in data records caused by white bands were made good by locating the forms with the contractor and keying corrections into a spreadsheet for combination with the final set of amendments. 12319 corrections were made to the affected records. (The ONS approach for black lines for many variables was that edit and imputation would take care of errors caused. In many cases a black line would cause that variable to be marked as 'multi-ticked' and hence susceptible to subsequent imputation. GROS believed that this was unacceptable because multi-tick rules wouldn't correct questions for which a multi-tick is permissible. Further, when no box at all had been ticked, the black line would force the capture of a single tick that would not be subject to imputation. No response in runs of households would become systematically a particular response.)

- **Missing forms Finally**, by comparing data from the contractor with enumeration records, sequences were identified of forms that had simply failed to get scanned. (This problem was related to another with the contractors: that of wrongly capturing the form number which had the effect of moving records about the country. See paragraph 3.4.10 in http://www.gro-scotland.gov.uk/files/report-datacapture-coding.pdf.) As the ONC would not deal with local clusters of missing forms, it was decided to fill these gaps by clerical means. In some cases the unscanned forms were located in the contractor's store, and others they were not. Where an ED had a sufficiently large shortfall, gaps of 5 or more within the ED were identified. 1644 missing forms were traced and a selection of variables (including occupation coded by GROS staff) keyed. This material was expanded to mimic full records and inserted into data files for loading. Where forms could not be found, 811 'skeleton' records were created from whatever information was available in the Enumeration Record. Missing values would be imputed later to give full data records. The name in each person record created in this way was entered as 'Made up person'. Care was required if any made up record was for a household in the CCS. In these cases the data made up was taken from the CCS record so as to ensure that the two records matched in the ONC matching process. The contractor also failed to supply records for an area of Edinburgh that had been used in a test of data capture. This shortcoming was made good using data from the test.

26. All of the above corrections were applied to the 'raw' data and the corrected version checked and re-loaded. In order to make these corrections, the 'raw' data was transferred from Titchfield and loaded into an Access database at Ladywell House. There was something of the

order of 4 staff years in this correction work. Given that many of the errors corrected in this exercise were localised, it was the right thing to have done this kind of work. Some of the corrections did not turn out to be as extensive as first feared so some scaling down was possible for later EAs without great damage to quality. Also some checks entailed looking at a large number of images for a relatively small number of corrections. These less productive checks were scaled down or dropped. Certainly the most labour intensive job – correcting errors caused by black lines – was essential in removing localised bias.

Other checks before final load

27. Communal establishments as enumerated were compared with a pre-Census register of CEs maintained by GROS. CEs from which forms were expected but not received were noted, but, no other action was possible. An analysis of the difference between CEs enumerated and as in the register is given in Appendix 1.

Load

Checking for abnormalities

28. Checks were carried out on the way filter rules had worked and, in particular, whether there were large numbers of cases being dealt with in an unexpected way. Generally, these checks showed that the rules were working satisfactorily although a few changes to the rules were proposed and made.

Data file amendments: black lines, etc

29. ‘Black line’ corrections not included in those made to the ‘raw data’ (paragraph 25, sixth bullet) were applied after the data had been loaded. Corrections for other faults (long ticks for qualifications causing more than one tick box to be registered, coding errors for industry and occupation, etc) were also included in the same batch of DFAs.

30. A few corrections were made following examination of error prints from the load process. For example, there were a few instances where a person record had a code for ethnicity that was not valid in Scotland (but was valid in England and Wales). The load program set the value to missing for later imputation, but the missing value was changed to value more fitting what could be seen on the form. These corrections were included in the DFA for black lines, etc.

31. Other checks following load included correcting the value for number of rooms when a multiple of 11. Such a value was liable to be generated when a single digit was written but captured from both boxes of the

double box space on the Census form.

32. In all some 86,000 corrections were made at this stage (26,000 of which due to black lines). The number of corrections to each variable is shown in the following table.

Field	Count
Qualifications	38175
Occupation	17003
Industry	12341
Carer	3964
Rooms	3842
Gaelic	1616
Current religion	1615
Religion of upbringing	1513
Ethnicity	1350
Relationship to person 1	941
Relationship to person 2	719
Other variables	3340

Edit and imputation, first application

General

33. The original strategy for edit and imputation was that gaps in a record (either as on the original form or created by the removal of a minimal number of inconsistent items) would be filled entirely by values from a donor record. Partly because there was not enough time to develop such a system and partly to speed up processing times, it was decided that some shortcomings in a record would be dealt with by the insertion of a suitable value according to a pre-set rule. For example, if a record contained values of under 16 for age and not single for marital status, and other items were consistent with the age being under 16, then marital status would be set to single without seeking that value from a donor. The system developed by ONS was called the *Edit and Donor Imputation System* (EDIS) indicating that imputation was by 'donor'. Details are given below in paragraph 41. Given that the developers would have liked more time to test the system – as it was it was available just in time - it performed its task pretty well.

34. EDIS provided the last elements of the range of checks on Census data after those built into were built into data capture (range checks and multi-tick rules) and load (filter rules). The edit part of EDIS applied

checks and, where appropriate, a value was inserted into a record according to an edit rule (e.g. for marital status as described above) or a 'missing' placeholder value would replace a value causing inconsistency. The placeholder would later be filled with a substantive value in the imputation stage of the program. In general, the minimum number of values was inserted to remove any inconsistencies.

35. In addition to the rules inserting specific or 'missing' values, there were a number of 'soft edits' that were aimed at quantifying the frequency of unlikely but possible combinations of values.

Pre-process

36. The first process in EDIS was 'pre-process' in which the data got a little preparatory grooming. For example, a value generated by a failure of a multi-tick rules would be replaced by a 'missing' placeholder value. Household records would be given grid references according to the postcode of enumeration; this was so that the distance between a record with gaps and a potential donor record could be calculated.

Consistency check and edits

37. The next process applied the edit rules inserting specific or missing values. Information about the number of times such rules were used can be found in the reports on Census questions on the GROS website (2001 Census variables).

38. There were a few exceptions to the general rule of 'minimum change' caused by a certain hierarchy among variables. First, variables that featured in filter rules (previously applied in Load) were not easily changed as this might have required changes to dependent variables. Second, when relationship variables were in conflict with other variables it was the relationship variables that were changed, even when changing, say, a single value of age would have resolved the inconsistency. There were no checks on the consistency of relationships among three persons taken together. It is recommended that the relative status of the checks and variables is re-assessed so that the values supplied on the form for the variable relationship are retained more frequently. Also, time should be taken to ensure that a full set of checks on relationship are included in the edit process.

39. Part of edit that did not change any data was the suite of 'soft checks'. These checks highlighted records with unlikely but possible combinations of variables. A selection of household and person records failing these was checked and forms inspected but none initially appeared to show a need to take any action. But counts given by soft checks seemed generally to underestimate the number with the given characteristic in the output database. This will partly be due to the

additional records added by the ONC. The soft checks done as part of EDIS are shown in the table below with the number of times each was triggered.

Soft check	Count
Accommodation rented from the council is unlikely to be rent free	34800
Brothers and sisters are unlikely to have an age difference > 30 years	18546
Two people living as partners are unlikely to be of the same sex	12334
“Not self-contained” is unlikely to have building type <> “part of converted or shared house”.	11806
A house, bungalow or purpose-built flat is unlikely not to be self contained	10899
It is unlikely that a mother will be 50+ older than her son/daughter	7028
Parent unlikely to be only 13 or 14 years older than child	7000
A stepchild is unlikely to be older than his step father/mother	6872
It is unlikely that person would have >2 mother/fathers and stepmother/fathers in household	5204
It is unlikely that a person would have more than one step-parent in a household	3910
A house, bungalow or purpose-built flat is unlikely not to have sole use of bath/WC	2841
Temporary accommodation with central heating	2375
Accommodation unlikely to be self-contained if there is no sole use of bath/WC	2354
Grandparent is unlikely to be only 26-29 years older than grandchild	1604
Person age 55+ are unlikely to be a student	1538
Persons aged <35 are unlikely to have Activity Last Week of "retired "	1052
Temporary accommodation rented from social sector	411
Persons age 16 or 17 are unlikely to be divorced	130
A person aged <2 is unlikely to be able to speak Gaelic	127
A person aged <3 is unlikely to be able to read or write Gaelic	99
It is unlikely that the oldest person in the household is less than 16	84

Person aged <16 and with Country of Birth "elsewhere" unlikely to have marital status <> single	26

40. One of the checks gave an unheeded warning of trouble later in the production of tables. This was that the check with the highest number of failures among soft check for households was “Accommodation rented from the council is unlikely to be rent free”. When tables on household tenure were produced, a reaction from many users was that many households who had ticked ‘living rent free’ in response to tenure were, in fact, receiving housing benefit. Although there was little that might have been done to change the data that was collected, we might have taken this likely form-filler error more into account when designing tables. In the event, supplementary tables on tenure had to be produced to help users interpret the tables originally produced (see paragraph 170). Another soft check that may have alerted us to a problem we had to deal with later was that showing the number of partners of the same sex. However, we had little idea of what the correct number might be, and it was only when checking the forms and data for households with couples of the same sex that it became clear that many were wrongly recorded as such (see paragraph 49). The soft checks built into EDIS were supplemented by others done in the DQMS on things such as the difference in the ages of a husband and wife.

Imputation

41. Imputation by donor was carried out by finding a donor household (in the same EA) that could supply all of the missing values required in a recipient household. The household and the persons within the household would have to match on a range of variables determined by those needing imputation. The values from the donor household were checked for consistency with non-missing items in the recipient household. Where several possible donor households were selected, the best was chosen by scoring each (positively) by how well it matched the recipient according to a second set of variables and (negatively) the extent to which the donor already been used. If there was still no single best donor, then the donor geographically closest to the recipient was chosen. (The sequence just described would be cut short with the first donor that passed the consistency checks, matched on the second set of variables, hadn’t been used already and was within 5 km of the recipient.) Imputation by donor had a number of fall-back processes if no suitable donor household could be found. First imputation would be attempted taking each individual separately, and if that didn’t produce donors then each variable would be treated separately. Failing this, then a fall-back ‘bank’ of ‘typical’ households would be used simply to replace the recipient in its entirety; this bank contained households of up to eight

persons. If the recipient contained more than eight people, then it would be edited clerically with corrections being made via a 'son of fallback' DFA (see paragraph 48).

42. The consistency checks for potential donors were on the same lines as those done in edit, and they suffered from the same shortcomings. For example, consistency checks were much stricter on parent-child relationships than on the equivalent step relationships. The rules would block imputation of natural parent-child relationship but accept step relationship. The real error could have been in another variable altogether probably age.

43. The search for donors was originally meant to take place in the whole of an EA. Hence the whole EA became a 'hot deck' for values for imputation (compared with the 1991 system where the last half dozen values encountered in processing the database were all that the imputation process had to choose from). In practice, since imputation was easily one of the slowest processes in downstream processing, the larger EAs were split geographically or according to how they were classified for ease of enumeration for ONC estimation or both.

44. The imputation part of EDIS imputed values for all Census variables except

- Religion, where since the questions were voluntary, it was inappropriate to impute something the form-filler had exercised the right not to supply an answer

- remote postcodes, as these items were to be imputed in a special process later on (see paragraph 78).

45. The proportion of records where a value was imputed in each category of each variable is shown in reports on Census questions on the GROS website(2001 Census variables) . The EDIS process generated some diagnostic counts that indicate that

- taking a household as the household record (including dummy records for vacant property etc) plus the records for the persons belonging to the household, some 65 per cent required some intervention by EDIS. This figure consisted of 19 per cent where edit rules were applied and 46 per cent for which one or more values were imputed without any prior application of edit rules.

- taking person records individually, 39 per cent of records required some intervention. This figure was made up of 13 per cent where edit rules were applied (including 11 per cent requiring subsequent imputation) and 26 per cent for which imputation alone sufficed.

- persons in communal establishments required some intervention by EDIS in 53 per cent of cases consisting of 32 per cent where edit rules were applied and 22 percent needing imputation alone.

(Percentages rounded independently.)

46. Imputation brought a few oddities that have been spotted by users. For example, imputation was done separately within the household and communal establishment populations. This led to the ability to speak Gaelic being imputed into records for the residents of an old persons' home in Wick (where Gaelic speaking is not prevalent) because the donors (in Eilean Siar) were in the same EA.

47. A piece of research that has not been done is to quantify how values were imputed into near empty records. There was a feeling during the enumeration of the 2001 Census that as long as the enumerator could induce a reluctant householder to complete enough of the form for everyone in the household to pass the '2 of 4' test, the imputation should do a good job of completing the rest of the form. It is recommended that there should be a check on how well the population of such households resembles the population at large – or resembles the data for households added by the ONC.

Data file amendments: son of fallback

48. DFAs were created to correct households failing the standard fall-back options in imputation. ONS also carried out a few corrections on behalf of GROS at this stage. Altogether only 718 values of variables were corrected at this stage, mostly to relationships in household.

Data file amendments: same-sex couples and other corrections

49. Investigation of household composition and relationships in households threw up that fact that a large proportion of couples given as same-sex couples after EDIS were not genuine. A number of errors had inflated the figure. Quantification of the error for same-sex couples in one EA showed that

- some 12 per cent of same sex couples were recorded as such because one member of an opposite-sex couple had ticked the wrong box for sex.

- 10 per cent stemmed from duplicate records that had not been removed by the de-duplication rule (see above, paragraph 22) because the form-filler had not duplicated the entries closely enough. The duplicate records were nevertheless identical in the variable of sex; and the relationship between some of these duplicates had been imputed as 'partner'.

- 6 per cent were genuinely distinct persons, usually related (e.g. mother and daughter), for whom the relationship of 'partner' had been (wrongly) imputed

- 9 per cent were other cases were the relationship of ‘partner’ had been imputed. For some of these cases, the donor household itself would have been in the first of these groups.

50. Other amendments made at this stage included

- corrections to type of communal establishment generated by further comparisons with the Communal Establishment Register (see paragraph 27).

- position of individual in communal establishment; where a member of staff in a CE had completed the I form for a person. The form-filler had often entered his or her own status in the CE rather than that of the person to whom the form related. It is recommended that the instructions on the I form clarify that ‘position in establishment’ relates to the person covered by the form and not the person who happens to complete it.

- further inconsistencies between postcode of enumeration and ED that came to light

- further corrections to records created from England forms collected from Scottish addresses. There were not enough of these to cause any great problem

- further corrections to occupation arising from cross-tabulating this item against qualifications, age, and industry, to various other items from cross-tabulating qualifications against age, marital status against age, and whether a carer against age,

- re-applying filter X after values for age, occupation, etc had been altered or imputed

51. DFAs submitted at this stage contained 90,000 corrected values. The variables corrected are shown in the following table. (At this stage of processing all of the separate variables of relationship of a person to other members of the household had been combined into a single item. The item consisted of a group or ‘string’ of characters, each character representing the relationship of the person to each other member of the household.)

Field	Count
Industry	21853
Occupation	17493
Position in establishment	10370
Relationships	7720
Qualifications	7691
Relationship to person 1	7210

Relationship group	7206
Travel destination	3133
Travel indicator	2544
Highest level of qualification	1570
Sex	1491
Marital status	244
Other corrections	1269

52. ONS submitted DFAs to Scottish EAs with 2217 corrections to occupation necessitated by errors in data capture. A few other corrections were made to the hours worked variable and ethnicity.

53. By the time these changes were prepared, the database had progressed to beyond ONC update, i.e. extra household and person records had been added from donors by the ONC. It was necessary for the update making these additions to be applied so that the ONC team could get a quick assessment of the outcome of the process. Nevertheless, the database for each EA was rolled back to pre-Update stage for the DFAs to be applied as we wished to correct any ONC donor records before they were replicated as recipient households and persons. Once the DFAs had been applied, the database went through the ONC Update stage a second time. More is said below (paragraphs 57 et seq) about the ONC process.

Other checks

54. To assess the affect of EDIS and the above DFAs, the DQMS was primed with the requisite version of the database for each EA so that comparisons could be made between Post-Load and Post-EDIS distributions. Such comparisons are built into the reports on Census questions on the GROS website (2001 Census variables) .

55. Analyses of the numbers of Armed Forces in various parts of the country were made and compared with other information held by GROS. It was decided that the shortfall in 2 parts of Scotland (Argyll & Bute and Moray) was too great not to be left untreated. Corrections were made as part of the ONC process (next section, paragraph 67).

ONS assessment of the edit and donor imputation system

56. The following paragraphs are extracted from the ONS evaluation of EDIS.

Although EDIS in general performed well and within planned running times, some aspects might have worked better if there had been more opportunity for testing. It did not prove possible to carry out a full run-through of the EDIS system on the data collected during the 1999 Dress Rehearsal. This would have afforded the opportunity of checking whether ideas which appeared sensible in theory would stand the test of practical application on live data. Rehearsal data needed to be delivered more rapidly, at least on a small scale, or the Rehearsal itself should have been brought forward so that it took place more than two years before Census day.

Inability to test the system meant that assumptions about how well the public would follow instructions, for example on answering or skipping certain questions, proved to be not entirely valid. This impacted most noticeably on imputation of age. Within EDIS, a number of assumptions were based on age being correct rather than other items. However, year of birth was occasionally mis-stated, not scanned correctly or given a wrong value during processing. Particularly when there was an error in the next to last digit of the year, EDIS may have imputed for a range of items where no value was needed, or conversely set reported data to 'no code required'. Further checks could have been tested if more time had been available to investigate real data from the Rehearsal: for example, identifying circumstances where age needed correcting rather than other fields, or querying large differences between the ages of spouses or partners. Nevertheless, a few households contained multiple errors which would have been difficult to resolve accurately by any automatic editing system.

A single edit and imputation system was designed to deal with the censuses in England, Wales, Scotland and Northern Ireland, which all had slightly different requirements. Variations in the design of the Census form and in editing requirements meant that great attention had to be devoted to ensuring that the processing for each country was carried out to the desired standards.

One Number Census

General process

57. The aim of the One Number Census was to make good any shortfall in enumeration to an agreed timetable. Users were willing to accept a delay to delivery of output as long as there was the expected improvement in quality and the lengthened timetable was adhered to. A guide to the number one census can be viewed from the ONS website. The following paragraphs give a brief outline of the ONC process and how it fitted into downstream processing.

58. After data from the Census and Census Coverage Survey (CCS) had been delivered by the contractor and loaded into databases, GROS staff

matched Census records that had been through EDIS with returns from the CCS. Data and images of forms for matching were held on servers at ONS Titchfield and were accessed using a secure communications link. The software for matching had been written by ONS.

59. Following matching Census and CCS records for an EA, the number and characteristics of households and persons missed by the Census were estimated using ONS software. These estimates were added to numbers of enumerated households and persons and the results scrutinised by a GROS quality assurance team (with, on occasion, the participation of members of the ONS team). The scrutiny of estimates required statistics from various sources including figures from the DQMS following ONC Update: populations for each council area by sex and single years of age; and particular sub-populations (prisoners, students, armed forces). Adjustments may be made to the estimates, and the results scrutinised again. When acceptable estimates had been produced, the requisite donor households and persons were identified. The products of the ONC process were data files with each record consisting of the location (in the form of a household or person identifier) where an additional record was to be put, and the identifier of the chosen donor record whose data was to be copied to the recipient. This copying process was called 'ONC Update'.

60. The result of the ONC process was the addition of

- 60,068 person records to enumerated households. This includes 1415 persons added so as to build up the armed forces population in enumerated households in Moray (see below, paragraph 69). Note that at most one person was added to an enumerated household. It is not clear why there was this restriction.

- 136,969 person records in 72510 wholly missed households. This includes 85 persons added as one-person armed forces households in Moray.

- 1950 added to the communal establishment at Faslane.

61. Once a particular postcode had been selected for the addition of one or more household records, the precise location of an added household record was any slot where a household had been recorded as absent or as having refused to take part in the Census. Once any such slots had been used up, there was a fall-back process of searching for gaps in the run of records where there was no household/CE record with a given form identifier. The selection of areas for additional records was done partly on the basis of numbers of absent and refusing households, so the fall-back was used infrequently. Also the work to fill gaps caused by lost forms (see paragraph 25) meant that the fall-back was even more rarely used extensively in a single area. This lack of recourse to the fall-back was fortunate because a fault in the program meant that donors used in the fall-back were not penalised against further use. There are one or two isolated instances where a donor household has been imputed

several times into the same area. It is recommended that broadly the same approach to find locations and donors for recipients be used but with a penalty method for restricting the re-use of donors. GROS should check the Update files for clustering of donors.

62. The ONC processes were written and run by ONS on behalf of all 3 Census Offices with some glitches caused by split postcodes in Scotland not recognised at outset. Also the ONC update in a few cases added a household with a postcode of enumeration invalid for the ED, which a check of the Update files would have spotted. It is recommended that a split character is adopted as standard throughout the whole of the UK even though it will be filled with a space for the most part. GROS should check the Update files for combinations of postcode and ED.

63. After the ONC adjustments had been made to the database, a fresh evaluation of data for the whole of Scotland considered that not enough persons aged 0 had been added although the total for all ages was satisfactory. Given that the Update process had already passed, it was decided to achieve the desired number of zero year olds by changing to zero the age of persons aged 1 to 9 added by the ONC process. This change was included in a DFA (see below, paragraph 82.)

64. The ONC process was generally successful in that no revisions to its estimates of Census day and associated mid-2001 population estimates were subsequently made.

Special geography for ONC

65. One of the principal sets of data with which to compare estimates of population produced by the ONC was of GROS' mid-year populations estimates rolled forward from 1991. These estimates were based on, among other things, data on migration from the National Health Service Central Register (NHSCR), whose principal geography was the health board area (HBA). These areas coincided with council area for all of Scotland except within the former Strathclyde region. The 8 EAs used for processing up to the stage of the ONC were based on council areas. For the ONC, the 3 that made up the Strathclyde area had to be re-apportioned into areas based on health boards. None of these 3 EAs could progress to its ONC stage until all 3 were ready for re-apportionment. The ONC process continued with HBA-based EAs. No EA could progress to post-ONC stages till all had gone through ONC and been re-apportioned back into council area-based EAs. It is recommended that no such switching between council area-based and HBA-based EAs is done. Population estimates using more detailed figures on migration than those from the NHSCR should help.

66. Another problem relating to HBAs that came to light very shortly before the Census was that, contrary to our understanding, the areas had never been changed since the boards were set up in 1975. All of GROS statistical work had been on the wrong assumption that the

boundaries of HBAs were changed in line with those of corresponding local authorities. So the population estimates for HBAs being used to assess the ONC were based on the wrong areas. It was decided that for Census processing, including the ONC, we should stick to the wrong areas (calling them 'operational' HBAs or some such name), but when it came to producing output, each Output Area (OA) would be assigned to a corrected HBA so that statistics for OAs would aggregated to the correct areas.

Armed forces adjustments

67. As stated above (paragraph 55), it was decided, after allowing for the form-filler not recording members of the armed forces as such (see paragraph 156) to make good an estimated shortfall in the enumeration of armed forces personnel in two areas. Because of controls on changes to the databases, there was only one place in downstream processing where records could be added, and this was at ONC update. The addition was made by amending the recipient-donor files that emerged from the ONC process (see above, paragraph 59).

68. The additional 1950 person records for Argyll & Bute were put into the Faslane naval base which had been enumerated and processed as a communal establishment. The age and sex characteristics of the recipients were determined and donors from the same CE found in the numbers required. The ONC Update file for the EA that contained Faslane was augmented accordingly before the Update program run.

69. The procedure for Moray was more complicated in that it was decided that the missed AF personnel should be added to households in the area around Kinloss and Lossiemouth in Moray. Having determined the age and sex mix of the 1500 persons to be added, it also had to be decided what mix of type of household they should be added to. To complicate things further, a household could not act as a recipient for additional person records if it included persons qualifying as members of the Longitudinal Study (as defined for England and Wales). For ease of programming this exclusion applied to Scotland. An Access database was set up with potential donor records and various queries run to produce the requisite number of recipient-donor pairs with suitable characteristics and locations. A record for each recipient-donor pair was added to the ONC Update file.

Edit and imputation, second application

70. Once the ONC Update process had been run, a second version of the EDIS process was run. This was primarily to give each person added to an enumerated household a relationship with each other

member of the household.

Disclosure control

71. The only method of disclosure control to be applied during downstream processing was that of record swapping. Other methods in the total package of measures used were applied as part of output production (see Disclosure Control). For record swapping, households in an area were matched according to number, age group and sex of residents, and a sample of matched households swapped. The process was carried out at the levels of the council area and EA; so some households could be swapped across council area boundaries. A similar process of matching, sampling and swapping was carried out for residents of communal establishments.

72. The proportions of records that were swapped varied slightly within Scotland to take account of the rate of imputation that had already been done on an area's data. The general levels of swapping so as to be comparable in effect to measures used for 1991 output. As part of the general strategy of disclosure control, the rates were not made public.

73. Record swapping was a particular form of modifying the results of the Census before tabulation. Among rejected options was randomly marking values of variables for imputation and assigning an imputed value to replace the original one. Such 'over-imputation' could be carried out at different rates according to variable and geographical area. One problem caused by record swapping is that swapping is ineffective for

- areas larger than the level chosen for the process. Output for such areas (such as the whole of Scotland) is unaffected by record swapping. This presented later challenges in devising disclosure controls for Samples of Anonymised Records (SARs, see paragraph 107).

- tabulations based on areas other than area of residence. Tables counting people by, say, area of workplace an unaffected even for very small areas.

Some of the measures eventually chosen for perturbing data records in the SARs closely resembled 'over-imputation. It is recommended that a more unified set of methods for disclosure control be used, preferably based on modifying the Census database in a way that does not additionally require any modification of tabulated results and that is sufficient for both tabular output and SARs.

Sundry other processes

Extract postcode counts

74. Following record swapping there was no process that could alter the numbers of residents and households. This was therefore the earliest (therefore the best) point at which to extract these counts that would be used to group postcodes into Output Areas (OAs) that met the minimum size of having both 20 or more households and 50 or more residents. These thresholds were one of the measures to ensure the confidentially of tables produced for OAs.

Early population counts

75. For the same reason (stability of counts), once the 8 EA databases had been loaded into the DQMS, counts could be produced by age and sex for council areas. This was a relatively easy task because a code for council area was built into the form identifier information. Corresponding counts with age calculated as at 30 June 2001 (instead of as at Census day) were also produced. These extractions formed the basis of the first report from the 2001 Census the 2001 Population Report (September 2002). No other geography was readily available in the DQMS, and age-sex counts for health boards (released 2 months later on the GROS website) had to be aggregated from counts based on postcode of enumeration. The original intention was to produce these counts from the full output database, with more detailed results following close behind. In the event, the data wasn't ready for loading into the output database at the time needed for the report. The main difficulty in using the DQMS for output included the fact that results had to be aggregated from eight databases.

76. Once the figures on age and sex had been released there was no possibility of amending either of these items if investigations found it to be wrong. Inconsistencies had to be removed by changing other items even when checks on household composition, occupation and qualification showed that would be the right correction.

Household composition algorithm

77. The household composition algorithm grouped individuals within households into families. The basic approach is first to put any pair of spouses or partners into a group and then link each unpaired person to his or her parent (who may or may not be a member of a couple). Thus family groups are formed alongside ungrouped individuals. One problem that was fixed just before operation was that the HCA would combine all concerned into a family when a child was recorded as the child of one person and the step-child of another and the older generation had no relationship recorded (because on different forms) or were recorded as 'not related'. This process would have wrongly created many same sex

couples. It was decided not to group people in these circumstances unless the two persons of the older generation were of the opposite sex. It was believed that an error in the relationship data (wrong recording of step-child) was more likely than that there was a same sex couple family with children. Other problems were identified but not all of them were fixed other than to ensure that the classifications later used in output took account of any remaining anomalies. The algorithm left small number of households and persons with an inconsistent set of household composition variables (see paragraph 171).

Postcode imputation

78. In response to user demand and led by GROS, the Census Offices included in downstream processing a process to impute missing remote postcodes. The postcodes may be entirely missing or perhaps a partial postcode may have been captured in which case the rest was imputed. The general principle was the same as that for EDIS i.e. for each record requiring imputation, a donor record was found that matched the recipient in a number of key variables. The process was specified by GROS and developed by ONS. GROS also produced test data (for UK) to check that all aspects of the process were checked; this proved to be a much bigger task than anticipated – even given the experience of providing test data for the HCA.

Dwellings

79. There was a requirement to group household spaces (occupied households plus properties that are vacant or second or holiday homes) into dwellings. The 2001 algorithm for this was devised so as to use information that was being collected in any case rather than, as in 1991, using extra information collected (expensively and not too well) for the purpose. The idea was to group into a 'dwelling' non-self-contained household spaces within partially converted accommodation at the same address. Information on address on the household form (or in the address database for pre-listed addresses) would have to be matched. The problem turned out to be that the data captured for non-pre-listed addresses was not sufficiently structured for it to be matched successfully to addresses for pre-listed addresses. To get round this ONS decided to use the grid-references of addresses to make these matches and amended the algorithm accordingly. This option was not open to GROS as grid referencing was not done at the address level. Instead it was decided to inspect clerically the outcome of the original version of the algorithm and make adjustments by means of DFA. (More information is available from the Identification of dwellings, 2001 (with comparisons with 1991) PDF) .

80. The quality of information on self containment and type of

accommodation was not evenly good especially in the accommodation most likely to be subject to being grouped into dwellings. In particular only 14 percent of households that are not self-contained were given as 'part of converted or shared house'.

81. The clerical process entailed

- Gathering data on all non-pre-listed addresses and all households recorded as not self-contained. Part of this process included fiddling with the different formats of the postcode and ensuring that split postcodes were properly recognised. It is recommended that a common format of postcode is used throughout Census operations.

- Examining this data (sometimes including images) for all households in any postcode with at least one non-self-contained household.

- Amending or adding to the dwelling codes generated by the automatic process

- Assembling data for the DFA.

Amend number of 0 year olds

82. As mentioned earlier (paragraph 63), it was decided to adjust the age distribution of those aged under 10. This was done by selecting a small number in each of the ages 1 to 9, and changing the age to zero with consequential changes to date of birth, student status and term-time address, Gaelic ability, whether a carer, address one year ago, and the travel items. About 1260 person records were subjected to this change.

Postcode validation and final check of input database

83. One of the final processes (needed before complex variables on migration and travel were derived) was to check the 'remote' postcodes: address one year ago and travel destination. A process to identify any postcode not recognised in the UK geography database was run for each EA. A few Scottish remote postcodes turned out at this stage to be invalid and needed correction. There were also invalid postcodes for outside Scotland that had to be referred to ONS and NISRA. Similarly Scottish postcodes on England and Wales and Northern Ireland databases were referred to GROS. Scottish postcodes on records for any part of the UK had to belong to a fixed or 'frozen' set used for data collection, processing and output. This 'freezing' of the postcode base of the Census enabled the geography of place of residence, residence one year ago, and travel destination to proceed smoothly, without the need to update the postcode base – index and boundaries. It is recommended that GROS continue to freeze its postcode base throughout the enumeration and processing of the Census.

84. A final check on postcode of enumeration showed a couple of cases where the postcode did not belong to the correct ED. It is believed these errors were introduced by the ONC Update.

85. The DQMS was used to investigate the outcome of the HCA. Analyses included looking at the numbers of unrelated people in households, the formation of step-families, treatment of households using more than one form, etc. Analyses were made of the new variables that the HCA had created. These were classifications of persons and households. The HCA also created a new type of record for each family it identified with variables describing the family. Unfortunately the DQMS did not support the new type of record and analyses at the family level were not possible. It is recommended that the DQMS should mirror the processing database as completely as possible, in particular, by including the record for the family created by the HCA.

86. The mainstream processes included one that did a final check consisting of a selection of filter rules and consistency checks. GROS added a few checks of the travel questions against age and of Year Last Worked and Ever Worked, plus a hunt for any missing values that should have been mopped up in imputation.

87. Finally, it was discovered that the enumeration of Persons Sleeping Rough (PSRs) had grouped all such cases within a council area into a single quasi-communal establishment with type of establishment coded as 'Other' instead of 'PSR'. Amendments to type of establishment were included in the final batch of DFAs. Even with this correction, the treatment of PSRs means that they can only really be presented in output tables at the level of council area. It is recommended that PSRs be recorded where they are enumerated.

Data file amendments for above

88. Several DFAs were produced at the tail end of downstream processing. The first was the one that concentrated on dwellings and adjusting the number of 0 year olds. These contained amendments as follows:

Field	Count
Family type (generation 1)	2393
Relationships	2168
Dwelling number	1812
Generation in family	1274
Method of travel	1264
Travel indicator	1264
Age group	1261

Age	1259
Date of birth	1259
Travel destination	1255
Carer	1253
Gaelic	1253
Student indicator	1253
Term-time address indicator	1253
Migration origin (5 variables)	1253
Travel destination indicator	1253
Year last worked	1228
Family number	1213
Family type (generation 2)	1139
Relationship to Person 1	844
Relationship group	844
Occupation	818
Other variables	280

89. There was also a set of some 64,000 corrections to the eastings and northings of postcode of enumeration. These corrections were not made correctly as they omitted the last character of each value which had the effect of putting the postcode several hundred kilometres too far south west of where they should be. This error materialised later when we were doing tabulations of distance travelled to place of study or work. The distance variable on the output database was corrected but the eastings and northings on which the distances are based may not have been – either in the 8 EA input databases or on the output database. It is recommended that these databases be checked and these items corrected, if in error.

90. A DFA was also made mopping up a small number (some 400) of non-Scottish postcodes needing correction, and around the same number of further corrections to occupation and industry and a few corrections to other variables. A few adjustments were made to correct data on persons sleeping rough, and to a derived variable on living arrangements that had been discovered to have errors in draft output tables.

Derived variables

91. It was early decided with ONS and NISRA, that as well as variables generated during downstream processing (e.g. by the HCA or dwellings algorithm), a range of additional variables should be derived should be

programmed and added to the EA databases and extracted for output with other data. The alternative would be to derive these additional variables as and when required for tables. The advantage of creating derived variables (DVs) during downstream processing were

- They would be done once and for all and subject to the same rigours of QA that prevailed for other downstream processing operations.

- There would be less danger of duplicate DVs being created by output staff on differing lines.

The advantages of the alternative - creating DVs during table production - were

- As source variables were updated for error or re-specified, the dependent DVs would automatically be updated

- The database would be kept more compact.

In practice, there has been a mix, with many ready-to-use DVs (prepared by ONS) included in the output database, and many others, created to meet particular demands, programmed in output processing.

92. The DVs created during downstream processing were divided into two main groups: those for migration and travel based on remote postcodes and the rest. A start was possible on the latter long before those on migration and travel were begun. (It had been decided to go ahead and create migration and travel DVs so that they could be included in the output database. An alternative would have been to produce the tables on migration and travel without DVs – which was theoretically possible by first creating an intermediate origin-destination product, see paragraph 95. However, while it was possible to test this approach, it was impossible to predict the run times of tabulation programs of the type required.) ONS took longer than expected on the migration and travel DVs because they were still sorting their non-frozen postcode geography in December 2002, and because they needed time to decide whether their 'remote' geography, for areas such as Parliamentary Constituencies and National Parks, should be based on aggregates of OAs or of postcodes. The latter would pass off as exact delineations of areas. GROS were a little bemused at this because even aggregations based on postcodes would only give an illusion, not the reality, of perfection. (Address referencing was only done for England and Wales for address of enumeration not for remote addresses, where, as for Scottish data, the postcode was the building brick.)

93. DVs for migration and travel were specified for each type of area in pairs:

- The first DV would in effect mirror area of enumeration so that where OAs of enumeration were grouped into, say, health board areas (HBAs), OAs for address one year ago would be similarly grouped

– with additional origins for non-UK origins and address not known, etc.

- The second would put each person into a small number of categories such as, for health board area, 'not a migrant, 'address one year ago not known', 'migrant in same HBA as current address', 'migrant in different HBA in Scotland', 'migrant from elsewhere in UK', 'migrant from outside UK'.

While there was some overlap within the pair of DVs for HBA, they allowed the identification of migrants into and out of each HBA for the tables on migration in the Area Statistics (more later, paragraph 152).

94. The Migration and Travel DVs were set up in output as 'hierarchical' variables with, say, a Scotland-HBA-OA hierarchy given among others within the geography group of variables. While this made the HBA geography easy to find and handle in simple tables, working with these variables made more complex queries harder and added to processing times. GROS specified a set of DVs with each type of area presented in a non-hierarchical way, but, in the event these were programmed but never fully tested (although they are listed in the SuperCROSS field menu).

95. Another approach might have been to generate tables counting migrants into and out of an area from double geography (i.e. origin-destination) statistics. Each table could be produced from an origin-destination 'matrix' with each cell of the matrix containing statistics on the relevant characteristics of migrants from one area to another. Instead of setting up derived variables, we would need the means of generating each matrix and then grouping the cells needed to give the flows into and out of each area. The matrices would constitute the products known as origin-destinations statistics. It is recommended that such an approach be considered, given possible improvements in tabulation and data-handling software.

Output

Introduction

96. The output produced from the 2001 Census depended largely on what users had stated as needed during consultation. The Census Act 1920 requires the laying of reports before the Scottish Parliament and the supply of 'abstracts' at the 'cost and request' of customers. In practice there was considerable overlap between the reports and pre-planned abstracts with reports for council areas or health board areas generally being printed versions of abstracts. Versions in machine-readable media were available for those types of area and all others in the Census range based on the 'Output Area' – which contained, on average, around 50 households. For 2001, there has also been a

considerable expansion of the service of providing tables specially commissioned by individual customers.

Area Statistics

97. Consultation on pre-planned abstracts concentrated on a particular set of products under the generic heading of *Area Statistics*. The consultation took place at both UK and Scotland levels, with general principles being discussed and agreed with representative users in a number of advisory groups (both UK and Scotland) with presentations on road-shows for all users. GROS conducted two rounds of road-shows on the Area Statistics. There was also a separate consultation exercise for products under the heading of *Origin-Destination Statistics* (O-D Statistics). GROS took the lead in this consultation on behalf of all three Census Offices and, unlike, the Area Statistics, there was a stronger pull towards maintaining a common UK line on what was, in its very nature, a UK product. ONS took the lead on the rather more specialised abstracts known as *Samples of Anonymised Records* (SARs).

98. At the end of the consultation on *Area Statistics* over 300 tables had been specified, around a half of which were in a set common throughout the UK and a half were specific to Scotland. The number of tables presented quite a challenge in production more below). An alternative approach might have been to produce only the simplest tables as standard and react to individual users' demands with ad hoc tables. This would be in accordance with a notion that table development proceeds 'organically' from exploratory univariate analyses to more complex tables that depend on the results of the initial tables. However, users justified such a large number of pre-defined tables because there exist groups of users (e.g. in local authorities) with extensive and similar needs. The Census Offices would also be able to control any risk of disclosure from producing tables that were similar but not exactly the same. The risk would be that small categories of person or household classifications would be exposed by 'differencing' similar tables. It is recommended that the approach to producing a wide range of standard tables be reviewed.

99. The *Area Statistics* fell into two main groups with several categories in each:

- Tables produced for the smallest of area, the *Output Area* (OA), and all areas recognised in output. These areas are constructed from the best approximation of OAs. For these tables there was a minimum threshold of 20 households with residents and 50 residents. The types of table were denoted *Census Area Statistics* (CAS), *Univariate Tables* (UV), *Key Statistics* (KS), and *Census Area Statistics Theme Tables* (CAST)

- Tables produced for larger areas, electoral wards or groups of wards, postcode sectors or groups of sectors, and larger areas such as

council areas, health board areas, and parliamentary areas of various kinds. For these tables there was a minimum threshold of 400 households with residents and 1000 residents. To meet these thresholds some wards and sectors had to be combined (and other adjustments made, see paragraph 116 below). The types of table were denoted *Standard Tables* (S), and *Theme Tables* (T).

There was also a set of 5 'profiles' for all types of areas devised especially for web dissemination with users new to the Census in mind.

100. The *Area Statistics* were produced using tabulation software called SuperCROSS, part of a suite of programs from an Australian company called Space-Time Research. Data for loading into SuperCROSS was assembled from extracts taken from the 8 EA databases created by downstream processing. Because of delays in creating the variables for Migration and Travel, two sets of extracts were taken. The first, without Migration and Travel, was taken and loaded into SuperCROSS for the production of most of the *Area Statistics* – all but Migration and Travel. Later a second extract provided the data required for the remainder.

101. Because of differing approaches to disclosure control in the 3 Census Offices, in particular concerning the modification of small numbers in cells of tables (small cell adjustment, SCA), each Office had a database containing records for the data collected in its area. Having separate databases gave each country greater control over its own production. Even with a single UK database, tables for each country would still have had to have been run separately – with or without small cell adjustment – destroying any advantage there might be in a single database. Nevertheless, because the Offices had to cooperate in the production of tables on migration and travel, exchanging programs and output (more below), it was necessary that the 3 databases had a common design. For example, variables relating to the place of origin of migration (e.g. local authority of residence one year ago) had to be common throughout the UK with a standard list of UK local authorities and non-UK areas. At first, responsibility for database design lay with GROS, but was transferred to ONS after a pilot stage.

102. To data 6 reports have been laid before the Scottish Parliament. Three of them contained printed versions of tables from the *Area Statistics*. The other three were

- The *2001 Population Report* containing counts by age and sex for council areas at Census day and at 30 June 2001

- The *RG's 2001 Census Report to the Scottish Parliament* containing tables on a wide range of Census topics with comparisons where possible with 1991 and with a commentary.

- The *Gaelic Report* giving detailed figures for this topic for areas grouped according to the Gaelic ability on their population with comparisons with earlier Censuses.

Two Occasional Papers have also been produced – on Inhabited Islands and on Migration – with another expected in September 2006 on Travel. The Census Act 1920 does not specifically cater for reports that are not laid before Parliament nor paid for, so it is recommended that the legislation be reviewed to ensure that it does cater for the full range of planned activities and outputs.

103. The Scottish Executive stood in on behalf of users generally and paid for

- The production of all *Area Statistics* other than the 3 printed reports. Since all tables appeared in the printed reports at least for Scotland as a whole (and for some tables for some areas within Scotland), the SE funding was seen as providing the marginal costs of producing these tables for the remaining areas for which output was produced.

- The dissemination of the *Area Statistics* via *Scotland's Census Results On-Line* (SCROL). SCROL took the form of a website (SCROL) and a series of CDs, later re-issued in a more compact DVD form. The CDs (and DVD) were supplied with software designed to read tables in the format produced by SuperCROSS. On some CDs (and on the DVD) tables were also supplied in a 'comma separated value' (csv) format for input into software of the user's choosing.

104. The *Area Statistics* were also delivered in a 'bulk supply' form. The format (also csv) was aimed at heavy users who would be loading the tables into, say, mapping software or at intermediaries who would supply the tables bundled with their own software.

105. It was decided that producing the *Origin-Destination Statistics* would be too challenging for SuperCROSS. The aggregated counts for agreed tables for the UK would instead be generated directly from the 112 EA Sybase databases. A process to apply SCA would be applied to the agreed sub-set of the products to mimic that provided in SuperCROSS. GROS were responsible for specifying the tables and ONS developed the programs to extract the counts.

Origin-Destination Statistics

106. The tables on workplace in the *Origin-Destination Statistics* had to be adapted to accommodate the extension of the relevant Census questions in Scotland to include travel to place of study. This adaptation (and others elsewhere in the output for Scotland) is based on the assumption that, because of the position of the questions on the form, it is workers who are not full-time students who would be those travelling to a place of work, and all other travellers would be travelling to a place of study.

Samples of Anonymised Records

107. The *Samples of Anonymised Records* (SARs) were especially commissioned by the Economic and Social Research Council on behalf of the academic community – but would be made available to any user on payment of a fee to the Cathie Marsh Centre for Census and Survey Research (CCSR). The customer specified a set of SARs that were more detailed than the 1991 equivalents but, in the event, a stricter confidentiality regime in the Census Offices meant that the SARs for public use were less detailed than specified. The SAR based on household records was restricted to England and Wales because of the relatively high risk of disclosure of records with Scottish or Northern Ireland data. Products more in line with the original specification are available in a 'safe setting' at 4 ONS sites. At the time of writing, they have not been made available at Edinburgh or Belfast. It is recommended that, if such products are available in future, they should be available at the outset at sites in all 3 Census Offices.

Eurostat and 'Focus on ...' reports

108. ONS also coordinated the production of tables for the UK for *Eurostat* (the Statistical Office of the European Union) and for non-Census staff in ONS. The latter have been producing the series of '*Focus on ...*' reports covering data from a range of sources on given topics.

Ad hoc commissioned tables

109. GROS staff continue to provide commissioned tables. There have been over 250 requests for ad hoc output containing over 500 tables. This has been one of the most successful elements of output production. The overwhelming majority of these tables have been produced from the SuperCROSS database used for *Area* Statistics. One major exception was a count of the people writing various religions in the write-in box for 'Other religion' in each of the two religion questions. This had to be done clerically inspecting the images of forms because these write-in answers had not been included in the contract for data capture. It is recommended that all write-in answers be coded.

110. At the time of writing, staff producing commissioned output have a set of precedents that govern whether a request table might breach confidentiality. It is recommended that these precedents be formally set out as rules for future work.

Preparation for live running

111. As for downstream processing, the activities of the Rehearsal did not penetrate as far as output, and we went into production with little testing of processes. It is recommended that Rehearsal activities include output production. However, given large differences in table design, and the 3-database approach caused by differences in disclosure control, the offices cooperated well especially over the complications of producing tables on migration and travel.

Disclosure control

112. A separate report has been prepared on disclosure control. However, there were severe problems caused by the fact that the 3 Census Offices had gone different ways on disclosure control. Chief among these differences was that the other two offices decided that they would apply small cell adjustment to all tabular output. GROS did not. This meant that

- Scottish tables were consistent with each other
- Scottish tables for OAs could be aggregated to higher area to give precisely the right figures; this meant that for some products we needed only to produce OA level output.
- Disclosure control didn't depend on the software producing tables and the data could be used indefinitely – including with future tabulation software.

The disadvantages of different approaches within the UK have been hinted at above. Not all these problems were foreseen. It is recommended that those taking decisions about disclosure control take full account of the implications for output. When challenged by a user about why different approaches were taken, the Census Offices agreed the following statement from GROS

> Initially all three Census Offices made a decision to rely on record swapping for disclosure control. This was supported by thresholding the size of areas used to produce results and reducing the amount of data detail in tables for small area types. The potential for deriving smaller areas as the difference between two overlapping areas was also a consideration. Different decisions were taken in different parts of the country. Scotland had smaller basic areas, but did not permit any overlaps. Precise geographies, with the potential overlap, were considered to be a priority for England and Wales.
>
> ONS and NISRA later reviewed their decision on disclosure control. When doing so they decided that they wished not to rely wholly on record swapping and decided to use small cell adjustment (SCA) in addition to record swapping, in order to give additional protection,

because of the perception of disclosure.

The decision to use SCA or not depends on what view you take about users' perceptions. If your view is that the user will see 1s and 2s in the cells of a table as disclosive then you may decide to introduce SCA because it removes them. If your view is that the user will see 0s as disclosive - in that a row or column with nothing but 0s except in one cell appears to be disclosive - then you may decide not to introduce SCA because it increases the number of 0s and hence increases the likelihood of leaving a row or column with one non-zero cell. The GROS view was that either approach to the perception issue could be supported. We were not convinced that there was a case for changing the earlier GROS decision in order to adopt the new ONS/NISRA line. Moreover, whatever the perception, the tables without SCA are not actually disclosive, because of record swapping.

113. A second difference within the UK was that the thresholds for OA-level output (see above, paragraph 99) were lower in Scotland. GROS also applied a threshold for *Standard Tables* using the more detailed 14-category classification of ethnicity. Thresholds. While 5-category tables were included in all sets of Standard Tables, 14-category tables were excluded for Standard tables for wards and postcode sectors with fewer than 50 persons in white or fewer than 50 persons in non-white ethnic groups.

114. Classifications in the various types of tables took account of the average size of the areas for which the tables were to be produced. In particular the number of categories in a classification in a CAS table was generally less than the number in the equivalent Standard Table.

115. Output Areas (OAs) in 2001 was based largely on OAs for the 1991 Census. Hence they averaged around 50 households. As in 1991, once OAs was produced, each was assigned to its ward, civil parish, etc. For output, each ward, civil parish, etc, was the aggregation of OAs assigned to it. Thus no output could be deduced by ‘differencing’ for any area that was not one or more output areas.

116. For standard tables, concern about thresholds and differencing was focussed on wards and postcode sectors. First, any ward or sector below the thresholds for standard tables (see paragraph 99) was merged with a neighbouring area of the same type until thresholds were met or exceeded. Second, the resulting two sets of merged wards and merged sectors were compared so as to identify any ‘slivers’ for which standard table output could be deduced by differencing. Any sliver was grouped with a neighbouring ward or group of wards.

117. Another problem arose about those tables in *Area Statistics* counting people according to the area where they worked (rather than where they lived). In fact there were two problems. The first was that tabulating workers by OA (or ward) of workplace could, in theory, disclose

information about 'employing establishments'. After a debate with ONS and NISRA about legal issues, the NS protocol and practices in other surveys, there was another parting of the ways. ONS and NISRA withdrew these tables for some area types. GROS continued according to the programme of planned *Area Statistics* agreed with users. Main justification for this was that industry and postcode of destination in the output database is not as collected on a Census form in around 30 per cent of cases and comparison with the Inter-departmental Business Register showed differences even when it was. We acknowledged that there were presentational problems in using this argument, as we have no wish to undermine the users' confidence in the data. Figures on data quality (by variable) have been put on the GROS website (census variables).

118. A similar problem presents itself for tables commissioned by SE to show the characteristics of households with children whose place of study was in certain postcodes (i.e. the educational qualifications of mothers of children attending each school in Scotland). This information was produced and used for statistics about the 'value added' by schools.

119. The second problem with the workplace tables in the *Area Statistics* was that record swapping (which, in effect blurs area of residence) offers no protection to output on area of workplace. Accordingly GROS had to resort to small cell adjustment for these tables. To do this, GROS had to get the SCA facility of SuperCROSS installed quickly – with the permission of the Australian Bureau of Statistics who had originally commissioned it from the SuperCROSS supplier.

120. One of the difficulties in conducting debates on these issues was that there was no longer a suitable forum for the purpose. It was difficult for GROS to engage with ONS Methodology Group to whom the issues had been referred by ONS Census. We dealt with ONS Census staff, who had their own difficulties with Methodology Group, and saw getting a reaction to GROS views as complicating a debate that was complex enough already. It is recommended that the committee structure for Census policy matters is maintained until all policy issues are dealt with and that membership is drawn from all interested parties. This did not happen for 2001.

121. Small cell adjustment was not applied to migration tables in the Scottish *Area Statistics* other than the SCA applied to components produced from the England and Wales and NI databases counting migrants from Scotland.

122. Another decision made late was to restrict greatly the range of tables on ethnicity and religion on the SCROL website. The Census Report *Key Statistics for Settlements and Localities Scotland* showed the same small number of Moslems and Pakistanis in a particular settlement. It was thought that the reader might wrongly assume that all Pakistanis were Moslem and vice versa (in fact, although there was a large overlap, the two groups were not the same). At the time of publication, military

action had started in Afghanistan and it was believed that statistics on these topics were too sensitive to be given a high profile by being included in tables on the SCROL website.

123. Finally, concerns about disclosure control were the cause of yet another late change to planned output. GROS, having been given the lead for specifying *Origin-Destination Statistics,* tried many times from May 2001 to get confirmation from all 3 Census Offices to proceed with output resulting from consultation with users. We warned that GROS would proceed on stated lines unless stopped. Many meetings had taken place with users and a lot of development of the product before we finally got a response (in September 2003). The outcome of the late intervention was:

- In Northern Ireland: local authorities were replaced with Westminster Parliamentary Constituencies for output at the highest area level. Travel output for NI was withdrawn at OA level. Other suggestions from NISRA have not been adopted (e.g. asymmetric flows).

- GROS decided that TV301 (an OA level table) had to be modified using SCA in line with the *Area Statistics* workplace tables so as not to undermine the protection given to the latter.

Eventually output for each country was produced (or not) with or without small cell adjustment as follows:

Product	Area level		Scotland	England and Wales	Northern Ireland
Special Workplace Statistics and, in Scotland, Special travel Statistics	Local authority		No SCA	SCA	SCA
		Ward and, in Scotland, Postcode sector	No SCA	SCA	SCA
		Output Area	SCA	SCA	Not produced
Special Migration Statistics	Local authority		No SCA	SCA	SCA
		Ward and, in Scotland, Postcode sector	No SCA	SCA	SCA
		Output Area	No SCA	SCA	SCA

124. The different decisions on SCA had implications for Samples of Anonymised Records. The main disclosure concern for SARs is the risk that a record in a SAR relates to a person or household that is unique in the population at large, especially if it can be shown that the person or household is unique. Since Scottish tabular output was not subject to SCA, it contained cells with 1s even for tables for the whole of Scotland. A major part of the disclosure control regime for Scotland for the SARS was the production of thousands of three-way cross-tabulations aimed at identifying 'population uniques' that, if in a SAR, might be disclosive if not dealt with. The two means of dealing with uniques were

- To reduce the number of categories in classifications
- For cases still unique after changing classification, to 'perturb' the unique record.

The outcome was that Scottish data, subjected to the above changes, was included in the 3% individual SAR. Scottish data was not included in the 1% household SAR because the necessary changes in classification were too drastic for users.

125. As can be seen there was hardly a unified treatment of disclosure control either among products or within the UK. It is recommended that a more unified approach be adopted for Area Statistics, Origin-Destination Statistics and SARs, preferably one based on a method of pre-tabular modification other than record swapping.

Geography

126. The counts of residents and households with residents generated during downstream processing were used to create OAs (see paragraph 74). OA geography products were produced and as soon as they were ready issued to users (October 2002) which was before the statistical output was ready. This advance supply of geographic information helped users familiarise themselves with OA geography and prepare their systems for statistical output. One of the key factors enabling GROS to supply geographic information relatively early was that we did not amend the geography database for the Census after it was frozen in December 2000.

127. Paradoxically perhaps, the number of residents and households with residents in each postcode or OA, although available, was not included in the OA geography products until the first detailed Census output was available in March 2003. This withholding of very basic Census counts was done to ensure that the releases in March were not pre-empted in any way.

128. There was unfortunately a mix in the way areas were represented on reference maps supplied with output as part of supporting information. An area could be shown either as it exists regardless of the

Census and before it was presented statistically as an aggregation of OAs (generally an approximation), or it could be shown as that aggregation of OAs. For example, civil parishes were shown in their 'pure' non-OA form. Localities, having been created from postcodes were shown as aggregations of postcodes. These depictions of areas were made despite the fact that statistics for civil parishes and localities were for aggregates of OAs. Alternatively, maps of wards were presented as groups of OAs. For some wards, not all the OAs belonging to the ward were a contiguous set and the ward had several detached OAs. It is recommended that a common practice be adopted. Because a depiction via OA can cause presentational problems (e.g. detached OAs for wards, or, in the case of localities, extra rural land that in population terms means little), it is recommended that 'pure' non-Census boundaries be adopted as the standard way for showing boundaries in supporting information.

129. Two types of area did not, as originally conceived, cover the whole of Scotland. This meant that these types of area had to be handled slightly differently from the rest. Eventually we had to produce output for the residual parts of Scotland anyway, and it would have been as well to have included these areas from the outset. It is recommended that, for Census processing, the area type 'settlements and localities' include an area 'rest of Scotland', and the area type 'inhabited islands' include the area 'mainland Scotland'.

130. Altogether some 15 area types were used the version of Area Statistics released on the SCROL website and CDs. These were supplemented for particular customers by area types such as those relating to local enterprise companies, ecclesiastical parishes, sheriff court districts, and the 2005 version of Westminster Parliamentary Constituencies (though the latter are to replace the 1997 version on the SCROL website. New area types would either be permanently incorporated into the output database or into 'recodes' that would be used to group areas on the database into those required for a specific table. Recodes could be a temporary measure pending the permanent addition to the database geography if a permanent addition is decided on.

131. There were a few adjustments made to our original set of postcodes. The original assignment of an OA to higher areas was made using an algorithm based on the number of households with residents in each postcode in the OA. Basing this process on the number of households rather than the number of residents resulted in one OA in Edinburgh containing a hall of residence being assigned to a ward that didn't contain the hall of residence. This in turn gave a misleading result for the population in each ward. Had we used the number of residents in the algorithm instead of the number of households, we would have got this right first time. We spotted the anomaly after we had produced the first batch of output. These results were re-run with a patch to correct the assignment of OAs to wards. Subsequent output was run with a

more permanent correction to the database. However, some output (the 'bulk supply') remained uncorrected for some time. It is recommended that the assignment of OAs to higher areas be done on the basis of population rather than households.

132. Another error had an OA misallocated between the mainland and the island of Bute. There was also an incorrect grouping of an area south of the River Forth with the settlement of Alloa. Both of these errors were corrected.

133. The provision of data on hectares (in Key Statistics Table KS01 and Univariate Table UV02) had to resolve a discrepancy between the figures available from Ordnance Survey for Council areas and those generated by GROS ultimately from the digitised boundaries of postcodes. Rather than attempt to reconcile the two by an elaborate pro rating exercise that would have had to take account of differing treatments of inland water and mean high water mark, it was decided to use both sets. The OS figures were used for Scotland, council areas, and a consistent set of figures was created for health board areas. GROS postcode-based figures were used for all other area types. A small error was made in the GROS calculations, which although put right quickly enough for most products, (as for the hall of residence adjustment, see above, paragraph 131) lingered in one product for longer than it should have. It is recommended that an inventory is kept of all products dependent on each 'upstream' product so that amendments are carried through comprehensively.

134. Some 2001 output (chiefly Key Statistics) contained figures from the 1991 Census. To produce these figures, area types to be used for the 2001 Census had to be added to the 1991 database. This was done by assigning each 1991 OA to each type of higher area by 'point in polygon'. The point for the 1991 OA was its centroid, and the polygon for a 2001 area was its 'pure' boundary (see paragraph 128 above). This link was not made for the 2001 OA as there were over 42, 000 of these but only 38,000 1991 OAs. A link made in this way would have meant over 4,000 2001 OAs appearing to have no 1991 data.

135. The figures from 1991 in the Key Statistics were either numbers that had been released unmodified in output from 1991, or were rates in the form of percentage change between the two Censuses – rates from which it would be hard to deduce the counts on which they were based. These count, if revealed in un-modified form, might have been a disclosure risk. Not including 1991 figures for 2001 OA also helped reduce the risk of disclosure.

136. Another piece of geographical preparation was to associate with an OA each 'offshore departure point' given in answer to the question on travel destination. This question had an answer category 'work on offshore installation, please … write in where you travel offshore from'. About 10 different points on the Scottish coastline where given in answer to this and each was associated with the OA containing the point.

137. One surprise was that there still seemed to be a demand for figures for the health board areas that had been shown to be wrongly defined (see above, paragraph 66). The areas were still wanted because much data had been produced for them by various agencies. It was decided to resist this pressure.

Databases

138. Before full-scale production started, data for one of the 8 Scottish EAs was extracted, copied to a server at Ladywell House, and loaded into a SuperCROSS database for table developers to start work on the *Area Statistics* tables.

139. A database for the whole of Scotland was created twice – once without DVs for migration and travel and later with these DVs. Once all the input databases had passed, first, the point where all derived variables (DVs) apart from those to do with migration and travel and, second, the point where all DVs where available, a file was extracted from the database for each EA and deposited on a sever at ONS Titchfield. The 8 files for Scotland were copied to a server at Ladywell House, combined, and loaded into SuperCROSS. After various checks on completeness and accuracy were done, it was then made available to table developers.

140. ONS Population Estimates Unit (PEU) commissioned their colleagues in ONS Census to produce tables on migration including from England and Wales to the rest of the UK. ONS Census decided to construct a UK database with a limited range of variables. Strictly speaking they should have sought the prior agreement of the other 2 Offices. However when the development came to light it was realised that the same UK database would provide tables required by GROS staff working on population estimates and accordingly no great issue was made of the matter. At first the tables for GROS were commissioned from ONS and several were delivered. But because of a few teething problems in commissioning these tables it was decided that GROS would obtain a copy of the UK database and produce tables ourselves. After some difficulties in transferring such a large file from Titchfield, the database was eventually installed at GROS and was used to produce the remaining commissioned output. Having this database also later proved useful in checking tables on migration in the Area Statistics. The use of SCA was an option for the database and when it was used to produce output on migration in the form of SuperCUBES two CUBES had to be created. One had no SCA and counted people enumerated in Scotland, the other had SCA and counted people enumerated in the rest of the UK.

141. Despite the fact that the UK migration database had proved of use, no other UK database was set up. The problems presented by SCA (fragmentation of tables on migration and travel) would have been much the same in any case. Also each Office could ensure that the respective

policies on using SCA or not using SCA would be followed. Each office has a good reason for ensuring that Scottish data is never rounded:

- If there were some output unrounded from GROS and the same output rounded from ONS, then we would get precisely the sort of confusion we want to avoid in our decision not to round (inconsistent tables, etc).

- ONS want to reveal as little as possible about the mechanics of rounding. The cleverer users could compare output from the two sources and work out how rounding operates to an extent beyond that which ONS would like.

142. A possibility might be to create a full UK database now for each Office to use for commissioned tables. Rules may be devised to ensure that each Office produces sub-country tables for its own country only. The other Offices would produce Scotland output at the Scotland level only. It might be thought that the number of small cells would be small, and hence the effect of applying SCA would be small. However, investigations for the SARs (into the number of cells in Scotland level table that would contain a count of 1) showed that around 5 and 3 per cent of cells would contain 1s and 2s respectively that would disappear under SCA. There would be too many of them to claim that applying SCA to that level of output was a trivial departure from current GROS policy on SCA.

143. An alternative might be to supplement each country's database with records from the other two countries for migrants and travellers. Each country might accept the 'wrong' SCA policy for these selected records. It is recommended that this possibility of extended country databases be considered.

Table design

144. The main work in table development however was to produce the wide range of tables in the *Area Statistics*. While under development several (particularly those originally common to the whole of the UK) underwent changes. These changes were initiated by formal *requests for change* (RFCs). Nevertheless, keeping track of RFCs became pretty much a full-time task for one of the team. The work was compounded by the fact that each table was held in various ways including: an Excel version that would, in due course, be filled with the results for Scotland and included in the *Reference Volume*; a development version in the main SuperCROSS format; a production version in another SuperCROSS format; a pdf version for inclusion in the consultation material on the GROS website; and an html form for the SCROL website. RFCs that were approved had to be recorded on all of these versions of a table. It is recommended that some way is found to minimise the number of versions of each table during the table development phase.

145. A particular problem with a changing base of tables was keeping track of table numbers, keeping them and all related information up to date. Also GROS made life more difficult for itself by not move to a zero-filled 3-digit number for tables S01 to S67. This meant that in any list of tables sorted automatically, tables S201 to S209 (for example) appeared between S20 and S21.

146. SuperCROSS was a good choice of tabulation software. New staff were able to pick up its essentials pretty quickly. With a core within the team who could tackle the more complex tables, we were able to develop the full range of tables in time to meet the publication deadlines. A rehearsal of table development would have speeded things up further and given us more time for checking tables and reducing the (small) number that were released with errors or to explore the data that had become available in an exploratory way that may have led to some unexpected but useful results. If no rehearsal data were available, an empty database might have enabled us to make a good start on table development. The table layouts produced from an empty SuperCROSS database might have replaced some of the table versions listed in the previous paragraph. Further, the tables may have been designed more in keeping with the constraints of the software (e.g. on the formatting of table headings). It is recommended that, since printed tables will form a tiny minority of the full output, table design needn't depart too far from whatever default options are available in the tabulation software.

147. Another source of difficulty in keeping track was the extensive use of footnotes that were to appear in versions of tables appearing in reports and on the SCROL website. The problem was solved – not too satisfactorily – by managing these footnotes in a separate database.

Production (Area Statistics)

148. The business of turning files extracted from 8 databases into the output database has been described briefly above (see paragraph 139). The details of the operation included obtaining 'geofiles' from the UK database at ONS Titchfield and attaching these to the SuperCROSS database so that a person's OA of residence, OA of place of work or study and OA of address one year ago were all assigned to the various geographical hierarchies. The second extract (including data on migration and travel) had to have the item on distance travelled corrected. (This error stemmed from a faulty DFA plied in downstream processing (see paragraph 89). The data had to be amended on the extract file rather than on the input databases via DFA. Difficulties arose because the person ID (ED-based) in analyses from the DQMS was different from the OA-based ID on the extract. To match an output person record with its corresponding input record required the postcode from the household record. It is recommended that, despite concerns over confidentiality, the input ID is retained on the output database. The link to input could still be made (albeit clumsily) without the input ID – and therefore not retaining it was ineffective.

149. Tables were developed and tested in SuperCROSS. When the database containing data for 1 EA was available, tables were checked against tables that had already been checked and against figures from the DQMS for the EA. Tables considered ready for production were converted to production format and assembled in batches for the *Table Production Module* (TPM). The TPM was a set of programs written by GROS to run batches of tables for a given set of area. The first stage of the TPM combined the tables in production format to the geography to create a set of table programs to be run in the second stage. The TPM served us well. Its only drawback turned out not to be critical and that was that it could only be used by one user at once. This meant that, in effect, the team divided into those on development who fed checked programs to the person who had become expert in running the TPM.

150. There were a few problems early on not having enough computer capacity. We hadn't fully anticipated the scale of exercise (which we may have been done if we had had a rehearsal). The problem was fixed by the acquisition of more IT storage.

151. The common UK database design had placed some variables (including geographical variables) in hierarchies, where the users may choose the level of the hierarchical appropriate to the table being developed. While convenient for general use, hierarchical variables were awkward when used in deriving new variables. It was convenient sometimes to have a 'flat' alternative to some hierarchical variables.

152. Generally, a table on Migration and Travel for an area was produced in 4 portions. Those for migration were:

- Non-migrants in area plus migrants into area (data held in records for residents in area)
- Migrants out of area living elsewhere in Scotland (data held elsewhere in Scottish database)
- Migrants out of area living in England and Wales (data held in England and Wales database)
- Migrants out of area living in Northern Ireland (data held in NI database)

Fragments of tables as above were developed and tested by GROS. Portions of the first two types were run against the Scottish database. The other two were sent to ONS and NISRA respectively and the results sent back. The procedure for tables run by ONS (and conversely the equivalents run by GROS for England and Wales migration tables) was to deposit programs and results in a reserved part of a server at Titchfield. The results were transferred and checked and then assembled into the final table. Some portions of tables from NISRA had to be converted from tables with areas represented as names to tables with areas represented as codes (see paragraph 159) – an extra job at a

busy time that was needed to ensure that the small cell adjustment in each version was identical.

153. A similar procedure was followed to generate tables on travel. Here the cross-border flows were different in that those out of Scotland included people travelling to a place of study with those travelling to a workplace. Travellers into Scotland were travelling to work only. Offshore departure points were treated as destinations and required special treatment.

154. Other tables that were developed in portions were those that contained figures on hectares (which were contained in a separate SuperCROSS database), figures from 1991 (which were in a database for that Census), and figures for 2001 counting a different entity from the bulk of the table (e.g. dwellings in a household table). The portions had to be joined, a process that worked only if the portions being joined matched exactly in a number of key respects. Getting portions to match ready for joining was a job requiring great attention to small details.

155. For various reasons there were times when there was a slackening off of activity. Chief of these was the time when most of the tables not involving migration and travel had been developed (if not run) and we were waiting for the database with the second extract of data to appear. In these times we turned our attention to producing commissioned output – requests for which had been building up. Also, ONS had started to send us production versions of tables for the Scotland portions of tables for Eurostat, and for tables for the 'Focus on …' reports (see paragraph 108).

Armed Forces tables

156. One part of output that, in the event, we abandoned was the set of tables on the armed forces. There were a number of problems. First, the tables would have been counting records with some doubts about the quality of the information on occupation and industry used to identify members of the armed forces. Special instructions were issued to army camps, naval bases, etc saying that members of the armed forces should complete the questions on occupation and industry in a certain manner that would make their identification pretty certain. For example an army cook would not say he or she was a cook but state (e.g.) 'army, NCO'. It appears that the special instructions did not filter to the people completing the forms because many army cooks wrote 'cook'. They were thus assigned to an occupation and industry to do with cooking rather than the armed forces. It is recommended that, if a count of the armed forces is required, appropriate instructions should be on the form itself – as in 1991. An attempt to correct this defect in enumeration was made in two areas where it was believed that faulty enumeration of armed forces had occurred with non-enumeration. Records were added to the input database during downstream processing (see paragraph 67).

157. The second problem was how to present tables on armed forces (AF). Before we abandoned that tables (which were to count AF by area of residence and workplace), we had got as far as identifying wards in which least 50 members of the armed forces lived and those (not the same) in which at least 50 worked. We believed that such a threshold was in keeping with the thresholds that we had adopted for other output. We hadn't decided how to deal with AF personnel who lived or worked outside these wards.

158. NISRA, for perhaps well known reasons, had no intention of releasing AF tables. ONS decided they would take an extract of records for people they suspected were members of the armed forces, investigate these cases by inspecting images of forms, etc, and make corrections to occupation and industry on the extracted database. It would then be used to create AF tables. But it would be inconsistent with the database used for the main output for England and Wales.

Formats

159. As we moved towards delivering output, we identified a number of formats for the supply of tables to users. These included

- The SuperCROSS development format, which could be read by using with SuperTABLE – available from the suppliers of SuperCROSS.

- Comma separated value with geographical areas represented by their names. This format was aimed at users who would load the tables into software such as Microsoft Excel and expect to see 'Angus' against the tables for the council area of that name. These users were those preferring to use software familiar to them rather than SuperTABLE.

- Comma separated value with geographical areas represented by their codes. Tables in this format were supplied to those working on the SCROL website. Once available in the development SCROL website, they tables were checked to ensure they had been loaded correctly before being moved to the production website. A version of tables in this format was also supplied to the heavy users as 'bulk supply' - but with the output chopped up into separate batches for each council area (more below, paragraph 160).

- SuperCUBE format. This format could be read by SuperTABLE in a more flexible way than the SuperCROSS development format and, in particular, ways in which, say, areas age grouped into areas defined by users can be transferred from one SuperCUBE to another. However, SuperCUBES could not be created for tables that had been joined from portions or contained items derived from other such as rates and percentages. Also the user had to be familiar with SuperTABLE. In the event, apart from the SuperCUBES disseminated (on the GROS website) containing migration data no tables were produced as SuperCUBES (see

paragraph 140).

160. A minor quirk of format was that one-dimensional tables such as the Key Statistics and Univariate *Tables* were presented vertically in some products and horizontally in others (yet another format). Another was that areas were stored in numerical order of code in the SuperCROSS database which was not always the order for publication. Codes for (say) council areas usually matched an obsolete alphabetic order of areas before names were amended. It is recommended that SuperCROSS is loaded with areas in publication order as default.

161. The format GROS chose for bulk supply was different from that offered by ONS. The difference was partly due to ONS being able to take advantage of a later version of SuperCROSS than GROS was using to develop an add-in facility to create bulk supply. GROS decided that we did not have the resources to write a program to re-format the output we got from our version of SuperCROSS. UK customers found this a nuisance. Some users bought their bulk supply direct from us but some chose to get theirs from the Greater London Authority (GLA) converted into SASPAC format. These latter users were in all likelihood intending to use the SASPAC software produced by GLA for analysing Census output. The bulk supply dataset for Scotland cost £3,200 from GROS, but from GLA converted to SASPAC format it cost £1,500. The format of SASPAC system files was more compact than that of the 'raw' data GROS supplied, and considerably more so than the SuperTABLE format. Formats supplied by GROS were intended to be used by unsophisticated users that had only software generally available at their disposal or software that came with the tables. It is recommended that more be done to accommodate the wishes of users about the format of bulk supply.

162. There were a few other demands created in meeting user demand for bulk supply. A general idea was that, with no SCA, tables need only be produced at the lowest area level and the heavy user (for whom bulk supply was designed) would be able to generate tables for all other area levels by aggregation. There were three types of table where this general rule did not apply. The first was *Key Statistics* that contained various rates and percentages that did not add up across areas. The second type was the set of tables on migration and travel where moves into and out of areas become moves within areas as areas are aggregated. The third type is the set of tables (on workplace) that have had SCA applied. The solution is to produce 'numbers' versions of Key Statistics that the user can aggregate and then re-calculate the required rates and percentages, and, for migration and travel – and workplace, is to supply tables for every area level rather than just at the lowest..

163. As tables were being re-supplied in various products we took the opportunity to provide new versions of 4 KS tables that included an average of one kind or another (average age, number of rooms, hours worked, household size). These had originally been calculated by including separately each value of the variable concerned, calculating

the average, and then hiding the separate values. Unfortunately, SuperTABLE could be used to reveal the separate values which for some area levels would be highly detailed and carry some risk of disclosure. In the revised versions of the tables the average was calculated instead by including the total of the variable and dividing by the number of instances and hiding these contributory items. It is recommended that averages are not calculated by deriving them from separate values of the variable being averaged.

164. A major challenge for the table development team was to keep track of all the tables in their various formats, area levels, and delivery routes. Altogether there were over 11 thousand tables taking account of these factors in combination. Other things to bear in mind included the fact that the SCROL website was not to offer tables on religion and ethnicity but did include the 5 'profile' tables. It is recommended that once the overall scheme of tables is known, a robust folder structure is agreed for storing them and all changes to that structure are agreed again by those using it.

Table acceptance

165. ONS had produced a range of 'integrity counts' from the input database for output staff to check totals appearing in tables. It was soon discovered that some of these counts were wrong. GROS replaced and extended these with counts generated from the DQMS. It was gratifying that the first counts from the SuperCROSS database on age and sex by council area were identical to those produced for the *Population Report* from the DQMS.

166. As mentioned earlier (paragraph 138), until the full database was available for the whole of Scotland, developers had a SuperCROSS database with data from one EA. This generally worked well except that some quirks in the data that required changes to table design were not contained in the data for the chosen EA. Some developers carried on with their work with the single-EA database because they were familiar with the counts it produced in tables. It is recommended instead that part-databases are abandoned as soon as something better comes along.

167. Migration and Travel tables presented a particular challenge for table checking. Fortunately we had the UK database on migration (see paragraph 140) which was extensively used to check migration tables in the *Area Statistics*. For some migration tables and for all travel tables we had no way of checking portions of tables produced by ONS and NISRA without their providing some simple output for the purpose. Using this output was a little complex because, with SCA, figures that should agree didn't always agree even when the table was actually right.

168. When the second extract with Migration and Travel was loaded into SuperCROSS, output from the new database was produced using only variables that were in the first database (without Migration and Travel) to

check consistency.

169. The general approach to checking a tables was

- Check that the figures for Scotland were OK i.e. to that totals agreed with figures in tables previously accepted or with figures from the DQMS.

- The rows and columns should add up, etc. With a tabulation package such as SuperCROSS that was generally a formality. One had to ensure that the categories of classification used where complete covering the whole population. Tables in general were design to meet this simple criterion. (This was not always the case in 1991, where smaller categories were sometimes omitted but still retained in totals. This caused confusion sometimes with users – and with Census staff.) An example from 2001 of including all cases in a classification that of 'all-minor households' (household with no adults). Such households didn't have an explicit slot in a household classification that assumed there were none. So one person 'all-minor households' were included with one non-pensioner person households and others (with two or more children) were included with 'one adult plus children'.

- Check that the totals for each area in the table agree with previously agreed figures. These can come from the geography database containing data on Output Areas. All areas in output were aggregates of OAs.

- Figures for variables that were recoded (categories combined) were checked against the original classifications.

- Variables derived in SuperCROSS were checked by cross-tabulating the contributory variables against each other. Categories of the derived variable corresponded to combinations of cells in the cross-tabulation and could be checked accordingly.

Data problems

170. From comparison with other sources, it became clear that form-fillers on housing benefit had been liable to tick *living rent-free* in response to question H8 on the Census form. They should have selected a rented category instead as rent was being paid on their behalf. It is recommended that the question on owning or renting be adapted accordingly. This realisation came too late to inform the way in which household tenure was grouped in output tables. For example, in many tables, 'living rent free' had unhelpfully been combined with the private rented group. Re-designing and re-producing the affected tables would have (a) risked disclosure because of differencing between slight different classifications and (b) taken more resources than we had at the time. So, it was decided to issue just 4 additional tables that would help users adjust the results of the main tables with living rent-free wrongly

grouped.

171. Data problems discovered while producing tables include

- 4 year old lone parent. Most output masked this by drawing up classifications to absorb this instance, but one table supplied to Eurostat tabulated age (detailed below 16) by person type (including whether or not a lone parent) so the case was revealed.

- 3 lone parents with no children.

- Two variables for families generated by the household composition algorithm were inconsistent. These variables (like others from the HCA) were not available in the DQMS. It is recommended that all variables generated in input processing are available to the DQMS.

- A derived variable on living arrangements failed to classify people in some households because when marital status was inconsistent with relationship. A DFA corrected the derived and contributory variables.

- There are some households in temporary accommodation not on the ground floor

- There is one person classified in one variable as ‘not working or studying’ and in another as a full-time student.

- In the variable of migration origin, there are records for people who migrated from inside the UK who were given a country code for one of the 4 UK countries for their origin of move rather than a UK postcode. The data capture contractor should instead have set the origin to missing; a UK postcode would then most likely have been imputed. Instead the count of migrants from outside UK wrongly includes several hundred persons. It is recommended that variables combining postcode and country exclude categories for the 4 UK countries.

- It appears that derived variables based on dwelling were created before the DFA was run to extend the output of the dwellings algorithm. These DVs are therefore wrong, and the DVs used for output were created in SuperCROSS instead.

- The DV to assign a highest qualification to each person aged 16-74 is out of line with the combination of qualifications from with it has presumably been derived. The HQ DV was used in output before this discrepancy was realised. A means of reverse-engineering individual qualifications to agree with the derived HQ has been recorded.

172. Some tables were produced with faults. Some totals were double the right number because sub-totals had been included as well as the basic counts. Some geographical variables did not include all areas.

These faults were posted on the GROS website and replacement tables produced.

Supporting information

173. Users were kept informed of progress with the release of products in a number of ways, chiefly by means of Census Updates on the main GROS website. The SCROL website also contains a News section. The Census Offices collaborated, with ONS in the lead, in the production of the publication *Census 2001, Definitions* which is also available on the web via Census Update no. 25. The Main website also provided a list of 'known errors' with information about how each had been dealt with.

174. Work on the *Definitions* volume unfortunately took lower priority than production of tables and hence was produced far later than originally intended. Some of the material in the volume had been released in one form or another – e.g. as supporting information on the SCROL website. However, it is recommended that ways be found systematically to put the material in the volume on the web as it becomes available rather than wait till it is all ready for publication in a single product. In particular, a table index in the form of an Excel spreadsheet should have been available at the outset of delivery of tables.

175. The Census Updates now need to be reviewed:

- To remove out-of-date material
- To provide links to key documents provided via the updates in a more direct way. For example, the *Definitions* volume should be referred to from higher up in the Census hierarchy of pages on the site.

It is recommended that these changes be made to the GROS website i.e. the material in the Updates be absorbed into the general structure for the 2001 Census. In future when a new Update is added, material in the previous update should be absorbed into the general structure.

176. Ancillary information on other websites is missing from the supporting information provided to users. It is recommended that links be provided to, e.g., the pages on the ONS website that explain the National Statistics – Socio-economic Classification.

Delivery

177. Six printed reports have been laid before Parliament:

- 2001 Population Report
- The Registrar General's 2001 Census Report to the Scottish Parliament

- Key Statistics for Council areas and Health Board areas Scotland

- Key Statistics for Settlements and Localities Scotland. The map section of the report contained some pages on which the boundaries of localities within settlements had been omitted. A replacement booklet containing the correct maps was printed.

- Scotland's Census 2001 Reference Volume Scotland. When printed, tables on migration and travel were not available and appear with blank cells. The main purpose of the volume was to provide the final version of table layouts; figures for Scotland would be provided if available. The version on the GROS website now contains figures in all tables. The scope of the volume is the full set of *Area Statistics* including the 5 'profiles' but excluding the Key Statistics as these appear in other volumes.

178. Four Occasional Papers have been published.

- Scotland's Census 2001 - Statistics For Inhabited Islands

- Scotland's Census 2001 - Statistics on Migration

- Scotland's Census 2001 – Statistics on Travel to Work or Study

- Scotland's Census 2001 – Gaelic Report

It is recommended that more Occasional Papers be produced – eg on relationships between migrants and non-migrants within households. It is also recommended that links to these Occasional Papers be placed in the 2001 Census part of the main website.

179. GROS contributed to the UK report on Westminster Parliamentary Constituencies that ONS published and laid before the Westminster Parliament. The areas for Scotland were those in force at the time of publication i.e. the 72 areas used for the 1997 and 2001 elections. Supplementary tables were sent on CD to the House of Commons Library for the 59 areas that were used in the 2005 election. Data for these areas are available from the GROS website (Standard Tables for Westminster Parliamentary Constituencies (2005))

180. Tables were added to the SCROL website as a primary means of dissemination. The main facilities available are the *browser, analyser, warehouse,* and *thematic map facility*. The site was funded as a means of providing the full range of *Area Statistics*. This original proposal was scaled down before the site was launched so that tables on ethnicity and religion were excluded (although available via other means, see below paragraph 187). It was also decided not to include the *origin-destination statistics* which was considered was a little specialised for the target audience of the SCROL website and a service would instead be provided by third parties (see below, paragraph 191). Third parties were also to offer a service to disseminate *Area Statistics* but these would

also be either limited to an academic audience or to those who paid the required subscription (see below, paragraph 190). It is recommended that the range of means of disseminating Census results by web be reviewed to see if some rationalisation of publicly funded services is possible.

181. SCROL Browser is an easy-to-use means of getting simple data presented as 5 'profiles' for a wide range of areas. The user can select a profile for a single area or for two areas for comparison. The Analyser delivers other tables in the *Area Statistics* with important exceptions (see paragraph 122). The approach of Analyser is to get the user first to specify which table he or she wants, then which areas within a chosen area type, and then view or download the selected tables. The warehouse delivers all the available tables for a selected area or areas. The thematic map facility is either launched from the main menu or becomes available if the user has selected a Key Statistics table from the Analyser. The user may select or calculate a quantity to be mapped for selected areas and may 'drill down' geographically by mapping areas within one of those already mapped.

182. Following publication of KS for Settlements and Localities, the Commission for Racial Equality (CRE) expressed concern that the settlement of Garelochhead contained the same percentage of Moslems as it did of Pakistanis and Bangladeshis. Although, as it happened, it would have been an incorrect inference, the CRE claimed that casual reader would conclude that all Pakistanis and Bangladeshis in the settlement were Moslems. Note that, at the time, the US were preparing to invade Afghanistan. Note also that SCA would have made no difference to the tables involved, with no change to the risk of an incorrect inference. Nevertheless, it was decided *not* to put similar and more detailed tables on ethnicity on the SCROL website as these topics were sensitive. The withheld tables would still be available on the SCROL CDs and DVD and on request. This created another complication for keeping track of tables and the versions that were to be supplied via each format.

183. The withdrawal of tables on ethnicity and religion from the SCROL website also means that the user is not able to produce thematic maps on those topics directly from a GROS source. It is recommended that some means of providing such maps be provided.

184. SCROL provides tables from the Area Statistics. To find a way of providing Origin-destination statistics on the web, discussions took place with the developers of a web-based system (WICID, see paragraph 192) for delivering origin-destination statistics to academic users. The proposal examined was that a copy of the system is provided by GROS for the general user. These discussions became stuck on funding and technical problems. The issue of funding is that widening the range of users of WICID might jeopardise the funding currently received from academic funding bodies such as the Economic and Social Research Council (ESRC).

185. Output from the 2001 Census has also appeared on the website Scottish Neighbourhood Statistics – usually in the form of percentages or univariate distributions.

186. Seven sets of SCROL CDs have been produced containing some 20 discs. Around half of the discs have now been re-issued because of errors affecting one or more tables on those originally supplied. It is recommended that some notification of which discs have been replaced be put on the GROS website. Despite efforts to ensure that users had an error-free set of discs, the replacement version of CD7 went into production with an error and had to be released with a correction CD. CDs 1 to 3 contain tables in both csv and SuperTABLE format, but the other discs contain SuperTABLE format only. This was partly to save space (CD4 runs to 8 discs) and partly because users could use generate csv versions of SuperTABLE files using software on the CD.

187. All of the material on the SCROL CDs – which includes all tables on ethnicity and religion - has been consolidated into a two-disc DVD set containing a full set of the *Area Statistics* in csv format on one disc and a full set in SuperTABLE format on the second. As the DVD will become, in time, the 'archive' version of the *Area Statistics*, care has been taken to ensure that the tables on it are the final correct versions. However, this checking has been a challenge because of the myriad of tables involved with and without various restrictions, thresholds, formats, area levels, etc.

188. It had been decided at the outset that the SCROL CDs and DVD would include SuperTABLE software. GROS held 6 or so well-attended classes in how to use the software. However, one major stumbling block was that the Scottish Executive could not use SuperTABLE without paying a fee to the third-party who provided their IT network. It is recommended that a review be carried out of the extent to which users will be able to use software such as SuperTABLE in future. It is also recommended that it should be checked whether the inclusion of GROS within the SE network system has overcome the barrier to the use of SuperTABLE by the SE.

189. One requirement expressed by users that we have not been able to meet, on the SCROL website or with SuperTABLE on the SCROL CDs, is that of being able to re-use a grouping of, say, output areas into school catchment areas. There was no perfect solution but possibilities include:

- Using SuperTABLE to re-arrange a table so that the output area was the only row variable and all the other variables were combined into columns; exporting the table to a database that also had a table linking OAs to the target areas; then using the database to aggregate to the target areas.

- Asking GROS to use an OA-target area link as a 'recode' for the SuperCROSS database. Such recodes exist for local enterprise

companies, sheriff court districts, etc. This solution would be difficult for OA-classifications of 'remote' geography (i.e. migration origin and travel destination).

- Asking GROS to produce the required tables in the SuperCUBE format, where the user can use SuperTABLE to create a 'recode' for one SuperCUBE which can then be saved and used for other SuperCUBES.

It is recommended that the SCROL website be amended so that users can use saved recodes of area classifications.

190. The final version of the 'bulk supply' was prepared for release including tables on migration and travel and corrected versions of tables. Tables are chopped up into batches one for each council area. The product was aimed at users who wished to load the Census results into software of their own choosing (e.g. a GIS) or at those who developed third-party software for handling the results. Examples of such third-party software include

- SASPAC
- CASWEB

Except those counting workers by area of work, tables were not modified after tabulation. Tables (UV, CAS, CAST) for OAs could be used to generate output for any areas built up from OAs. Hence tables only for these OAs were supplied in this product. A similar approach was adopted for ST and Theme tables where tables for only the bottom level in any hierarchy were included in the product.

191. Work is still being considered to supply various products ancillary to the bulk supply (see paragraph 162).

192. As for *Area Statistics* there are a couple of routes to the *Origin-Destination Statistics* provided by third parties. SASPAC contains modules for analysing the product. Academic users have a service, *Web-based Interface to Census Interaction Data* (WICID) provided by the University of Leeds with ESRC funding.

Possible future developments

193. Despite differences on the use of small cell adjustment (SCA), the Census Offices have been discussing the possibility of setting up a combined UK database. One of the difficulties will be to ensure that no output from any country has the treatment for SCA (applied or not applied) contrary to the policy for that country (see paragraph 112).

194. Time inevitably brings changes to existing geographies and new geographies altogether. A recent example is that of the Westminster Parliamentary Constituencies. Ideally new or changed areas should be

incorporated into the SuperCROSS database, and *Area Statistics* produced and added to the SCROL website. It is recommended that these processes be reviewed so that they can easily be used as required.

195. Smaller scale changes in the geographical landscape have their effect even without changes in the boundaries of the areas for which output is produced. For example, the SCROL user may enter a postcode in order to find out which ward or civil parish he or she wants statistics for. New postcodes are introduced continually and the user may enter one that SCROL doesn't recognise as it uses the set of postcodes frozen for Census enumeration and processing. A decision still has to be taken as to whether SCROL should continue to use the postcodes frozen for Census, current postcodes, or both somehow. It is recommended that an index linking current postcodes to 2001 Census areas be added to the SCROL website. Perhaps the index should include all postcodes that have existed since the 2001 Census.

Data quality and the data quality management system

General

196. There were a variety of means of assessing the quality of the data collected in the Census before, during and after it was processed and turned into output. For example, Census Quality Surveys were conducted in conjunction with the Census Rehearsal in 1999. A data quality management system (DQMS) was used during and after processing to quantify the amount of edit and imputation and to produce data that could be compared with other sources. Other activities under the heading of assessing data quality were the follow-up survey of vacant property – whose main purpose was to provide information on vacant accommodation not collected in the main Census.

The data quality management system and data quality strategy

197. The DQMS provided analyses of versions of the data for each EA as the data went through the various stages of downstream processing. The copy of the database for the EA taken at each key stage could be loaded into a 'bucket' that could be analysed using an SQL-generator called EasyAsk that placed the results of a query into a spreadsheet. Also accessible via EasyAsk were the 'tick and text' files supplied by the data capture contractor, and 'audit' files that recorded the types of changes made to each data item by each process of downstream processing.

198. Any analysis would generally take the form of producing results from each of 8 'buckets' and putting them together to produce results for Scotland. It is recommended that any similar facility uses all of the data for Scotland in a single database. Apart from simplifying the task of

checking the quality of any Census item, the larger the area used the better for checks on migration and travel. EasyAsk had certain limitations e.g. one could not easily derive combined categories of variables, create complicated variables (which usually had to be constructed with extracted data in other software), nor do complex (e.g. 3-way) tables - other than by concatenating variables before tabulation. One often wished one had the software (and data) that was used for output. It is recommended that the same package that is used for output is also used for a DQMS. There would only be one package to learn, one would have a full range of tabulation tools, and a full range of data records including that for family. DVs such as those relating to household composition would be more easily checked against other sources.

199. The original approach to data quality using the DQMS was to automate the process of checking quality as far as possible with as much analysis as possible programmed in advance. Queries producing univariate distributions and cross-tabulations would be pre-programmed and the results compared against data from other sources. Discrepancies above given tolerances would be highlighted for investigation. The data from other sources would be gathered in advance and placed in a database for the purpose. As required, the DQ team could add their own tabulations to the pre-prepared set.

200. The three key processing stages at which data was investigated were after load, after EDIS, and after postcode imputation (by which stage the data would have also been augmented by addition ONC records and have been swapped for disclosure control). The results for each of these stages could be compared with those for earlier stages, though the comparison had to take account of the fact that record swapping swapped only the data records and not the audit or tick and text records.

201. Unexpected values in the data could be investigated by examining the images of forms. ONS wrote software to aid the process of selecting and examining the images of those forms selected for investigation. One could specify a single page of a form for examination or submit a file of page or form references generated from an EasyAsk query. The software, called ‘Imageviewer’, worked very satisfactorily (apart from the lengthy time it took to start).

202. At first only those images were available on-line for EAs that were thought to be needed for examination at a given time. At soon became clear that images for all 8 EAs were needed at any time. The other Census Offices also found difficulties juggling sets of images on and off servers. Before processing was complete, ONS decided to acquire sufficient on-line capacity to have images for all 112 EAs available at once. It is recommended that all images are made available as soon as they are received from data capture and that they remain available while downstream processing and work on data quality is in progress.

203. The strategy based on pre-planned comparisons had been devised originally by ONS. They reviewed how the strategy was working out after several England and Wales EAs had been checked. They decided to streamline the process for two reasons. First, the outcome of comparisons with the 'secondary data' usually was that the non-Census source was inadequate in some way for the purpose. Data from the 1991 Census was of course 10 years old and data from surveys had non-response bias. So discrepancies tended to be routinely explained away. The second reason for change was that the process was taking too long. The revised approach was partly to group the checks so that action was only required if a batch of checks contained more than a given number of 'failures'. Previously processed 2001 Census results would be the secondary data for comparison. GROS followed suit to some extent in that we built in comparisons of results for any stage with results for earlier stages. (The results of these comparisons are, in effect, now summarised by the information on data quality, see census variables). We received training from ONS in the streamlined techniques of grouping checks. However, we were already concentrating on particular lines of investigation that we felt were the most fruitful (i.e. would lead to finding data that needed changing and then changing it).

204. Other work at GROS included simple investigations of univariate distributions and of the effectiveness of edits, e.g. the soft edits in EDIS (see paragraph 39). Work was divided among the team by groups of related Census topics. The team assembled a series of documents containing comments on findings. These are partly stored in spreadsheets on a server at Ladywell House. The comments have yet to be collated into a single reference document but points requiring investigation were dealt with as they came up. It is not expected that the collated document would contain adverse comments about data quality except where further work was done. It is recommended that these comments be collated.

Data Quality Review Procedure

205. A UK procedure to resolve issues of data quality was set up with issues being logged on a special database at Titchfield. This procedure was generally put to good use so that all could contribute to discussions and track how problems had been dealt with. It is certain that this process was a little too bureaucratic once the pace quickened. Also there was no forum in which to resolve issues not sorted at working level (see paragraph 120).

Comparisons with 1991

206. Several of the tables in the Registrar General's 2001 Census Report to the Scottish Parliament contained comparisons (mostly at the Scotland level) of 1991 and 2001 figures. This gave an opportunity to seek explanations for differences. For example, the number of people

with Higher National Certificate (HNC), equivalent or higher qualifications (qualification groups 3 and 4) almost doubled between 1991 and 2001. Broken down by council area there is a strong correlation between the percentages with qualifications in the 2 Censuses. After consulting colleagues in the Scottish Executive, it was decided to publish the comparison in Table 24 of the Report.

Comparisons with other 2001 sources

207. Comparisons included:

- *Inter-Departmental Business Register* After various problems in getting IDBR data and preparing it for comparison by validating postcodes and so on, we were able to check the Census items on industry and workplace. Data from the Census had to be assembled from EAs into a single database for checks on workplace to be sensible (but, even so, the Census count of workers excluded those enumerated outside Scotland). The following table compares the number of workers in three area types according to the two sources. Most of the figures deal with the absolute differences between the two figures (i.e. disregarding sign). The average difference is dominated by a few sizeable outliers so the median difference is given as well. Once the average number of workers in each area is taken into account the comparison can be seen to be an improving one as the size of area increases. So, there is noise in the data on workplace but it appears to be manageable at the output area where, if the IDBR were taken as the benchmark, there is an error of around 9 percent (using the median absolute difference). It is recommended that fuller results are prepared for public release.

Area type	Average absolute differ-ence	Median absolute differ-ence	Average mismatch: (b)÷ average workers per area	Average mismatch: (c)÷ average workers per area	Correl-ation coefficient
(a)	(b)	(c)	(d)	(e)	(f)
Postcode	15.6	3	1.18	0.23	0.71
OA	20.4	4	0.47	0.09	0.93
Postcode sector	409.8	153	0.21	0.08	0.95

- *Schools data* Once we had data from the full Census we did some

checking against data on school rolls - and this was developed further when carrying out tabulations for HMIE to whom the outcome was satisfactory. Generally the numbers compared well enough for the HMIE commission to be completed satisfactorily.

- *Travel data and students* One of the problems with extending the question on travel to include place of study was discontinuity with the previous Censuses over the travel of students. Data on travel (to work) of students was a problem in 1991 anyway given that they were counted for that Census as resident at their vacation address. Since students were generally numerated during term-time, their 1991 workplace would have been somewhere to which they didn't travel starting from their 'residence' but from their term-time address. So even if we hadn't extended the question, the workplace data wouldn't have been comparable for students 1991-2001. For 2001-2011, extending the question in England and Wales may have to be done in a way that distinguishes a little more directly between types of destination, and that allows 2001 data to be re-tabulated in a comparable way. We may to some extent have 'future-proofed' the 2001 travel statistics by making output for Scotland as comparable as we could with the rest of the UK. This was done by including columns for full-time students in the table designs. See, for example, TV301 (for Scotland) and the equivalent W301 (for the rest of the UK) in the migrants final layout pdf.

- *Scottish Household Survey (1999 and 2000 combined)* Comparisons were made for all variables common to both sources. The Scottish executive also provided comparisons with 2001/2002 Scottish Household Survey.

- **Migration** Between and within council migration was checked against sources used for population estimates.

- **Population** *(rolled forward 2001 Mid-Year Estimates)* Comparisons were an intrinsic part of the QA of the One Number Census.

- **Census Coverage Survey** The CCS also provided some distributions e.g. in the complex area of the relationship matrix for which there were few other sources for comparison.

- **Register of communal establishments** Checks were made of CEs found in the Census against those listed in the register GROS maintained for providing information to enumerators and other field staff. The results are given in Appendix 1. One error emerged after output had been produced and that was that a CE (a hall of residence) had been included in the wrong OA. The reason was that the postcode for the CE had been wrongly captured but was still within the correct enumeration district. However, the postcode (with the CE wrong included) was included in an OA that did not include the correct postcode for the CE. The error did not emerge

during checks against the register because it was common for halls of residence to have wrong postcodes on the register; so slight mismatches weren't always investigated. The user (from Fife Council) who had pointed out the error was happy to have his suspicion confirmed but did not require the data to be re-processed. Further checks on whether the same thing had happened to other CEs revealed about a dozen or so large CEs that had similarly gone astray. It is recommended that a short note describing the error is put on the GROS website. Postcodes of large CEs should be carefully checked before enumeration and during processing.

Other checking

208. As stated above (e.g. at paragraph 204), certain checks outside the pre-planned range were done. The scale of these was fairly wide to begin with, particularly on a sample of data from the first EA. The scale was reduced as it became clearer which checks showed no cause for concern and which were less re-assuring.

209. Data quality work found errors such as those reported under output (see paragraph 171). Some more arcane errors were unearthed such as one in the 'relationships' variables (a 'string' concatenating all the relationships of a person to each other person in the household). The string should contain an X in the nth place for the nth person in the household. For example Person 2 in a 4 person household should have a string such as '1X66'. For some persons the X was not in the expected place, nevertheless the household algorithm seems to have made the correct grouping of individuals into families.

210. Because of interest in devising a new classification of ethnicity that would in time be used for the 2011 Census, a couple of very specific checks were carried out. These were to classify the text written in write-in boxes in the 2001 Census question on ethnicity for persons who were assigned to the output categories of 'Any mixed background' and 'Other ethnic group'. The results are on the GROS website, 2001 Census Ethnicity Reports

211. One check was suggested at the time the Census Offices decided to impute remote postcodes was to see whether students in same household had same address one year ago imputed. Unfortunately there has not been time to do this check. It is recommended that the check be done.

212. A general problem in tracking forms for possible errors or assessment of quality was the lack of a person number on the form for individuals in communal establishments. The processing system assigned each person record a number but having ascertained that number there was no easy way of finding the corresponding image because the images were stored without a person number and not necessarily in the same order as that generated in processing. The only

way of finding a required image was to look at the name on each until the right one came up. It is recommended that individual forms for people in communal establishments are given person numbers at the stage they are enumerated.

Missingness

213. By analysing the proportion of 'missing' values for each variable (after the filter rules had been applied in Load, see paragraph 22), an initial assessment was made of how well each question had been answered. The results were placed on the GROS website (missingness excel file). ONS website has similar results for England and Wales, Question non-response rates. Similar results are seen in the two analyses. Exceptions include

- The better response to the question on marital status in England and Wales, perhaps because enumerators south of the border were instructed to ensure that there was a response to at least the first 4 questions for each person.

- The better response to the questions on 'work last week' in Scotland, perhaps due to the better layout and position of the question on the form.

214. A fuller analysis is now available that, in effect, examines each value of each variable in the output database. The values are those appearing in output tables such as, to take the simplest example, 'male' and 'female' of the variable Sex. The analysis quantifies the extent to which each value in the output database is as collected on a Census form, or was inserted or altered by an edit rule, was imputed, or was on a record added by the One Number Census process. The results are to be found on the GROS website, census variables mentioned above several times.

The Census Quality Surveys

215. Both ONS and GROS conducted Census Quality Surveys to coincide with the 1999 Census rehearsal. ONS selected a country-wide sample and issued each selected household with a Census form and then later revisited the household to interview household members. The interview would elicit 'correct' responses to Census questions. GROS took a different approach. We sampled households that had taken part in the 1999 rehearsal and interviewed them. The Scottish version concentrated on the questions that were not in the England and Wales survey, concentrating on those expected to be unique in the 2001 Census in Scotland. In the event several questions were either changed or introduced between the 1999 Rehearsal and the 2001 Census. It was decided not to conduct a CQS in 2001 because the results from 1999 should, in general, meet the need for such information and that non-Census field activities in 2001 should concentrate on the CCS.

However full results from neither exercise have yet appeared, though 'gross agreement rates' for each Census variable appear in the ONS 2001 Census Quality report for England and Wales pdf. It is recommended that the results of the Census Quality Surveys are published.

Vacant Follow-Up Survey

216. As stated above (paragraph 196), GROS conducted a survey of property identified as vacant during enumeration. The survey was carried out by those staff who had acted as supervisors of enumerators and who had opted to do this extra work. The results were eventually released on the website of the Scottish Executive, who had commissioned the survey, Post Census Vacant Survey. Among other things the survey showed that some 84 per cent of property assessed as vacant by enumerators, and where a response was obtained in the survey, were confirmed to have been vacant.

Recommendations

217. A list of recommendations extracted from the above follows. It is recommended that:

The Census programme

- Rehearsal activities should include output production (paragraph 111).

- a rehearsal should be held in good time for all downstream processing and output production to be rehearsed before live production (paragraph 11).

Management

- the UK committee structure for Census policy matters should be maintained until all policy issues are dealt with and that membership is drawn from all interested parties (paragraph 120).

Data collection

- there should be a space on the individual form for an individual sequence number so that enumerators aren't tempted to use the form number for that purpose. An individual number would also be very useful in data quality work (see paragraphs 25 and 212).

- the instructions on the I form should clarify that 'position in establishment' relates to the person covered by the form and not the person who happens to complete it (paragraph 50).

- if a count of the armed forces is required, appropriate instructions should be on the form itself – as in 1991 (156).

• the question on owning or renting should be adapted to ensure the proper treatment of households on housing benefit (paragraph 170).

• persons sleeping rough should be recorded where they are enumerated (paragraph 87).

Data capture

• all write-in answers should be coded (paragraph 109).

• variables combining postcode and country should exclude categories for the 4 UK countries (paragraph 171).

Downstream processing, general

• Given that it is impossible to guarantee that automatic processing can do all of the data cleaning needed, there should be an easy-to-use clerical edit process that incorporates the checking of the edited records (paragraph 5).

• A 'process control' system to which GROS staff have read-only access should be set up for next Census no matter how it is processed (paragraph 7).

• the relative status of the checks on variables should be re-assessed so that the values supplied on the form for the variable relationship are retained more frequently. Also, time should be taken to ensure that a full set of checks on relationship are included in the edit process (paragraph 38).

• an inventory should be kept of all products dependent on each 'upstream' product so that amendments are carried through comprehensively (paragraph 133).

• the input ID should be retained on the output database, despite concerns over confidentiality. For 2001, the link to an input record could still be made – with difficulty - without the input ID, therefore not retaining it was ineffective (paragraph 148).

One number Census

• when adding records as part of ONC imputation, broadly the same approach as for 2001 to find locations and donors for recipients should be used (i.e. concentrating on where the enumerator recorded absent or refusing households) but with a penalty method for restricting the re-use of donors (paragraph 61).

• no switching between council area-based and HBA-based EAs should be done in ONC processes. Population estimates using more detailed figures on migration than those from the NHSCR should help (paragraph 65).

Disclosure control

- those taking decisions about disclosure control should take full account of the implications for output (paragraph 112).

- a unified set of methods for disclosure control should be used, for Area Statistics, Origin-Destination Statistics and SARs, based on modifying the Census database in a way that does not additionally require any modification of tabulated results and that is sufficient for both tabular output and SARs (paragraphs 73 and 125).

Output

- the Census Act 1920 should be reviewed to ensure that it caters for the full range of planned activities and outputs, as it does not specifically cater for reports that are not laid before Parliament nor paid for such as Occasional Papers (paragraph 102).

- generating tables on migration and travel in the areas statistics from origin-destination matrices should be considered, given possible improvements in tabulation and data-handling software (paragraph 95).

- producing a smaller range of standard tables should be considered (paragraph 98).

- if products in a ‘safe setting’ are available in future, they should be available at the outset at sites in all 3 Census Offices (paragraph 107).

- the number of versions of each table during table development should be minimised (paragraph 144).

- table design should not depart too far from whatever default options are available in the tabulation software, since printed tables will form a tiny minority of the full output, (paragraph 146).

- more should be done than in 2001 to accommodate the wishes of users about the format of bulk supply (paragraph 160).

- arrangements should be made to ensure that tabulation software tabulate areas in publication order as default (paragraph 160).

- averages should be not calculated by deriving them from tabulated separate – and detailed - values of the variable being averaged (paragraph 163).

- once the overall scheme of tables is known, a robust folder structure should be agreed for storing them and all changes to that structure should be agreed by those using it (paragraph 164).

- databases for part of the country set up for testing purposes should

be abandoned as soon as full databases are available (paragraph 166).

- supporting information should be put on the web as it becomes available rather than wait till all is ready for publication in a single product (paragraph 174).

- when an 'Update' with new information for users is added to the website, material in the previous updates should be deleted and still current material in them be absorbed into the general structure (paragraph 175).

- the range of means of disseminating Census results by web should be reviewed to see if some rationalisation of publicly funded services is possible (paragraph 180).

- a review should be carried out of the extent to which users will be able in practice to use software theoretically generally available (paragraph 188).

- the website to disseminate Census results should allow users to save recodes of area classifications (paragraph 189).

- the website to disseminate Census results should permit the addition or replacement of area types (paragraph 194).

- the website to disseminate Census results should allow users to locate statistics geographically using not only postcodes on the set frozen for the Census but postcodes created in the years following the Census (paragraph 195).

- If a database for the UK is not available to each Census Office, the possibility should be considered of country databases extended to include, for Scotland, records enumerated in the rest of the UK for migrants with origin in Scotland and travellers with destination in Scotland (paragraph 143).

Data quality

- the system to check data quality should mirror the processing database as completely as possible, in particular, by including the record for the family created by the HCA (paragraph 85).

- all variables generated in input processing should be available for checks on data quality (paragraph 171).

- any system to check data quality should use all of the data for Scotland in a single database. The larger the area used the better for checks on migration and travel (paragraph 198).

- all images of forms should be made available as soon as they are

received from data capture and that they remain available while downstream processing and work on data quality is in progress (paragraph 202).

Geography

- given inaccuracies with postcode of migration origin and travel destination, when such a postcode is identified as one that GROS records as split, the A split be used invariably – unless there are some key B splits containing large employing establishments (paragraph 18).

- a split (eighth) character for a postcode should be adopted as standard throughout the whole of the UK even though it will be filled with a space for the most part (paragraph 62).

- a common format of postcode should be used throughout Census operations (paragraph 81).

- GROS should continue to freeze its postcode base throughout the enumeration and processing of the Census (paragraph 83).

- a common practice for presenting reference maps for Census geographies should be adopted. Because a depiction via aggregations of OAs can cause presentational problems (e.g. detached OAs for wards, or, in the case of localities, extra rural land that in population terms means little), it is recommended that boundaries depicting areas as exactly as possible be adopted as the standard way for showing boundaries in supporting information (paragraph 128).

- for Census processing, the area type 'settlements and localities' should include an area 'rest of Scotland', and the area type 'inhabited islands' should include the area 'mainland Scotland' (paragraph 129).

- the assignment of OAs to higher areas should be done on the basis of population rather than households (paragraph 131).

- Postcodes of large CEs should be carefully checked before enumeration and during processing (paragraph 207).

2001 Census loose ends

- databases should be checked and eastings and northings of postcode of enumeration corrected, if in error (89).

- the possibility should be considered of country databases extended to include, for Scotland, records enumerated in the rest of the UK of migrants with origin in Scotland and travellers with destination in Scotland (paragraph 143).

- material placed on the website in ‘Updates’ should be absorbed into the general structure for the 2001 Census (paragraph 175).

- links should be provided on the website to, e.g., the pages on the ONS website that explain the National Statistics – Socio-economic Classification (paragraph 176).

- more Occasional Papers should be produced – eg on relationships between migrants and non-migrants within households (paragraph 178).

- links to Occasional Papers and Gaelic Report based on the 2001 Census have be placed in the 2001 Census part of the main website (paragraph 178).

- some means of providing such thematic maps on ethnicity and religion should be provided as for other topics (paragraph 183).

- some notification of which discs have been replaced should be put on the GROS website (paragraph 186).

- remaining comments on data quality should be collated (paragraph 204).

- fuller results comparing statistics on workplace with the interdepartmental business register are prepared for public release (paragraph 207).

- a short note describing errors where large communal establishments have been included in the wrong Output Area should be put on the GROS website (paragraph 207).

- The check should be done on imputation of migration postcodes for students at the same address (paragraph 211).

- The results of the Census Quality Surveys should be published (paragraph 215).

- precedents built up on confidentiality of commissioned tables should be formally set out as rules for future work (paragraph 110).

- it should be checked whether the inclusion of GROS within the SE network system has overcome the barrier to the use of SuperTABLE by the SE (paragraph 188).

- the SCROL website should be amended so that users can use saved recodes of area classifications (paragraph 189).

- processes to add or replace areas to SCROL should be reviewed so that they can easily be used as required (paragraph 194).

- an index linking current postcodes to 2001 Census areas should be added to the SCROL website. Perhaps the index should include all postcodes that have existed since the 2001 Census (paragraph 195).

GROS Census Division March 2007

Appendix 1 Comparison of CEs enumerated with those on pre-Census register

The following describes how enumeration of communal establishments (CEs) went for all types. The comparison is with the CEs we (that is geography section) had put on to the register we set up of CEs to aid Data Collection in the first instance. The register contains an item meant to record the capacity of each establishment.

The table below compares the number of persons enumerated with capacity. CEs are divided among

- those that were on the register that were, in the event, enumerated mostly but not always at the location we expected; some were enumerated at the expected location but as households (these are included in the table below)

- those we enumerated that were not on the register

- those CEs on the register that we didn't enumerate; geography followed these up (concentrating on categories other than hotels).

	CEs on register enumerated		CEs not on register enumerated	CEs on register not enumerated		Excess	
Type of CE	Capacity	Persons enumer-ated	Persons enumerated	Capacity	% for which explan-ation obtained	Persons enumer-ated - capacity	as % of capacity
General Hospital	9497	1483	45	2731	44	-10700	-88
Psychiatric Hospital or Home	6563	3268	585	302	100	-3012	-44
Other Hospital	4189	2809	116	32	41	-1296	-31
Nursing Home	23532	20362	1101	530	15	-2599	-11
Residential Care Home	18114	13758	2599	1842	26	-3599	-18
Children's Home (inc. Secure Units)	1251	811	148	187	27	-479	-33
Other Medical and Care Home	1223	898	550	163	31	62	4
Defence	9722	2991	0	1265	73	-7996	-73
Prison and Young Offenders	3657	4084	25	0	-	452	12
Educational Establishment	24376	17096	4258	2194	61	-5216	-20
Hotel, Boarding House, Guest House	22703	2198	1493	17038	2	-36050	-91
Hostels (inc. Youth and Homeless)	5031	1452	1088	1533	28	-4024	-61
Civilian Ship	0	0	2	0	-	2	-
Other	4900	998	1448	0	-	-2454	-50
All types	134758	72208	13458	27817	19	-76909	-47

The table shows

- Hotels have the greatest shortfall (91%) but this would be expected given the transient nature of their client group; similarly hostels show a large shortfall (61%). It is a little worrying that some two-fifths of hotel capacity didn't get enumerated. However, if the resident to capacity ratio of the enumerated hotels is anything to go by we should have missed only around 2,000 people.

- General hospitals have the greatest shortfall (88%) after hotels but this would also be expected ; other types of hospital show smaller percentage shortfalls; the 41% for psychiatric hospitals may be explained by a move to shift clients 'into the community'.

- Defence establishments show the next largest % shortfall (under investigation)

- Educational establishments (boarding schools and hall of residence) - 20% shortfall and a possible under enumeration of 5,000

- Nursing and residential care homes - perhaps 6,000 missed; however the latter category in particular can have a high turnover of establishments and it is difficult to keep the register up to date.