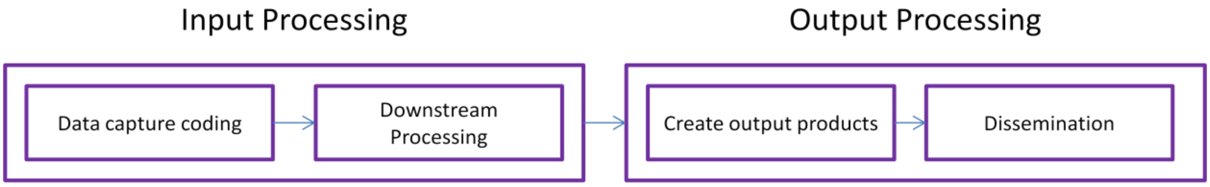
5 Data processing

Introduction

5.1 Before outputs from the census could be produced, responses on the 2011 Census questionnaires and Census Coverage Survey had to be captured. Responses were converted into coded data and then validated and cleaned so that the outputs were of high quality.

5.2 As was the case in 2001, the 2011 Census was processed in phases.

- Input processing - comprising two stages
  - the main data capture and coding stage,and
  - downstream processing - the subsequent process to clean, adjust, validate and protect the data (including edit and imputation, coverage assessment and adjustment process, and statistical disclosure control)
- Output processing - comprising the creation of an outputs database, from which census output products were produced and subsequently disseminated



5.3 This chapter deals with the data processing stages up to and including the coverage assessment and adjustment, and summarises the quality assurance processes built in to the data processing operations. Statistical disclosure control is covered in chapter 6 and the tabulation process is described in chapter 7.

Data capture and coding

5.4 As with the 2001 Census, it was decided that better value for money could be obtained from contracting-out the main scanning, data capture and coding services. Paragraphs 2.328 to 2.335 report on the strategy that was adopted for the 2011 Census. ONS carried out an 'open options' procurement exercise in 2005 to select a contractor to provide services for the printing and processing of the census questionnaire and other support services. The contract was awarded in August 2008 to Lockheed Martin (UK), which sub-contracted the operational elements of the data capture and coding process to UK Data Capture Ltd. The processing was carried out at a specially commissioned and secure site in Trafford Park, Manchester. More than 24 million census questionnaires were handled between March and November 2011, with processes to capture and code all of the ticks and texts on these questionnaires.

5.5 The paper census questionnaires were securely stored at the processing centre until an electronic archive copy was made for retention for 100 years as a historical record. The paper questionnaires were then destroyed (shredded) in a secure, controlled manner, witnessed and verified by ONS census staff. The electronic archive copy was copied to microfilm for retention at a secure ONS site and will be transferred to the National Archives for eventual release in 2112. The captured and coded data was securely transferred to ONS systems for further processing and validation as part of the downstream processing stage.

Data capture

5.6 Questionnaire processing began by scanning the questionnaires and capturing the data in a four-stage process:

1. Scanning – to obtain images of the completed questionnaires

2. Image checking – to check the quality of the images produced from scanning and to prepare them for data capture. This comprised:

   a) automated image quality assurance (AIQA) to check the size of the image and that the expected barcodes were present
   b) data lift and registration (DLR) to undertake additional quality checks and prepare the images for data capture, and
   c) document analysis where an interactive user could verify or reject the quality of captured images if either of the two previous components (AIQA or DLR) had identified an error

3. Recognition – to capture automatically the data from the questionnaires.

   This was achieved via:
   a) optical mark recognition (OMR) to capture the tick box data
   b) optical character recognition (OCR) to capture the characters from text boxes and numeric responses, and
   c) contextual analysis (CTX) to ensure the captured data was contextually logical, with the expected type of text entered in the appropriate sections of the questionnaires

4. Keying – to capture manually the fields that could not be recognised automatically with sufficiently high accuracy

5.7 Accuracy rates were used to report on the quality of the captured data, and were measured against a set of targets. To calculate the accuracy rates a sample of data was presented to a team of keyers. The values obtained were then compared with the original values, and agreement between the two values was considered to be an accurate result. In cases where the two values differed, the record was passed to another keyer. Table 5.1 shows the possible outcomes of the verification.

Table 5.1 Possible outcomes of verification

Original value	Keyer 1	Keyer 2	Outcome
Value 1	Value 1	-	Pass
Value 1	Value 2	Value 1	Pass
Value 1	Value 2	Value 2	Fail
Value 1	Value 2	Value 3	Inconclusive

5.8 The accuracy rates were calculated as the percentage of passes out of the total number of cases sampled. The sample sizes used depended on the number of people expected to provide responses for the field being sampled, the expected accuracy, and the acceptable bounds for error in the sample. Inconclusive outcomes were excluded from accuracy calculations, but were monitored to ensure that their volume remained within acceptable limits.

5.9 All of the targets set for data capture accuracy were exceeded. These are set out in Table 5.2 (with accuracy achieved in 2001 for comparison). For all of the field types, the accuracy achieved for the 2011 Census was higher than the targets set and was broadly in line with achieved accuracy rates for the 2001 Census.

Table 5.2 Accuracy results for paper questionnaire data capture, 2011 and 2001

		Achieved accuracy (%)	
Field type	**Target (%)**	**2011**	**2001**
Marks	99.30	99.85	99.84
Numeric	98.00	98.93	99.75
Alpha numeric	95.00	97.58	98.99
Year of birth	99.95	100.00	99.93
Sex*	99.50	99.96	-
Marital status*	99.50	99.92	-

*Accuracy rates for these questions was not separately reported for the 2001 Census

Coding

5.10 The data were then loaded into a database and validated to ensure that the values for each question were within the range specified in the relevant coding frame.

5.11 The coding process assigned numerical values to written text and ticked boxes. This involved applying coding rules and standardised national coding frames, such as SIC07 (Standard Industrial Classification 2007) and SOC2010 (Standard Occupational Classification 2010), which allow data from different sources to be easily compared. The data were loaded into a database and validated to ensure that the values for each question were within the range specified in the relevant coding frame. The text responses provided on both the paper and online questionnaires were converted into coded data using this coding process.

1. Automatic coding – the first step for all responses was to attempt to match them to the appropriate reference data and assign a code automatically

2. Frontline coding – responses that could not be automatically coded were assigned to a team of coders who attempted to code the response using defined business rules

3. Expert coding – responses that could not be coded at frontline coding were referred to another team of coders, who had additional reference materials available to code the response,and

4. Welsh expert coding – responses in Welsh were assigned to bilingual coders who used the same process as expert coders to code the response

5.12 Table 5.3 shows that the targets set for coding accuracy in the 2011 Census were exceeded and, again, are broadly similar to the accuracy achieved in the 2001 Census. The majority of records found to have been coded incorrectly were only minor inaccuracies, for example, a primary school teacher coded as a secondary school teacher.

Table 5.3 Accuracy results for coding, 2011 and 2001

		Achieved accuracy (%)	
Field type	**Target (%)**	**2011**	**2001**
Country of birth	96.00	99.87	99.80
Ethnic group	96.00	98.80	98.60
Religion	96.00	99.11	98.80
Citizenship (passport)	96.00	99.82	*
Language	96.00	99.58	*
National Identity	96.00	99.66	*
Occupation	88.00	94.14	91.10
Industry	88.00	93.11	89.10
Workplace address	85.00	93.28	94.10
Address one year ago	96.00	98.33	98.10
Visitor address	96.00	98.63	*
Second address	96.00	98.65	*

*Data not collected on these topics in 2001

Completion of upstream processing

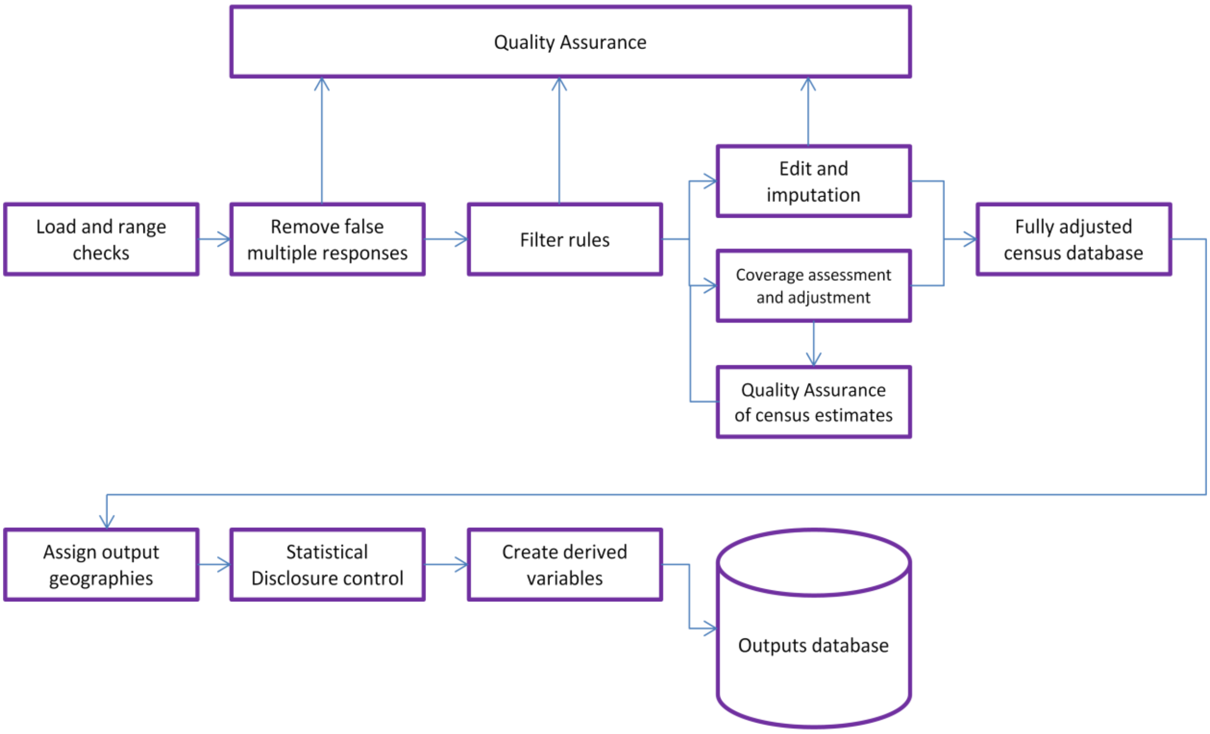
5.13 After completing these processes the captured and coded data was then securely transferred for loading on to ONS systems for further processing and validation (downstream processing). The data was sent in encrypted form for increased security and in batches (processing units) to facilitate processing. The physical archive data (recorded on microfilm) was transferred by secure transport to a secure ONS site.

Downstream processing

5.14 The downstream processing (DSP) project provided a set of IT systems capable of carrying out the subsequent processing of all 2011 Census and Census Coverage Survey data. The project was responsible for the live running of the data through the downstream process and providing operational support during live running.

5.15 The whole process started with the loading of the data and ended with a disclosure control process, before the production of outputs (see the main steps in figure 5.1). The process control centre for DSP monitored the movement of each processing unit (PU) through the system, with validation and checks at the completion of each stage. Each process could handle multiple areas at a given time. For example, in item imputation up to four PUs could be run simultaneously and this was later upgraded to eight for coverage imputation.

Figure 5.1 The main steps in downstream processing


Quality Assurance
Load and range checks
Remove false multiple responses
Filter rules
Edit and imputation
Coverage assessment and adjustment
Quality Assurance of census estimates
Fully adjusted census database
Assign output geographies
Statistical Disclosure control
Create derived variables
Outputs database

Range checks

5.16 The range checks process checked that the value of each variable was within the valid range for that variable. For example, there were four valid values for the sex variable: male, female, missing or multi-tick. The range checks process verified that all values for the sex variable were one of these four valid values. If an invalid value was found it was set to 'missing' or 'not required', so that all values were valid for the statistical processes that would be applied to the data later. The missing or multi-tick values were then imputed as part of the edit and imputation process.

5.17 The range checks process also cleaned up all postcode fields by removing any invalid characters. At the data capture stage the strings of text in postcode fields were captured without any validation of the text being performed. Because statistical processes carried out on the data at a later stage required postcode fields to contain only valid values, any invalid characters were changed to blank during the range checks.

Removing false person records

5.18 As part of the data capture process, a person record was created during the recognition phase each time at least one mark was detected in any of the person questions. But such records could be created in error if, for example: there was dust on the scanners that was incorrectly interpreted as a mark; or where respondents crossed through whole pages of the questionnaire as not being relevant and this had been identified as a response; or where respondents may have accidentally skipped pages, completing their response over two different person records. A process was developed to identify genuine person records, after analysing data from the 2001 Census to establish which combinations of key variables were most often present on genuine responses.

5.19 For a person record to be counted as a genuine response and kept in the data the following information had to be present on the record:

- name (from individual questions or household members table) or date of birth, and
- at least one other item, different from the above filter, from: name (from individual questions), date of birth, sex, marital status, or name (from household members table)

If a person record did not meet these requirements then it was considered to be a false person and was flagged as an invalid person record.

5.20 The removal of false persons process removed a total of 982,400 person records (1.8 per cent), compared with the removal of 3,297,800 person records (6.3 per cent) in the 2001 Census. It should be noted, however, that a large proportion of the false person records identified in the 2001 Census had been created by processing errors rather than respondent errors: marks on forms that were the result of printing quality and handling, as well as dust settling on the scanners, had been captured as responses. Fewer records had to be removed in the 2011 Census because of improvements in the processing as a result of the lessons learned from 2001, and because a modification to the rule used to identify false persons minimised the number of genuine responses removed.

Resolving multiple responses

5.21 There was an increased likelihood of multiple responses from the same household occurring in the 2011 Census compared with previous censuses, because of the introduction of online completion and the post-back of paper questionnaires. Multiple responses at the same address could be created in a number of ways: for example, both a paper and an online response being returned for the same address effectively created a multiple household response; or a person being included on the same questionnaire more than once could create a multiple individual response. A new process was therefore developed to resolve both household and individual multiple responses at the same address.

5.22 Multiple household responses were identified by looking for more than one response for an address ID, and matching the people on the different responses to determine if they related to the same or different households. All individuals on one questionnaire were matched to all individuals on all other household questionnaires returned for the same address. Name, date of birth, and sex were the variables used for matching.

5.23 Initially the following criteria were used to determine whether individuals were a match:

- first name and surname matched exactly
- date of birth matched on day and month, or month and year, and
- sex matched, or was missing or multi-ticked on one or more of the records

5.24 If a match was not found using the above criteria, additional matching was carried out for individuals aged 30 or over using the following criteria to determine a match:

- ‘soundex’ of both first and last names matched, or name was missing on one or more of the records
- date of birth matched exactly, and

- sex matched, or was missing or multi-ticked on one or more of the records

('Soundex' is an algorithm for indexing names by sound, which allows for names that are spelt differently to be matched. The algorithm converts names into a four digit code by retaining the first letter of the name and assigning a code to the consonants in the rest of the name. Similar sounding consonants are assigned the same code, therefore similar sounding names will match on 'soundex'.)

5.25 The second set of matching criteria was not applied to individuals aged under 30, to minimise the risk of identifying twins as matching individuals.

5.26 If any matching individuals were found the multiple responses were considered to relate to the same household and were resolved into one response. If no matching individuals were found the multiple responses were considered to relate to different households and were left in the data, unless one of the following applied:

- all of the usual residents on one of the responses were aged under 16
- one of the responses was on a Welsh language questionnaire, or
- one of the responses was on an online questionnaire

These responses are likely to be a continuation of another response, and therefore even if there were no matching individuals in these cases, multiple responses were considered to relate to the same household and were resolved into one response.

5.27 When multiple responses relating to the same household or individual were identified, the records were merged to leave just one record for the household or individual. The most complete response was kept, with any missing variables being filled in from the other response(s) if possible. In the case of multiple individual responses, a response on an individual questionnaire was given priority over a response on a household questionnaire.

5.28 Any multiple communal establishment responses found for the same address were assumed to relate to just one communal establishment and were resolved into one response. Again, the most complete response was kept, with any missing variables being filled in from the other response(s) if possible.

5.29 Addresses where both a household and communal establishment response had been returned were assumed to relate to a communal establishment, and the household response was deleted after any individuals on the household response had been moved to the communal establishment response.

5.30 The process also dealt with multiple responses involving dummy records. Dummy questionnaires were completed by field staff for addresses where no census questionnaire had been returned. The dummy questionnaire collected basic information about the property. If there was more than one dummy record for an address the records were resolved into one by starting with the most complete record and filling in any missing variables from the other record(s) if possible.

5.31 For some addresses both a dummy record and household or communal establishment record existed. In these cases the dummy record was deleted, but information that was missing on the household record but present on the dummy record was first copied on to the household record.

5.32 Although the main purpose of the resolving multiple responses (RMR) process was to identify and deal with any duplicate responses, the process also had a secondary function of moving records to ensure that all individuals from individual or continuation questionnaires were included on the household or communal establishment record to which they belong.

5.33 Individual questionnaires could be requested by anyone in a household who did not want to answer their individual questions on the main household questionnaire. They were also issued to all usual residents of communal establishments, because the main communal establishment questionnaire collected only information about the establishment and not personal information about the residents. Continuation questionnaires were used for households with more than six usual residents, where there was not enough space on the main household questionnaire for all residents to answer the individual questions.

5.34 Household or communal establishment records were also created for addresses where only a dummy record existed, or 'orphan addresses' where only individual or continuation responses existed without the main household or communal establishment response.

5.35 Table 5.4 shows the number of records that were removed as part of the RMR process, while table 5.5 shows the number of records that were created as part of this process. Individual records were only deleted during the RMR process; none were created.

5.36 Table 5.6 shows the overall change in the number of records as a result of the RMR process. Note that records removed and created may not sum to the net change because of rounding. However, the work to assess overcount (see paragraph 5.60(c)) identified that in future the RMR process should also consider resolving duplicates or multiple responses in the same postcode as well as at the same address.

Table 5.4 Records removed by the resolve multiple responses process

	Records removed	
Record type	Number	Per cent
Individual	237,200	0.44
Household	181,300	0.78
Communal establishment	300	0.71

Table 5.5 Records created by the resolve multiple responses process

	Number of records created	
Record type	From dummy responses	From orphan addresses
Individual	1,466,500	5,500
Communal establishment	6,900	5,700

Table 5.6 Overall change in the number of records during the resolve multiple responses process

Record type	Number of records Net change	 Percentage Change
Individual	-237,200	-0.44
Household	+1,290,600	+5.32
Communal establishment	+12,300	+26.45

Filter rules

5.37 A further process provided a consistency check that reconciled contradictory responses arising from instances where a questionnaire filter had been ignored; for example:

- when a respondent had misunderstood the second address question and re-entered their enumeration address instead, or
- when a child aged 14 had been recorded as being in full-time employment

Edit and imputation

5.38 As with all social surveys, the 2011 Census data contained item non-response and inconsistent responses to the census questionnaire. Typically, item non-response refers to an event where a respondent does not know or refuses to answer a particular question in an otherwise completed questionnaire. Inconsistent responses are relationships between recorded values for two or more variables that are clearly invalid, such as a parent being younger than their child. Item non-response and inconsistent responses can have a detrimental impact on the utility of the census data in three basic ways.

- Missing and/or inconsistent data can lead to a reduction in the precision of population estimates
- If the characteristics of the non-respondents differ from the respondents, population estimates may also be biased. This is referred to as a non-response bias, and
- Users of census data may try to account for item non-response and inconsistencies in the data in different ways, leading to disparity in population estimates derived by different analysts

5.39 Imputation is a widely recognised statistical framework that serves to minimise these risks. The census imputation strategy had one overarching objective to replace all missing and inconsistent data with imputed values. This is done by using a robust statistical method that estimates the distributional properties of the missing/inconsistent data as accurately as possible.

5.40 To meet this aim, several objectives served to underpin two key aspects of the imputation system.

The baseline statistical methodology used to impute the census data should:

- resolve inconsistencies with minimal change to the observed data
- implement a consistent approach to the imputation of all census variables, and
- focus on estimating accurately the multivariate joint distributions in the data. This means imputing accurately the relationships between variables such as age by gender by marital status where one or more of these variables are missing, rather than imputing the variables independently from each other

5.41 To ensure that the imputed data had a beneficial impact on the utility of the census data, the statistical performance of the system during live processing should also:

- avoid introducing bias or inconsistency into the census data through the imputation process
- adjust for non-response bias where appropriate

5.42 Development of the 2011 Census imputation strategy began in 2005 with a review of the 2001 Census methodology and an evaluation of alternative processing platforms. In 2001 the UK Census Office developed the edit and imputation system (EDIS) for resolving inconsistencies and imputing for item non-response. From the review, the Canadian Census Edit and Imputation System (CANCEIS) (Bankier, Lachance, Poirier 1999; Canceis, 2009[72]) was identified as a potential alternative.

5.43 Both EDIS and CANCEIS implement a donor-based/minimum-change imputation strategy (Fellegi & Holt 1976[73]), widely recognised as a methodological standard for imputing census and social survey data. In this approach inconsistencies are identified by a set of pre-defined edit rules specifying invalid relationships between variables and identifying how they could be resolved causing the minimum amount of change to observed data. Missing values are replaced by drawing an observed value from another record in the data, referred to as a donor. A donor is selected from a small pool of potential donors with characteristics similar to the record currently being imputed. Similarity is measured by comparing the differences between the record needing imputation and each potential donor across a set of key demographic and other predictive matching variables.

5.44 CANCEIS was better designed to optimise the statistical advantages of a donor-based approach (Rogers & Wagstaff, 2006[74]). Amongst others, significant optimisation strategies included:

- simultaneous multivariate processing of inconsistent and missing data under edit constraints. For resolving inconsistencies, this allowed all plausible solutions from every potential donor to be evaluated and only those leading to minimum change in the record needing imputation to be included in the potential donor pool. For imputing missing data, it also served to ensure a more accurate imputation of the relationships between variables with missing data. Imputing under edit constraints meant that invalid relationships between variables belonging to an individual and between people in a household did not arise through the imputation process

- staged near donor search strategies. This contributed to the accuracy of the imputation by ensuring that imputed values were drawn from donors living in close geographic proximity to the record being imputed
- stratification by household size. This ensured that donor selection was not only based on the observable characteristics of the person having data imputed, but also on the composition and structure of other people living at the same address, and
- a soft editing strategy. Soft edits were employed to identify records in the data with valid but unique or unusual characteristics. The soft edit strategy allowed such records to remain in the data but did not allow the characteristics associated with them to be propagated. This served to preserve the quality of the observed data and help minimise the risk of introducing bias into the data through the imputation process

5.45 Early design and development of the end-to-end imputation processing strategy and parameterisation of CANCEIS was conducted through a systematic empirical research programme (Rogers & Wagstaff, 2006[74]). A synthetic census data set consisting of fully observed records from the 2001 Census was created and perturbed in a way consistent with the item non-response patterns also observed in 2001. Optimal tuning was based on analyses focusing on how well the system recovered the observable statistical properties of the perturbed data. Based on this research, the recommendation that CANCEIS was the most appropriate platform for the 2011 Census imputation strategy was approved through an independent quality assurance process by leading academics at the University of Southampton.

5.46 Research and development directed at optimising the performance of the 2011 Census imputation system continued up to and throughout live processing. This ensured that the structure and characteristics of the 2011 Census data that may have differed from these in 2001 were included in the fine tuning of the CANCEIS system parameters.

5.47 A detailed report of the development and final design of the 2011 imputation strategy can be found online[38].

5.48 Table 5.7 provides some key post-census processing measures comparing how well the 2001 and 2011 imputation systems met the statistical objectives of the baseline methodology.

5.49 In general, the investment in the design and development of the 2011 Census imputation baseline methodology led to some significant improvements over that applied in 2001. The most notable improvements are clearly linked to the statistical objectives for this aspect of the imputation system. The baseline methodology ensured that:

- inconsistencies were always resolved with minimal change to the observed data
- almost all of the 18.6 million people and 2.8 million households needing at least one value imputed were treated consistently using the same processing method, and
- the imputed data for all household records and a high proportion of person records (82 per cent) were drawn from an implicit multivariate model of all plausible values specific to each particular record that needed imputing while taking into account within and between person edit constraints

5.50 Typically, records that were not imputed using the standard baseline methodology had unusual characteristics such as extremely young parents or extremely young people reporting a duty of care to someone else in the household. In most cases, these records were imputed by passing them through the same system but with slight adjustments to some of the parameters in CANCEIS. For a very small number of records where this did not provide a solution, inconsistent and missing data was edited based on a record by record domain expert review.

5.51 Overall, the methodology implemented in the imputation system was successful in meeting the main aims and objectives of the strategy. A complete and consistent database was achieved within the timescales available in the downstream processing timetable.

Table 5.7 Operational comparisons of CANCEIS and EDIS

		EDIS (2001)[a]	CANCEIS (2011)[b]
Persons			
Records processed		49.4 million	53.5 million
Average number of records in processing unit		500,000	530,000
Average time to take to impute a processing unit		48 hours	12 hours
Persons needing at least one question imputed		13.8 million[c]	18.6 million
Percentage		28%[c]	35%
-	Percentage imputed as household taking into account multivariate joint distributions between persons and between questions	34%	82%
-	Percentage imputed as individuals	72%	18%
-	Percentage imputed using alternative methods to that implemented in the primary imputation system	3%	0.10%
-	Persons imputed by more than one method	Over 1 million	Under 300
Households			
Records processed		22.3 million	24.3 million
Households requiring at least one item imputed		2.5 million	2.8 million
Percentage		11%	9.5%
-	Percentage imputed taking into account multivariate joint distributions between questions	97%	100%

a Census 2001 Review and Evaluation Report[40]
b Data derived through the 2011 CANCEIS system diagnostics
c Excludes overlap with deterministic applied to 11.8 million persons

Coverage assessment and adjustment

Introduction

5.52 Most census-taking countries carry out some form of coverage assessment and adjustment, often using a post-enumeration survey (PES). Measured undercount levels have, on the whole, been increasing over the past few decades. More importantly, the differential nature of the undercount has worsened with, for example, young males in inner city areas becoming increasingly difficult to enumerate. This has led to increasing priority and focus on the methods for measuring this differential undercount.

5.53 The coverage assessment and adjustment (CAA) process was designed to identify and adjust for the number of people and households not counted in the 2011 Census. The extent of this under-enumeration was identified using a large survey covering approximately 340,000 households, the Census Coverage Survey (CCS) (see chapter 4). Standard statistical estimation techniques were then used to produce an adjusted database from which the final census results were produced. These results also formed the new 2011 base for the mid-year population estimates produced by the ONS. The overriding strategy was to build on the 2001 One Number Census (ONC) framework, using it as a platform to develop an improved methodology.

The 2001 One Number Census

5.54 For the 2001 Census the ONC project had the goal of providing a methodology and processes to identify and adjust for the number of people and households not counted (see Brown *et al* 1991, Holt *et al* 2001). The ONC estimated the undercount in the 2001 Census to be 6.1 per cent of the total population in England and Wales.

5.55 The ONC was a big step forward. Both the Statistics Commission and the Local Government Association published reviews that concluded that the methodology used in 2001 was the best available and no alternative approach would have produced more reliable results overall. However, there were some issues with the results which led to further studies and adjustments. The lessons taken from these were that:

- the ONC had not been able to make robust adjustments in all situations, particularly when there were pockets of poor census response
- engagement with stakeholders was critical to facilitate user acceptance of the methodology
- the methodology needed to be robust to failures in underlying assumptions and in particular to have inbuilt adjustments for such failures – for example, any lack of independence between the census and CCS
- two of the weaknesses of the were not having additional sources of data to complement the CCS, and the perception that it would solve all ‘missing data’ problems
- the measurement of over-count required greater attention, and
- the balance of 'measurement' resource between easier-to-count and harder-to-count areas needed careful consideration

2011 methodology

5.56 Accordingly the strategy for the programme of coverage assessment and adjustment in the 2011 Census aimed to develop an improved methodology, by not only addressing the lessons from 2001 but taking into account changes to the 2011 Census design. The programme had a number of specific objectives.

- Gain acceptance of the methodology from users. This was important because users, particularly local authorities, would not trust their census population estimates if they were not confident about the methodology used to derive them
- Develop simple methods where possible, to aid communication of the methodology
- Measure the extent of each of these, permitting more transparent adjustments. (there are a number of ways in which undercount can occur (such as missing a whole household or missing a person from a counted household).
- Provide local authority and age-sex level population estimates with minimal variation of precision, therefore ideally being the same relative precision across all
- Target precision rates (for sampling errors only) of 95 per cent confidence intervals of 0.2 per cent around the national population estimate (ie plus or minus 120,000 persons) and 2 per cent for a population of half a million (ie plus or minus 10,000 persons), and
- There should be no local authorities with a precision worse than the worst that was achieved in 2001, and to improve the worst 5 per cent of areas (ie there should be no relative confidence interval for a local authority total population that is wider than 6.1 per cent, and a 5 per cent confidence interval is the desirable upper bound)

5.57 The methodological improvements necessary to meet the objectives were developed in the years leading up to 2011 when the methods were finalised. The methodology had wide scrutiny and peer review during development, by:

- Census Design and Methodology Advisory Committee
- GSS Methodology Advisory Committee
- ONS Statistical Policy Committee, and
- International and other academic peer review (such as the RSS)

5.58 In addition, the CAA and QA process (see paragraphs 5.65 to 5.78) were subject to an independent review of the methods. This review, led by Professor Ian Plewis at the University of Manchester, reported initially in February 2011, and subsequently in June 2011. The review made a number of recommendations, all of which ONS accepted and addressed, and the review team concluded that:

‘We would like to put on record our belief that many lessons have been learned from the Census in 2001 (which was itself a considerable improvement over the 1991 Census). We have been impressed by the scope and depth of the methodological investigations initiated by ONS, by their willingness to discuss with a wide range of interest groups concerns about coverage and Quality Assurance (QA), and by the procedures that are in place to use field staff flexibly. We are reasonably optimistic that, having taken account of our recommendations to develop, document and consult on specific aspects of methodology, the 2011 Census in England and Wales will provide population estimates that can guide resource allocation and social policy

in the right direction for the next ten years. It must, however, be recognised that the target 95 per cent confidence intervals set by ONS for the population counts – a maximum interval of ± 3 per cent for all LAs – are entirely contingent on achieving local as well as national targets for non-response.' [45]

'We are delighted to learn from ONS that early indicators of response to the Census suggest that the targets will be met. If confirmed, this will be a considerable achievement at a time of falling response rates to official enquiries.'

5.59 The full review and the ONS response to the recommendations can be found on the ONS website[44].

5.60 The methodology implemented is summarised below and the corresponding stages are shown graphically in figure 5.2:

(a) As noted in chapter 4, a Census Coverage Survey (CCS) was conducted independently of the 2011 Census. The survey was designed to estimate the under-enumeration (undercount) in the census. A sample of output areas (OAs) was drawn from each local authority in England and Wales, stratified by a hard-to-count (HtC) index. The HtC index was a proxy for non-response in the census. The sample included approximately 17,400 postcodes (around 340,000 households)

(b) The CCS records were matched with those from the census using a combination of automated and clerical matching

(c) A large sample (around 5 per cent) of census individuals were checked to see if they were duplicated within the region and within Wales, and the CCS data were used to help estimate the levels of over-count in the census by broad age-sex groups and region

(d) The undercount was estimated within groups of geographically contiguous (neighbouring) local authorities called estimation areas (EAs) to ensure that CCS sample sizes were of a sufficient size to produce estimates that met the target precision rates above. The matched census and CCS data were used within a dual system estimator (DSE) to estimate the population in the areas sampled in the CCS. The DSEs were then used within a simple ratio estimator to derive population estimates for the whole of the estimation area. As the data were processed, various modifications were made to the DSE and ratio estimation process to ensure that the estimates were robust and to reduce variability where appropriate. This included in some cases collapsing HtC groups, collapsing age-sex groups and removing CCS sample postcodes with no data

(e) The DSEs were assessed for any bias at household level using an alternative household estimate (AHE) from the census field process. The assumption of independence for individuals within households was explored using social survey data

(f) The sample was assessed for balance, which would affect the ratio estimator, using the dummy questionnaire data from the census field process

(g) The population within communal establishments (CEs), which were defined as managed accommodation, was assessed for under-enumeration using both the CCS (for CEs with less than 100 bed spaces) and administrative data and local information (for CEs with more than 100 bed spaces). Adjustments were made to the CE population where these checks highlighted significant undercount

(h) The national population estimates were assessed for quality and plausibility by comparisons with sex ratios from alternative sources

(i) A synthetic estimator (a robust statistical methodology for estimating small areas) was used to estimate the local authority population, using the patterns observed at EA level

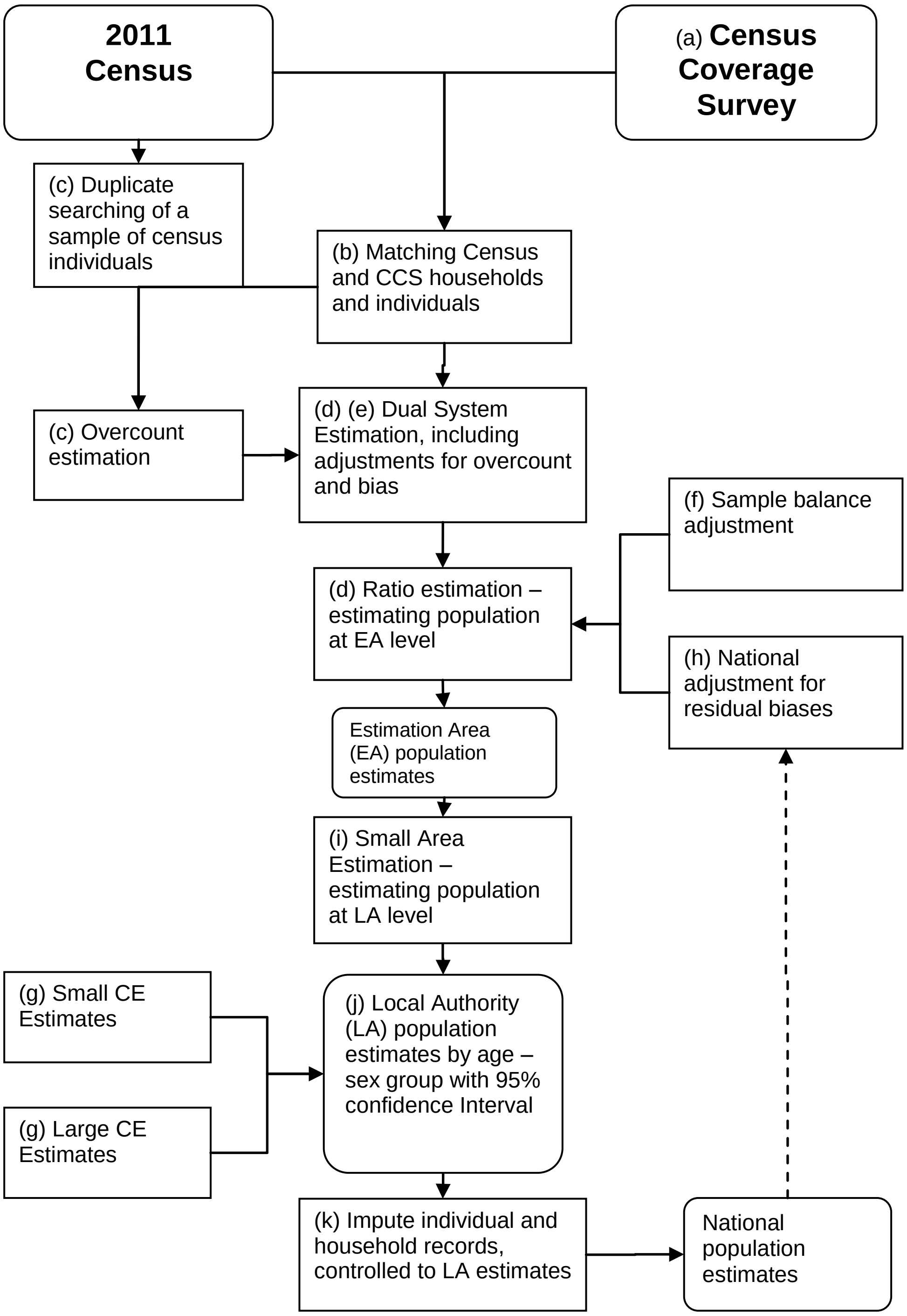
(j) To provide a measure of the variability in the estimates, 95 per cent confidence intervals were calculated for the EA and LA estimates by age-sex group using a ‘bootstrapping’ statistical technique, a re-sampling method for calculating the variance of a complex estimator

(k) Households and individuals estimated to have been missed from the census were imputed onto the census database, after reducing the measured under count by the estimated level of over count. This process copied a subset of characteristics from real households and individuals to create the imputed households, and imputed individuals estimated to have been missed. Information recorded on dummy forms was used to impute households and persons into geographical locations across the whole EA and LA

5.61 The above process from step (c) onwards was iterative. Some of the processes could not be carried out until all data had been processed at least once through the basic estimation process at step (d). For example, the national population estimates could not be assessed until all estimation areas had completed the estimation process. Once the national adjustment had been defined, then all areas were rolled back to the appropriate stage and re-estimated using new parameters.

5.62 Following each iteration of the coverage assessment process, all the population estimates were quality assured using demographic analysis, survey data, qualitative information, administrative data and local information to ensure the estimates were plausible (see paragraphs 5.65 to 5.78) The quality assurance results were examined by a quality assurance panel, which recommended acceptance of the estimate or asked for further work to explore the estimates or the comparator data.

Figure 5.2 Overview of the 2011 Census coverage assessment and adjustment process


2011 Census
(a) Census Coverage Survey
(c) Duplicate searching of a sample of census individuals
(b) Matching Census and CCS households and individuals
(c) Overcount estimation
(d) (e) Dual System Estimation, including adjustments for overcount and bias
(f) Sample balance adjustment
(d) Ratio estimation – estimating population at EA level
(h) National adjustment for residual biases
Estimation Area (EA) population estimates
(i) Small Area Estimation – estimating population at LA level
(g) Small CE Estimates
(j) Local Authority (LA) population estimates by age – sex group with 95% confidence Interval
(g) Large CE Estimates
(k) Impute individual and household records, controlled to LA estimates
National population estimates

5.63 An assessment of the various stages of the 2011 design, together with some conclusions and lessons learned, is given in chapter 10. More information on the different components of the CAA process as shown in figure 5.2 (such as matching rates, CE adjustments, overcount) are available on the ONS website[32].

5.64 The coverage assessment process estimated some 3.79 million people and 2.26 million households were missed in the 2011 Census, and subsequently included in the census outputs database through the adjustment process. This adjustment for usual residents is broken down into its several components (rounded to the nearest thousand in table 5.8 for both the 2011 and 2001 Censuses.

Table 5.8 Components of the census estimates of usual residents, 2011 and 2001, England and Wales

	2011		2001	
Component	**Number**	**%**	**Number**	**%**
Census (enumerated) count	52,639,000	93.9	48,843,000	93.4
Change due to				
Estimation and sample bias	+2,805,000	+5.0	+2,919,000	+5.6
Bias adjustment	+583,000	+1.0	+253,000	+0.5
Over-count adjustment	-352,000	-0.6	0	0
National adjustment	+303,000	+0.2	+27,000	+0.1
Communal establishment adjustment	+98,000	+0.5	0	0
Total changes	**+3,436,000**	**+6.1**	**+3,199,000**	**+6.2**
Total published census estimate	**56,075,900**	**100.0**	**52,042,000**	**n/a**
Adjustment made to mid-year estimate after the census as a result of estimation inaccuracies	n/a	-	+275,000	+0.5
Total census estimate after post-census adjustment	**n/a**	**-**	**52,317,000**	**100.0**

Quality assurance

5.65 Quality assurance procedures were built into all stages of data processing, and the 2011 Census estimates were subject to a rigorous QA process prior to their release. The overall aim was to provide confidence in the estimates by using comparator data sets and by conducting a series of vital checks.

5.66 The QA process was the subject of wide consultation with a variety of stakeholders, including academics, statisticians, demographers and expert census users. The process was designed to:

- ensure 2011 Census estimates were fit for purpose
- use comparator sources to identify discrepancies with census estimates
- use contingencies, where required, to improve census estimates
- ensure census population characteristics were accurate, and
- build user confidence through transparency in the methods

5.67 Key steps in the process were:

- a range of quality assurance panels reviewed estimates at varying levels of detail, including different geographic levels
- a range of evidence was considered, including comparison with administrative data sources

- the quality assurance process checked persons and their key characteristics (for example, students, armed forces, ethnicity)
- estimates of households occupied by usual residents were also quality assured
- identifying issues which were adjusted for in the data processing, and
- further analysis to explain inconsistencies with the comparator data against which census estimates were evaluated

5.68 After the coverage assessment and adjustment step, census population estimates for all 348 local authorities were compared with upper and lower tolerance bounds derived from administrative and survey data sources. These gave a range of plausible values within which census estimates were expected to fall. The tolerance bounds were designed to reflect known differences between alternative sources and census estimates in terms of definitions, accuracy and timing. The main comparator data sets were:

- birth registrations (ONS)
- school census (Department for Education, Welsh Government)
- social security information (Department for Work and Pensions)
- mid-year population estimates (ONS)
- GP NHS patient register (National Health Service)
- census address register (ONS)
- household projections (Department for Communities and Local Government)
- council tax data (Department for Communities and Local Government, Valuation Office Agency)
- local authority supplied council tax data
- integrated household survey (ONS)
- population estimates by ethnic group (ONS)
- migrant worker scan (HM Revenue and Customs)
- short-term migration estimates (ONS)
- students in higher education (Higher Education Statistics Agency)
- further education data (Department for Business, Innovation and Skills, Welsh Government),and
- armed forces data (Defence Analysis Service Agency, United States Air Force)

5.69 While these administrative data sources were used extensively to quality assure the 2011 Census estimates, direct comparisons between these datasets and the census should be treated with caution. This is because there are differences in definitions, recording practices and data quality; and because these datasets were set up for specific administrative purposes they are unlikely to measure the same population. A paper summarising the strengths and limitations of each source in relation to these topics is available on the ONS website[46].

5.70 The following indicators from the census estimates were routinely compared against the comparator data sources:

- age and sex
- household numbers
- household size
- ethnicity
- international migration

- identifying issues which were adjusted for in the data processing, and
- further analysis to explain inconsistencies with the comparator data against which census estimates were evaluated

5.71 After the coverage assessment and adjustment step, census population estimates for all 348 local authorities were compared with upper and lower tolerance bounds derived from administrative and survey data sources. These gave a range of plausible values within which census estimates were expected to fall. The tolerance bounds were designed to reflect known differences between alternative sources and census estimates in terms of definitions, accuracy and timing. The main comparator data sets were:

- short-term UK residents
- students, and
- armed forces

In addition, demographic analyses such as fertility and mortality rates (using census estimates as the denominator), and the ratios of males to females were examined to see if they were in line with historical time series.

5.72 Checks were also developed to validate census estimates by topic against 2001 Census and/or ONS survey data. The broad topic areas covered included:

- demography
- ethnicity, identity, language and religion
- health
- education
- labour market
- travel/transport, and
- households/housing

ONS and other topic experts were periodically invited to review the checks for their specialist subject areas and identify instances where census estimates deviated from expectations. Anomalies were investigated and the process refined as appropriate.

5.73 Other information taken into account during the quality assurance process included:

- operational intelligence compiled from the main census and Census Coverage Survey (CCS) field operations, such as return rates, new addresses identified and addresses deactivated
- local authority intelligence (where provided) such as locally held council tax data, identification of new builds or demolitions in particular small areas, areas or populations which were particularly difficult to enumerate, and
- the profile of the local authority, which included such information as its hard-to-count and multiple deprivation index, and any enumeration challenges identified during the field operation.

5.74 All checks were routinely undertaken at the local authority level. In addition, population and occupied household estimates by LSOA were compared with patient register counts. Data that was significantly out of line with other information (outliers) were explored in detail to ensure discrepancies could be explained and people/households had not been missed. Where comparisons highlighted discrepancies between the census and alternative data sources, more detailed investigations were carried out. This frequently involved drawing on locally provided intelligence, analysis below local authority level, cross referencing with additional data sources, and an assessment of the accuracy of comparators, in particular mid-year population estimates. Where necessary and where data were available at record level, anomalies were resolved by data matching.

5.75 Quality assurance panels were central to the process. Four panels reviewed the evidence provided for all 348 local authorities and made recommendations about whether local authority estimates should be accepted. These were:

- an internal QA steering group which reviewed coverage-adjusted estimates by sex and five-year age group and sex ratios against comparators and tolerance bounds at local authority level
- main QA panels - which included representatives from a range of disciplines within ONS and the Welsh Government. It reviewed estimates for all the checks against comparators and bounds at local authority level, and within each local authority (there were a number of panel groups to cover the high number of estimation and LA areas)
- a high level QA panel which included census/demographic experts and individuals independent of the census process. It also included academic expert membership, an expert former user and representatives from the Welsh Government and devolved administrations (National Records of Scotland, and Northern Ireland Statistics and Research Agency). This panel was responsible for reviewing census estimates for the whole of England and Wales, and separately for the English regions and Wales. It also considered estimates for specific issues and groups raised by the main QA panel on topics such as babies, students, armed forces and international migrants. In addition it reviewed the proposed methods and evidence for making adjustments to census estimates, providing input into and agreement for any methodological changes needed, and
- an executive QA panel was responsible for final agreement to publish the census estimates. This panel included the National Statistician, ONS Director General, ONS executive management and executive management representation from the Welsh Government. This panel considered the England and Wales census population estimates and local authority estimates where inconsistencies with comparator data were greatest. It also reviewed the quality assurance evidence at England and Wales level. This panel was accountable for the final sign-off of the national and local census population estimates ahead of publication

5.76 As expected, there were instances where estimates fell outside the bounds set. For the majority of cases further investigation and analysis was able to explain differences between census and comparator sources. For a small number of cases issues were identified which resulted in adjustments to the data prior to publication. This included the correction of a small number of communal establishments that had been misclassified as households. Some communal establishments had been enumerated correctly but needed to be moved to the correct geographical area in the census data. Single year of age ‘spikes’ (which occurred when a particular donor was used several times during the imputation process), were also identified and resolved.

5.77 This was the final process in agreeing the census population estimates that were published on 16 July 2012. Information from both the CAA and the quality assurance process was published alongside the census population estimates to help users understand the quality of the estimates (such as how much adjustment was applied by area or response rates) and place the estimates in the context of other administrative data sources. This package of supporting information included:

- response rates by local authority and by age and sex
- 95 per cent confidence intervals by local authority
- the size of the household bias adjustment, overcount and CE adjustments
- census estimates against other sources for each local authority, such as patient register, school census and child benefit, but also showing the tolerance bounds for each area, and
- how the estimates were built from their count, quantifying the effects of the various processing steps.

5.78 Overall, the process to quality assure the results has been highly successful and met its key objectives. Most importantly, the methods and data sources used were transparent and gave users confidence in the process and hence the census population estimates. This was a significant improvement on 2001, when the estimates for 15 LAs were adjusted after the census results had been published.

Remaining processing steps

5.79 After quality assurance, the data then went through a number of further processing steps to prepare the data for outputs. These include:

- assigning output geographies, where each person and household record has a number of geographies assigned to it based on the address information collected in the census (such as usual residence, workplace address, second address). These can then be used to allocate the records to any particular output geography, such as output area or workplace zones
- applying statistical disclosure control routines to protect the confidentiality of the standard outputs
- creating derived variables. Some outputs use variables derived from more than one census question; for example, age is derived from date of birth, and distance travelled to work is estimated from the location of the addresses of the place of usual residence and the place of work

5.80 The whole process from data capture to completion of an outputs database took about 18 months – an improvement on the 2001 timetable. The length of time reflects the vast quantity of information to process (24 million household questionnaires with 56 million people) involving some very complex computing such as the CAA and edit and imputation processes. Although lessons have been identified for individual processing steps and methods, the main challenge for a future census (see chapter 11) is not only to maintain similar levels of quality but to complete processing more quickly. Certainly, higher volumes of online completions will help because this significantly reduces some of the lengthy early steps involved in data capture from paper questionnaires; and good design of the online questionnaire will help minimise the level of missing variables and improve the quality of the data recorded.