Scotland's Census 2011
Shaping our future

Estimation and Adjustment Strategy

November 2012

Table of Contents

1. Introduction

1.1 Purpose of this document

1.1.1 This document describes the strategy for estimating the level of undercount in Scotland's Census 2011, and adjusting the final database to reflect this undercount. From this database it will be possible to produce tables for any variable, and at any level of geography, that are consistent with each other and accurately reflect Scotland's population as at census day (27 March 2011). 95% confidence intervals will be calculated and published for the main estimates.

1.1.2 The main part of this document describes the estimation (section 3) and adjustment (section 4) elements of the coverage process. The other elements are described more briefly - those prior to estimation in section 2 and those following adjustment in section 5. Elements of downstream processing other than coverage are outside the scope of this document.

1.2 UK context

1.2.1 The strategy, which builds upon the One Number Census methodology used in 2001, is almost identical to that used in the rest of the UK. The only differences are to reflect variations in the way the census is taken in each country, and other local factors: for instance, in Scotland a six-month residence base was used in census enumeration compared to three months in the rest of the UK; and output areas in Scotland are approximately half the size of those in England, Wales and Northern Ireland.

1.2.2 The systems to carry out the processes have been developed by the Office for National Statistics (ONS). However the actual processing of Scottish data will be carried out by National Records of Scotland (NRS), using ONS's systems adapted as appropriate to handle Scottish data. Methodological support will be provided by ONS.

2. Coverage processes prior to estimation

2.1 Census Coverage Survey

2.1.1 The Census Coverage Survey (CCS) was a follow-up survey that took place over a five-week period beginning six weeks after census day. It differed from the census in a number of respects, the main ones being:

- it surveyed a sample of small areas, covering approximately 1.5% of households in Scotland
- it was conducted by trained interviewers rather than as a self-completion exercise
- only a subset of census questions was asked
- it was not compulsory.

2.1.2 The primary purpose of the CCS was to provide an alternative list of households and residents that could be matched against the census to assess coverage levels. The methodology used requires statistical independence between the census and CCS

(paragraph 3.2.2).To help achieve this, the CCS used a different design and methodology from the census, and census field staff were not permitted to also work on the CCS.

2.1.3 The CCS had a clustered design. Most sampling units, otherwise referred to as clusters, consisted of a single 2001 Output Area (OA), but those OAs whose population had changed significantly in the last 10 years were split or merged as appropriate to maintain a consistent cluster size. The target size for a cluster was approximately 50 households.

2.1.4 The sample was stratified by Local Authority (LA), and by the Hard to Count (HtC) index. The HtC index is calculated at datazone level, which is the level of geography above OAs. Each datazone, and therefore each OA within it, is assigned to one of five levels depending on the predicted difficulty of obtaining a response in the census. It takes into account a number of factors known to affect response rates, including the proportion of students and privately rented dwellings, and the datazone's ranking in the Scottish Index of Multiple Deprivation.

2.1.5 The sampling fraction was higher in those LAs expected to have a poorer census response. Additional sample was also put into those LAs that showed a large variation in response rates between areas in 2001. Within each LA more sample was put into the hardest to count areas.

2.2 Census/CCS matching

2.2.1 The coverage estimation process needs to know, within each estimation stratum, the number of people and households in the census and the CCS, and the number that were counted in both. To achieve this, a matching exercise between the census and CCS will be carried out before estimation begins for each of the ten Processing Units (PUs) into which Scotland has been divided for census purposes.

2.2.2 The exercise will use a combination of automatic and clerical matching in order to achieve the highest possible accuracy rate. The output for each PU will be a list of census people not matched to a census record, a list of CCS people not matched to a census record, and a list of paired census and CCS people where it has been possible to establish that they are the same person.

3. Estimation of the population

3.1 Overview

3.1.1 The process to estimate the number of households and their population consists of three stages: the application of Dual System Estimation (DSE) to estimate the level of undercount in CCS areas; the extension of this to non-sampled areas using ratio estimation; and the use of small area modelling to derive LA totals. A separate, simpler methodology is used to estimate the population of communal establishments.

3.1.2 Once the initial estimates have been calculated, a number of processes are carried out to ensure that the final estimates are as accurate as possible. These take into account known weaknesses in the methodology.

3.2 Stage 1 – Dual System Estimation (DSE)

3.2.1 Dual system estimation is firstly used to estimate the population within the CCS areas. At this stage people are stratified by age and sex, and households by tenure.

3.2.2 The use of DSE requires a number of conditions to be met to ensure the minimisation of error in the estimates. These include:

- statistical independence between the census and CCS - dependence is likely to cause bias in the final estimate. The census and CCS were operationally independent and used different methodologies (section 2.1) in order to help achieve this.
- a closed population - it is assumed that households do not move in between the census and CCS. Clearly this will not be the case, although the CCS was conducted soon enough after census day that the numbers involved should be small.
- homogeneity - within a cluster, the chance of a person being in the census or CCS is assumed to be the same across all people within the stratum. This is a reasonable assumption since clusters are small and contain similar types of people (the OAs on which they are based were designed to be internally homogenous with respect to the population). In addition, the main stratification variables group together people and households who would be expected to have similar response characteristics.
- perfect matching - the matching strategy is designed to achieve a very high level of accuracy.

3.2.3 After matching between the census and the CCS, a 2×2 table of counts of people or households can be derived. This is shown in Table 1.

Table 1 - 2×2 Table of Counts of People (or households)

		CCS		
		Counted	*Missed*	**Total**
Census	*Counted*	n_{11}	n_{10}	$\mathbf{n_{1+}}$
	3.2.4 Missed	n_{01}	n_{00}	$\mathbf{n_{0+}}$
	Total	$\mathbf{n_{+1}}$	$\mathbf{n_{+0}}$	$\mathbf{n_{++}}$

3.2.5 This output from the matching process will be used to estimate the undercount for each of the sampled clusters. Given the assumptions, DSE combines those people counted in the census and/or CCS and estimates those people missed by both using a simple formula to calculate the total population as follows:

$$DSE = n_{++} = \frac{n_{1+} \times n_{+1}}{n_{11}}$$

3.2.6 The formula assumes that the proportion of CCS responders that were also counted in the census is identical to the proportion of CCS non-responders who were in the census (this is the independence assumption). This is equivalent to saying that, assuming independence, the odds of being counted in the CCS among those

counted in the census should be equal to the odds of being counted in the CCS among those not counted in the census.

3.2.7 The standard DSE formula has been shown to be unstable when the numbers are small, as they will be in some cases. In addition, it does not work in cases where no individuals were matched between the census and CCS. A small adjustment, known as the Chapman correction, is therefore made to the formula as follows:

$$DSE = n_{++} = \frac{(n_{1+} + 1) \times (n_{+1} + 1)}{(n_{11} + 1)} - 1$$

3.2.8 For the main estimates, people will be stratified by sex and five-year age group within each cluster. Households will be stratified by tenure and (where applicable) type of landlord. Estimates will also be produced for several other characteristics, including ethnic group and activity last week.

3.2.9 The output from stage 1 of the estimation process will be estimates of the true population of households and people population for the CCS sampled areas, by a number of characteristics.

3.3 Stage 2 – Estimation area estimation

3.3.1 The DSEs calculated in stage 1 only produce population estimates for the areas in the CCS sample. In the second stage a statistical model is used to obtain estimates for the non-sampled areas.

3.3.2 Of the 32 LAs in Scotland, only Glasgow has sufficient CCS sample to allow a direct LA-level estimate with an acceptable level of precision. The remaining LAs will be grouped together at the estimation stage into Estimation Areas (EAs), which will be the main level at which estimation takes place.

3.3.3 The EAs correspond to the Processing Units in which census data was delivered to National Records of Scotland and within which matching will take place. In 2011 Scotland is divided into 10 EAs, each of which (apart from Glasgow, which is an EA in itself) consists of two or more LAs which are geographically contiguous (neighbouring) or as near contiguous as possible for the island areas.

3.3.4 Within each EA, a simple ratio estimator (which uses a straight line of best fit through the origin) will be used to estimate the relationship in the sample between the census count and the DSE. The estimate will be calculated separately for each age-sex group within each HtC stratum. This relationship is then used to estimate the total EA-level population of the age-sex-HtC stratum. The variance of the estimate will be estimated by a standard method called the bootstrap.

3.3.5 The output from this process will be estimates of the population for each EA by sex and five-year age group, together with 95% confidence intervals. A similar methodology will be used to calculate an estimate of the number of households by tenure. All of the subsequent stages described below will be consistent with these estimates.

3.4 Stage 3 – Local Authority Estimation (LAE)

3.4.1 Although the CCS was designed at LA level, practical considerations mean that, with the exception of Glasgow, none contains sufficient CCS sample to enable an accurate direct estimate of the population to be made. Nevertheless, estimates of the population (by age and sex) and household count are needed for each LA. The third stage of the estimation process produces LA-level estimates from the EA-level figures.

3.4.2 A small area estimation technique will be applied to produce LA-level population estimates that have lower variances (i.e. smaller confidence intervals) than those that would be produced by just using the sample specific to each LA. The technique uses information from the whole EA to model coverage within the LAs. It makes the assumption that coverage levels across the LAs are the same within each HtC and age-sex stratum. The resulting population (and household) estimates will then be calibrated to the EA estimates to ensure consistency, and their accuracy can also be calculated to provide confidence intervals around the LA population estimates.

3.5 DSE bias adjustments

3.5.1 The application of the DSE at the cluster level is relatively robust to small violations of the assumptions given at paragraph 3.2.2. However, violation of the assumptions can sometimes result in significantly biased estimates of the population. Experience from 2001 indicates that there is likely to be some residual bias in the DSE due to failure of some of these assumptions.

3.5.2 Bias is most likely to result from failure of the independence and homogeneity assumptions. Although there is no way of distinguishing between these two sources of bias, it is possible to make an overall adjustment accounting for both.

3.5.3 Bias can occur both between and within households. Between-household bias refers to a systematic inaccuracy in the estimated number of households, and by extension the total population (since most people are contained within households).

3.5.4 Within-household bias is a systematic inaccuracy in the estimated number of people in each household, This will produce a biased estimate of the total population, even if the estimated number of households is unbiased.

3.5.5 The methodology to correct for between-household bias involves calculating an aggregate number of households for the CCS areas in each HtC stratum of the EA using information from the fieldwork exercise. For instance, account will be taken of addresses where no return was received despite the enumerator believing it was occupied.

3.5.6 The DSE estimate of households is calibrated to this aggregate figure. Person-level DSE adjustments are obtained by modelling from the household estimates.

3.5.7 Office for National Statistics have conducted an exercise to measure within-household bias for England and Wales. This involved using social surveys to provide an external estimate of within-household coverage, and comparing it to that measured by the CCS. No within-household bias was detected in this exercise, and

since Scotland can be expected to show a similar pattern, no specific adjustment will be made for this type of bias.

3.6 Estimates for other characteristics

3.6.1 In order to control the adjustment system (section 4), estimates of the population other than for the age-sex totals are required. The DSE, ratio and LA estimation methodology outlined above will be used to estimate the population by five year age-sex group, ethnicity, activity last week and tenure. As the total person or household estimate will differ slightly for each variable, each set of estimates will be calibrated so that they are consistent with the age-sex total (for person variables) or the tenure total (for household variables).

3.6.2 A different methodology will be used to estimate the household size distribution. This will use a probability matrix approach to estimate the likelihood that a household enumerated in the census as size X is actually size Y. The probabilities are then used to estimate the true distribution. This will be calibrated to the household and person totals (as it can be used to derive both).

3.7 Estimation for Communal Establishments (CEs)

3.7.1 The CCS collected information from communal establishments (CEs) that have less than 100 bed spaces. As the CCS is not specifically designed to count CEs, and the CE population is much smaller than the household population, relatively few CEs will have been captured by the CCS within each EA. Estimation of this population will use a DSE and ratio methodology, but this will be applied across all CEs in Scotland rather than at EA level. This should provide enough sample to produce robust estimates.

3.7.2 Coverage within CEs is most closely related to the type of CE and the age of the residents. CE estimation is therefore stratified by establishment type and broad age-sex groups, the assumption being that all CEs of a particular type have the same coverage regardless of which LA they fall in.

3.7.3 As noted above, the CCS did not collect information from CEs with 100 or more bed spaces. Coverage within these large CEs (such as prisons and university halls of residence) will be measured using both data collected by the census field staff and specific administrative data. This will lead to specific adjustments for specific CEs where there is evidence to show that the census enumeration failed to count the population (e.g. administrative data shows there should be 200 students in a hall of residence and the census field staff issues 195 questionnaires, but only 100 returns were received).

3.8 Additional adjustments for estimation

3.8.1 NRS will investigate whether an overcount adjustment is necessary and/or feasible.

3.8.2 No national adjustment will be made at this stage to the Scottish data due to the lack of availability of a robust comparator source. Were this to be needed once suitable comparator data becomes available, it would be taken forward as part of the mid year estimate work.programme.

3.9 Methods to use when assumptions fail

3.9.1 In the event of the estimates being considered implausible by either estimation analysts or the quality assurance panel, a number of predefined adjustment and improvement strategies are available. The one likely to be used most often is to collapse strata together (e.g. combining adjacent HtC levels or age-sex groups) where one stratum has a small sample size and/or implausible adjustment level. It is also possible to drop individual clusters from the sample if they are having an undue influence on the final estimates.

3.9.2 Another option is to re-estimate using different strata. For instance, the EA grouping could be changed or the totals could be calibrated to different characteristics (e.g. ethnic group instead of age and sex).

3.9.3 Ultimately, if quality assurance indicates a major failure of the methodology, external data – where it is available and of a suitable quality - can be used to make adjustments.

4. Adjustment

4.1 Overview

4.1.1 Following the production of the population estimates at all levels, the census database will be adjusted to take account of the undercount (and, if applicable, overcount). Synthetic households will be created to allow for those missed by the census, and synthetic people will be imputed into counted households to allow for within-household undercount.

4.1.2 The database thus created will represent our best estimate of the entire population, whether counted by the census or not. This adjusted database will be used to generate all statistical outputs from the census.

4.1.3 The outputs from the estimation process define the number of households (by tenure) and people (by age and sex) to be imputed along with basic information about coverage patterns for some other characteristics. However, it is important that we produce plausible values for all characteristics of those households and people missed by the census. The adjustment process can be summarised in three stages.

4.2 Stage 1 – Modelling characteristics

4.2.1 The first stage of the process is to model the likelihood of households and persons, with their characteristics, being missed from the census. These models use the matched CCS/census data to predict (for example) p, the probability of being counted in the census for a 20-24 year old male who is single, white, and living in a privately rented house in the hardest to count stratum.

4.2.2 The models provide the probabilities of both types of undercount: wholly missed households and persons missed from counted households. The variables that are

included in the models are those which are controlled explicitly by the adjustment process, which must be characteristics collected by the CCS.

4.2.3 These predicted probabilities are then converted into coverage ‘weights’ (by taking the reciprocal of p). These weights will be calibrated precisely to the population estimates at LA level described in section 3.4, as these population estimates are the higher quality benchmark.

4.3 Stage 2 – Creation of missed households and people

4.3.1 The second stage of the process will impute people into counted households, and then impute wholly missed households including the people they contain. The coverage weights will determine the characteristics of the synthetic records.

4.3.2 The weights are allocated to each census person corresponding to the likelihood of persons of that type being missed by the census. The census people are ordered by these weights and cumulative actual and weighted counts calculated. The cumulative counts are compared and, if the weighted count exceeds the unweighted count by more than 0.5, a synthetic person is created with the characteristics of the current person.

4.3.3 The characteristics will be limited to those used by the models and those which need to be controlled. Thus a number of ‘skeleton’ records are created that have certain characteristics. The remaining information will be completed by using the census item imputation system CANCEIS (section 5.2). This ensures the final data is consistent, preserving marginal distributions.

4.3.4 The process for imputing person records is exactly equivalent to that for households.

4.3.5 Census questionnaires received after the beginning of CCS fieldwork are not used in the estimation process, as the presence of CCS interviewers may have affected census response rates in the areas covered.

4.4 Stage 3 – Placement of synthetic households and people

4.4.1 In stage 3 synthetic households (and the people within them) are each placed into a location within the LA, and synthetic people are placed into counted households.

4.4.2 Where possible, synthetic households will be placed into addresses where a household is believed to exist but there is not a complete census record. These include addresses with placeholder forms where the enumerator believed the address was occupied, census questionnaires returned blank, and census questionnaires where only the household questions were completed. The households to be imputed will be compared against the potential locations and scored based on their similarity (including accommodation type and estimated number of residents) to provide the best placement possible.

4.4.3 Where there is no suitable location, a synthetic household record will be created. It will be allocated into a postcode to give it a geographical reference.

4.4.4 The people to be imputed into counted households will be placed into relevant household types: for instance, a baby missed from a four-person household also containing a mother, father and young child would be imputed into an existing three-person household of this type. The relationship information will be modified to ensure consistency.

4.4.5 The final output of the adjustment process is a set of skeleton records to be added to the census database.

5. Coverage processes following adjustment

5.1 Addition of skeleton records to census database

5.1.1 Following adjustment, a further process will add the skeleton records to the counted households and people in the census database.

5.2 Coverage imputation

5.2.1 In order to create a complete and consistent database, all variables must have a plausible value. This is achieved by running the post-adjustment database through the item imputation tool CANCEIS. This treats the skeleton records as if they were real people and households who omitted to answer many of the census questions. It imputes values by copying them from a similar donor record, and applies edit rules to ensure that illegal variable combinations are not created.

5.2.2 Before any results can be made public, a further process called Statistical Disclosure Control (SDC) is applied to ensure that no personal data is disclosed in the published figures. This process is outside the scope of this document.

5.2.3 The final output is an individual-level database that represents the best estimate of what would have been collected had the 2011 Census not been subject to undercount (and possibly overcount). Tabulations derived from this database will automatically include compensation for such enumeration issues for all variables and all levels of geography, and will be consistent with the census estimates.

6. Further information

More detail on the coverage methodology for the 2011 Census can be found on the Office for National Statistics website.