2011 Census

Release 1B - How the 2011 Census population estimates were obtained

March 2013

Contents

1. Introduction 3
2. The estimation process 4
2.1 Overview 4
2.2 The Census Coverage Survey 6
2.3 Matching the CCS to the Census 6
2.4 The Hard-to-Count index 7
2.5 Census/CCS linkage 7
2.6 Dual System Estimation 8
2.7 Tuning the estimation 8
3. Further adjustments 8
3.1 Dual system estimation bias adjustment 9
3.2 Overcount 10
3.3 Communal Establishments 11
3.4 Sample Balance 13
3.5 National Adjustment 13
4. Confidence intervals 13

How the 2011 population estimates were obtained – Release 1B

1. Introduction

Scotland's Census took place on Sunday 27 March 2011 with the purpose of providing an accurate population count. Although every effort is made to ensure everyone is included in the census, inevitably some individuals are missed. This under-counting does not usually occur uniformly across all geographical areas or across other sub-groups (for example, by age and sex) of the population.

To fill the gap, the National Records for Scotland (NRS) implemented a coverage assessment process to estimate the population that was missed. In addition, this process identified and adjusted for the people who were counted more than once or who were counted in the wrong place as well as correcting for underpinning assumptions that were not realised in practice. Carrying out this work allowed a census estimate of the entire population to be obtained. This paper outlines the key stages in the estimation process and provides links to other relevant published papers.

The estimation process was based on a number of assumptions and at each stage these assumptions were checked and adjustments made when necessary, based on evidence, to ensure that the estimates are of the highest quality. This paper describes these adjustments and outlines how they fit together within the estimation framework. The adjustments were all modifications to the basic estimation process. Each was intended to address forms of bias in the estimates that resulted from the underpinning assumptions not being realised in practice.

The methods were largely based on those developed by the Office for National Statistics (ONS). The ONS systems were also implemented although adapted as necessary to cope with Scotland specific data. ONS have produced a full suite of methodology papers detailing the statistical theory and practical application of the methodology. They can be found here: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/methods/coverage-assessment-and-adjustment-methods/index.html

It was not always practical or appropriate to replicate exactly what was done for the rest of the UK due to differences in fieldwork processes, data capture and processing and also the availability of comparator data sources. The ONS documentation should be read bearing in mind there were small differences between Scotland and the rest of the UK[1].

This document provides an overview of the estimation and adjustment process used to produce census population estimates for Scotland. Divergence from ONS methodology is highlighted where necessary.

[1] The main differences between Scotland and the rest of the UK were that:
- Scottish fieldwork relied mostly on hand delivery of questionnaires with post out to more remote areas only; ONS implemented a full post out delivery method;
- Scotland employed different contractors to undertake certain aspects of the census. This means that data capture and processing solutions will inevitably vary between countries;

2. The estimation process

2.1 Overview

The census coverage assessment and adjustment methodology used is summarised below. The key stages are shown in Figure 1.

a) A Census Coverage Survey (CCS) was undertaken, independently of the 2011 Census. The survey was designed to estimate the under-enumeration (undercount) in the census. A sample of postcodes was drawn from each local authority, stratified by a hard to count (HTC) index. The HTC index was a proxy for non-response in the census. The CCS in Scotland included around 40,000 households.

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

c) All census individuals were deterministically matched see if they (the individuals) were duplicated within Scotland, and the CCS data were used to help estimate the levels of overcount in the census by broad age-sex groups.

d) The undercount was estimated within groups of geographically neighbouring local authorities (called processing units) to ensure that CCS sample sizes were adequate. 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 to derive population estimates for the whole of the processing unit.

e) The DSEs were assessed for any bias at household level using an alternative household estimate (AHE) from the census field process.

f) The sample was assessed for balance, which would affect the ratio estimator, using the placeholder data from the census field process. No extreme samples were detected.

g) The population within communal establishments (CEs), which were defined as managed accommodation, was assessed for under-coverage using both the CCS (for small communal establishments) and administrative data and local information (for large communal establishments). Adjustments were made to the communal establishment population where these checks highlighted significant undercount.

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

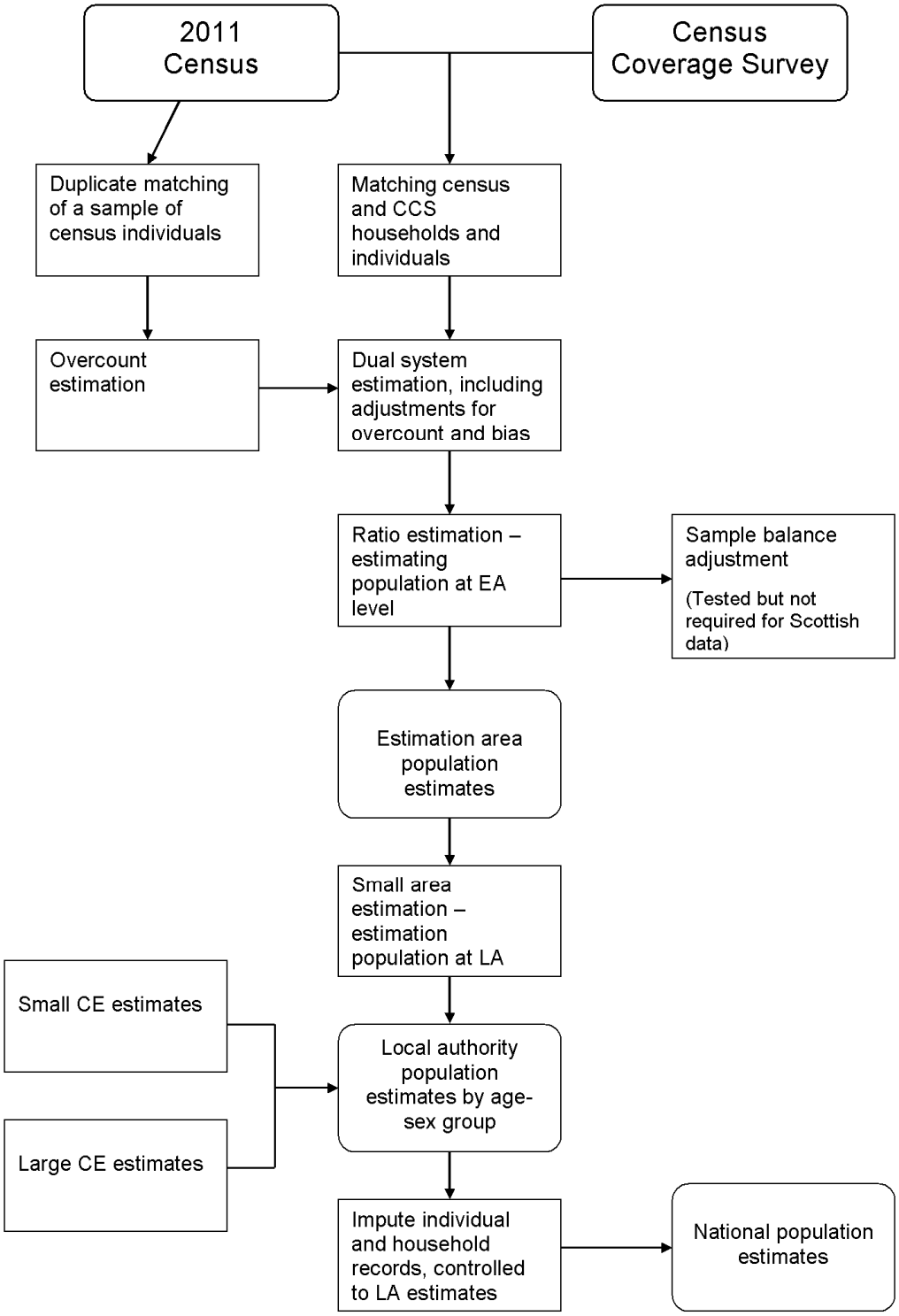
i) To provide a measure of variability in the estimates, 95 per cent confidence intervals were calculated for the estimation area and local authority population estimates by age-sex group using a bootstrapping technique.

j) Households and individuals estimated to have been missed from the census were imputed on to the census database, after reducing the measured undercount by the estimated level of overcount. 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. The households and persons were imputed into geographical locations across the whole estimation area and local authority.

After 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.

Figure 1 – The 2011 coverage assessment and adjustment process overview



2.2 The Census Coverage Survey

The primary source of 'missingness' was addressed using a **Census Coverage Survey** (CCS) in which about 40,000 households across Scotland were visited by trained survey interviewers. The CCS was a sample survey independent to the census, conducted six weeks after census day. Participation in it was voluntary and it provides an alternative list of households and residents which can be matched with census returns.

The sample was designed to ensure it was spread across all local authorities with a clustered, stratified sample of postcodes drawn from each processing unit[2] (PU). The sample was stratified according to a Hard to Count (HtC) index. The HtC index is calculated at datazone level, which is the level of geography above Output Areas (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.

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.

To minimise interview time, the questionnaire contained only a set of key social and demographic questions from the census as well as some additional questions, to maximise coverage, on household members who were likely to be counted elsewhere.

The overall achieved response rate[3] from the CCS was 87 per cent. The Excel document 2011 Census response rate and sample size.xls shows the response rates for each processing unit and council area in Scotland. This is available to download as a PDF.

A fuller explanation of the methodology behind the CCS is given in the ONS paper: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/census-coverage-survey-summary.pdf

2.3 Matching the CCS to the Census

The Census Coverage Survey records were probabilistically matched with those from the census using a combination of automated and clerical matching. The individuals were matched first, and then the households. Any discrepancies in the households, such as a household splitting or reforming between the census and CCS, were fully investigated and resolved in a reconciliation phase which followed the automatic and clerical matching phases.

[2] A processing unit (PU) is made up of one or more neighbouring council areas (CAs) and consisted of approximately 500,000 respondents. CAs were grouped in to PUs for practicalities around data processing etc. In ONS papers the term Delivery Group (DG) is used instead of PU.
[3] Response rate is defined as the number of valid responses achieved divided by the number of occupied households found by either the census or the CCS.

The characteristics used for matching are:

- First name
- Surname
- Date of birth
- Sex
- Address

2.4 The Hard-to-Count index

The Hard-to-Count index was derived in advance of the census, in order to facilitate the sampling strategy to be used for the Census Coverage Survey. It was derived by combining a range of datazone level indicators that were as up-to-date as possible. This ensured that there was minimal dependence on 2001 Census results. The Hard-to-Count index for 2011 included the following factors:

- % f/t students (aged 18-24) in Higher Education
- level of deprivation
- % privately rented accommodation
- % school children with English as second home language
- % occupied dwellings classed as flats

These variables were standardised and ranked, to produce a Hard-to-Count index. Finally, the index was classified into 5 categories according to a 2%, 8%, 10%, 40%, 40% (hardest – easiest) distribution.

2.5 Census/CCS linkage

The main purpose of the Census Coverage Survey was to establish a measure of the underenumeration in the census. This was achieved by closely monitoring the patterns of enumeration in specially sampled areas, as detailed in the CCS section above. Results from the census enumeration were compared with those from the CCS using data linkage methods including deterministic and probabilistic matching, as well as using household structure, address information, detailed investigation and searching at postcode level. A system was specifically designed to incorporate the most efficient combination of automated and clerical methods, thus ensuring an accurate and comprehensive linkage.

The linkage therefore provided information on results that were:

- confirmed by the CCS (linked)
- in the Census only (not linked - missed by the CCS)
- in the CCS (not linked - missed by the census)

Across Scotland, around nearly around 94% of the persons counted in CCS were linked, and therefore identified as being present in the census enumeration.

2.6 Dual System Estimation

Having matched responses from both the census and the CCS, the known number of responses from each of these were used in a statistical technique called Dual System Estimation (DSE) to determine the number of respondents missed for each of the sample areas in the CCS. Dual system estimation works on the basis of capture/recapture. For an explanation of how the dual system estimation process works see http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/census-coverage-survey/trout--catfish-and-roach---the-beginner-s-guide-to-census-population-estimates.pdf.

Further details of how this is worked out can be found in: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/coverage-assessment-and-adjustment-process.pdf

2.7 Tuning the estimation

Where there was insufficient data available to provide robust estimates for a particular group of people (e.g. within an age-sex or ethnicity group), these were collapsed with another group as appropriate. The same collapsing rules and guidelines as ONS were used to ensure consistency between both geographical areas within Scotland as the rest of the UK.

Where no data was collected for a particular postcode within the sample or the data collected was thought to be unrepresentative, it was removed from the sample as leaving it in would bias the results. A total of 20 postcodes were removed from the initial sample of 2153 postcodes.

For full details on the collapsing of categories and postcode removal within Scotland, see the Excel document: Tuning 2011 census coverage estimation.xls. This is available to download as a PDF.

3. Further adjustments

The estimation process was based on a number of assumptions:

- the census and the CCS were independent and capture probabilities homogeneous
- there was no overcount in the census
- the sample was representative and balanced with respect to coverage
- there was no overcount in communal establishments[4]
- the national and local estimates were plausible

It was not always possible to meet the assumptions of the estimation process for practical reasons. Therefore, these assumptions were checked and, where they weren't met, necessary adjustments were made to the estimation process to account for any resultant bias.

[4] A communal establishment is a managed residential accommodation with ten or more bed spaces.

3.1 Dual system estimation bias adjustment

The DSE method makes the assumption of independence between the CCS and census so an individual's likelihood to respond to the CCS is not influenced by how they responded to the census. However, this assumption will not always hold due to the fact that i) households that are less willing and likely to respond to the census will also be less willing and likely to respond to the CCS; and ii) a household's likelihood of responding to one may be affected by whether or not it responded to the other. This assumption applied to both the process of counting households and individuals within households – these were assessed separately. The violation of the assumption of independence will cause bias in the estimates. The two main types of bias are between household bias and within household bias.

An example of between-household bias would be when a household would respond to neither the census nor the CCS. To help overcome this bias an alternative count of occupied households was calculated, based primarily on information gathered during the census field operation. These alternative household estimates (AHE) were calculated for the CCS postcode clusters within each hard to count stratum of each PU.

The AHE were constructed using the following components:

- Returned questionnaires from households containing usual residents
- Proportion of households where a placeholder form was completed

It should be noted that ONS considered further elements for the AHE that were thought not to be relevant for Scotland[5].

For some placeholders, misclassification was identified for the following reasons:

- placeholder indicated an address was occupied, but the returned questionnaire showed no usual residents
- placeholder indicated an address was unoccupied, but the questionnaire had been completed by its usual residents
- placeholder indicated that no contact was made at an address, and the returned questionnaire showed no usual residents

The proportions of these misclassifications were calculated for each PU to determine how many placeholder questionnaires to include in the AHE. For the whole of Scotland 78% of occupied, 19% of unoccupied and 91% of non-contact placeholder forms were estimated to be occupied.

These estimates then provided a measure of bias at household level which was used to estimate the bias at an individual level. The alternative count was fed in to the ratio estimation and was used (if necessary) as an inflation factor for the DSE that calculated people missed by both the census and the CCS.

[5] 1. Due to the hand delivery of the majority of questionnaires in Scotland, it was felt that additional addresses were more easily identified during fieldwork and so such an adjustment would be too small to include; 2. differences in processing meant it was not possible to include households who returned a blank questionnaire as these were not retained; and 3. a separate field exercise was not undertaken to quantify unaccounted for addresses.

The rounded AHE for every hard to count group in every PU can be found in the spreadsheet: 2011 census alternative household estimate.xls. This is available to download as a PDF.

An example of within-household bias would be when a person would always be excluded from a counted household in both the census and CCS. Both ONS and NISRA (Northern Ireland Statistics and Research Agency) undertook exercises to link census data with social survey data to test for within-household bias and found no evidence from this method that such bias existed. As a result, no adjustment was made for England, Wales and Northern Ireland. Due to the amount of work involved in such a linking exercise for Scotland, combined with the assumption that it would be highly unlikely to show a different result from ONS and NISRA, no such adjustment was made for Scottish data.

For further information on bias adjustments, see the ONS papers:
http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/within-household-bias-adjustment.pdf

http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/household-bias-adjustment.pdf

3.2 Overcount

Particular groups of people may be included on one or more census questionnaires due to the following:

- duplicate returns at the same location (for example, an internet and paper return for the same household) (type 1)
- duplicate returns from different locations (for example students who completed a questionnaire at their term-time address but were also included in the one returned by their parents for the family home) (type 2)
- individuals being counted in the wrong location (for example, a student which is incorrectly included on their parents household questionnaire and did not return one for their term-time address) (type 3)
- erroneous returns (type 4)

Multiple responses for the same household (type 1) were removed during the early stages of processing. Overcount from erroneous returns (type 4) is not able to be determined without further fieldwork. Therefore, an overcount adjustment focussed on duplicate returns from different locations (type 2) and individuals being counted at the wrong location (type 3).

Due to the relatively small size of the Scottish population, it was possible to undertake a full overcount matching exercise i.e. matching the entire census database to both the CCS and itself. Type 3 estimates were calibrated with results from type 2 (census-census). The calibration component identified cases of type 2 overcount in the census-census

match, and produced a multiplier based on a ratio of type 2 overcount estimates from both matching exercises.

It was not feasible to impose probabilistic matching with clerical checking given the size of such a database. The identification of duplicates therefore relied solely upon deterministic matching, accepting only exact matches i.e. where names and dates of births were exactly the same. This then resulted in conservative 'dampening' factors for broad population groups which modified the DSE and adjusted the census estimate.
For more background on overcount see the ONS paper: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/overcount-estimation-and-adjustment.pdf

3.3 Communal Establishments

Communal establishments (CEs) are managed residential accommodation with ten or more bed spaces. Examples include student halls of residence, care homes, prisons, boarding schools and homeless hostels. Dedicated census takers were employed to enumerate CEs and provided each CE with a questionnaire about the establishment as a whole, to be completed by the manager, and individual person questionnaires, to be completed by each resident.

For the purpose of the census these were split in to 'small' (those with between 10 and 99 bed spaces) and 'large' (those with 100 or more bed spaces). Small CEs were within the scope of the CCS, although the sample design did not account for them and therefore there was no control over the size of the small CE sample. Large CEs were not within the scope of the CCS and a separate piece of work was undertaken to investigate the likely number of persons missed in the census.

Small (CEs) were treated in the same way as individual respondents i.e. using a DSE to determine the missing people (this method measured coverage within CEs but not coverage of CEs). Due to the sample size, it was not possible to produce DSEs at PU or CA level and so this exercise was undertaken for Scotland as a whole. For similar reasons, extensive collapsing was required across age and sex groups as well as for establishment types. These were:

- males and females aged 0 to 59
- males and females aged 60+
- all establishment types

A total of 2,178 additional persons were added in to small CEs.

For large communal establishments checks were undertaken to identify those which required further investigation. These checks were:

Response rate 1: Percentage returns less than 75% (CE capacity[6] as the denominator)
Response rate 2: Percentage returns less than 75% (Active forms[7] as the denominator)
Actual non-response: Active forms – Returns > 50

[6] As recorded on the CE address register as collated by NRS geography department.
[7] The number of active forms is the number of individual questionnaires issued by the CE manager as recorded on the CE questionnaire.

Comparison with alternative sources: if the census count was low in comparison

For CEs that failed one or more of the above checks, the number of missing residents by gender and 5 year age group were identified using appropriate comparator data[8]. In cases where the comparisons between the active forms, census returns and comparator count were inconclusive, the establishment was contacted to confirm the number of people usually resident in the CE on census night and the adjustment was made in line with the information given. This applied to age and sex distributions only, other characteristics were imputed.

It should be noted that no adjustments were made for hospitals and travel and leisure establishments irrespective of the outcome of the above checks. These tend to have few usual residents according to census definitions and also no suitable comparator sources were obtained on which to base adjustments.

Of the 268 large CEs, 84 were considered for further investigation by one or more of the above checks.

The table below shows the adjustments made for large CEs by establishment type.

Table 1 - Adjustment of usual residents in large CEs by establishment type

Estnature	Type of CE	Number of large CEs	Number that fail checks	Population Adjustment
1	General hospital	35	NA	NA
2	Psychiatric hospital	14	NA	NA
3	Other hospital	12	NA	NA
5	Care home with nursing	23	6	142
8	Other medical and care establishment	1	NA	NA
9	School	10	1	0
10	Halls of residence	136	53	2381
11	Other educational establishment	3	3	0
13	Prison	15	8	1413
15	Immigration/Detention/Asylum centre	1	0	NA
17	Hotel, guest house, B&B, youth hostel	2	NA	NA
19	Hostel or shelter for homeless	2	NA	NA
21	Religious establishment	1	NA	NA
23	Other establishment	2	2	0
24	Armed forces bases	10	10	0
25	Other armed forces establishment	1	1	0

[8] HESA data for single year of age and sex for the academic year 2010-2011 was provided by the Scottish Funding Council for students; the prison statistics branch in the Scottish Government provided a count of prisoners on census day by single year of age, sex and length of sentence.

Total		**268**	**84**	**3936**

3.4 Sample Balance

The CCS sample would be expected to be an accurate representation of the overall population for the sampled areas. However, with every sampling process there is a risk that a sample may be an outlier amongst all possible samples. For example, the chosen CCS sample could have, by chance, drawn postcodes in an area where the census had managed to count everyone.

The number of placeholder questionnaires was used to evaluate whether any of the CCS samples were outliers i.e. 'unbalanced'. These data were believed to be the best possible proxy for coverage as they represented households from which a return was not received. The placeholder response rates within sample areas were compared to the corresponding response rates across the whole processing unit and if the two were significantly different the sample was an outlier. If this was found to be the case an adjustment would be made based on the ratio of placeholder coverage to the equivalent for the sample.

However, when this work was undertaken for the Scottish sample, it showed that no adjustments were required to correct for sample balance.

The ONS paper provides in-depth methodology and theoretical information as well as quantifying the adjustments made within England and Wales: http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/ccs-sample-balance-adjustment.pdf

3.5 National Adjustment

ONS assessed the national census estimates for England and Wales against alternative demographic sources for plausibility. This evidence showed that, for 20-49 year olds, the number of females was plausible but the number of males was not. Sex ratios from alternative sources strengthened this evidence and a target sex ratio was derived from the Longitudinal Study and used to define an adjustment to the male age group census estimates.

Due to a lack of suitable alternative sources with which to compare the Scottish census estimates it was not possible to prove that such a problem with sex ratios existed or to correct for these if they did. Evidence from the 2001 Census implies that this may not actually be as large a problem for Scotland as other parts of the UK anyway. We shall look at this within the context of our ongoing work programme.

4. Confidence intervals

The population figures produced by NRS are estimates. A basic requirement of any estimate is a measure of its precision or uncertainty.

A 95 per cent confidence interval, which provides a measure of accuracy, can be interpreted as the interval within which the true value being estimated will lie 95 per cent of the time if the sample was repeated a number of times.

The census estimate of the total Scotland population (5,295,000) had a 95 per cent confidence interval width of plus or minus 0.44 per cent (i.e. plus or minus 23,000 people). The confidence intervals around the census estimates for each Local Authority can be found in the spreadsheet: 2011 Census Confidence intervals.xls. This is available to download as a PDF.

A bootstrap methodology (Efron & Tibshirani, 1993) was used to estimate the variance of the census population estimates. This method was developed to estimate the variance and confidence intervals of complex estimators associated with multistage survey samples. An advantage of the bootstrap method is that it provides an easily understood methodology for the non-technical user.

The method draws a large number of bootstrap sample replicates, sampling with replacement from the observed sample, using exactly the same sample design. Each bootstrap replicate therefore has the same sample size as the original sample, but could (due to random chance) contain only data from one sample point (because of sampling with replacement). Each replicate is then used to construct an estimate using the usual estimation process. The result is a series of estimates of the population across the replicates. The empirical variance can then be taken from the distribution of those estimates.

For further information on the calculation of confidence intervals see:

http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-data/2011-first-release/first-release--quality-assurance-and-methodology-papers/confidence-intervals-for-the-2011-census.pdf.