Scotland's Census

Quality assurance and dealing with non-response in the Census

Cecilia Macintyre and Ali Greig

June 25th 2014

National Records of Scotland

Scotland's Census 2011
Shaping our future

Overview

▶ Quality assurance approach

▶ Documentation of quality assurance

▶ The Estimation System in Census and its Accuracy

Development of methods

▶ Carried out agreed series of simple univariate checks at early stages.

▶ Benefits of early sight of data was that feedback could be provided to processing team

▶ Developed systems and tools to be used throughout process and for dissemination of quality information

What did we do about it?

▶ Carried out more in-depth checks , prioritising key data used in first release

▶ Analysed data for issues which would cause problems in later processes, in particular edit and imputation

▶ Recoded some text responses including ethnic group and language

▶ Sometimes nothing – but will need to report quality to users

QA panels

▶ Met with internal quality assurance working group to discuss approach to quality assurance

▶ External panel
  - provided knowledge and comparator data
  - provide a source of local contact
 - provide insights to NRS on final results

Metadata available online

Quality Assurance Pack

▶ To accompany the first release of population and household statistics, NRS published detailed data used in the quality assurance process

▶ The following slides are extracts from the pack

All persons - West Lothian

Age	2011 Census Counts[1]	2011 Census Estimates[1]	Rolled forward estimates[1]	NHSCR 2011[2]	Child Benefits 2011[3]	Super Older Persons 2011[4]	Response Rates
0-4	10,500	11,600	11,300	11,100	11,100		90%
5-9	9,800	10,700	10,600	10,700	10,500		92%
10-14	10,200	10,800	10,700	10,800	10,700		94%
15-19	10,200	11,100	11,100	11,100			92%
20-24	9,200	10,400	11,000	11,000			89%
25-29	9,500	10,800	10,500	11,400			89%
30-34	10,000	10,900	10,200	11,800			92%
35-39	12,000	12,900	11,900	13,400			93%
40-44	14,200	14,700	14,600	15,400			96%
45-49	13,900	14,500	14,000	14,700			96%
50-54	11,900	12,300	12,300	12,300			97%
55-59	10,100	10,300	10,400	10,200			98%
60-64	10,100	10,400	10,200	10,100			98%
65-69	7,800	8,000	8,000	7,700		7,800	97%
70-74	6,100	6,300	6,300	6,100		6,300	97%
75-79	4,400	4,600	4,600	4,400		4,500	97%
80+	4,800	4,900	5,100	4,700		4,800	97%
Total	**164,700**	**175,100**	**172,900**	**177,000**	**32,300**	**23,300**	**94%**

Source: National Records of Scotland[1], National Health Service[2], HM Revenue and Customs[3], Department for Work and Pensions[4]

Descriptive metadata for tables in this pack is provided in the Explanatory Materials tab

---

All persons - West Lothian
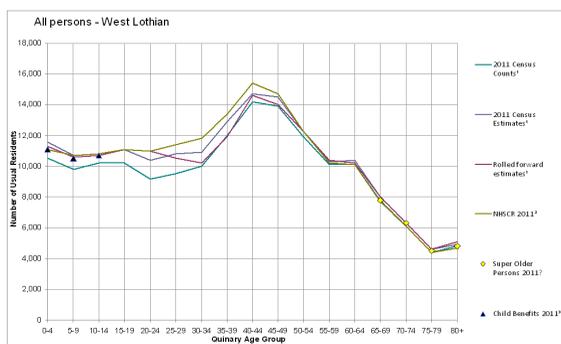
Quality Topic Report Format

▶ 1: Questions & Variables Covered
▶ 2: Tracking Missing Data
▶ 3: Data Changes through process
▶ 4: Internal Analysis
▶ 5: External Analysis
▶ 6: Known Quality Issues (may only be relevant for some variables)
▶ 7: Definitions and references
▶ 8: Documentation

Current work and next steps

▶ Quality assurance of migration and workplace flow data

▶ Investigation of issues arisen following publications

   Impact of approach to dealing with overlapping areas

   Use of microdata to investigate household compositions

▶ Planning for documentation and quality products, QA papers, enhanced metadata, item level imputation rates and deterministic edit rates

---

Further information

▶ All data available at: www.scotlandscensus.gov.uk
▶ Also sign up there for our e-newsletter

▶ Media enquiries: 2011Comms@gro-scotland.gsi.gov.uk
▶ General enquiries: Customer@gro-scotland.gsi.gov.uk

Questions?

The Estimation System in Census and its Accuracy
A Quick Guide

Quick Question

Does anyone know the census estimation methodology?

Fundamentals of Estimation System

▸ **Key goal:** estimate census <u>non-response</u>.
  Quantify the number of people that did not complete a census questionnaire.

▸ This is primarily achieved through a Census Coverage Survey
  1.5% sample of Scottish postcodes
  Stratified two-stage cluster sampling

Estimation Modelling Framework

Capture-Recapture Modelling

▸ Using the CCS and census, the probability individuals and households were missed on the census can be estimated for different groups.

▸ This is used to estimate the true population for CCS areas. These are then used to derive weights which are then applied nationally.

Key methodological issues

▸ Although 1.5% is a relatively big sample, standard sampling issues apply to the CCS. And standard issues around questionnaire design apply to both the CCS and census.

▸ **Independence of CCS and Census**
  The probability of not responding to the census and not responding to the CCS need to be independent.
  e.g., if, in an extreme example, a particular group does not respond to either the census or CCS – we will be unable to estimate the probability that individuals from that group are missed on the census.
  Some remedial steps are taken.

How accurate is the estimation system?

Theoretically complex to estimate.

▸ Sample size of CCS / Stratification
▸ Independence assumption
▸ Edit & Imputation assumptions
  Based on nearest neighbour algorithm
▸ Data processing and other adjustments
▸ Symmetric sampling distributions
▸ Note on small area population estimates

Different ways to estimate accuracy

▸ Imputation / Non-response rates

▸ Bootstrapping to derive confidence intervals

▸ Theoretical development of confidence intervals.

Imputation Response Rates

▸ The number of values/people which are synthetic.
▸ Imputation rates (or response rates) are a very useful indication of data quality and easy to interpret.
▸ But… the indicator is not an error rate (i.e., variables with higher imputation rates are not necessarily the most inaccurate).

Bootstrap methodology

▸ Basically, 'shuffle' CCS responses within each strata across Scotland.
▸ Estimation system is rerun with 'new' PU-level data, and new estimates generated.
▸ Provides an indicator which can be interpreted as a confidence interval.
▸ Not really a conventional confidence interval (i.e., it asks "what variance is expected from the estimation system?").

Theoretical Confidence Interval

▸ Produce a confidence interval using statistical theory.
▸ We tried this using a Bayesian Approach.
▸ Can investigate independence assumption, and produce consistent confidence interval less reliant on responses.
▸ … but… depends on the extent of dependence
▸ … and this is difficult to measure.

Current thoughts

▸ Best approach would likely be a combination of bootstrap or theoretical approach and imputation rates, although theoretical approaches are useful in learning about estimation system.

▸ **General conclusion**: Demographic subpopulations with relatively larger numbers of responses (in CCS areas) will produce good data.

Further information

▸ All data available at: www.scotlandscensus.gov.uk
▸ Also sign up there for our e-newsletter

▸ Media enquiries: 2011Comms@gro-scotland.gsi.gov.uk
▸ General enquiries: Customer@gro-scotland.gsi.gov.uk

Questions?

Contact Details

▶ Cecilia.Macintyre@gro-scotland.gsi.gov.uk

▶ Alastair.Greig@gro-scotland.gsi.gov.uk

Scotland's
Census 2011
Shaping our future