Scotland's Census

Downstream Processing Operational Outline

Head of Downstream Processing Unit
November 2012

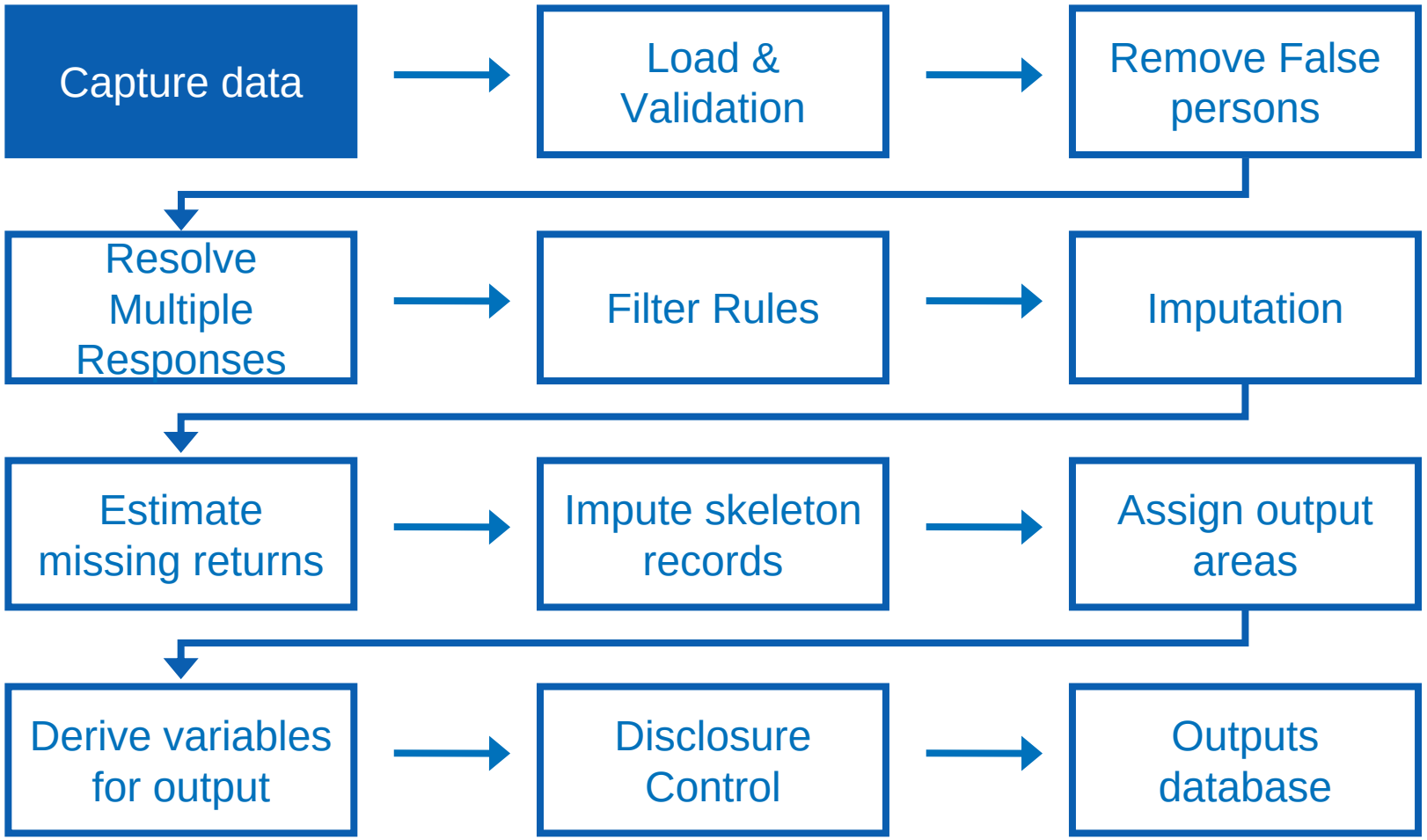
National Records of Scotland

Scotland's Census 2011
Shaping our future

Overview

- Census taken on 27 March 2011
- Roughly 80% paper returns, 20% internet.
- To arrive at a population figure we:
  - Capture and clean the data
  - Impute missing characteristics
  - Estimate the returns we didn't get
  - Derive variables for output
  - Assign output areas
  - Apply Disclosure Controls

Flow of data


Capture data
Load & Validation
Remove False persons
Resolve Multiple Responses
Filter Rules
Imputation
Estimate missing returns
Impute skeleton records
Assign output areas
Derive variables for output
Disclosure Control
Outputs database

Development of methods

- Developed in close consultation with Office for National Statistics (ONS), Welsh Assembly Government (WAG) and Northern Ireland Statistics and Research Agency (NISRA)
- Allows harmonised outputs
- Implementation by National Records of Scotland (NRS), but making use of ONS algorithms and code where possible
- Benefits & Issues

Capture and Coding

- Scanning / Operators
  - All tick boxes and text fields captured as text
  - Questionnaires guillotined and scanned
  - Hundreds of operators
  - Questionable fields flagged to operators
  - Quality assurance samples drawn and checked

Data Cleaning – Initial Validation

- Load and Validation – right types of values/ranges etc
  - Check data received as expected
  - Load into Small Area Statistics (SAS) database
  - Referential integrity
  - Range checks

- Remove false Persons – (2 of 6 rule)
  - Occur due to: crossings out/mistakes or dust on scanner
  - Reject person records without a response to at least 2 of:
    - name
    - sex
    - marital/civil partnership status
    - date of birth

Data Cleaning – Multiple Responses

- Can occur due to:
  - Internet & paper returns from same household
  - Two paper returns from same household
  - person filling in details twice
  - person on both household and individual forms
- Identify which case then
  - Decide which is ‘best’ response (rules)
  - merge data where appropriate

Data Cleaning – Filter rules

- Not everyone should answer every question, e.g.own accommodation (skip landlord question), born in UK (skip date of arrival) under 16 (skip employment questions)
- Resolve inconsistent responses
- Deterministic
- Which response do we believe?

Imputation (1)

- Some records have missing/inconsistent data
- Probabilistic approach
- Missing and inconsistent responses
- Requires complex relationships between members of the household to be analysed – triangulation of relationships

Imputation (2)

- CANCEIS – Canadian Census Edit and Imputation Software
- Donor imputation
- Minimum change
- Decision Logic Tables (DLT)
- Deterministic edits?

Coverage matching and estimation

- Missing households and people
- Census Coverage Survey (CCS)
- Match Census and CCS records - automatic and clerical
- Dual systems estimation
- Regression estimator
- Age-sex groups by local authority
- Overcount?
- Estimates quality assured against admin sources

Coverage adjustment

- Produce consistent individual level database
- Add missed households and individuals
- Use known gaps where possible
- Maintain consistency with surrounding area
- ‘Skeleton records’

Post-Coverage Imputation

- We need to fill out realistic characteristics for the skeleton records
- Again using Canadian Census Edit and Imputation Software (CANCEIS) and preserving variable distributions

Derive complex variables

- Remaining variables for outputs, e.g.
  - household composition algorithm
  - dwellings
  - occupation
  - industry

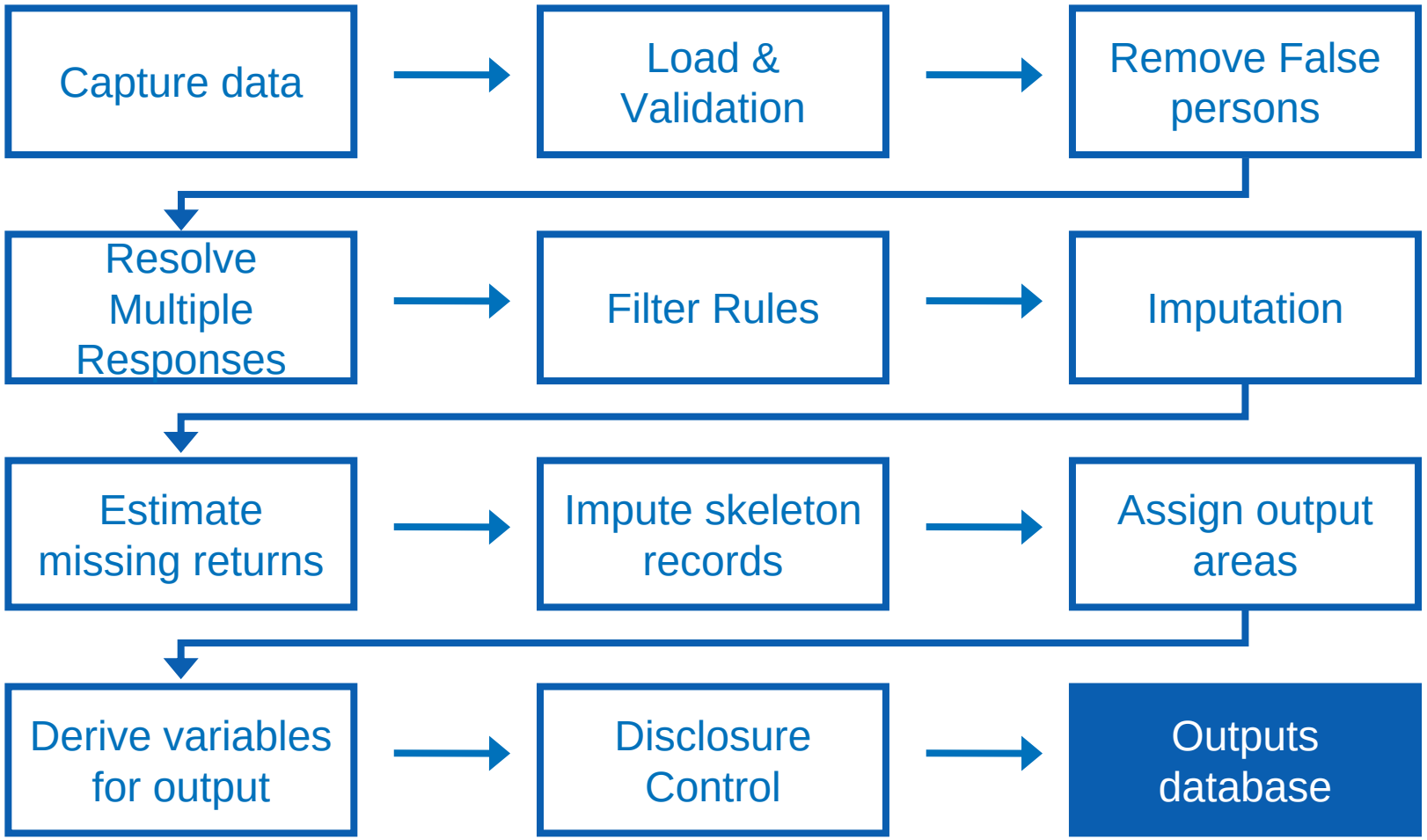
Output area creation

- Lowest geographical level of unrestricted data release
- Working on a principle of minimum change from 2001
- Working closely with National Records of Scotland (NRS) Geography

Disclosure control

- Protect individual-level data by introducing uncertainty
- Assuming pre-tabular either over-imputation or record swapping
- Level to be decided (and not made public)
- Balance between protection and utility

Flow of data


Capture data
Load & Validation
Remove False persons
Resolve Multiple Responses
Filter Rules
Imputation
Estimate missing returns
Impute skeleton records
Assign output areas
Derive variables for output
Disclosure Control
Outputs database

Publication and Dissemination

- Phased releases
- Increasing detail
- Thematic outputs etc

Thank you

National Records of Scotland

Scotland's Census 2011
Shaping our future