Scotland's Census 2021: Statistical Methods

National Records of Scotland

Scotland's Census
Shaping our future
A' dealbhadh ar n-àm ri teachd

Introduction
What is Scotland's Census?

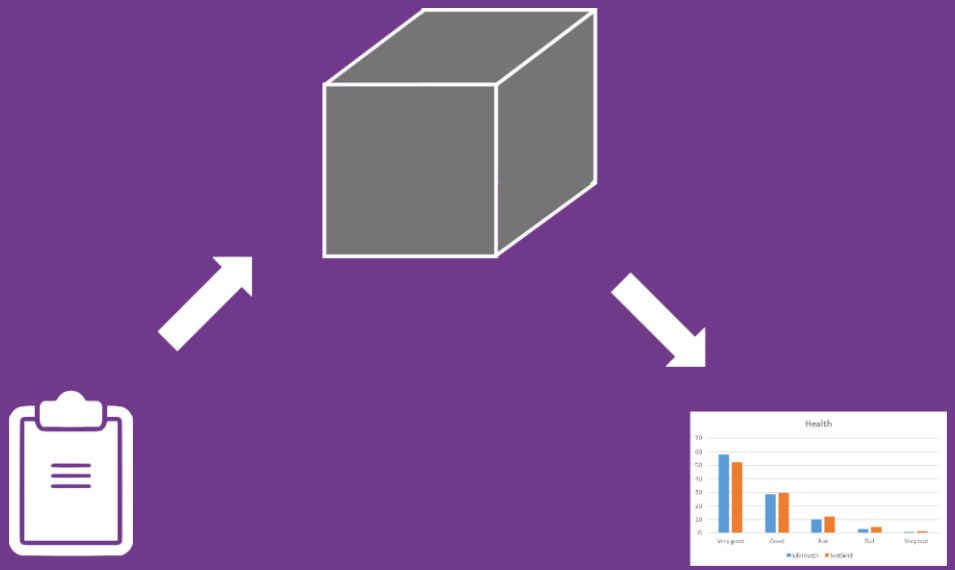
- Scotland's next census will be on 21 March 2021 as will other censuses across the UK.
- It aims to collect information to provide a snapshot of the nation and where we live.
- The Registrar General for Scotland is responsible for conducting the census in Scotland.
- We have one chance to get it right

Key Census Messages

- The 2021 Census will be conducted primarily online, making best use of technology and digital services.
- Everyone's personal census information is protected by law and National Records of Scotland will keep it confidential for 100 years.
- Success will require the support and contribution of many others, particularly participation by the people of Scotland.

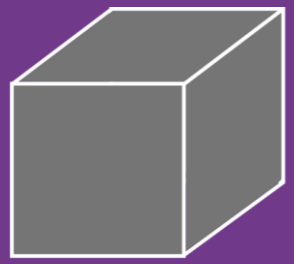
Statistical Methodology

- Today we will walk you through what happens to census data from when you submit your questionnaire, through to when we produce the 2021 census outputs.



Data Processing

Overview



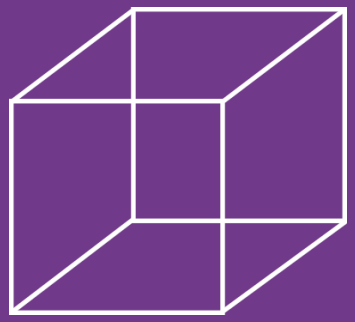
Our Overall Objective:

- “To transform census questionnaire returns into a correct, complete and consistent dataset suitable for outputs.”

- We are aiming to publish first outputs one year after Census day.

Data Processing

Overview



- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation (E&I)
- Census Coverage Survey (CCS)
- Estimation & Adjustment (E&A)
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

(Paper) Capture

“To convert responses into electronic data”

Capture

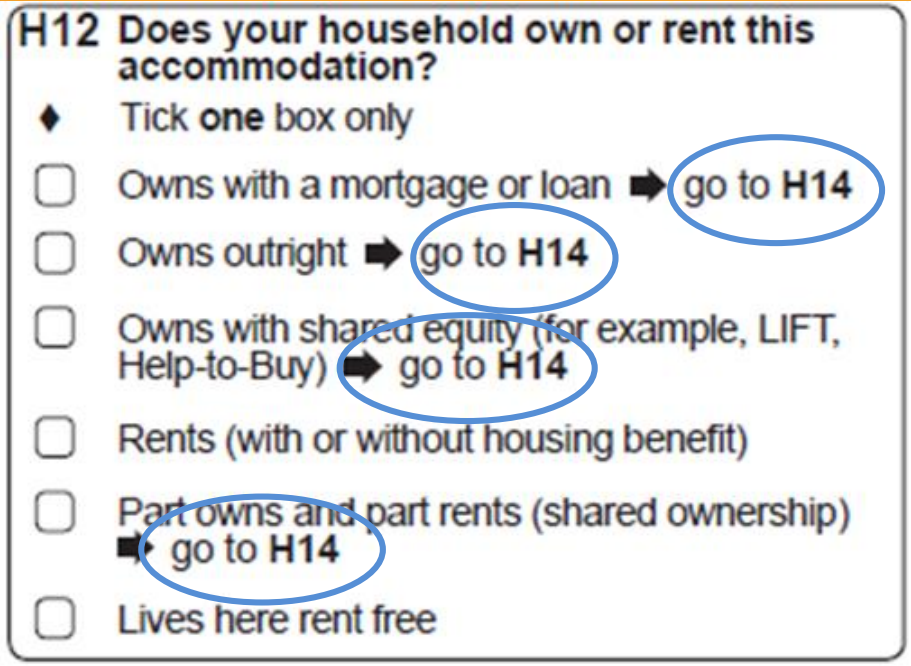
- Two ways of Capturing census data
  - Paper forms, scanning and automated coding
  - Online capture from online census forms

- Investing lots of time upfront to improve our online coding and automated coding

Capture

What do the forms look like?

Paper

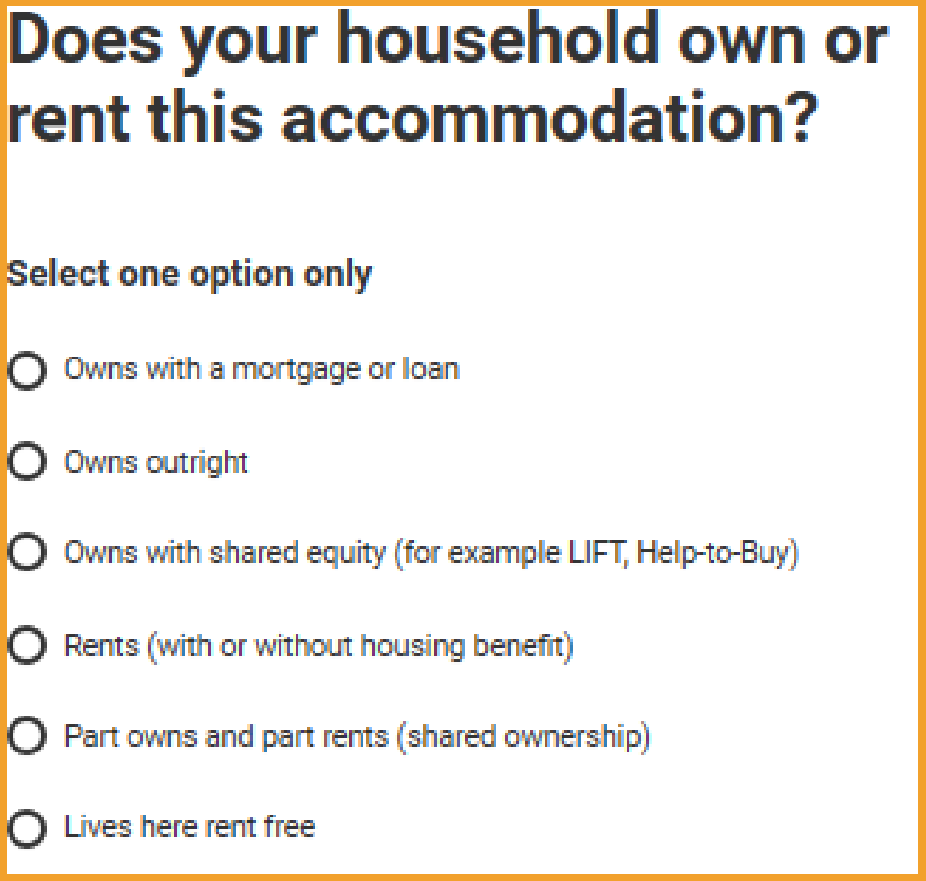


H12 Does your household own or rent this accommodation?

♦ Tick **one** box only

- ☐ Owns with a mortgage or loan ➡ go to **H14**
- ☐ Owns outright ➡ go to **H14**
- ☐ Owns with shared equity (for example, LIFT, Help-to-Buy) ➡ go to **H14**
- ☐ Rents (with or without housing benefit)
- ☐ Part owns and part rents (shared ownership) ➡ go to **H14**
- ☐ Lives here rent free

Online



Does your household own or rent this accommodation?

Select one option only

- ○ Owns with a mortgage or loan
- ○ Owns outright
- ○ Owns with shared equity (for example LIFT, Help-to-Buy)
- ○ Rents (with or without housing benefit)
- ○ Part owns and part rents (shared ownership)
- ○ Lives here rent free

Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Coding

“To classify data consistently in a suitable format ”

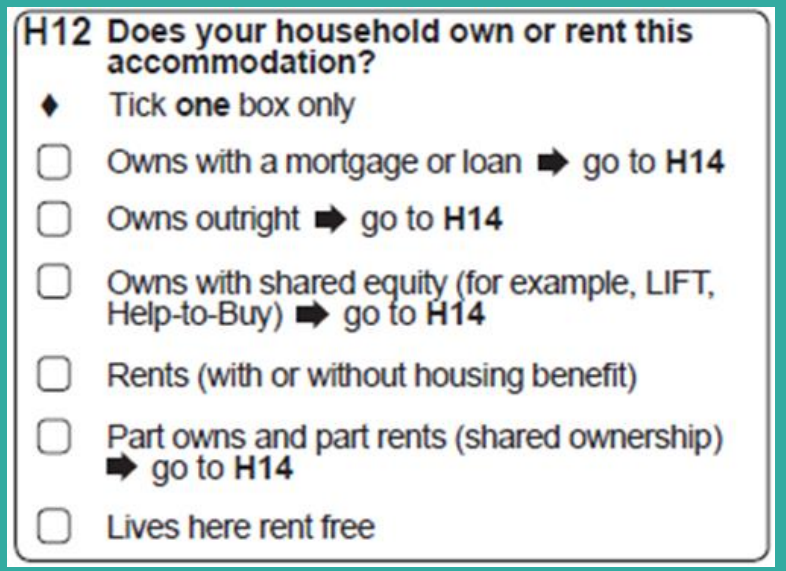
Coding
Overview

- **Definition:** Coding is the process by which responses given by an individual or household are assigned a recognised code.
- Need consistently coded data from all types of returns in order to undertake the rest of data processing and provide categorised outputs.

Coding
Tick boxes

- Single tick = response matched to a code in the classification index.

Box Ticked	Accommodation Type	Code
1	Owns with a mortgage or loan	1
2	Owns outright	2
3	Owns with shared equity	3
4	Rents	4
5	Part owns and part rents	5
6	Lives here rent free	6


H12 **Does your household own or rent this accommodation?**

♦ Tick **one** box only

☐ Owns with a mortgage or loan ➡ go to **H14**

☐ Owns outright ➡ go to **H14**

☐ Owns with shared equity (for example, LIFT, Help-to-Buy) ➡ go to **H14**

☐ Rents (with or without housing benefit)

☐ Part owns and part rents (shared ownership) ➡ go to **H14**

☐ Lives here rent free

- Sometimes we cannot match

Coding
Response types

- Several different types of response to be coded:
  - Tick Box
  - Number
  - Text
  - Address
  - Combinations of the above.

Coding
Classifications

- Classification Index - is a list of codes with their corresponding categories from a statistical classification (a 1:1 relationship).

- For example from the 2011 Country of Birth question.

Category	Code
Northern Ireland	922
Wales	924

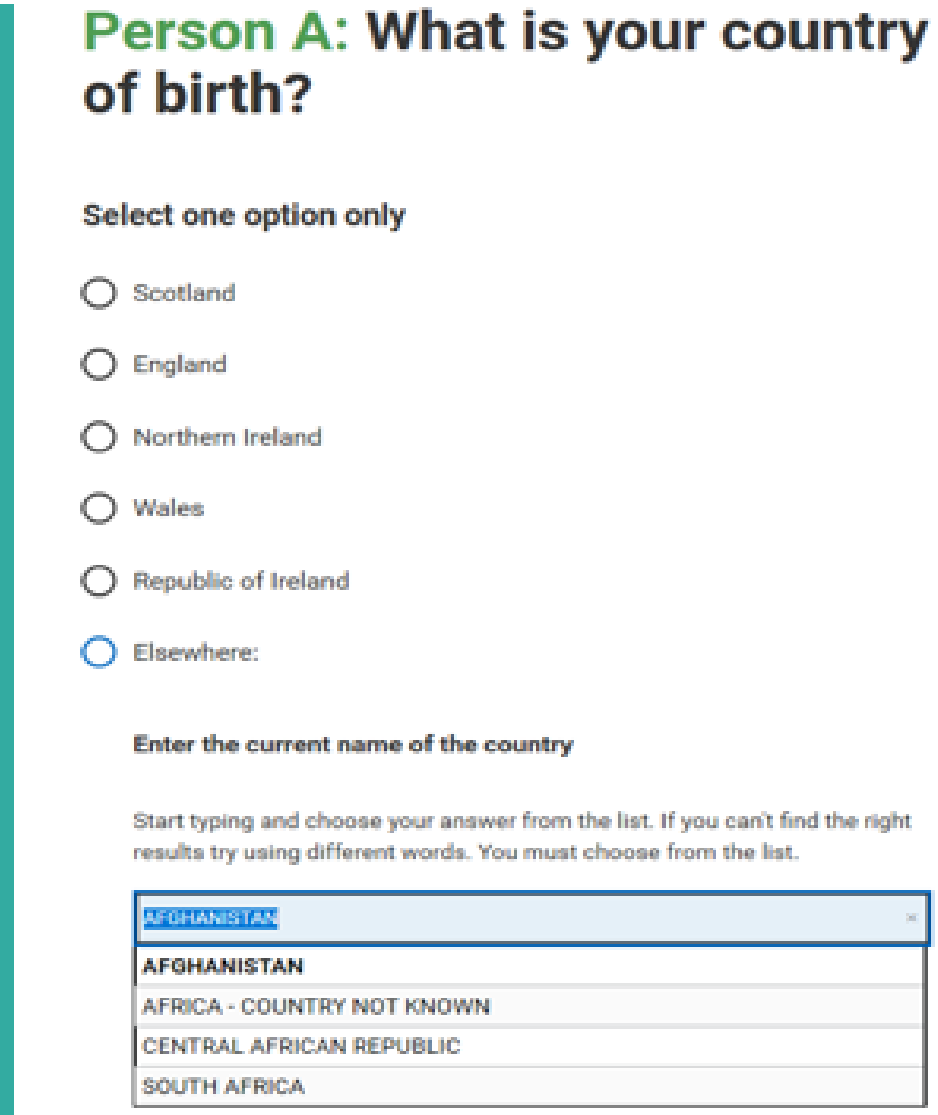
Coding
Classifications

- A synonym list allows known variations of responses to be matched to entries in the classification index.

Synonym	Category	Code
Belfast	Northern Ireland	922
Cardiff	Wales	924
Edinburgh	Scotland	923
Glasgow	Scotland	923
Scottish	Scotland	923

Coding

Type ahead functionality



Coding

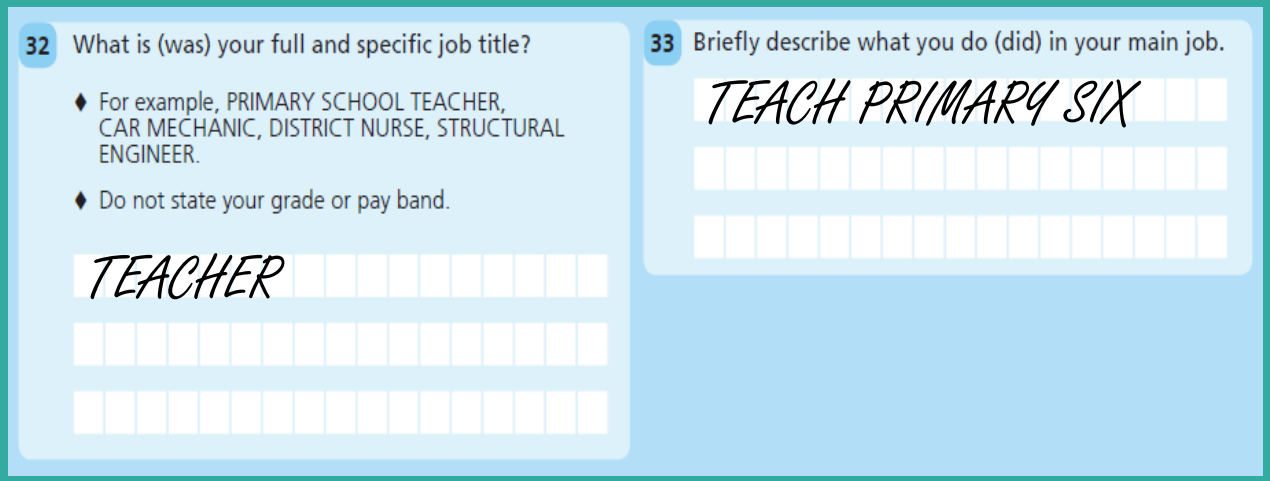
Occupation

Coding
Text boxes

- In 2011, automatic coding of OCCUPATION was < 5%.

Standard Occupation Classification (SOC) index, which has >28000 job titles.

- SOC has hierarchical structure.



- For example,

If the response is TEACHER, code =“231-”

Looking at question 33, code = “2314” for Primary Teacher

Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Data Cleansing

"A collection of processes we apply to Census data to account for specific errors, and prepare the data so it's suitable for later statistical processes"

Data Cleansing

What Is Data Cleansing?

- Analysing data for possible errors
  - Is this a real response?
  - Duplicate households or individuals
  - Answering questions you don't have to

- 3 Primary Methods in 2021:
  1. Remove False Persons
  2. Resolving Multiple Responses
  3. Filter (Routing) Rules

Remove False Persons (RFP)

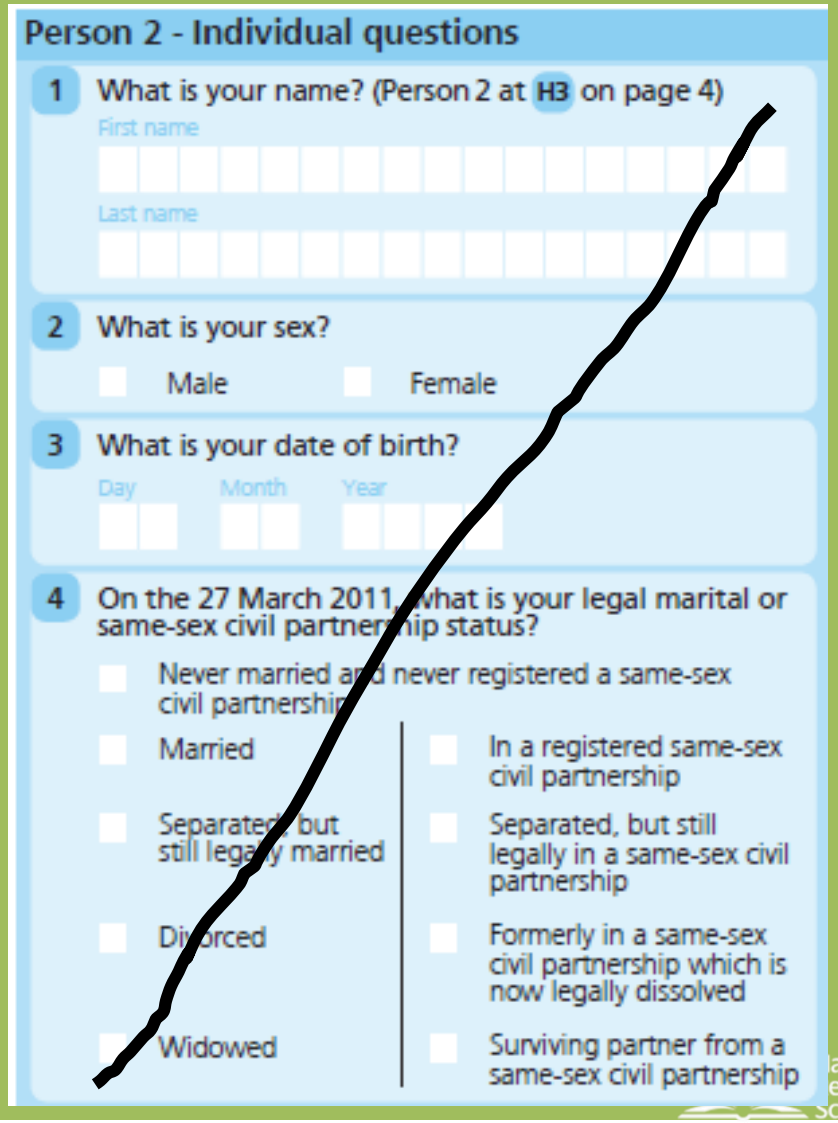
Data Cleansing

"False persons" are records in the census data which don't correspond with a genuine respondent.

Remove False Persons (RFP)

Data Cleansing

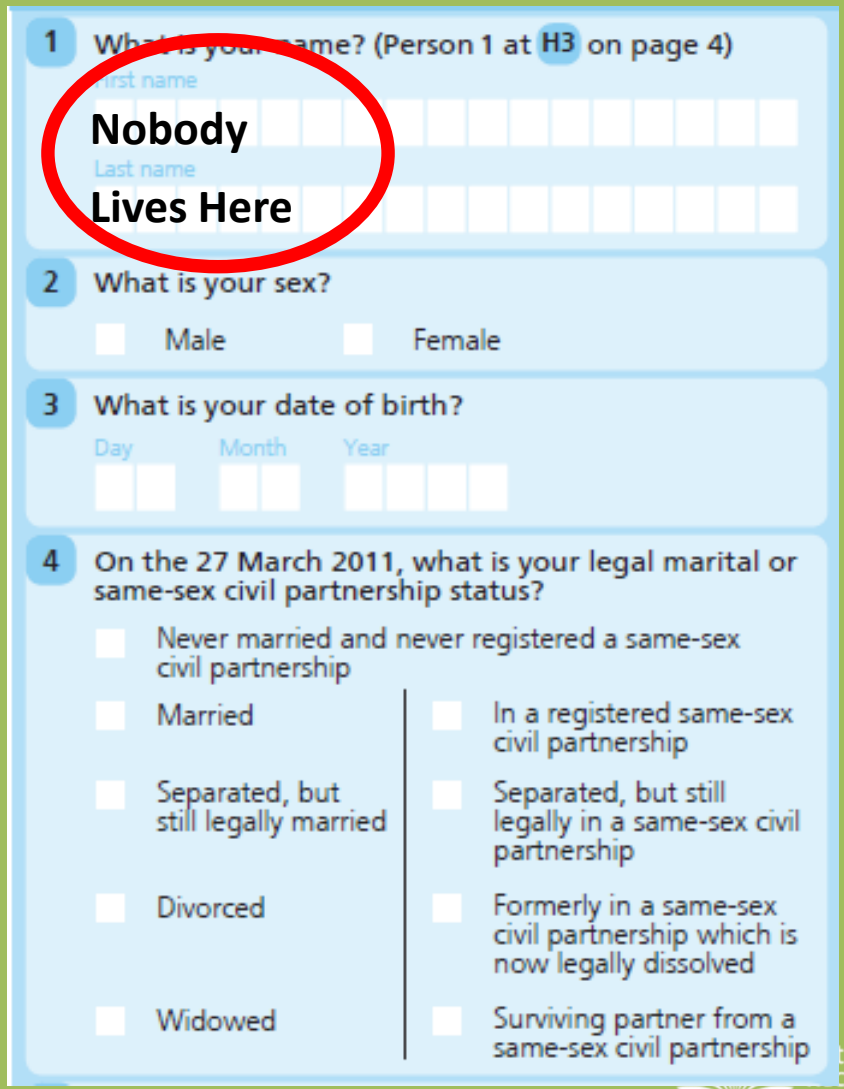
- A person living alones scores out the remaining person pages on their questionnaire



Remove False Persons (RFP)

Data Cleansing

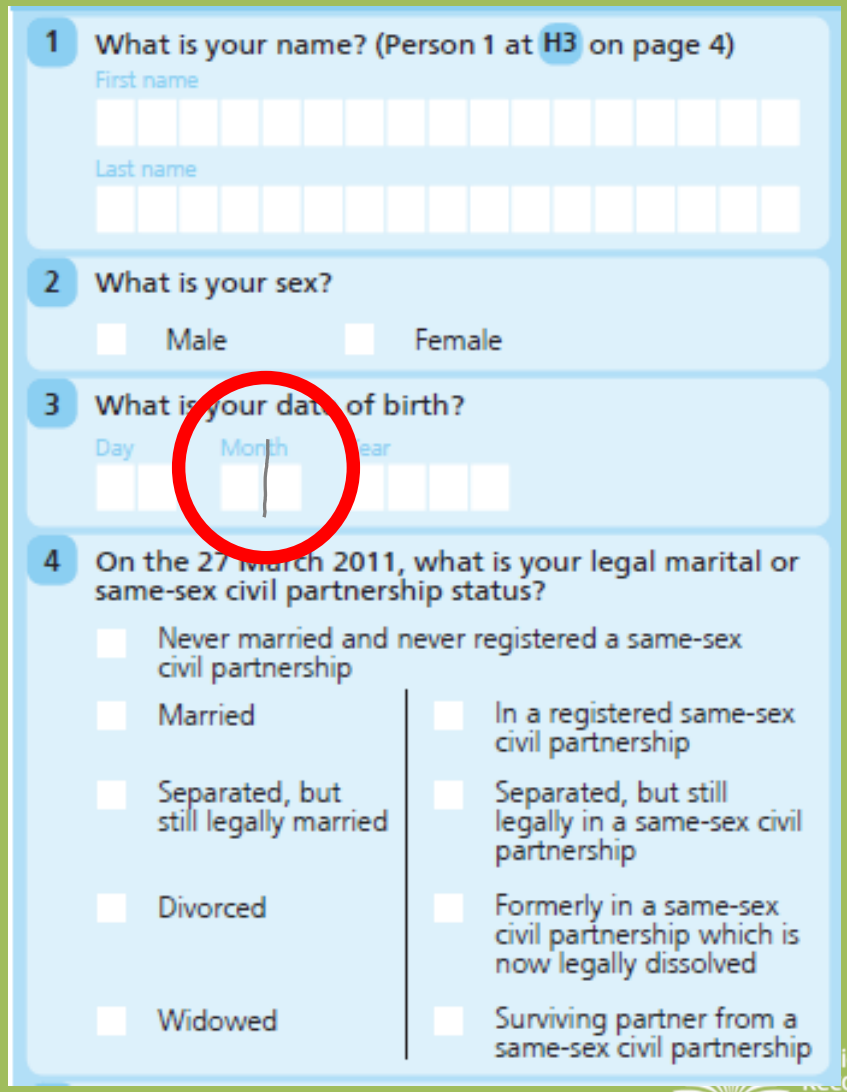
- The owner of an unoccupied house tried to respond correctly, but their response is captured and coded



Minimum Requirements Filter(s)

Data Cleansing

- An accidental mark on the page is captured and coded, resulting in a response



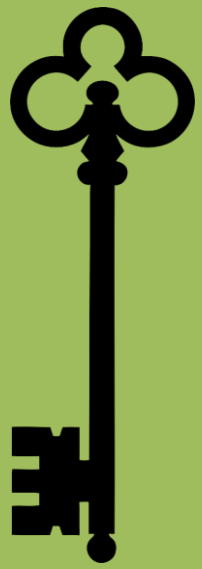
Remove False Persons (RFP)

Data Cleansing

- over·count | \ ˌō-vər-ˈkau̇nt \
- **overcounted; overcounting**
- **Definition of *overcount***
- **:** to count more of (people or things) than is accurate

Remove False Persons (RFP)

Data Cleansing

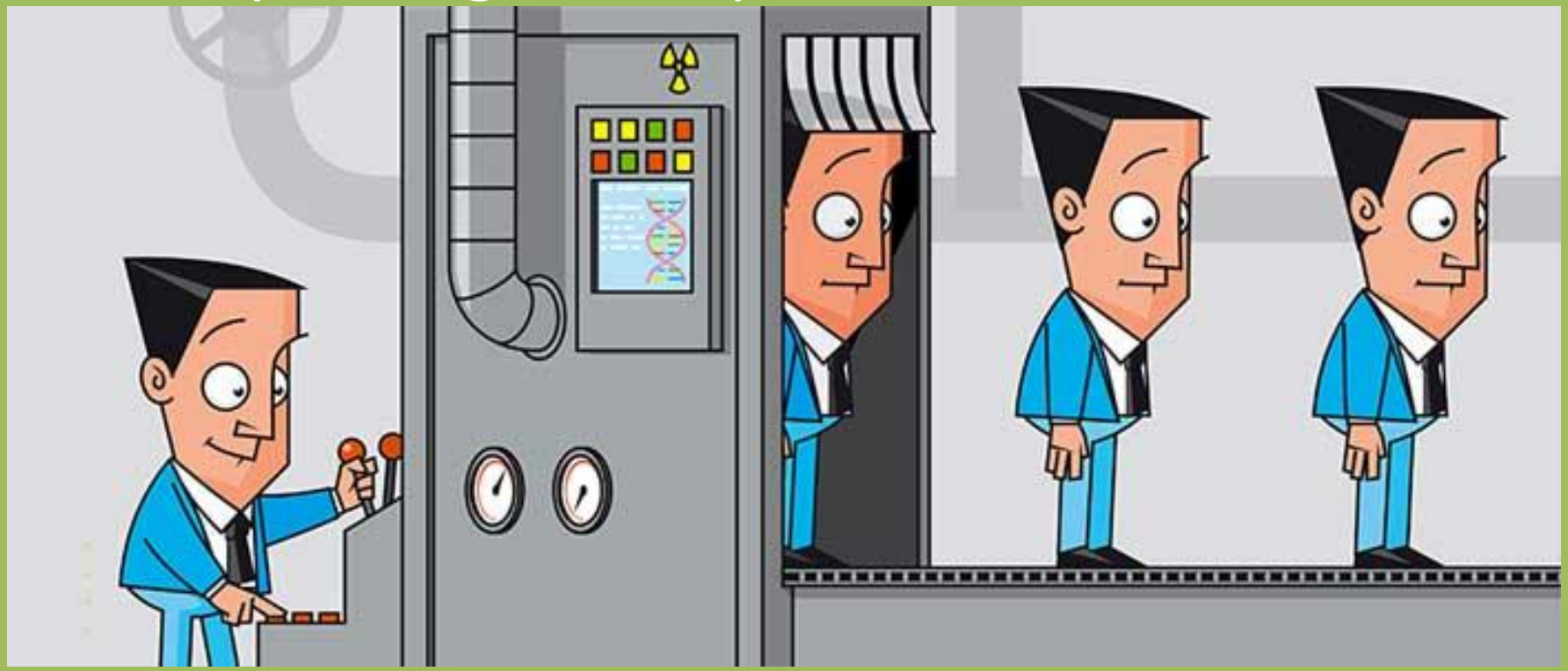


- **Name**
  - Name on the household section, and/or
  - Name on the person section
- **Date of Birth**
- **Sex**
- **Marital Status** (where appropriate)
- **Relationship within household** (where appropriate)

Resolving Multiple Responses

Data Cleansing

Multiple response - two or more Census returns corresponding to one person or household



Resolving Multiple Responses

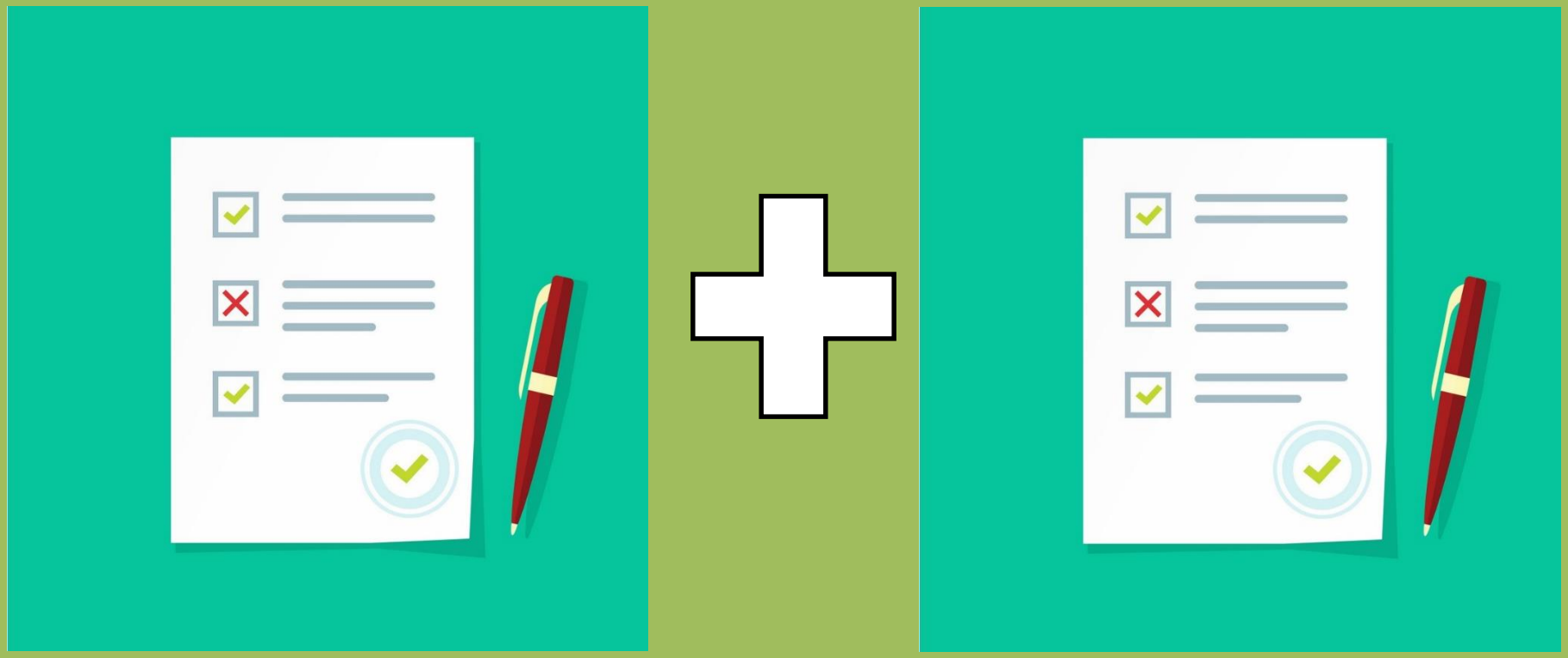
Data Cleansing

- A household providing both a paper return and an online return

Resolving Multiple Responses

Data Cleansing

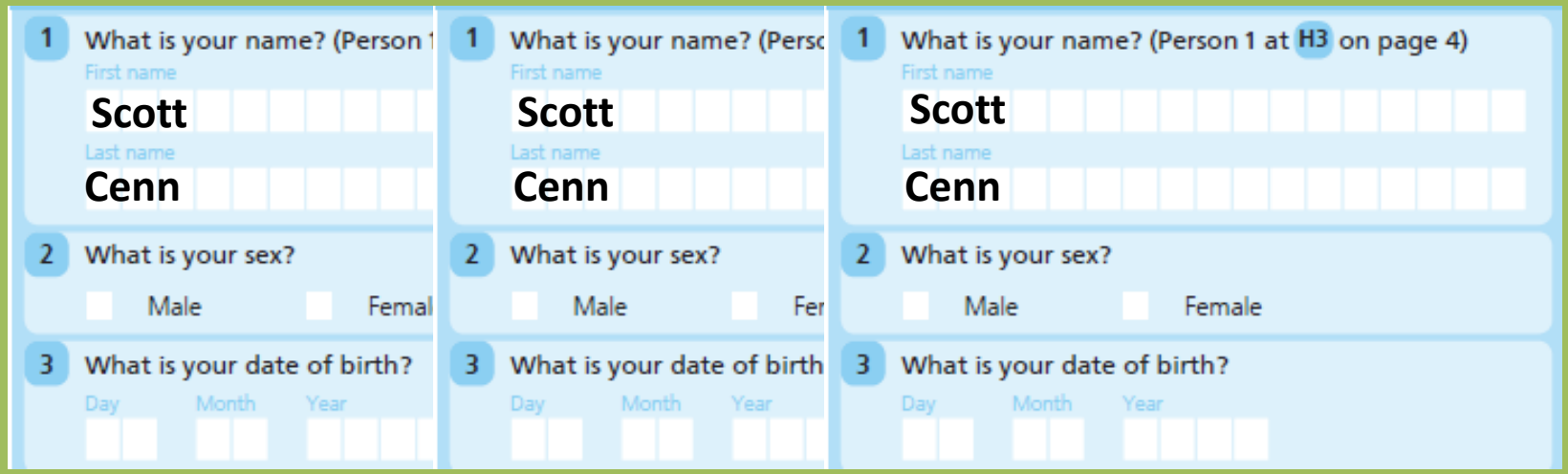
- A household receives two questionnaires and returns both



Resolving Multiple Responses

Data Cleansing

- One person fills the same information multiple times in their household return



Resolving Multiple Responses

Data Cleansing

1 What is your name? (Person 1 at H3 on page 4)

First name: Scott

Last name: Cenn

2 What is your sex?

☐ Male ☐ Female

3 What is your date of birth?

Day: 05 Month: 03 Year: 1976

4 On the 27 March 2011, what is your legal marital or same-sex civil partnership status?

☐ Never married and never registered a same-sex civil partnership

☐ Married

☐ Separated, but still legally married

☐ Divorced

☐ Widowed

☐ In a registered same-sex civil partnership

☐ Separated, but still legally in a same-sex civil partnership

☐ Formerly in a same-sex civil partnership which is now legally dissolved

☐ Surviving partner from a same-sex civil partnership

1 What is your name? (Person 1 at H3 on page 4)

First name: Scott

Last name: Cenn

2 What is your sex?

☐ Male ☐ Female

3 What is your date of birth?

Day: 10 Month: 10 Year: 2001

4 On the 27 March 2011, what is your legal marital or same-sex civil partnership status?

☐ Never married and never registered a same-sex civil partnership

☐ Married

☐ Separated, but still legally married

☐ Divorced

☐ Widowed

☐ In a registered same-sex civil partnership

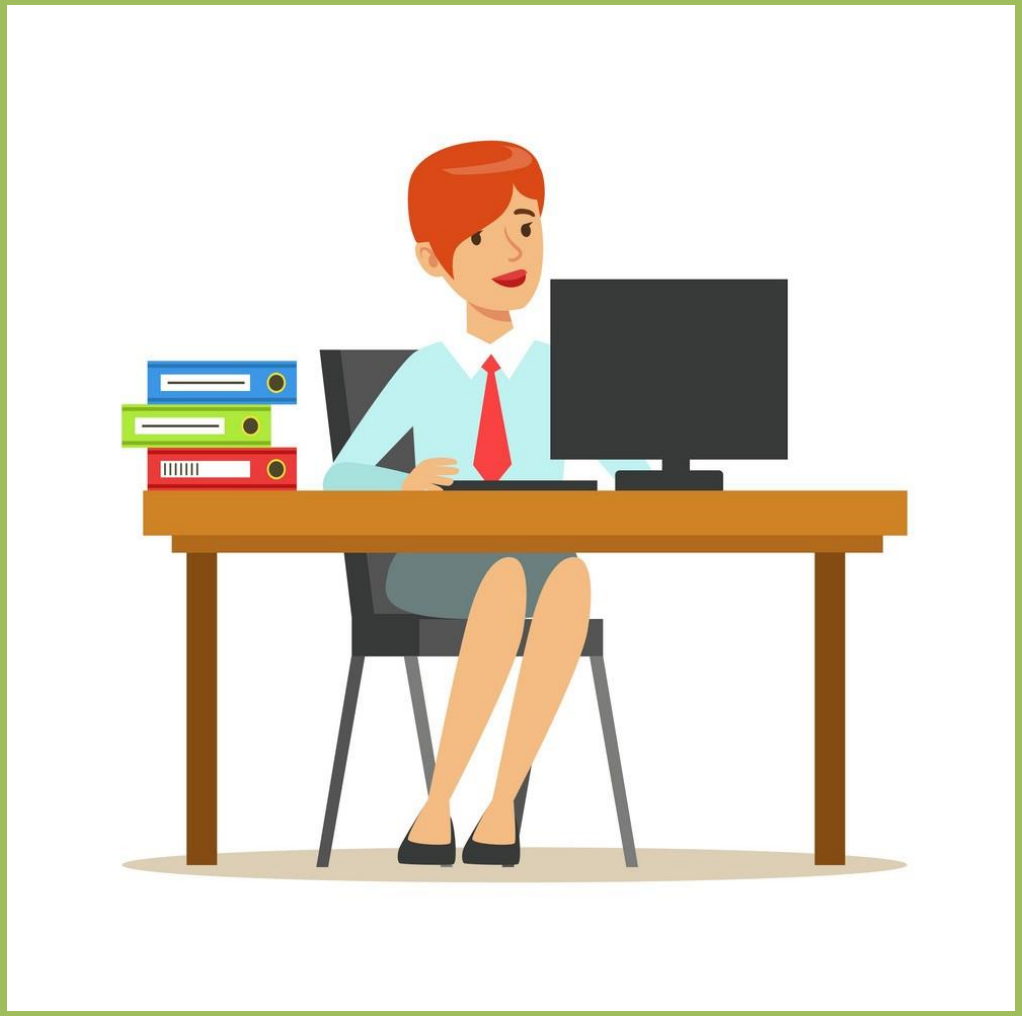
☐ Separated, but still legally in a same-sex civil partnership

☐ Formerly in a same-sex civil partnership which is now legally dissolved

☐ Surviving partner from a same-sex civil partnership

Resolving Multiple Responses

Data Cleansing



Resolving Multiple Responses

Data Cleansing

1. Identify cluster of duplicates
2. Select a primary record to retain
3. Combine the records, where information on the primary record takes priority
    - Based on specific criteria unique, methodologically applied to each situation

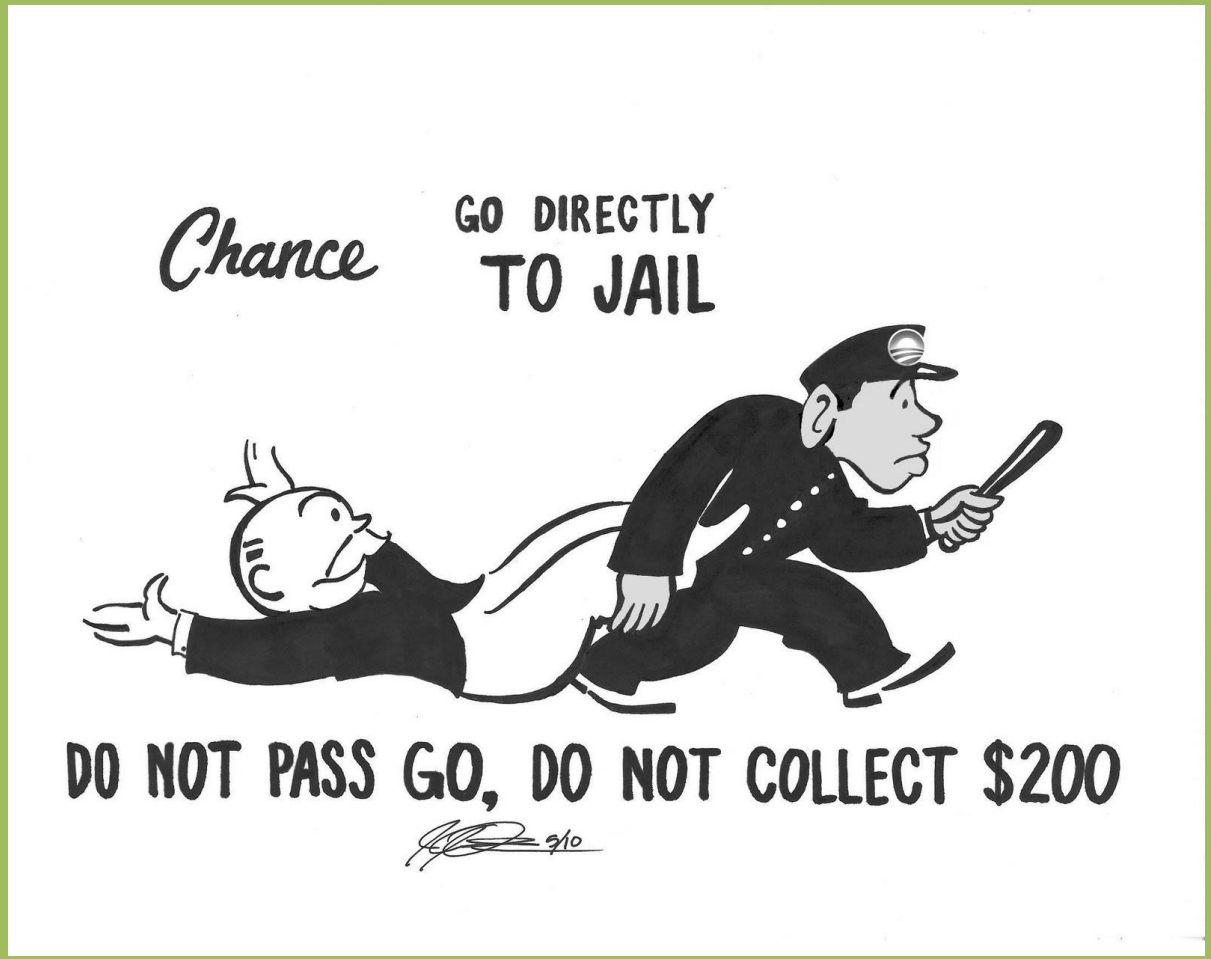
Resolving Multiple Responses

Data Cleansing

Individual responses, which is where someone wants to provide a response but keep it private from their householder, are prioritised regardless of other factors

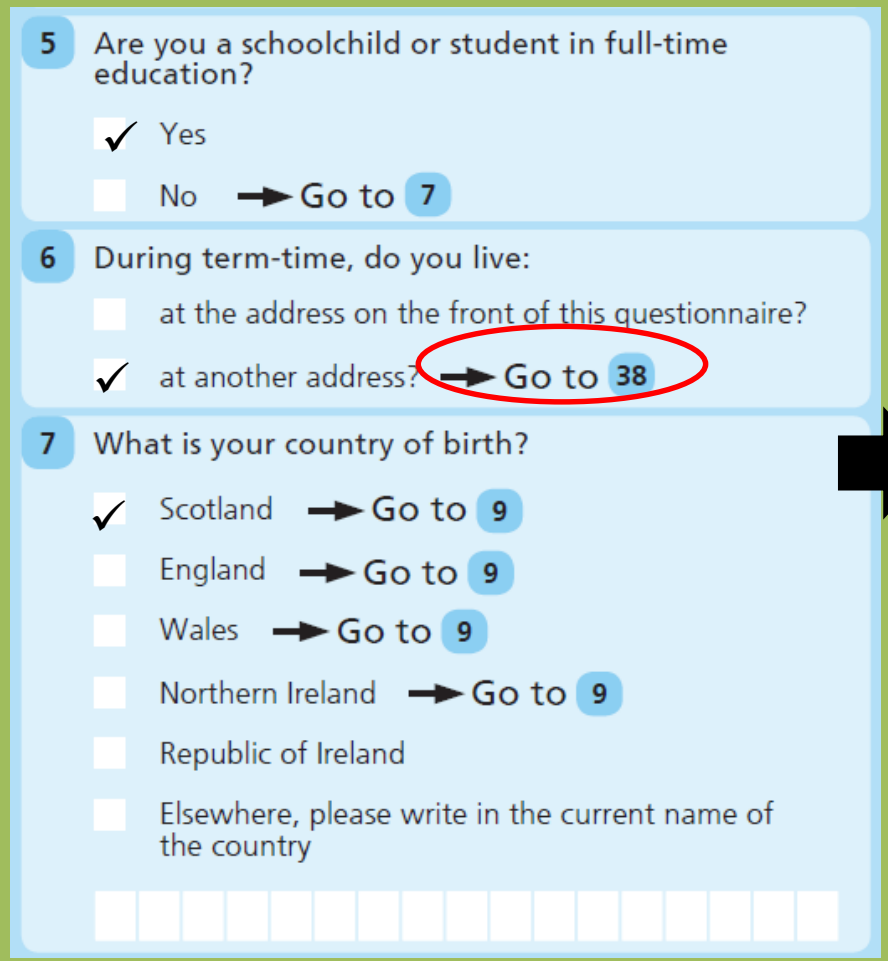
Filter (Routing) Rules

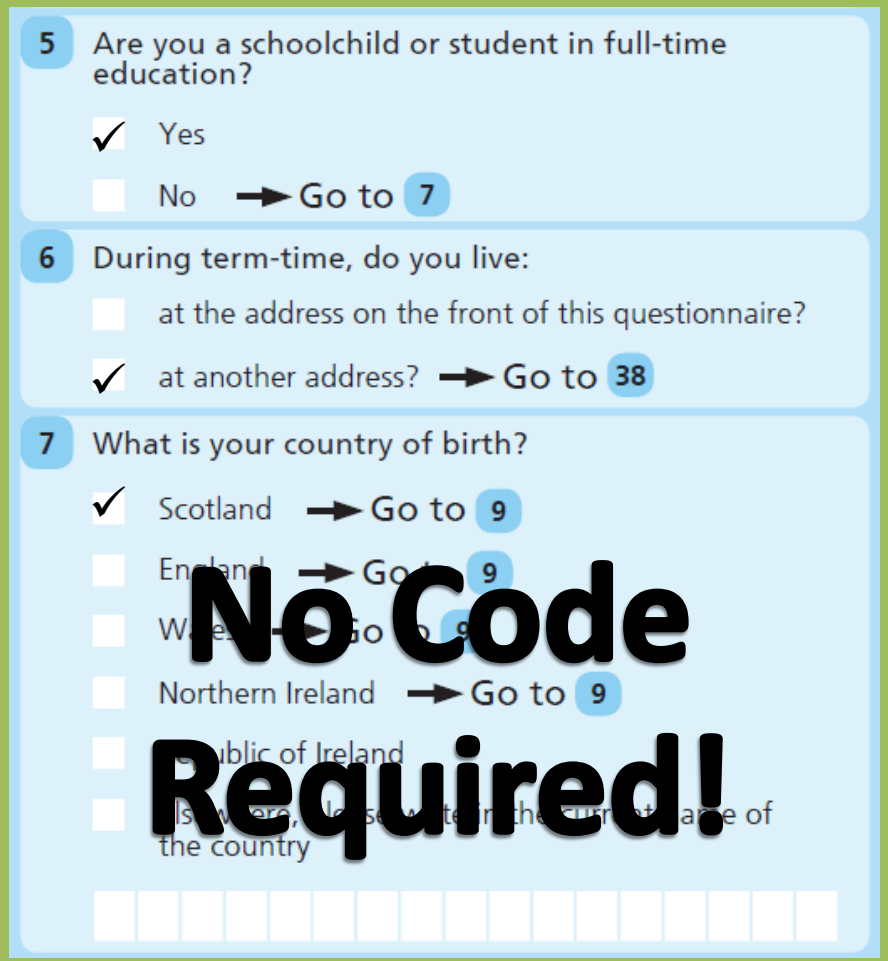
Data Cleansing



Filter (Routing) Rules

Data Cleansing





Summary

Data Cleansing

1. False records removed
2. Multiple responses identified and resolved
3. Missing questions identified for further processing (we'll come back to this)

Data Processing

Overview

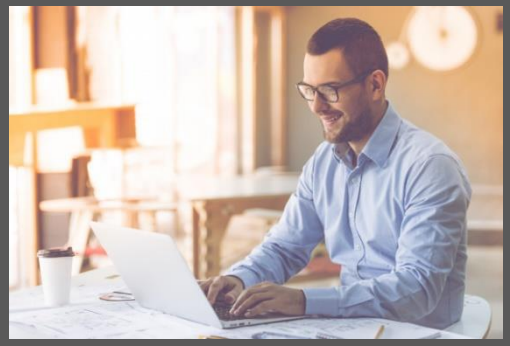
- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Administrative Data

"To improve the quality of the data"

Admin Data

What Is Admin Data?



- “Administrative data refers to information collected primarily for administrative reasons (not research). This type of data is collected by government departments and other organisations for registration, transactions and record-keeping, usually when delivering a service. Administrative data are often used for operational purposes and their statistical use is secondary”. - UKSA

- Paper records, Computer files, from online questionnaires

Admin Data

Different Types of Admin Data

- Counts of characteristics :

  For example- from Scottish Summary Statistics for Schools: Number of pupils

	2013	2014	2015	2016	2017	2018	2019
Primary	377,382	385,212	391,148	396,697	400,312	400,276	398,794
Secondary	289,164	284,762	281,939	280,983	281,993	286,152	292,063
Special	6,956	6,940	6,871	6,668	6,654	6,823	7,132
Total	673,502	676,914	679,958	684,348	688,959	693,251	697,989

- Individual records : Mock Census Records

Capture	HH ID	Person Number	Age	Gender	Marital Status	Student	Term Address	Country of Birth
Online	5	1	38	Male	Married	No	-	UK
Online	5	2	39	Female	Married	No	-	UK
Online	5	3	18	Female	Single	Yes	Elsewhere	France

Admin Data

Why Use Admin Data in the Census?

- Quality Assurance
- To check that the data we are receiving is what we expect :
  - Data Completeness?
  - Data Quality?
  - Missing Data?
  - We can use admin data to help with this

Admin Data
Counts

- Example of under counts in census returns :

  School Pupil Census Vs Census Count output

	2019 School Pupil Census	Hypothetical Census Count from our outputs	Difference
Primary School Pupils	398,794	406,794	8,000
Secondary school Pupils	292,063	296,063	4,000

- Highlights something is wrong – indicating that our code has dropped people or the underlying data is wrong.
- Action – check the data or the computer code

Admin Data

Individual Data

- Linking independent dataset together – why this will give us more accurate data.
- RFP – Stops potentially real people being removed from the census dataset, allowing them to be counted.
- RMR – Same person on form multiple times. Linking to admin allows us to minimise over counting the population.

Admin Data

Can you just link individual datasets together?

- Legislation/Ethics
- Security
- Privacy consideration
- GDPR

Admin Data

Example of how linking dataset helps the census

At University in Edinburgh

ADMIN DATA

Home in Inverness

Scotland's Census
Shaping our future

National Records of Scotland

Admin Data

Main aim of Admin Data ?

- To help identify over or under counts in geographic areas
- Which will feed through to the ‘Edit and imputation’

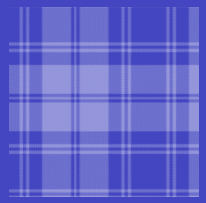
Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Edit and Imputation

"To ensure the data are complete and consistent at record level"



Edit & Imputation (E&I)

Overview

We detect and correct

- **Missing values** (e.g. skipped questions)
- **Invalid values** (e.g. age out of range)
- **Inconsistencies** (e.g. arrived in UK before born)

Edit & Imputation (E&I)

Overview

2011 item-level response rates

Highest:

- Date of birth (99.3%)
- Sex (99.2%)

Lowest:

- Year last worked (83.2%)
- Long-term health condition (86.8%)

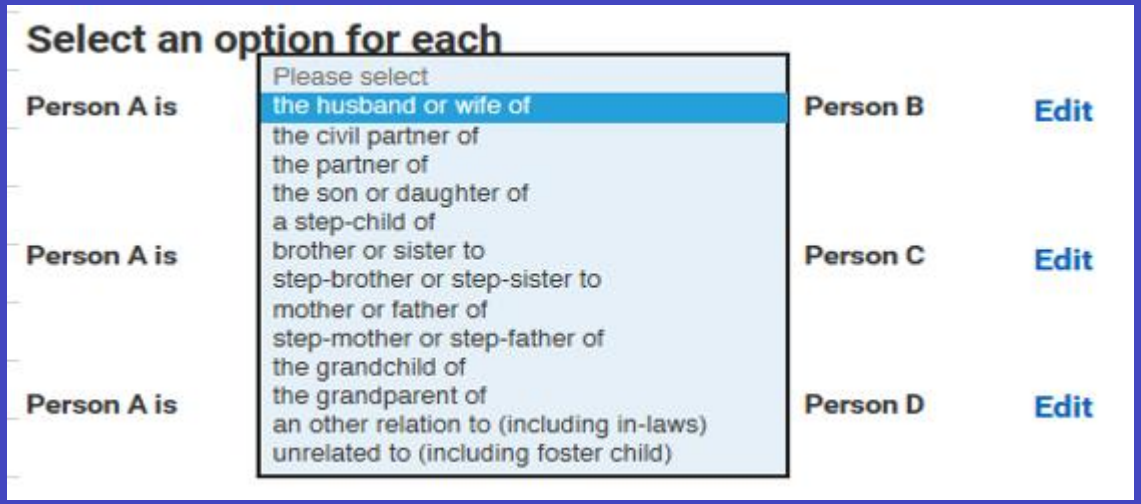
Edit & Imputation (E&I)

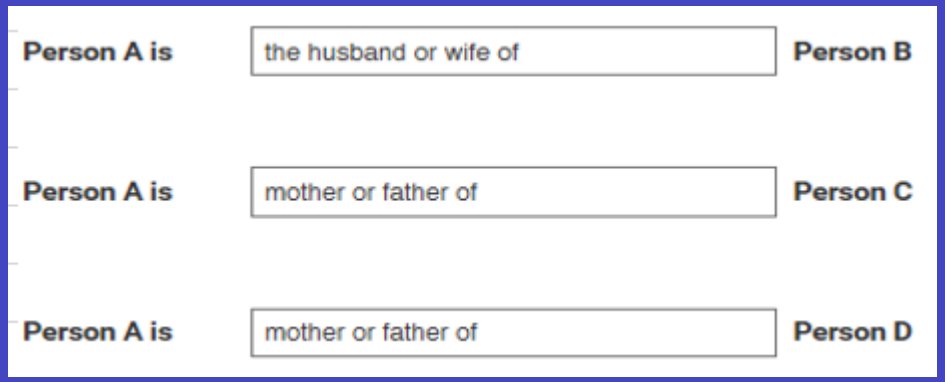
Relationship Algorithm 1

- We ask respondents how people in their household are related to each other
- Required for outputs (e.g. to group people into families within households)
- The nature of the question can make it challenging to complete – easier online than on paper

Edit & Imputation (E&I)

Relationship Algorithm 1





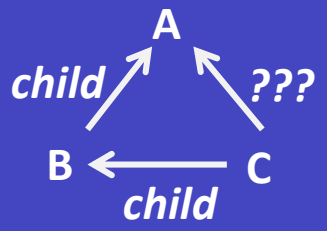
Rehearsal 2019 version

Edit & Imputation (E&I)

Relationship Algorithm 1

Relationship Algorithm 1 is a process we use to tidy up the relationship variables before donor imputation.

- Can fix some common respondent errors
- Can back-fill relationships, by triangulating them



Edit & Imputation (E&I)

Donor Imputation - CANCEIS

CANCEIS

- Specialist piece of software, developed by Statistics Canada for use on census data
- Used internationally for census imputation (inc. E&W, NI)
- Designed to implement nearest-neighbour hot-deck donor imputation

Edit & Imputation (E&I)

Donor Imputation - CANCEIS

- Can program in edit rules to prevent creating inconsistencies when imputing
- Highly customisable imputation and system parameters
- Especially good for categorical data, cross-distributions
- Detailed output files with audit trail

Edit & Imputation (E&I)

Donor Imputation

- An **edit check** identifies errors, omissions, and inconsistencies.

- Records are **flagged** as a ‘pass’ or a ‘fail’

- Donor Imputation:

Where a record ‘fails’, we pick a **donor** from a selection of similar records, and **copy and paste** the donor’s responses to fill in the gaps.

Edit & Imputation (E&I)

Donor Imputation

What do we mean by similar?

- Depends on which variable we're imputing
- Some variables are particularly good at predicting others
- We might be more interested in the household as a whole
- Can add weights so that some variables have more influence as predictors

Edit & Imputation (E&I)

Donor Imputation

Measuring similarity:

Group variables into modules for imputation

Demographics	age, sex, marital status…
Culture	language, ethnicity…
Health	disability, health conditions…
Labour Market	occupation, hours worked…
Household	tenure, number of cars…

Edit & Imputation (E&I)

Donor Imputation – examples

Two simplified examples:

- Imputing as individuals (labour market)
- Imputing as households (culture)

Edit & Imputation (E&I)

Donor Imputation – Example 1

ID	SEX	AGE	QUALIFICATIONS	INDUSTRY	OCCUPATION	EDIT
1	M	21	1	Retail	Assistant	Pass
2	F	41	3	Transport	Pilot	Pass
3	F	24	1	Retail		Fail
4	F	40	2	Retail	Store Manager	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation – Example 1

ID	SEX	AGE	QUALIFICATIONS	INDUSTRY	OCCUPATION	EDIT
1	M	21	1	Retail	Assistant	Pass
2	F	41	3	Transport	Pilot	Pass
3	F	24	1	Retail		Fail
4	F	40	2	Retail	Store Manager	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation – Example 1

ID	SEX	AGE	QUALIFICATIONS	INDUSTRY	OCCUPATION	EDIT
1	M	21	1	Retail	Assistant	Pass
2	F	41	3	Transport	Pilot	Pass
3	F	24	1	Retail		Fail
4	F	40	2	Retail	Store Manager	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation – Example 1

ID	SEX	AGE	QUALIFICATIONS	INDUSTRY	OCCUPATION	EDIT
1	M	21	1	Retail	Assistant	Pass
2	F	41	3	Transport	Pilot	Pass
3	F	24	1	Retail	Assistant	Pass
4	F	40	2	Retail	Store Manager	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation – Example 2

HH	ID	MARITAL STATUS	AGE	ETHNICITY	LANGUAGE	EDIT
1	1	Married	46	White - Scottish	English only	Pass
1	2	Married	38	White – Scottish	English only	Pass
1	3	Single	6	White - Scottish	English only	Pass
2	1	Married	35	Asian - Pakistani	Punjabi	Fail
2	2	Married	36	Asian – Pakistani	Punjabi	Fail
2	3	Single	6			Fail
3	1	Married	56	Asian - Pakistani	Punjabi	Pass
3	2	Married	51	Asian – Pakistani	Punjabi	Pass
3	3	Single	20	Asian – Pakistani	Punjabi	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation – Example 2

HH	ID	MARITAL STATUS	AGE	ETHNICITY	LANGUAGE	EDIT
1	1	Married	46	White - Scottish	English only	Pass
1	2	Married	38	White – Scottish	English only	Pass
1	3	Single	6	White - Scottish	English only	Pass
2	1	Married	35	Asian - Pakistani	Punjabi	Fail
2	2	Married	36	Asian – Pakistani	Punjabi	Fail
2	3	Single	6			Fail
3	1	Married	56	Asian - Pakistani	Punjabi	Pass
3	2	Married	51	Asian – Pakistani	Punjabi	Pass
3	3	Single	20	Asian – Pakistani	Punjabi	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation – Example 2

HH	ID	MARITAL STATUS	AGE	ETHNICITY	LANGUAGE	EDIT
1	1	Married	46	White - Scottish	English only	Pass
1	2	Married	38	White – Scottish	English only	Pass
1	3	Single	6	White - Scottish	English only	Pass
2	1	Married	35	Asian - Pakistani	Punjabi	Fail
2	2	Married	36	Asian – Pakistani	Punjabi	Fail
2	3	Single	6			Fail
3	1	Married	56	Asian - Pakistani	Punjabi	Pass
3	2	Married	51	Asian – Pakistani	Punjabi	Pass
3	3	Single	20	Asian – Pakistani	Punjabi	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation – Example 2

HH	ID	MARITAL STATUS	AGE	ETHNICITY	LANGUAGE	EDIT
1	1	Married	46	White - Scottish	English only	Pass
1	2	Married	38	White – Scottish	English only	Pass
1	3	Single	6	White - Scottish	English only	Pass
2	1	Married	35	Asian - Pakistani	Punjabi	Fail
2	2	Married	36	Asian – Pakistani	Punjabi	Fail
2	3	Single	6			Fail
3	1	Married	56	Asian - Pakistani	Punjabi	Pass
3	2	Married	51	Asian – Pakistani	Punjabi	Pass
3	3	Single	20	Asian – Pakistani	Punjabi	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation – Example 2

HH	ID	MARITAL STATUS	AGE	ETHNICITY	LANGUAGE	EDIT
1	1	Married	46	White - Scottish	English only	Pass
1	2	Married	38	White – Scottish	English only	Pass
1	3	Single	6	White - Scottish	English only	Pass
2	1	Married	35	Asian - Pakistani	Punjabi	Pass
2	2	Married	36	Asian – Pakistani	Punjabi	Pass
2	3	Single	6	Asian – Pakistani	Punjabi	Pass
3	1	Married	56	Asian - Pakistani	Punjabi	Pass
3	2	Married	51	Asian – Pakistani	Punjabi	Pass
3	3	Single	20	Asian – Pakistani	Punjabi	Pass

Not real data

Edit & Imputation (E&I)

Donor Imputation

- Item (question) level response rate is important for imputation quality

Higher response rate:
More potential donors
Better matches between failed records & donors
More accurate imputation, better variance

Quality in => Quality out

- NB: Voluntary questions not imputed

Edit & Imputation (E&I)

Summary

- We have detected **missing**, **invalid**, and **inconsistent** values
- We have corrected them using **donor imputation** in CANCEIS software
  - Find similar record to act as “donor”
  - Copy values onto failed record
- Dataset is now complete and consistent at record level

Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Census Coverage Survey (CCS)

"To provide a secondary source of data to the census which allows us to estimate the total Scottish population"

Census Coverage Survey
Objective

- Not everyone completes a Census questionnaire and some people are missed
- Does not occur uniformly across all geographical areas or demography
- Individuals or whole households can be missed

Census Coverage Survey

Objective

- To produce total population estimates for Scotland, we require a separate, independent survey to the census.

- Then we can apply Dual System Estimation (also known as capture-recapture)

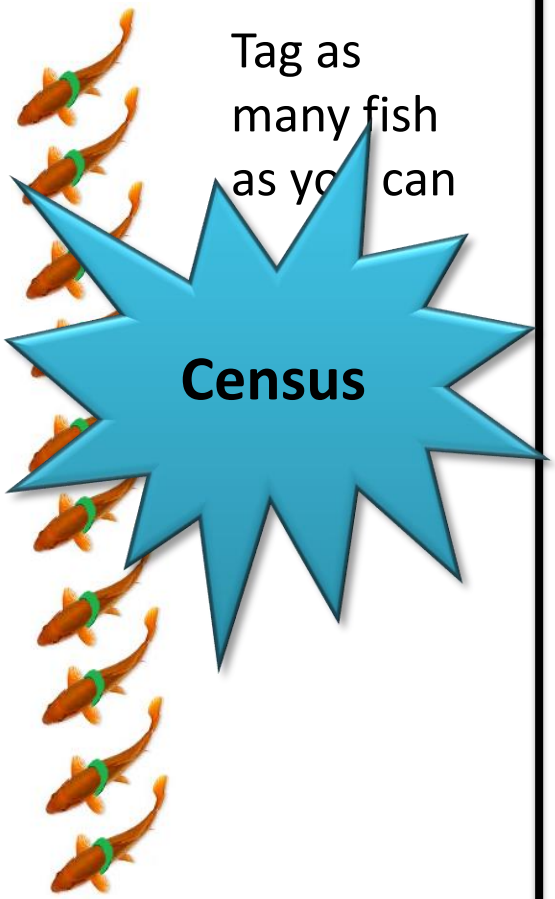
A Pond Full of Fish

Imagine you have inherited a pond full of fish

Perhaps you know that the fish are worth a lot of money

But you don't know how many fish you have

Day 1



Tag as many fish as you can

Census

10 Fish Tagged

Day 2

Note the number of tagged and untagged

CCS

5 Tagged
5 Untagged

We can now estimate the total number of fish in the pond!

$$\frac{\text{Number in Catch1} \times \text{Number in Catch2}}{\text{Number tagged in both}}$$

20 fish in the pond

National Records of Scotland

The CCS is…

- Sample survey
- Includes approx. 1.5% of the population
- Conducted approx. 6 weeks after census
- Subset of the census questionnaire
- Door-step interview, face-to-face
- Voluntary participation

Constraints for the Census Coverage Survey

Independence – Participation in the census or CCS should not trigger participation in the other

- No overlap between Census and CCS collection periods
- CCS uses its own address register
- Ensure Census fieldworkers do not work in the same CCS postcode

Census Coverage Survey
Summary

- Census Coverage Survey is a sample survey taken 6 weeks after Census
- Provides a secondary source of information to assess Census coverage and produce population estimates

Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Estimation and Adjustment

"To provide a dataset with records for the estimated total population of Scotland"

Estimation and Adjustment

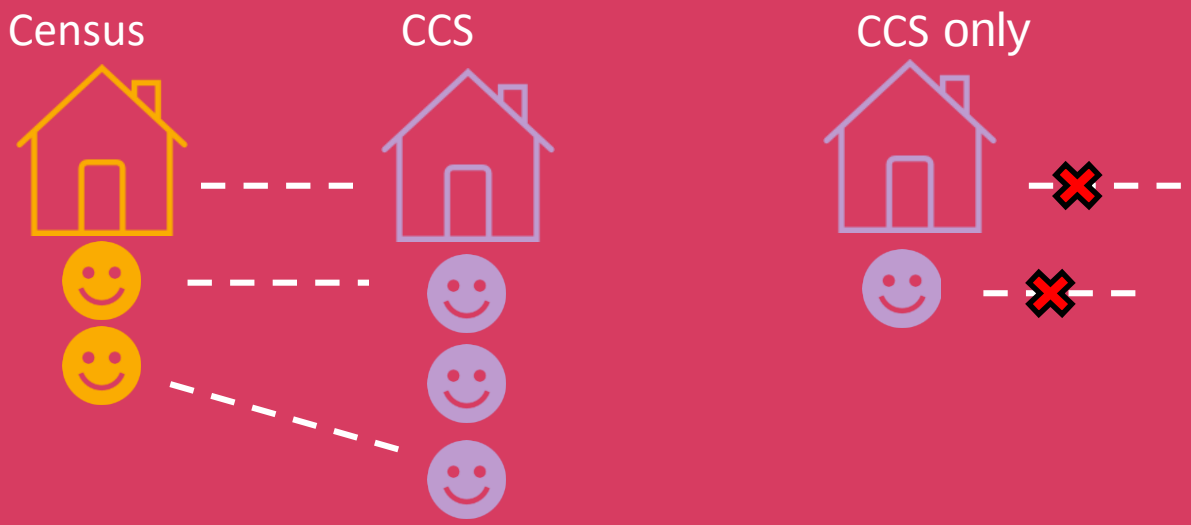
Intro

- Estimation produces overall population and household estimates
- Adjustment creates new records for the missed population
- Gives Census dataset for whole population

Estimation and Adjustment

Matching Census to CCS

- Matching between census records and CCS records



Gives counts for Dual System Estimation

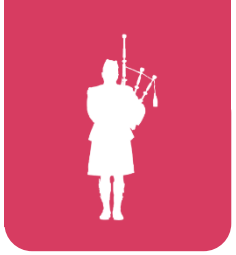
Estimation and Adjustment

Estimation

- We get estimated totals for people by:
  - Age-sex group (5 year age banded)
  - Activity last week
  - Ethnicity
  - Household tenure
- We get estimated totals for households by:
  - Tenure
  - Household size

Estimation and Adjustment

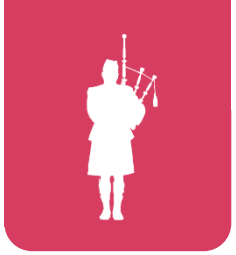
Estimation Corrections



- Estimation predominantly for people missed but overcount does exist
  - e.g. children between separated parents
- Some breaking of independence between Census and CCS needs bias correction
  - People who really don't want to complete census unlikely to complete CCS

Estimation and Adjustment

Adjustment



- Create records for those missed in order to get complete Census dataset
- Predicts likelihood of type of persons or households being missed in Census
- Then used to select records to be “donated” to form new records

Estimation and Adjustment

Adjustment

- People are added to existing households, and entirely new households are created.

Estimation and Adjustment

Adjustment

- These skeleton records are created with a subset of original variables

ID	SEX	AGE	QUALIFICATIONS	ACTIVITY LAST WEEK	OCCUPATION	ETHNICITY	LANGUAGE
1	M	21	---	Working	---	White	---

- The missing variables are added in post-coverage Edit and Imputation.

Edit & Imputation (E&I)

Post-Coverage Imputation

- Here the edit & imputation process is rerun on skeleton records only, to fill in their missing information.

- We also impute voluntary questions for these records (“no response” is a valid value to impute)

Edit & Imputation (E&I)

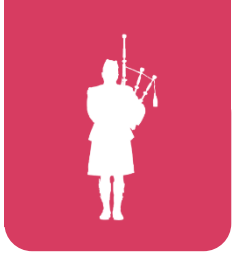
Post-Coverage Imputation

- Here the edit & imputation process is rerun on skeleton records only, to fill in their missing information.

- We also impute voluntary questions for these records (“no response” is a valid value to impute)

Estimation and Adjustment

Summary



- We produce estimates for the population from the Census and CCS
- People and households are added to the dataset to account for missed people
- After second imputation, have a complete dataset for the whole of Scotland

Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation (again)
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Statistical Quality Assurance

"to prevent, reduce or limit the occurrence of errors and therefore, to get it right first time."

National Records of Scotland

Scotland's Census
Shaping our future

Statistical Quality Assurance

- how we will assess and measure the level of quality being achieved throughout the processing of Census data and the production and dissemination of statistical outputs.

Statistical Quality Assurance

- Quality Assurance (QA) as the 'anticipation and avoidance of problems'.
- Quality Control (QC) as 'responding to observed problems'.
- Quality Management as the 'encompassing approach to quality'.

Statistical Quality Assurance

- Critical Success Factors (CSFs) describe what success will look like and are aligned to Scotland's Census 2021 objectives.

How we will achieve high quality results?
We will maximise our overall person response rate
We will ensure a minimum level of response with every local authority in Scotland
We will maximise the accuracy of our national population estimates

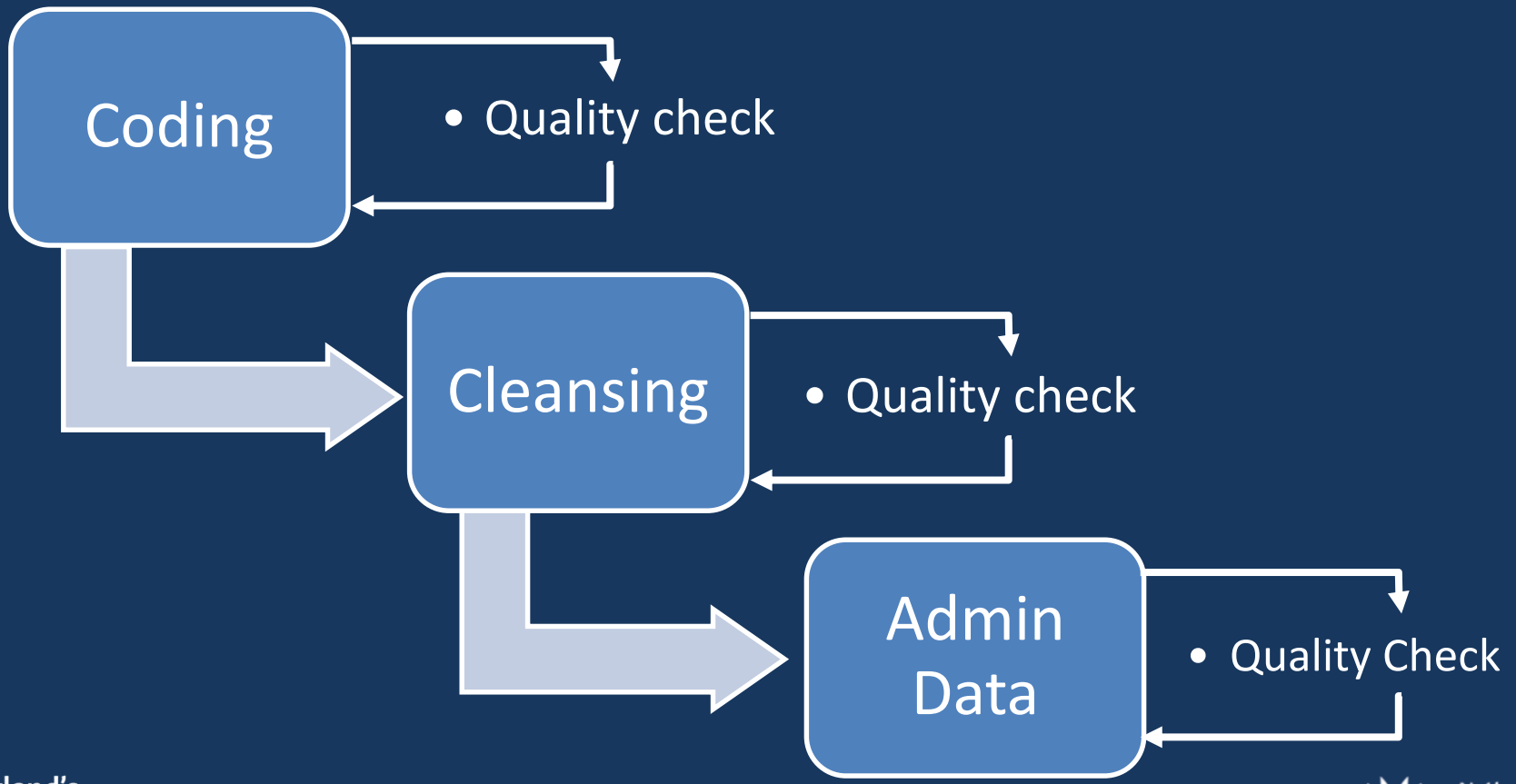
Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation (again)
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Statistical Quality Assurance

- Assurance of processes

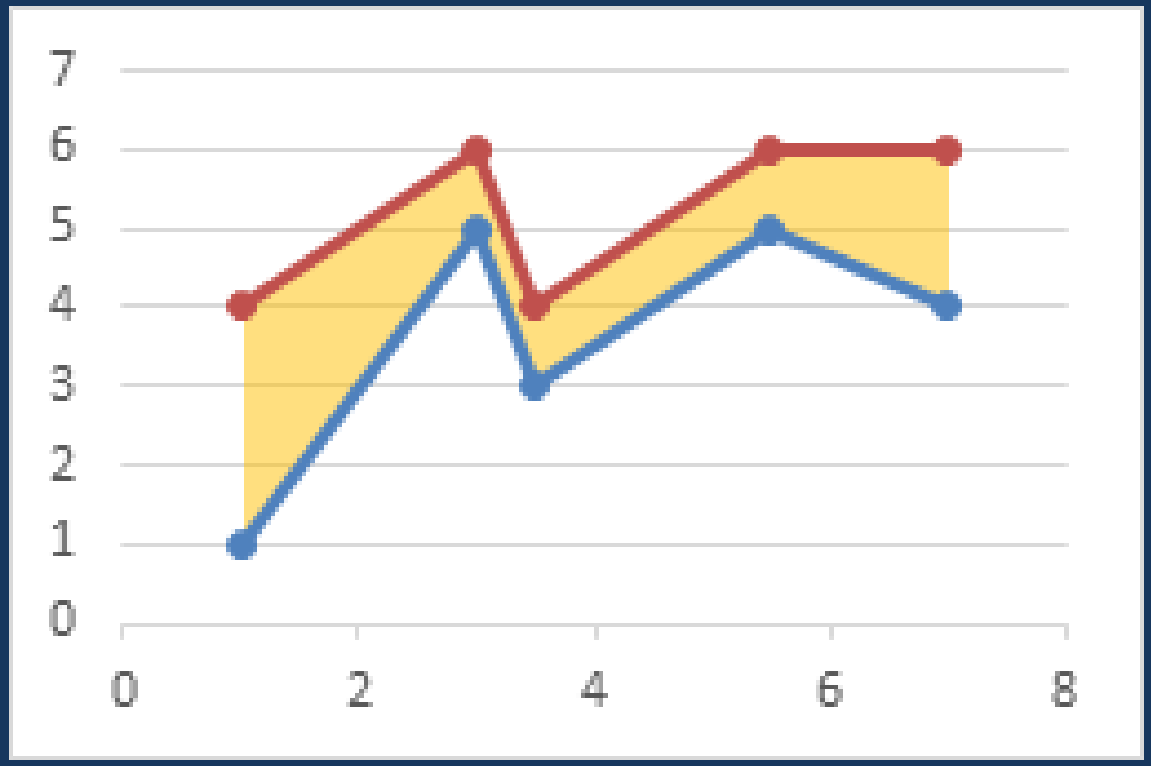


Statistical Quality Assurance

- Peer review
- Internal Peer Review
- External Methodology Assurance Panels

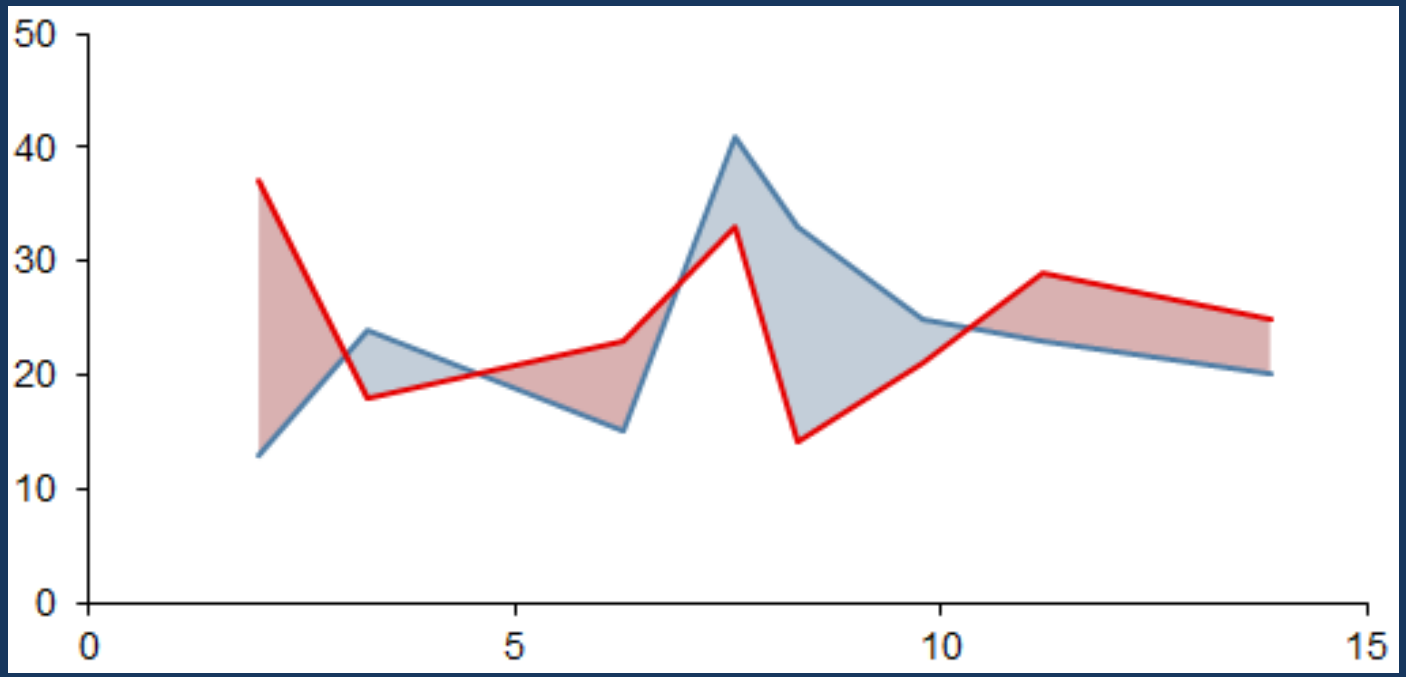
Statistical Quality Assurance

- Validation of Population Estimates
- Do our results look as expected?



Statistical Quality Assurance

- Validation of Population Estimates
- Do our results look as expected?



Statistical Quality Assurance

- Topic-based Analysis
- Do our results look as expected?

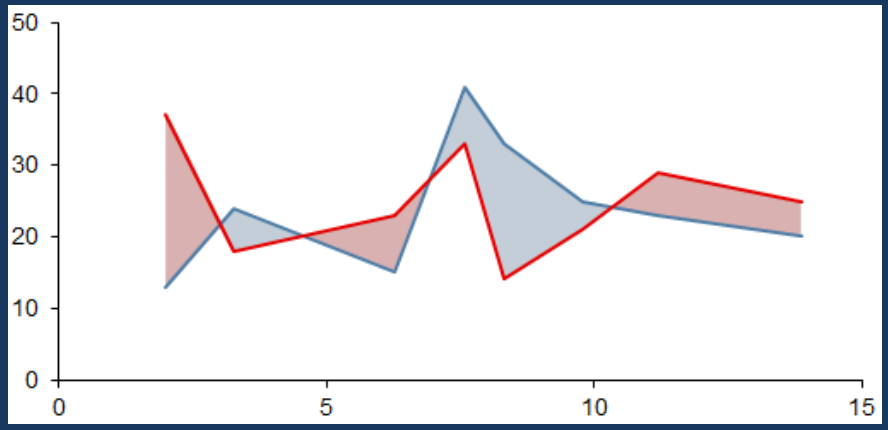
Housing

Health

Ethnicity

Occupation

Language



Statistical Quality Assurance

- Quality Assurance Panels
- Local authorities review QA packs in secure environment
- Provide advice on population estimates prior to Output release.

Statistical Quality Assurance

- National Statistics Accreditation
- All activities and outputs assessed against Code of Practice for Statistics
- Trustworthiness, quality and value
- https://www.scotlandscensus.gov.uk/national-statistics-accreditation

Statistical Quality Assurance Summary

- Assure the data processes
- Validate population estimates
- Invite rigorous peer review
- Retain National Statistics Accreditation

- Get it right first time to provide trustworthy, high quality census data that is of value to the people of Scotland.

Data Processing

Overview

- (Paper) Capture
- Coding
- Data Cleansing
- Admin Data
- Edit & Imputation (again)
- Census Coverage Survey (CCS)
- Estimation & Adjustment
- Statistical Quality Assurance
- Statistical Disclosure Control & Outputs

Census 2021: Statistical Disclosure Control and Outputs methodology

National Records of Scotland

Scotland's Census
Shaping our future

What is statistical disclosure control?

Statistical Disclosure Control (SDC) is required to:

- Prevent the release of any information about an individual, household or enterprise that involves or could lead to:
  - their identification, or
  - the disclosure of confidential information about them

What is statistical disclosure control?

SDC involves

- either:
  - Introducing sufficient ambiguity/uncertainty into, or reducing the level of detail of published statistics so that the risk of disclosing confidential information is reduced to an acceptable level
- and / or:
  - Controlling access to data

What is statistical disclosure control?

Why do we need SDC?

- Legal Census Act 1920
  - Census (Confidentiality) Act 1991
  - Data Protection Act 1998
  - The Census (Scotland) Regulations 2010
- GDPR
- UK Stats Authority Code of Practice

What is statistical disclosure control?

What we did for the 2011 Census

- Targeted record swapping
  - Targeted to “risky” records
- Table redesign
  - Criteria of % 1s and attribute disclosures that are “real”

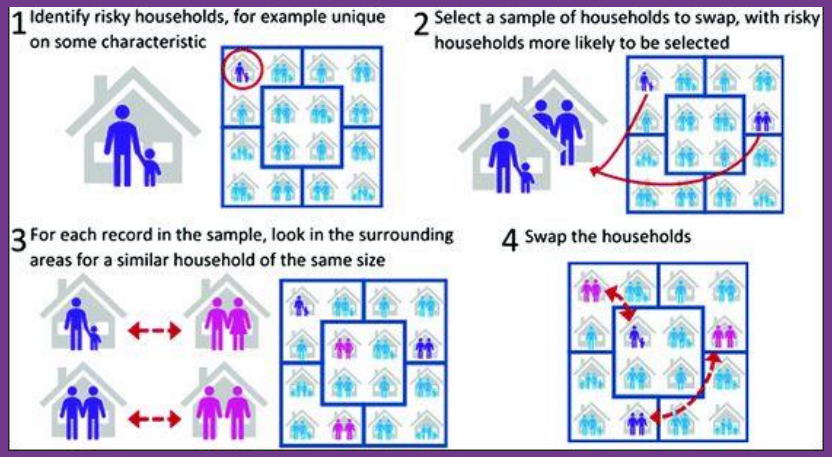
Every table had to be checked for disclosure

Timing was affected

What is statistical disclosure control?

Targeted Record Swapping

- Select a sample of records/households to be swapped
- Using a set of variables, find a match for each sample record/household
- Swap the geographic variables of sample record/household with that of matched record/household



What is statistical disclosure control?

What did we learn from the 2011 Census?

- The process of SDC for the 2011 Census was not as efficient as it could have been:
  - Each table had to be manually checked for disclosure risks, taking up a lot of time an resources
  - Manual manipulation of all tables was required to ensure no disclosure issues occurred
  - At times, this caused lengthy delays in the release of publications and the release of data for customers

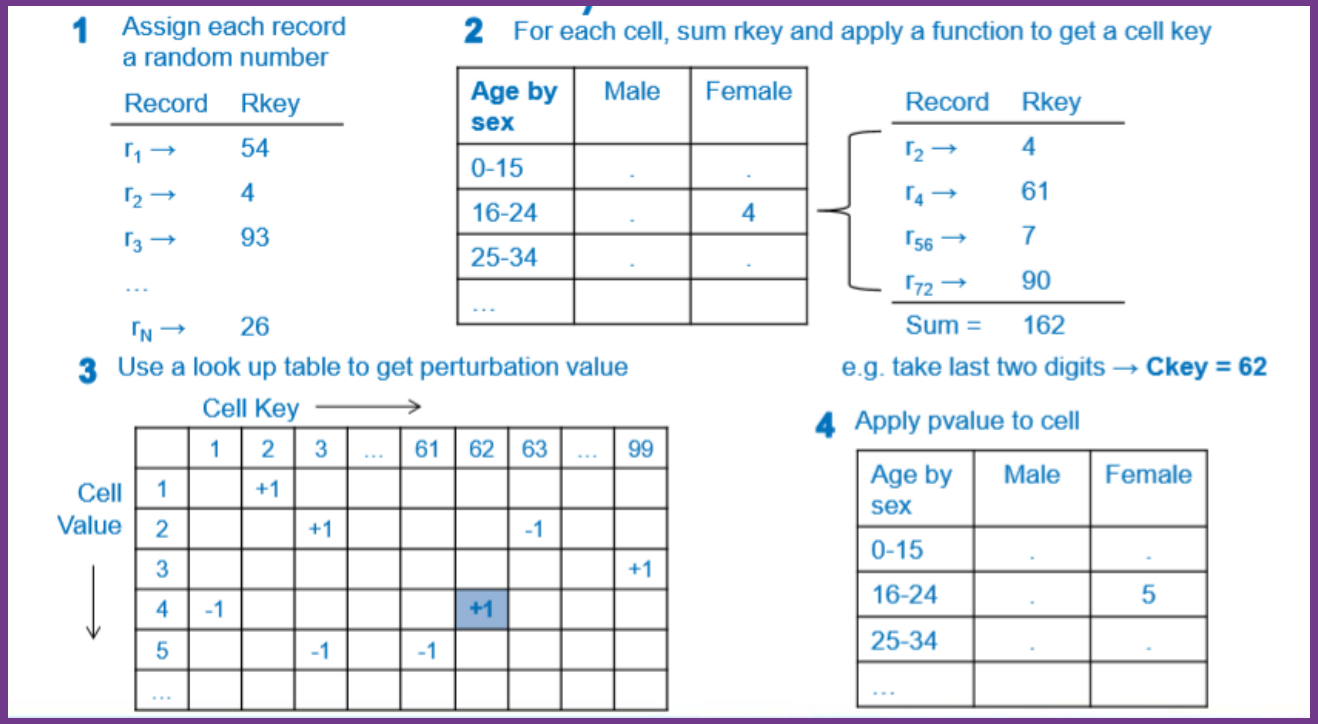
What is statistical disclosure control?

2021 Census

- For the 2021 Census a we are planning to build on what was done in 2011
- Planning to use targeted record swapping
- We are planning on using a method of Cell Key Perturbation
- This should:
  - Reduce the time from gathering the data to when we publish data
  - Reduce the need of manually checking tables which would speed up the publication process
- Aim to release some tables at higher geographies unperturbed

What is statistical disclosure control?

Cell Key Perturbation



What is statistical disclosure control?

Notes for Cell Key Method

- Adapted from “Australian Bureau of Statistics (ABS) method”
- Primarily a protection against ‘differencing’
- Looking at a light touch (record swapping still the primary approach)
- Introduces another layer of uncertainty for intruder
- Consistency in same cell across tables

Outputs

Dissemination

- The website will be the main platform for results dissemination and will be redesigned to make it more user friendly
- Metadata will be incorporated into all outputs
- The number of standard tables will be reduced (change from 2011)
- A flexible table builder to allow users to create their own tables (new for 2021)
- Change in the order outputs are released as a result of the introduction of the flexible table builder and the cell key perturbation.

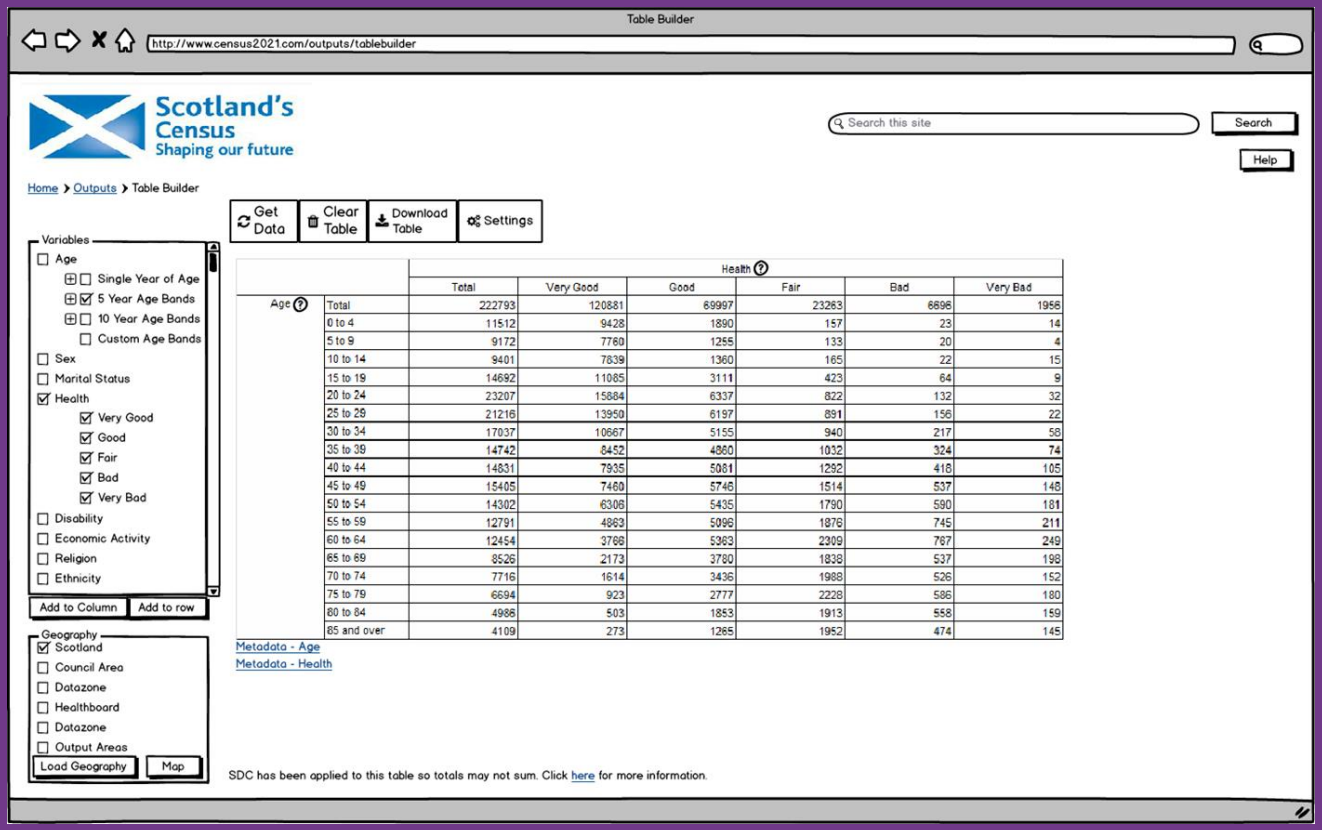
Outputs

Flexible Table Builder

- Allows users to create their own tables
- Reduce the need for standard tables and commissioned requests.
- We will still have a commissioned table service for things that cannot be created using the flexible table builder

Outputs

Flexible Table Builder



Outputs

Standard Tables for 2021

- Aiming to produce less in 2021 as users will be able to create their own tables
- Aim to produce standard tables for most output variables by age and sex
- Small number of other cross tabulations
- Over 400 standard tables in 2011

What is statistical disclosure control?

Outputs schedule

Date	Topic
March 2022	First release: Summary rounded population table by age, sex and Council Area
Autumn 2022	Unrounded and potentially unperturbed population statistics by a range of topics and Council area. These variables also released in the flexible table builder but only for higher level geographies
Winter 2022/2023	Predefined tables by topic, age and sex for unchanged/largely unchanged questions for all standard geographies down to output area. These variables also released in the flexible table builder and available to standard geographies down to output area.
March 2023	Final predefined outputs for new questions by topic, age and sex for all standard geographies down to output area. These variables also released in the flexible table builder and available to standard geographies down to output area.
Autumn 2023	Microdata and origin destination data
2024	Workplace and daytime

Exact timings and content subject to statistical disclosure control and UK harmonisation considerations

Statistical Disclosure Control & Outputs Summary

- A combination of Statistical Disclosure Control methods will be used to protect the confidentiality of 2021 Census respondents
  - Targeted Record Swapping
  - Cell key perturbation
- A flexible table builder tool will be available on the Scotland's Census 2021 website which will allow users to create their own tables
  - Reduced need for standard output tables
  - New SDC methodology designed to speed up the publication process

UK Harmonisation

- Three UK censuses:
  - NRS = Scotland
  - NISRA = Northern Ireland
  - ONS = England and Wales

- Proposed for 21 March 2021
- UK-wide census outputs will be produced through strong collaboration

National Records of Scotland

Scotland's Census
Shaping our future

UK Harmonisation

- Various working groups meet regularly
- Share lessons learned and methodology across offices and internationally
- Align where possible
- All contribute to UK outputs

- Are you a UK data user?
  - Please get in touch to share your needs scotlandscensus@nrscotland.gov.uk

National Records of Scotland

Scotland's Census
Shaping our future

Thank You!

We thank you for taking the time to come to our event today.

Please fill in a feedback form to help us plan future events.

If you have any questions, please email scotlandscensus@nrscotland.gov.uk