Coverage Adjustment Methodology

Census Division
General Register Office for Scotland

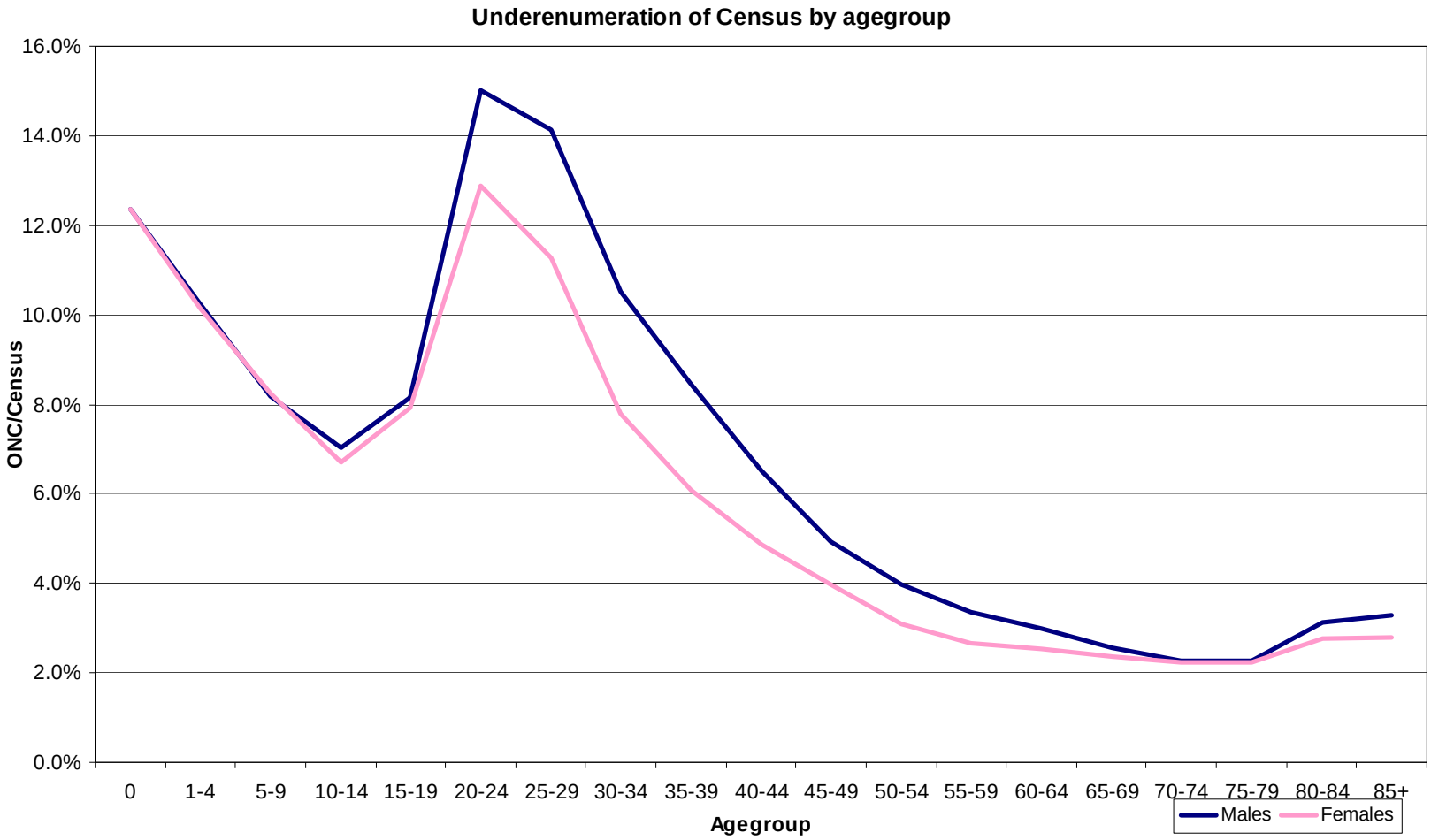
Coverage

- **Some households and persons will be missed by the Census**
- **Need to adjust the census to take account of this**
- **Produce estimates by Local Authority (LA) and age-sex**
- **Why?**
  - **In 2001, ~70,000 households estimated missed**
  - **200,000 persons (4%) estimated missed (mostly, but not all, from missing households)**
  - **this varies by age-sex and geography**

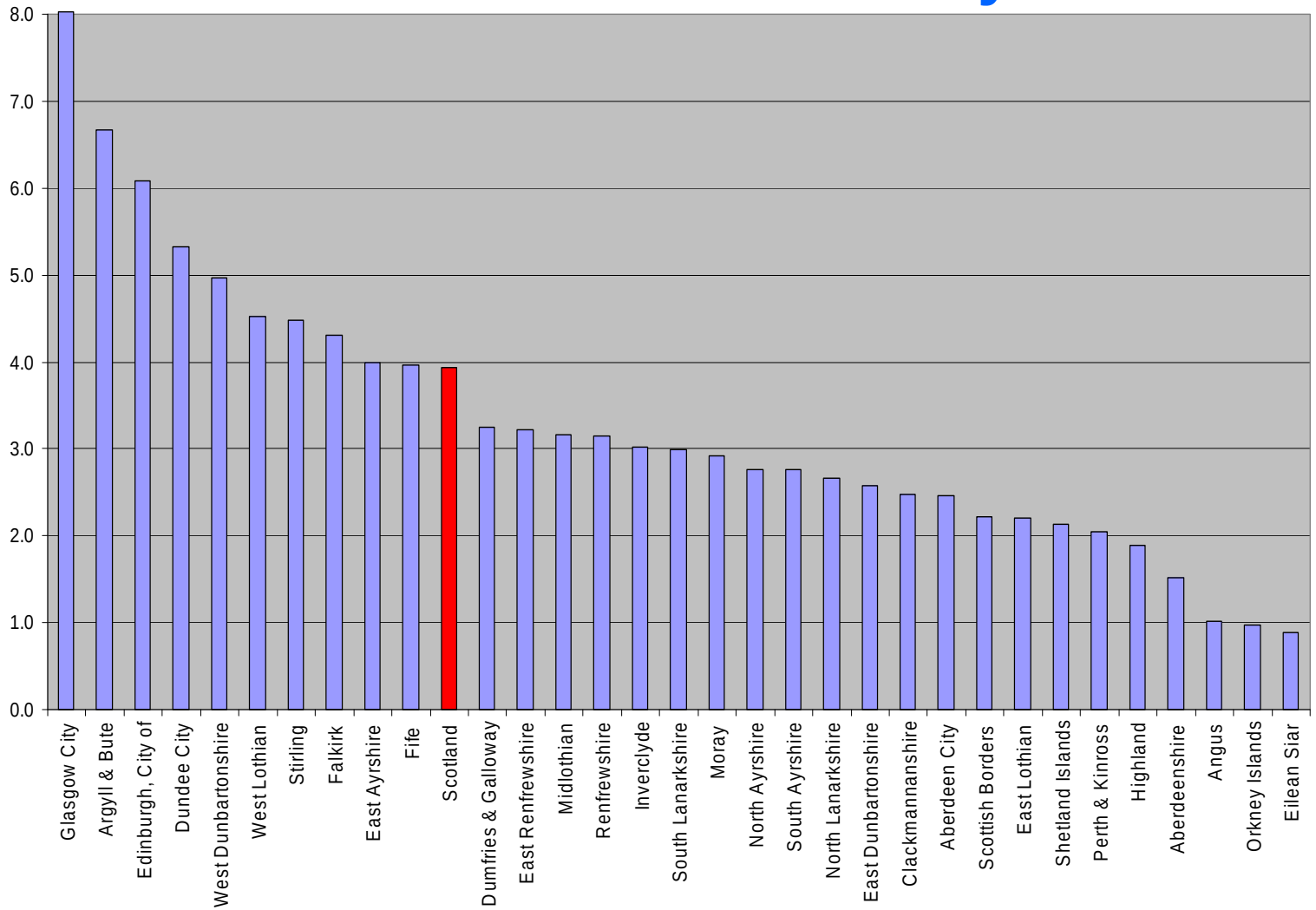
Coverage

- **Coverage assessment:**
  - **Method for estimating what and who is missed**
  - **Based on a Survey**
  - **Uses standard statistical techniques**
  - **Produces estimates of population**
  - **Output database is adjusted by adding households and persons**
- **Quality assurance (not covered here)**
  - **Checking plausibility of estimates and outputs**

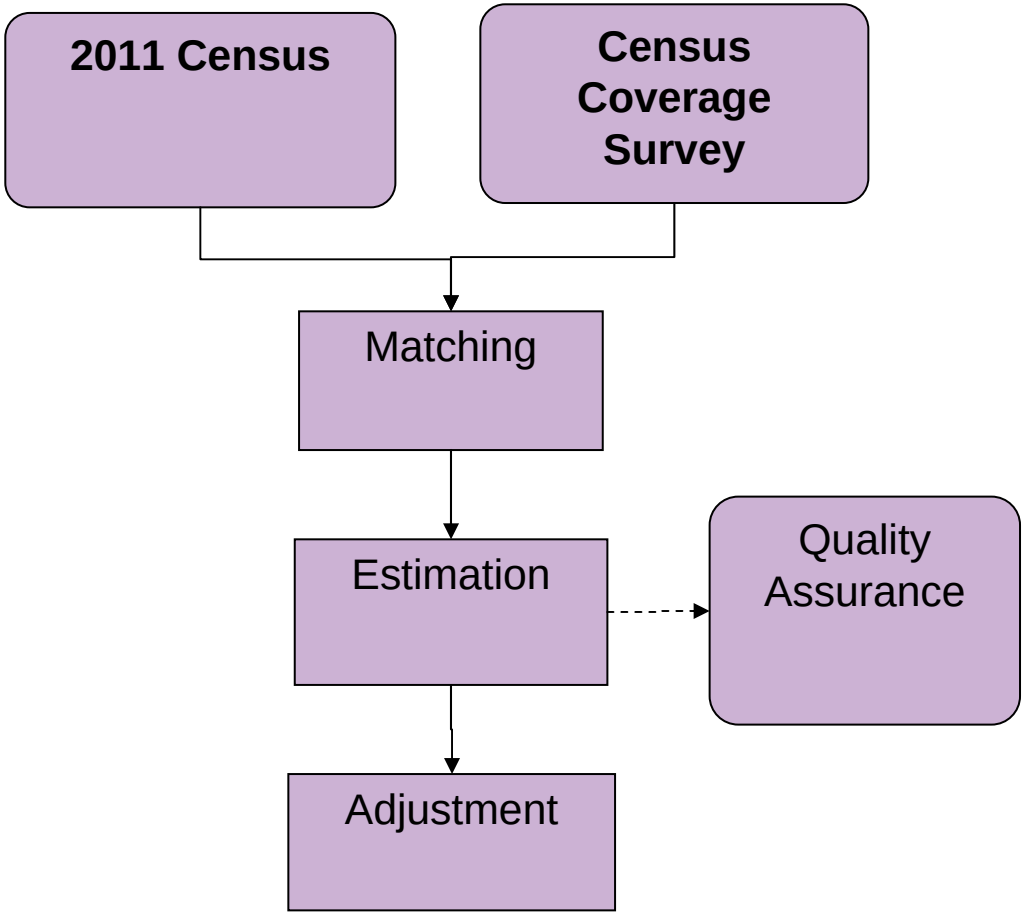
2001 Census Undercount by Age-sex



2001 Census Undercount by Area



Coverage Assessment Process Overview



The Census Coverage Survey (CCS)

- **Key tool for measuring coverage**
- **Features:**
  - Sample of postcodes
    - Measure coverage of households and persons
    - Postcodes cover whole country
  - Large - 40,000 Households
  - 6 weeks after Census Day
    - Fieldwork starting 7th May 2011
  - Voluntary survey

The Census Coverage Survey (CCS)

- **Features:**
  - **Independent of census process**
    - No address listing
    - Operationally independent
  - **Interviewer based**
    - Not self completion
    - Better coverage within households
    - Application of definitions
    - Persuasion/Persistence
  - **Short questionnaire**
    - Variables required to measure coverage
    - Low burden on public

The CCS Sample Design

- **Objective: design survey to be able to estimate LA coverage**

- **Sample selection:**
  - **Divide Scotland into clusters of ~50 households**
    - **Most clusters are a whole Output Area (OA)**
  - **Select sufficient clusters (~800) to achieve sample size**
  - **Sample all postcodes within each selected cluster**
- **How are the clusters selected?**
  - **Grouped by Local Authority**
    - **expect coverage to vary by LA**
  - **Then Hard to count index within each LA**
    - **expect coverage to vary within LA by 'area characteristics'**

The Hard to Count (HtC) Index

- **Designed to predict census coverage**
- **Nationally consistent**
- **Based on model of 2001 response patterns to predict non-response for Datazones**
- **Uses up to date data sources:**
  - Deprivation index, private rented, flats, Higher Education students, schoolchildren with English as second language
- **Split into 40%, 40%, 10%, 8%, 2% distribution**
  - **Easiest lowest 40%, hardest top 2%**
- **Assume OAs/clusters have same HtC in Datazones**
- **Most LAs have about 3 levels**

CCS Sample

- **How big a sample in each LA?**
- **Allocation uses 2001 coverage information**
- **With some minimum and maximum constraints**
  - Min 1 cluster per LA/HtC stratum
  - Max clusters depending on size of LA
- **Drivers of sample size:**
  - Population size
  - Large undercoverage in 2001
  - Variability in 2001 coverage
  - If HtC patterns changed since 2001

Matching

- **Estimation based on dual system estimation**
  - More on this later
- **Requires individual level matching**
  - Both households and persons
  - Identifies those counted by both, those missed by census and those missed by CCS
  - Accuracy is very important
  - Want to minimise ‘missed matches’

Matching

- **Features that permit high quality matching:**
  - Census and CCS designed to allow matching
    - Collect postcode, accommodation type, address, names, dates of birth
    - Data collected on same basis (reference date and definitions)
  - High coverage in both census and CCS (expect to have a match)
  - Good data quality

Matching

- **Mixture of methods – Automatic and clerical**
- **As expect many matches, and data quality high, can reduce clerical effort using probabilistic techniques**
  - Use algorithm to derive ‘probability’ that two records relate to the same entity
  - And then set threshold over which we accept match
- **Remainder have to be viewed by clerical staff**
  - Use a structured workflow in order to ensure a high accuracy rate of matches
  - Sample of matches reviewed at every stage by experts

Automatic Matching

- **Automatic matching an iterative process**
  - It is data driven
  - Might need more than one pass
- **Outcome dependent on a number of key components:**
- **Blocking**
  - reduces number of comparisons (usually postcode)
- **Matching variables**
  - Name, year of birth, month of birth, house number, accommodation type
- **Comparison functions**
  - spelling distance, soundex, token algorithm
  - distance matrices

Clerical Review

- **Takes in the ‘likely’ matches that the automatic system is not allowed to make a decision on (i.e. those under the threshold)**
- **Clerical review of these potential matches**
  - Matcher sees the data
  - And can view images
- **Matches presented in descending score order (household, then individual)**
  - Matcher can defer to a supervisor
- **Supervisor must make a decision for all remaining pairs to complete the resolution**

Examples

- **Exact Match**

Census			CCS		
House number	Surname of HoH	Acccom Type	House number	Surname of HoH	Acccom Type
15	DONEGAN	3	15	DONEGAN	3

Census			CCS		
Person number	Name	DOB	Person number	Name	DOB
1	NICOLA MARY DONEGAN	19121966	1	NICOLA MARY DONEGAN	19121966
2	PHILLIP ANDREW DONEGAN	1111988	2	PHILLIP ANDREW DONEGAN	1111988
3	JACK ANTHONY DONEGAN	18041992	3	JACK ANTHONY DONEGAN	18041992
4	CHLOE MARIE DONEGAN	6011995	4	CHLOE MARIE DONEGAN	6011995

Examples

- **High probability matches**

Census			CCS		
House number	Surname of HoH	Acccom Type	House number	Surname of HoH	Acccom Type
15	DONEGAH	3	15	DONEGAN	3

Census			CCS		
Person number	Name	DOB	Person number	Name	DOB
1	NICOLA MARY DONEGAH	19121966	1	NICOLA DONEGAN	19121966
2	PHILLIP ANDREW DONEGAN	1111988	2	PHILIP DONEGAN	1111988
3	JACK ANTMONY DONEGAN	18041992	3	JACK DONEGAN	18041992
4	CHLOE MARIE DONEGAH	6011995	4	CHLOE DONEGAN	6011995

Examples

- **Low probability matches**

Census			CCS		
House number	Surname of HoH	Acccom Type	House number	Surname of HoH	Acccom Type
15	DONEGAH	4	Sunnyside	DONEGAN	3

Census			CCS		
Person number	Name	DOB	Person number	Name	DOB
1	NICOLA MARY DONEGAH	19121966	1	NICOLA DONEGAN	19121966
			2	PHILIP DONEGAN	1111988
2	JACK ANTMONY DONEGAN	18041992	3	JACK DONEGAN	18041992
3	CHLOE MARIE DONEGAH	missing	4	CHLOE DONEGAN	6011995

Data After Matching

- **We have for the sampled areas (about 800 clusters), household and person data:**
  - **Those seen by both (i.e. matched)**
  - **Those seen ONLY by the census**
  - **Those seen ONLY by the CCS**
  - **The total census count**

Estimation

- 3 parts of the estimation process:
- Dual System Estimation
  - What is the true population in the sampled areas?
- Ratio Estimation
  - How do we estimate for the non-sampled areas?
  - How do we get enough sample to be able to make robust estimates?
- Local Authority Estimation
  - How do we get LA level estimates after getting Estimation Area level estimates?

Dual System Estimation

- Dual System Estimation (DSE)
  - Used mainly for wildlife applications
  - Requires two counts of the population
- Assumptions vital to the DSE
  - Matched data with no matching errors
  - Closed population
  - Independence
  - Homogeneity
  - Non zero probabilities
- Applied at very low level to approximate assumptions
  - ‘cluster’ of postcodes
  - Age-sex group

Dual System Estimation

- **DSE estimates adjustment for those missed in both Census and CCS in each cluster by age-sex group**

		Counted By CCS		
		Yes	**No**	**TOTAL**
Counted	**Yes**	n_{11}	n_{10}	n_{1+}
By Census	**No**	n_{01}	n_{00}	n_{0+}
	TOTAL	n_{+1}	n_{+0}	n_{++}

- **The DSE count for an age-sex group in a cluster is**

$$n_{++} = n_{1+} \times n_{+1} \div n_{11}$$

Dual System Estimation

- **DSE estimates adjustment for those missed in both Census and CCS in each cluster by age-sex group**

		Counted By CCS		
		Yes	**No**	**TOTAL**
Counted	**Yes**	**6**	**3**	**9**
By Census	**No**	**2**	n_{00}	n_{0+}
	TOTAL	**8**	n_{+0}	n_{++}

- **The DSE count for an age-sex group in a cluster is**

$$n_{++} = n_{1+} \times n_{+1} \div n_{11}$$

Dual System Estimation

- **DSE estimates adjustment for those missed in both Census and CCS in each cluster by age-sex group**

		Counted By CCS		
		Yes	**No**	**TOTAL**
Counted	**Yes**	**6**	**3**	**9**
By Census	**No**	**2**	n_{00}	n_{0+}
	TOTAL	**8**	n_{+0}	n_{++}

- **The DSE count for an age-sex group in a cluster is**

$$n_{++} = 8 \times 9 \div 6$$

Dual System Estimation

- **DSE estimates adjustment for those missed in both Census and CCS in each cluster by age-sex group**

		Counted By CCS		
		Yes	**No**	**TOTAL**
Counted	**Yes**	**6**	**3**	**9**
By Census	**No**	**2**	n_{00}	n_{0+}
	TOTAL	**8**	n_{+0}	n_{++}

- **The DSE count for an age-sex group in a cluster is**

$$n_{++} = 8 \times 9 \div 6 = 12$$

Dual System Estimation

- **DSE estimates adjustment for those missed in both Census and CCS in each cluster by age-sex group**

		Counted By CCS		
		Yes	**No**	**TOTAL**
Counted	**Yes**	**6**	**3**	**9**
By Census	**No**	**2**	1	3
	TOTAL	**8**	4	12

- **The DSE count for an age-sex group in a cluster is**

$$n_{++} = 8 \times 9 \div 6 = 12$$

Ratio Estimation

- **DSE gives an estimate of the population within each sampled cluster by age-sex**
- **But not for the non-sampled areas**
- **Need to make an adjustment for the undercount outside of sampled areas**
- **Ratio estimation is used to do this**
  - **a standard technique used in a lot of surveys**
  - **Used when you have data for everywhere that is highly correlated with your survey outcome**

  (e.g. use height to predict weight)
  - **We have a census count that is highly correlated with our DSE**

Ratio Estimation

- **Step 1: Find the relationship between the DSE and census count in our sample**
  - **Expect the relationship to be different by age-sex**
  - **And by the HtC index**

- **Step 2: assume the relationship holds across the non-sampled areas and predict using relationship**

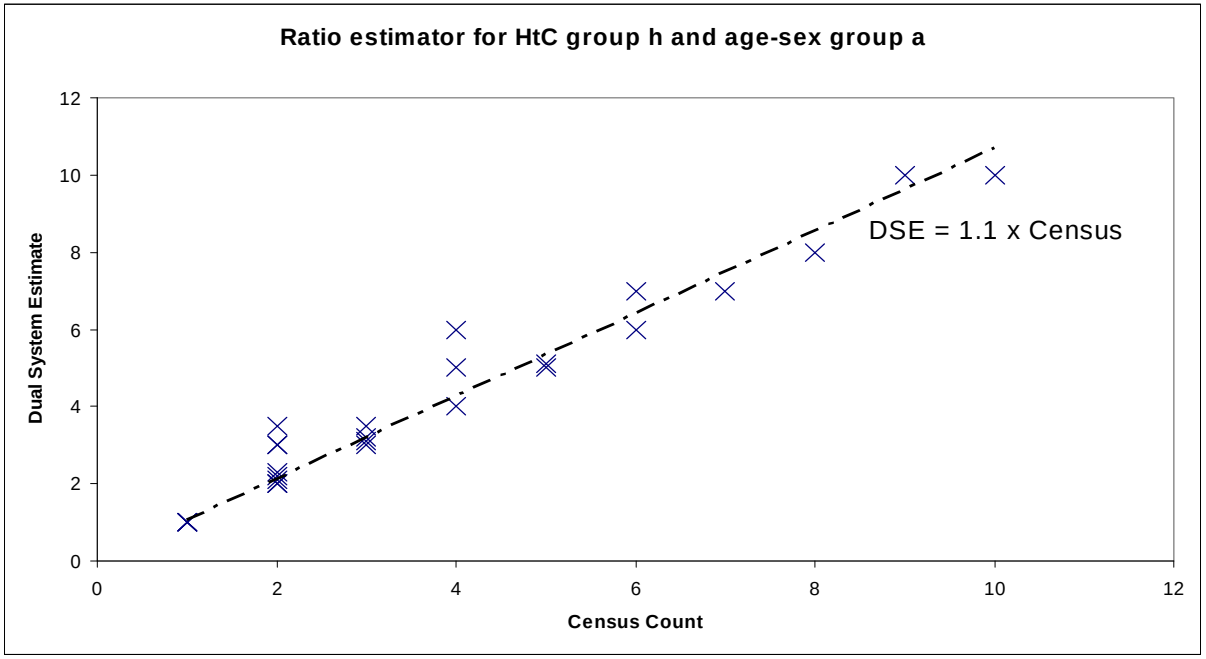
Estimation Areas (EAs)

- **Step 1: Find the relationship between the DSE and census count in our sample**
  - **generally not enough clusters in most LAs by HtC to get a robust measure of the relationship (need about 7 in a LA by HtC)**
  - **Solution is to put LAs into groups called Estimation Areas until have enough clusters – about 70 or more in total**
  - **Glasgow only LA in Scotland with enough sample to be an EA in itself**
  - **EAs are formed from contiguous LAs**
    - **But we reserve option to make changes during processing**

Ratio Estimation

- **Relationship is obtained by ratio between DSE and census count across the clusters**
  - **sum of the DSE divided by sum of the census counts for each postcode cluster**

  (slope of the line of best fit through the origin)
  - **Interpreted as 'coverage weight' or adjustment factor**
  - **Should be greater than 1 (as we are expecting the Census to undercount the "truth")**
  - **Multiply by census count in non-sampled clusters**



x Each point marks the DSE population and the Census count for an age-sex group in a cluster of postcodes within a hard-to-count stratum for an Estimation area.

LA Estimation

- **Ratio estimator gives EA population estimates**
- **How to get to LA totals?**
- **Use ‘synthetic’ estimator**
- **Assumes the relationship at EA level holds across the LAs**
  - **Within HtC and broad age-sex group**
  - **Hence if measure coverage to be 95% for 40-44 yr old males in HtC 2 stratum**
  - **Assume 95% coverage for all 40-44yr old males in HtC 2 in all LAs within the EA**
  - **Essentially applies the adjustment factors from the ratio estimator to the LA census counts**

Estimation - DSE Bias

- We noted a number of assumptions for DSE
  - key ones are independence and homogeneity
- If these are violated, it causes bias in the DSE
  - essentially, the estimates for the cluster are, on average, too low
  - the adjustment factors in the ratio estimator are then too low
- Solution – bring in additional data
  - We adjust the DSEs so that they are consistent with an estimate of the number of households for the cluster

Coverage Adjustment

- Add in the records estimated to have been missed
  - Imputing missed households and the persons in them
  - Imputing persons missed from counted households
- Estimation process gives LA numbers
- For imputation want detailed characteristics
- First step is to get this from modelling CCS data
  - Model persons and households missed by census
- Models include those questions included on CCS
- Only imputing key characteristics (age, sex, alw, ethnic etc)
  - Creating ‘skeleton’ records
  - Non-controlled variables imputed by item imputation process

Coverage Adjustment

- Now that have weights can impute records
  - Should get close to key totals at LA level
  - Impute types of households and persons CCS found were missed
- What about getting it right locally?
  - Key to this is geographical placement
  - Solution: Use identified non-responders on address register (‘Dummy’ questionnaires) or late returns
- We place households into these spaces using a best fit approach
  - E.g. use try to use same accommodation type and ‘copy’ records from nearby